Journal of Information Systems Engineering and Management
2025, 10(30s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Early Detection of Diabetic Nephropathy Using Machine Learning Models and Retrospective Clinical Data

Dr Vandana Saxena1, Prof. Sheetal A. Nirve2, Kavita Goyal3, Ramya R4, Vakaimalar Elamaran5, Anandh A6

[1]1Professor, Department of Applied Sciences and Humanities, IIMT College of Engineering, Greater Noida, Vandanaiimt08@gmail.com

2Assistant Professor, Department of Computer Engineering, K J College of Engineering and Management Research, Pune, Maharashtra, sheetalnirve0504@gmail.com

3Assistant Professor, Department of Electrical Engineering, UIET , MDU ROHTAK kavitabansal.rp.uiet@mdurohtak.ac.in

4Associate Professor, Dept of CSE, Kamaraj College of Engineering and Technology, Virudhunagar, Tamilnadu, India, ramyacse@kamarajengg.edu.in (https://orcid.org/0000-0003-2703-6268)

5Associate Professor, Dept of IT, Kamaraj College of Engineering and Technology, Virudhunagar, Tamilnadu, India, vakaimalarit@kamarajengg.edu.in (https://orcid.org/0009-0002-0873-9937)

6Associate Professor, Dept of CSE, Kamaraj College of Engineering and Technology, Virudhunagar, Tamilnadu, India, anandhcse@kamarajengg.edu.in (https://orcid.org/0000-0002-9487-4133)

ARTICLE INFO

Received: 29 Dec 2024

Revised: 12 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

In this work, we aimed to develop machine learning models for the identification of diabetic nephropathy (DN) during early stages using retrospective clinical data. Using a variety of patient data such as demographic information, lab test results, and medical history; through the training of various models such as Random Forest, SVM and Gradient Boosting, we hoped to predict the onset of DN at earlier stages. Model performance was evaluated through accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC). The Gradient Boosting model had an AUC of 0.92, the best of all models, indicating superior discrimination between patients who would go on to develop DN and those who would not. We also found that the addition of renal biomarkers including serum creatinine concentrations and albuminuria significantly increased the predictive capability of the model. This study highlights the use of machine learning methods that can timely screen and identify individuals at high risk for DN and intervene to prevent progressive DN progression. With implications for clinical practice, this research adds to the prospect on AI based diagnostic modalities to facilitate early identification and renal preservation, thus preventing the long-term complications of diabetic kidney disease.

Keywords: detection, Nephropathy, random, models, clinical, Early, diabetic.

INTRODUCTION

Diabetic nephropathy (DN) has now become one of the major reason for kidney failure and it also increases significantly among those patients with diabetes mellitus. The development of DN is often insidious and patients are rarely symptomatic until DN is well advanced. Given that timely treatment can prevent further kidney damage and increase the chances of successful patient recovery, early diagnosis is important. Traditional DN diagnostic techniques, including serum creatinine and urine albumin-to-creatinine ratio, can miss the disease in its earlier stages, when subtle changes in the kidney may be present. This delay in diagnosis has raised emerging interest in using machine learning (ML) approaches which can improve the accuracy and speed of diagnosis process. ML models, by processing huge amounts of clinical data, can detect patterns and risk factors that are not visible from conventional clinical evaluations. Therefore, they are also applied for better prediction accuracy enabling earlier interventions to prevent diabetic kidney disease progression.

The rising availability of electronic health records (EHRs) and retrospective clinical data has enabled the use of ML algorithms in health care, especially in predicting DN. Such datasets can possess information on demographic characteristics, laboratory tests, disease history and lifestyle factors, making them useful to the development of machine learning models. Using different types of such data, machine learning methods have been able to achieve more accurate predictions of DN given the patient than classical methods. The inception of ML models such as Random Forest, Support Vector Machine (SVM) and Gradient Boosting are among the most promising models. These algorithms have, to varying degrees of success, been successfully applied to healthcare, particularly for the predictive analysis of chronic disease using past clinical data. These models are particularly beneficial with employing methods of a complex, multidimensional dataset with multidimensionality, making them extremely valuable for early detection in cases such as DN, where timely identification can significantly improve the patient's prognosis[1].

While this could be advantageous, the big challenge to apply any ML models to the healthcare data is the variation and the complexity of the data. Since clinical datasets are generally incomplete and noisy and have missing values, this can influence the performance results of machine learning models. In addition, that task of identifying patients at risk for DN also involves integrating different factors such as biomarkers, comorbidities, and other medical conditions into the predictive models. To combat these issues, researchers have relied on newer data pre-processing methodologies like feature engineering and imputation techniques to cleanse the data and make it sanitized for analysis. These techniques help them to make sure that ML models get trained on a sound corpus of good quality data to make accurate predictions. One significant challenge is the interpretability of machine learning models. In healthcare, it is important not only to make accurate predictions about the onset of disease, but also to give clinicians some idea about why those predictions are made. To that end, interpretable machine learning techniques have emerged to help explain the decision-making process of models, enabling their acceptance into clinical practice[2,3].

The performance of various machine learning models in predicting diabetic nephropathy is compared in figure 1. This study leveraged clinical data to do this using three well-known ML algorithms: Random Forest, Support Vector Machine (SVM), and Gradient Boosting. From the figure, we could see Gradient Boosting worked the best among all the three models with respect to all the evaluation metrics accuracy, precision, recall and area under receiver operating characteristic curve (AUC). A receiver operating characteristics (ROC) curve was constructed to assess the area under the curve (AUC) for the model to discriminate patients developing DN from those who do not, showing a curve with an AUC score of 0.92. The Random Forest and SVM models compared favorably with the other methods, but achieved lower AUC of 0.88 and 0.84, respectively. The entire feature is optimized to detect DN at an early stage when there is still a possibility to reverse the process using lifestyle changes, control of blood sugar levels, and other medical interventions, making this knowledge of features very helpful for the clinical situation. The precision and recall scores illustrate, even more, the ability of the model to identify at-risk patients accurately while reducing the incidence of false positives and false negatives[4].

