Volume 3, Issue 1, January –2024
PP 520-525

International Journal of Futuristic Innovation in Arts, Humanities and Management (IJFIAHM)

# Comparative Study of Machine Learning Algorithm to Predicting Diabetics Kidney Disease

Shweta Yadu (MTech)
Computer science and engineering
Raipur Institue of Technology
*(CSVTU)*
Raipur, Chhattisgarh
y.shweta1316@gmail.com

Mahadev Bag (Head Of Dept.)
Computer science and engineering
Raipur Institue of Technology
*(CSVTU)*
Raipur, Chhattisgarh
bagmahadev1010@gmail.com

**Abstract**— Diabetic Kidney Disease (DKD) poses a significant health concern globally, impacting 20-40% of individuals affected by diabetes. This research presents a comprehensive comparative study aimed at predicting DKD onset using machine learning algorithms. The main goals of the project are to use medical data to identify people who are at risk and to construct a software system that can predict the chance of a disease developing in the future. Several machine learning classification methods, such as IBK, Random Tree, Random Forest, Naive Bayes, and adaBoostM1, are used in the study technique. By utilizing the WEKA machine learning software, these algorithms are put through a rigorous comparison process in order to ascertain their prediction effectiveness. The study finds that IBK and Random Tree classification are the best-performing techniques using 10-fold cross-validation. They show an accuracy of 93.6585% and a higher K value (0.8731). This comparative research demonstrates how machine learning approaches may be used to predict DKD onset accurately, which can lead to proactive intervention options.

***Keywords***- *Machine Learning, Classifier, Random Tree, adaBoost, IBK, Diabetic Kidney Disease Prediction*

## 1. Introduction

Diabetic Kidney Disease (DKD) remains a prevalent and critical complication affecting a significant proportion of individuals afflicted with diabetes globally. The escalating prevalence of diabetes has amplified concerns regarding associated complications, with DKD emerging as a major cause of end-stage renal disease (ESRD) and a substantial contributor to cardiovascular morbidity and mortality.

The imperative to predict, detect, and mitigate the progression of DKD has prompted a surge in research exploring predictive modeling techniques, notably employing machine learning algorithms. These computational methodologies offer promising avenues to harness the power of extensive medical data for prognostication and early intervention. This research embarks on a comprehensive exploration aimed at evaluating and comparing various machine learning algorithms for their efficacy in predicting the onset and progression of diabetic kidney disease. By leveraging a diverse array of classification algorithms including IBK, Random Tree, Random Forest, Naive Bayes, and adaBoostM1, this study endeavors to decipher the optimal model for accurate prognostication.

The primary objective is not only to identify individuals at heightened risk of DKD onset but also to devise a robust software system capable of foreseeing the temporal trajectory of the disease, aiding in informed clinical decision-making and targeted interventions. Through an intricate comparative analysis utilizing the WEKA machine learning software and employing rigorous 10-fold cross-validation, this study seeks to unravel the algorithmic nuances and performance metrics critical for precise DKD prognostication. Furthermore, this research aims to bridge existing gaps in the domain of predictive modeling for diabetic kidney disease, contributing valuable insights that could potentially revolutionize clinical strategies, improve patient outcomes, and reduce the burden of DKD on healthcare systems. The subsequent sections delve into the methodology employed, present findings, comparative analysis, and discuss implications, aiming to provide a comprehensive understanding of machine learning's role in predicting diabetic kidney disease.

As information technology has developed, vast volumes of data have been produced. Innovations in healthcare information management systems have also led to an abundance of medical databases. The handling of diverse data and the extraction of valuable knowledge from it constitute a major area of study in data mining. Identifying unique, useful, legitimate, and logical patterns in data is the goal of this technique [1]. There are two major categories of data mining approaches: supervised and unsupervised learning procedures. When an

algorithm is trained using data that is neither labeled nor categorized, it may operate on it without human oversight. This process is known as unsupervised learning. Classification and association are the two kinds of unsupervised learning algorithmsUsing well-labeled data to instruct or train a machine is known as supervised learning [2]. This indicates that some data has already been assigned the right response. Regression, logistic regression, classification, Naive Bayes classifiers, K-NN (k closest neighbors), decision trees, support vector machines, and regression are some examples of supervised learning methods. [3].

