SMIJ
SUSTAINABLE MACHINE INTELLIGENCE JOURNAL

# A Comparative Analysis of Machine Learning Models for Prediction of Chronic Kidney Disease

**Nariman A Khalil[1], Mohamed Elkholy[2,*], and Mohamed Eassa [2]**

[1] Faculty of Information Systems and Computer Science, October 6th University, Cairo, 12585, Egypt; narimanabdo.csis@o6u.edu.eg
[2] Faculty of Information Systems and Computer Science, October 6th University, Cairo, 12585, Egypt; mohamed.elkholy.cs@o6u.edu.eg
[3] Faculty of Information Systems and Computer Science, October 6th University, Cairo, 12585, Egypt; mohamed.eassa.cs@o6u.edu.eg
* Correspondence: mohamed.elkholy.cs@o6u.edu.eg

**Abstract: Prediction of chronic kidney disease (CKD) has emerged as a useful technique for early detection of at-risk persons and the introduction of appropriate management strategies. Machine learning and data-driven methods have been used in predictive modeling to examine massive databases of patient demographics, medical histories, test findings, and genetic information. These cutting-edge methods allow for the profiling of high-risk patients and the tailoring of healthcare administration approaches. Patient outcomes, complication rates, and healthcare system efficiency may all benefit greatly from CKD screening and prediction. Responsible use of CKD prediction algorithms, however, requires resolving issues with data availability, integration, and ethics. The area of medicine has benefited greatly from the use of Machine Learning (ML) methods, which have played an increasingly central role in illness prediction. In this study, we use a strategy that makes use of ML methods to construct effective tools for predicting the development of CKD. Multiple ML models are trained, and their results are compared using a variety of criteria. We applied five ML methods such as logistic regression (LR), Decision tree (DT), random forest (RF), support vector machine (SVM), and k-nearest neighbor (KNN). The LR and KNN have the highest accuracy with 99%.**

| Event | Date |
|---|---|
| Received | 01-07-2023 |
| Revised | 21-09-2023 |
| Accepted | 17-10-2023 |
| Published | 25-10-2023 |

**Keywords: Machine Learning Models, Chronic Kidney Disease, Logistic Regression, Support Vector Machine, Random Forest, KNN, Decision Tree**

## 1. Introduction

Millions of individuals throughout the globe are dealing with Chronic Kidney Disease (CKD), making it a major issue in public health. Chronic kidney disease causes kidney function to deteriorate over time, increasing the risk of cardiovascular disease and kidney failure in advanced stages. To better manage CKD, slow its progression, and improve patient outcomes, it is important to diagnose and anticipate the onset of the illness as early as possible [1], [2].

Various clinical, demographic, and laboratory markers are used to identify those at high risk for the development and progression of CKD. More precise risk stratification and individualized healthcare treatment are possible with the use of machine learning and data-driven techniques, which have recently emerged as useful tools in CKD prediction [3], [4].

Patients' demographics, medical histories, laboratory test results, imaging data, and genetic or proteomic information are all part of the massive datasets that must be analyzed for

CKD prediction. To determine whether people are at risk for developing CKD or experiencing a rapid decline in kidney function, sophisticated methods such as artificial intelligence and predictive modeling algorithms may be used in these massive databases [5], [6].

To delay the course of the illness and lessen the likelihood of complications, CKD patients should be closely monitored and have their kidney function regularly tested. Predicting chronic kidney disease (CKD) may also help healthcare providers minimize costs by prioritizing high-risk individuals for therapies and directing preventative efforts [7], [8].

Predicting chronic kidney disease (CKD) has potential, but it faces obstacles. Predictions may be impacted by the quantity and quality of data available, such as electronic health records and longitudinal follow-up. There are also technological and analytical hurdles associated with making sense of and combining different types of data. Further, for CKD prediction algorithms to be implemented morally, concerns about patient privacy, data security, and appropriate model usage must be addressed [9], [10].

