2025 7th International Conference on Sustainable Technologies for Industry 5.0 (STI), 18-19 December, Dhaka

Chronic Kidney Disease Prediction Using Machine Learning

Sayla Srabony*, Abir Hasan*, and Md Anwarul Islam *
Dept of Computer Science and Engineering, Netrokona University, Netrokona, Bangladesh
Email: shayla.srabony@gmail.com, abir.neu@gmail.com, anwar.abir@neu.ac.bd

***Abstract*—People today try to stay healthy, but they don't pay attention to it until they get sick because they're so busy. On the other hand, chronic kidney disease (CKD) is a silent and progressive disease that sometimes does not have symptoms. So, it is challenging to guess, find, and stop this disease, which can hurt your health in the long run. However, machine learning (ML) can be used to make predictions and analyze this disease. In this paper, we suggest nine ML methods: gradient boosting, naive Bayes, AdaBoost, kernel SVM, random forest, KNN, logistic regression, SVM, and decision tree. We got the data from Kaggle.com [1]. We used the models to see how accurate each one was. In addition, this study examined how well these models worked in comparison to each other. The decision tree model helps us predict kidney disease more accurately than before, with an accuracy rate of 99.02%.**

***Index Terms*—Chronic kidney disease (CKD), Machine learning technique, classification algorithm, prediction of kidney disease**

I. Introduction

The kidneys are incredibly important because they remove waste from the blood and send it out through urine. This helps keep the body balanced inside. The progressive kidney disease known as chronic kidney disease (CKD) affects kidney function over time. It affects millions of people throughout the world and puts a lot of burden on healthcare systems because it typically does not show any indicators until it is too late. End-stage renal disease (ESRD), a serious disease that requires dialysis or a kidney transplant to survive, can develop from chronic kidney disease (CKD) if treatment is not received. Finding CKD early is crucial to minimize significant problems, reduce medical expenses, and make patient lives better overall [2]. Serum creatinine levels, glomerular filtration rate (GFR), and urine albumin concentration are some of the most essential lab tests that use classic diagnostic approaches [3].These are standard ways to find CKD, although they don't function very well in the early stages. The kidneys also make hormones that affect how the other organs in the body work.

There are five stages of CKD, which are as follows:

- Stage 1: GFR is normal or high (GFR >90 mL/min)
- Stage 2: Mild CKD (GFR = 60 to 89 mL/min)
- Stage 3: Moderate CKD (GFR = 30–59 mL/min)
- Stage 4: Severe CKD (GFR = 15–29 mL/min)
- Last stage: End Stage (GFR <15 mL/min)

Chronic kidney disease is a disorder that slowly makes the kidneys stop working over time. About 14% of people are thought to have chronic renal disease [4] . More people die from chronic renal illness than from breast or prostate cancer together. But more than two million people suffer renal failure and need dialysis or a new kidney. The kidneys make a hormone that controls a lot of things in the body, like making red blood cells, blood pressure, and calcium metabolism. Age, sex, race, and creatinine levels are only a few of the things that can alter eGFR [5].

Machine learning (ML), on the other hand, is a novel and promising technique to discover diseases early and precisely. It uses more and more clinical datasets and novel ways of doing things with computers.

ML is highly useful in healthcare since it can detect and sort a lot of various medical problems, such as heart disease, breast cancer, stroke, and renal disease [6], [7], [8]. ML looks at high-dimensional, complicated datasets, such as electronic medical records (EMRs), using algorithms [9] to help plan early, low-cost interventions and generate intelligent predictions. This is a useful technique to improve the diagnosis of CKD.

The major purpose of this project is to use machine learning to uncover patterns in raw medical data that can assist in predicting CKD. We want to make diagnosis more accurate and faster so that we can act sooner and better control the risk factors that come with chronic renal disease.

II. Literature review

Researchers are particularly interested in predicting CKD. Most of them use ML algorithms and conventional techniques to improve classification accuracy. Moreover, researchers employ a variety of hybrid methods and algorithms to categorize kidney diseases from secondary sources found on UCI and Kaggle.

