Chronic Kidney Desease (CKD) Prediction Using Random Forest and XGBoost Machine Learning Models

Nadine Beatricia
Department of Computer Science, School of Computer Science, Bina Nusantara University
Jakarta, Indonesia
nadine.beatricia@binus.ac.id

Antony Wijaya
Department of Computer Science, School of Computer Science, Bina Nusantara University
Jakarta, Indonesia
antony.wijaya@binus.ac.id

Laurentius Brandon Vikario
Department of Computer Science, School of Computer Science, Bina Nusantara University
Jakarta, Indonesia
laurentius.vikario@binus.ac.id

Annabelle Fevriane
Department of Computer Science, School of Computer Science, Bina Nusantara University
Jakarta, Indonesia
annabelle.fevriane@binus.ac.id

Chingtya Tjoeng
Department of Computer Science, School of Computer Science, Bina Nusantara University
Jakarta, Indonesia
chingtya.tjoeng@binus.ac.id

Samantha Niandra
Department of Computer Science, School of Computer Science, Bina Nusantara University
Jakarta, Indonesia
samantha.niandra@binus.ac.id

Jessica Debora
Department of Computer Science, School of Computer Science, Bina Nusantara University
Jakarta, Indonesia
jessica.debora@binus.ac.id

***Abstract*— Chronic Kidney Disease (CKD) is a long-term condition that severely impacts a patient's physical and emotional well-being. This research applies supervised machine learning techniques—Random Forest and XGBoost—to predict whether a patient has CKD or not using clinical features from a CKD dataset. The dataset undergoes preprocessing, including handling missing values, normalizing numerical attributes, and encoding categorical variables. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied. Models are trained and evaluated using a split dataset approach, consisting of training, validation, and test sets. Hyperparameter tuning is performed using RandomizedSearchCV to optimize each model. The expected outcome is to determine the most accurate predictive model and identify key clinical attributes that influence CKD patients. This data-driven approach aims to enhance healthcare decision-making and improve the overall management of CKD.**

Keywords*— *Chronic Kidney Disease, Machine Learning, Random Forest, XGBoost, Regression, Clinical Prediction, Healthcare Analytics

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a progressive condition that significantly impacts patients' overall well-being and daily functioning. CKD is characterized by kidney damage that is measured by the Glomerular Filtration Rate (GFR) and is frequently caused by diabetes, hypertension, as well as other conditions. Cases of CKD significantly grew by 89% from 1990 to 2016 with death increased by 98% [1].

Early detection of CKD is critical in preventing further complications, such as kidney failure and cardiovascular diseases. However, CKD is often asymptomatic in its early stages, making diagnosis challenging through conventional methods alone. Therefore, there is a pressing need for predictive models that can assist in early identification and risk stratification of CKD.

Artificial intelligence has already been used in the healthcare field for medical imaging, natural language processing, and genomics [2]. These technologies allow the analysis of large, complex datasets and the development of predictive models that support data-driven decision-making.

In this study, two machine learning models — Random Forest (RF) and XGBoost — are applied to predict the presence of CKD using clinical features from the publicly available "Chronic Kidney Disease" dataset on Kaggle. RF is a popular technique for its predictive approach that has high prediction accuracy. On the other hand, XGBoost is known for its high efficiency performance, its approach helps reduce bias and variance. Both are widely used in medical research and healthcare analysis, making it suitable for predicting patients diagnose of CKD in this study.

This paper aims to compare the performance of these two models in accurately classifying patients with CKD. Through this comparison, we seek to identify the most reliable approach for supporting early diagnosis and aiding healthcare providers in delivering timely treatment interventions.

II. LITERATURE REVIEW

A. *Overview of CKD*

CKD is a progressive kidney failure and will be worsened through time. CKD affects both the function and structure of the kidney. Since CKD is irreversible, CKD patients have a higher risk to experience complications especially cardiovascular-related. The main cause of CKD includes diabetes, hypertension, chronic glomerulonephritis, chronic pyelonephritis, chronic use of anti-inflammatory medication, autoimmune diseases, polycystic kidney disease, Alport disease, congenital malformations, and prolonged acute renal disease [3]. CKD is categorized into five stages based on the Glomerular Filtration Rate (GFR), with early stages often being asymptomatic. As the disease progresses, patients face increased risks of complications, and in the final stage (End-Stage Renal Disease), dialysis or kidney transplantation becomes necessary.

