Biomedical Materials & Devices
https://doi.org/10.1007/s44174-022-00027-y

ORIGINAL PAPER

Chronic Kidney Disease Prediction Using Machine Learning Techniques

Saurabh Pal[1]

Received: 20 April 2022 / Accepted: 16 August 2022

Abstract
Chronic kidney disease (CKD) is a life-threatening condition that can be difficult to diagnose early because there are no symptoms. The purpose of the proposed study is to develop and validate a predictive model for the prediction of chronic kidney disease. Machine learning algorithms are often used in medicine to predict and classify diseases. Medical records are often skewed. We have used chronic kidney disease dataset from UCI Machine learning repository with 25 features and applied three machine learning classifiers Logistic Regression (LR), Decision Tree (DT), and Support Vector Machine (SVM) for analysis and then used bagging ensemble method to improve the results of the developed model. The clusters of the chronic kidney disease dataset were used to train the machine learning classifiers. Finally, the Kidney Disease Collection is summarized by category and non-linear features. We get the best result in the case of decision tree with accuracy of 95.92%. Finally, after applying the bagging ensemble method we get the highest accuracy of 97.23%.

Keywords Chronic kidney disease · Decision Tree · Support Vector Machine · Logistic Regression and Bagging Ensemble Method

Introduction

Engineers and medical researchers are trying to develop machine learning algorithms and models that can identify chronic kidney disease at an early stage. The problem is that the data generated in the health industry is large and complex, making data analysis difficult. However, we can process this data into a data format using data mining technology, and then this data can be translated into machine learning algorithms.

A combination of estimated glomerular filtration rate (GFR), age, diet, existing medical conditions, and albuminuria can be used to assess the severity of kidney disease, but requires more accurate information about the risk to the kidney is required to make clinical decisions about diagnosis, treatment, and referral [1].

The purpose of this model is to develop and validate predictive models for chronic kidney disease. The main goal will be to evaluate kidney failure, which means the need for kidney dialysis or kidney transplant first [2].

These models also teach the patient how to live a healthy life and help the doctor see the risk and severity of the disease, as well as how to proceed with the treatment in the future. It may be possible to identify patterns of data collection using ANN, mining methods, and the future occurrence of certain diseases that may cause harm can be predicted in advance [3].

The purpose of the proposed model is to predict whether the patient will suffer or develop chronic kidney disease in the future if he continues their lifestyle. This information can be used to determine whether the kidney disease is using eGFR (glomerular filtration rate), which helps the doctor plan the appropriate treatment. Estimated glomerular filtration rate (eGFR) defines the degree of kidney disease and measures kidney function [4].

The main function of the kidney is to filter the blood in the body. Kidney disease is a silent killer because it can cause kidney failure without causing any symptoms or concern. Chronic kidney disease is defined as a decline in kidney function over a period of months or years. Kidney disease is often caused by diabetes and high blood pressure. Chronic kidney disease is a major health problem that affects people worldwide. Not getting the right treatment for chronic kidney disease can have serious consequences, affecting people who can't afford it. Glomerular filtration rate (GFR) is

✉ Saurabh Pal
drsaurabhpal@yahoo.co.in

[1] Department of Computer Applications, VBS Purvanchal University, Jaunpur, India

Published online: 31 August 2022

the most accurate test to determine your kidney function and the degree of chronic kidney disease. Blood creatinine level, age, gender, and other characteristics can be used to calculate it. In most cases it is better to get sick early. Therefore, it is possible to prevent serious illness [5].

Data mining methods are specifically used in models proposed to predict kidney disease. The database can be extended with more information than the existing chronic kidney disease model. That is, more information obtained from patients with chronic kidney disease can be added (although the information must be reliable), which will increase the accuracy of the prediction. In addition, with the help of experts, research can be done to identify new features that cause chronic kidney disease and then add these features as features to the fabric paper, which will increase the accuracy of the prediction. The design process is as follows: First, the best algorithm to classify the data as CKD or NOT_CKD is selected, and then the data is classified as CKD or NOT_CKD. The CKD_EPI equation will be used to determine the eGFR value if the classification is chronic kidney disease. We will be able to calculate the patient's current status using this eGFR measurement.

