Application of Machine Learning for Predicting the Occurrence of Nephropathy in Diabetic Patients

Benjamin Lartey
Dept. of Electrical and Computer Engineering
North Carolina A&T State University Greensboro, NC

Kelvin Adrah
Joint School of Nanoscience and Nanoengineering
University of North Carolina Greensboro
Greensboro, NC

Frederick Adrah
Joint School of Nanoscience and Nanoengineering
University of North Carolina Greensboro
Greensboro, NC

Joan Isichei
Dept. of Industrial and Systems Engineering
North Carolina A&T State University Greensboro, NC

ABSTRACT
This paper presents an in-depth technical analysis and comparison of various machine learning models for predicting the occurrence of nephropathy in diabetic patients. The models evaluated in this study encompass a wide range of algorithms, including logistic regression, support vector machines, decision trees, random forest, naive Bayes, k-nearest neighbors, gradient boosting machines, and fully connected neural network. The performance of these models is evaluated using accuracy, precision, and recall metrics. The findings from this extensive evaluation provide valuable insights into the strengths and limitations of each model, facilitating informed decision-making for selecting the most appropriate algorithm for predicting the occurrence of nephropathy in diabetic patients. The experimental results indicated that random forest exhibited an excellent performance whereas naive bayes algorithm performed poorly.

General Terms
Application of Machine Learning, Diabetes

Keywords
Nephropathy, diabetic patients, machine learning

1. INTRODUCTION
Nephropathy, also known as kidney disease has been known to affect over 10% of the global population, representing over 800 mil- lion worldwide [14]. Aside other risk factors, diabetes is regarded as the most notable cause of end-stage nephropathy [21]. The dis- ease has a huge public health and socio-economic impact on both developed and developing countries. Timely diagnosis can expedite patients receiving better disease management and yield a higher chance of positive prognosis [12]. However, there are mostly limited or no visible symptoms in the early stages of the disease. This means patients fail to notice the disease, which can aggravate other organs and cause lasting damage in some cases [18]. The cost of treating nephropathy in general is high, and preventive measures must be taken to avoid getting to the end stage [18, 8].

Earlier research done as far back as 2013 used artificial neural net- work (ANN) to classify patient status to determine nephropathy. The classifiers were trained using data sampled from the University of Bari (Italy) over a period of 38 years and the assessment was done based on precision, recall and F-measure [1]. ANN has proven to be an effective machine learning model for medical data analysis with the ability to map out non-linear relationships [13]. In recent times, machine learning, advanced data analytics and predictive modeling tools have been employed to improve the early detection of nephropathy in patients with some known clinical data [18, 19]. Risk factors attributed to kidney disease can be derived from patients' vital statistics such as Body Mass Index (BMI), Blood Pressure (BP), age and urine protein levels, among others [17]. Machine Learning algorithms can be used in tandem with software systems to monitor and predict kidney diseases. Re-searchers have found out that using integrated models such as a combination of logistic regression and random forest could achieve an accuracy as high as 99.83% [18, 23].

A more effective and efficient mechanism is to use machine learning and comparative data mining techniques to predict the presence of early-stage nephropathy in patients using clinical data collected from patients in a hospital. In this paper, six (6) parameters that are some of the topmost risk factors in diabetic patients are considered while using machine learning models to forecast nephropathy in patients. The data was collected from the Ho Municipal Hospital in Ghana after a 5-month survey.

Blood sugar is an important determinant in healthcare due to its direct relation with cardiovascular disease, obesity and kidney dis- ease [15]. The normal range for blood sugar in healthy individuals is between 3.9-5.6 mmol/L, beyond which a person is considered diabetic. Glycated haemoglobin/HbA1C is a simple test used to measure blood sugar over a period of 2-3 months. It is a simple cost-effective test that has been regarded as the gold standard for diabetes diagnosis and management; with a normal range between 4.0-5.6% [5]. Data for individuals who are either smokers or non- smokers was also collected. Research indicates smoking as one of the risk factors of kidney disease. Evidence suggests that as much as 25 million diabetic patients worldwide have smoking alone as the main attributable causative factor [11, 4].The Body Mass Index (BMI) range for healthy individuals is between 18.5-24.9 kg/m2 with obese persons found in the range of 30-39.9 kg/m2. Studies have shown a correlation between obesity and chronic kidney disease using multivariable logistic regression, making BMI a key parameter for determining the presence of nephropathy [3]. Similar studies have shown that most of the risk factors that cause nephropathy are in elderly patients, usually over the ages of 45, with 1 in 7 adults globally having some form of nephropathy [7, 24].