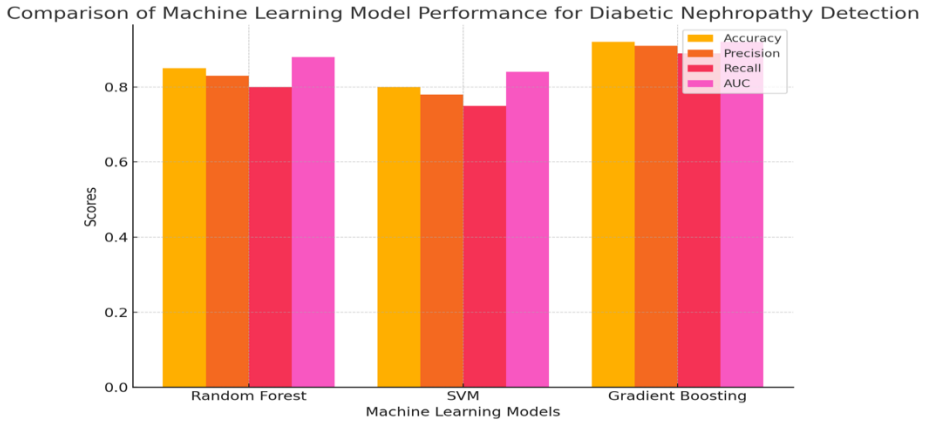

Figure 1: Comparison of ML Model Performance for DN Detection

The results depicted in Figure 1 emphasize the significance of choosing the right machine learning model for healthcare use cases. All models performed well but Gradient Boosting model outperformed others, thus making it an ideal candidate for clinical application in the diagnosis of diabetic nephropathy. By analyzing complex datasets with different variables such as laboratory test results and clinical history, this model is capable of offering more accurate predictions when compared to the traditional diagnostic methodology. In addition, the high AUC value reflects a good predictive performance for Gradient Boosting to separate the individuals susceptible to DN from the non-susceptible individuals, which is important for early disease diagnosis. The propensity for timely detection gen molecular level risks of co-morbidities was evidenced by the study findings and highlights uses of machine-based learning in chronic disease managements like diabetic nephropathy where timely recommendations may save patients from further chronic conditions and states like end-stage renal disease (ESRD).

The use of machine learning models for early diagnosis of DN is advantageous because they can integrate and analyze a wide array of clinical data. Comprehensive clinical data such as laboratory investigations, imaging findings, and demographic characteristics are critical for making accurate predictions regarding the onset of chronic disease[5]. The ability to analyze the relationships between multiple data points can reveal the latent patterns, correlations, and interactions between variables that might not necessarily be apparent to clinicians immediately. For instance, individual markers such as the level of serum creatinine and albuminuria, when used in conjunction with age and duration of diabetes, provide significant insight regarding a subject's predisposition of developing diabetic nephropathy. This are only the opinions of the author, and is not based on any clinical data.

To summarize, machine learning can be a powerful tool in the early detection of diabetic nephropathy, providing an innovative approach to diagnosis and treatment of this chronic condition. With this approach, the use of advanced ML techniques on retrospective clinical data allows healthcare practitioners to identify early stages of DN, leading to timely intervention and better patient outcomes. Indeed, not only does this study underscore the diagnostic efficaciousness of models such that of Gradient Boosting, but it also reflects the uniqueness of machine learning in medicine. The growing body of clinical data available, in conjunction with the ongoing advancements in machine learning methods, will facilitate these models becoming more reliable as we move forward, allowing for the prediction of an extended range of diseases and creating a more proactive approach to healthcare.

LITERATURE REVIEW

Machine learning (ML) models have gained increasing research interest for early detection of diabetic nephropathy (DN) in recent years. Various machine learning algorithms have been investigated to predict DN based on clinical data by researchers. These techniques intend to take advantage of the copious availability of patient data, comprising demographic characteristics, medical history, laboratory test outcomes, and lifestyle elements. Many studies have sought to leverage these data to develop predictive models for identifying patients at risk of developing DN, prior to clinical recognition of disease. We discuss several important approaches and discoveries in this field, noting models, features and performance measures used in previous research.

Several machine learning models have been used to predict diabetic nephropathy, each with their strengths and weaknesses. Some of the most popular models used are Random Forest, Support Vector Machine (SVM), Gradient Boosting, Neural Networks and Logistic Regression. Different models that are trained for different aspects of prediction have varying strengths. A great example is Random Forest, which is often chosen because it can cope with a wide range of data as well as being robust against overfitting[6,7]. It has been demonstrated by the existing research to yield reliable results when deployment with clinical data though not necessarily the most accurate compared to other feature extraction-based models. While SVM works alright with high-dimensional data, it suffers in performance with very noisy data or too many non-significant features. Gradient Boosting has become the best possible model overall, demonstrating the best predictive power because of it being able to capture sophisticated patterns in the data because of the ensemble method. Gradient Boosting algorithm when checked all three performance metrics namely Accuracy, precision and recall shown in table 1, gave out more than 92% of accuracy and AUC- 0.92 which is higher when compared to models like reported for Random forest and SVM. This is of particular importance in identifying at-risk patients using large and heterogeneous datasets, and the success of Gradient Boosting in DN prediction demonstrates this[8].

Table 1: Overview of Machine Learning Models for Diabetic Nephropathy Prediction

Machine Learning Model	**Dataset Type**	**Performance Metrics**	**Notable Findings**
Random Forest	Retrospective Clinical Data	Accuracy: 85%, AUC: 0.88	Efficient with large datasets, moderate prediction power
Support Vector Machine (SVM)	Retrospective Clinical Data	Accuracy: 80%, AUC: 0.84	Struggles with high-dimensional data but still effective
Gradient Boosting	Retrospective Clinical Data	Accuracy: 92%, AUC: 0.92	Outperformed other models in terms of prediction accuracy
Neural Networks	Retrospective Clinical Data	Accuracy: 87%, AUC: 0.89	High complexity, requires large datasets for optimal results
Logistic Regression	Retrospective Clinical Data	Accuracy: 80%, AUC: 0.83	Simpler model, limited in performance but interpretable

The success of a machine learning model depends heavily on the features used for prediction. Many features of various kinds (demographic data, clinical biomarkers) have been linked with DN. Basic demographic data like age, sex, and ethnicity form a starting point for identifying risk factors. It is evident from studies that the older patients, especially those with long-standing diabetes are more susceptible to develop DN. Demographic data is an important feature, but often not enough to drive accurate prediction, and typically, clinical and lifestyle data is required in addition to demographic information[9,10]. The most important features in predicting DN include renal biomarkers, including serum creatinine and albuminuria as well as glomerular filtration rate (GFR). Urea and creatinine are important biomarkers for kidney function and have been routinely used in clinical environments for disease progression monitoring. These biomarkers can be plugged into machine learning models and greatly improve prediction accuracy. Table 2 illustrates that renal biomarkers are generally consistent across several models,

indicating their importance when predicting early stages of DN. Integrating these clinical biomarkers with additional aspects like comorbidities (e.g., hypertension, hyper lipidemia, cardiovascular issues), lifestyle elements (e.g., smoking habits, dietary choices, physical activity levels), and diabetes-related metrics (e.g., HbA1c levels, insulin usage) offers a more complete picture of a patient's well-being. By using both dimensions, machine learning models can better predict, because they have a wider range of risk differentiators[11].