By facilitating early diagnosis, risk stratification, and focused therapies, CKD prediction has the potential to greatly improve patient outcomes. To better identify CKD risk factors and direct individualized healthcare treatment, we may integrate modern data-driven approaches like machine learning and predictive modeling. While there are still obstacles to overcome, continued research and cooperation between healthcare practitioners, researchers, and data scientists can drive the development of accurate and reliable CKD prediction models, ultimately leading to better patient care and better allocation of healthcare resources [11], [12]s.

To describe the process by which computers can automatically process and classify new data based on old data and information, the term "machine learning" was first coined by statistician Arthur Samuel in 1959; today, it is considered to be a subset or subpart of Artificial Intelligence (AI), associated with algorithms that permit processors or computers to do so. Computers may make predictions and decisions on their own without any programming by using mathematical models built by these machine learning algorithms using training data (the current sample data set)[13], [14]. It is not necessary to design and write the code for the entire problem to predict the outcome of a given complex problem statement; rather, it is sufficient to serve the algorithm with the available information, at which point the machine may construct a mathematical model or logic to make the prediction. Further, machine learning may be divided into three major categories: Learning paradigms include unsupervised, supervised, and reinforcement [15], [16].

Training a machine entails feeding it examples of data with labels so that it may learn to make predictions about new data. After this is completed, the machine's accuracy is checked using a set of randomly generated inputs[17], [18]. The concept of "supervision" lies at the heart of the theory behind supervised learning, which seeks to establish a connection between the input data and the output data. Although a lot of manual labor is required to build the model, this approach ultimately allows for quicker execution of a time-consuming process. The subfield of Machine Learning known as "Supervised Machine Learning" has seen widespread implementation [19], [20].

This research will provide an ML-based strategy for evaluating CKD. These are some of the most important results from using this methodology: Effective classification models for predicting the probability of CKD incidence need a data preparation phase to guarantee the dataset cases are distributed in a balanced fashion. Common measures like Precision, Recall, F-measure, and Accuracy are used to provide a comparative analysis of the performance of different models. All of the models were able to achieve outstanding results in a performance assessment.

## 2. Machine Learning Models

This section introduces the data description, then how to process and deal with these data, and finally we introduce the machine learning models to apply these data to it.

### 2.1 Data Description

This part introduces the describes the criteria of the dataset. This data has thirteen criteria and one target name class. Table 1 shows the part of the dataset.

Table 1. The first five rows in the dataset.

| | **CKCF$_1$** | **CKCF$_2$** | **CKCF$_3$** | **CKCF$_4$** | **CKCF$_5$** | **CKCF$_6$** | **CKCF$_7$** | **CKCF$_8$** | **CKCF$_9$** | **CKCF$_{10}$** | **CKCF$_{11}$** | **CKCF$_{12}$** | **CKCF$_{13}$** | **CKCF$_{14}$** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Row$_1$ | 80.0 | 1.020 | 1.0 | 0.0 | 1.0 | 36.0 | 1.2 | 137.53 | 4.63 | 15.4 | 7800.0 | 5.20 | 1.0 | 1 |
| Row$_2$ | 50.0 | 1.020 | 4.0 | 0.0 | 1.0 | 18.0 | 0.8 | 137.53 | 4.63 | 11.3 | 6000.0 | 4.71 | 0.0 | 1 |
| Row$_3$ | 80.0 | 1.010 | 2.0 | 3.0 | 1.0 | 53.0 | 1.8 | 137.53 | 4.63 | 9.6 | 7500.0 | 4.71 | 0.0 | 1 |
| Row$_4$ | 70.0 | 1.005 | 4.0 | 0.0 | 1.0 | 56.0 | 3.8 | 111.00 | 2.50 | 11.2 | 6700.0 | 3.90 | 1.0 | 1 |
| Row$_5$ | 80.0 | 1.010 | 2.0 | 0.0 | 1.0 | 26.0 | 1.4 | 137.53 | 4.63 | 11.6 | 7300.0 | 4.60 | 0.0 | 1 |

We obtain the statistical analysis of the dataset as shown in Table 2.