Related Work

Nephropathy, or kidney damage, is called kidney disease. People with kidney disease have kidney failure, which can lead to kidney failure if not treated quickly. According to the National Kidney Foundation, chronic kidney disease affects 10% of the world's population, and millions of people die each year due to inadequate treatment. Recent advances in ML and DL-based kidney disease testing may bring hope to countries that cannot manage kidney disease testing.

Bemando et al. investigated the relationship between blood-related diseases and their features utilising classifier methods such as Gaussian NB, Bernoulli NB, and Random Forest. These three algorithms anticipate and offer statistical findings in a variety of ways. In this experiment, we discovered that Nave Bayes estimated accuracy was higher than that of other algorithms [6].

Kumar and Polepaka devised a technique for illness prediction in the medical field. They employed Random Forest and CNN as well as other machine learning methods. For illness dataset classification, precision, recall, and F1-score, these algorithms deliver better results. In this experiment, Random Forest outperformed other algorithms in terms of accuracy and statistical performance [7].

Sing et al. developed a technique for predicting medical-related illness datasets. For improved prediction, they utilised a support vector machine classifier. The accuracy ranged from 73 to 91 percent, and the author eventually improved accuracy to 91 percent [8].

Desai et al. devised a technique for illness prediction in the medical field. The author employed back-propagation NN and LR classification algorithms in this study. These two strategies provide distinct outcomes, with statistical analysis and logistic regression yielding a more accurate model than other algorithms [9].

Patil et al. created a database for ECG arrhythmia-related medical conditions. On a disease dataset, the authors employed machine learning approaches such as Support Vector Machine and Cuckoo search optimised Neural Network, and support vector machine estimated 94.44 percent improved accuracy [10].

Observed illness dataset for statistical analysis by Liu et al. They estimated superior findings for specificity, sensitivity, positive predictive value, and negative predictive value using machine learning approaches such as support vector machines [11].

For better statistical analysis outcomes, Acharya et al. reviewed medical linked illness dataset. They employed several machine learning techniques, such as CNN, and applied machine learning algorithms to the ECG dataset, achieving a classification accuracy of 94 percent [12].

Wasle et al. devised a statistical analysis technique. For the examination of the Chronic Kidney Disease dataset, the authors employed a variety of machine learning approaches. They used Nave Bayes, Decision Trees, and Random Forest to improve prediction, and they discovered that Random Forest computed greater classification accuracy than the other algorithms [13].

On the Kidney disease dataset, Nithya et al. developed a method for categorization and cluster-based analysis. On diverse sets of photos, the authors utilized the K-Means clustering technique to collect the closest familiar images. They calculated 99.61 percent classification accuracy using Artificial Neural Networks for Kidney Disease Image Prediction [14].

Al Imran et al. examined the use of machine learning techniques to analyze datasets for chronic renal disease. For statistical analysis such as F1-score, Precision, Recall, and AUC, the authors employed Logistic Regression and Feed forward Neural Network and generated better results than previous algorithms [15].

Navaneeth and Suchetha devised a method for predicting chronic renal disease using a dataset. They employed machine learning methods such as CNN and SVM. Authors estimated greater accuracy, sensitivity, and specificity findings after the prediction [16].

Brunetti et al., used a system or method for chronic kidney disease dataset. Authors used CNN machine learning technique and calculated 95% classification accuracy for disease dataset [17].

Methodology

In this research paper, we have applied three machine learning classifiers logistic regression, decision tree and support vector machine on chronic kidney diseases dataset collected from UCI machine learning repository. The developed model was evaluated by training and testing using 80% as training dataset and 20% as the testing dataset. tenfold cross validation method is applied for training the classifiers. After evaluating the three machine learning classifier, we applied bagging ensemble method to improve the performance of the developed model and then final results have been evaluated. The proposed methodology was described in Fig. 1.

