Chronic Kidney Disease Detection Using Machine Learning: From Analysis to Framework Development

Bobbinpreet Kaur[1]*, Bhawna Goyal[2], Ayush Dogra[3], Sonam Ramshankar[4], Devendra Singh[5] and Ahmed Alkhayyat[6]

[1]ECE, Chandigarh University, Mohali, India.
[2]Faculty of Engineering, Sohar University, Sohar, Oman.
[3]Chitkara Institute of Engineering and Technology, Chitkara University, Punjab, India.
[4]IES Institute of Pharmacy, IES University, Bhopal, Madhya Pradesh India 462044.
[5]Uttaranchal Institute of Technology, Uttaranchal University, Dehradun-248007, India.
[6]College of Technical Engineering, The Islamic University, Najaf, Iraq.
*Corresponding Author E-mail: ayush123456789@gmail.com

Considering the aspects of sustainable development goals, Good health and well-being ensure the development of a nation. Chronic kidney disease (CKD) is a progressive and irreversible condition characterized by the gradual loss of kidney function over time. One of the major diseases, CKD affecting 10-15% population globally needs to be detected at early stages to reduce morbidities and mortalities. Majorly the risk factors include Diabetes, Hypertension, Age, Hereditary, and Ethnicity which need to be screened on regular intervals to ensure the timely detection of the disease. The primary hurdle for detection is asymptomatic behavior during the early stages. Machine learning (ML) based models are majorly governing various sectors and applications. The models have capabilities to serve as assistance to the medical practitioners for effective CKD detection at early stages. This paper demonstrates the development of a framework for early detection considering various parameters.

Keywords: CKD; Ensemble learning; Good Health; Improving Mortality; Machine Learning; Medical assistance Well-being.

In recent years, the number of patients suffering from CKD is elevating. CKD is emerging as one of the major cause for causalities worldwide. As per the recent reports, 324 million people are suffering from CKD across the globe[1]. The normal functioning of the kidney ensures the balanced amount of various salts, ions and other electrolytes in the human body. Kidneys filtrates the blood passing through human body and removes the toxins from the blood stream. One of the major disease effecting the normal operation of the kidney is CKD. The epidemiology of this infectious disease Fluctuates across sections and inhabitants, with inclining prevalence in the past few years. The major risk causing factors for CKD are diabetes, hypertension, obesity, smoking, and a family history of kidney disease[2]. As per statistics available through various reports published by World Health Organization (WHO) 10% of the world's population is estimated affected from CKD, with some proportion requiring renal replacement cure, such as dialysis or kidney transplantation, continuing to rise. Moreover, because of CKD there is an increased risk of cardiovascular disease, mortality, and reduced quality of life[3].

Clinically, the major hurdle in detection of CKD is its asymptomatic nature in the early stages and majorly the symptoms show in the later stages after substantial damage has already occurred. Preliminary diagnosis for evaluating kidney health can be done through laboratory tests, including serum creatinine, estimated glomerular filtration rate (eGFR), and urine analysis[4]. Majorly these tests are conducted by assessing the kidney functioning over a window of few months which measures the trend in reduction of kidney functioning.

Multiple risk factors contribute to the development and progression of CKD. Diabetes mellitus and hypertension are the leading causes of CKD, accounting for a substantial proportion of cases. Other risk factors include obesity, smoking, older age, family history of kidney disease, and certain genetic disorders. The complex interplay between these factors and the underlying pathophysiological mechanisms of CKD presents a multifaceted challenge in understanding, diagnosing, and managing the disease.

The clinical diagnosis and management of CKD is primarily dependent upon the glomerulus, a filter present inside the kidney. Majorly the diagnosis is governed by the fact that the symptoms related to non-functional glomerulus become pre-dominant after acute loss have been done[5]. Estimated glomerular filtration rate (eGFR) is one of the major metric which is measured to define the functionality and progression of CKD. As shown in figure 1 with progression in years of incidence of CKD the eGFR value dips with sharp edge on delay in diagnosis. Therefore it is mandatory to develop methodologies and support system to detect the disease at early stages[6, 7].