Table 2. Some statistical analysis in the dataset.

| | **CKCF$_1$** | **CKCF$_2$** | **CKCF$_3$** | **CKCF$_4$** | **CKCF$_5$** | **CKCF$_6$** | **CKCF$_7$** | **CKCF$_8$** | **CKCF$_9$** | **CKCF$_{10}$** | **CKCF$_{11}$** | **CKCF$_{12}$** | **CKCF$_{13}$** | **CKCF$_{14}$** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 400.000 | 400.000 | 400.000 | 400.000 | 400.0000 | 400.0000 | 400.000 | 400.000 | 400.000 | 400.000 | 400.0000 | 400.000 | 400.0000 | 400.0000 |
| **mean** | 76.455 | 1.0177 | 1.01500 | 0.39500 | 0.882500 | 57.4055 | 3.07235 | 137.529 | 4.627 | 12.5269 | 8406.0900 | 4.708275 | 0.369350 | 0.625000 |
| **std** | 13.476 | 0.0054 | 1.27232 | 1.04003 | 0.322418 | 49.28597 | 5.61749 | 9.204 | 2.819 | 2.716 | 2523.2199 | 0.840315 | 0.482023 | 0.484729 |
| **min** | 50.000 | 1.0050 | 0.000000 | 0.00000 | 0.000000 | 1.50000 | 0.40000 | 4.5000 | 2.5000 | 3.100 | 2200.0000 | 2.100000 | 0.000000 | 0.000000 |
| **25%** | 70.000 | 1.0150 | 0.000000 | 0.00000 | 1.000000 | 27.00000 | 0.90000 | 135.000 | 4.000 | 10.875 | 6975.0000 | 4.500000 | 0.000000 | 0.000000 |
| **50%** | 78.000 | 1.0200 | 1.000000 | 0.00000 | 1.000000 | 44.00000 | 1.40000 | 137.530 | 4.6300 | 12.530 | 8406.0000 | 4.710000 | 0.000000 | 1.000000 |
| **75%** | 80.000 | 1.0200 | 2.000000 | 0.00000 | 1.000000 | 61.75000 | 3.07000 | 141.000 | 4.800 | 14.625 | 9400.0000 | 5.100000 | 1.000000 | 1.000000 |
| **max** | 180.000 | 1.0250 | 5.000000 | 5.00000 | 1.000000 | 391.0000 | 76.0000 | 163.000 | 47.000 | 17.800 | 26400.0000 | 8.000000 | 1.000000 | 1.000000 |

### 2.2 Data preprocessing

We used the normalization process to normalize the dataset, due to the gap between the values of the dataset. So, we used the standard scalar algorithm to obtain the normalization dataset.

### 2.3 Machine Learning Models

Machine learning (ML) is a subfield of AI that uses a collection of methods to automatically discover patterns from data while making no assumptions about the data's structure. Methods in machine learning include the well-known neural network, SVM, RF, and GB machines. These methods excel because they can account for nonlinear correlations in the data and the interplay between different variables.

#### 2.3.1 Logistic Regression (LR)

LR was the pioneering model in machine learning (ML). Linear Regression (LR) originates in the Linear Model with a binary outcome variable [21], [22]. The log-odds of the probability p are modeled in an LR with k features.

$$\log\left(\frac{p}{1-p}\right) = \sum_{j=1}^{k} \varphi_j x_j \qquad (1)$$

$$p = \frac{1}{1-e^{-\sum_j \varphi_j x_j}} \qquad (2)$$

Where $\varphi = \varphi_1, \varphi_2, \ldots \ldots \varphi\_k$ refers to the maximum likelihood

#### 2.3.2 Decision Tree (DT)

The Decision Tree method may be used for both classification (with non-continuous output values) and regression (with continuous output values) issues, since it is a supervised machine learning methodology. The structure of a tree was the inspiration for the name of this technique; the characteristics (or conditions) are the branches, and the class labels are the leaves. The DT method's greatest strength is that it can be easily grasped, interpreted, and visualized. The Decision Tree may also be expanded to include more decision-making methods. This approach may also be used to model datasets in which the connection between the output and the input variables is very nonlinear. Its overfitting tendencies and trouble with labeling various output classes are two of its downsides [23], [24].