Almustafa et al. [10] classified a CKD dataset using various classifiers. They developed techniques to classify the CKD dataset using some classifiers, including Decision Tree (DT), Random Tree (RT), Statistical Gradient Descent (SGD), J48, KNN, and Bayesian Naive. Furthermore, they developed a prediction model based on the selection of features that can accurately predict the presence of CKD symptoms. The results showed that the J48 and decision tree techniques outperformed other techniques with an accuracy rate of 99%.

Yashfi et al. [11] developed a method for CKD risk analysis by combining data from Khulna City Medical College with data from the UCI Machine Learning Repository of 455 patients. Random Forest (RF) and Artificial Neural Network

(ANN) models are trained by the 10-fold cross-validation method. According to their results, the RF model gave 97.12% accuracy, and the ANN achieved a slightly lower accuracy of 94.5%

S. Gopika et al. [12] developed a technique for predicting CKD using cluster analysis. The main objective of developing their technique is to detect kidney failure using the clustering method. The fuzzy C algorithm performed well in the experiment, as it got 89% of the answers correct.

Siddheshwar Tekale et al. [13] analyzed 14 different features related to CKD patients and provided the expected accuracy for several ML processes such as decision trees and support vector machines. The study revealed that the decision tree algorithm provides 91.75% accuracy and the SVM provides 96.75% accuracy. The prediction process is less time-consuming.

Punia RC et al. [14] showed that proper feature selection is crucial in CKD detection. They worked with 400 samples and 24 features and selected important features using KNN, ANN, SVM and Naïve Bayes algorithms, as well as Recursive Feature Elimination (RFE) and the Chi-square test. When the logistic regression model was optimized with chi-square-based features, it gave a maximum accuracy of 98.75%.

Sonon and Daniel (2021) [15] studied how well different machine learning (ML) methods work in finding and predicting the advancement of chronic kidney disease (CKD). They classified patients into CKD and non-CKD categories using random forest, SVM, logistic regression, and naive Bayes models using clinical datasets. The study emphasized missing value imputation and feature normalization as part of data processing. According to the results, random forest and SVM achieved about 98% accuracy, which is also consistent with other contemporary studies. The study shows that ML technology can play an important role in improving the early detection and treatment planning of CKD patients.

TABLE I: Summary of some related work

Author	Technique Applied	Accuracy	Limitation
Tekale et al. (2023) [13]	SVM, DT	SVM: 91.75%, DT: 96.75%	Dataset size small; intensity predictions not included; only categorical outcomes used.
Yashfi et al. [11]	ANN, RF	RF: 97.12%, ANN: 94.5%	Small dataset; no prediction stages included.
Punia R.C. et al. (2022) [14]	SVM, ANN, NB, KNN, LR (with Chi-Square, RFE)	SVM: 97.5%, ANN: 65%, KNN: 66.25%, NB: 95%, LR: 97.5%	Small dataset; missing prediction stages; needs accuracy improvement.
Almustafa (2021) [16]	DT, RT, SGD, J48, k-NN, Bayesian Naive	DT: 99%, RT: 95.5%, SGD: 98.25%, J48: 99%, k-NN: 95.75%, BN: 95%	No data scaling; limited cross-validation; lacks time complexity analysis.
Sawhney et al. (2023) [17]	ANN, LR, SVM	ANN: 100%, LR: 96%, SVM: 82%	Missing value handling and scaling were skipped; no time complexity evaluation.

III. Methodology

Machine learning is a type of artificial intelligence that learners use without having to process the data. It focuses on creating computer programmers who can make changes in response to new information. It can be either supervised or unsupervised [18]. The most important thing is to combine the right parts to make a framework that works for the right reason. Some of the work in this area includes multidimensional and multi-class classifications, predictive clustering, and parametric modeling.[19].