With the use of past data and machine learning, a computer system may be explicitly trained to make predictions or execute certain actions. Massive amounts of structured and semi-structured data are used in machine learning in order to provide reliable results or predictions based on the data [4]. Analyzing an algorithm for data extraction automatically is part of machine learning. In order to build models of what is happening behind specific information and forecast future outcomes, machine learning makes use of data mining techniques as well as another learning algorithm[5]. A set of machine learning algorithms for data mining activities is called Weka (Waikato Environment for Knowledge Analysis). Tools for data preparation, classification, regression, clustering, mining association rules, and visualization are all included in it[6]

Controlling and treating high blood pressure and diabetes are key components in treating diabetic kidney disease (DKD). Medication, exercise, dietary modifications, and prescription drugs are all part of the treatment. Blood pressure and blood sugar control may help to avoid or postpone renal problems and other consequences [7]. Pharmaceuticals: -When diabetic nephropathy is first developing, medications to control the following conditions may be part of your treatment: heart rate. Angiotensin 2 receptor blockers (ARBs) and angiotensin-converting enzyme (ACE) inhibitors are medications used to treat high blood pressure. blood sugar. For those with diabetic nephropathy, medications can help regulate elevated blood sugar levels. They include insulin and other older diabetic medications. "Newer medications, such as Metformin (available under brands like Fortamet, Glumetza, among others), glucagon-like peptide 1 (GLP-1) receptor agonists, and SGLT2 inhibitors, are emerging in diabetes treatment. Consulting your healthcare provider regarding the suitability of SGLT2 inhibitors or GLP-1 receptor agonists for your condition is advisable, as these therapies exhibit potential in safeguarding the heart and kidneys from diabetes-related damage.

Additionally, managing high cholesterol often involves statins, medications designed to mitigate high cholesterol levels and decrease urinary protein excretion. For diabetic nephropathy's kidney scarring, the use of Finerenone (marketed as Kerendia) shows promise in reducing tissue scarring. Studies indicate its potential to mitigate the risk of kidney failure, diminish the likelihood of heart-related fatalities, including heart attacks, and reduce hospitalizations due to heart failure in adults managing chronic kidney disease associated with type 2 diabetes [8]. If you take these medicines, you'll need regular follow-up testing. The testing is done to see if your kidney disease is stable or getting worse [9].

## 2. RELATED WORK

Logistic regression has been utilized by Mishra et al. to forecast diabetes. The dataset they utilized has several characteristics, including gender, body mass index, HBA1C, and others. Using IBM SPSS 20.0 software, they were able to analyze data with a likelihood of 78.556% correctness [10].

PIMA data and the WEKA software tool were utilized by Gnana et al. They have employed a number of algorithms, including MLP, Naive-Bayes (NB), and Random Forest (RF), as well as a number of testing techniques, including UTD, PS, and FCV. For prediction, the author has examined both the preprocessed and unprocessed data. They have attained 100% accuracy by using the UTD approach with the RF algorithm [11].

To create a prediction model based on several machine learning algorithms, Chowdhury et al. examined the data from the epidemiology of diabetes treatments and complications clinical trials. There were 19 characteristics and 1375 type 1 diabetic individuals. The results showed that the light gradient-boosted machine came in second (95%) and the random forest model (96%), as the best [12].

B. Boukenze et coll. [1] The chronic failing illness prediction system that has been proposed uses a variety of machine learning methods, including SVM, KNN, MLP/ANN, C4.5, and Bayesian networks. Compare the algorithm's performance accuracy based on several metrics, such as execution time, sensitivity specificity, and f-

measure, after incorporating weak platform tools. A decision tree with a C4. 5 score has high accuracy. It was used to forecast the illness chronic kidney failure with a 63% accuracy rate.

