Institut Riset dan Publikasi Indonesia (IRPI)
IJATIS: Indonesian Journal of Applied Technology and Innovation Science
Journal Homepage: https://journal.irpi.or.id/index.php/ijatis
Vol. 1 Iss. 2 August 2024, pp: 68-74
ISSN(P): 3032-7466 | ISSN(E): 3032-7474

Performance Comparison of Classification Algorithms for Chronic Kidney Disease Prediction

Farin Junita Fauzan[1*], Celine Mutiara Putri[2], Prita Laura[3]

[1,2]Department of Information Systems, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[3]Department of Performing Arts, College of design, Shute University, Taiwan

E-Mail: [1]12050323918@students.uin-suska.ac.id,
[2]12050327108@students.uin-suska.ac.id, [3]laurazizi98@gmail.com

Received Dec 27th 2023; Revised Jun 15th 2024; Accepted Jul 14th 2024
Corresponding Author: Farin Junita Fauzan

Abstract

Chronic Kidney Disease (CKD) is an abnormal kidney function or failure of the kidneys to filter the bloodstream and remove metabolic waste that progresses over months or years. Chronic kidney disease is asymptomatic in its early stages. It has no age limit, and if you already suffer from chronic kidney disease, the likelihood of a sudden decline in kidney function increases. The medical record data of chronic kidney disease patients can be utilized to make predictions and can be processed using machine learning to classify the risk of death. This research will use Ensemble Learning, which combines Decision Tree, XGBoost, and Extra Trees algorithms. In the pre-processing stage, value filling is carried out using the random sampling method. It was concluded that the highest accuracy value in Extra Trees was 96%. In comparison, the Decision Tree was 94%, and the XGBoost method obtained 95% accuracy so that Pathologists can use it in developing a program to predict chronic kidney disease.

Keywords: Chronic Kidney Disease, Decision Tree, Ensemble Learning, Extra Trees, XGBoost

1. INTRODUCTION

Chronic Kidney Disease (CKD) refers to the atypical functioning of the kidneys or their inability to effectively filter the blood and eliminate metabolic waste, which develops gradually over a period of months or years [1]. Frequently, the diagnosis of chronic kidney disease occurs through the screening of individuals who are identified as being at risk of kidney complications, such as those with hypertension or diabetes, as well as those with family members affected by CKD [2]. According to the Global Burden of Disease survey in 2010, chronic renal disease was rated 18th among fatal diseases, impacting 10% of the global population [3].

Chronic kidney disease is asymptomatic in its early stages and is not limited by age. However, having chronic kidney disease increases the likelihood of experiencing an abrupt reduction in kidney function. Hence, it is crucial to promptly identify the disease in order to enhance the likelihood of impeding or halting its advancement during its initial phase and mitigate the escalation of the condition. The risk prediction can be made using the medical record data of patients with chronic renal disease. Machine learning can be employed to analyse medical record data and accurately categorise the mortality risk of patients diagnosed with cardiovascular disease.

Machine Learning is a subdivision of artificial intelligence that specifically emphasises the capacity of systems to acquire knowledge and skills from data. Machine learning systems possess the ability to enhance their functionalities autonomously, without necessitating frequent manual programming by humans [4]. This learning process enables the system to see patterns, generate forecasts, and make judgements based on the given data. Machine Learning employs various models, including Logistic Regression, K-Nearest Neighbours, Support Vector Machine (SVM), Decision Tree, and Naïve Bayes. Machine Learning encompasses two primary concepts: supervised learning and unsupervised learning.

Supervised learning operates under the assumption that a teacher or supervisor is present to add class information or labels to the training instances. Within this framework, every training example is categorised into a certain class. Unsupervised learning techniques are frequently employed for the purposes of clustering and reducing dimensionality. Clustering is a process where data is grouped into clusters based on the similarity

of their features. Conversely, in the context of dimensionality reduction, the objective is to decrease the intricacy of the data while preserving essential information.

In Machine Learning modeling, there are two types, namely Single Model and Ensemble Learning. Ensemble learning itself is a supervised learning algorithm because it can be trained and used to make predictions. Ensemble learning combines one or more of the single model algorithms so that it has high flexibility because it has a combined use of the single model [5]. Ensemble Learning is one of the algorithms in supervised learning. It is used to make predictions based on training data that has been labeled with a class. Ensemble Learning involves combining one or more single-model algorithms to improve the quality of predictions.