This paper aims to provide a comprehensive overview and development of a framework for CKD detection through various machine learning methods, exploring its epidemiology[8]. By enhancing our understanding of CKD, its risk factors, and its impact on individuals and society, we can develop targeted interventions and policies that alleviate the burden of this silent epidemic. Additionally, ongoing research efforts and advancements in the machine learning based diagnosis field offer hope for improved supportive diagnostics, novel therapies, and preventive measures that will reshape the future of CKD management[9].

Related Work

The enormous number of applications of Machine learning algorithms nowadays made it possible to act as a decision and diagnosis support system for health care sector as well[10]. A big set of diseases can be predicted with the help of effectively trained model thereby making it possible to save the precious lives of the patients[11]. One of the underlying advantages rises from the fact that these models (if effectively trained) can detect the disease at early stage of incidence where even the clinical pathology tests fails to detect the variation in the normal parameters. While extensively reviewing the literature, a number of researchers have been found working towards the
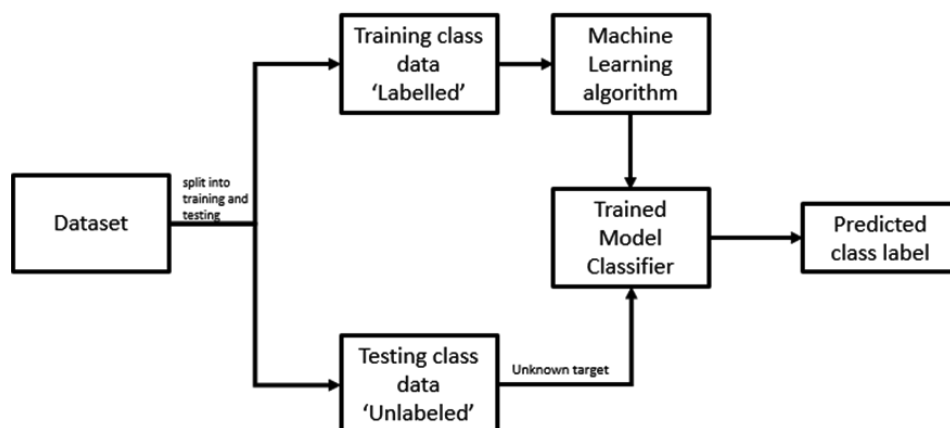
Fig. 1. The prediction process through Machine learning model

development of computer aided diagnostic support systems for CKD detection.

In[12] a development of an integrated model is proposed. Firstly an individual set of machine learning algorithms were used for the diagnosis development from the sample dataset. Then the optimally performing classifiers were selected on the basis of correct judgments to form an integrated approach. The output class primarily belongs to ckd or notckd as predictors. After rigorous analysis and testing the final model was develop by integration of Regression based model and Tree based model. The accuracy thus obtained after integration is 99.83%. Therefore this method may serve as an efficient diagnostic support system in real time applications.

An approach utilizing the principles of data mining and machine learning is presented in[13]. Several classifiers were tested on 400 samples containing 24 variables and 1 predictor. The least performing classifier is found to be Naïve Bayes and the best performing came out to be Random Forest classifier. An overall accuracy of 99% is obtained using this approach.

Utilizing different models of machine learning and deep learning is presented in[14]. The researchers tested various models and a deep learning model and found it over performing the machine learning models. Feature optimization methodology based on 3 feature selection algorithms was developed in order to select the most optimum features to ensure higher levels of accuracy. Majorly the division of training and testing data is done as 50%, 50%. Synthetic Minority Oversampling Technique (SMOTE) in hybridization with regression model gives the optimum results.

A method developed based on ensemble learning is proposed in[15] to predict the risk level associated with CKD. Major researchers have published their methods based upon dataset containing 400 samples. The number of samples for the proposed model in [15] are 1 million samples which validates the performance of the proposed model over good figures of training and testing samples. The stage of CKD depends upon the value of parameter glomerular filtration rate (GFR). The prediction of creatinine based on regression model is appended to 23 features to form a set of 24 predictor variables for evaluating risk/stage of CKD.

