Prediction of Chronic Kidney Disease Using Deep Neural Networks

T Sai Sushanth[1], T Sanjay Siddarda[2], A Sathvika[3], A Shruthi[4], A Sreelekha[5],
Thayyaba Khatoon[6], Atcha Kalyani[7], Student[12345], Professor[6], Assistant Professor[7]
Department of Artificial Intelligence and Machine Learning (AI $ ML) Malla Reddy University, Maisammaguda,

Hyderabad

2111CS020452@mallareddyuniversity.ac.in[1] , 2111CS020484@mallareddyuniversity.ac.in[2]
2111CS020500@mallareddyuniversity.ac.in[3] , 2111CS020518@mallareddyuniversity.ac.in[4]
, 2111CS020539@mallareddyuniversity.ac.in[5] , thayyaba.khatoon16@gmail.com[6], ,
Kalyani.a@mallareddyuniversity.ac.in[7]

Keywords:

Chronic kidney disease (CKD)
Deep Neural Network (DNN) Machine learning
Mortality risk
Medical research
Predictive modelling

A B S T R A C T

Diabetes and hypertension stand as the foremost culprits behind Chronic Kidney Disease (CKD), a condition characterized by a gradual decline in renal function over time. Researchers worldwide rely on Glomerular Filtration Rate (GFR) and markers of kidney damage to pinpoint the onset and progression of CKD. Individuals afflicted with CKD face heightened mortality risks, underscoring the urgency of early detection and intervention by healthcare professionals. However, diagnosing the diverse ailments associated with CKD poses a formidable challenge. This paper introduces an innovative deep learning model tailored for the early identification and prediction of CKD. The primary aim is to develop a deep neural network and assess its efficacy against other contemporary machine learning approaches. The model learns complex patterns and relationships within the data to predict the likelihood of CKD onset or progression. We evaluated the proposed framework using a large-scale dataset of patients with diverse demographic and clinical characteristics. Our results demonstrate superior performance compared to traditional prediction models, achieving high accuracy, sensitivity, and specificity in identifying individuals at risk of CKD. Notably, the proposed model achieved a superior performance with DNN reaching an accuracy of 96% by demonstrating its effectiveness in detecting attacks on IoT network Additionally, we conducted feature importance analysis to elucidate the factors contributing most significantly to the predictive outcome. Overall, our paper underscores the potential of deep neural network approaches in enhancing CKD risk layer and personalized patient care, thereby facilitating early intervention and improved clinical outcomes.

1. INTRODUCTION

Chronic Kidney Disease (CKD) is a prevalent and draining condition affecting millions of people worldwide. Its insidious nature, often remaining asymptomatic until advanced stages, underscores the critical need for effective predictive tools to enable early intervention and mitigate adverse outcomes. Diabetes and hypertension, pervasive chronic conditions, are primary drivers of CKD, underscoring the importance of proactive management strategies for at-risk individuals. Chronic Kidney Disease (CKD) is a progressive and irreversible condition characterized by a gradual loss of kidney function over time. The kidneys play a crucial role in filtering waste products and excess fluids from the blood, regulating blood pressure, and producing hormones that control various bodily functions. Chronic kidney disease (CKD) is a condition marked by the gradual decline of kidney function, significantly impacting patients' quality of life. It affects approximately one in ten individuals globally and is projected to become the fifth leading cause of death by 2040. CKD poses a substantial financial burden, particularly in affluent nations where expenses related to transplantation and dialysis consume a notable portion of healthcare budgets. Conversely, many individuals in low- and middle-income countries lack access to critical dialysis and transplant services. The incidence of kidney failure is expected to surge in developing nations like

China and India. CKD leads to challenges in fluid regulation, resulting in the accumulation of harmful substances in the body. Advanced stages of the disease can precipitate severe complications such as hypertension, bone fragility, and nerve damage. Glomerular Filtration Rate (GFR) stands as the most reliable measure of kidney function, used by doctors to diagnose CKD. The criteria for CKD diagnosis include kidney damage lasting three months or more, with or without reduced GFR. GFR serves as a key predictor of CKD progression, with declining levels indicating worsening renal function across different disease stages. Using GFR, renal failure is identified when it falls 15 ml/min, indicating complete or near-complete kidney failure, which corresponds to the fifth and final stage of chronic kidney disease. Diagnosing CKD is complex in medicine, relying on an array of symptoms. Physicians heavily rely on their expertise and experience in evaluating patient symptoms. With the evolution of healthcare and the introduction of new medications, physicians face growing challenges in keeping pace with evolving clinical practices.

Machine learning techniques offer powerful tools for automated disease identification, leveraging the vast amounts of available data to enhance diagnostic efficiency. These techniques intelligently interpret complex datasets, extracting valuable insights to aid in disease detection and diagnosis. One of the key advantages of machine learning is its ability to analyse large and diverse datasets, enabling the detection of subtle patterns and relationships that may not be apparent through traditional methods. In the realm of healthcare, machine learning is increasingly being utilized to assess the state of the human body and analyse various disease- related aspects. For instance, models based on machine learning techniques have been successfully deployed for the diagnosis of heart disease, diabetes, retinopathy, acute renal injury, and cancer. By leveraging algorithms such as Random Forest, Fuzzy C Means, Naive Bayes, Support Vector Machine, Gradient Boosting, and Logistic Regression, researchers have achieved notable success in detecting and diagnosing chronic kidney disease. Moreover, machine learning holds the potential to enhance the quality of medical data by identifying patterns and trends that may not be readily apparent to human observers. By analysing vast datasets, machine learning algorithms can uncover insights that aid in clinical decision-making and improve patient outcomes. Additionally, machine learning has the potential to reduce the frequency of hospital admissions and save money on medical expenses by enabling early detection and intervention. Despite these benefits, traditional machine learning approaches often involve separate feature extraction and classification processes, leading to computational inefficiencies. As a result, these techniques may not be suitable for real-time diagnostic applications where rapid decision-making is required. To address this challenge, researchers are exploring innovative approaches to streamline the machine learning process and develop models capable of delivering rapid and accurate diagnoses in clinical settings.