A. Data collection

The dataset used in this study was collected from the open-source 'Chronic Kidney Disease Prediction'[1]. on the Kaggle online platform. The dataset includes a total of 400 samples and specifies 24 different features. Of these, 23 features are predictive, and 1 is a target variable. Of these 24 features, 11 can be classified as numerical and 13 as categorical variables. The categorical target variable determines whether a patient has CKD (Chronic Kidney Disease). This category has two values—'ckd', which means the disease is detected, and 'notckd', which indicates the disease is not present. In this dataset, 250 samples are CKD, and 150 samples are not CKD.non-CKD. Table II depicts the attributes found in the dataset. Some of the notable features and symptoms used to detect kidney-related diseases in the dataset are hypertension, coronary artery disease, edema, diabetes mellitus, anemia, specific gravity, hunger status, age, blood sugar and albumin levels, red blood cell status, presence of pus cells and bacteria, clumps, blood urea, potassium levels, hemoglobin, serum creatinine, sodium, white and red blood cell counts, and packed cell volume. outcome. Fig. 1 describes the overall scenario of the CKD dataset after PCA, in addition to the original CKD dataset. This study included both numerical and categorical features in the analysis to enhance the accuracy and effectiveness of disease detection during the prediction process.

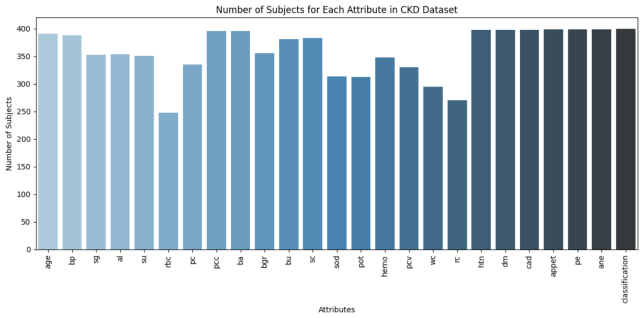

Fig. 1: CKD Dataset

B. Data Preprocessing

According to the proposed method, preprocessing starts after the data collection is completed. The first step in data pre-preparation was to identify and handle missing values.

TABLE II: In-depth descriptions of each feature in the main CKD dataset

Attribute	Meaning	Category	Scale/Unit	Missing
age	Age	Numerical	Years	9
bp	Blood pressure	Numerical	mm/Hg	12
sg	Specific gravity	Nominal	1.005 to 1.025	47
al	Albumin	Nominal	0 to 5	46
su	Sugar	Nominal	0 to 5	49
rbc	Red blood cells	Nominal	Abnormal, Normal	152
pc	Pus cell	Nominal	Abnormal, Normal	65
pcc	Pus cell clumps	Nominal	Not present, Present	4
ba	Bacteria	Nominal	Not present, Present	4
bgr	Blood glucose random	Numerical	mgs/dl	44
bu	Blood urea	Numerical	mgs/dl	19
sc	Serum creatinine	Numerical	mgs/dl	17
sod	Sodium	Numerical	mEq/L	87
pot	Potassium	Numerical	mEq/L	88
hemo	Hemoglobin	Numerical	gms	52
pcv	Packed cell volume	Numerical	Pcv	71
wc	White blood cell count	Numerical	cells/cumm	106
rc	Red blood cell count	Nominal	millions/cmm	131
htn	Hypertension	Nominal	No, Yes	2
dm	Diabetes mellitus	Nominal	No, Yes	2
cad	Coronary artery disease	Nominal	No, Yes	2
appet	Appetite	Nominal	Poor, Good	1
pe	Pedal edema	Nominal	No, Yes	1
ane	Anemia	Nominal	No, Yes	1
classification	Class label	Nominal	CKD, Not CKD	0

The remaining missing values were handled using appropriate techniques. Missing values were filled in using the mean or median, depending on the data distribution for numeric variables. For categorical variables, we used the mode, which represents the most frequent category. Little's MCAR (missing completely at random) test was performed to assess the randomness of missing data. The results helped determine whether the missing data were randomly distributed or exhibited a pattern. If the data were found to be MCAR, simple imputation methods were justified. Otherwise, more careful imputation techniques or model-based methods were considered. After handling missing values, the dataset was transformed into a machine-readable format. Some numerical columns were originally stored as object types, so these were converted to the float64 data type to allow for numerical analysis and model compatibility. Categorical attributes were encoded using label encoding. This method assigns a unique integer to each category in a feature. For instance, a feature like "Appetite" with values such as "good" and "poor" would be encoded as 0 and 1, respectively. This step was essential for algorithms that require numerical input.

c. Validation Process

Choosing an effective validation strategy is crucial when working with a particular dataset. The classification algorithms used in this study include Gradient Boosting, Naïve Bayes, AdaBoost, Kernel SVM, Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), and Decision Tree [20]. We followed the hold-out validation approach to train and evaluate the dataset. This method assesses the accuracy of kidney disease predictions and model performance. Hold-out validation is generally considered a suitable approach for large-scale datasets, as it provides reliable results with low complexity. In this study, we assigned the dataset 30% of the data was allocated for model testing, while the remaining 70% was used for model training. This validation technique evaluated the performance of each machine learning algorithm and calculated performance indicators such as accuracy, recall, and F1-score, which helped in selecting the best model. The overall picture of the proposed methodology workflow is presented in Fig. 2.

