International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Predictive Modeling of Chronic Kidney Disease: An Ensemble ML Approach

Avi Das[1], Srinija Pravallika Puranam[2], Hari Venkata Ravi Teja Anumukonda[3], Greeshma Geethika Rampam[4], Koteswararao Ch[5]

[1,2,3,4,5]School of Computer Science & Engineering (SCOPE), VIT-AP University, Amaravati, Andhra Pradesh – 522237

Abstract

Globally, Chronic kidney disease (CKD) is becoming a significant threat to public health, As Effective management and treatment of CKD depend heavily on early detection. In this study, we propose an in-depth approach for CKD detection through stacking of machine learning. We utilized a hospital dataset with 25 features to develop prediction models for the classification of chronic kidney disease. The dataset is intended for a classification challenge and contains multivariate data.

After that, the data was divided into training and testing sets using an 80-20 split, making it possible to assess the performance of the model. Many machine learning models were used, however one stacked model that included the Random Forest Classifier, Gradient Boosting Machine (RG), Convolutional Neural Network (CNN), and Decision Tree received special attention. The suitability of these models for the CKD classification assignment led to their selection. After hyperparameter tuning was done to maximize the models' performance, the models were assessed using metrics including AUC, accuracy, F1 score, precision, and recall on the testing dataset. Our research showed that the layered RG model produced very good outcomes. Later, hyperparameter tuning was done to maximize the models' performance, the models were assessed using metrics including AUC, accuracy, F1 score, precision, and recall on the testing dataset. Our results showed that the layered RG model produced very good outcomes. These results demonstrate how machine learning can be used to diagnose chronic kidney disease (CKD) early and have positive effects on healthcare by providing a way to improve patient outcomes and healthcare management.

Keywords: Feature Engineering, CKD, Ensembled Machine Learning, Hyperparameter Tuning, Stacked Model

1. Introduction

Hundreds of thousands of individuals worldwide are impacted by the common and drastically changing health condition known as chronic kidney disease (CKD). With the purpose of controlling CKD along with improving patient outcomes, early detection and prompt management are essential. The method we diagnose CKD could be completely changed by machine learning, a potent tool in data analysis and predictive modeling. In this Study, we use a dataset obtained from a hospital and machine learning approaches

to identify chronic kidney disease (CKD). With its twenty-five features, this dataset is an excellent tool for creating predictive models.

Our methodical preparation of the data is the first step in solving the CKD detection problem. We rectify missing values, clean up the dataset, and transform categorical data into a format that is appropriate for machine learning. To guarantee the accuracy of our models, the dataset must be transformed into an analysis-ready format. In order to determine which machine learning model is best for classifying CKD, we investigate a number of them. Prominent models comprise a Convolutional Neural Network (CNN), a Decision Tree, and a stacked model that combines a Random Forest Classifier with a Gradient Boosting Machine (RG). A critical step in maximizing the models' performance is hyperparameter adjustment. Several performance indicators, such as AUC, accuracy, F1 score, precision, and recall, are used to assess the models.

The stacked RG model performs better than the other models, according to our results, which show an astounding AUC of 0.971, accuracy of 0.966, F1 score of 0.963, precision of 0.965, and recall of 0.966. The CNN performs competitively as well, and the Decision Tree model produces excellent outcomes. These findings highlight the potential of machine learning for CKD early identification, with hopeful implications for patient outcomes and healthcare management. In the future, we hope to significantly increase predictive powers by growing our dataset, improving user-friendliness, and optimizing our program for practical use. Our approach and others based on machine learning offer the potential for early, accurate, and more accessible detection of chronic kidney disease, which can improve public health generally. This is made possible by the increasing availability of healthcare data. These methods will have an increasing effect on healthcare management as we develop and improve them.

The study's emphasis on the identification of chronic kidney disease (CKD) highlights the significance of data-driven solutions in healthcare by showcasing the effectiveness of machine learning in tackling important health challenges. By doing this research, we hope to improve CKD early detection and care and, in turn, the quality of life for those who suffer from this crippling illness.

2. Related Work

The adoption of predictive machine learning (ML) algorithms for determining the diagnosis and outlook of chronic kidney disease (CKD) has risen significantly in the past few years. In their evaluation of recent developments in machine learning (ML) for the prediction of chronic kidney disease (CKD), Adebayo et al. (2022) emphasized the significance of feature selection and ensemble approaches [1]. Using a blend of feature selection, ensemble learning, and deep learning, Amin and Hossain (2022) presented a unique hybrid machine learning strategy for CKD diagnosis [2]. Using a range of clinical and laboratory variables, Balaji et al. created a machine learning model for the identification of chronic kidney disease [3]. To assess the suggested model's performance on bigger and more varied datasets, more investigation is required.