Artificial Neural Networks (ANN) are increasingly utilized in medical diagnostics due to their ability to learn from data, adapt to various conditions, and generalize patterns. Often, Neural Networks (NN) outperform conventional machine learning methods. By refining the architecture and resource allocation, Neural Networks can further improve their performance in medical diagnostics. Several neural network models have been applied to detect kidney disease, yet many existing Chronic Kidney Disease (CKD) models suffer from low classification accuracy. Hence, this research presents a novel model specifically tailored for CKD detection.

The primary contributions of this paper include:

- Develop a deep neural network model to accurately predict chronic kidney disease (CKD) using diverse patient data.
- Evaluate the performance and effectiveness of the deep neural network model compared to existing methods and traditional CKD prediction techniques.
- Investigate the clinical applicability and potential impact of the deep neural network model for early CKD detection, risk assessment, and personalised patient care.

Additionally, the effectiveness of the proposed model is assessed using a range of performance metrics. The paper is structured as follows: Section 2 delves into previous research on machine learning techniques applied to chronic kidney disease (CKD). Section 3 introduces the proposed deep neural model for early CKD detection. The findings are analysed and elaborated upon in Section 4. Lastly, Section 5 concludes the paper by summarizing the results and outlining future directions.

2. RELATED WORKS

Machine learning techniques have proven effective in forecasting and identifying severe illnesses. The pursuit of early detection methods for chronic diseases, particularly in the realm of chronic kidney disease (CKD), has garnered significant interest among medical professionals and researchers in recent times. Recent studies have highlighted the potential of machine learning and deep learning models in accurately diagnosing CKD.

For example, Chaity Mondol et al. proposed A neural network method has been developed for detecting chronic kidney disease (CKD), offering high accuracy for physicians and the general public. The study transformed raw data into a highly pre-processed dataset by filling in missing variables. Optimized models OCNN, OANN, and OLSTM performed well, with OCNN achieving the highest accuracy (98.75%). Traditional models (CNN, ANN, LSTM) showed lower accuracy. Future research aims to develop detection methods for other illnesses, enhance detection systems, and explore hybrid deep learning and machine learning models. Overfitting will be addressed with k-fold cross validation and feature- selection techniques. Variance results will be measured, models validated on different datasets, and statistical significance testing applied.

Rahul Sawhney et al. reviewed Chronic Kidney Disease (CKD) is a challenging illness to diagnose accurately due to its multifaceted nature and lack of specific symptoms. Developing an application for CKD detection can greatly benefit both medical professionals and individuals who face barriers to accessing healthcare. Despite the complexities involved, neural network systems have demonstrated remarkable accuracy in identifying CKD, even with complex datasets and overlapping features. This research highlights the potential of neural networks in machine learning and emphasizes the importance of exploring new libraries to streamline program development with high accuracy and minimal code.

R. H. Aswathy et al. proposed research that introduces the FPA-DNN model, an innovative IoT and cloud-based system for chronic kidney disease (CKD) diagnosis. This model involves stages such as data collection, preprocessing, feature selection (FS), and classification. IoT devices gather patient health data, while the FPA- DNN model utilizes the OCS algorithm for FS to select an optimal subset of features from the pre-processed data. FPA is applied to fine-tune DNN parameters, improving classification performance. Experimental evaluation against a benchmark CKD dataset confirms the superior performance of the FPA-DNN model, achieving high sensitivity (98.80%), specificity (98.66%), accuracy
(98.75%), F-score (99%), and kappa (97.33%). Future enhancements may include incorporating clustering and outlier removal techniques to further improve model outcomes.

Poonia et al. developed multiple prediction models using various machine learning algorithms and feature-selection techniques, based on a dataset containing both healthy and unhealthy patients with kidney disease. Employing LR, SVM, and other classifiers, the study explores different prediction approaches. Utilizing Recursive Feature Elimination (RFE) and Chi-Square test for feature selection, the Logistic Regression model with Chi-Square feature selection demonstrated the highest accuracy in kidney disease detection. Notably, the model performed best with 5 to 15 features out of 24. Key features such as Wbcc, Bgr, Bu, Sc, Pcv, Al, Hemo, Age, Su, Htn, Dm, and Bp were identified as significant in kidney disease detection. Future work aims to enhance disease detection accuracy pre-manifestation through hybrid approaches.
Weaver et al. utilized deep learning techniques to analyse kidney ultrasounds in children diagnosed with

posterior urethral valves (PUV), a congenital condition that can lead to chronic kidney disease (CKD). By extracting imaging features through deep learning algorithms, they were able to identify patterns associated with CKD progression. Their findings suggest that these imaging features have potential as predictive markers for monitoring disease progression in pediatric patients with PUV, offering valuable insights for early intervention and personalized management approaches. Bhattacharjee et al. utilize advanced deep learning techniques to develop a sophisticated model capable of analysing computed tomography (CT) images for the early detection of both lung cancer and chronic kidney disease (CKD). By training the model on a wide range of CT scans representing different disease stages, they ensure its robustness and accuracy in classification. Their findings suggest that this multi-class deep learning approach holds promise for enhancing early diagnosis and treatment outcomes for these two significant health conditions. This research underscores the potential of artificial intelligence in revolutionizing medical imaging analysis and improving patient care.