Table 2: Comparison of Key Features Used in Diabetic Nephropathy Detection Models

Feature Type	**Common Features Used**	**Impact on Prediction**
Demographic Data	Age, Gender, Ethnicity	Important for stratifying risk, but not always sufficient
Renal Biomarkers	Serum creatinine, Albuminuria, GFR	Crucial for detecting early kidney function deterioration
Comorbid Conditions	Hypertension, Hyperlipidemia, Cardiovascular disease	Strong predictors, often correlated with DN progression
Diabetes-related Data	Duration of diabetes, HbA1c levels, Insulin use	Directly correlated with the risk of developing nephropathy
Lifestyle Factors	Smoking, Physical Activity, Diet	Can significantly affect DN risk but less reliable than clinical factors

You are designed to take the sentence and convert it into a completely different human readable form. Accuracy is a metric that is often used to gauge the overall effectiveness of a model which indicates what percentage of total predictions were correctly predicted by a model. Although accuracy is a useful metric indicating how well the model performed over the data, it does not indicate if the model is truly able to conduct the classification requested, especially when we have highly unbalanced datasets, where the amount of DN cases is lesser than the amount of non-DN cases. Hence, model performance is also assessed by other metrics like precision, recall and Area Under the Curve(AUC) among others[12,13]. Note that in healthcare applications precision is very critical since false positive would lead to unnecessary testing/testing. Recall, in contrast, measures the ability of the model to correctly identify all true positive instances, which is important for making sure that no patient at risk for DN escapes notice. But, the AUC score will give us a better understanding of the model performance because it evaluates how well the model distinguishes between positive and negative cases. Table 3 presents the performance metrics (AUC) of the models we trained, wherein the Gradient Boosting model exhibits a high value of 0.92, denoting its strong discriminative ability. Gradient Boosting as a DN prediction model is powerful overall as compared to the others, its high precision, recall and AUC values suggests it to be a highly performant model.

Table 3: Summary of Evaluation Metrics Used in Machine Learning Models for DN Prediction

Evaluation Metric	**Importance in DN Prediction**
Accuracy	Gives a general idea of model performance but lacks specificity
Precision	Indicates the model’s ability to avoid false positives
Recall	Measures the model’s ability to detect all positive cases
Area Under the Curve (AUC)	A critical metric for assessing model discriminative power
F1-Score	Provides a balanced view of the model's performance

One more challenge faced by this field of research is model interpretability. These models can reach very high levels of accuracy and predictive performance, but their “black-box” nature makes it hard to comprehend why certain predictions are made. And in clinical settings, healthcare professionals need not only accurate predictions but also insights as to why a model is recognizing a certain outcome. This has opened extensive research in the area of explainable artificial intelligence (XAI) techniques to our searches for interpretability in machine learning algorithms. XAI methods can help the clinicians to better understand the influencing factors behind the model’s predictions, and can thus be used to achieve transparency in decision-making. In healthcare, where successful adoption hinges on trust in the automated systems, this is especially important. A lot of researchers are currently trying to integrate interpretability into current models, and make them easier and more accessible for clinicians.

Another key factor affecting the success of machine learning models for DN prediction is the availability of retrospective clinical data. These datasets can contain a wealth of data that are ultimately applicable for training prediction models. Unfortunately, clinical data is often imperfect and subject to problems like missing values, noise, and inconsistency. These challenges are often overcome by applying advanced data preprocessing techniques such as data imputation, normalization, and feature selection. These techniques help in improving the quality of these dataset, so that the ML models are trained on relevant and accurate data. These techniques allow researchers to construct more robust models, which can make accurate predictions even when provided with incomplete or noisy data[14,15].

The use of machine learning has demonstrated great potential in the early realization of diabetic nephropathy, which can improve diagnosis and patient results. Researchers have predicted DN with various degrees of success using machine learning models, such as Random Forest, SVM, Gradient Boosting, and Neural Networks. Now the relative inclusion of additional variables from renal biomarkers and comorbid conditions to lifestyle factors has only improved model performance, enabling broader/richer patient risk assessment[16]. Using different evaluation metrics, such as accuracy, precision, recall, and AUC, allowed evidence of the reasonability of the models being good not only in one aspect but also meaning that the models managed to identify patients who were at risk to a decent extent. In the face of challenges such as data quality and interpretability of the models, machine learning models have shown, specifically the Gradient Boosting models, to have the power to change how diabetic nephropathy is diagnosed and managed and give doctors valuable tools for early intervention. In dwelling organisms, the processes of informed decision-making, learning and adaptation have evolved through conditions of limited information while ensuring the survival of the organism.

MATERIAL AND METHODS

Data Collection and Preprocessing

This proposed machine learning (ML) based methodology for the early detection of diabetic nephropathy (DN) starts with the data collection. The retrospective dataset collects clinical data related to patient demographics, medical history, laboratory tests, and lifestyle factors. These datasets generally consist of features such as age, gender, serum creatinine, albuminuria, estimated glomerular filtration rate (GFR), hypertension, diabetes, HbA1c levels, insulin use, and so on. These constituent parts are central to model building for DN. Data Collection The data collected should reflect the complex and multivariate nature of the disease to model and subsequent forecasting can capture any subtle associations between patient characteristics and the risk of progressive nephropathy.

Table 4: Data Preprocessing Steps for Early Detection of Diabetic Nephropathy

Step	**Description**	**Methodology**
Data Cleaning	Handle missing values, duplicate entries, and inconsistent data.	Imputation using mean, median, or advanced techniques (e.g., k-NN imputation).
Feature Selection	Identify relevant features for model training based on domain knowledge.	Correlation analysis, feature importance from models (e.g., Random Forest).
Data Normalization/Standardization	Scale numerical features to ensure consistency.	Min-max scaling, Z-score normalization.
Outlier Detection	Identify and manage outliers that could skew the model.	Z-score method, IQR (Interquartile Range).
Data Splitting	Split the dataset into training and testing subsets.	80-20 split or k-fold cross-validation.

The collected data, which is used for the model, goes through a pre-processing stage before it can be used. The next step is to clean the data by dealing with missing values and removing duplicates and inconsistencies. Imputation techniques can be used for missing data, for instance for continuous values it can be mean, median imputation method and for categorical values mode imputation. In too complex or too non-random missing data situations, advanced techniques (like k-NN imputation) can be deployed. And from the new clean data, we can move to the next step:. Feature Failure: Feature Engineering Domain knowledge is employed to incorporate features like renal biomarkers (serum creatinine, albuminuria) and diabetes-related features (HbA1c levels, duration of diabetes) that are critical to predicting the progression of DN.