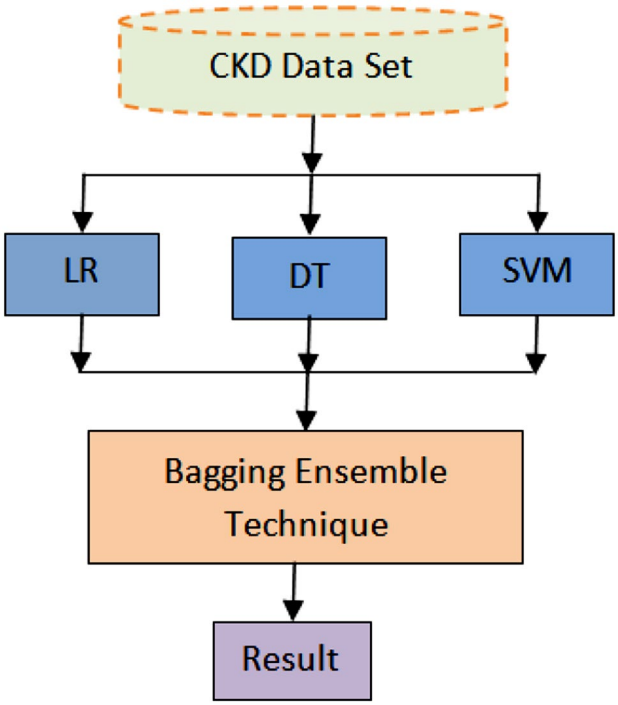

Fig. 1 Proposed Chronic Kidney Disease prediction model

Data Description

The database used in this study was collected from the UCI machine learning repository. There are 400 cases in the dataset (250 CKD, 150 NOT_CKD). Table 1 shows 11 non_categorical attributes and 14 categorical attributes [18]. The categorical and non_categorical chronic kidney disease characteristics are shown in Table 1 with null counting values and data types as attribute values. Chronic kidney disease has a total of 25 attributes (categorical and non_categorical) and 400 instances. Data on chronic kidney disease are gathered as an electronic medical record from the UCI machine learning repository. “1” and “0” are the two values of the target variable. The “1” denotes normal instances, whereas the "0" denotes sickness [19].

Machine Learning Classifiers

Logistic Regression Classifier

The logistic function was earlier named "Logistic," and Pierre François Verhulst developed it as a model of population growth in the years between the 1830s and 1840s. Later

Table 1 Categorical and non_categorical attributes of chronic kidney disease dataset

Non_categorical attributes				Categorical attributes			
#	Column	Non-null count	Dtype	#	Column	Non-null count	Dtype
0	Id	400 non-null	int64	0	bp	388 non-null	float64
1	Age	391 non-null	float64	1	Sg	353 non-null	float64
2	bgr	356 non-null	float64	2	Al	354 non-null	float64
3	Bu	381 non-null	float64	3	su	351 non-null	float64
4	Sc	383 non-null	float64	4	Rbc	248 non-null	object
5	Sod	313 non-null	float64	5	Pc	335 non-null	object
6	Pot	312 non-null	Float64	6	Pcc	396 non-null	object
7	Hemo	348 non-null	float64	7	Ba	396 non-null	object
8	Pcv	330 non-null	object	8	Htn	398 non-null	object
9	Wc	295 non-null	object	9	Dm	398 non-null	object
10	rc	270 non-null	object	10	Cad	398 non-null	object
11	classification	400 non-null	object	11	Appet	399 non-null	object
dtype: float64(7), int64(1), object(4)				12	pe	399 non-null	object
				13	Ane	399 non-null	object
				14	classification	400 non-null	object
				dtype: float64(4), object(11)			

many researchers worked on this function development. The multinomial model called "logit" was introduced by Cox and Theil in 1966 and 1969. Daniel McFadden had linked the multinomial function logit to the theory of discrete choice, showing that it came out from the assumption of independence of irrelevant alternatives as relative preferences. It laid a theoretical foundation for the new logistic regression concept. Logistic regression is useful for many applications such as Machine learning, medical, Social sciences. This technique is also helpful for engineering, especially for predicting a system or model's success or failure [20]. A brief explanation of logistic regression is as below:

Logistic regression predicts the probability in two values only, while a linear regression predicts the values outside the range of (0 to 1).