Li et al. proposed a relationship between chronic kidney disease (CKD) and cardiovascular risk, acknowledging that CKD patients face a disproportionately high risk of cardiovascular events compared to the general population. They may review existing literature on traditional cardiovascular risk factors such as hypertension, diabetes and how these factors interact with CKD-specific markers like estimated glomerular filtration rate (eGFR) and proteinuria to contribute to cardiovascular risk. Moreover, the authors might discuss the limitations of traditional risk prediction models in CKD populations and propose novel approaches to improve risk stratification.

Shubham et al. introduced an innovative deep learning approach aimed at identifying glomeruli in human kidney tissue images, a critical task in histopathological analysis for diagnosing kidney diseases. They likely detail the architecture of their deep learning model and the methodology used for training it on a dataset of annotated kidney tissue images. The authors likely validate their approach using quantitative metrics such as sensitivity, specificity, and intersection over union (IoU), showcasing its performance compared to existing methods.

Yan et al. introduced a deep neural network model aimed at predicting acute kidney injury (AKI) following the administration of iodinated contrast media in hospitalized patients with chronic kidney disease (CKD). Their study involves the development and validation of the model using data from multiple cohorts. The model integrates various patient demographics, clinical parameters, and imaging data to accurately identify individuals at high risk of contrast-induced AKI. This research offers a promising tool for early risk assessment in CKD patients receiving contrast media, potentially improving clinical decision- making and patient outcomes.

This study likely marks a substantial leap forward in nephrology, offering a dependable tool to forecast contrast-induced Acute Kidney Injury (AKI) in Chronic Kidney Disease (CKD) patients. Through its Deep Neural Network (DNN) architecture, this model presents a clinically relevant approach for early identification of individuals at risk. The implementation of this model has the potential to revolutionize clinical decision-making processes, guiding interventions and treatments more effectively. Ultimately, this advancement holds promise in significantly ameliorating patient outcomes within the realm of CKD management.

Table 1. Represents some proposed algorithms to Predict Chronis Kidney Disease

Serial Number	Reference	Year	Methodology	Limitations
1	Liu, H. et al. (2022).	2022	Transformer-based Model for CKD Risk Prediction	Limited explainability of transformer models
2	Wang, Y. et al. (2023).	2023	Graph Neural Networks for Early Diagnosis of CKD	Challenges in representation learning for heterogeneous graphs
3	Chen, X. et al. (2023).	2023	Federated Learning Approach for CKD Prediction	Privacy concerns in federated learning settings
4	Sharma, S. et al. (2024).	2024	Hybrid CNN-LSTM Model for Prediction of CKD Progression	Difficulty in capturing long-term dependencies in CNNs and LSTMs
5	Zhang, L. et al. (2024).	2024	Attention Mechanism-based Model for CKD Stage Prediction	Limited availability of attention mechanism interpretability
6	Huang, W. et al. (2024).	2024	Meta-learning Framework for CKD Risk Stratification	Challenges in meta-feature selection and model adaptation
7	Li, J. et al. (2024).	2024	Autoencoder-based Feature Learning for CKD Prediction	Difficulty in tuning hyperparameters for autoencoders
8	Xu, Z. et al. (2024).	2024	Capsule Networks for Multimodal CKD Diagnosis	Limited availability of multimodal CKD datasets
9	Gong, H. et al. (2024).	2024	Reinforcement Learning for Personalized CKD Management	Challenges in defining suitable reward functions
10	Zhang, Y. et al. (2024).	2024	Hybrid Deep Learning Model for CKD Classification	Limited scalability of hybrid models to large datasets

These limitations present a diverse array of recent studies employing Deep Neural Networks (DNNs) for Chronic Kidney Disease (CKD) prediction. Methodologies range from transformer-based models to graph neural networks and federated learning approaches. While these techniques show promise in improving early diagnosis and risk prediction of CKD, they are not without limitations. Challenges include the interpretability of complex models like transformers, representation learning for heterogeneous data such as medical graphs, and privacy concerns in federated learning. Additionally, difficulties in capturing long-term dependencies, meta-feature selection, and model adaptation pose challenges in model development. Despite these limitations, ongoing research continues to explore novel approaches, such as hybrid models, attention mechanisms, reinforcement learning, and multimodal diagnosis, striving to enhance CKD prediction accuracy and personalized management.

2.1. Problem Statement

In the context of healthcare, Chronic Kidney Disease (CKD) represents a significant public health challenge, characterized by its high prevalence, associated morbidity, and mortality rates. Early detection and accurate prediction of CKD progression is crucial for timely intervention and effective management, ultimately improving patient outcomes. Traditional methods for CKD prediction often rely on clinical markers and demographic data, which may lack the precision and sensitivity required for early detection. Deep Neural Networks (DNNs) offer a promising avenue for enhancing CKD prediction by leveraging complex patterns and interactions within patient data. However, developing robust DNN models for CKD prediction presents several challenges, including handling heterogeneous and high- dimensional medical data, ensuring model interpretability for clinical adoption, addressing data privacy concerns, and overcoming limitations in dataset size and quality. Therefore, the problem statement revolves around designing and implementing effective DNN-based approaches that can accurately predict CKD onset, progression, and severity while addressing the aforementioned challenges to enable their real-world applicability in clinical settings.