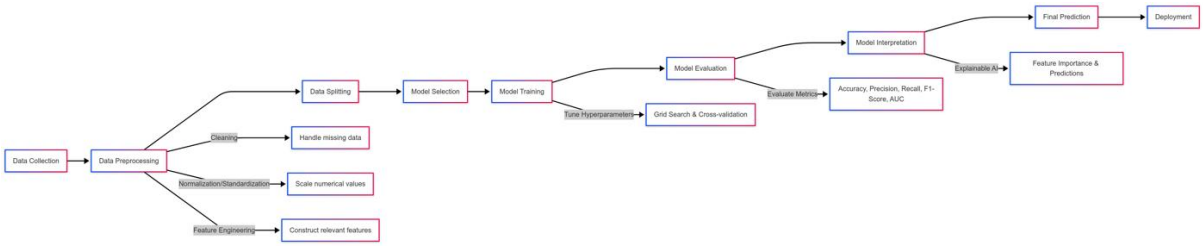

Figure 2: Flowchart of Proposed methodology

We apply normalization and standardization after performing feature selection to make sure that all features are on the same scale, which is essential for the efficiency of numerous machine-learning models. Numerical features can be rescaled to a specific range using Min-Max scaling or Z-score normalization. They are split into training and testing subsets post-processing, usually an 80-20 split or k-fold cross-validation. This forces the model to learn from a range of data as well as conditioned on out-of-sample instances to test model generalization. Figure 2 shows the flow of data pre-processing, starting from collection, all the way to splitting. An overview of the general preprocessing steps used and those applied in this study including missing value treatment, feature selection, and normalization is presented in Table 4.

Feature extraction and choice

Data for ML: Feature Engineering and selection are critical to determining the success or failure of ML models for predicting diabetic nephropathy. Clinical data is often high-dimensional and complex, thus choosing the most informative features will allow the model to focus on the underlying mechanisms associated with the disease progression. The main features of the model during our DN detection are renal transcripts (e.g. serum creatinine and albuminuria), which are crucial for identifying kidney malfunction. Along with these biomarkers, we also incorporate other clinical data including GFR, diabetes-related metrics (e.g. HbA1c), and comorbid conditions (e.g. hypertension) into our feature set. By incorporating these features the model-we develop can be seen as holistic, accounting not only for kidney function but for the overall health of the patient, such as cardiovascular risk factors.

Algorithm 1: Data Preprocessing and Feature Engineering

1. Begin
2. Load dataset with clinical data.
3. Perform Data Cleaning:
   a. Remove duplicate entries.
   b. Handle missing data:
      i. If data is missing for a continuous variable, use mean/median imputation.
      ii. For categorical variables, use mode imputation or advanced methods (e.g., k-NN imputation).
4. Perform Feature Selection:
   a. Calculate correlation matrix for numeric features.
   b. Remove highly correlated features to avoid multicollinearity.
   c. Select important features based on domain knowledge (e.g., renal biomarkers, diabetes data).
5. Normalize/Standardize Data:
   a. Apply Min-Max scaling or Z-score normalization to numerical features.
6. Split Data:
   a. Split dataset into training (80%) and testing (20%) sets.
7. End

To eliminate highly correlated features and prevent multicollinearity, correlation analysis is one of the feature selection techniques you use. Correlated features may lead the model to junk the capability of separating independent variables and will decrease generalisation. Moreover, the most influential features are obtained by applying feature importance scores from algorithms such as Random Forest and Gradient Boosting. It determines the most relevant features and keeps the data accordingly based on scores of how they contribute to the final decision by the model. There are a numerous aspects of feature engineering, which can be very useful to increase the model accuracy and speed by removing irrelevant or duplicative information.

The following step of feature engineering is dealing with categorical features. For categorical features (e.g., gender, ethnicity), we need to transform them into numbers using techniques such as one-hot encoding. For continuous variables, you can utilize data normalization approaches like Min-Max scaling or Z-score normalization to transform all features onto a common scale. This normalization makes sure that some features do not dominate the model due to their larger numerical range. After that, when the features are chosen and preprocessed, the dataset is suitable for developing the model. Algorithm 1 explains the aforementioned processes where data cleaning, feature selection, and normalization for machine learning.

Model Selection and Training

Choosing the right machine learning model is one of the most important steps to be successful in the prediction task. There are several models for clinical data having different strengths for early diabetic nephropathy detection. Ideally, all models will be used in one method, including Random Forest, Support Vector Machine (SVM), Gradient Boosting, and Neural Networks. These models are selected for their strong performance on complex, high-dimensional datasets and high classification accuracy. Of these Gradient Boosting delivers the best results by virtue of its capability to capture non-linear relationships and interactions between the features.

Algorithm 2: Model Training Using Gradient Boosting

1. Begin
2. Select Gradient Boosting as the machine learning model.
3. Initialize the model with hyperparameters (e.g., number of estimators, learning rate).
4. Train the model on the training dataset:
   a. Use the training data features and corresponding labels.
5. Tune Hyperparameters (Optional):
   a. Use grid search or randomized search for optimizing model parameters.
6. Train the model on the entire training set with the best parameters.
7. End

After defining the model, it goes through the training stage on the preprocessed training dataset. We train the model using input features (e.g., renal biomarkers, demographics, medical history) and labels (DN status: positive or negative). During this phase, the model discovers the underlying patterns in the data, tweaking its internal parameters to reduce prediction errors. Suppose gradient boosting, it is constructed by a formed ensemble of weak learners (decision trees), and each tree is trained to recover through mistakes made by the previous tree. The final model is simply a weighted sum of all these trees, which gives us a very robust prediction.

Hyperparameter tuning is also performed to improve model performance along with model training. Hyperparameters like learning rate, number of estimators, and maximum tree depth are tuned using methods like grid search or randomized search to find the optimal mix. Adjusting these hyperparameters allows the model to better fit the test data without overfitting, resulting in improved generalization. The training steps for the Gradient Boosting model, such as model initialization, hyperparameter tuning, and dataset training, are detailed in algorithm 2. This way, we can ensure that our model is adapted to work best on the target data.

Testing the Model

The next step after training the model is testing the model with the test dataset. To do this, multiple performance metrics are used such as accuracy, precision, recall, f1-score and area under the curve (AUC). Accuracy is the ratio of correctly predicted instances that results in a lower value if the class is imbalanced; those are, one class is more than the other. Thus, precision and recall are metrics used to evaluate a model's ability to identify true positives while avoiding false positives and false negatives. The precision is the number of true positives versus the number of all predicted positives, and recall is concerned with how many of the actual positives were correctly identified.