Machine learning models can enhance a more proactive healthcare. Medical personnel can use these predictions to better manage patients by implanting preventive measures such as lifestyle changes or early treatment. These models can also empower patients by alerting them about their risks of developing nephropathy. This education enables them to make

more informed decisions and better self-care. By harnessing the power of data-driven techniques, this body of research seeks to enhance the predictive capacity of the occurrence of nephropathy in diabetic patients based on vital clinical parameters.

In summary, the contributions of this work are:

(1) Employ machine learning models to predict the occurrence of nephropathy in diabetic patients. The models considered include logistic regression, support vector machines, k-nearest neighbors, naive bayes, decision tree, random forest, gradient boosting machines and fully connected neural network.

(2) Conduct extensive experiments to evaluate the effectiveness and efficiency of the models.

The remaining parts of this paper are organized as follows. Section 3 presents the problem definition. Section 4 describes the various machine learning models considered. In Section 5, we provide a detailed description of the experimental setup and results. We then conclude our studies and future work in Section 6.

2. RELATED WORK

Existing research done in this field using Machine Learning (ML) algorithms focused on using the least subset of features from datasets obtained from the University of California, Irvine (UCI) data repository. The four ML algorithms used in the modeling stage are support vector machines (SVM), logistic regression (LR), gradient boosting (GB) and random forest (RF). These algorithms were modelled to assess the ability of the datasets to detect nephropathy in patients. At the end of the testing, all the models displayed a tremendous performance (> 97%) for detecting nephropathy. The gradient booster displayed the highest accuracy of 99% [1]. Similar works using Python programming and the scikit-learn library has been employed to predict the presence of IgA nephropathy. In addition to SVM, RF and ANN, Gaussian Naïve Bayes Classifier and K-nearest neighbor classifiers were developed. Data was collected from the Wroclaw Medical University in Poland, focusing on patients with biopsy-proven Immunoglobulin A Nephropathy (IgAN) [13]. Due to the small amount of data used, it limits the effective use of ML on a wider scale. Researchers find the use of larger datasets a more efficient way of validating results of smaller datasets [1]. However, this research established that there is no such thing as "one effective model", but rather a group of effective models. Furthermore, it was demonstrated that ML tools and techniques are capable of fishing out important data from a cluster of datasets for predicting the presence of nephropathy [18, 13, 19]. This is a continually evolving field, with researchers exploring new techniques to enhance the accuracy of predictions. A consolidation of contemporary research work done in this field supports ML as an important tool in healthcare to advance and sup- port nephropathy diagnostics. The choice of input data and the right model or combination of models have a direct correlation on the performance of the output [13].

3. PROBLEM STATEMENT

Given a medical data $\boldsymbol{X} = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\}$ with $\boldsymbol{N}$ samples, an output label $\boldsymbol{y} = \{\boldsymbol{0}, \boldsymbol{1}\}$ where $\boldsymbol{1}$ means there is the occurrence of nephropathy while $\boldsymbol{0}$ means otherwise. We employ a machine learning model $\boldsymbol{f}(\boldsymbol{X})$ to the data $\boldsymbol{X}$ to predict whether there is the occurrence of nephropathy $(\boldsymbol{y} = \boldsymbol{1})$ or not $(\boldsymbol{y} = \boldsymbol{0})$.

4. METHODOLOGY

In this section, we first provide a brief description of the following machine learning models: random forest, decision tree, sup- port vector machines, naive bayes, logistic regression, fully connected neural network, k-nearest neighbors and gradient boosting machines. Then, we present the approach considered for predicting the occurrence of nephropathy in diabetic patients.

4.1 Random Forest

Random forest is an ensemble learning method that consists of a collection of decision trees. Each tree is built based on a distinctive arbitrary subset of the input features, and the ultimate prediction of the ensemble is the majority vote among the trees [2]. In addition, each tree plays the part of performing a "nonlinear mapping from complex input spaces into continuous output spaces' [10]. The non- linearity is accomplished by partitioning up the initial problem into a miniature one that can be solved with less complex models [10]. Overfitting is more likely to occur for a single standard decision tree whereas an ensemble of randomly trained trees possesses high generalization power. Random Forests are especially practical for the current study as they have great classification performance and can be used for multi-label and binary classification.