Current diagnostic approaches rely on laboratory tests including serum creatinine levels, estimated GFR (eGFR), and urine albumin levels. However, these methods often detect CKD only after substantial kidney damage has occurred. As a result, there is increasing interest in utilizing advanced computational methods, such as machine learning, to enhance early prediction and classification of CKD based on a broader range of clinical features [15].

B. *Medical Approaches to Chronic Kidney Disease Diagnosis*

CKD is traditionally diagnosed using a mix of clinical evaluations and lab tests. Doctors usually start by running a urinalysis to check for protein or blood in the urine, which can be early signs of kidney problems. Blood tests are also commonly used, especially to measure creatinine levels, which help calculate the estimated Glomerular Filtration Rate (eGFR)—a key indicator of kidney function. Other tests might include Blood Urea Nitrogen (BUN) levels, which show how well the kidneys are removing waste from the blood. In some cases, imaging tests like ultrasound are used to examine the size and shape of the kidneys, and if needed, a kidney biopsy might be performed to get a closer look at the type of damage present.

Even though these traditional methods are widely used, they do have limitations. One of the biggest challenges is that CKD often doesn't show clear symptoms until it's in a more advanced stage, which can lead to late diagnosis. These methods also rely heavily on the doctor's interpretation, which can sometimes lead to inconsistent results. Most importantly, traditional approaches are more focused on diagnosing the current condition rather than predicting how the disease might progress. With the rise of digital health technologies and access to large amounts of medical data, these limitations are pushing healthcare professionals to explore newer, more predictive methods—like machine learning—which can help catch CKD earlier and with more accuracy [3][16].

C. *Application of Machine Learning in Healthcare*

Machine Learning (ML) and the medical field are two opposites fields. Recently the developments of ML have opened new opportunities to collaborate. To process medical data, identify health patterns, and make predictions, the computer needs ML. The components that are needed in the process are models, training, and evaluation [20]. RF, Support Vector Machine, Neural Networks, and XGBoost are machine learning models that are frequently used in the medical field [4]. Therefore, predicting disease progression is where machine learning plays a crucial role. In example, machine learning can predict the risk of diabetes in patients by analyzing attributes that affects diabetes. ML surpasses traditional methods by accurately identifying risk factors and predicting diseases with greater precision. This capability allows for early intervention, personalized treatment, and improved patient outcomes [5][17].

D. *Previous Studies on Machine Learning for Chronic Kidney Disease*

In recent years, the use of machine learning models, particularly Random Forest and XGBoost,

has been widely explored in the prediction of CKD. These models are favored for their high accuracy, robustness, and ability to handle complex and heterogeneous clinical data.

Random Forest has proven effective in classifying CKD by identifying key predictive features. In a study by Mendapara Shreyas, Ghate Anagha, Guha Debargha, and Natarajan Anjana, Random Forest was applied to gene expression data for CKD classification, achieving an accuracy of 94% and an AUC of 0.990 after extensive cross-validation and external testing. This study underscores the model's strength in biomedical contexts, especially where high-dimensional data is involved [11][14].

On the other hand, XGBoost has gained popularity due to its performance efficiency and reduced bias-variance trade-off. Li Yuhan, Kumar Manish, and Nguyen Tuan used XGBoost to predict the progression of CKD to end-stage renal disease (ESRD) using administrative claims data. The model demonstrated strong predictive power, and with the integration of SHAP (SHapley Additive exPlanations), it offered high interpretability by highlighting the most influential features in the decision-making process [12][18].

Additionally, hybrid approaches have shown promising results. Wang Haoyu, Liu Jing, and Zhao Fang combined Random Forest and XGBoost with the SMOTE oversampling technique to address class imbalance in clinical datasets. This ensemble model achieved an AUC of 0.99 in predicting CKD progression, illustrating the potential of combining multiple models to enhance prediction reliability [13][19].

These studies collectively indicate that both Random Forest and XGBoost are effective and reliable tools in the early prediction of CKD, offering a balance between accuracy and interpretability, and making them highly suitable for deployment in clinical decision-support systems.

III. METHODOLOGY

The algorithms used in this project consist of RF and XGBoost machine learning model. Each model is trained and tested to classify data into distinct categories.

A. Datasets and Data Pre-processing

In this study, we utilize a dataset related to CKD to train and also evaluate the machine learning models. The datasets consist of various medical attributes, including laboratory test results and clinical indicators, which are used to predict continuous values associated with kidney function.