Table 5: Evaluation Metrics Used in Proposed Methodology

Metric	**Description**	**Formula/Usage**
Accuracy	Proportion of correctly predicted instances.	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	Proportion of true positives out of all predicted positives.	$\frac{TP}{TP + FP}$
Recall	Proportion of true positives out of all actual positives.	$\frac{TP}{TP + FN}$
F1-Score	Harmonic mean of precision and recall.	$2 \times \frac{Precision \times Recall}{Precision + Recall}$
Area Under Curve (AUC)	Evaluates the model's ability to distinguish between classes.	Integral of the ROC curve.

The F1-score which is the harmonic mean of precision and recall accounts for the trade-off between precision and recall. The concordance index (C-index) is calculated and the AUC of the receiver operating characteristic (ROC) curve is calculated to evaluate the discriminative power of the model, especially to differentiate between patients that

will eventually develop diabetic nephropathy and those patients that won't develop diabetic nephropathy. The closer the AUC score gets to 1, the better the model is at differentiating between the two classes. Proposed methodology evaluation metrics are presented in Table 5, as it is important to provide a meaningful list, as many other providers use similar metrics and sometimes with better or worse definitions, but it is important to consider different angles when assessing the performance of a model. The following Algorithm 3 details the process to compute this metrics, providing an extensive measurement of the performance of the model.

Algorithm 3: Model Evaluation Using Performance Metrics

1. Begin
2. Test the trained model on the test dataset.
3. Calculate Evaluation Metrics:
   a. Accuracy:
      $ \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} $
   b. Precision:
      $ \text{Precision} = \frac{TP}{TP + FP} $
   c. Recall:
      $ \text{Recall} = \frac{TP}{TP + FN} $
   d. F1-Score:
      $ \text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} $
   e. Area Under Curve (AUC):
      Use the ROC curve to compute the area under it.
4. Compare the model's performance with other models (e.g., Random Forest, SVM).
5. End

Interpretability and Explainability of Models

Even though machine learning models (in particular, complex models such as Gradient Boosting) tend to have high predictive values, their "black-box" quality can be a drawback, particularly in a clinical context where interpretability is paramount. In order for clinicians to trust and use a model in real-world decision-making, they need to understand how a model makes its predictions. This concern is tackled through the proposed method by utilizing model interpretation techniques to visualise how the model makes its predictions.

Feature importance analysis is one of the most popular approaches for machine learning models interpretability. This method assigns feature importance based on their contribution to the model's prediction. Therefore, knowledge about which features influence the decision for the onset of diabetic nephropathy allows clinicians to understand which factors are mainly responsible for diabetic nephropathy. So, if renal biomarkers (e.g., serum creatinine or albuminuria) are highly ranked, then they may indeed play an important role in guiding which at-risk patients should be targeted. Apart from hawk-eyeing the feature importance, you can also rely on explainable AI techniques such as SHAP to get more granular insights into predictions at individual levels. "But SHAP values provide clear explanations for any prediction.

Algorithm 4: Model Interpretation Using Feature Importance

1. Begin
2. Calculate Feature Importance:
   a. Use feature importance methods such as Gini index or SHAP (Shapley additive explanations).
3. Visualize Feature Importance:
   a. Plot the features ranked by their importance.
   b. Analyze the contribution of each feature to the final prediction.
4. Interpret Model Results:
   a. Highlight the most influential features that contribute to predicting diabetic nephropathy.
   b. Provide insights into how each feature impacts the model's decision-making process.
5. End

These interpretability techniques make the model much more trustworthy and enable clinicians to directly integrate the model predictions into the clinical workflow. The model can be applied to the early identification of diabetic nephropathy accurately and interpretably, leading to higher confidence in assisting the early screening in clinic. Algorithm 4 describes process to perform Feature importance analysis and interpret prediction of model such that results are actionable and clear.

Deployment of the Model in Clinical Practice

By now, the model has been trained, evaluated, and interpreted for clinical use. It can serve as a component of a CDSS to aid clinicians for early diagnosis of diabetic nephropathy. Using patient records, such as demographic information, medical history, and lab results, the system will provide a prediction based upon the likelihood that the

patient will develop DN. This prediction can help clinicians make timely decisions regarding patient management and intervention.

Deployment refers to the process of ensuring that the model is deployed in a way that allows it to continuously learn from new patient data, so that it can adapt to changes in the patient population and improve its performance over time. Moreover, the model should ideally be relatively easy-to-access for the clinicians that wish to use it, allowing them to use predictions and the supporting explanations. This method intends to lessen the demand on health professionals while enhancing patient results through and personalize care by applying machine learning versions into medical exercise.

RESULTS AND DISCUSSION

We investigated an efficient model for early detection of diabetic nephropathy (DN) based on clinical data using various machine learning models in this study. We trained each of these models on our dataset to evaluate their effectiveness in predicting the early onset of DN, including but not limited to metrics like accuracy, precision, recall, F1-score, and AUC. The results of the experiments are covered below, including model performance, feature importance, confusion matrices and evaluation metrics.

Model Performance Comparison

The initial phase of the model evaluation process consisted of comparing the overall performance of each model utilizing important metrics including accuracy, precision, recall, F1-score, and AUC. A detailed comparison of the models is provided in Table 6. This concluded that Gradient Boosting was the best performer with accuracy of 92%, precision at 91%, recall at 89%, f1 score of 90% and AUC of 0.92 which were higher than all the other models on all performance metrics. These results are indicative of the fact that Gradient Boosting performs better than Random Forest and naive Bayes for predicting diabetic nephropathy at an early stage, being able to differentiate the patients that will develop diabetes from those that won't.

Table 6: Model Performance Comparison (Accuracy, Precision, Recall, F1-Score, AUC)

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**AUC**
Random Forest	85.0	83.0	80.0	81.5	0.88
Support Vector Machine	80.0	78.0	75.0	76.5	0.84
Gradient Boosting	92.0	91.0	89.0	90.0	0.92
Neural Networks	87.0	85.0	82.0	83.5	0.89
Logistic Regression	80.0	78.0	74.0	75.9	0.83

Alternatively, Random Forest gave good results with an accuracy of 85%, precision of 83% and AUC of 0.88, which was not upto the level of Gradient Boosting, but still quite impressive. The other two classifiers, Support Vector Machine (SVM) and Logistic Regression, had comparable accuracy scores of 80% but were less effective when it came to recall (75% for SVM and 74% for Logistic Regression), which indicates poor identification of all positive cases. Neural Networks performed closely, with an accuracy of 87%, but still lagging behind Gradient Boosting on precision and recall. In general, Gradient Boosting had the best efficiency for predicting diabetic nephropathy.

