Chronic Kidney Disease Prediction Using Machine Learning

Chamandeep Kaur [1], M. Sunil Kumar [2], Afsana Anjum [1], M. B. Binda [3], Maheswara Reddy Mallu [4], and Mohammed Saleh Al Ansari [5,*]

[1] Department of Computer Science & Information Technology, Jazan University, Jizan, Saudi Arabia; Email: cgourmeat@jazanu.edu.sa (C.K.), aisrar@jazanu.edu.sa (A.A.)
[2] School of Computing, Department of CSE, Mohan Babu University & Sree Vidyanikethan Engineering College Tirupati, AP, India; Email: sunilmalchi1@gmail.com (M.S.K.)
[3] Traffic Signal Division, Keltron Communication Complex, Thiruvananthapuram, Kerala, India
[4] Department of Biotechnology, Koneru Lakshmaiah Education Foundation, Vaddeswaram-522302, Guntur, Andhra Pradesh, India; Email: mahesh_bt@kluniversity.in (M.R.M.)
[5] Department of Chemical Engineering, University of Bahrain, Bahrain
*Correspondence: malansari.uob@gmail.com (M.S.A.A.)

***Abstract*—The occurrence of Chronic Renal Disease (CRD), is also referred to as Chronic Kidney Disease (CKD). It depicts a medical condition that harms the kidneys and has an impact on a person's overall health. End-stage renal disease and the patient's eventual mortality can result from improper disease diagnosis and treatment. In the field of medical science, Machine Learning (ML) techniques have become a valuable tool and play a significant role in disease prediction. The development and validation of a predictive model for the prognosis of chronic renal disease is the aim of the proposed study. A dataset on chronic kidney disease with 400 samples was taken from the UCI Machine Learning Repository. Three machine learning classifiers—Logistic Regression (LR), Decision Tree (DT), and Support Vector Machine (SVM)—were used for analysis, and the bagging ensemble method was used to enhance the model's performance. The machine learning classifiers were trained using the clusters of the dataset for chronic renal disease. The Kidney Disease Collection is then compiled using non-linear features and categories. The decision tree produces the best results, with an accuracy of 95%. Finally, we achieve the greatest accuracy of 97% by using the bagging ensemble approach.**

***Keywords*—chronic renal disease, classification algorithms, random forest classifier, machine learning**

I. Introduction

Chronic Kidney Disease (CKD) poses a significant risk to both one's physical well-being and overall quality of life. When the glomerular filtration rate (GFR) drops, it is possible to treat the complications that arise, which can reduce the likelihood of developing cardiovascular disease and boost the chances of survival. Laboratory tests that are performed regularly can be used to diagnose and treat chronic kidney disease. Treatments for a decreased GFR and the complications that come with it can help delay the onset of the disease, stop it altogether, or even prevent it entirely. The use of tobacco, the adoption of unhealthy eating patterns, the failure to get enough sleep, and a host of other risk factors can all play a role in the development of Chronic Kidney Disease. In the year 2016, this disease affected more than 700 million people all over the world, with 417 million of those victims being female and 336 million of those victims being male. Kidney failure may develop as the disease worsens. Creatinine levels in the serum and urine analyses are used in the diagnostic process currently being utilized. This is accomplished through the use of a wide range of medical procedures, some of which include screening and ultrasound methods [1]. Before any tests are carried out on a patient, screening is performed to check for hypertension, a previous history of cardiovascular disease, an active illness, and a family history of kidney disease. Using this method, it is possible to estimate GFR based on the Albumin-to-Creatinine Ratio (ACR) of a urine sample taken first thing in the morning, in addition to the serum creatinine level of the model. This research uses machine learning methods such as decision trees, random forests, and K-Nearest Neighbor (KNN) to improve the accuracy of prediction. This is accomplished by reducing the total number of features while simultaneously selecting the most essential features.

The kidney is composed of two organs, each roughly the size of a human fist. One kidney is located in each of the rib cages. For the kidneys to be able to produce one to two quarts of urine per day, they have to filter between 120 and 150 quarts of blood every single day [2]. Urination is the primary means by which the kidneys carry out their primary function, which is to rid the body of waste and excess fluid. Urine is produced through processes that involve the excretion of excess water and waste as well as the reabsorption of waste. This procedure is essential to preserve the body's delicate chemical equilibrium. The kidneys are responsible for a

Manuscript received September 19, 2022; revised October 12, 2022; accepted December 21, 2022; published April 26, 2023.

significant portion of the regulation of the levels of acid, potassium, and salt that are found within the body. In addition, the kidneys are responsible for the production of hormones that influence how the body's other organs function.