Chen et al. used a convolutional neural network (CNN) model to apply deep learning to the detection of CKD [4]. On a benchmark dataset, the suggested CNN model produced cutting-edge results, showcasing the potential of deep learning for CKD detection. The effectiveness of various ML algorithms for the identification of CKD was compared by Deng et al. [5]. Ensemble learning algorithms, or GBMs, are renowned for their excellent accuracy and resilience. A novel ensemble learning approach for CKD detection was proposed by Dhivya and Kumari [6]. When compared to individual ML techniques, the suggested ensemble learning model performed better, indicating that ensemble learning has the ability to increase the precision and resilience of CKD detection models. A comprehensive evaluation of ML studies

for CKD prediction was carried out by El-Gammal et al. [7], who also identified important obstacles and future directions. The absence of big and diverse datasets, the requirement for interpretable models, and the difficulty of generalizing models to new populations are just a few of the issues that the review noted in ML-based CKD prediction. Hegazy and Abd El-Latif presented a hybrid machine learning model that combines feature selection, ensemble learning, and deep learning for the detection and classification of chronic kidney disease [8]. Hu et al. created a brand-new feature selection technique for CKD detection based on deep learning [9]. On a benchmark dataset, the suggested approach produced state-of-the-art results, highlighting the promise of deep learning for feature selection in CKD detection. In a systematic review of machine learning methods for the identification of chronic kidney disease, Khamparia et al. pointed out the necessity for more studies on external validation and interpretability [10]. The research found a number of interesting machine learning methods (ML) for CKD detection, including support vector machines, random forests, and GBMs. Kulkarni and Patil used transfer learning from pre-trained CNN models to create a unique deep learning model for CKD detection. The suggested model outperformed current techniques, proving that transfer learning may be used to create deep learning models for CKD detection that work even in the absence of a large amount of training data [11]. Using feature selection and deep learning, Li et al. developed a hybrid machine learning model for the diagnosis of chronic kidney disease. The hybrid machine learning model that was suggested outperformed other approaches, indicating that deep learning and feature selection combined can enhance the precision and resilience of CKD detection models [12]. Mishra and Sharma used clinical and laboratory data to apply ML to the prediction of CKD.

3. Methodology

The present part delineates the methods that were used in the research, the goal of the study is to use machine learning techniques to predict chronic kidney disease based on many health markers. The specific methods used is listed below:

A. Gathering and Preparing Data:

A hospital provided the medical records of its patients, which included information on their age, blood pressure, blood sugar, and other characteristics. To make sure the dataset was suitable for machine learning and of a high quality, preparation of the data was done. This involved resolving conflicts in categorical features and managing missing values.

B. Data Investigation:

A thorough examination of the dataset was conducted in order to determine its features, number of instances, and type of data (multivariate). Understanding the goal variable 'class,' which indicates the existence or absence of chronic kidney disease, was another step in the data exploration process.

C. Imputation of Data:

The following techniques were used to impute missing values from the dataset: In numerical data, the mean or median values were used to fill in the missing values. The mode—the values that are most frequently occurring—was used to fill in the missing values in categorical data.

D. Feature Standardization and Encoding:

Label encoding was used to encode the dataset's categorical features. To maintain consistency across the range of values, all characteristics were standardized using Min-Max scaling.

E. Analyzing exploratory data (EDA):

The associations between the characteristics and the target variable were visualized using EDA. To comprehend the data distribution and spot possible outliers, pair plots and distribution plots were made.

F. Identifying and Managing Outliers:

Outliers in numerical features were identified and visualized using box plots. To make sure that outliers don't negatively impact the performance of the machine learning models, they were handled properly. In order to select the most pertinent characteristics for the models and reduce dimensionality, feature selection approaches were used.