The performance evaluation of multiple classifiers for early detection of CKD is presented in[16]. The performance of classifiers is tuned using adjustment of Hyper-parameters thereby constructing ensemble model with ranking weight. The accuracy achieved using the mentioned approach is 98.75%.

A deep learning based approach is presented in[17]. The researchers' developed model based upon selection of most prominent features thereby improving accuracy. Recursive Feature Elimination (RFE) method is deployed to select the optimum feature set with respect to class label. The deep learning method gives high range of accuracy and can serve as supportive tool for CKD detection by nephrologists.

Another model based on ensemble method is proposed in[18]. Random subspace ensemble method is found to be the most optimally performing classifier as compared to individual classifier. The implementation is done on the dataset available through UCI repository containing 400 samples. The data pre-conditioning is performed optimally by preserving the samples rather than deleting the samples for missing values. The accuracy achieved through the proposed model is found to be 1.00 while other parameters also in the higher ranges.

Majorly the methodologies presented in the literature relies upon dataset containing clinical parameters. Few of the researchers have worked upon utilizing 2D Ultrasound images as dataset. The prediction of the risk factors associated with CKD depending upon the values of GFR is also presented in the different articles [19-22].

MATERIAL AND METHODS

The methodology is divided into three parts: Data pre-processing, Model development followed by model performance evaluation through quality metrics as shown in figure 3.

Dataset and its processing

To carry out the research work the dataset available through UCI Machine learning repository is utilized[23]. The dataset contains 400 instances of CKD patients. The total number of attributes (predictors) are 25 wherein 11 belongs

to numeric set and 14 belongs to nominal set. The target class corresponds to binary classification case containing 2 possible values ckd or non ckd. While analyzing the dataset, few samples containing missing values are witnessed. Therefor data pre-processing is an integral step to ensure unbiased prediction. In order to complete the missing values the corresponding value available in the next row for sample is backfilled[24]. For data split we have tested on 2 scenarios like 70:30, 80:20. The ratio selection 80:20 is performing optimally. 320 instances correspond to training and 80 instances correspond to testing portion of the data as shown in figure 3. A 5 cross validation methodology is deployed for validation purposes. Clean and preprocess the collected data to ensure its quality and compatibility with the machine learning algorithms.

Figure 4 provides information about the attributes available in the dataset. In total 24 attributes along with their nominal ranges/ units are described through Table 1. For handling nominal values as they belongs to different nominal scales a standardized procedure is to be applied to get the values on the same scale. The original value are mapped to standardized scale.

Model Development

While making choice of an appropriate machine learning algorithm or integration of algorithms to train the kidney disease detection model, a rigorous set of experimentation is required in order to get best performance. The selection of the optimum model depends primarily on the nature of the problem (classification, regression), availability of dataset (size and type) and the computational resources availability[25]. Machine learning models are extensively being applied in medical diagnosis applications due to their ability to analyze and process complex patterns present in the data and make predictions for unknown cases based on large amounts of data [26, 27].

From the literature we have studied a ppopular set of algorithms for classification tasks include decision trees, random forests, support vector machines (SVMs), and deep learning models like convolutional neural networks (CNNs) or rrecurrent neural networks (RNNs). The following section discuss the predominantly used classifiers for CKD detection:

Logistic Regression (LR)