The logistic regression can be written mathematically as shown below:

$$p = \frac{1}{1 + e^{-(+b_1x_1+b_2x_2+\ldots+b_px_p)}} \tag{1}$$

Decision Tree Classifier

A decision tree is a supervised learning-based predictive modeling tool created by J. R. Quinlan at the University of Sydney and published in his book Machine Learning. This tool works on the principle of multivariate analysis, which can help predict, explain, describe, and classify the outcome. It splits the dataset based on multiple conditions, describing beyond one cause cases and describing the condition based on numerous influences. Quinlan created the Iterative Dichotomiser version 3 (ID3) algorithms, which is used to generate decision trees. He then expanded his research based on ID3 and created C4.5, an improved version of ID3. A feature-rich and enhanced version over C4.5, which is being sold by Quinlan, is C5.0 under GPL [21]. A decision tree is generated from the root following a top-down approach that involves the partitioning of data. Entropy and Gini index are calculated using the formula given below.

$$Gini = 1 - \sum_{i=1}^{C}(p_i)^2 \tag{2}$$

$$Entropy = \sum_{i=1}^{C} -p_i log_2(p_i) \tag{3}$$

There are many algorithms used for the generation of decision trees.

- Classification And Regression Tree (CART)
- ID3
- CHAID
- ID4.5

Support Vector Machine Classifier

Support Vector Machine Classifier is a supervised learning tool that is useful for regression and classification. The central theme of working of SVM is that it's a binary classification algorithm that separates the data points to find a hyperplane in case of many possible inputs. It handles the outliers efficiently and, in the case of high dimensional spaces, works well. It uses decision functions, also known as support vectors, to perform classification. At least four kernel types are used for classification SVC with linear kernel, Linear SVC, SVC with RBF kernel, and SVC with the polynomial kernel [22]. The mathematical representation of the classifier algorithm is as under.

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i\left(w \cdot x_i - b\right)\right)\right] + \lambda||w||^2 \tag{4}$$

Bagged Decision Trees (Bagging) Classifier

Bagging classifier is named because it creates an ensemble of decision trees, and it is used for classification or regression. It is also known as bootstrap aggregation. Input data replica is drawn independently in every tree to grow. To handle regression problems, it supports mean and quantile regression. Bagging ensemble technique is used when the goal is to reduce the variance of another base learner. The goal is to create some subset of the data from a randomly selected and replaced training set. Each subset of the data set is used to train its base learner. As a result, we have a collection of different models. Using the average of all predictions from different base learners, it is more reliable than one base learner [23, 24].

Results and Discussion

In the present paper, the machine learning classifiers used in two stages, at first the logistic regression classifier is used, second the Decision Tree Classifier is used, then the Support Vector Machine classifier. All these classifiers are described in the methodology section above.

A confusion matrix is a table-like structure in which the performance of a classifier is described or evaluated. This description of the performance is performed on a dataset for which the true values are already known. Although the terms of the matrix may look confusing, the confusion matrix is usually found to be simple and easy to understand [25]. The terms of the confusion matrix are described below.

Table 2 Confusion Matrix metric formula table

Type of Metric	Formula	
Accuracy	$ACC = \frac{tp+tn}{tp+fp+tn+fn}$	(5)
Recall	$Recall = \frac{tp}{tp+fn}$	(6)
Precision	$Precision = \frac{tp}{tp+fp}$	(7)
F1-score	$F = 2.\frac{precision.recall}{precision+recall}$	(8)
Specificity	$Specificity = \frac{tn}{tn+fp}$	(9)

Table 3 Classification report with accuracy from experiment 1: base classifiers

Classifier used	Precision	Recall	F1-score	Accuracy achieved (%)
Logistic regression	0.96	0.96	0.96	93.28
Decision tree	0.99	0.98	0.98	95.92
Support vector machines	0.96	0.96	0.96	94.80

True Positives (TP): In this case, the prediction is yes (they are Phishing URLs).

True Negatives (TN): In this case, the prediction is no (they are not Phishing URLs).

False Positives (FP): In this case, the prediction is yes, but it is a false prediction i.e., they are not the Phishing URLs (It is called a Type I error).

False Negatives (FN): In this case, the prediction is no, but they are actuallythe phishing URLs (It is called a Type II error).