Related research is Breast Cancer Classification using XGBoost[6] achieves an accuracy of 94.74% and a recall of 95.24% on the Wisconsin Breast Cancer (Diagnostic) dataset.. [7] The research also indicates that twelve different machine learning classifiers were evaluated in a supervised learning setting, with the highest performance metrics achieved by the XgBoost classifier, including an accuracy of 0.983, precision of 0.98, recall of 0.98, and F1-score of 0.98. Other relevant studies focus on a hybrid machine learning model for predicting chronic kidney disease. [8] Gradient boosting achieves approximately 99% accuracy, random forest achieves 98%, the decision tree classifier achieves 96% accuracy, and our suggested hybrid model performs the best, achieving 100% accuracy on the same dataset. Various machine learning algorithms were also tested on the chronic kidney disease dataset in the research. [9], and their performance was compared.
their performance was compared. The Extra Tree Classifier achieved the highest accuracy of 99% in predicting chronic kidney disease. In predicting chronic kidney disease. This study concludes that machine learning approaches, specifically the Extra Trees Classifier, can be a useful tool for early diagnosis of chronic kidney disease, which could potentially provide better outcomes for patients through timely intervention.

In the study titled "Chronic Kidney Disease Prediction Using the Naive Bayes Classifier Algorithm Based on Particle Swarm Optimization [10]; the Naive Bayes classification results enhanced by Particle Swarm Optimization achieve a confusion matrix accuracy of 98.75% and an AUC of 99%. In comparison, Naive Bayes without Particle Swarm Optimization achieves a confusion matrix accuracy of 97.00% and an AUC of 99.8%. Other related research Comparison Of Svm And Nn Data Mining Methods For Classification Of Chronic Health Diseases [11]. From the research results obtained, the neural network method produces an accuracy value of 93.36% and SVM with a value of 95.16%. Other related research

From the background of the problem and related research that has been described above, this research will use Ensemble Learning, which combines Decision Tree algorithms, XGBoost, and Extra Trees. In this study, classification using Ensemble Learning is carried out because Ensemble learning is a powerful tool in improving the quality of kidney disease prediction and can provide significant benefits in medicine for early diagnosis and better treatment.

2. MATERIAL AND METHOD

2.1. Data Mining

Data mining, as described in the book "Data Mining Concepts and Techniques," refers to the systematic exploration and analysis of extensive data sets with the objective of discovering patterns, models, and other forms of valuable knowledge [12]. Hartono describes data mining as a systematic process that utilizes statistical, mathematical, artificial intelligence, and machine learning techniques to extract valuable information and identify its relationships from large databases. Data mining is not an entirely new field [13]. Data mining, according to another definition, refers to a set of procedures aimed at extracting previously unknown information of value from a database. The information is derived by extracting and identifying significant or intriguing patterns from the database's data [14].

2.2. Classification

Classification involves evaluating data objects and assigning them to certain classes from a set of accessible classes. Classification involves the creation of a model using pre-existing training data, which is subsequently utilised to categorise fresh data. Classification is the process of training a target function that translates sets of qualities (features) to certain class labels [15]. The objective of classification is to infer the category of an unlabeled object. The possible models utilised include if-then rules, decision trees, and neural networks [16].

2.3. Algorithm

In general, an algorithm is a clear sequence of steps to solve a problem. In computer science and mathematics, an algorithm is a sequence of steps to perform calculations or can also be used to solve problems that are written sequentially [17].

2.4. Ensemble learning

Ensemble Learning is a machine learning technique in which multiple machine learning models are combined to improve prediction accuracy and consistency. In the context of the classification of health effects due to air pollution, Ensemble Learning can be used to integrate several different machine learning models, such as Decision Trees, Random Forests, Support Vector Machines (SVM), and Neural Networks (Gokul et al. 2023).