For binary classification applications like the identification of CKD, logistic regression is a common solution. The target variable in the dataset possess 2 values ckd or notckd. Thereby making LR an optimum choice for the problem under consideration. Based on the input feature vector, this algorithm determines the probabilities of class membership with respect to target class variable. The assumption behind logistic regression is that the target variable's log-odds and features are linearly related.
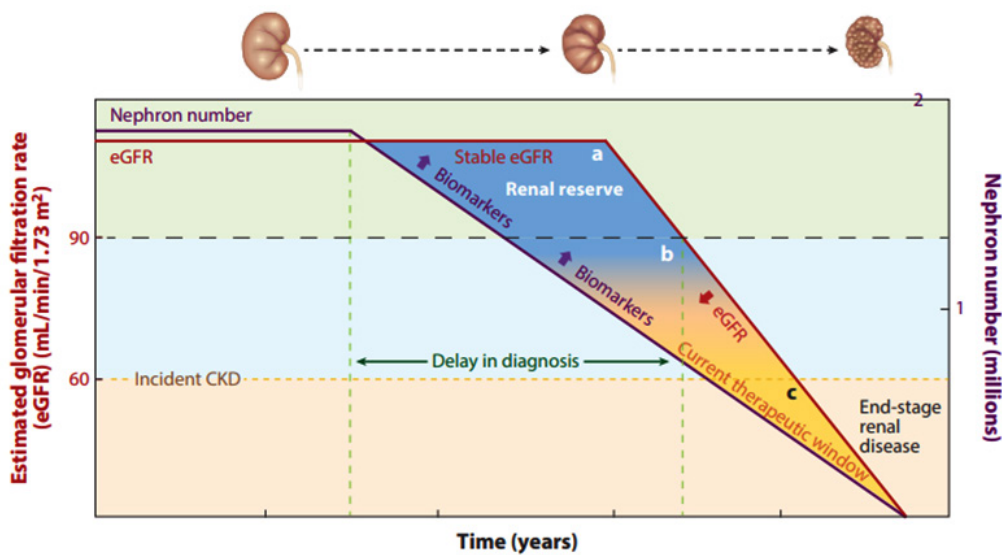

Fig. 2. Stages of CKD w.r.t eGFR[5]

Support Vector Machines (SVMs)

The flexible models known as SVMs are capable of handling both linear and non-linear classification tasks. They operate by locating an ideal hyperplane in a high-dimensional feature space that maximally separates the classes. When dealing with non-linear relationships or with tiny datasets, SVMs can be quite useful.

Random Forest

An ensemble learning technique called random forests combines various decision trees to produce predictions. They are renowned for their robustness against overfitting and are capable of handling both classification and regression problems. Random forests are appropriate for chronic renal disease identification where several
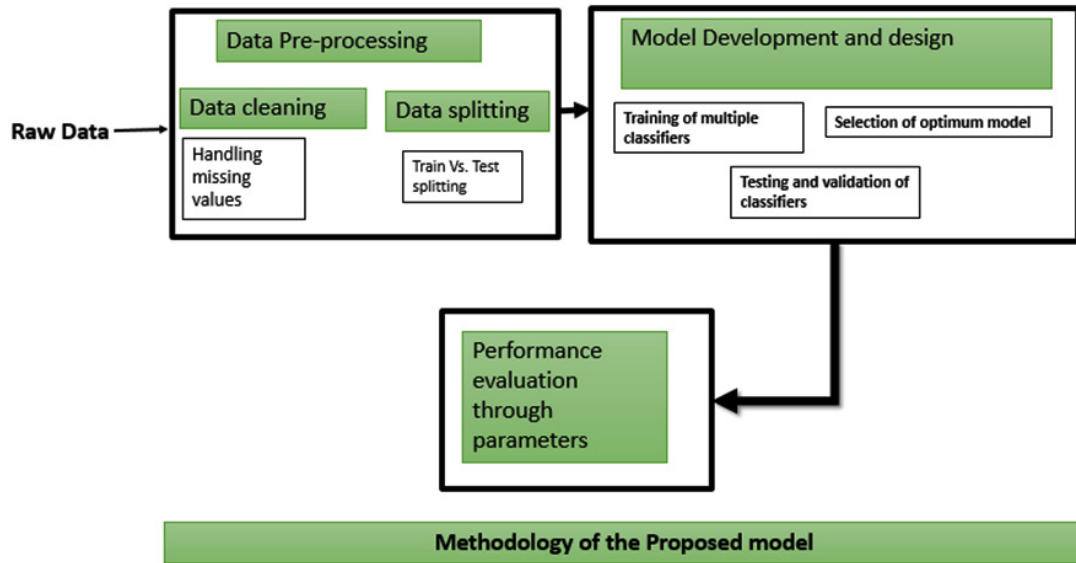