Comparison research on the diagnosis of thyroid illness utilizing naïve bays and KNN to categorize thyroid disease data set was proposed by Chandel, K. et al. Following the thyroid dataset's implementation on Rapid Miner, the experimental outcome demonstrates that KNN outperformed naïve bays in terms of performance accuracy. In comparison to KNN, Naïve Bays fared poorly, scoring 22.56% as opposed to 93.44% for KNN [13].

In order to forecast the transitional period of renal illness, particularly stages 3 to 5, Pangong, P. et al. presented by building classification models by applying data mining, classification techniques such KNN, ANN, DT, and Naïve Bays. The chosen or reduced collection of characteristics was classified in order to create these classification models. Accuracy and performance were guaranteed to be approximately 85% after using a balanced classifier or attribute lowered by the feature selection approach [14].

## 3. METHODOLOGY

For this investigation, clinical and biochemical information from DKD patients was collected. The risk factors for diabetic kidney disease are displayed in Figure 1.

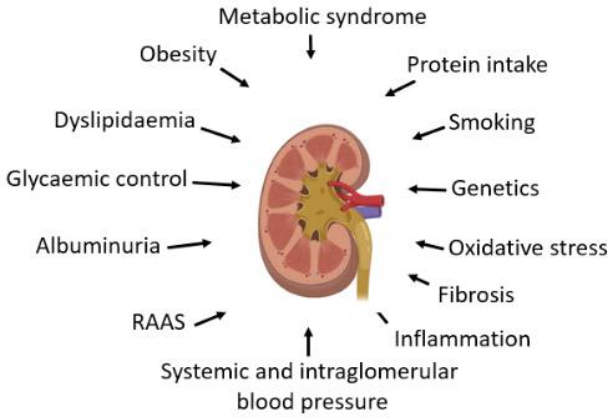


Figure1 The risk factors for diabetic kidney disease

The collected data underwent a transformation into ARFF file formats, short for Attribute-Relation File Format (ARFF), an extension similar to CSV files but with an added header containing metadata about the data categories within columns. This conversion was necessary to make the dataset compatible with WEKA, a process initiated using the "Reviewer" tool under the "Tools" menu within WEKA. Initially in CSV format from Microsoft Excel, the data was saved with an ARFF extension to enable its utilization in the WEKA environment. This transition was pivotal for the subsequent analysis. Prior to any WEKA analysis, the dataset underwent a critical 10-fold cross-validation.

Dataset: The dataset for diabetic kidney disease was sourced from UCI [18], comprising 18 attributes—14 numeric and 4 nominals—spanning 410 cases. Key features such as age (years), gender (male/female), serum albumin (mg/dL), sodium (mmol/L), potassium (mmol/L), urea (mg/dL), glucose (mg/dL), creatinine (mg/dL), HbA1c (%), Hb (g/dL), white blood cell counts (WBCs) (109/L), red blood cell counts (RBCs) (1012/L), Hb (%), platelet counts (109 /L) (M/µl), systolic BP in sitting condition (mmHg), diastolic BP in sitting condition (mmHg), hypertension (yes/no), and retinopathy (yes/no) constituted the dataset's features.

Preprocessing: WEKA's primary GUI Chooser window provided access to four interfaces. Loading the DKD dataset into the WEKA explorer window revealed Figure 4, visualizing the data with color-coded distinctions (blue and red) within the visualization area. To streamline analysis and evaluation, WEKA findings were categorized into multiple subitems. This initial phase involved segregating data into numerical and percentage values based on accurate and erroneous categorizations. Subsequent steps encompassed computing Kappa statistics, meaning absolute error, and root mean squared error for numerical data.

Classification: Classification, a fundamental data mining technique, determines outputs for new data instances. For evaluating overall performance and selecting the optimal classifier for DKD prediction, diverse classifiers were applied to the DKD dataset. Metrics including accuracy, instances correctly or erroneously classified, error rates, and execution time were compared across these applied algorithms.
Applied Algorithms: The study embraced distinct classification approaches, including...