There are five stages of CKD as follows:

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

Chronic kidney disease is a condition in which the function of the kidneys gradually declines over time. It is estimated that chronic kidney disease affects 14 percent of the population [3]. Chronic renal disease is responsible for the deaths of more people than breast or prostate cancer combined. This is even though over two million people have kidney failure and require dialysis or a kidney transplant. A hormone that is produced by the kidneys is responsible for controlling many bodily functions, including the production of red blood cells, blood pressure, and the metabolism of calcium.

eGFR can be affected by several factors including age, gender, race, and creatinine levels [4]. eGFR is used as the primary metric for stage classification in CKD. There are five distinct stages of function that the kidneys go through [5]. Nevertheless, stage 3 accounts for the vast majority of cases. Both Stage 1 and Stage 2 exhibit a moderate decline in their respective abilities to perform. Several distinct data sets have been utilized in the research on machine learning algorithms for kidney disease prediction. The development of sensor networks, communication technologies, data science, and statistical processing has made ML techniques crucial tools for several health-related applications, including the early diagnosis of several chronic conditions, the development of pervasive (assisted) living environments (smart homes) based on the Internet of Things (IoT), the detection of elderly falls, and others. The following diseases have some of the following characteristics: COVID-19, Hypertension, Stroke, Diabetes, Cholesterol, Chronic Obstructive Pulmonary Disease (COPD), Acute Liver Failure, Cardiovascular Diseases (CVDs), Acute Lymphoblastic Leukemia, Cancer, etc.

The classification and regression algorithms that are used in machine learning are two of the most important aspects of this field [6]. Using machine learning, it is possible to accurately predict the stages of CKD as well as the presence of CKD. One of them is the UCI dataset [7], which is also sometimes referred to as the UCI dataset. This particular research makes use of the standard dataset that was discussed earlier, just like the vast majority of other related studies. Clinical data for CKD should be analyzed for missing attributes, and the best way to handle them is determined by how randomly they were overlooked. This is because missing attributes can be caused by a variety of factors. In addition, the sample size for the UCI data collection is relatively small, totaling 400 cases, and there are 25 features in total [8]. In this particular instance, the most likely explanations involve either a lack of completeness in the data collection or redundant (closely related) characteristics.

Several prediction models, including Random Forest (RF), Decision Tree, Logistic Regression (LR), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) were compared in this study. This study also addresses the difficulties that arise when attempting to evaluate CKD data when there are missing values, and it does so by employing a novel method and conducting a comparison of multiple methods using the dataset from the UCI. This study highlights the importance of statistical analysis and feature-specific domain knowledge when making predictions about chronic kidney disease based on clinical data. Specifically, the study focuses on the relationship between the two.

II. Related Work

In the system that is detailed in, data mining strategies like Random Forest and Back Propagation neural networks have been utilized. A comparison of the two approaches reveals that the Back Propagation method, which makes use of a supervised learning network, is the more successful of the two approaches. In the beginning, Mohammed Elhoseny put forward the idea of treating CKD with an approach that depends on density-based feature selection and ACO. Developing a machine learning system should make use of methods such as Support Vector Machines, K Nearest Neighbor, Logistic Regression, Naive Bayes, Random Forested, Decision Trees, and Multi-Layer Perceptron [9]. The authors also suggested using these methods. Examining the results of each method's recall, accuracy, and precision helps determine how successful each one is. In the end, Random Forest is what's used to put the system into action.

In predicted illness using Boosting Classifiers, Ant-Miner, and Decision Trees [10]. Recall, accuracy, and precision are the three components that go into determining how effective each is. The Random Forest algorithm is then used to construct the system. One of the primary goals of the study is to identify Chronic Kidney Disease, and the other is to establish correlations between the many different CKD characteristics. During the tests, it was discovered that Logit Boost performed significantly better than AdaBoost.

In Shinde and Rajeswari's work, CKD can be predicted by making use of an extreme learning machine in conjunction with an ACO [11]. Because ELM has limitations when it comes to optimization, classification is done with a MATLAB program rather than using ELM. This tactic functions more effectively with SLFNs that have a sigmoid additive structure. An artificial intelligence system that is based on a decision tree and an SVM algorithm was described [12]. When compared to the other method, the support vector machine performs significantly better. As a direct result of this technique's predictive capabilities, medical professionals can complete patient evaluations in a shorter amount of time.