4.2 Decision Tree

Decision Tree is a non-parametric supervised learning model that learns decision rules from the input data to make predictions or classify instances. The algorithm builds a tree-like model of decisions and their possible consequences, where each internal node represents a feature or attribute, each branch represents a decision rule, and each leaf node represents a class label or a predicted value. The key steps of the decision tree algorithm is summarized as follows: attribute selection, splitting data into subsets recursively until a certain stopping criteria is reached, assigning a class label or a predicted value to each leaf node based on the majority class and finally pruning to avoid overfitting [20].

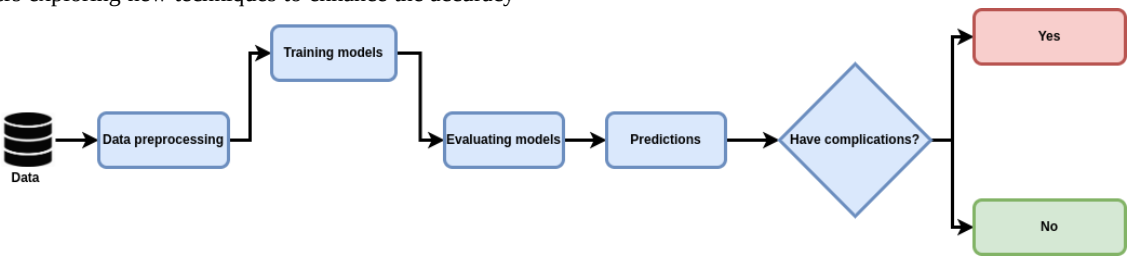

Fig. 1. Flow chart for predicting the occurrence of nephropathy in diabetic patients

4.3 Support Vector Machines

Support vector machines (SVM) is an extremely effective algorithm for classification problems. SVMs possess computational advantages and a good generalization performance. A noteworthy characteristic of SVM is that it is absolutely insensitive to the comparative number of data points in each class, since it does not try to curtail the error rate [25].

VMs work by finding a hyperplane that isolates the information sample data such that it is divided into two classes. The selected hyperplane is one in which the distance between the hyperplane and the closest data points on either side is maximized [25]. This maximizes the edge of division between the two classes, thus improving the precision of the model. SVM is a competent algorithm that can be utilized to solve an assortment of classification problems, hence its use in this work.

4.4 Naive Bayes

The Naive Bayes algorithm is a machine learning algorithm commonly used in classification tasks. It assumes that the characteristics in the data set follow a Gaussian (normal) distribution and assumes independence of the characteristics. Applying Bayes' theorem, the algorithm calculates the probability of a given class identifier for a set of features [16]. It achieves this by estimating the mean and variance of each feature for each class of training data. In the prediction phase, the algorithm uses these estimates to calculate the probability of a given class identifier for a new case [16]. Naive Bayes is efficient and works well under the assumption of feature independence, making it a popular choice for text classification and other areas where that assumption is feasible.

4.5 Logistic Regression

Logistic regression is a machine learning algorithm that is used for classification tasks. The algorithm estimates the probabilities of an instance belonging to each class using the SoftMax function, which guarantees that the predicted probabilities sum up to one. By com- paring these probabilities, the algorithm assigns the instance to the class with the highest probability. In other words, it utilizes a decision boundary to assign the most likely class label based on the highest predicted probability.

4.6 Fully Connected Neural Network

Fully Connected Neural Network (FCNN) are typically used for ap- plications in which there are unknown relationships between input and output variables, making them a formidable and flexible modelling tool for predictive purposes. A neural network is a sequence of nodes/neurons. Each node comprises a set of inputs, weight, and a bias value. The weights are the parameters in an FCNN that trans- form input data within the network's hidden layers. The operation of the FCNN is an iterative process that consists of an input arriving at the node, and then it is multiplied by a weight value. The weighted inputs are summed at the processing component. If the sum crosses a threshold value, it goes as input to other neurons passes out as an output from the model. FCNN makes use of an activation function, a transfer function that is used to decide whether a neuron is activated or not or to achieve the target output for the problem.