#### 2.3.3 Random Forest (RF)

The simplest explanation for RF is that it is a collection of Decision Trees (DTs) that forecast the desired output by averaging their individual predictions or by selecting the value with the most votes. To be more specific, RF may be thought of as a method that merges Bagging with Random feature selection via the use of several Decision Trees. In 1995, Ho presented a method for random decision forests, the foundation of this technique. Breiman provided a method in another study that combined his bagging notion with the random feature selection given by Ho, Amit, and Geman. When a big database is utilized for training, the RF method's findings are very accurate, which is one of its numerous benefits over other ML-based approaches. In addition, it may be put into action quickly and easily [23], [24].

#### 2.3.4 Support Vector Machine (SVM)

Alexey ya. Chervonenkis and Vladimir N. Vapnik developed support vector machines (SVM) in 1963. Since the development of Support Vector Machines, this method has found widespread use in the resolution of image, hypertext, and text categorization issues. These sophisticated algorithms have applications in both handwritten text recognition and protein sorting in the lab. They have numerous additional applications as well, like autonomous vehicles, chatbots, facial recognition, etc. The Support Vector Machine (SVM) algorithm is one of the most popular

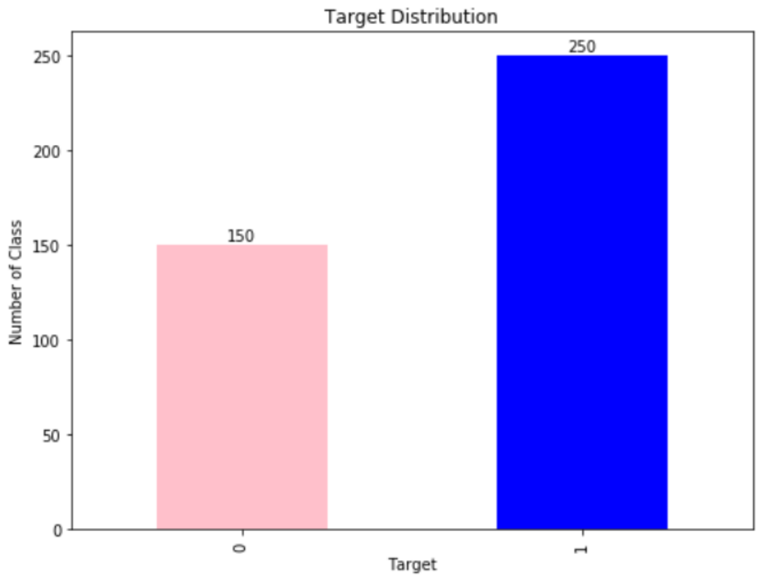

**Figure 1. The target class distribution.**

supervised learning methods and is designed to solve regression and classification issues. The goal of the support vector machine (SVM) method is to generate the best possible decision boundary, or hyperplane, that divides the n-dimensional space into multiple classes, making it simple to assign a new point to the appropriate category. To generate a suitable hyperplane, the SVM algorithm selects extrema in the form of vector points known as Support Vectors[25], [26].

#### 2.3.5 K-Nearest Neighbor (KNN)

When the provided data is somewhat ambiguous, K-nearest-neighbor (K-NN) has the potential to become the major option for implementation as it is one of the most critical and successful algorithms for data segregation. When deciding the probability densities through parametric estimate proved difficult, Evelyn Fix and Joseph Hodges developed this approach in 1951 for use in discriminant analysis. In addition, several properties of this method were determined in 1967; for instance, if 'k' = 1 and 'n' goes to infinity, then the K-NN classification fallacy or mistake is constrained above by double the error rate of Bayes. After establishing such specific attributes and traits, researchers and experimenters labored over time to develop fresh rejection strategies, Bayes error rate improvements, distance-based soft computing processes, and other ways. The K-NN algorithm is one of the most accessible methods in Machine Learning, and it falls under the umbrella of supervised learning. Classification is its primary use, while it may also be used for regressing data. It's a very useful method for resampling data and filling in any blanks. This method for a given data set makes an educated guess as to how the new information relates to the existing information and then assigns the information to the most common classification that seems to fit. As a result, the K-NN algorithm may confidently be used to categorize newly collected data. The new information is ranked according to how its neighbors are arranged in the database. K-NN is often called the lazy learner algorithm since it only stores the data set once and does not begin learning from the training data set until there is a need for classification or prediction of a new data set. In addition, K-NN is non-parametric, meaning that there is no assumed connection between input and output [25], [26].