An example of a prediction system that makes use of backpropagation neural networks was presented by Nilesh Borisagar. Levenberg, Scaled Conjugate, Bayesian Regularization, and Robust Back Propagation Algorithms are some of the topics that are covered in this investigation. To complete this project, Matlab R2013a was utilized. It has been discovered that scaled conjugate gradient and resilient backpropagation require less time for training than Levenberg and Bayesian regularization. Dua and Graff proposed a method to forecast CKD by making use of the data mining capabilities of Hadoop [12]. Two different data mining classifiers are built on top of SVM and KNN respectively. In this particular instance, data columns are chosen explicitly by hand for the prediction analysis.

In this particular system, the accuracy of the KNN classifier is superior to that of the decision tree classifier. The author suggested a method that would automatically evaluate and compute the results of a patient's renal illness [13]. In this scenario, a prediction method based on rules is utilized. Mathematical calculations were performed using a neuro-fuzzy method to arrive at the results. A clustering method that makes use of multiple pheromone tables and is based on ACO was proposed by Kai-Cheng Hu in the year 2015. This challenge could be broken down into some distinct patterns, each of which was based on a different set of criteria [14]. The use of two different pheromone tables is required to keep track of both positive and negative information. One table is used for keeping track of positive information, while the other table is used for storing negative information.

III. Dataset and Methods

A method of artificial intelligence called machine learning enables learners to process information without having to be explicitly programmed. It focuses on producing computer programmers who can change in response to fresh data. It can be classified as either supervised or unsupervised [15]. It all comes down to combining the proper characteristics to create frameworks that achieve the proper objectives. Examples of these tasks include multi-dimensional and multi-classification, predictive clustering, and parametric modeling [16].

Three main steps are involved in the proposed methodology: preprocessing of the data, training of the models, and model selection (Fig. 1).

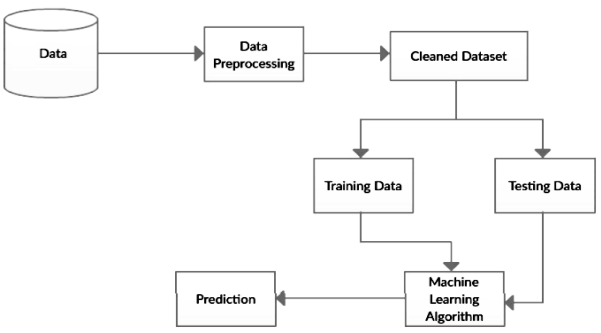

Figure 1. Proposed methodology.

A. Dataset

TABLE I. Features Listed in the CKD Dataset

#	Column	Non-null Count	Datatype
0	id	400 non_null	int64
1	age	391 non_null	float64
2	dias_blood_pressure	388 non_null	float64
3	ur_specific_gravity	353 non_null	float64
4	ur_albumin	354 non_null	float64
5	ur_sugar	351 non_null	float64
6	red_blood_cells	248 non_null	object
7	ur_pus_cell	335 non_null	object
8	ur_pus_cell clumps	396 non_null	object
9	ur_bacteria	396 non_null	object
10	blood glucose random	356 non_null	float64
11	blood urea	381 non_null	float64
12	serum creatinine	383 non_null	float64
13	sodium	313 non_null	float64
14	potassium	312 non_null	float64
15	hemoglobin	348 non_null	float64
16	packed cell volume	330 non_null	object
17	white blood cell count	295 non_null	object
18	red blood cell count	270 non_null	object
19	hypertension	398 non_null	object
20	diabetes	398 non_null	object
21	coronary artery disease	398 non_null	object
22	appetite	399 non_null	object
23	pedal edema	399 non_null	object
24	anemia	399 non_null	object
25	class	400 non_null	object
Datatypes: float64(11), int64(1), object(14)			
Memory usage: 81.4+ KB			

It is possible to use machine learning to make predictions about chronic renal disease by downloading a dataset from the Kaggle competition. The dataset contained information on a total of 400 different patients' records. The ages of the people involved, bacteria, serum creatinine, white blood cell count, potassium, albumin, and red blood cell count are also included on the list of 25 factors. Patients frequently exhibit erratic and unpredictable patterns regarding their blood glucose and urea levels, as well as their classification, appetite, and packed cell volume. Diabetes and high blood pressure are the two primary contributors to chronic kidney disease (CKD) [17]. We should prepare ourselves for high blood sugar levels as a natural consequence of the damage that diabetes causes to our many organs. It is of the utmost importance that the patient's condition is ascertained as quickly as possible. Within the scope of this study, several different approaches to machine learning were modified to forecast the illness.