2.5. Decision Tree

Decision Tree is one of the machine learning algorithms used for classification and regression problems. This algorithm generates a model in the form of a tree structure, where each internal node represents a decision based on data features, and each leaf represents a class label (for classification) or predicted value (for regression) [18]. The advantage of a Decision Tree is its high interpretability. Decision tree models can be described visually so that humans can easily understand them. In addition, Decision Tree can also handle categorical and numerical data without the need for complex data preprocessing [18]. In general, it can be shown in equation 1.

$$\text{Entropy}(t) = -\sum_{i=1}^{K} P_i \text{Log}_2(P_i) \tag{1}$$

2.6. XGBoost

Xgboost is a machine-learning library that can be used to predict or classify based on decision trees [19]. XGBoost is composed of several decision trees that use boosting techniques in the construction of the algorithm [20].This algorithm allows optimization 10 times faster than other GBMs and has a better ability to fight overfitting problems [21]. In general, it can be shown in equation 2.

$$\text{Obj}^{(t)} = \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)} + f_i(x_i) \right) \right] + \Omega(f_t) \tag{2}$$

2.7. Extra Trees

The Extra Trees Classifier, also known as the Extremely Randomized Trees Classifier, is an ensemble algorithm that falls within the family of decision tree methods[22]. It is similar to Random Forest, but there is a key difference in the way the trees are constructed. The advantages of the Extremely Randomized Trees Classifier include higher training speed compared to Random Forest due to the use of randomized divisors[23], which reduces the complexity of the calculations. However, on the other hand, the interpretation of the model may be more difficult because the trees are built randomly. In general, it can be shown in equation 3.

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} h_m(x) \tag{3}$$

2.8. Confusion Matrix

A confusion matrix is a performance evaluation tool in machine learning, representing the accuracy of a classification model. It displays the number of true positives, true negatives, false positives, and false negatives. This matrix aids in analyzing model performance, identifying mis-classifications, and improving predictive accuracy. The confusion matrix can be show as figure 1 [26].

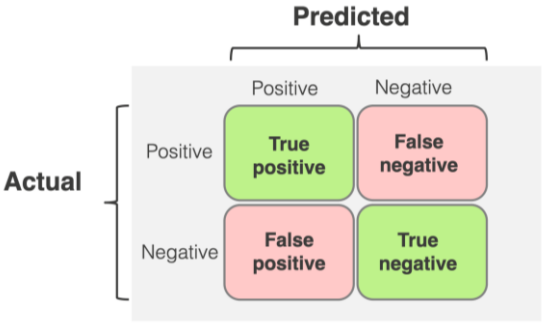

Figure 1. Confusion Matrix

Let's decipher the matrix:

1. The target variable has two values: Positive or Negative
2. The columns represent the actual values of the target variable
3. The rows represent the predicted values of the target variable

2.9. Research Method

The research consists of several main processes from data collection, pre-processing to algorithm evaluation. In general, the research methodology can be shown in Figure 2.

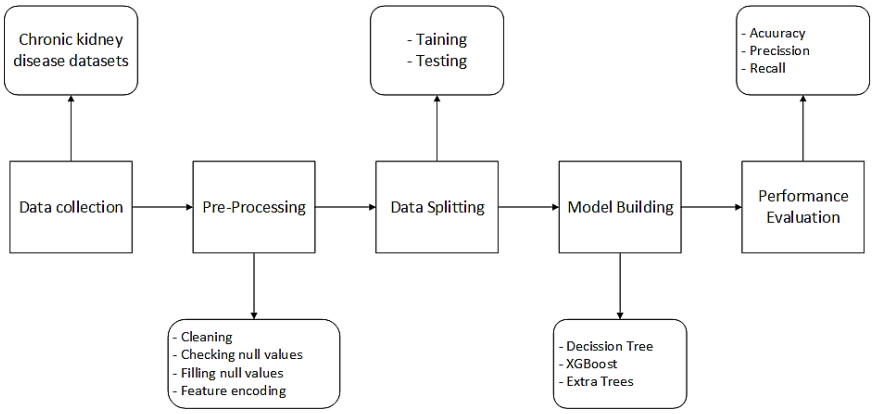

Figure 2. Research Methodology

The stage of data collection is fundamental and early in the research process. In this study, the dataset consisted of 400 records with 25 attributes focused on Chronic Kidney Disease (CKD) as the target class. The data was obtained from UCI and is accessible via www.kaggle.com. It includes both numerical and nominal data types, featuring attributes such as age, blood pressure, specific gravity, albumin, sugar, and various other health indicators like red blood cells, pus cells, and medical conditions such as hypertension, diabetes mellitus, and coronary artery disease. Below is a table 1 describing the attributes of the dataset.