## 3. Results and discussions

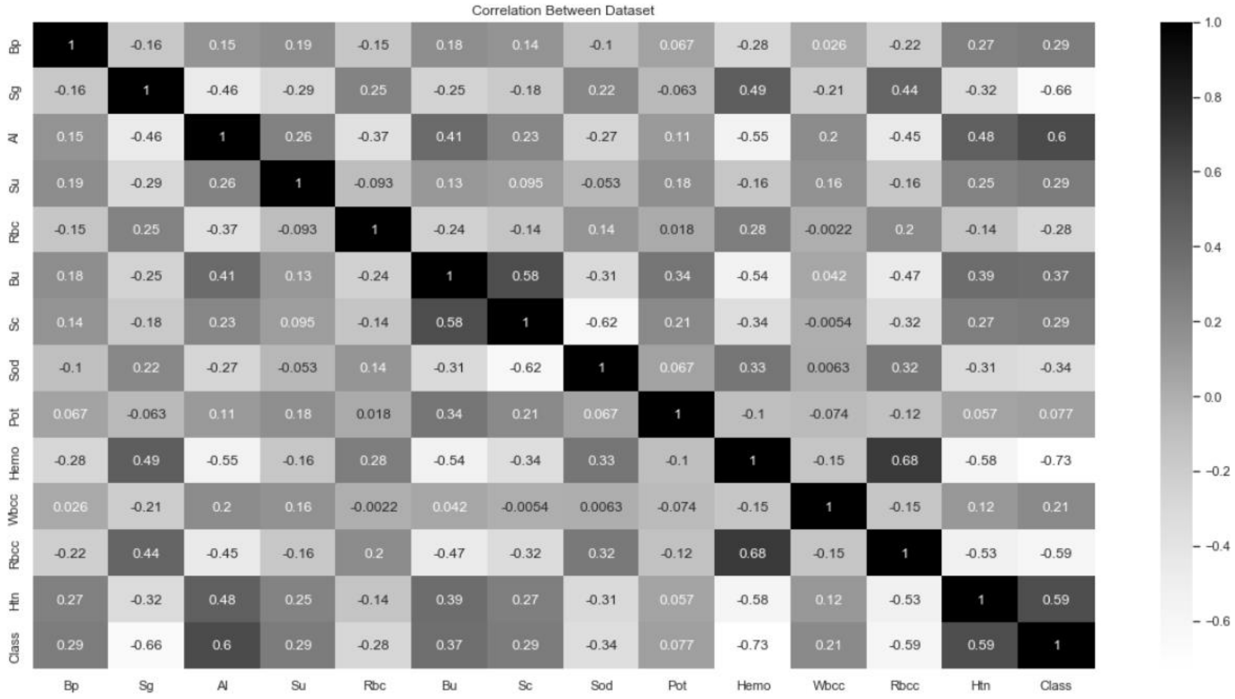

**Figure 2. The correlation matrix between criteria of dataset.**

In this section, we applied the five suggested methods to the dataset to compute the accuracy, recall, and f-measure for every algorithm. We visualize the dataset by introducing the feature on it. Figure 1 shows the count of the target class on 1 and 0. We found that 150 people had the disease and 25 people had the disease. Then we obtain the correlation between all criteria and the target class shown in Figure 2.

We plot the distribution between datasets as shown in Figure 3. Distribution plots give a visual evaluation of the distribution of sample data by comparing the empirical distribution of the data to the theoretical values predicted from a specific distribution. Distribution plots may be used with traditional hypothesis testing to determine the data sample's likely origin. Distribution plots are a graphical representation of the relationship between the observed data and the theoretical distribution parameters. To check whether the data in your sample follows a certain distribution, use distribution plots in addition to more traditional hypothesis testing.

We split the dataset into a train set and a test set. The train set has 85% of the dataset and the test set has 15% of the dataset. We applied the five ML techniques to datasets such as LR, DT, RF, KNN, and SVM.

Precision, recall, accuracy, F-Measure, and area under the curve (AUC) are some of the most often used metrics for evaluating ML models. We may use each indicator to further assess the models.