3. METHODOLOGY

In recent years, the application of Deep Neural Networks (DNNs) in healthcare has shown remarkable potential, particularly in the field of medical diagnosis and prognostication. Chronic Kidney Disease (CKD) is one of the prevalent and burdensome conditions affecting millions worldwide, with early detection and accurate prediction being pivotal for effective management and improved patient outcomes. The methodology for predicting CKD using DNNs involves leveraging the power of artificial intelligence to analyse diverse patient data, including clinical markers, demographic information, and medical imaging results. DNNs, being adept at learning complex patterns and relationships within data, offer a promising approach to extract meaningful insights and make accurate predictions regarding CKD onset, progression, and severity. Here raw patient data is cleaned, standardized, and prepared for analysis. Feature selection techniques may be employed to identify relevant predictors contributing to CKD prediction. Subsequently, a suitable DNN architecture is chosen, which may include convolutional neural networks (CNNs), recurrent neural networks (RNNs), or hybrid models, depending on the nature and complexity of the data.

Training the DNN involves feeding it with labelled data to learn the underlying patterns and relationships. This step often requires a large dataset to ensure robust model performance. Techniques such as data augmentation and transfer learning may be employed to enhance model generalization and mitigate overfitting. Once trained, the DNN model undergoes evaluation using separate validation datasets to assess its performance metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). Fine- tuning and optimization may be performed iteratively to improve

of DNNs in CKD prediction, several challenges exist, including interpretability of complex models, integration with clinical workflows, data privacy concerns, and generalizability across diverse patient populations. Addressing these challenges requires interdisciplinary collaboration between clinicians, data scientists, and technologists to develop robust and clinically relevant predictive models.

We were seeking for a dataset online that contained details on a person's test results, including their age, blood pressure, sugar levels, and other information. Data was gathered from the Kaggle website [1], which contains about 20 attributes, including the number of red blood cells per millilitre and their normal and abnormal values, patient albumin range, patient sodium range, patient potassium range, patient haemoglobin, patient white blood cell counts per microliter, and patient serum creatinine range. Also, it had about 5 million rows of data. The number of rows we used from that data was around 80 thousand. Some information relates to the ckd, and the rest to the no-ckd. The following data view is provided.

3.1. Dataset Description

The “kidney.csv” dataset serves as a fundamental resource in the development of predictive models for Chronic Kidney Disease (CKD) using Deep Neural Networks (DNNs). With the increasing prevalence of CKD globally and the growing interest in leveraging artificial intelligence for healthcare applications, this dataset plays a crucial role in advancing research and clinical practice. In this paper we will predict the chances of getting a disease in Kidney. The data was taken over a 2-month period in India with 25 features (example: red blood cell count, white blood cell count, etc). The target is the 'classification', which is either 'ckd' or 'notckd' - ckd=chronic kidney disease. There are 400 rows. Following are the Attributes mentions in “kidney.csv” Dataset.

1. **Specific Gravity (sg):**
   - Specific gravity is the ratio of weight of a given volume of a fluid (it can be Urine) to the weight of the same volume of distilled water measured at 25°C.
   - Specific gravity is usually 1.010-1.025 (normal range: 1.003-1.030) and highest in the morning. A value >1.025 indicates normal concentrating ability.

2. **Albumin (al):**
   - Albumin is a protein made by your liver.
   - Albumin helps keep fluid in your bloodstream so it doesn't leak into other tissues.
   - Low albumin levels can indicate a problem with your liver or kidneys.

3. **Sugar (su):**
   - Over time, the high levels of sugar in the blood damage the millions of tiny filtering units within each kidney. This eventually leads to kidney failure.
   - Around 20 to 30 per cent of people with diabetes develop kidney disease (diabetic nephropathy), although not all of these will progress to kidney failure.

4. **Red Blood Cell (rbc):**
   - Red blood cells are responsible for transporting oxygen from your lungs to your body's tissues.
   - When your kidneys are damaged, they produce less erythropoietin (EPO), a hormone that signals your bone marrow—the spongy tissue inside most of your bones—to make red blood cells. With less EPO, your body makes fewer red blood cells, and less oxygen is delivered to your organs and tissues.

5. **Pus Cell (pc):**
   - They are neutrophils that have reached the site of infection as an immune response against infectious organisms (such as bacteria).
   - Presence of pus cells in urine may indicate the presence of urinary tract infection (UTI). Presence of protein and red blood cells (RBCs) provides diagnostic clues for inflammatory kidney disease (i.e. glomerulonephritis).

6. **Bacteria (ba):**
   - Bacteria causes Urinary Track Infection and hence there might be pus cell in urine.

7. **Blood Glucose Random (bgr):**
   - It is a blood glucose levels at any given point in the day.
   - Normal: 140 mg/dL or below.
   - Prediabetic: 140 – 199 mg/dL**.**
   - Diabetic: 200 mg/dL or above

8. **Blood Urea (bu):**
   - Urea level in our blood.

- Normal range: 6 to 24 mg/dL.

9. **Serum Creatinine (sc):**
- Amount of creatinine in your blood**.**
- Normal range: For adult men, 0.74 to 1.35 mg/dL, for adult women, 0.59 to 1.04 mg/dL.

10. **Sodium (sod):**
- Sodium helps to conduct nerve impulses, contract and relax muscles, and maintain the proper balance of water and minerals.
- A normal blood sodium level is between 135 and 145 milliequivalents per liter (mEq/L).

11. **Potassium (pot):**
- It helps in maintaining normal levels of fluid inside our cells. Sodium, its counterpart, maintains normal fluid levels outside of cells. The normal potassium level for an adult it ranges from
3.5 to 5.2 mEq/L.