B. Data Processing

In this particular investigation, the process of data preparation was broken down into two stages. To get started, we got rid of all of the attributes that had more than twenty percent of their data missing (see Table II). As a direct consequence of this fact, this particular set of

characteristics is not explored in the research. During the second stage of the data preparation process, we completed the task of filling in the values that were absent from the remaining data. During the pre-processing phase, they are required to manage missing data by their distributions. This is done to ensure an acceptable level of accuracy. During this inquiry, Little's MCAR test was utilized to demonstrate that the missing numbers exhibited erratic behavior. Depending on what took place, this can either be a positive or a negative bias. To evaluate the analytical approaches that can be utilized to complete the missing information in multivariate quantitative data, a chi-square test of MCAR [18] is utilized. investigates the possibility that the means of the various missing-value patterns are, in fact, quite different from one another.

TABLE II. AFTER DATA PROCESSING

#	Column	Non-null Count	Datatype
0	id	366 non_null	int64
1	age	366 non_null	float64
2	dias_blood_pressure	366 non_null	float64
3	ur_specific_gravity	366 non_null	float64
4	ur_albumin	366 non_null	float64
5	ur_sugar	366 non_null	float64
6	blood glucose random	366 non_null	float64
7	blood urea	366 non_null	float64
8	serum creatinine	366 non_null	float64
9	sodium	366 non_null	float64
10	potassium	366 non_null	float64
11	hemoglobin	366 non_null	float64
12	packed cell volume	366 non_null	float64
13	white blood cell count	366 non_null	float64
14	red blood cell count	366 non_null	float64
15	diabetes	366 non_null	uint8
16	anemia	366 non_null	uint8
17	CKD	366 non_null	uint8
18	pedal edema	366 non_null	uint8
19	poor	366 non_null	uint8
20	hypertension	366 non_null	uint8
21	coronary artery disease	366 non_null	uint8
22	abnormal_red_blood_cells	366 non_null	uint8
23	abnormal_ur_pus_cell	366 non_null	uint8
24	ur_pus_cell_clumps_present	366 non_null	uint8
25	Ur_bacteria_present	366 non_null	uint8
Datatypes: float64(14), int64(1), uint8(11)			
Memory usage: 49.7 KB			

Data that has been converted into a format that a machine can understand can be comprehended quickly and easily by the machine. The term "dataset" is used to refer to a collection of individual data elements [19]. Criteria such as the mass or time at which an event is guaranteed can be used to facilitate the identification and assurance of fundamental properties of data items. There is a good chance that the dataset contains missing values; these can either be calculated or removed. The value of the mean, median, or mode of the associated characteristics can be used to fill in the blanks when dealing with missing data [20]. This is the most common method for dealing with missing data. A conversion from object-typed numerical numbers to float 64 values is required before analysis can be performed. When dealing with categorical attributes that contain null values, the value that appears in the attribute column the most frequently is substituted for the null value. The transformation of categorical data into numeric properties can be accomplished through the use of label encoding [21]. This involves giving each attribute value its integer value. As a direct consequence of this, an int data type will be generated immediately. Calculations are made in advance to determine the mean values of each column, and those values are then used to fill in any gaps in the respective attribute column. It is possible to calculate the mean value for each column by utilizing the classifier function. After the data has been replaced and encoded, it needs to go through the processes of training, verification, and testing. Our algorithms acquire the knowledge necessary to construct a model through the process of learning from the data that we provide for them. The validation portion of the dataset is utilized by us to check the accuracy of the multiple model fits that we have created and to enhance the model [22].

C. Feature Selection

Because our output or prediction variable has the most significant computational impact on our feature selection process, we look for features that have the most significant impact and prioritize them accordingly. In this particular study, it was utilized to determine which dataset attributes were of the utmost significance. The Ant Colony Optimization (ACO) algorithm was implemented [21]. It is a method for solving computational problems by locating paths through graphs that are both effective and efficient. Artificial Ants and other local search algorithms are typically used in conjunction with one another in modern graph-based optimization projects. The term "Artificial Ants" refers to multi-agent systems that attempt to simulate the behavior of ants [24]. Pheromones are the primary mode of communication utilized by biological ants; this is the standard practice. A classification approach is used to evaluate the performance of the subsets, which is known as the wrapper evaluation function. This is because pheromone intensities are evaluated at each iteration rather than cumulatively. Using this strategy, we will first select the most productive ants, and then we will update a subset of their attributes.