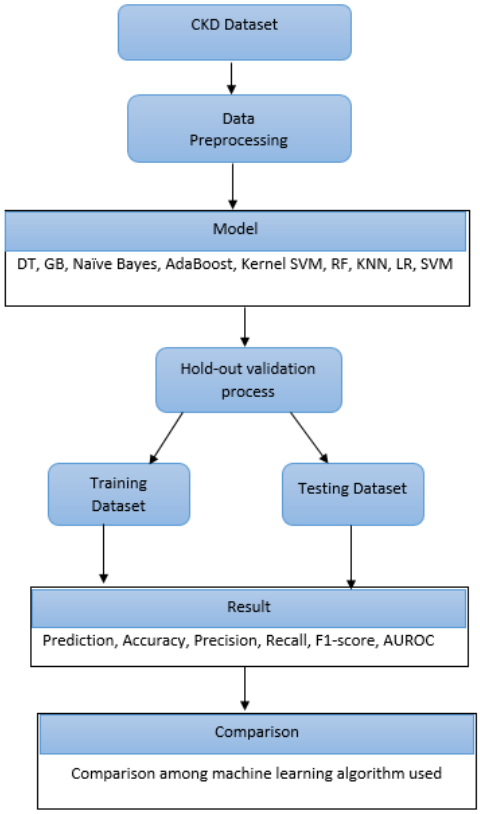

Fig. 2: Proposed methodology workflow

IV. Result Analysis

Table III We compare nine machine learning methods using several performance metrics, such as area under the accuracy, precision, recall, F1-score, and ROC curve (AUC). Although accuracy is still a commonly used performance indicator, it is insufficient on its own to evaluate a model's performance in its entirety. AUC is a crucial metric since it analyzes the

true positive and false positive rates across various probability thresholds to gauge a model's capacity to discriminate between classes.

TABLE III: Comparison of Nine Machine Learning Algorithms

Machine Learning Model	Accuracy	Precision	Recall	F1 Score	AUROC
Random Forest Algorithm	0.975%	0.95%	1.00%	0.98%	1.00
AdaBoost Classifier	0.967%	0.95%	0.98%	0.97%	1.00
Gradient Boosting	0.975%	0.95%	1.00%	0.98%	1.00
Logistic Regression	0.925%	0.93%	0.92%	0.93%	0.998
Naïve Bayes	0.950%	1.00%	0.90%	0.95%	1.00
KNN	0.933%	0.95%	0.92%	0.93%	0.754
SVM	0.942%	0.95%	0.94%	0.94%	0.997
Kernel SVM	0.900%	0.89%	0.92%	0.90%	0.674
Decision Tree	0.992%	1.00%	0.98%	0.99%	0.993

Fig. 3 shows that Decision Tree has the highest accuracy of 99.2%, while kernel SVM hast the lowest accuracy of 90%. The AdaBoost Classifier, Gradient Boosting, Logistic Regression, Naïve Bayes, KNN, SVM accuracy are 96.7%, 97.5%, 92.5%, 95%, 93.3%, 94.2% respectively.