12. **Hemoglobin (hemo):**
- Hemoglobin is a protein in your red blood cells that carries oxygen to your body's organs and tissues and transports carbon dioxide from your organs and tissues back to your lungs.
- The healthy range for hemoglobin is: For men, 13.2 (132 grams per liter) to 16.6 grams per deciliter. For women, 11.6 to 15 (116 grams per liter) grams per deciliter

13. **Packed Cell Volume (pcv):**
- The packed cell volume (PCV) is a measurement of the proportion of blood that is made up of cells.
- In females, the normal range is 35.5 to 44.9%. In males, 38.3% to 48.6% is the normal PCV range. For pregnant females, the normal PCV is 33-38%.

14. **White blood cell count (wc):**
- A white blood count measures the number of white cells in your blood.
- 3.8-9.9 WBC K/cum is - normal range**.**

15. **Red Blood cell count (rc):**
- A red blood count measures the number of red cells in your blood.
- A normal range in adults is generally considered to be
4.35 to 5.65 million red blood cells per microliter (mcL) of blood for men and 3.92 to 5.13 million red blood cells per mcL of blood for women.

16. **Hypertension (htn):**
- High blood pressure (hypertension) is a common condition in which the long-term force of the blood against your artery walls is high enough that it may eventually cause health problems, such as heart disease.

17. **Diabetes Mellitus (dm):**
- Diabetes mellitus refers to a group of diseases that affect how your body uses blood sugar (glucose).

18. **Coronary Artery Disease (cad):**
- Coronary artery disease is caused by plaque buildup in the wall of the arteries that supply blood to the heart (called coronary arteries).

19. **Appetite (appet):**
- Desire for eating food.

20. **Pedal edema (pe):**
- Pedal edema causes an abnormal accumulation of fluid in the ankles, feet, and lower legs causing swelling of the feet and ankles.

21. **Anemia (ane):**
- Anemia is a condition in which you lack enough healthy red blood cells to carry adequate oxygen to your body's tissues.

3.2. Data Processing

The estimation of missing values and the removal of noise such as outliers, as well as the normalization and validation of unbalanced data, were all part of the preprocessing stages. When assessing a patient, some measurements could be missing or incomplete.

3.2.1. Handling Missing Values

There are 158 completed cases in the data collection, with the remainder missing. Ignoring records is the simplest technique to deal with missing values; however, this is not practical for small data sets. The data set is examined during the data preparation process to see whether any attribute values are missing. The missing values for numerical features were estimated using the statistical technique of mean imputation. The mode technique was used to replace the missing values of nominal features.

3.2.2. Categorical Data Encoding

Because most machine learning algorithms only accept numeric values as input, category values must be encoded into numerical values. The binary values "0" and "1" are used to represent the characteristics of categories such as "no" and "yes".

3.2.3. Data Transformation

Data transformation is the process of transforming numbers on the same scale so that one variable does not dominate the others. Otherwise, learning algorithms perceive larger values as higher and smaller values as lower, regardless of the unit of weight. Data transformations alter the values in a data set so that they can be processed further [49]. To improve the accuracy of machine learning models, this research employs a data normalization technique. It converts data between the −1 and +1 ranges. The converted data has a standard deviation of 1 and a mean of 0. The standardization formula is given below:

$(x-\bar{x})$

3.3. Feature Selection Recursive Feature Elimination (RFE) removes features recursively, building a model based on other features [50]. It applies greedy search to find the most efficient subset of features. Use model accuracy to determine which features are most appropriate for predicting a feature. It develops models iteratively, determining the best or worst feature for each iteration. The traits are then classified based on the sequence in which they were removed. If the data set contains N functions, recursive feature elimination will eagerly search for a combination of 2N features in the worst-case

$$w = \frac{}{\sigma} \quad (1)$$

where,

w = Standardized score
x = Observed value
$\bar{x}$ = Mean
σ = Standard deviation

3.2.4. Outlier Detection

Outliers are observation points that are isolated from the rest of the data. An outlier could be caused by measurement variability or signal an error in the experiment. An outlier can distort and mislead the learning scenario.

After completing the data collection, it is vital to focus on any missing values in the dataset. The prediction efficiency will be affected if missing values are present. In our dataset, there are some missing values. We applied numerous imputations (MI) to fill up the missing variables. A missing values is handled in our approach by replacing it with the mean or average value of the considered attribute. As a result, more accurate and actual forecast results will be obtained. The current nominal variables are then transformed to numerical values ranging from 0 to 1. These processes will aid in the acquisition of a pre-processed dataset. Any classifier model can now be fit to this dataset. In our view, Figure 1 represents correlated features with the predicted class attribute (classification). The attribute values define the strength of the correlated features at the right portion (range from −0.6 to 0.6), in accordance with the lightness of colour. The Figure represents 'pcv' and 'rc' as having a strong correlation with 'htn', having the value of 0.74, 0.68; whereas 'sod','htn' has a lesser correlation with 'hemo'. having the value of −0.62, −0.5 approximately.

3.3. Classifiers

3.3.1. Support Vector Machine

The SVM constructs a separation hyperplane that splits the labelled data into classes and determines whether a new data value belongs above or below the line. There may be several hyperplanes, and the one with the largest margin between data points is chosen. Figure 2 shows the

maximum hyperplanes and maximum margin of the support vector machine. The equation of hyperplane that separates two classes is given by:

$$D(x) = w_0 + w_1 a_1 + w_2 a_2 \quad (2)$$

However, the equation of the maximum-margin hyperplane can be written

$$x = b + \sum \alpha_i y_i a(i) \times a \quad (3)$$

Here, i is the support vector, and y is the training instance a (1) class value. The learring algorithm determines the numeric value b and a_i, respectively.