4.7 K-Nearest Neighbors

K-nearest neighbors' algorithm (KNN) is a non-parametric method used for classification. The k-nearest neighbor is used to establish the category of test data points according to K-neighboring samples. The so-called adjacent sample is the K samples closest to it in the initial data, which are computed by calculating the Euclidean distance between them and all known data points [6]. The K ad- jacent data points with the smallest distance from the sample to be tested are selected to confirm the proportion of each category in these K data points. The category with the highest proportion is used as the prediction classification of this sample.

4.8 Gradient Boosting Machines (GBoost)

GBoost is an algorithm that is extensively used due to its efficacy. It is an ensemble learning algorithm because it learns while adding a sequence of weak learner decision trees additively so that the loss function decreases steadily, thus minimizing the prediction errors of the weak learners [9]. It possesses a fast training speed, high efficiency, supports parallel learning, and can handle large-scale data [9]. The gradient-boosting model consists of an ensemble of weak learners. As previously mentioned, the GBoost algorithm is highly effective in modeling intricate, nonlinear relationships and in handling high-dimensional data as in the case of Electronic Health Records (EHRs). It was selected for use in this study because its use is well entrenched in the medical literature [22]. Additionally, its flexibility allows us to make explicit comparisons between it and other models used in this work.

4.9 System Overview

Figure 1 provides a high-level description of the steps involved in predicting the occurrence of nephropathy. The raw input data is first cleaned to remove missing values and outliers. A data trans- formation technique is applied to scale the data values in a specified range. The preprocessed data is split into training and testing set using sklearn's train test split() library. The training set is employed to train each of the eight machine learning models considered in this work. The testing set is used to evaluate the prediction performance of the models based on some evaluation metrics. The models predict whether a diabetic patient will have complications due to diabetes or not.

5. EXPERIMENTAL STUDY

In this section, we provide the details of our experiment and analyze our results.

5.1 Dataset Description

The data was collected from the Ho Municipal Hospital in Ghana after a 5-month survey. The data comprises five features and over 3000 samples. The binary class label is the occurrence of nephropathy which indicates whether a diabetic patient will experience nephropathy or not. The features are briefly described below:

(1) Body Mass Index (BMI): BMI is a measure of body fat based on an individual's weight and height. It is calculated by dividing the weight in kilograms by the square of the height in meters. The BMI parameter in the dataset provides valuable insights into the patients' overall body composition, which is known to influence nephropathy risk.

(2) Age: It represents the age of each patient in years. Age is a crucial factor that contributes to the development of various dis- eases, including nephropathy. Understanding how age interacts with other parameters can help identify at-risk individuals.

(3) Fasting Blood Glucose (mmol/L): Fasting blood glucose levels were measured after an overnight fast and are reported in millimoles per liter (mmol/L). Elevated blood glucose levels are a hallmark of diabetes and are known to contribute to the development and progression of nephropathy.

(4) Smoking History: This categorical parameter captures the smoking habits of the patients, with values indicating whether a patient is a smoker (yes) or a non-smoker (no). Smoking is a well-established risk factor for the

progression of nephropathy in individuals with diabetes.

(5) Glycated Haemoglobin: Glycated haemoglobin, often referred to as HbA1c, represents the average blood glucose levels over a prolonged period. It is expressed as a percentage. High HbA1c levels indicate poor glucose control and are associated with an

5.2 Simulation Details

We preprocessed the data to remove missing values and outliers. The data was normalized to ensure that all parameters lie within a similar scale. The data was split into training set and testing set. 70% of the data samples were used to train the models and 30% for testing. We run the experiment 20 times and computed the average to guarantee consistent results.

5.3 Hyperparameter Settings

To ensure simplicity and a fair comparison of models, we used the default hyperparameter values for all models.