D. Classifiers

1) Decision tree

The most essential components of a decision tree are the tree's trunk, its nodes, and its branches. It is a graphical representation of a particular decision situation that is included in predictive models. In fields of medicine with a large number of factors to take into consideration, the use of decision trees has become increasingly common. Out of all the different machine learning techniques, decision trees are by far the most effective [25]. These unmistakably reflect important

facets of the data collection process that took place earlier. They also have the potential to produce the characteristic that has the greatest influence on the lives of the vast majority of people. Entropy is the foundation upon which the decision tree is constructed, and the information gained from the dataset demonstrates just how essential it is. The use of decision trees comes with a variety of drawbacks, the most notable of which is overfitting and a greedy strategy [26]. Because it required a large number of nodes to divide the data, using a decision tree to split datasets aligned to axes led to overfitting. This was because the tree required a large number of nodes. According to J48, it is impossible to generate exponentially more trees using a dynamic approach as opposed to a greedy approach because of the practical difficulties involved [27].

2) Random forest classifier

During the training phase of random forests, which are also referred to as random decision forests or RDF for short, a significant number of decision trees are produced. As a consequence of this, the vast majority of trees will choose the appropriate category when confronted with classification problems [28]. In regression tasks, random choice forests determine the average or mean forecast for each tree. Decision trees have the propensity to produce results that are too good for their training data, so this approach is appropriate [29]. When compared to random forests, gradient-boosted trees perform significantly better than decision trees in the majority of instances. It is possible, on the other hand, that the data attributes will have varying degrees of usefulness.

3) KNN classifier

The K-Nearest Neighbors (KNN) algorithm can make predictions about the values of new data points by comparing those points to the data in the training set and determining how similar they are [30]. The following is a list of the steps that need to be taken for us to figure it out.

Step 1: Any algorithm must be implemented with a dataset, hence the initial step of KNN loads both testing and training the data.

Step 2: The following phase involves choosing the K value, or the closest data points. Any integer, K, may exist.

Step 3: Perform the actions listed below for each test data point.

- Use Euclidean, Manhattan, or Hamming distances to determine the distance between each row of training and test data. This method is commonly used for distance calculation.
- Sort the test points in ascending order using the distance value.
- Next, select the top K rows from the array.
- Based on these rows, classify the test points.

E. Prediction

1) Prediction using decision tree

Using the CKD dataset for training, the Decision tree model predicts the following Table V:

TABLE III. DECISION TREE MODEL

Classification Report:-				
	precision	recall	f1-score	support
0	0.93	0.99	0.96	72
1	0.98	0.90	0.93	48
accuracy			0.95	120
macro avg	0.96	0.94	0.95	120
weighted avg	0.95	0.95	0.95	120

2) Prediction using random forest

Using the CKD dataset for training, the Random Forest model predicts the following Table IV:

TABLE IV. RANDOM FOREST MODEL

Classification Report:-				
	precision	recall	f1-score	support
0	0.95	1.00	0.97	72
1	1.00	0.92	0.96	48
accuracy			0.97	120
macro avg	0.97	0.96	0.96	120
weighted avg	0.97	0.97	0.97	120

3) Prediction using KNN

Using the CKD dataset for training, the KNN model predicts the following Table V:

TABLE V. KNN MODEL

Classification Report:-				
	precision	recall	f1-score	support
0	0.79	0.78	0.78	72
1	0.67	0.69	0.68	48
accuracy			0.74	120
macro avg	0.73	0.73	0.73	120
weighted avg	0.74	0.74	0.74	120

IV. RESULT AND DISCUSSION

These findings served as the basis for selecting the methods that provided the highest level of accuracy across all three datasets in Tables III–V. There are many different kinds of classifiers, including KNN, decision trees, and random forests. Following the conclusion of the analysis of the significance of the selected features for each type of prediction, a choice needs to be made. A calculation of the standard deviation of the importance of features is performed for every algorithm. The preferences of the algorithm are presented in Fig. 2 for the various characteristics. Comparing the decision tree classifier to the random forest classifier reveals that the decision tree classifier has the least amount of feature bias.