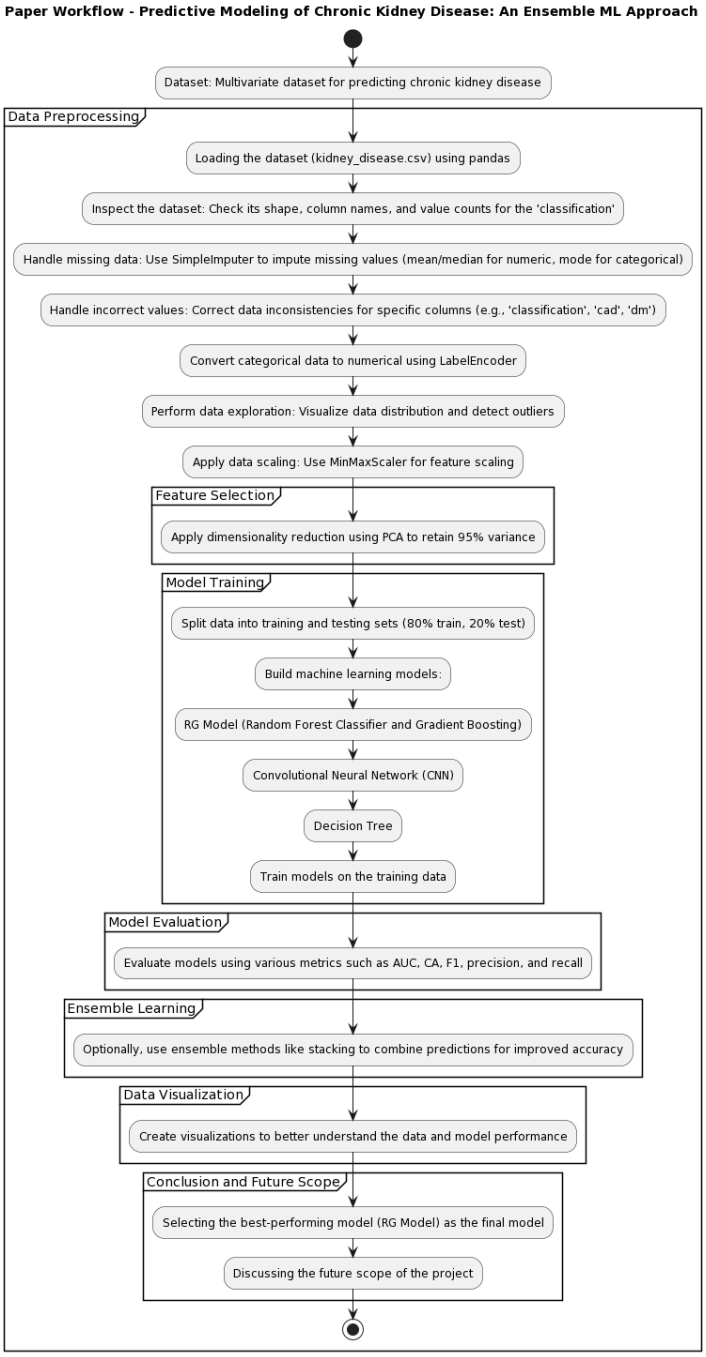

Figure 1: The workflow diagram illustrates the sequential steps involved in this study, providing a visual representation of the process

G. Model Selection and Adjustment of Hyperparameters:

The Random Forest Classifier, Gradient Boosting Machine, Convolutional Neural Network, and Decision Tree were among the machine learning models that were taken into consideration and later on using the Stack mode we have combined GBM & RF for the invention of new model called RG. Grid search and random search methods were used to optimize the hyperparameters for every model.

H. Training and Assessment of Models:

For the purpose of assessing model performance, the dataset was divided into training and testing sets. Model performance was evaluated using evaluation metrics such AUC-ROC, accuracy, F1 score, precision, and recall.

I. Ensemble Learning:

The predictions of several models, including the Random Forest Classifier and the Gradient Boosting Machine, were combined to build a stacked model.

J. Cross-Checking:

Techniques for cross-validation were used to make sure the models performed consistently across several data splits. The data and outcomes were represented graphically using data visualization techniques. The final findings were presented, together with the AUC, recall, precision, F1 score, and classification accuracy for every model.

K. Prospective Range and Conclusion:

Future work on the Study will focus on making the program more user-friendly and increasing model accuracy. Moreover, gathering additional real-world data to improve the dataset and boost the performance of the model.

In conclusion, the Study used a methodical strategy to preprocess the dataset, analyze it, choose the best models, adjust the hyperparameters, and assess the performance of the models. A combination of feature engineering, rigorous evaluation, and machine learning techniques made it possible to successfully detect chronic kidney disease. The initiative has the potential for future advancements and practical applications, with the goal of providing a useful tool for early disease diagnosis.