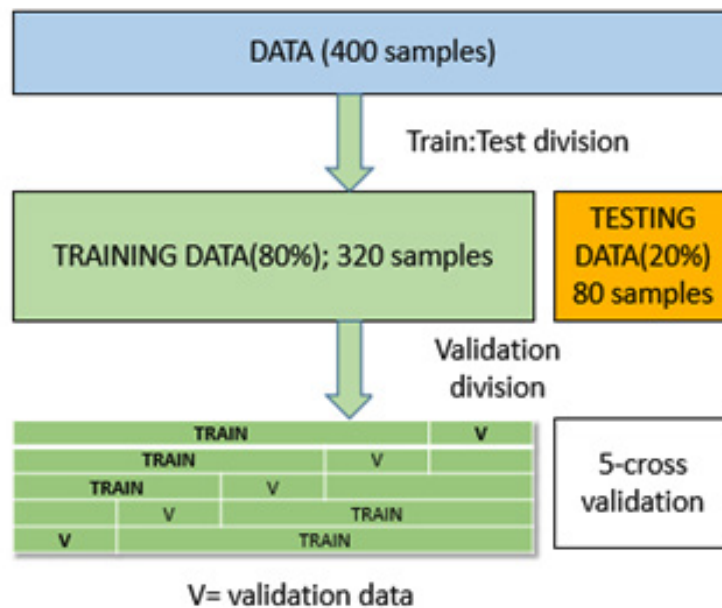**Fig. 3.** The Proposed model for CKD detection

Fig. 4. Data and its split up

data types may be involved since they can handle a wide range of feature types, including numerical and categorical variables.

Gradient Boosting Models

The high predictive performance of gradient boosting models like Gradient Boosting Machines (GBM), XGBoost, and LightGBM has made them popular. These models sequentially create a group of ineffective learners (often decision trees), with each new model fixing the flaws of the prior one. It is well known that gradient

boosting models may detect intricate connections and patterns in data.

Deep Learning Models

Convolutional and recurrent neural networks (CNNs) in particular have demonstrated promising outcomes in a number of medical applications. CNNs are excellent at analyzing spatial patterns, which makes them useful for image-based methods of detecting kidney disease, such as examining kidney MRI or ultrasound pictures. For sequential data, such as time-series

Table 1. Predictors and target variables present in data

Attribute Name	Type	Measurement Unit/Range	Attribute category
Patient age (age)	Numerical	Years	24 Predictors
Blood Pressure (bp)	Numerical	mm/Hg	
Specific Gravity (sg)	Nominal	(1.005,1.010, 1.015,1.020,1.025)	
Albumin (al)	Nominal	(0,1,2,3,4,5)	
Sugar (su)	Nominal	(0,1,2,3,4,5)	
Red Blood Cells (rbc)	Nominal	normal,abnormal	
Pus Cell(pc)	Nominal	normal,abnormal	
Pus Cell clumps (pcc)	Nominal	present,notpresent	
Bacteria (ba)	Nominal	present,notpresent	
Blood Glucose Random (bgr)	Numerical	mgs/dl	
Blood Urea (bu)	Numerical	mgs/dl	
Serum Creatinine (sc)	Numerical	mgs/dl	
Sodium (sod)	Numerical	mEq/L	
Potassium (pot)	Numerical	mEq/L	
Hemoglobin (hemo)	Numerical	gms	
Packed Cell	Numerical		
White Blood Cell Count (wc)	Numerical	cells/cumm	
Red Blood Cell Count (rc)	Numerical	millions/cmm	
Hypertension (htn)	Nominal	yes,no	
Diabetes Mellitus (dm)	Nominal	yes,no	
Coronary Artery Disease (cad)	Nominal	yes,no	
Appetite (appet)	Nominal	good,poor	
Pedal Edema (pe)	Nominal	yes,no	
Anemia (ane)	Nominal	yes,no	
Class	Nominal	ckd, notckd	1 Target with two response classes

Table 2. Performance evaluation scenarios

Name	Training set instances	Testing set instances	Training Vs. testing+ Validation ratio
Scenario 1(sc1)	280	120	70:30
Scenario 2 (sc2)	320	80	80:20

measurements or text from medical records, RNNs are effective.