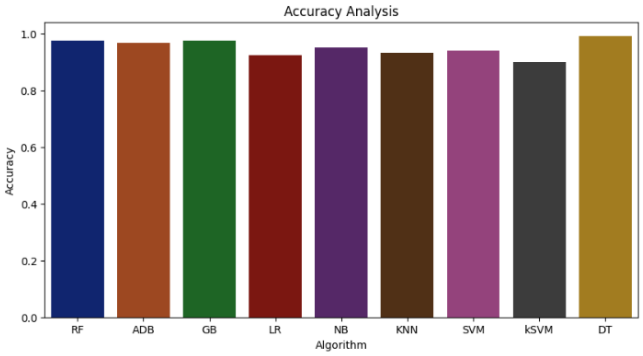

Fig. 3: Accuracy Analysis using Machine Learning models

Fig. 4 shows the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUROC) for the several machine learning models that were tested on the CKD detection dataset. The ROC curve shows the True Positive Rate (TPR) and the False Positive Rate (FPR), and the AUROC score shows how well the model can tell the difference between the two groups (CKD and non-CKD). A higher AUROC score means that the model works better. A perfect classifier would have an AUROC of 1.0. The ROC curve shows that logistic regression (AUROC = 0.998), gradient boosting (AUROC = 1.0), Naïve Bayes (AUROC = 1.0), AdaBoost (AUROC = 1.0), random forest (AUROC = 1.0), and decision tree (AUROC = 1.0) all have perfect discriminative ability, which makes them great choices for classifying CKD. Linear SVM (AUROC = 0.997) also does a great job, but it's a little behind the other models. KNN (AUROC = 0.754) and Kernel SVM (AUROC = 0.668) have much lower AUROC scores, which means they are not as good at telling the classes apart. The Kernel SVM does especially badly since it only does a little better than random prediction. The random prediction line (AUROC = 0.5) is there to show what a classifier that produces random predictions looks like. It is the starting point for comparison.

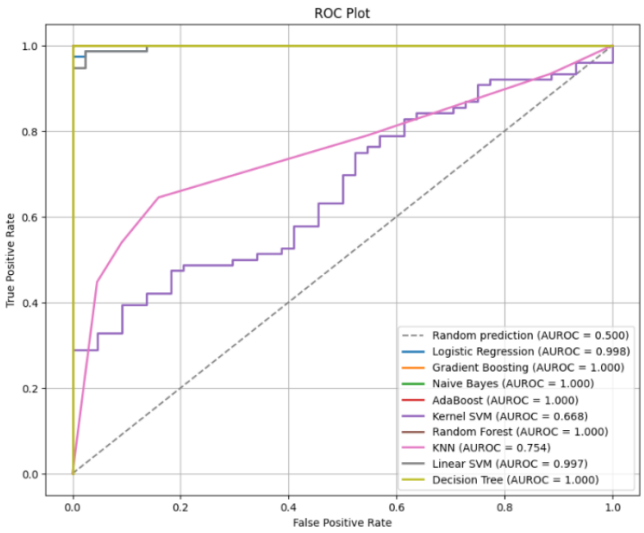

Fig. 4: ROC Curve

V. Conclusion and Future Work.

Chronic Kidney Disease (CKD) presents significant challenges in the field of healthcare, particularly because early diagnosis can play a crucial role in reducing patient suffering. In this study, a globally recognized and reliable CKD dataset was used to explore predictive methods for early detection. Our proposed model achieved a high accuracy of 99.02%, indicating its potential in identifying CKD at an early stage. To develop an effective predictive model, various machine learning algorithms were utilized, including Gradient Boosting, Logistic Regression, Support Vector Machine, Naïve Bayes, AdaBoost, Kernel SVM, Random Forest, K-Nearest Neighbors (KNN), and Decision Tree. Among these techniques, the decision tree classifier delivered superior results in terms of accuracy, precision, and F1-score when compared to the others. The outcomes of this study indicate that machine learning approaches have strong potential in the medical field, particularly for early detection and monitoring of chronic kidney disease. These models can play a critical role in proactive healthcare, supporting timely We will focus on the diagnosis and regular observation of patients who are at risk. In future work, we plan to use a larger, more representative dataset to enhance the model's generalization ability and allow for the detection of CKD at various stages of progression.

ACKNOWLEDGMENT

I would like to express my gratitude to my supervisor, Md. Anwarul Islam, who guided us throughout this project.

REFERENCES

[1] M. Iqbal, "Chronic kidney disease dataset," https://www.kaggle.com/datasets/mansoordaku/ckdisease, 2020, accessed: 2025-07-08.

[2] D. B, V. K. R. Rao, D. N. Rampure, P. Prajwal, D. G. Gowda *et al.*, "Early prediction of chronic kidney disease by using machine learning techniques," *American Journal of Computer Science and Engineering Survey*, vol. 8, no. 2, p. 07, 2020.