AdaBoost- AdaBoost, an abbreviation for Adaptive Boosting, stands as an ensemble machine learning algorithm proficient in addressing an extensive spectrum of regression and classification tasks. This supervised learning technique amalgamates numerous weak or base learners, such as decision trees, to collaboratively construct a robust learner capable of effectively categorizing diverse datasets.
IBK: The k-nearest-neighbor algorithm is that. It belongs to the lazy class in Weka and is known as instance-based learning with parameter k (IBk). Let's access the dataset for glass. Navigate to Classify and choose IBk, the lazy classifier.
Bayes' Theorem: is the foundation for a group of classification algorithms known as Naïve Bayes classifiers. It is a family of algorithms rather than a single method, and they are all based on the same principle—that is, each pair of characteristics being categorized stands alone.
Random Tree: The branches of the dataset create a tree when the full dataset is classified. The reason this approach is named random trees is that it creates several decision trees by classifying the dataset multiple times using a random sub selection of training pixels. A vote is cast by each tree to determine the winner.
Randon Forest: One well-known machine learning method that is a part of the supervised learning approach is Random Forest. It may be applied to ML issues involving both classification and regression. Its foundation is the idea of ensemble learning, which is the act of merging several classifiers to solve a challenging issue and enhance the model's functionality.
10-fold cross-validation is the accepted approach for assessing various machine learning strategies. Ten equal subsets of the dataset were created, one for training and one for testing. This process was maintained until every subset had been put through testing. As seen in Figures 5–8, we used the 10-fold cross-validation test to assess how well various classifiers performed. The "Classifier Output" tab in WEKA then displays the predictions for each test case. To determine the optimal classification strategy through comparison, many models, preprocessing techniques, and feature selection strategies were learned using WEKA machine learning software.

## 4. RESULT ANALYSIS

Several research communities worldwide have produced excellent work on DKD prediction. Several subcategories of this study have been mentioned below:

| Reference No. | Study by | Technique Used | Obtain Accuracy |
|---|---|---|---|
| [1] | Boukenze, B | SVM | 62.5% |
| [2] | Mishra | KNN | 98.78% |
| [3] | Sobrinth | J48 | 95.0% |
| [4] | Khanna and Simon | ANN | 88.57% |
| [5] | Senaet | SVM, decision tree | 96.67% |

The following phases comprise the system's general methodology:
1. The UCI machine learning repository provided the dataset.
2. Preprocessing is the process of removing noise and outliers from the dataset, replacing missing values, or both.
3. Choose the best feature subset and use feature selection techniques to reduce the dimensionality of the feature.
4. Step 3 of the data reduction procedure is repeated until high-performance accuracy is achieved.
5. Model building with several classification methods, such as Random Tree, Random Forest, and KNN.

## 5. CONCLUSION AND FUTURE SCOPE

This work aims to explore how various diabetic kidney disease classification algorithms, feature selection techniques, and data mining are applied to assess and forecast various diabetes-related kidney diseases.
The experimental outcome will show us how feature selection, classification, and data mining techniques have been applied to identify, analyze, and forecast diabetic kidney illnesses. To increase the algorithm's performance accuracy, several researchers have experimented with various classification algorithms, including KNN, ANN, Naïve Bays, SVM, Decision trees (J48, C4.5), and feature selection.
We are now focusing on improving the accuracy of prediction systems in the future by combining several classifier methods.
.

**References**

[1] B. Boukenze, A. Haqiq, and H. Mousannif, "Predicting Chronic Kidney Failure Disease Using Data Mining Techniques," in Advances in Ubiquitous Networking 2, Springer, Singapore, 2017, pp. 701–712.

[2] S. Mishra, P. Chaudhury, B. K. Mishra, and H. K. Tripathy, "An Implementation of Feature Ranking Using Machine Learning Techniques for Diabetes Disease Prediction," in Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, New York, NY, USA, 2016, p. 42:1–42:3.

[3] A. Sobrinho, A. C. M. D. S. Queiroz, L. D. Da Silva, E. D. B. Costa, M. E. Pinheiro, and A. Perkusich, "Computeraided diagnosis of chronic kidney disease in developing Countries: a comparative analysis of machine learning techniques," IEEE Access, vol. 8, pp. 25407–25419, 2020.

[4] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, no. 4, pp. 432–439, 2021.