Overall CKD characteristics like hunger, anemia and pedal edema are over-represented in the data even though the distribution covers the entire CKD spectrum. Despite the ease with which this data set can be used to accurately forecast, the recall column in Tables III–V shows that

doing so may lead to false positives in a broader context. Because the missing data were completely lost at random, it was impossible to achieve perfect accuracy without replacing them with values from a collaborative imputer rather than a constant. Depending on the stage of a patient's development, some features have a weaker link to medical value than others. Model accuracy is strongly influenced by the training process. Except for serum creatinine, all of the selected attributes have a clear class differentiation, which the distribution of the data set can be used to support. Finally, as shown in Fig. 2, certain trained models favor certain features when selecting the algorithm. There are many more options than CKD to consider when you take into account the factors that changed their nominal values. Consequently, the use of an additional tree classifier encourages decision-makers to consider multiple factors rather than just one, which is why it was chosen.

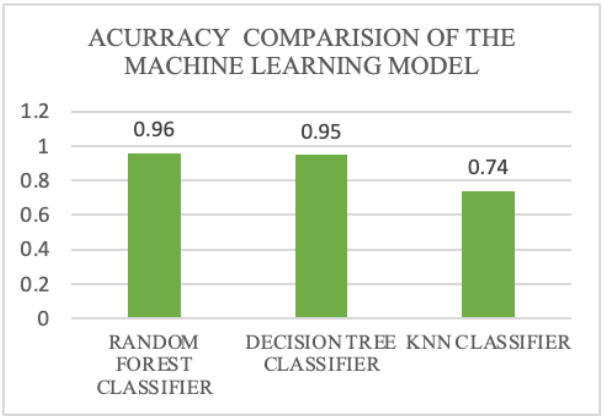

Figure 2. Accuracy comparison based on CKD prediction.

V. Conclusion

Progressive loss of kidney function over time is a feature of chronic renal disease. Since the majority of victims show no symptoms, it is a quiet illness. The medical community has severe difficulty in the early detection and treatment of CKD, and they turn to machine learning theory to develop an effective answer. People would be able to detect it early and receive treatment with the least amount of risk and expense if they could accurately predict it with one hundred percent certainty. More than 14 percent of the world's population is affected by CKD.

Random forest classifiers can reduce the number of features in the prediction algorithm, which could lead to fewer medical tests being required by filling in missing values and combining other variables. This could be achieved by combining other variables and filling in missing values. This new methodology includes a variety of components, including the preparation of data, the handling of missing values, the selection of features, and the prediction of the CKD status based on the features. Random forests and decision trees are two examples of superior algorithms; both of these types of algorithms have a high level of accuracy and are influenced solely by the characteristics being considered. This study demonstrates that having domain knowledge is essential for correctly interpreting clinical data related to CKD.

As a result, it might be worthwhile to conduct research into the application of a Random forest model to manage missing values in datasets in the future that are related to a variety of diseases. In addition, by including data on genetics, water consumption, and the types of food consumed in the research, we can learn more about CKD.

Conflict of Interest

The authors state that they are aware of no personal or financial conflicts that might have appeared to have an impact on the research provided in this study.

Author Contributions

Chamandeep Kaur wrote and revised the final manuscript. M. Sunil Kumar and Afsana Anjum Literature Review. M. B. Binda implemented. Maheshwara Reddy Mallu analyzed the data. Mohammed Saleh Al Ansari data collection. All authors had approved the final version.

References

[1] K. B. Naidu, B. R. Prasad, S. M. Hassen, *et al*., "Analysis of Hadoop log file in an environment for dynamic detection of threats using machine learning," *Elsevier Measurement: Sensors*, vol. 24, pp. 1–5, 2022.

[2] E. Dritas and M. Trigka, "Machine learning techniques for chronic kidney risk prediction," *Big Data and Cognitive Computing*, pp. 1–15, 2022.

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

[4] Your Kidneys & How They Work. NIDDK. [Online]. Available: https://www.niddk.nih.gov/health-information/kidneydisease/kidneys-how-they-work

[5] Kidney disease: The basics. (Aug. 2014). [Online]. Available: https://www.kidney.org/news/newsroom/factsheets/KidneyDiseaseBasics

[6] Global facts: About kidney disease. [Online]. Available: https://www.kidney.org/kidneydisease/global-facts-about-kidneydisease/

[7] Facts about chronic kidney disease. (May 2020). [Online]. Available: https://www.kidney.org/atoz/content/about-chronic-kidney-disease

[8] J. Xiao, R. Ding, X. Xu, H. Guan, X. Feng, T. Sun, S. Zhu, and Z. Ye, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," *Journal of Translational Medicine*, vol. 17, no. 1, p. 119, 2019.