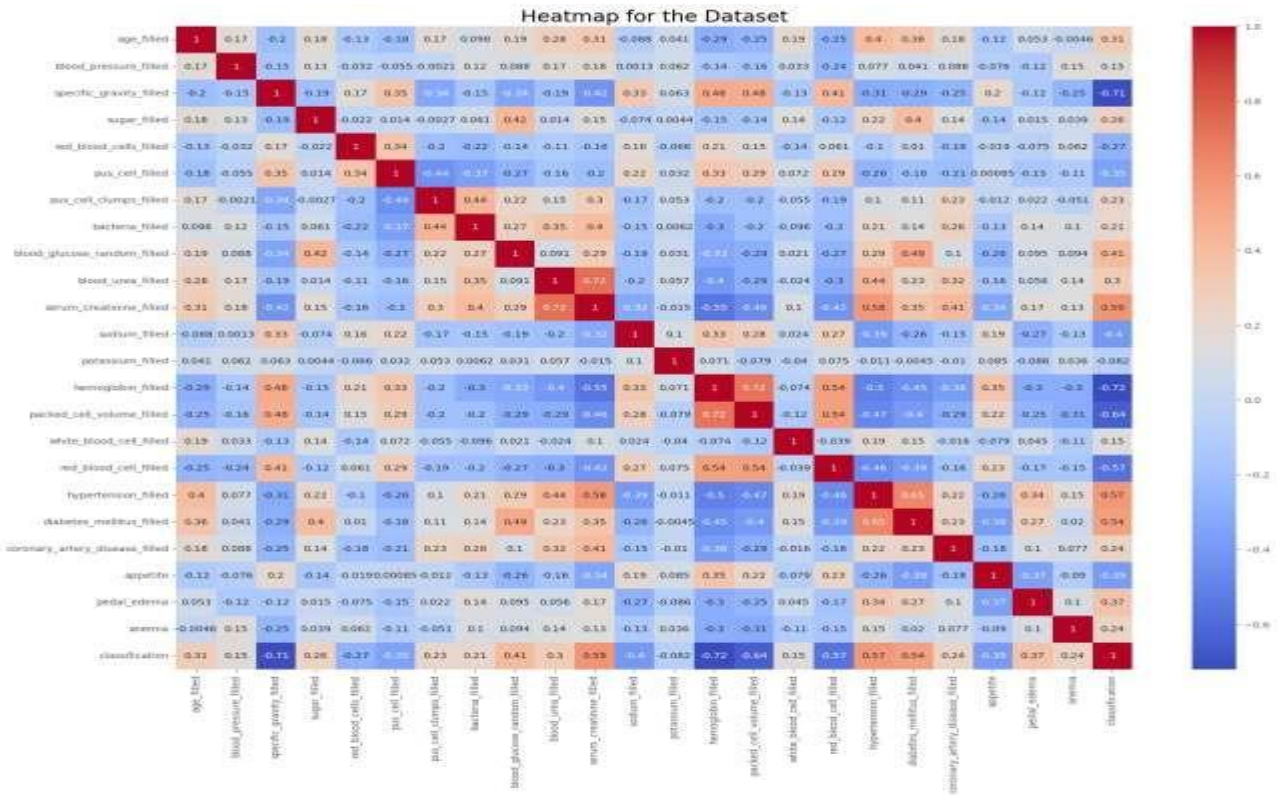

Fig 1: Highly correlated features of CKD dataset

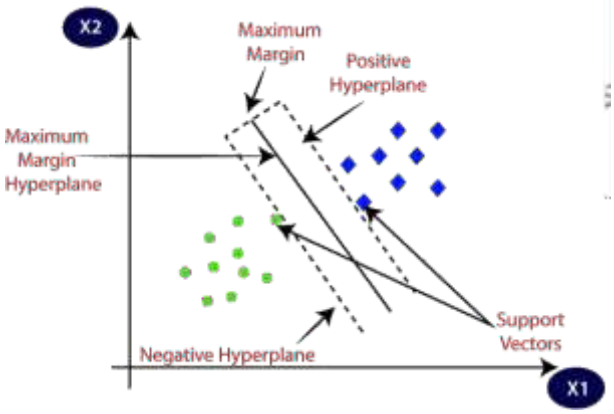

Fig 2: Support Vector Machine

3.3.2. K-Nearest Neighbour

The K-Nearest Neighbors (KNN) algorithm operates by identifying similarities between new and historical data points to classify new test points into existing groups. Unlike parametric methods, KNN doesn't rely on fitting specific parameters but rather learns directly from the training dataset, which can lead to slower learning. The algorithm determines classification by considering the K nearest neighbors, where K represents the number of neighboring points examined. Euclidean distance is commonly used to measure the distance between the new observation and the stored training points. Figure 3 visually represents the K-Nearest Neighbor classification process, showcasing the impact of different K values on classification outcomes. This method's reliance on neighboring data points makes it a versatile approach suitable for various classification tasks.

$$d_{ij} = \sum_{t=1}^{n} (x_{it}^{test} - x_{jt}^{train})^2 \quad (4)$$

KNN algorithm searches t training data set with minimum distance to the testing set