The dataset contains patient records with multiple features such as blood pressure, blood glucose levels, hemoglobin, and serum creatinine, among others. These attributes play a crucial role in assessing kidney health and disease progression. The data is preprocessed to handle missing values, normalize numerical attributes, and encode categorical variables where necessary. Table 1 describes the attributes of the dataset that has correlation with 'Diagnosis' for more than 0.02.

TABLE I. ATTRIBUTES DESCRIPTION

Attribute Name	Description
SocioeconomicStatus	The socioeconomic status of the patients
EducationalLevel	The education level of the patients
BMI	Body Mass Index of the patients, ranging from 15 to 40.
Smoking	Smoking status
PhysicalActivity	Weekly physical activity in hours, ranging from 0 to 10.
DietQuality	Diet quality score, ranging from 0 to 10.
SleepQuality	Sleep quality score, ranging from 4 to 10.
FamilyHistoryKidneyDesease	Family history of kidney disease, where 0 indicates No and 1 indicates Yes.
FamilyHistoryHypertension	Family history of hypertension, where 0

	indicates No and 1 indicates Yes.
UrinaryTractInfections	History of urinary tract infections, where 0 indicates No and 1 indicates Yes.
SystolicBP	Systolic blood pressure, ranging from 90 to 180 mmHg.
DiastolicBP	Diastolic blood pressure, ranging from 60 to 120 mmHg.
FastingBloodSugar	Fasting blood sugar levels, ranging from 70 to 200 mg/dL.
HbA1c	Hemoglobin A1c levels, ranging from 4.0% to 10.0%.
SerumCreatinine	Serum creatinine levels, ranging from 0.5 to 5.0 mg/dL.
BUNLevels	Blood Urea Nitrogen levels, ranging from 5 to 50 mg/dL.
GFR	Glomerular Filtration Rate, ranging from 15 to 120 mL/min/1.73 m^2.
ProteinInUrine	Protein levels in urine, ranging from 0 to 5 g/day.
SerumElectrolytesSodium	Serum sodium levels, ranging from 135 to 145 mEq/L.
SerumElectrolytesPotassium	Serum potassium levels, ranging from 3.5 to 5.5 mEq/L.
HemoglobinLevels	Hemoglobin levels, ranging from 10 to 18 g/dL.
CholesterolTotal	Total cholesterol levels, ranging from 150 to 300 mg/dL.
CholesterolHDL	High-density lipoprotein cholesterol levels, ranging from 20 to 100 mg/dL.
Diuretics	Use of diuretics, where 0 indicates No and 1 indicates Yes.
Edema	Presence of edema, where 0 indicates No and 1 indicates Yes.
NauseaVomiting	Frequency of nausea and vomiting, ranging from 0 to 7 times per week.
MuscleCramps	Frequency of muscle cramps, ranging from 0 to 7 times per week.
Itching	Itching severity, ranging from 0 to 10.
QualityOfLifeScore	Quality of life score, ranging from 0 to 100.

We recognized an issue of class imbalance in the dataset between patients diagnosed having or not having CKD, this could affect the results of biased model performance. To address this issue, we used Synthetic Minority Over-sampling Technique (SMOTE).

For model training, the dataset is split into training, test, and validation sets. The training set is used to learn patterns in the data, the test set evaluates the model's performance on unseen

samples, while the validation set is used to ensure that the model does not only learn from the training and test data but also generalizes well by avoiding overfitting.

B. ML Model

1) Random Forest (RF)

RF is a supervised machine learning method that creates multiple decision trees for the prediction model, where each tree is trained on randomly selected subset of the available training data [6][7]. Random Forest is a versatile algorithm due to its simplicity and diversity, as can be applied to classification and regression problems. RF can handle datasets with both continuous and categorical variables [8].

In this study, the RF model is trained using the CKD dataset, with hyperparameters tuning performed to optimize its performance. The evaluation metrics used to assess the model's performance include accuracy, precision, recall, and F1-score, which are commonly used in classification analysis.

2) XGBoost

XGBoost is an optimized gradient boosting framework designed for efficiency and scalability. It improves upon traditional boosting by introducing sparsity-aware algorithms for handling missing data and weighted quantile sketching for efficient tree construction. Additionally, regularization techniques in the objective function help prevent overfitting, making XGBoost more robust than standard gradient boosting models. These optimizations, along with parallel computing, allow XGBoost to train faster and handle massive datasets efficiently. [9] Beyond efficiency, XGBoost maintains model consistency by iteratively improving tree predictions, unlike RF, which averages independent trees. It also performs feature selection dynamically and reduces computational cost while improving quality. Compared to RF, XGBoost better handles high-dimensional data and employs shrinkage and column subsampling to mitigate overfitting, which makes it highly effective for complex machine learning tasks [10].