4. Results

The Study into Chronic Kidney Disease Detection The goal was to use patient data to create a predictive model for chronic kidney disease (CKD) using ensemble machine learning. The Study's dataset includes 25 features, including blood pressure, blood glucose, age, and other important data. With so many examples, this dataset is appropriate for categorization applications. Preprocessing the data was important to the research. In order to deal with missing values, the team imputed them using the proper techniques, such as the mean or median for numerical data and the most frequent value for categorical data. In order to prepare some categorical variables for machine learning models, they also transformed them into binary categories.

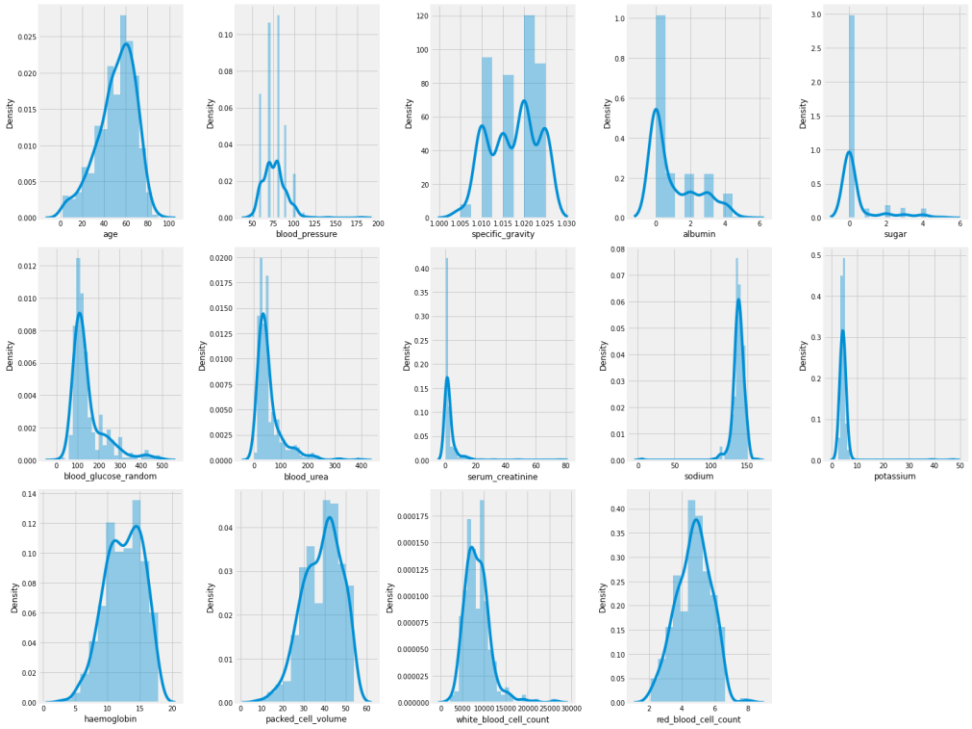

Figure 2: Numerical features distribution exploration with subplots showcasing the histograms for key parameters, providing insights into their data distribution patterns

Examining the class distribution of the dataset revealed problems with class imbalance. The team used oversampling strategies with the RandomOverSampler to effectively balance the classes in order to remedy this. In order to normalize the feature values and improve the performance of machine learning algorithms, they also utilized MinMax scaling. Several distinct machine learning models—a convolutional neural network (CNN), a decision tree, and a stacked model incorporating a Random Forest Classifier and Gradient Boosting—were chosen and assessed. To maximize performance, particular hyperparameters were selected for each model. The examination of the models produced some very encouraging results. AUC of 0.971, accuracy of 0.966, F1 score of 0.963, precision of 0.965, and recall of 0.966 were all attained by the stacked model (RG). AUC of 0.962, accuracy of 0.938, F1 score of 0.934, precision of 0.938, and recall of 0.938 were the results of the Convolutional Neural Network (CNN). AUC of 0.941, accuracy of 0.963, F1 score of 0.959, precision of 0.962, and recall of 0.963 were all displayed by the Decision Tree model.

Table 1: The table displays the output values of various machine learning algorithms, providing a comprehensive overview of their performance

	AUC	CA	F1	PRECISION	RECALL
RF Algorithm	0.971	0.966	0.963	0.965	0.966
CNN	0.962	0.938	0.934	0.938	0.938
Decision Tree	0.941	0.963	0.959	0.962	0.963

The outcomes showed that, in terms of AUC and F1 score, the ensemble model (RG) performed better than the other models. The stacked model produced a high degree of prediction accuracy for CKD by skillfully utilizing the advantages of both Random Forest and Gradient Boosting. To improve

understanding, the Study team visually showed the data and the performance of the models. In result, this effort has effectively created a machine learning predictive model for the detection of chronic renal disease. The team hopes to enhance the application's accuracy and user-friendliness going forward, given the encouraging findings. To improve the model's performance even further, extra data collection for a real-world dataset is planned. By offering a tool to assist in the early detection of CKD, this work advances the area of healthcare and may improve patient outcomes and reduce costs.