Table 1. Description of Attributes in the Dataset

Attribute	Description
Age	Age of the patient
Blood Pressure	Blood pressure measurement
Specific Gravity	Specific gravity of urine
Albumin	Albumin content in urine
Sugar	Sugar content in urine
Red Blood Cells	Presence of red blood cells in urine
Pus Cell	Presence of pus cells in urine
Pus Cell Clumps	Presence of clumps of pus cells in urine
Bacteria	Presence of bacteria in urine
Blood Glucose Random	Random blood glucose measurement
Blood Urea	Blood urea nitrogen level
Serum Creatinine	Serum creatinine level
Sodium	Sodium level in blood
Potassium	Potassium level in blood
Hemoglobin	Hemoglobin level in blood
Packed Cell Volume	Volume occupied by packed red blood cells in blood
White Blood Cell Count	Count of white blood cells
Red Blood Cell Count	Count of red blood cells
Hypertension	Presence of hypertension
Diabetes Mellitus	Presence of diabetes mellitus
Coronary Artery Disease	Presence of coronary artery disease
Appetite	Appetite condition
Pedal Edema	Presence of pedal edema (swelling in legs)
Anemia	Presence of anemia

Next, the dataset will undergo cleaning and preprocessing. This involves checking each column for missing values, which will be addressed through imputation. Categorical data will be encoded into numeric values. The dataset will then be split into training and testing sets, with the testing set comprising 30% of the dataset. Various classification algorithms discussed earlier will be applied to the training set. To assess and compare their performance, metrics such as accuracy, sensitivity, and specificity will be utilized [24].

3. RESULTS AND DISCUSSION

The dataset that has been processed is then divided into training data and testing data. Then calculations are carried out with the Decision Tree, XGBoost, and Extra Trees methods, the confusion matrix results are obtained as shown in Table 2.

Table 2. Confusion Matrix Decision Tree

	True CKD	True Not CKD	Class precision
Pred, ckd	60	3	95%
Pred,notckd	4	44	94%
Class recall	96%	92%	

Table 2 shows the confusion matrix for Decission Tree model, our model achieves the following:

1. 60 instances of TP, That is, the model correctly predicted that the patient had CKD.
2. There were 3 cases that should have been predicted as not CKD but were incorrectly predicted as CKD FP.
3. 4 instances of false FP, This means the model erroneously predicted that the patient had CKD when they did not.
4. 44 TN, This means the model correctly predicted that the patient did not have CKD.

Table 3. Confusion Matrix XGBoost

	True CKD	True Not CKD	Class precision
Pred, ckd	72	0	94%
Pred,notckd	5	43	100%
Class recall	100%	90%	

Table 3 shows the confusion matrix for XGBoost model, our model achieves the following:

1. 72 instances of TP, That is, the model correctly predicted that the patient had CKD.
2. No prediction of FN, This means that the model correctly predicted that the patient did not have CKD.
3. 5 instances of FP, This means the model erroneously predicted that the patient had CKD when they did not.
4. 43 TN, This means the model correctly predicted that the patient did not have CKD.

Table 4. Confusion Matrix Extra Trees

	True CKD	True Not CKD	Class precision
Pred, ckd	72	0	95%
Pred,notckd	4	44	100%
Class recall	100%	92%	

Table 4 shows the confusion matrix for Extra Trees model, our model achieves the following:

1. 72 instances of TP, That is, the model correctly predicted that the patient had CKD
2. No prediction of FN, This means that the model correctly predicted that the patient did not have CKD.
3. 4 instances of FP, This means the model erroneously predicted that the patient had CKD when they did not.
4. 43 TN, This means the model correctly predicted that the patient did not have CKD.

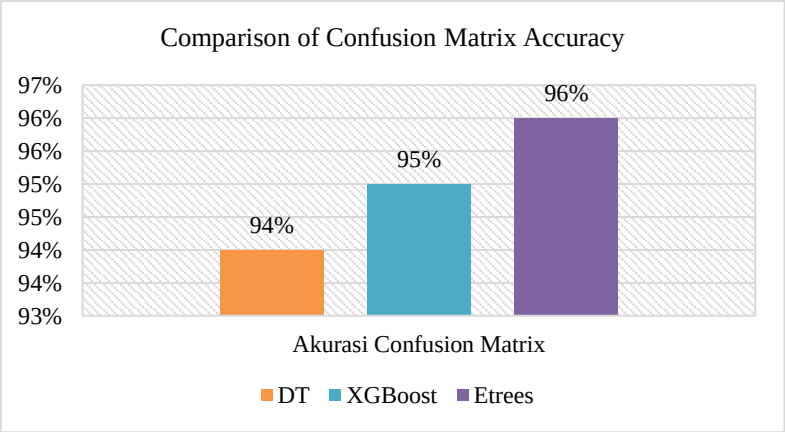

Figure 2. Accuracy Comparison

Overall, all models tended to provide great predictions for both classes, with fairly high precision and recall values. However, there were some undetected cases of CKD, which may require further attention. To facilitate the comparison of accuracy differences from the Confusion Matrix results between Decision Tree, XGBoost, and Extra Trees methods, a graphical representation is necessary [25]. Figure 3 is a comparison graph displaying the Confusion Matrix for Decision Tree, XGBoost, and Extra Trees.