5.4 Evaluation Metrics

We evaluated the performance of all the models based on their ac- curacy, precision, and recall values. The mathematical representation of these metrics are shown below:

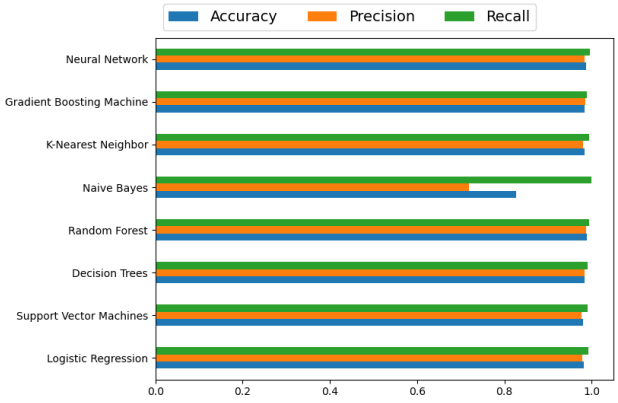

Fig. 2. Performance evaluation of models

$$Accuracy = \frac{T\,P\ +\ T\,N}{T\,P\ +\ T\,N\ +\ FP\ +\ FN}$$

$$Precision = \frac{T\,P}{T\,P\ +\ FP}$$

$$Recall\ = \frac{T\,P}{T\,P\ +\ FN}$$

where T P, T N, FP, FN represent true positive, true negative, false positive and false negative respectively.

Table 1. Performance evaluation of models.

Logistic Regression	0.9824	0.9786	0.9927
Support Vector Machines	0.9802	0.9769	0.9909
Decision Trees	0.9846	0.9840	0.9910
Random Forest	0.9901	0.9893	0.9946
Naive Bayes	0.8265	0.7193	0.7283
K-Nearest Neighbors	0.9846	0.9804	0.9945
Gradient Boosting	0.9846	0.9858	0.9893
Machines Neural Network	0.9879	0.9840	0.9964

5.5 Results and Discussions

Figure 2 and Table 1 present the performance of Logistic Regression, Support Vector Machines, Decision Trees, Random Forest, Naive Bayes, K-Nearest Neighbors, Gradient Boosting Machines, and Fully Connected Neural Network based on their accuracy, precision and recall. Random Forest achieved the highest accuracy of 0.9901 and highest precision of 0.9893 whereas Neural Network has the highest recall of 0.9964. It implies that Random Forest is capable of capturing the non-linear relationship in the data. Also, it exhibits an excellent prediction of the occurrence of nephropathy in diabetic patients. Naive Bayes has the least accuracy of 0.8265, precision of 0.7193, and recall of 0.7283. Naive Bayes performed poorly because of its underlying assumption of independent predictors. This assumption limits its performance since we considered a real medical data in this work which contains dependent predictors. With the exception of Naive Bayes, all models have a precision value above 0.9500 which means the models are able to correctly classify positive instances. Similarly, the recall values are above 0.9500 which indicates that there were few false negatives and therefore, the models are reliable in terms of classifying the occurrence of nephropathy.

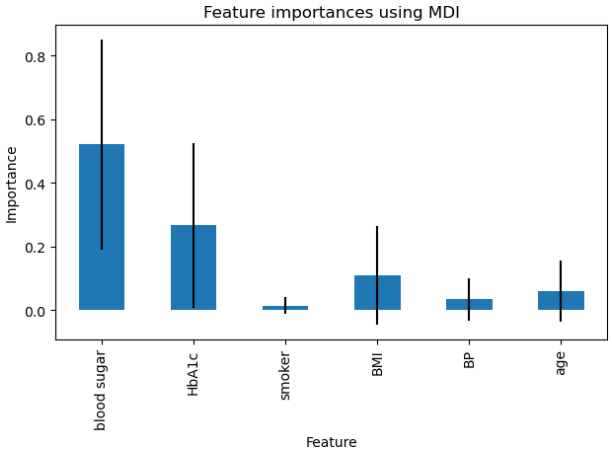

Fig. 3. Feature importance

We also analyzed the importance of each attribute in predicting the occurrence of nephropathy using mean decrease in impurity (MDI). Figure 3 indicates that blood sugar contributes the most in predicting whether a patient will experience nephropathy or not whereas smoking contributes the least.

5.6 Runtime Analysis

All experiments are run on the same machine with windows 10 operating system, 2.40GHz processing speed, 12 CPU cores and 16GB RAM. Table 2 shows the runtime of all the models. Naive Bayes model is the most efficient with a runtime of 0.01s whereas Neural Network is the least efficient with a runtime of 4.32s. Similarly, K-Nearest Neighbors is less efficient with a runtime of 3.32s. Also, Decision Trees, Support Vector Machines and Logistic Regression have an impressive efficient runtime. Random forest and Gradient Boosting Machines have comparable runtime of 0.77s and 0.34s respectively.