[3] M. T. James, B. R. Hemmelgarn, N. Wiebe, N. Pannu, B. J. Manns, S. W. Klarenbach, and M. Tonelli, "Glomerular filtration rate, proteinuria, and the incidence and consequences of acute kidney injury: A cohort study," *The Lancet*, vol. 376, no. 9758, pp. 2096–2103, 2010. [Online]. Available: https://doi.org/10.1016/S0140-6736(10)61271-8

[4] U. Ekanayake and D. Herath, "Chronic kidney disease prediction using machine learning methods," in *Proceedings of the 2020 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2020, pp. 260–265.

[5] National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), "Your kidneys & how they work," https://www.niddk.nih.gov/health-information/kidneydisease/kidneys-how-they-work, n.d., accessed: 2025-07-09.

[6] R. C. Das, M. C. Das, M. A. Hossain, M. A. Rahman, M. H. Hossen, and R. Hasan, "Heart disease detection using ml," in *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, 2023, preprint.

[7] S.-J. Sammut *et al.*, "Multi-omic machine learning predictor of breast cancer therapy response," *Nature*, vol. 601, no. 7894, pp. 623–629, 2022.

[8] R. S. Chaulagain, D. Kim, N. Jayasena, A. Bhatele, and D. Tiwari, "Achieving the performance of global adaptive routing using local information on dragonfly through deep learning," in *ACM/IEEE SC Technical Poster*, 2020.

[9] M. G. Bastos and G. M. Kirsztajn, "Chronic kidney disease: importance of early diagnosis, immediate referral and structured interdisciplinary approach to improve outcomes in patients not yet on dialysis," *Jornal brasileiro de nefrologia*, vol. 33, no. 1, p. 93—108, March 2011. [Online]. Available: http://europepmc.org/abstract/MED/21541469

[10] K. M. Almustafa, "Prediction of chronic kidney disease using different classification algorithms," *Informatics in Medicine Unlocked*, vol. 24, p. 100631, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914821001210

[11] S. Yashfi, M. Islam, Pritilata, N. Sakib, T. Islam, M. Shahbaaz, and S. Pantho, "Risk prediction of chronic kidney disease using machine learning algorithms," 07 2020, pp. 1–5.

[12] G. S and V. Muthuraman, "Survey on prediction of kidney disease by using data mining techniques," *IJARCCE*, vol. 6, pp. 198–201, 01 2017.

[13] S. Tekale, P. Shingavi, and S. Wandhekar, "Prediction of chronic kidney disease using machine learning algorithm," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 7, no. 10, pp. 92–96, 2018.

[14] R. C. Poonia, M. K. Gupta, I. Abunadi, A. A. Albraikan, F. N. Al-Wesabi, M. A. Hamza, and T. B, "Intelligent diagnostic prediction and classification models for detection of kidney disease," *Healthcare*, vol. 10, no. 2, 2022. [Online]. Available: https://www.mdpi.com/2227-9032/10/2/371

[15] N. Sonone and A. Daniel, "Early prediction and progrssion of chronic kidney disease using machine lerning techniques," in *2024 2nd International Conference on Networking and Communications (ICNWC)*, 2024, pp. 1–6.

[16] M. Almasoud and T. E. Ward, "Detection of chronic kidney disease using machine learning algorithms with the least number of predictors," *International Journal of Soft Computing and Its Applications*, vol. 10, no. 8, 2019.

[17] R. Sawhney, M. Sharma, M. Alazab, M. Tariq, M. Shoaib, and M. N. Mohd, "A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease," *Decision Analytics Journal*, vol. 6, 2023.

[18] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, "When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts," *BMC Medical Research Methodology*, vol. 17, no. 1, p. 162, 2017. [Online]. Available: https://doi.org/10.1186/s12874-017-0442-1

[19] S. Arasu and R. Thirumalaiselvi, "Review of chronic kidney disease based on data mining techniques," *International Journal of Applied Engineering Research*, vol. 12, pp. 13 498–13 505, 2017.

[20] A. Alnuaimi and T. Albaldawi, "An overview of machine learning classification techniques," *BIO Web of Conferences*, vol. 97, p. 00133, Apr. 2024.