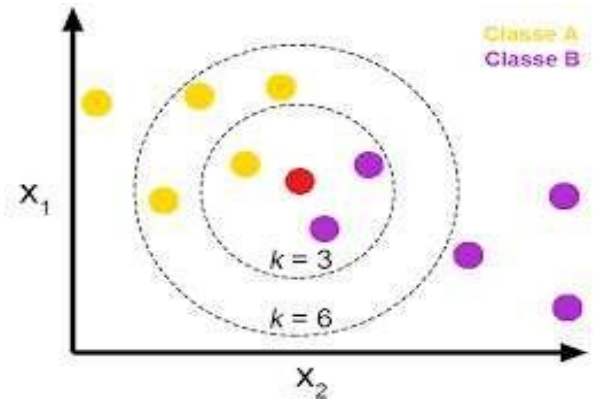

Fig 3: K-Nearest Neighbour

3.3.3. Decision Tree Classifier

Decision trees are a nonparametric method of supervised learning [51]. This is a classified structured tree that defines the characteristics of a data set. It represents internal rules for decision-making through internal nodes and tree branches. It has two types of nodes, the decision, and the leaf nodes. The decision nodes take some decisions, and the outcomes of such decisions are leaf nodes. A decision tree has presented in Figure 4.

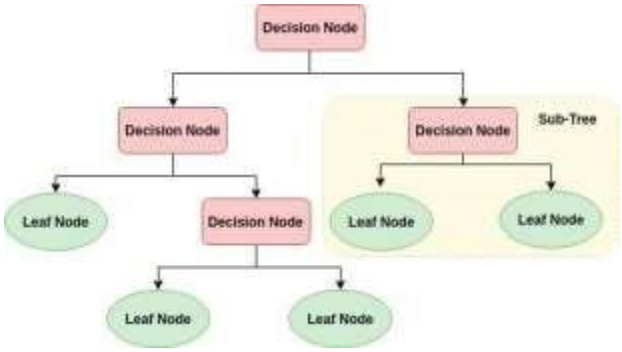

Fig 4: Decision Trees

3.3.4. Random Forest Classifier

The random forest algorithm is based on ensemble learning, improving the model's performance, and solving complex problems by combining several classifiers. A classifier named after the algorithm that contains multiple decision trees averaged over a database subset to improve predictions. In the forecasting process, it does not rely on a single decision tree, and the random forest algorithm creates forecasts from each decision tree that predicts the conclusion based on the majority of decision votes. The usage of several trees decreases the possibility of the model overfitting. To predict the classes in the database, the algorithm includes many decision trees, some of which can predict the proper outcome while others cannot. As a result, there are two assumptions regarding the prediction's accuracy. To forecast a more accurate outcome than an estimate, the algorithm must first includethe actual value of the feature variable. Second, there must be an extremely low correlation between the forecasts for each tree. As a result, there are two requirements for high forecast accuracy. Figure 5 shows a Random Forest Classifier.

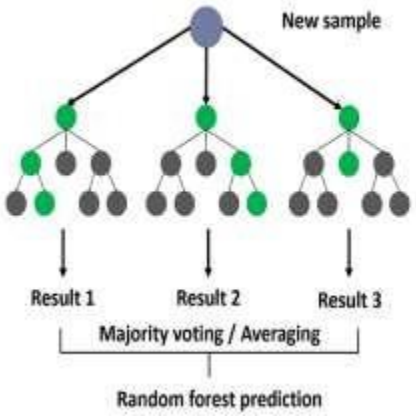

Fig 5: Random Forest

3.4. Model Development

In the model's framework depicted in Figure 6, there are three main phases: preprocessing, model hyper-tuning, and classification. Among these, preprocessing is deemed crucial due to potential noise and redundant values in the dataset. Various methods are employed in this phase, including handling missing data, encoding categorical variables, transforming data, eliminating outliers and extreme values, and selecting relevant features. For instance, in this study using the CKD dataset, only a subset of features is chosen via Recursive Feature Elimination (RFE), a process that assesses the significance of each feature to streamline processing complexity and filter out irrelevant characteristics. After preprocessing, the dataset is split into training and testing sets. The selected features are then utilized to train the learning model. Figure 6 displays the pseudo-code outlining the methodology. Initially, the data undergoes preparation and standardization processes, followed by further processing steps. After standardizing the data, additional processing steps are applied to

refine it further. These steps involve assessing feature importance and filtering out redundant or unrelated characteristics. The Recursive Feature Elimination (RFE) algorithm plays a key role here, as it evaluates the significance of each feature and aids in reducing computational complexity. By selecting only the most relevant features, the model is better equipped to learn and make accurate predictions. Finally, the processed data is ready to be fed into the learning model for training and evaluation.

There are 12 layers in the proposed model architecture: an input layer, five dense layers, five drop layers, and an output dense classifier layer. In Figure 6, the layered architecture's exact specifications are depicted. Each dense layer is connected directly in a feed-forward method in this architecture. The layer is built in such a way that the outputs of its activation maps are handed on to all following levels as input. A dropout layer is placed between two dense layers in this model, with drop rates of 0.5, 0.4, 0.3, 0.2, and 0.1. Figure 6 presents the layered architecture of the proposed model. The CNN model has several hyperparameters that need to be optimized. The optimal hyperparameters selection process is experimental; however, it is time-consuming and difficult. Adam [52,53] optimizer initiates hyperparameters with smaller parameters during the training phase. Adam uses adaptive assessment to determine individual learning rates for various hyperparameter grades ranging from first to second-order gradients. Stochastic Gradient Optimization (SGD) [54] is less efficient than Adam. It necessitates minimal learning time and memory. The classification performance is enhanced by the CNN correct activation function. Neural network's standard activation functions are sigmoid, tan, Rectified Linear Unit (ReLU) [55], Exponential Linear U.