C. Training Hyperparameters

To minimize the variables and chances of inconsistency in the results, all the machine learning models used in this paper are trained and tested using the same hyperparameters searched by RandomizedSearchCV for each model as elaborated in Table 2 and 3

1) Random Forest (RF)

TABLE II. HYPERPARAMETERS SETTING

N Estimators	157
Max Depth	25
Min Samples Split	5
Min Samples Leaf	1

2) XGBoost

TABLE III. HYPERPARAMETERS SETTING

N Estimators	163
Max Depth	9
Learning Rate	0.04993475643167195
Col Sample By Tree	0.8925879806965068
Subsample	0.73338144662399

IV. RESULTS

In this study, RF and XGBoost were implemented to predict whether a patient has CKD using clinical features from patients dataset. Both models were trained to perform classification task.

A. Confusion Matrix

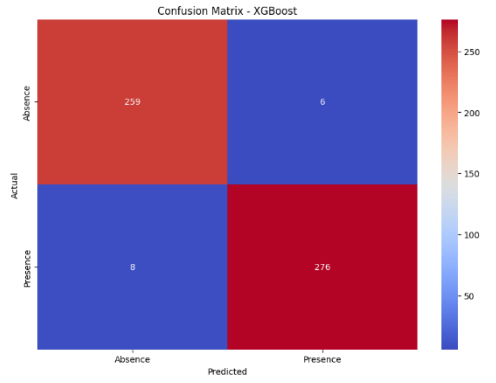
A confusion matrix shows the performance of a classification model by comparing the predicted labels with the actual labels. The structure is:

	Predicted: Absence	**Predicted: Presence**
Actual: Absence	True Negative (TN)	False Positive (FP)
Actual: Precense	False Negative (FN)	True Positive (TP)

Where TN means correctly predicted as absence of CKD. FP means predicted CKD presence when actually absent. FN means predicted absence of CKD when actually present. TP means correctly predicted presence of CKD.

IMAGE I. RF CONFUSION MATRIX

IMAGE II. XGB CONFUSION MATRIX



Based on the confusion matrix on both models, we can compute the evaluation metrics for each model:

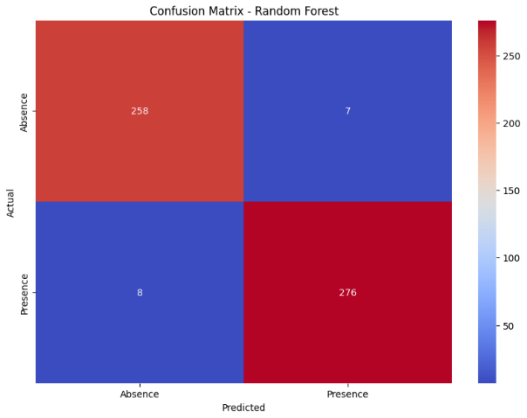
1) Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Model	Accuracy Test Set
Random Forest	97.27%
XGBoost	97.45%

2) Precision for Chronic Kidney Disease Presence

$$Precision = \frac{TP}{TP + FP}$$



B. *Evaluation Results*

The evaluation metrics used to assess both models performance include accuracy, precision, recall, and F1-score. The evaluation results of both models on each evaluation metrics could be seen down below:

TABLE IV. ACCURACY SCORE

Model	Accuracy	
	Validation Set	Test Set
Random Forest	97.7049%	97.2678%
XGBoost	97.3770%	97.4499%

TABLE V. PRECISION SCORE

Model	Precision	
	Class 0 (Non CKD)	Class 1 (CKD)
Random Forest	96.9925%	97.5265%
XGBoost	97.0037%	97.8723%

TABLE VI. RECALL SCORE

Model	Recall	
	Class 0 (Non CKD)	Class 1 (CKD)
Random Forest	97.3585%	97.1831%
XGBoost	97.7358%	97.1831%

Model	Precision Class 1 (CKD)
Random Forest	97.53%
XGBoost	97.87%

TABLE VII. F1-SCORE

Model	F1-score	
	Class 0 (Non CKD)	Class 1 (CKD)
Random Forest	97.1751%	97.3545%
XGBoost	97.3584%	97.5265%

TABLE VIII.MACRO AND WEIGHTED AVERAGE

Model	Macro F1	Weighted F1
Random Forest	97.2650%	97.2684%
XGBoost	97.4483%	97.4510%

After training and testing the both of the classification model using the datasets, RF has better accuracy (97.7049%) using the validation set compared to XGBoost (97.3770%). It is important to note that RF is more robust when evaluated on a dataset similar to training data. However, XGBoost (97.4499%) outperforms RF (97.2678%) while using the test set, meaning that XGBoost might be more effective in making predictions on new or unseen data.