4. CONCLUSION

In this study, Decision Tree, XGBoost, and Extra Trees modeling were carried out using chronic kidney disease datasets taken from the UCI Repository. Researchers conducted data processing to obtain which mode has a higher accuracy value for chronic kidney disease datasets. It is known from the research results that the Decision Tree method obtained an accuracy value of 94%, the XGBoost method obtained an accuracy value of 95%, and the Extra Trees method obtained an accuracy value of 96%.

The results obtained are included in the type of excellent classification. So, it can be concluded that Decision Tree, XGBoost, and Extra Trees have good performance performance for processing chronic kidney disease datasets. Furthermore, from the research results it is known for chronic kidney datasets that the Extra Tress method produces higher accuracy values than the Decision Tree and XGBoost methods. So, it can be used by Pathologists in making programs to predict chronic kidney disease. Based on the results of the research that has been done, the researchers propose to conduct experiments using other methods such as AdaBoost, Random Forest, Gradient Boosting, or optimization algorithms such as genetic algorithm and ant colony optimization.

REFERENCES

[1] A. Ogunleye and Q. G. Wang, "XGBoost Model for Chronic Kidney Disease Diagnosis," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 17, no. 6, pp. 2131–2140, 2020, doi: 10.1109/TCBB.2019.2911071.

[2] Parul Sinha and Poonam Sinha, "Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM," Int. J. Eng. Res., vol. V4, no. 12, pp. 608–612, 2015, doi: 10.17577/ijertv4is120622.

[3] D. Baidya, U. Umaima, M. N. Islam, F. M. J. M. Shamrat, A. Pramanik, and M. S. Rahman, "A Deep Prediction of Chronic Kidney Disease by Employing Machine Learning Method," 2022 6th Int. Conf. Trends Electron. Informatics, ICOEI 2022 - Proc., no. April, pp. 1305–1310, 2022, doi: 10.1109/ICOEI53556.2022.9776876.

[4] R. Diana, H. Warni, and T. Sutabri, "Penggunaan Teknologi Machine Learning Untuk Pelayanan Monitoring Kegiatan Belajar Mengajar Pada SMK Bina Sriwijaya Palembang," J. Tek. Inform., vol. 11, no. 1, 2023.

[5] C. Sanjaya and S. H. Supangkat, "Predictive Analitycs Menggunakan Machine Learning Untuk Memprediksi Waktu Keterlambatan Berdasarkan Penyebab Keterlambatan Pada PT. Kereta Api Indonesia," J. Sist. Cerdas, vol. 3, no. 1, pp. 165–180, 2020.

[6] Rahmanul Hoque, Suman Das, Mahmudul Hoque, and Mahmudul Hoque, "Breast Cancer Classification using XGBoost," World J. Adv. Res. Rev., vol. 21, no. 2, pp. 1985–1994, 2024, doi: 10.30574/wjarr.2024.21.2.0625.

[7] M. A. Islam, M. Z. H. Majumder, and M. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms," J. Pathol. Inform., vol. 14, p. 100189, 2023, doi: https://doi.org/10.1016/j.jpi.2023.100189.

[8] H. Khalid, A. Khan, M. Zahid Khan, G. Mehmood, and M. Shuaib Qureshi, "Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease," Comput. Intell. Neurosci., vol. 2023, no. 1, p. 9266889, Jan. 2023, doi: https://doi.org/10.1155/2023/9266889.

[9] M. Imran et al., "Predictive Modeling of Chronic Kidney Disease Using Extra Tree Classifier: A Comparative Analysis with Traditional Methods," vol. 06, no. 02, 2024.

[10] T. Arifin and D. Ariesta, "Prediksi Penyakit Ginjal Kronis Menggunakan Algoritma Naive Bayes Classifier Berbasis Particle Swarm Optimization," J. Tekno Insentif, vol. 13, no. 1, pp. 26–30, 2019.