Table 2. Runtime of models

	Runtime (s)
Logistic Regression	0.22
Support Vector Machines	0.02
Decision Trees	0.02
Random Forest	0.77
Naive Bayes	0.01
K-Nearest Neighbors	3.32
Gradient Boosting Machines	0.34
Neural Network	4.32

6. CONCLUSION

In this paper, we presented a detailed analysis of the prediction performance and time efficiency of various machine learning models. We investigated the performance of the models for predicting the occurrence of nephropathy in diabetic patients. The models considered include Logistic Regression, Support Vector Machines, Decision Trees, Random Forest, Naive Bayes, K-Nearest Neighbors, Gradient Boosting Machines and Fully Connected Neural Network. Our extensive experimentation demonstrated that Random Forest algorithm achieves the highest performance with regards to predicting the occurrence of nephropathy whereas Naive Bayes performed poorly. Also, Naive Bayes, Support Vector Machines and Decision Trees attained the fastest time. Therefore, these models can be employed in the medical field to accurately predict the occurrence of nephropathy in diabetic patients. This will enable medical practitioners to take appropriate measures and make informed decisions in order to curtail this issue. Further investigations and experimentation with larger datasets and different evaluation metrics can enhance our understanding of these models' capabilities in predicting the occurrence of nephropathy.

7. ACKNOWLEDGMENTS

The authors would like to express their profound gratitude to the Ho Municipal Hospital in Ghana.

8. REFERENCES

[1] Marwa Almasoud and Tomas E Ward. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. International Journal of Soft Computing and Its Applications, 10(8), 2019.

[2] Timothy C Au. Random forests, decision trees, and categorical predictors: the" absent levels" problem. The Journal of Machine Learning Research, 19(1):1737–1766, 2018.

[3] Bjorn Kaijun Betzler, Rehena Sultana, Riswana Banu, Yih Chung Tham, Cynthia Ciwei Lim, Ya Xing Wang, Vinay Nangia, E Shyong Tai, Tyler Hyungtaek Rim, Mukharram M Bikbov, et al. Association between body mass index and chronic kidney disease in asian populations: a participant- level meta-analysis. Maturitas, 154:46–54, 2021.

[4] D Campagna, A Alamo, A Di Pino, C Russo, AE Calogero, F Purrello, and R Polosa. Smoking and diabetes: dangerous liaisons and confusing relationships. Diabetology & metabolic syndrome, 11(1):1–12, 2019.

[5] Haleh Chehregosha, Mohammad E Khamseh, Mojtaba Malek, Farhad Hosseinpanah, and Faramarz Ismail-Beigi. A view beyond hba1c: role of continuous glucose monitoring. Diabetes Therapy, 10:853–863, 2019.

[6] Zewei Chen, Xin Zhang, and Zhuoyong Zhang. Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. International urology and nephrology, 48:2069–2075, 2016.

[7] Dervla M Connaughton, Claire Kennedy, Shirlee Shril, Nina Mann, Susan L Murray, Patrick A Williams, Eoin Conlon, Makiko Nakayama, Amelie T van der Ven, Hadas Ityel, et al. Monogenic causes of chronic kidney disease in adults. Kidney international, 95(4):914–928, 2019.

[8] Tommaso Di Noia, Vito Claudio Ostuni, Francesco Pesce, Giulio Binetti, David Naso, Francesco Paolo Schena, and Eugenio Di Sciascio. An end stage kidney disease predictor based on an artificial neural networks ensemble. Expert systems with applications, 40(11):4438–4445, 2013.

[9] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, pages 1189–1232, 2001.

[10] Kai Hakala, Suwisa Kaewphan, Jari Bjo¨rne, Farrokh Mehryary, Hans Moen, Martti Tolvanen, Tapio Salakoski, and Filip Ginter. Neural network and random forest models in protein function prediction. IEEE/ACM Transactions on Compu- tational Biology and Bioinformatics, 19(3):1772–1781, 2020.