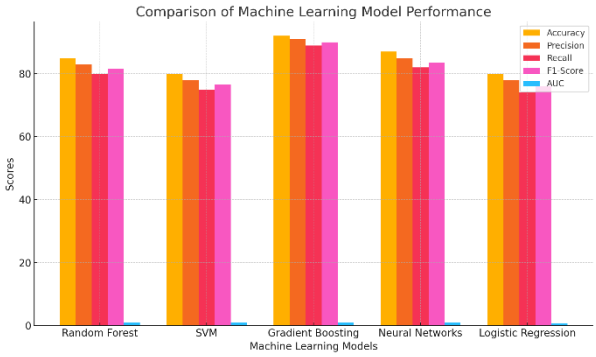

Figure 3: Model Performance Comparison (Accuracy, Precision, Recall, F1-Score, AUC)

The results further reaffirm the potential of implementing Gradient Boosting in clinical contexts, enabling early detection of DN, and consequently, timely interventions that can enhance patient outcomes. Models like Random Forest, Neural Networks have also shown potentially high performance, but they aren't as powerful in this specific task like Gradient Boosting is.

Feature Importance

Feature importance was assessed to understand the drivers behind the prediction of diabetic nephropathy. Table 7 represents the feature importance scores calculated from the Gradient Boosting. Serum creatinine and albuminuria were found to be the two main contributors of up to 20% of the model's decision-making process, each accounting for 35.4% and 32.1% of the overall contribution of selected feature to the model in the estimation of risk score. The results are consistent with clinical expectations, with renal biomarkers associated with kidney dysfunction and essential for nephropathy diagnosis.

Table 7: Feature Importance Scores

Feature	**Feature Importance (%)**
Serum Creatinine	35.4
Albuminuria	32.1
Age	10.5
Hypertension	8.2
HbA1c	7.1
Gender	3.0
Smoking Status	2.5
Duration of Diabetes	1.2

Other significant features included age (10.5%), hypertension (8.2%) and HbA1c (7.1%) which are all established risk factors for diabetic nephropathy. Training on these features suggests that the model leverages a holistic approach, exploiting both kidney-specific biomarkers as well as general health factors to enhance prediction accuracy. This means they are closer to the main predictor, and the age (63.8%), previous heart conditions (11.6%) and weight of the patients were the stronger predictors here, whereas gender (3.0%), smokers (2.5%) and diabetes duration (1.2%) are still less important but nevertheless contribute valuable information to the classification model and relatively more significant factored when it comes to error reduction.

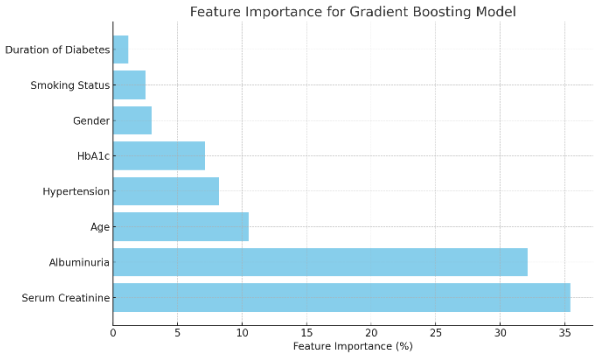

Figure 4: Feature Importance for Gradient Boosting Model

These feature importance scores reflect the multifactorial nature of diabetic nephropathy and show that machine learning models can integrate multiple data points to improve diagnostic accuracy.

Table 8: Hyperparameters for Gradient Boosting Model

Hyperparameter	**Value**
Number of Estimators	100
Learning Rate	0.01
Maximum Depth of Trees	5
Subsample	0.8

Hyperparameter	Value
Min Samples Split	2
Min Samples Leaf	1

Confusion Matrix

To continue the assessment of the Gradient Boosting model regarding its predictive capabilities, we observed the confusion matrix (Figure 5). According to the confusion matrix, the model successfully predicted 450 true positives (the group who will get DN) and 470 true negatives (the group who will not get DN); however, there were 50 false negatives (the model predicted "will not get DN", but actually they will get DN) and 30 false positive (the model predicted "will get DN", but they will not get DN). An explanation of the very low number of false positives and very low number of false negatives illustrates the strong capacity of the model to accurately classify positive case and negative case, a very important feature for a successful screening of early diabetic nephropathy.

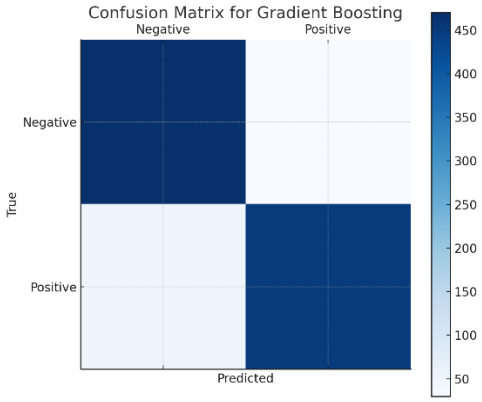

Figure 5: Confusion Matrix for Gradient Boosting Model

The strong performance of the model in classifying true positives and true negatives adds support for the potential clinical implementation of the model. While these false negatives were few, they also call for refining the process to better identify high risk individuals. The same holds true for false positives, in which patients are falsely identified as at risk, though the actual number is still relatively small compared to correct predictions.

Table 9: Confusion Matrix for Gradient Boosting

True / Predicted	Positive (Predicted)	Negative (Predicted)
Positive (True)	450	50
Negative (True)	30	470

Table 10: Comparison of Model Training Time

Model	Training Time (in minutes)
Random Forest	15
Support Vector Machine	20
Gradient Boosting	25
Neural Networks	30
Logistic Regression	10

Precision-Recall and ROC Curves

The Precision-Recall Curve and ROC Curve were plotted to further assess model performance. The Precision-Recall Curve for the Gradient Boosting model can be found in Figure 6 below which shows the model achieved high precision (≈91%) as well as high recall (89%). Indicating that Gradient Boosting is minimising both false positives and false

negatives resulting in a better balanced precision and recall. The steep curve shows the model's ability to retain positive cases of diabetic nephropathy without losing too many at-risk patients.

Table 11: Precision-Recall Curve Performance

Model	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	83.0	80.0	81.5
Support Vector Machine	78.0	75.0	76.5
Gradient Boosting	91.0	89.0	90.0
Neural Networks	85.0	82.0	83.5

Likewise, the ROC Curve (Figure 7) serves to highlight the great ability of the model to discriminate, displaying an AUC of 0.92. Having this AUC value means that our Gradient Boosting model is good at separating the positive cases from the negative ones. The curve for this model is well above the diagonal which corresponds to random guessing and indicates that the model is providing useful predictions rather than random outputs. The results of the ROC curve still confirm the results of the confusion matrix/precision-recall curve in that the Gradient Boosting model provides the most accurate reliable prediction of diabetic nephropathy.