[11] H. Amalia, "Perbandingan Metode Data Mining Svm Dan Nn Untuk Klasifikasi Penyakit Ginjal Kronis," J. PILAR Nusa Mandiri, vol. 14, no. 1, pp. 1–6, 2018.

[12] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques - Jiawei Han, Jian Pei, Hanghang Tong - Google Books, Fourth Edi. 2022.

[13] R. Hartono, Y. Sumaryana, and A. Nurfaizi, "Analisa Perbandingan Kinerja Algoritma Klasifikasi Untuk Prediksi Penyakit Kanker Payudara," J. Teknol. Inf., vol. 7, no. 1, pp. 116–124, 2023.

[14] A. P. Sandi and V. W. Ningsih, "Implementasi Data Mining Sebagai Penentu Persediaan Produk Dengan Algoritma Fp-Growth Pada Data Penjualan Sinarmart," J. Publ. Ilmu Komput. dan Multimed., vol. 1, no. 2 SE-Articles, May 2022, doi: 10.55606/jupikom.v1i2.343.

[15] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," J. Media Inform. Budidarma, vol. 4, no. 2, p. 437, 2020, doi:

10.30865/mib.v4i2.2080.
[16] I. Ikko, M. Rizky, S. Yusuf, and I. Sriyanto, “Perbandingan Kinerja Algoritma Naive Bayes, Support Vector Machine dan Random forest untuk Prediksi Penyakit Ginjal Kronis Imaniar Ikko Mulya Rizky 1a* , Suhendro Yusuf Irianto 2b. Sriyanto 3c,” vol. 18, pp. 139–151, 2023.
[17] M. P. Putri et al., Algoritma Dan Struktur Data. 2022.
[18] B. Sunarko et al., “Penerapan Stacking Ensemble Learning untuk Klasifikasi Efek Kesehatan Akibat,” Edu Komputika J., vol. 10, no. 1, pp. 55–63, 2023, doi: https://doi.org/10.15294/edukomputika.v10i1.72080.
[19] Y. Jiang, G. Tong, H. Yin, and N. Xiong, “A Pedestrian Detection Method Based on Genetic Algorithm for Optimize XGBoost Training Parameters,” IEEE Access, vol. 7, pp. 118310–118321, 2019, doi: 10.1109/ACCESS.2019.2936454.
[20] M. Syukron, R. Santoso, and T. Widiharih, “Perbandingan Metode Smote Random Forest Dan Smote Xgboost Untuk Klasifikasi Tingkat Penyakit Hepatitis C Pada Imbalance Class Data,” J. Gaussian; Vol 9, No 3 J. GaussianDO - 10.14710/j.gauss.9.3.227-236, Aug. 2020.
[21] B. Jange, “Prediksi Harga Saham Bank BCA Menggunakan Prophet,” J. Trends Econ. Account. …, vol. 2, no. 1, pp. 1–5, 2021, doi: 10.47065/arbitrase.v3i2.495.
[22] R. Shafique, A. Mehmood, and G. S. Choi, “Cardiovascular disease prediction system using extra trees classifier,” 2019.
[23] J. Abdollahi, B. Nouri-Moghaddam, and M. Ghazanfari, “Deep Neural Network Based Ensemble learning Algorithms for the healthcare system (diagnosis of chronic diseases),” arXiv Prepr. arXiv2103.08182, 2021.
[24] M. Vakili, M. Ghamsari, and M. Rezaei, “Performance analysis and comparison of machine and deep learning algorithms for IoT data classification,” arXiv Prepr. arXiv2001.09636, 2020.
[25] P. Carmona, A. Dwekat, and Z. Mardawi, “No more black boxes! Explaining the predictions of a machine learning XGBoost classifier algorithm in business failure,” Res. Int. Bus. Financ., vol. 61, p. 101649, 2022.
[26] M. A. Fayyad, V. Kurniawan, M. R. Ahugrah, B. H. Estanto, and T. Bilal, "Application of Recurrent Neural Network Bi-Long Short-Term Memory, Gated Recurrent Unit and Bi-Gated Recurrent Unit for Forecasting Rupiah Against Dollar (USD) Exchange Rate", Public Research Journal of Engineering, Data Technology and Computer Science, vol. 2, no. 1, pp. 1–10, Apr. 2024, doi: 10.57152/predatecs.v2i1.1094.