Accuracy, in particular, is a summary measure of the classification task's success; it is the fraction of data instances that were properly predicted. The percentage of chorine kidney disease cases that were accurately labeled as chorine kidney disease is captured by the recall metric. Accuracy shows the percentage of chorine kidney disease patients that fall into this category. The F-measure is a summary statistic for a model's prediction accuracy; it is calculated as the harmonic mean of the recall and precision statistics. The following is a definition of the metrics:

$$Precision = \frac{A}{A+C} \tag{3}$$

$$Recall = \frac{A}{A+D} \quad (4)$$

$$Accuracy = \frac{A+B}{A+B+C+D} \quad (5)$$

$$F1 - Score = 2\frac{Precision \cdot Recall}{Precision+Recall} \quad (6)$$

Where $A$, $B$, $C$, and $D$ denote $True\ Positive$, $True\ Negative$, $False\ Positive.$ $False\ Negative,$ respectively.

**Table 3. The ML performance analysis**

| | Accuracy | Recall | F1-Score | Precision |
|---|---|---|---|---|
| LR | 0.99 | 0.99 | 0.99 | 0.99 |
| DT | 0.9833 | 0.975 | 0.9876 | 0.99 |
| RF | 0.9833 | 0.975 | 0.9876 | 0.99 |
| SVM | 0.9833 | 0.975 | 0.9876 | 0.99 |
| KNN | 0.99 | 0.99 | 0.99 | 0.99 |

We applied the five ML models to the chorine kidney disease to obtain the evaluation matrices as shown in Table 3. The LR and KNN have the highest accuracy scores. The LR and KNN have the highest recall scores. The LR and KNN have the highest f1-score.

## 4. Conclusions

CKD prognosis has the potential to dramatically alter the way this common disease is treated and managed. To slow the course of CKD and lessen the likelihood of problems, doctors may use cutting-edge methods like machine learning and data-driven approaches to pinpoint patients most in need of individualized treatment. Better patient outcomes and more efficient use of healthcare resources may result from detecting and predicting CKD early on, which allows for prompt treatments like changing patients' diets and dosages of their medications.

The current study describes a supervised learning-based technique to develop accurate models for forecasting the likelihood of CKD recurrence, with a particular emphasis on probabilistic, tree-based, and ensemble learning-based models. We also compared LR, DT, RF, SVM, and KNN methods. The LR and KNN have the highest accuracy, precision, recall, and f1-score with a value of 0.99.

## Author Contributions

All authors contributed equally to this work.

## Funding

This research was conducted without external funding support.

## Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

## Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

## Data Availability Statement

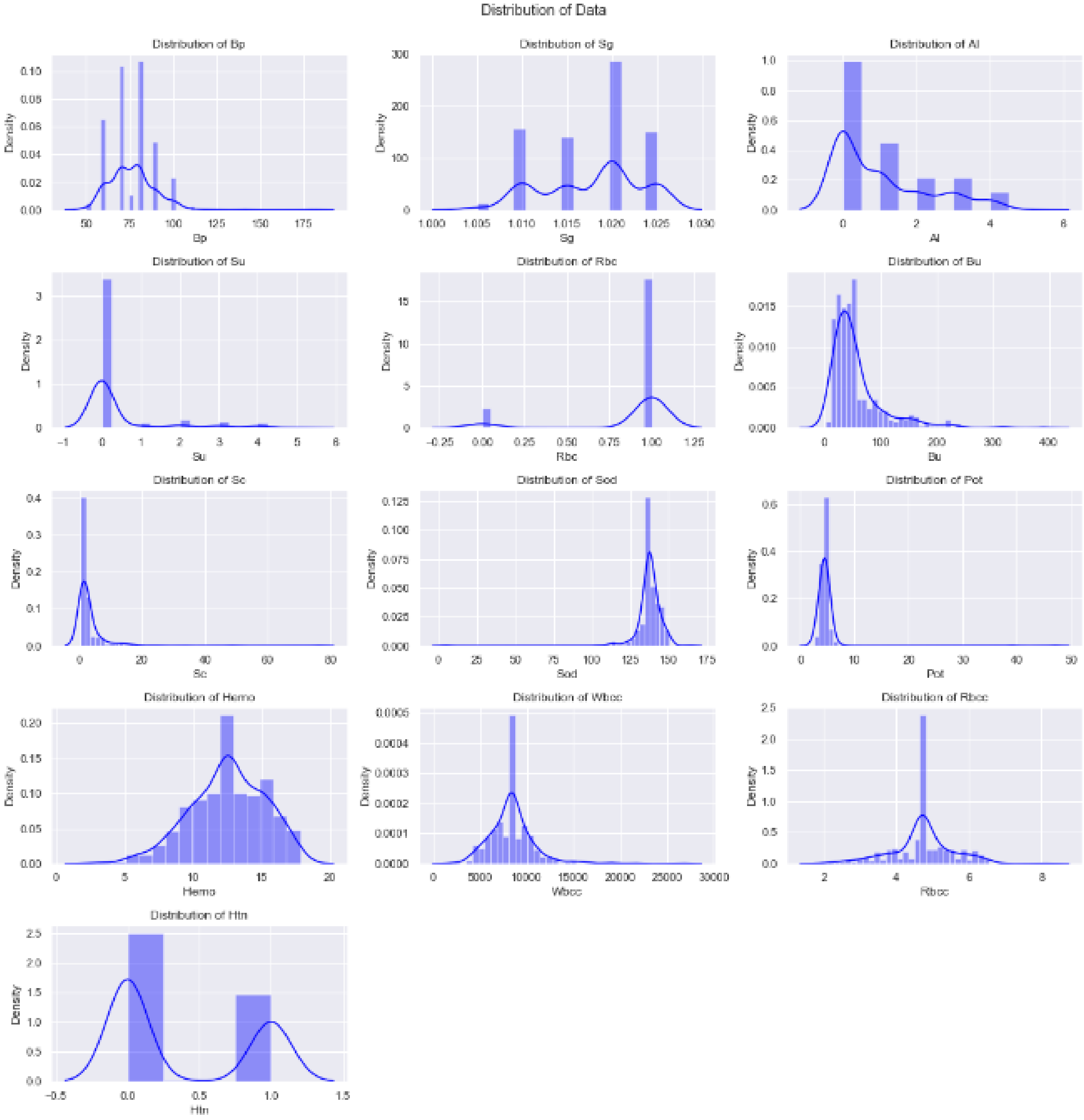

**Figure 3. The distribution plot of dataset.**

All data supporting the findings of this study are included in the paper.

## References

[1]. S. Mihai et al., “Inflammation-related mechanisms in chronic kidney disease prediction, progression, and outcome,” Journal of immunology research, vol. 2018, 2018.

[2]. P. Sinha and P. Sinha, “Comparative study of chronic kidney disease prediction using KNN and SVM,” International Journal of Engineering Research and Technology, vol. 4, no. 12, pp. 608–612, 2015.

[3]. I. A. Pasadana et al., “Chronic kidney disease prediction by using different decision tree techniques,” in Journal of Physics: Conference Series, IOP Publishing, 2019, p. 12024.

[4]. S. Revathy, B. Bharathi, P. Jeyanthi, and M. Ramesh, “Chronic kidney disease prediction using machine learning models,” International Journal of Engineering and Advanced Technology, vol. 9, no. 1, pp. 6364–6367, 2019.

[5]. P. Yildirim, “Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction,” in 2017 IEEE 41st annual computer software and applications conference (COMPSAC), IEEE, 2017, pp. 193–198.

[6]. P. Chittora et al., “Prediction of chronic kidney disease-a machine learning perspective,” IEEE Access, vol. 9, pp. 17312–17334, 2021.

[7]. I. U. Ekanayake and D. Herath, “Chronic kidney disease prediction using machine learning methods,” in 2020 Moratuwa Engineering Research Conference (MERCon), IEEE, 2020, pp. 260–265.

[8]. N. Tangri et al., “A predictive model for progression of chronic kidney disease to kidney failure,” Jama, vol. 305, no. 15, pp. 1553–1559, 2011.