In order to reduce the biases of individual models and enhance overall prediction performance, ensemble models make use of the diversity of numerous models. They are able to manage complex patterns, lessen overfititng, and provide predictions that are more reliable. It is significant to note that compared to individual models, ensemble models may offer more complexity and computing needs, therefore factors like computational resources and model interpretability should be taken into account.

Experimental setup for Model Testing and Analysis

For the validation and performance evaluation of the proposed model for CKD detection, dataset consisting 400 samples with 150 ckd instances and 250 notckd instances is utilized. The ratio of target to test split is explicitly taken up for experimentation as 70:30 and 80:20 so as to obtain optimum performance through rigorous testing and experimentation.

We have evaluated the performance for different models considering both scenarios so as to select optimum train test ratio and select the suitable classifier model for CKD detection. Figure 6 defines the structure of the confusion matrix and the reference terminology referring to presence and absence of disease.

The metric used for performance evaluation and comparison of various machine learning classifiers is represented through equation 1. Accuracy is a direct measure of performance so we have focused on calculation of accuracy for model selection.

$$ACC = \frac{n_{ckd \to ckd} + n_{notckd \to notckd}}{n_{ckd \to ckd} + n_{notckd \to notckd} + n_{notckd \to ckd} + n_{ckd \to notckd}}$$

$$...(1)$$

Input	Have the disease	Does not have the disease
Have the disease (ckd)	*True positive(TP)-$n_{ckd \to ckd}$*	*False Positive (FP)- $n_{ckd \to notckd}$*
Does not have the disease (notckd)	*False negative (FN)- $n_{notckd \to ckd}$*	*True Negative (TN)- $n_{notckd \to notckd}$*

Fig. 5. Confusion Matrix for CKD detection

Table 3. Model performance terms of accuracy

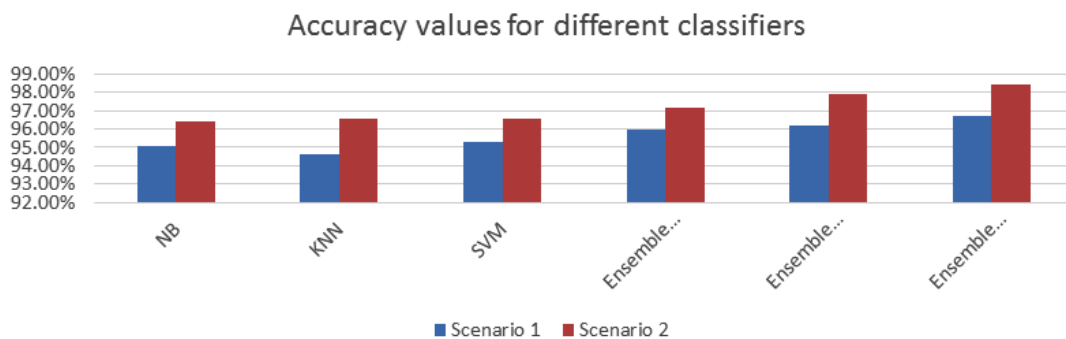| Feature vector	NB	KNN	SVM	Ensemble (Boosted Trees)	Ensemble (Subspace /KNN)	Ensemble (Subspace/ discriminant)
Scenario 1	95.1%	94.6%	95.3%	96%	96.2%	96.7%
Scenario 2	96.4%	96.6%	96.6%	97.2%	97.9%	98.4%

Fig. 6. Accuracy comparison for different classifiers

Table 3 is depicting the values of maximum accuracy obtained with different set of classifiers for two different scenarios as mentioned in table 2. By analyzing this table we can deduce that for this dataset Ensemble (Subspace/discriminant) is giving best results in all the scenarios developed for testing.

Figure 7 presents comparative performance of different classifiers in terms of accuracy value. As evident from the results Ensemble model is giving best performance as the underlying principle behind its operation is majority voting rule. The purpose of the proposed method is to make the system design in efficient manner so the disease detection process through automation is capable of replacing human efforts.