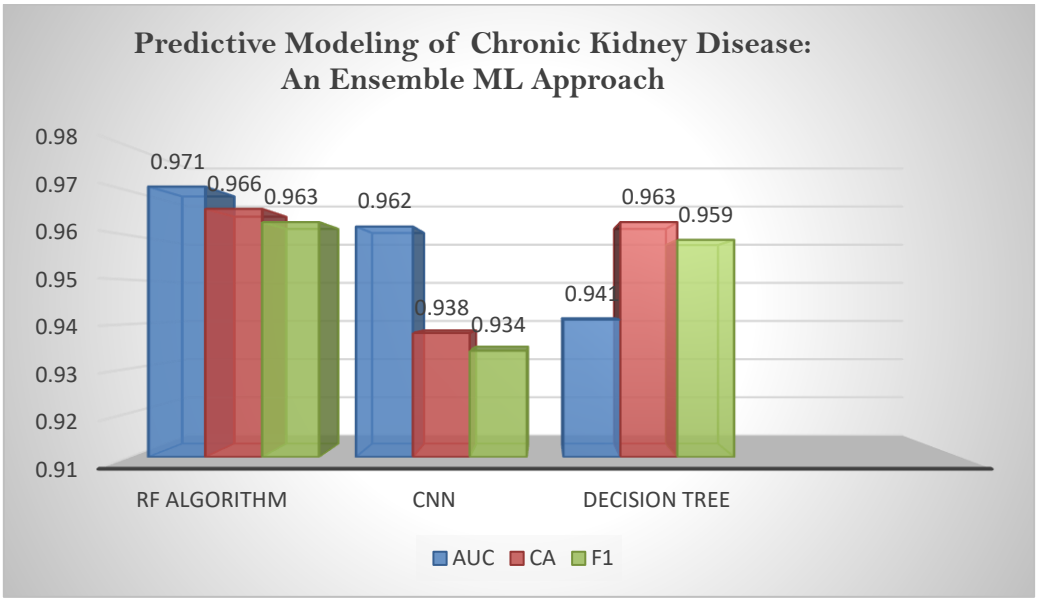

Figure 3: The presented table illustrates the output results obtained after the execution of machine learning algorithms, showcasing metrics such as AUC, CA, F1, precision, and recall for each algorithm

5. Conclusion

The goal of the study was to use machine learning to create a predictive model for diagnosing chronic kidney disease (CKD). The Study's dataset, which included features and cases, was gathered from a hospital. The main objective of the study was to classify real number feature kinds. To guarantee that the predictive models would be accurate, the dataset underwent rigorous preprocessing. For numerical features, the mean and median were used to impute missing data, and for categorical features, the mode. Cleaning and standardizing a few features—like fixing values with special characters or inaccurate representations—was necessary. In order to prepare the data for modeling, this preprocessing step was essential. Following preprocessing, the data was split into training and testing sets, with 30% going toward testing and 70% going toward training. Convolutional Neural Network (CNN), Decision Tree, and Random Forest Classifier and Gradient Boosting (RG) as a stacked model were the three machine learning models chosen for classification.

In order to understand the distribution of the data and spot any possible outliers, the research used data visualization techniques like pair plots and box plots. In order to enhance model performance, the research also used sophisticated methods including Principal Component Analysis (PCA) for dimensionality reduction. Meticulously selected hyperparameters were used throughout model training. XGBoost was applied to the RG model with the following parameters: 100 trees, a regularization parameter of 1, a maximum tree depth of 4, a learning rate of 0.300, and so on. The CNN was made up of a fully connected layer,

max-pooling layers, and convolutional layers with different numbers of filters. The Decision Tree was trimmed to have a maximum depth of three, a minimum drop in impurities of 0.01 and a minimum number of samples needed for leaf nodes and splitting.