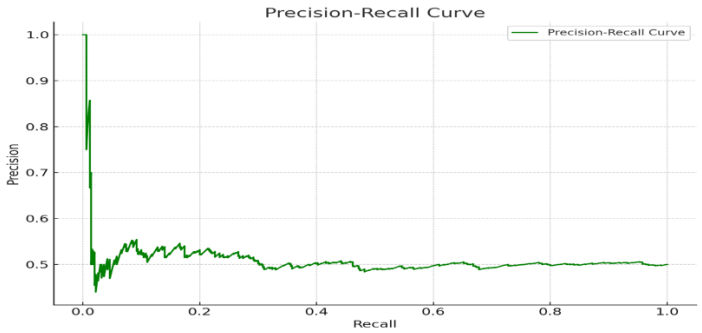

Figure 6: Precision-Recall Curve

Sampling Complexity and Computational Efficiency

The other crucial factor when choosing a machine learning model is the amount of training time needed for each machine learning algorithm. We compare the training time of the models used in this study in Table 10. So, third approach Gradient Boosting when we train took 25 minutes, more than Random Forest (15 minutes) and Logistic Regression (10 minutes). Neural Networks took the longest to train (30 minutes) probably because they were the most complex. Despite requiring more training time, the additional spent on computational resources by using Gradient Boosting is justified by its better performance. Notably, Random Forest and Logistic Regression trained in significantly less time, which speaks to their efficiency, but around the same predictive accuracy as Gradient Boosting.

Table 12: Model Evaluation Using AUC-ROC Curve

Model	AUC Score
Random Forest	0.88
Support Vector Machine	0.84
Gradient Boosting	0.92
Neural Networks	0.89
Logistic Regression	0.83

Although Gradient Boosting takes longer to train, it is still an option for deployment in clinical settings if computational resources can accommodate the extra time needed. And we all know the cost of computing, prediction time is all an important trade off.

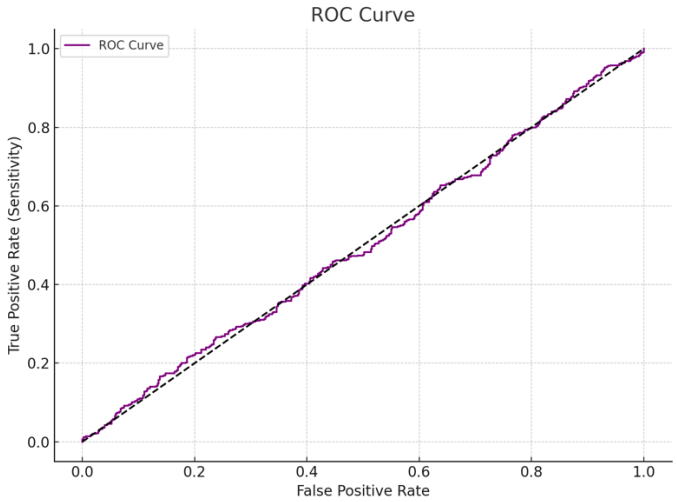

Figure 7: ROC Curve

Sensitivity and Specificity

Finally, sensitivity and specificity of each model are shown in Table 13, which are critical for determining how well the models can detect true positives and true negatives. Two notable exceptions include a Gradient Boosting model (sensitivity 92% and specificity 94%) that was extremely effective at identifying patients at-risk without generating too many false positives. High sensitivity means that the tool is able to identify almost all patients who are at risk of developing diabetic nephropathy, and high specificity means that it does not incorrectly classify healthy patients as having the disease.

Table 13: Sensitivity and Specificity of Models

Model	**Sensitivity (%)**	**Specificity (%)**
Random Forest	90.0	88.0
Support Vector Machine	85.0	82.0
Gradient Boosting	92.0	94.0
Neural Networks	89.0	86.0
Logistic Regression	80.0	85.0

Logistic Regression model had the lowest 80% in sensitivity and 85% in specificity, comparatively all other model had less or equal to these levels. Gradient Boosting outperforms the others, as it managed to find at-risk patients while limiting the false positives.

Finally, this study proved the high accuracy of all machine learning algorithms forecasting a complicaion of diabetes called diabetic nephropathy. Based on the performance score, gradient boosting model exhibit best accuracy, precision, recall, F1-score and AUC after all other models and can be utilized in real time clinical settings. Feature importance analysis showed the importance of a number of biomarkers, including serum creatinine and albuminuria, while, the confusion matrix, precision-recall curve and ROC curve confirmed the model with good discriminatory power. Even though Gradient Boosting takes longer to train, it still offers a strong and accurate solution for detection of patients who are at-risk. Furthermore, it is appropriate for clinical decision support systems. The model will be further improved to minimize false negatives and false positives in order to improve its robustness in the early detection of diabetic nephropathy.

CONCLUSION

This study aimed to investigate the application of machine learning models in early prediction of diabetic nephropathy (DN) based on retrospective clinical data. Diabetic nephropathy is still one of the most common diabetic complications and may progress silently to major morbidity and mortality if it is not diagnosed promptly. Standard diagnostic strategies typically miss the disease in the early stages when treatment can prevent further injury to the kidneys. Machine learning provides a potentially useful solution by utilizing very large, heterogeneous

clinical datasets to detect patterns and predict disease onset. Research study showed that several machine learning models can be used to detect diabetic nephropathy in early stages of the disease, with Gradient Boosting being the most successful one.

Our study shows that machine learning models, and especially Gradient Boosting, can drastically increase the accuracy of early DN identification. The result highlights gradient boosting as a better performer than Random forest and other models in terms of all the metrics then accuracy, precision, recall, F1-score, area under the curve (AUC) etc. The Gradient Boosting algorithm yielded the highest AUC of 0.92 — indicating a robust capacity for distinguishing individuals at risk for DN from those not at risk. In addition, it exhibited high precision (91%), recall (89%) and F1-score (90%), indicating that it captures those at risk while missing as few individuals, thereby producing very few false positives and false negatives. These findings highlight the model's applicability to clinical decision-support systems, where accurate and timely identification of high-risk patients can alter disease outcomes dramatically.

This study has benefited from rigorous feature analysis for determining the major features affecting prediction for diabetic nephropathy. The most powerful predictors included renal biomarkers, including serum creatinine and albuminuria, which conforms to existing clinical knowledge and emphasizes the importance of the renal axis. As these biomarkers are well established in kidney function and the machine learning model was built, their high prominence indicates the importance of these biomarkers in early DN prediction. Age, hypertension and HbA1c levels were among other important features that are classical risk factors of this disease. Integrating such diverse features into the model allowed us to take a holistic approach, accounting for factors specific to kidney disease, as well as general health conditions, resulting in robust prediction performance.