[9] E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informatics in Medicine Unlocked*, vol. 15, 100178, 2019.

[10] J. Aljaaf, D. Al-Jumeily, H. M. Haglan, M. Alloghani, T. Baker, A. J. Hussain, and J. Mustafina, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in *Proc. 2018 IEEE Congress on Evolutionary Computation (CEC)*, 2018, pp. 1–9.

[11] S. A. Shinde and P. R. Rajeswari, "Intelligent health risk prediction systems using machine learning: A review," *IJET*, vol. 7, no. 3, pp. 1019–1023, 2018.

[12] D. Dua and C. Graff, *UCI Machine Learning Repository*, 2017.

[13] W. Gunarathne, K. Perera, and K. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)," in *Proc. 2017 IEEE*

17th International Conference on Bioinformatics and Bioengineering (BIBE), 2017, pp. 291–296.
[14] P. Yildirim, "Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction," in *Proc. 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, 2017, vol. 2, pp. 193–198.
[15] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, "When and how should multiple imputation be used for handling missing data in randomised clinical trials—A practical guide with flowcharts," *BMC Medical Research Methodology*, vol. 17, no. 1, p. 162, 2017.
[16] S. D. Arasu and R. Thirumalaiselvi, "Review of chronic kidney disease based on data mining techniques," *International Journal of Applied Engineering Research*, vol. 12, no. 23, pp. 13498–13505, 2017.
[17] S. Sharma, V. Sharma, and A. Sharma, *Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis*, July 18, 2016.
[18] A. Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," in *Proc. IEEE International Conference on Healthcare Informatics (ICHI)*, Oct. 2016.
[19] S. Ramya and N. Radha, "Diagnosis of chronic kidney disease using machine learning algorithms," in *Proc. International Journal of Innovative Research in Computer and Communication Engineering*, 2016.
[20] S. Vijayarani, S. Dhayanand, *et al.*, "Data mining classification algorithms for kidney disease prediction," *International Journal on Cybernetics & Informatics (IJCI)*, vol. 4, no. 4, pp. 13–25, 2015.
[21] L. Rubini. (2015). Early stage of chronic kidney disease UCI machine learning repository. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease
[22] Estimated Glomerular Filtration Rate (eGFR). (Dec. 2015). [Online]. Available: https://www.kidney.org/atoz/content/gfr
[23] S. Nair, S. V. O'Brien, K. Hayden, B. Pandya, P. J. G. Lisboa, K. J. Hardy, and J. P. H. Wilding, "Effect of a cooked meat meal on serum creatinine and estimated glomerular filtration rate in diabetes-related kidney disease," *Diabetes Care*, vol. 37, no. 2, pp. 483–487, Feb. 2014.
[24] C. Li, "Little's test of missing completely at random," *The Stata Journal*, vol. 13, no. 4, pp. 795–809, 2013.
[25] F. E. Murtagh, J. Addington-Hall, P. Edmonds, P. Donohoe, I. Carey, K. Jenkins, and I. J. Higginson, "Symptoms in the month before death for stage 5 chronic kidney disease patients managed without dialysis," *Journal of Pain and Symptom Management*, vol. 40, no. 3, pp. 342–352, 2010.
[26] D. C. Yadav and S. Pal, "Prediction of thyroid disease using decision tree ensemble method," *Human-Intell. Syst. Integr.*, vol. 2, no. 1, pp. 89–95, 2020.
[27] G. M. Ifraz, M. H. Rashid, T. Tazin, S. Bourouis, and M. M. Khan, "Comparative analysis for prediction of kidney disease using intelligent machine learning methods," *Comput. Math. Methods Med.*, 2021.
[28] C. Bemando, E. Miranda, and M. Aryuni, "Machine-learning-based prediction models of coronary heart disease using naïve bayes and random forest algorithms," in *Proc. 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*, 2021, pp. 232–237.
[29] R. P. R. Kumar and S. Polepaka, "Performance comparison of random forest classifer and convolution neural network in predicting heart diseases," in *Proc. the Third International Conference on Computational Intelligence and Informatics*, Singapore, 2020.
[30] R. S. Walse, G. D. Kurundkar, S. D. Khamitkar, A. A. Muley, P. U. Bhalchandra, and S. N. Lokhande, "Effective use of naïve Bayes, decision tree, and random forest techniques for analysis of chronic kidney disease," in *Proc. International Conference on Information and Communication Technology for Intelligent Systems*, Singpore, 2020.