A classification report is constructed using many constituent variable parameters; these parameters are used to show the values of the parameters used for the calculations of an accuracy score. Recall value that is also known as hit rate, True Positive Rate (TPR), or sensitivity, is obtained by using the formula as given below in the table. True Negative Rate (TNR), selectivity also called specificity, is obtained by using the formula as given below in the table. Positive Predictive Value (PPV), also known as precision, is obtained by using the formula given below in the table f1-score or balance f-score, also called traditional f-measure, is actually the harmonic mean of sensitivity & precision [32–34]. Precision is calculated by using the formula as given below in Table 2. The detailed description of the results obtained from base classifiers are described in Table 3.

The results are plotted using a boxplot using python code that clearly shows the accuracy scores and the outliers. The boxplot showing the comparison between all the base classifiers used for prediction is shown in the boxplot below in Fig. 2.

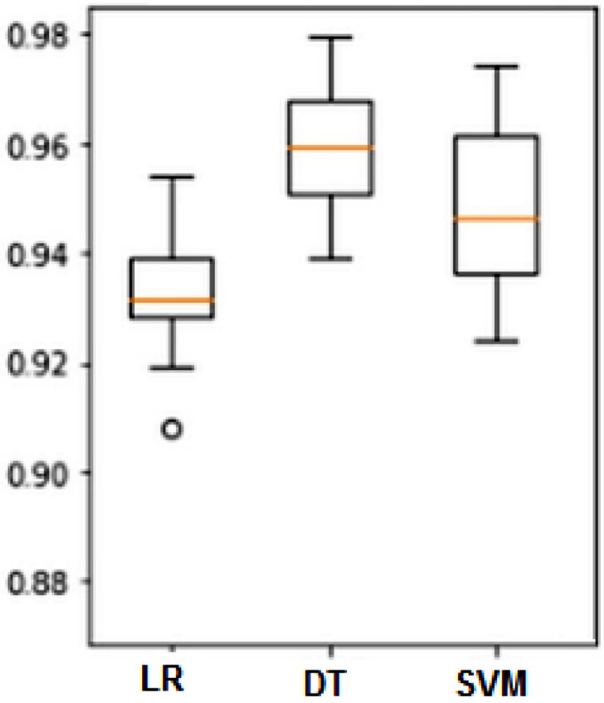

Fig. 2 Boxplot representation showing Performance of base classifiers

Table 4 Accuracy achieved by different Ensemble techniques

Ensemble Methods	Accuracy (%)		
	LR	DT	SVM
Bagging	94.53	97.23	95.70

We have used bagging ensemble techniques to enhance the results obtained by base learners and different predication accuracy is found which are shown in Table 4.

From the Table 4, it is clear that the accuracy of the base machine learning classifiers has increased. In case of logistic regression bagging ensemble method has increase the accuracy from 93.28% to 94.53%, respectively. The best chronic kidney disease prediction model is decision tree which gives the 97.23% of accuracy.

From Tables 5, it can be seen that the proposed methodology improves the performance of the otherwise independent models and achieves comparable or better performance compared to the models proposed in previous studies.

Conclusion

In this research, we have used chronic kidney disease dataset collected from UCI machine leaning repository. We have developed a chronic kidney disease prediction model using

Table 5 Comparison with other models

Model	Accuracy (%)	Precision	Recall	F-score
Krishnamurthy [27]	78.2	.46	.48	.47
Rehman [28]	88.3	.66	.86	.74
Gazi [26]	71	.72	.72	.71
Rady [30]	77.29	.44	.13	.21
Han [29]	87.79	.10	.87	–
Dong [31]	76.8	.86	.68	.74
HU [20]	57.9	.59	.60	.58
Proposed Model				
LR	93.28	.96	.96	.96
DT	95.92	.99	.98	.98
SVM	94.80	.96	.96	.96

three machine learning classifiers Logistic Regression, Decision Tree and Support Vector Machine to measure the performance of the prediction model. The performance of the model depends upon various performance matrices like sensitivity, precision, recall, f1-score, support, confusion matrix etc. The developed chronic kidney disease prediction model has been trained by categorical and non_categorical chronic kidney disease dataset attributes. After applying the base classifiers we find decision tree classifier obtained better results in terms of Accuracy, Precision, Recall, F-score as 95.92%, 0.99, 0.98, and 0.98, respectively. The decision tree classifier, perform better compare to logistic regression and support vector machine. In second step, we have applied bagging ensemble method to improve the performance of base classifiers and find the highest accuracy of 97.23% in case of decision tree. This can help the medical practitioners and patients for the early prediction of chronic kidney disease to save a life. In the future, the model can be further tuned by applying feature selection methods to increase the performance of the prediction.