[9]. F. Aqlan, R. Markle, and A. Shamsan, “Data mining for chronic kidney disease prediction,” in IIE Annual Conference. Proceedings, Institute of Industrial and Systems Engineers (IISE), 2017, pp. 1789–1794.

[10]. R. Misir, M. Mitra, and R. K. Samanta, “A reduced set of features for chronic kidney disease prediction,” Journal of pathology informatics, vol. 8, no. 1, p. 24, 2017.

[11]. L. Antony et al., “A comprehensive unsupervised framework for chronic kidney disease prediction,” IEEE Access, vol. 9, pp. 126481–126501, 2021.

[12]. D. A. Debal and T. M. Sitote, “Chronic kidney disease prediction using machine learning techniques,” Journal of Big Data, vol. 9, no. 1, pp. 1–19, 2022.

[13]. M. A. Islam, M. Z. H. Majumder, and M. A. Hussein, “Chronic kidney disease prediction based on machine learning algorithms,” Journal of Pathology Informatics, vol. 14, p. 100189, 2023.

[14]. N. Tangri et al., “Risk prediction models for patients with chronic kidney disease: a systematic review,” Annals of internal medicine, vol. 158, no. 8, pp. 596–603, 2013.

[15]. L. J. Rubini and P. Eswaran, “Generating comparative analysis of early stage prediction of Chronic Kidney Disease,” International Journal of Modern Engineering Research (IJMER), vol. 5, no. 7, pp. 49–55, 2015.

[16]. P. G. Scholar, “Chronic kidney disease prediction using machine learning,” International Journal of Computer Science and Information Security (IJCSIS), vol. 16, no. 4, 2018.

[17]. J. Rysz, A. Gluba-Brzózka, B. Franczyk, Z. Jabłonowski, and A. Ciałkowska-Rysz, “Novel biomarkers in the diagnosis of chronic kidney disease and the prediction of its outcome,” International journal of molecular sciences, vol. 18, no. 8, p. 1702, 2017.

[18]. J. Singh, S. Agarwal, P. Kumar, D. Rana, and R. Bajaj, “Prominent features based chronic kidney disease prediction model using machine learning,” in 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), IEEE, 2022, pp. 1193–1198.

[19]. R. Devika, S. V. Avilala, and V. Subramaniyaswamy, “Comparative study of classifier for chronic kidney disease prediction using naive bayes, KNN and random forest,” in 2019 3rd International conference on computing methodologies and communication (ICCMC), IEEE, 2019, pp. 679–684.

[20]. A. Maurya, R. Wable, R. Shinde, S. John, R. Jadhav, and R. Dakshayani, “Chronic kidney disease prediction and recommendation of suitable diet plan by using machine learning,” in 2019 International Conference on Nascent Technologies in Engineering (ICNTE), IEEE, 2019, pp. 1–4.

[21]. N. Srimaneekarn, A. Hayter, W. Liu, and C. Tantipoj, “Binary response analysis using logistic regression in dentistry,” International Journal of Dentistry, vol. 2022, 2022.

[22]. A. Bailly et al., “Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models,” Computer Methods and Programs in Biomedicine, vol. 213, p. 106504, 2022.

[23]. H. Dabiri, V. Farhangi, M. J. Moradi, M. Zadehmohamad, and M. Karakouzian, “Applications of Decision Tree and Random Forest as Tree-Based Machine Learning Techniques for Analyzing the Ultimate Strain of Spliced and Non-Spliced Reinforcement Bars,” Applied Sciences, vol. 12, no. 10, p. 4851, 2022.

[24]. A. F. Ibrahim, A. Abdelaal, and S. Elkatatny, “Formation resistivity prediction using decision tree and random forest,” Arabian Journal for Science and Engineering, vol. 47, no. 9, pp. 12183–12191, 2022.

[25]. M. Bansal, A. Goyal, and A. Choudhary, “A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning,” Decision Analytics Journal, vol. 3, p. 100071, 2022.

[26]. E. Utami, “Comparison Naïve Bayes Classifier, K-Nearest Neighbor And Support Vector Machine In The Classification Of Individual On Twitter Account,” Jurnal Teknik Informatika (JUTIF), vol. 3, no. 3, pp. 673–680, 2022