CONCLUSION

This work aims to develop significant contribution for designing computer aided support system for nephrologists. The system thus developed aims at providing decision support to the medical practitioners. The prevalence of CKD is one of the major threat to human population as it causes serious health effects. We have tested the performance of base classifiers as well as the ensemble approaches over the same dataset considering two scenarios by varying test train split ratio. From our experimentation we have found the better performance of ensemble methods over individual classifiers. In order to improve classification performance across a variety of domains, including medical diagnosis, the subspace discriminant ensemble approach offers a framework for merging feature subspace selection and ensemble learning. This approach utilizes the characteristics of different classifies and develop an ensemble based approach. It's important to remember that the specific problem and dataset characteristics can influence the feature subspace selection method, classifier, and combination strategy selected. We have achieved an accuracy value of 98.4% with train vs. test ratio selected as 80:20 and 96.7% with train vs. test ratio of 70:30. Also we have tested different classifiers for 2 different scenarios considering different training and testing division. This clearly indicates the importance of training the models with higher number of samples containing labelled data. In conclusion, ensemble learning techniques are quite successful at detecting chronic kidney disease (CKD). Ensemble models can take advantage of each model's advantages by integrating the predictions of several different individual models, which boosts overall predictive performance.

REFERENCES

1. Bhaskar, N.; Suchetha, M.; Philip, N.Y. Time Series Classification-Based Correlational Neural Network With Bidirectional LSTM for Automated Detection of Kidney Disease. *IEEE Sens. J.* 2021, *21*, 4811–4818.
2. Singh, Vijendra, Vijayan K. Asari, and Rajkumar Rajasekaran. "A deep neural network for early detection and prediction of chronic kidney disease." *Diagnostics* 12, no. 1 (2022): 116.
3. Kovesdy CP. Epidemiology of chronic kidney disease: an update 2022. Kidney Int Suppl (2011). 2022 Apr;12(1):7-11. doi: 10.1016/j.kisu.2021.11.003. Epub 2022 Mar 18. PMID: 35529086; PMCID: PMC9073222.
4. Drew, David A., Daniel E. Weiner, and Mark J. Sarnak. "Cognitive impairment in CKD: pathophysiology, management, and prevention." *American Journal of Kidney Diseases* 74, no. 6 (2019): 782-790.
5. Zhang, William R., and Chirag R. Parikh. "Biomarkers of acute and chronic kidney disease." *Annual review of physiology* 81 (2019): 309-333.
6. Ma, Fuzhe, Tao Sun, Lingyun Liu, and Hongyu Jing. "Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network." *Future Generation Computer Systems* 111 (2020): 17-26.
7. Almasoud, Marwa, and Tomas E. Ward. "Detection of chronic kidney disease using machine learning algorithms with least number of predictors." *International Journal of Soft Computing and Its Applications* 10, no. 8 (2019).
8. M. Patricio et al., "Using resistin, glucose, age and BMI to predict the presence of breast cancer,"

BMC CANCER, vol. 18, Jan. 2018.

9. R. J. Kate et al., "Prediction and detection models for acute kidney injury in hospitalized older adults," Bmc. Med. Inform. Decis., vol. 16, Mar. 2016.

10. Y. Chen et al., "Machine-learning-based classification of real-time tissue elastography for hepatic fibrosis in patients with chronic hepatitis B," Comput. Biol. Med., vol. 89, pp. 18-23, Oct. 2017.

11. N. Park et al., "Predicting acute kidney injury in cancer patients using heterogeneous and irregular data," Plos One, vol. 13, no. 7, Jul. 2018. [22] [23] X. Wang et al., "A new effective machine learning framework for sepsis diagnosis," IEEE Access, vol. 6, pp. 48300-48310, Aug. 2018.

12. Qin, Jiongming, Lin Chen, Yuhua Liu, Chuanjun Liu, Changhao Feng, and Bin Chen. "A machine learning methodology for diagnosing chronic kidney disease." IEEE Access 8 (2019): 20991-21002.