Acknowledgements Author thanks to Veer Bahadur Singh Purvanchal University, Jaunpur for providing the support for conducting this research work as a part of minor project "Analysis of Hidden Pattern and Discover Real Fact of Medical Diseases using Integrated Machine Learning Techniques.

Declarations

Conflict of interest Author declares no conflict of interest.

References

1. Aljaaf, A.J. 2018 Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics. In Proceedings of the IEEE Congress on Evolutionary Computation (CEC). Wellington. New Zealand
2. A. Nishanth, T. Thiruvaran, Identifying important attributes for early detection of chronic kidney disease. IEEE Rev. Biomed. Eng. **11**, 208–216 (2018)
3. A. Ogunleye, Q.-G. Wang, XGBoost model for chronic kidney disease diagnosis. IEEE/ACM Trans. Comput. Biol. Bioinform. **17**, 2131–2140 (2020)
4. F. Aqlan, R. Markle, 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
5. N. Borisagar, D. Barad, P. Raval, Chronic kidney disease prediction using back propagation neural network algorithm. Proce. Int. Confe. Commun. Netw. **19–20**, 295–303 (2017)
6. C. Bemando, E. Miranda, M. Aryuni, "Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms," in *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*, (IEEE, 2021), pp. 232–237
7. R.P. Ram Kumar, SanjeevaPolepaka, Performance comparison of random forest classifier and convolution neural network in predicting heart diseases, in *Proceedings of the Third International Conference on Computational Intelligence and Informatics*. ed. by K. SrujanRaju, A. Govardhan, B. PadmajaRani, R. Sridevi, M. Ramakrishna Murty (Springer, Singapore, 2020)
8. H. Singh, N. V. Navaneeth, G. N. Pillai, "Multisurface proximal SVM based decision trees for heart disease classification," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, (IEEE 2019), pp. 13–18
9. S.D. Desai, S. Giraddi, P. Narayankar, N.R. Pudakalakatti, S. Sulegaon, *Backpropagation neural network versus logistic regression in heart disease classification in advanced computing and communication technologies* (Springer, Singapore, 2019)
10. D.D. Patil, R.P. Singh, V.M. Thakare, A.K. Gulve, Analysis of ecg arrhythmia for heart disease detection using svm and cuckoo search optimized neural network. Int. J. Eng. Technol. **7**(217), 27–33 (2018)
11. N. Liu, Z. Lin, J. Cao, Z. Koh, T. Zhang, G.-B. Huang, W. Ser, M.E.H. Ong, An intelligent scoring system and its application to cardiac arrest prediction. IEEE Trans. Inf Technol. Biomed. **16**(6), 1324–1331 (2012)
12. U. Rajendra Acharya, Oh. Shu Lih, Y. Hagiwara, J.H. Tan, M. Adam, A. Gertych, R.S. Tan, A deep convolutional neural network model to classify heartbeats. Comput. Biol. Med. **89**, 389–396 (2017)
13. R.S. Walse, G.D. Kurundkar, S.D. Khamitkar, A.A. Muley, P.U. Bhalchandra, S.N. Lokhande, Effective use of naïve bayes, decision tree, and random forest techniques for analysis of chronic kidney disease, in *International Conference on Information and Communication Technology for Intelligent Systems*. ed. by T. Senjyu, P.N. Mahalle, T. Perumal, A. Joshi (Springer, Singpore, 2020)
14. A. Nithya, A. Appathurai, N. Venkatadri, D.R. Ramji, C.A. Palagan, Kidney disease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound images. Measurement (2020). https://doi.org/10.1016/j.measurement.2019.106952
15. Abdullah Al Imran, Md Nur Amin, and Fatema Tuj Johora. Classification of chronic kidney disease using logistic regression, feedforward neural network and wide & deep learning. In 2018 International Conference on Innovation in Engineering and Technology (ICIET), pages 1–6. IEEE, 2018.
16. B. Navaneeth, M. Suchetha, A dynamic pooling based convolutional neural network approach to detect chronic kidney disease. Biomed. Signal Proce. Control **62**, 102068 (2020)
17. A. Brunetti, G.D. Cascarano, I. De Feudis, M. Moschetta, L. Gesualdo, V. Bevilacqua, Detection and segmentation of kidneys