XGBoost has a higher precision for Class 1 (CKD) (%) compared to RF (97.8723%) meaning that RF is more conservative and careful in diagnosing patients with CKD. XGB also performs better in precision for Class 0 (non-CKD) (97.0037%) compared to RF (96.9925%), meaning that XGBoost is more precise in identifying healthy individuals (non-CKD patients).

Recall score measure how many actual positive cases of CKD are correctly identified. A higher recall means fewer false negative. XGBoost has the same recall for Class 1 (CKD) (97.1831%) compared to RF (97.1831%) meaning that both model is better at catching more CKD cases. While, XGBoost has slightly higher recall for Class 0 (non-CKD) (97.7358%) compared to RF (97.3585%), making XGBoost slightly better at identifying healthy individuals (non-CKD patients).

Based on the F1-scores (Table VII), XGBoost display a slightly more balanced performance between precision and recall because XGBoost has higher F1-scores for both classes than RF. It indicates that XGBoost balances between detecting CKD patients and avoiding false diagnosis better than RF. As well as the macro f1 and weighted f1, XGBoost performs better accross both balanced (macro f1) and imbalanced (weighted f1) evaluation scenarios compared to RF, showing that XGBoost handles both CKD and non-CKD classes effectively. These metrics confirms that XGB provides more stable and consistent results.

C. Top 20 Feature Importance

1) Random Forest

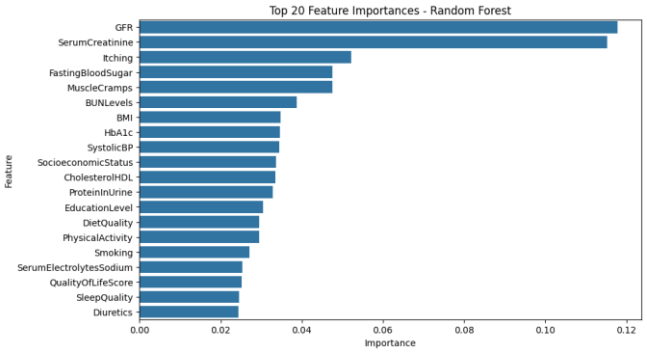
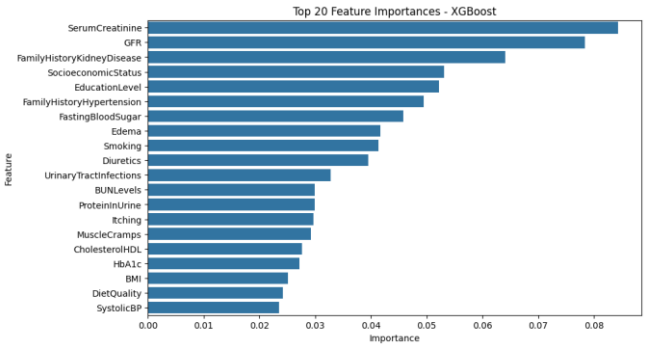


IMAGE III. RF TOP 20 FEATURE IMPORTANCE

2) XGBoost



REFERENCES

[1] Raihan, M. J., Khan, M. A. M., Kee, S. H., & Nahid, A. al. (2023). Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-33525-0

[2] Kim, H. W., Heo, S.-J., Kim, J. Y., Kim, A., Nam, C.-M., & Kim, B. S. (2021). Dialysis adequacy predictions using a machine learning method. Scientific Reports, 11(15417). https://doi.org/10.1038/s41598-021-94964-1

[3] Ammirati, A. L.. (2020). Chronic Kidney Disease. Revista Da Associação Médica Brasileira, 66, s03–s09. https://doi.org/10.1590/1806-9282.66.S1.3

[4] Furizal, Ma'arif, A., & Rifaldi, D. (2023). Application of Machine Learning in Healthcare and Medicine: A Review. In Journal of Robotics and Control (JRC) (Vol. 4, Issue 5). https://doi.org/10.18196/jrc.v4i5.19640