[5] E. M. Senan, M. H. Al-Adhaileh, F. W. Alsaade et al., "Diagnosis of chronic kidney disease using Effective classification algorithms and recursive feature Elimination techniques," Journal of Healthcare Engineering, vol. 2021, p. 1004767, 2021.

[6] S. Zeynu, A. Professor, and S. Patil, "Survey on prediction of chronic kidney disease using data mining classificationtechniques and feature selection," Shruti Patil, vol. 118, no. 8, pp. 149–156, 2018.

[7] Y. Jian, M. Pasquier, A. Sagahyroon, and F. Aloul, "A machine learning Approach to predicting diabetes complications," Healthcare, vol. 9, no. 12, 2021.

[8] K. R. Tan, J. J. B. Seng, Y. H. Kwan et al., "Evaluation of machine learning methods developed for prediction of diabetes complications: a systematic review," Journal of Diabetes Science and Technology, p. 193229682110569, 2021.

[9] V. Rodriguez-Romero, R. F. Bergstrom, B. S. Decker, G. Lahu, M. Vakilynejad, and R. R. Bies, "Prediction of nephropathy in type 2 diabetes: an analysis of the ACCORD trial applying machine learning techniques," Clinical and Translational Science, vol. 12, no. 5, pp. 519–528, 2019.

[10] V. Mishra, C. Samuel, and S. S.K, "Use of Machine Learning to Predict the Onset of Diabetes," Int. J. Recent Adv. Mech. Eng., vol. 4, no. 2, pp. 9–14, 2015, doi: 10.14810/ijmech.2015.4202.

[11] A. Gnana, E. Leavline, and B. Baig, "Diabetes Prediction Using Medical Data," J. Comput. Intell. Bioinforma., vol. 10, no. January, pp. 1–8, 2017.

[12] N. H. Chowdhury, M. B. Reaz, F. Haque et al., "Performance analysis of Conventional machine learning algorithms for identification of chronic kidney disease in type 1 diabetes mellitus patients," Diagnostics, vol. 11, no. 12, 2021

[13] K. Chandel, V. Kunwar, S. Sabitha, T. Choudhury, and S. Mukherjee, "A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques," CSI Trans. ICT, vol. 4, no. 2–4, pp. 313–319, Dec. 2016.

[14] P. Panwong and N. Iam-On, "Predicting transitional interval of kidney disease stages 3 to 5 using data mining method," in 2016 Second Asian Conference on Defence Technology (ACDT), 2016, pp. 145–150

[15] Allen A, Iqbal Z, Green-Saxena A, et al. Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus. BMJ Open Diab Res Care 2022;10:e002560. doi:10.1136/ bmjdrc-2021-002560

[16] Gonzalez CD, Carro Negueruela MP, Nicora Santamarina C, Resnik R, Vaccaro MI. Autophagy Dysregulation in Diabetic Kidney Disease: From Pathophysiology to Pharmacological Interventions. Cells. 2021 Sep 21;10(9):2497. doi: 10.3390/cells10092497. PMID: 34572148; PMCID: PMC84698.

[17] Zeynu, Sirage & Patil, Shruti. (2018). Survey on prediction of chronic kidney disease using data mining classification techniques and feature selection. International Journal of Pure and Applied Mathematics. 118. 149-155.

BIOGRAPHY

**Shweta Yadu** is currently pursuing her MTech degree in Computer Science and Engineering from Raipur Institute of Technology affiliated to Chhattisgarh Swami Vivekanand Technical University, Bhilai, Chhattisgarh, India. She has completed BE in Computer Science & Engineering from ITGGU, Bilaspur Chhattisgarh, India in 2010. Her research interests are Data Science, Artificial intelligence, and Machine Learning.

**Mahadev Bag** is currently working as an HOD in the Computer Science and Engineering Department at Raipur Institute of Technology, affiliated to Chhattisgarh Swami Vivekanand Technical University, Bhilai, Chhattisgarh, India. He has 18 years of experience in teaching. He has published more than 12 research papers in Scopus and SCI. Research interests are machine learning and data science.