Chamandeep Kaur has been working as a lecturer in the Department of Computer Science and Information Technology at Jazan University, Saudi Arabia. She received her Ph.D. in computer science and engineering from JJT University, Rajasthan, India in 2022, and her MCA from Punjab Technical University, Jalandhar, India in 2006. She has over 15 years of experience in teaching, research, education, industry, and consulting.
Her research areas include IoT, cloud computing, artificial intelligence, computer networks, data security, big data, and machine learning. She has made several innovative and outstanding contributions to academic research. She has contributed over 25 publications in Scopus, SCIE, Elsevier, WoS, and International peer-reviewed impact journals, several patents, and books. She is a member of the Institute of Electrical and Electronics Engineers (IEEE) and the International Association of Engineers. Her most recent publication is "Recognition of Copy Move Forgeries in Digital Images using Hybrid Optimization and Convolutional Neural Network Algorithm" and "Artificial Intelligence – A Modern Approach (Book)".

M. Sunil Kumar has completed a Ph.D. in computer science and engineering at S.V. University, Tirupati; MTech in computer science from JNT University; B.Tech in computer science & information technology from JNT University. He is currently working as a professor and head of the Department of CSE, Mohan Babu University, (erstwhile Sree Vidyanikethan Engineering College), A. Rangampet, Tirupati, A.P.
His main research interest includes software engineering, software architecture, information retrieval, machine learning, deep learning, and database management systems.

Afsana Anjum has been working as a lecturer in the Department of Computer Science and Information Technology at Jazan University, Saudi Arabia. She received her M.Tech. in Maharashi Dayanand University, Rohtak, Haryana, India in 2011. She has over 12 years of experience in teaching, research, education, and consulting.
Her research areas include IoT, cloud computing, computer networks, data security, machine learning, blockchain, and big data. She has made several innovative and outstanding contributions to academic research. She has contributed over 10 publications in Scopus, and International peer-reviewed impact journals and patents. Her most recent publication is "Role of AI in the curative field: A review".

M. B. Binda working as a senior engineer in the Department of Traffic Signal Division at Keltron Communication Complex, Monvila, Kulathoor, Thiruvananthapuram, Kerala. She graduated in electrical & electronics engineering at NSS College of Engineering, Palakkad, Kerala, India. She secured a master of technology in electrical & electronics engineering at the College of Engineering, Thiruvananthapuram, Kerala, India. She did Ph.D. in the field of control systems at Sathyabama Institute of Science and Technology, Chennai, India. She was in the teaching profession for more than 14 years. She is in the industrial profession for more than 4.5 years. She has 9 Indian Patent Publications and 1 German Patent Publication in the research field of expertise. She has presented several papers in National and International Journals, conferences, and Symposiums. Her main area of interest includes control systems, artificial intelligence, electrical machines, digital signal processing, power systems, analog & digital communications, computer architecture, power electronics, electrical drives & control, optical fibre communication, digital electronics, digital signal processing, high voltage engineering, robotics, soft computing techniques, genetic algorithms, machine learning and internet of things.

Maheswara Reddy Mallu completed his Ph.D. in biotechnology. He is currently working as an assistant professor in biotechnology, Koneru Lakshmaiah Education Foundation (Deemed to be University).

His main research interest includes bioprocess engineering, formulation of therapeutic drugs, and recombinant products. He has given an outstanding contribution to academic research.

Mohammed Saleh Al Ansari has been working as an associate professor in the College of Engineering, Department of Chemical Engineering at the University of Bahrain. He earned his bachelor's degree in chemical engineering from King Saud University in Riyadh in 1983. In 1987, he received his MSc degree in desalination engineering, Desalination Technology, Glasgow. In 1998, he received his Ph.D. in corrosion and protection science and engineering. He has over 35 years of experience in teaching, research, and industry.

His research areas include chemical engineering, machine learning, complex system, data sciences, sustainability, and energy. He has made several innovative and outstanding contributions to academic research. He has contributed over 40 publications in Scopus, and International peer-reviewed impact journals and patents. Her most recent publication is "Recognition of Copy Move Forgeries in Digital Images using Hybrid Optimization and Convolutional Neural Network Algorithm".