[5] Kaur, M., Dhalaria, M., Sharma, P. K., & Park, J. H. (2019). Supervised machine-learning predictive analytics for national quality of life scoring. Applied Sciences (Switzerland), 9(8). https://doi.org/10.3390/app9081613

[6] K. I. Islam, E. Elias, K. C. Carroll, and C. Brown, "Exploring Random Forest Machine Learning and Remote Sensing Data for Streamflow Prediction: An Alternative Approach to a Process-Based Hydrologic Modeling in a Snowmelt-Driven Watershed," Remote Sens., vol. 15, no. 16, p. 3999, 2023, doi: 10.3390/rs15163999.

[7] R. Meenal, P. A. Michael, D. Pamela, and E. Rajasekaran, "Weather prediction using random forest machine learning model," Indonesian Journal of Electrical Engineering and Computer Science, vol. 22, no. 2, p. 1208, May 2021, doi: https://doi.org/10.11591/ijeecs.v22.i2.pp1208-1215.

[8] M. Muhasshanah, M. Tohir, D. A. Ningsih, N. Y. Susanti, A. Umiyah, and L. Fitria, "Comparison of the Performance Results of C4.5 and Random Forest Algorithm in Data Mining to Predict Childbirth Process", CommIT (Communication and Information Technology) Journal, vol. 17, no. 1, pp. 51-59, Mar. 2023.

[9] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, *13-17-August-2016*. https://doi.org/10.1145/2939672.2939785

[10] Scornet, E., Biau, G., & Vert, J. P. (2015). Consistency of random forests. *Annals of Statistics*, *43*(4). https://doi.org/10.1214/15-AOS1321

[11] S. Mendapara, A. Ghate, D. Guha, and A. Natarajan, "CKD biomarker identification using gene expression and random forest classification," *Frontiers in Genetics*, vol. 15, pp. 1–11, 2024. https://doi.org/10.3389/fgene.2024.1409755

[12] Y. Li, M. Kumar, and T. Nguyen, "XGBoost-based early prediction of CKD progression using claims data and explainable AI," *arXiv preprint arXiv:2409.12087*, 2024. https://arxiv.org/abs/2409.12087

[13] H. Wang, J. Liu, and F. Zhao, "Predicting chronic kidney disease progression using hybrid ensemble models with SMOTE," *Diagnostics*, vol. 12, no. 10, p. 2454, 2022. https://doi.org/10.3390/diagnostics12102454

[14] Boucekine, M., Loundou, A., Baumstarck, K. et al. Using the random forest method to detect a response shift in the quality of life of multiple sclerosis patients: a cohort study. BMC Med Res Methodol 13, 20 (2013). https://doi.org/10.1186/1471-2288-13-20

[15] Kaplan, R. M., & Hays, R. D. *Health-related quality of life measurement in public health*. *Annual Review of Public Health*, 43, 355–373. (2022) https://doi.org/10.1146/annurev-publhealth-052120-012811

[16] Al Salmi, Issa, et al. *Kidney Disease-Specific Quality of Life among Patients on Hemodialysis*, *International Journal of Nephrology*, 2021, 8876559. https://doi.org/10.1155/2021/8876559

[17] Faridah, V. N., Ghozali, M. S., Aris, A., Sholikhah, S., & Ubudiyah, M. (2021). [7] Indonesian Journal of Community Health Nursing, 6(1), 28. https://doi.org/10.20473/ijchn.v6i1.26660

[18] M. Savić, V. Kurbalija, M. Ilić, M. Ivanović, D. Jakovetić, A. Valachis, S. Autexier, J. Rust, and T. Kosmidis, "The application of machine learning techniques in prediction of quality of life features for cancer patients," Computer Science and Information Systems, vol. 20, no. 1, pp. 381–404, 2022. doi: 10.2298/CSIS220227061S.

[19] S. Fatima, A. Hussain, S. B. Amir, S. H. Ahmed, and S. M. H. Aslam, "XGBoost and Random Forest Algorithms: An In-Depth Analysis," Pakistan Journal of Scientific Research (PJOSR), vol. 3, no. 1, pp. 26–31, 2023. 10.57041/pjosr.v3i1.946

[20] Aracena, C., Villena, F., Arias, F., & Dunstan, J. (2022). Applications of machine learning in healthcare. Revista Medica Clinica Las Condes, 33(6). https://doi.org/10.1016/j.rmclc.2022.10.001