13. Emon, Minhaz Uddin, Rakibul Islam, Maria Sultana Keya, and Raihana Zannat. "Performance analysis of chronic kidney disease through machine learning approaches." In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pp. 713-719. IEEE, 2021.

14. Chittora, Pankaj, Sandeep Chaurasia, Prasun Chakrabarti, Gaurav Kumawat, Tulika Chakrabarti, Zbigniew Leonowicz, Micha³ Jasiñski et al. "Prediction of chronic kidney disease-a machine learning perspective." *IEEE Access* 9 (2021): 17312-17334.

15. Wang, Weilun, Goutam Chakraborty, and Basabi Chakraborty. "Predicting the risk of chronic kidney disease (ckd) using machine learning algorithm." *Applied Sciences* 11, no. 1 (2020): 202.

16. Srivastava, Swapnita, Rajesh Kumar Yadav, Vipul Narayan, and Pawan Kumar Mall. "An Ensemble Learning Approach For Chronic Kidney Disease Classification." *Journal of Pharmaceutical Negative Results* (2022): 2401-2409.

17. Singh, Vijendra, Vijayan K. Asari, and Rajkumar Rajasekaran. 2022. "A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease" *Diagnostics* 12, no. 1: 116. https://doi.org/10.3390/diagnostics12010116

18. Jongbo, Olayinka Ayodele, Adebayo Olusola Adetunmbi, Roseline Bosede Ogunrinde, and Bukola Badeji-Ajisafe. "Development of an

ensemble approach to chronic kidney disease diagnosis." *Scientific African* 8 (2020): e00456.

19. Revathy, S., B. Bharathi, P. Jeyanthi, and M. Ramesh. "Chronic kidney disease prediction using machine learning models." *International Journal of Engineering and Advanced Technology* 9, no. 1 (2019): 6364-6367.

20. Xiao, Jing, Ruifeng Ding, Xiulin Xu, Haochen Guan, Xinhui Feng, Tao Sun, Sibo Zhu, and Zhibin Ye. "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression." *Journal of translational medicine* 17, no. 1 (2019): 1-13.

21. Kriplani, Himanshu, Bhumi Patel, and Sudipta Roy. "Prediction of chronic kidney diseases using deep artificial neural network technique." In *Computer aided intervention and diagnostics in clinical and medical images*, pp. 179-187. Springer International Publishing, 2019.

22. Jena, Lambodar, Bichitrananda Patra, Soumen Nayak, Sushruta Mishra, and Sushreeta Tripathy. "Risk prediction of kidney disease using machine learning strategies." In *Intelligent and Cloud Computing: Proceedings of ICICC 2019, Volume 2*, pp. 485-494. Springer Singapore, 2021.

23. UCI machine learning repository. (2015):Chronic kidney disease dataset. Accessed February 2023, fromhttp://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_disease

24. Jongbo, Olayinka Ayodele, Adebayo Olusola Adetunmbi, Roseline Bosede Ogunrinde, and Bukola Badeji-Ajisafe. "Development of an ensemble approach to chronic kidney disease diagnosis." *Scientific African* 8 (2020): e00456.

25. Sharma, Vivek Kumar, Thakur Gurjeet Singh, Nikhil Garg, Sonia Dhiman, Saurabh Gupta, Md Habibur Rahman, Agnieszka Najda et al. "Dysbiosis and Alzheimer's disease: a role for chronic stress?." *Biomolecules* 11, no. 05 (2021): 678.

26. Kaushal, Chetna, Shiveta Bhat, Deepika Koundal, and Anshu Singla. "Recent trends in computer assisted diagnosis (CAD) system for breast cancer diagnosis using histopathological images." *Irbm* 40, no. 4 (2019): 211-227.

27. Shukla, Prashant Kumar, Jasminder Kaur Sandhu, AnamikaAhirwar, DeepikaGhai, PritiMaheshwary, and Piyush Kumar Shukla. "Multiobjective genetic algorithm and convolutional neural network based COVID-19 identification in chest X-ray images." Mathematical Problems in Engineering 2021 (2021): 1-9.