[11] Edgar A Jaimes, Ming-Sheng Zhou, Mohammed Siddiqui, Gabriel Rezonzew, Runxia Tian, Surya V Seshan, Alecia N Muwonge, Nicholas J Wong, Evren U Azeloglu, Alessia Fornoni, et al. Nicotine, smoking, podocytes, and diabetic nephropathy. American Journal of Physiology-Renal Physiology, 320(3):F442–F453, 2021.

[12] Naseer Ullah Khan, Jing Lin, Xukun Liu, Haiying Li, Wei Lu, Zhuning Zhong, Huajie Zhang, Muhammad Waqas, and Lim- ing Shen. Insights into predicting diabetic nephropathy using urinary biomarkers. Biochimica et Biophysica Acta (BBA)- Proteins and Proteomics, 1868(10):140475, 2020.

[13] Andrzej Konieczny, Jakub Stojanowski, Magdalena Krajew- ska, and Mariusz Kusztal. Machine learning in prediction of iga nephropathy outcome: A comparative approach. Journal of Personalized Medicine, 11(4):312, 2021.

[14] Csaba P Kovesdy. Epidemiology of chronic kidney disease: an update 2022. Kidney International Supplements, 12(1):7– 11, 2022.

[15] Wei Li, Weixiang Luo, Mengyuan Li, Liyu Chen, Liyan Chen, Hua Guan, and Mengjiao Yu. The impact of recent devel- opments in electrochemical poc sensor for blood sugar care. Frontiers in Chemistry, 9:723186, 2021.

[16] Subramani Mani, Yukun Chen, Tom Elasy, Warren Clayton, and Joshua Denny. Type 2 diabetes risk forecasting from emr data using machine learning. In AMIA annual symposium proceedings, volume 2012, page 606. American Medical Infor- matics Association, 2012.

[17] Vijayakumar Natesan and Sung-Jin Kim. Diabetic nephropathy–a review of risk factors, progression, mechanism, and dietary management. Biomolecules & therapeutics, 29(4):365, 2021.

[18] Jiongming Qin, Lin Chen, Yuhua Liu, Chuanjun Liu, Chang- hao Feng, and Bin Chen. A machine learning methodology for diagnosing chronic kidney disease. IEEE Access, 8:20991– 21002, 2019.

[19] Violeta Rodriguez-Romero, Richard F Bergstrom, Brian S Decker, Gezim Lahu, Majid Vakilynejad, and Robert R Bies. Prediction of nephropathy in type 2 diabetes: an analysis of the accord trial applying machine learning techniques. Clinical and translational science, 12(5):519–528, 2019.

[20] S Rasoul Safavian and David Landgrebe. A survey of deci- sion tree classifier methodology. IEEE transactions on sys- tems, man, and cybernetics, 21(3):660–674, 1991.

[21] Hanny Sawaf, George Thomas, Jonathan J Taliercio, Georges Nakhoul, Tushar J Vachharajani, and Ali Mehdi. Therapeutic advances in diabetic nephropathy. Journal of Clinical Medicine, 11(2):378, 2022.

[22] Hiroe Seto, Asuka Oyama, Shuji Kitora, Hiroshi Toki, Ryohei Yamamoto, Jun'ichi Kotoku, Akihiro Haga, Maki Shinzawa, Miyae Yamakawa, Sakiko Fukui, et al. Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. Scientific Re- ports, 12(1):15889, 2022.

[23] Alvaro Sobrinho, Andressa CM Da S Queiroz, Leandro Dias Da Silva, Evandro De Barros Costa, Maria Eliete Pinheiro, and Angelo Perkusich. Computer-aided diagnosis of chronic kidney disease in developing countries: A comparative anal- ysis of machine learning techniques. IEEE Access, 8:25407– 25419, 2020.

[24] Katherine R Tuttle, Radica Z Alicic, O Kenrik Duru, Cami R Jones, Kenn B Daratha, Susanne B Nicholas, Sterling M McPherson, Joshua J Neumiller, Douglas S Bell, Carol M Mangione, et al. Clinical characteristics of and risk fac- tors for chronic kidney disease among adults and children: an analysis of the cure-ckd registry. JAMA network open, 2(12):e1918169–e1918169, 2019.

[25] Ginny Y Wong, Frank HF Leung, and Sai-Ho Ling. Predict- ing protein-ligand binding site using support vector machine with protein properties. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10(6):1517–1529, 2013.