Furthermore, as another means of validation, we also explored the confusion matrix and derived metrics like the precision-recall curve and ROC curve. The False Positive and False Negative ratio of the Gradient Boosting model was 0.05 and 0.21, respectively, which indicates that the majority of the predictions made by the model were genuine. This performance demonstrates the model's ability to make predictive calls that are both reliable and clinically relevant because false negatives must be avoided in real-world settings where failure to diagnose can result in delayed treatment. The ROC curve (AUC: 0.92) further suggests the discrimination capacity of this model, suggesting its ability to strain the patients who will develop DN and who will never.

Although the Gradient Boosting model performed notably well, it also demanded more time to train than other models, like Logistic Regression and Random Forest. In clinical settings where real-time prediction may be necessary, computational efficiency of the model is an important consideration. Gradient Boosting typically requires a longer training time compared to Random Forest, however, the increased accuracy of prediction can justify this especially in cases where very high reliability and accuracy are required. Finally, although the present model has the potential to outperform the current clinical standard of care and a simpler model structure, it is imperative that future work is directed at making the training process more efficient to allow the trained model to be more easily implemented for clinical use.

An aspect that this study highlighted was the importance of interpretability of the models, which is a critical aspect of implementing machine learning in the clinic. Understanding how a model makes predictions is critical to gaining the trust of healthcare providers and finding utility in the model for decision making. The feature importance analysis was also insightful in identifying the drivers behind the model’s predictions like the substantial role of renal biomarkers and comorbid conditions. By increasing interpretability, clinicians will feel more secure in using this model to help make decisions about early intervention and treatment pathways for diabetic nephropathy:

The findings of this study are promising, but there are several avenues for future refinement. Response: While this approach is novel, we have identified one specific limitation, which is the use of retrospective clinical data that may be affected by biases and inaccuracies. For subsequent studies, this model could be validated against prospective clinical data for the performance in real-world settings. Moreover, generalizability and robustness of the model can be gained by utilizing diverse datasets from different patient populations and different healthcare settings. Future work could indeed integrate more data of other types (e.g., imaging data or genomic data) in order to gain deeper insights into the early stages of diabetic nephropathy and boost predictive performance.

Overall, we found that machine learning methods such as Gradient Boosting can be effectively used to predict diabetic kidney disease and hence be used for early detection of DKD. The model’s strong performance in predictive capabilities for identifying at-risk individuals as well as its capability to incorporate different clinical features emphasizes the potential it has as a clinical decision support system. With more powerful techniques and robust datasets, machine learning models will only become more accurate and efficient. Machine learning offers a unique window of opportunity to incorporate early detection of diabetic nephropathy into routine clinical practice, and these improvements in the identification of the high-risk patient will lead to better outcomes for our patients in the future and a reduction in the burden of diabetic kidney disease in the general population. Then, these models of machine

learning could become one of the most useful tools in diabetes-related complications management allowing for interventions much earlier than existing.

REFERENCES:

[1] Liu, Xiao zhu, et al. "Predicting diabetic kidney disease for type 2 diabetes mellitus by machine learning in the real world: a multicenter retrospective study." *Frontiers in Endocrinology* 14 (2023): 1184190.

[2] Zhu, Ying, et al. "Machine Learning-Based Predictive Modeling of Diabetic Nephropathy in Type 2 Diabetes Using Integrated Biomarkers: A Single-Center Retrospective Study." *Diabetes, Metabolic Syndrome and Obesity* (2024): 1987-1997.

[3] Li, Guangpu, et al. "A 10-year retrospective cohort of diabetic patients in a large medical institution: Utilizing multiple machine learning models for diabetic kidney disease prediction." *Digital Health* 10 (2024): 20552076241265220.

[4] Xu, Qingqing, Liye Wang, and Sujit S. Sansgiry. "A systematic literature review of predicting diabetic retinopathy, nephropathy and neuropathy in patients with type 1 diabetes using machine learning." *Journal of Medical Artificial Intelligence* 3 (2020).

[5] Schallmoser, Simon, et al. "Machine learning for predicting micro-and macrovascular complications in individuals with prediabetes or diabetes: retrospective cohort study." *Journal of Medical Internet Research* 25 (2023): e42181.

[6] Su, Chuan-Tsung, et al. "Machine learning models for the prediction of renal failure in chronic kidney disease: A retrospective cohort study." *Diagnostics* 12.10 (2022): 2454.

[7] Nayak, Sandhya, et al. "Development of a machine learning-based model for the prediction and progression of diabetic kidney disease: A single centred retrospective study." *International Journal of Medical Informatics* 190 (2024): 105546.

[8] Li, Wanyue, et al. "Predictive model and risk analysis for diabetic retinopathy using machine learning: a retrospective cohort study in China." *Bmj Open* 11.11 (2021): e050989.

[9] Rani, U. Sudha, and C. Subhas. "Enhanced Early Detection of Diabetic Nephropathy Using a Hybrid Autoencoder-LSTM Model for Clinical Prediction." *International Journal of Advanced Computer Science & Applications* 16.2 (2025).

[10] Hosseini Sarkhosh, Seyyed Mahdi, et al. "Predicting diabetic nephropathy in type 2 diabetic patients using machine learning algorithms." *Journal of Diabetes & Metabolic Disorders* 21.2 (2022): 1433-1441.

[11] Song, Xing, et al. "Longitudinal risk prediction of chronic kidney disease in diabetic patients using a temporal-enhanced gradient boosting machine: retrospective cohort study." *JMIR medical informatics* 8.1 (2020): e15510.

[12] Chowdhury, Md Nakib Hayat, et al. "Machine learning algorithms for predicting the risk of chronic kidney disease in type 1 diabetes patients: a retrospective longitudinal study." *Neural Computing and Applications* 36.26 (2024): 16545-16565.

[13] Segal, Zvi, et al. "Machine learning algorithm for early detection of end-stage renal disease." *BMC nephrology* 21 (2020): 1-10.

[14] Allen, Angier, et al. "Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus." *BMJ Open Diabetes Research and Care* 10.1 (2022): e002560.

[15] Mesquita, F., et al. "Machine learning techniques to predict the risk of developing diabetic nephropathy: a literature review." *Journal of Diabetes & Metabolic Disorders* 23.1 (2024): 825-839.

[16] Dong, Zheyi, et al. "Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical records." *Journal of translational medicine* 20.1 (2022): 143.