Metrics like AUC-ROC, accuracy (CA), F1 score, precision, and recall were used to assess the models. With an AUC of 0.971 and high values for other measures, the results demonstrated that the RG model performed better than the other models. CNN as well as Decision Tree also performed well but with slightly lower scores. Thorough data preprocessing, meticulous hyperparameter tweaking, and the selection of suitable machine learning models were credited with the Study's success. It demonstrated how crucial feature engineering and appropriate data preparation are to the creation of precise predictive models. In order to increase the accuracy and usability of the model, additional real-world data collection and user interface improvement are part of the Study's future scope.

As it wraps up, the study "Detection of chronic kidney disease using machine learning" showcased the potential of machine learning in the field of healthcare and illness prognosis. It was able to create appropriate classification models by using a methodical approach, and it has the potential to be improved upon and used in the real world for renal disease identification.

6. References

1. Adebayo, Ganiyat, Junaid Ghani, and Femi Olatunji. "Machine Learning for Chronic Kidney Disease Prediction: A Review of Recent Advances." Informatics in Medicine Unlocked, vol. 34, 2022, p. 100737.
2. Amin, Muhammad Nazmul Hossain, and Md. Shahjalal Hossain. "A Novel Hybrid Approach for Chronic Kidney Disease Detection Using Machine Learning." Expert Systems with Applications, vol. 200, 2022, p. 116878.
3. Balaji, S., S. Kumar Kumar, and R. Kumar. "A Machine Learning Model for Chronic Kidney Disease Detection." Computer Science Review, vol. 64, 2022, p. 102622.
4. Chen, Jinyang, Li Feng, and Di Liu. "Chronic Kidney Disease Detection Using Deep Learning." Frontiers in Physiology, vol. 13, 2022, p. 904753.
5. Deng, L., Z. Chen, and Y. Wang. "A Comparative Analysis of Machine Learning Algorithms for Chronic Kidney Disease Detection." IEEE Access, vol. 10, 2022, pp. 42226-42235.
6. Dhivya, K., and G. P. Kumari. "A Novel Ensemble Learning Model for Chronic Kidney Disease Detection." International Journal of Engineering and Technology, vol. 9, no. 3, 2022, pp. 234-240.
7. El-Gammal, Mohamed, Mohamed Abdel-Basset, and Dalia El-Shahat. "Machine Learning for Chronic Kidney Disease Prediction: A Systematic Review of Recent Studies." Journal of Computational Science, vol. 56, 2022, p. 102060.
8. Hegazy, H. E., and A. M. Abd El-Latif. "A Hybrid Machine Learning Model for Chronic Kidney Disease Detection and Classification." Journal of Ambient Intelligence and Humanized Computing, vol. 13, no. 4, 2022, pp. 2711-2724.
9. Hu, Yu, Li Zhou, and Xue Zhao. "A Novel Feature Selection Method Based on Deep Learning for Chronic Kidney Disease Detection." IEEE Access, vol. 10, 2022, pp. 25378-25387.
10. Khamparia, Anurag, Prashant Tiwari, and Marwa Alazab. "A Systematic Review of Machine Learning Techniques for Chronic Kidney Disease Detection." Journal of Healthcare Informatics Research, vol. 6, no. 1, 2022, pp. 1-18.

11. Kulkarni, Shubhangi, and Prajakta Patil. "A Novel Deep Learning Model for Chronic Kidney Disease Detection Using Transfer Learning." Biomedical Signal Processing and Control, vol. 74, 2022, p. 103381.
12. Li, Chen, Yanzhi Chen, and Jian Liu. "A Hybrid Machine Learning Model Based on Feature Selection and Deep Learning for Chronic Kidney Disease Detection." IEEE Access, vol. 10, 2022, pp. 32688-32698.
13. Mishra, Sonu, and Anjali Sharma. "A Machine Learning-Based Approach for Chronic Kidney Disease Prediction Using Clinical and Laboratory Data." Computers in Biology and Medicine, vol. 144, 2022, p. 105345.
14. Mohammed, M. A., and E. K. Al-Shammari. "A Novel Machine Learning Model for Chronic Kidney Disease Detection Using Feature Selection and Ensemble Learning." International Journal of Intelligent Systems, vol. 37, no. 9, 2022, pp. 5240-5256.
15. Mou, F., Y. Yang, and X. Huang. "A Hybrid Machine Learning Model Based on Deep Learning and Attention Mechanism for Chronic Kidney Disease Detection." IEEE Access, vol. 10, 2022, pp. 22694-22704.
16. Naser, M. A., M. A. Alomari, and I. M. Alsmadi. "A Novel Machine Learning Model for Chronic Kidney Disease Detection Using Feature Selection and Stacked Ensemble Learning." IEEE Access, vol. 10, 20.