from magnetic resonance images in patients with autosomal dominant polycystic kidney disease, in *International Conference on Intelligent Computing*. ed. by D.-S. Huang, K.-H. Jo, Z.-K. Huang (Springer International Publishing, Cham, 2019)
18. D. Ramos et al., Using decision tree to select forecasting algorithms in distinct electricity consumption context of an office building. Energy Rep. **8**, 417–422 (2022)
19. H.E. Song et al., Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. Chemosphere **290**, 133388 (2022)
20. H.U. Rongyao et al., Multi-task multi-modality SVM for early COVID-19 diagnosis using chest CT data. Inf. Proc. Manag. **59**(1), 102782 (2022)
21. X.U. Ankun et al., Artificial neural network (ANN) modeling for the prediction of odor emission rates from landfill working surface. Waste Manag. **138**, 158–171 (2022)
22. D.C. Yadav, S. Pal, *An Ensemble Approach on the behalf of Classification and Prediction of Diabetes Mellitus Disease Emerging Trends in Data Driven Computing and Communications* (Springer, Singapore, 2021)
23. D.C. Yadav, S. Pal, Performance based evaluation of algorithms on chronic kidney disease using hybrid ensemble model in machine learning. Biomed. Pharmacol. J. **14**(3), 1633–1646 (2021)
24. D.C. Yadav, S. Pal, Discovery of Thyroid Disease Using Different Ensemble Methods with Reduced Error Pruning Technique, in *Computer-aided Design and Diagnosis Methods on the behalf of Biomedical Applications*. ed. by G.R. Varun Bajaj, V.B. Sinha, G.R. Sinha (CRC Press, Boca Raton, 2021)
25. A. Zoda et al., Inferring genetic characteristics of Japanese Black cattle populations using genome-wide single nucleotide polymorphism markers. J. Animal Genet. **50**(1), 3–9 (2022)
26. G.M. Ifraz, M.H. Rashid, T. Tazin, S. Bourouis, M.M. Khan, Comparative analysis for prediction of kidney disease using intelligent machine learning methods. Comput. Math. Methods Med. (2021). https://doi.org/10.1155/2021/6141470
27. S. Krishnamurthy, K.S. Kapeleshh, E. Dovgan, M. Luštrek, B.G. Piletič, K. Srinivasan, Y.C. Li, A. Gradišek, S. Syed-Abdul, "Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan." medRxiv. (2020).https://doi.org/10.1101/2020.06.25.20139147
28. Z.U. Rehman, M.S. Zia, G.R. Bojja, M. Yaqub, F. Jinchao, K. Arshid, Texture based localization of a brain tumor from MR-images by using a machine learning approach. Med. Hypotheses **141**, 109705 (2020)
29. X. Han, X. Zheng, Y. Wang, X. Sun, Y. Xiao, Y. Tang, W. Qin, Random forest can accurately predict the development of end-stage renal disease in immunoglobulin a nephropathy patient. Annals Transl. Med. (2019). https://doi.org/10.21037/atm.2018.12.11
30. E.H.A. Rady, A.S. Anwar, Prediction of kidney disease stages using data mining algorithms. Inform. Med. Unlocked (2019). https://doi.org/10.1016/j.imu.2019.100178
31. Z. Dong, Q. Wang, Y. Ke, W. Zhang, Q. Hong, C. Liu, X. Chen, Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical records. J. Transl. Med. **20**(1), 1–10 (2022)
32. D.C. Yadav, S. Pal, Prediction of thyroid disease using decision tree ensemble method. Human-Intell. Syst. Integr. **2**(1), 89–95 (2020)
33. V. Chaurasia, S. Pal, Applications of machine learning techniques to predict diagnostic breast cancer. SN Compu. Sci. **1**(5), 1–11 (2020)
34. Chaurasia, V., & Pal, S. (2014). Performance analysis of data mining algorithms for diagnosis and prediction of heart and breast cancer disease. Review of research. 3(8).