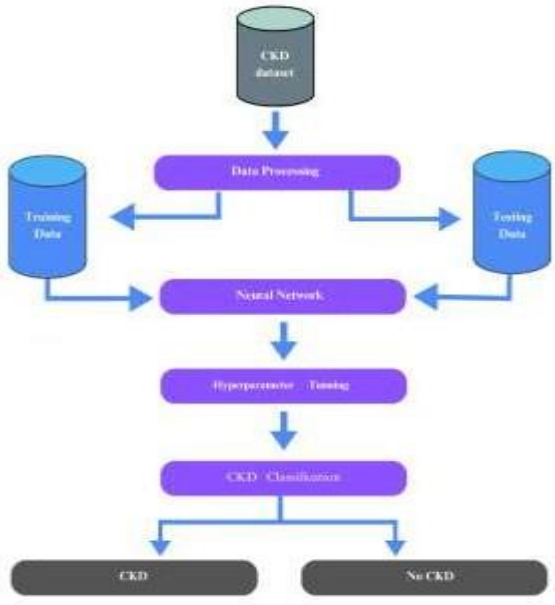

Fig 5: Framework for proposed model

PSEUDO-CODE

BEGIN

1) Importing a CKD Dataset.
2) Fill in the blanks and eliminate outliers.
3) Convert text into numeric values
4) Adjust the data scale
5) Make use of the RFF 'Feature Selection.'
6) FOR FEATURE SELECTION METHODS
7) Choose the most important features
8) Make a list of the most important features
9) Count the number of times each feature appears. END FOR
10) Provide the proposed model with a list of features
11) Set the model hyperparameters
12) Assign different validation scores to the scoring.
13) Train using the proposed model
14) Sort data into CKD and non-CKD categories.
15) Validation score END

4. RESULTS AND DISCUSSION

The proposed model accuracy was calculated by making the CKD class value positive and the not-CKD class value negative. The confusion matrix was utilized to evaluate the performance by using True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) [58]. According to TP, CKD samples have been accurately categorized. The findings of the FN test show that CKD samples were misclassified. The not-CKD samples were not accurately identified, as indicated by a false-positive result (FP). True negative (TN) samples have been accurately categorized as not CKD. The findings of the proposed model are presented in this section. The CKD data sets are split into 75% training and 25% test data sets. The hyperparameter settings for the proposed model. The confusion matrices are shown in Figure 6. It demonstrates that the suggested model correctly identified all genuine positive and true negative events. The CKD class reports recall, precision, sensitivity, F1 score, and accuracy.

The proposed model is compared with other classifier algorithms, including logistic regression, KNN, SVM, Decision tree, and Random forest. No

parameter adjustments were made for these algorithms to show the improved performance of the proposed model. Therefore, the default value for a parameter was used in scikit-learn. All models are evaluated using the F1-score. Table 2 showed experimental results when the proposed model was tested on CKD data sets.

The findings of the proposed model are presented in this section. The CKD data sets are split into 75% training and 25% test data sets. The hyperparameter settings for the proposed model are shown in Table 3. The confusion matrices are shown in Figure 6. It demonstrates that the suggested model correctly identified all genuine positive and true negative events. The CKD class reports recall, precision, sensitivity, F1 score, and accuracy.

4.1. Evaluation Process

In assessing model performance, several metrics are employed, including accuracy score, precision, recall, and F1 score, as outlined in reference [19]. Accuracy measures the proportion of correct predictions made by the model, providing a straightforward indicator of its overall performance. Precision, on the other hand, delves into the quality of positive predictions, reflecting the model's capability to accurately identify relevant instances within the dataset. Meanwhile, recall evaluates the model's sensitivity to detecting relevant data points, indicating its ability to capture all pertinent information. The F1 score

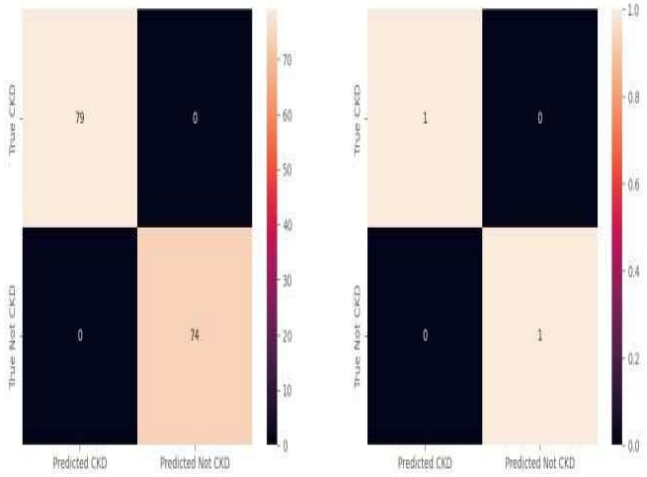

Fig 6: Confusion of Proposed Model

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1 - \text{Score} = 2\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right)$$

The false positive rate (FPR) computes the model's incorrect predictions of positive values contrary to the false negative rate (FNR), determining the incorrect predictions of negative values. The negative predict value (NPV) determines which negative prediction is in fact negative. The false discovery rate (FDR) measures the number of false predictions determined as true.

$$FPR = \frac{FP}{FP + TN}$$

$$FNR = \frac{FN}{FN + TP}$$

$$NPV = \frac{TN}{TN + FN}$$

$$FDR = \frac{FP}{FP + TP}$$

Here, true positive (TP) means that values classified as true in theory are also true in reality. False positive (FP) occurs when false results are incorrectly labelled as true. False negative (FN) denotes a value that is positive but incorrectly recognized as negative. When a value is correct but incorrectly labeled as negative, it is designated as true negative (TN).

S No	Model	Score
1	Random Forest	98
2	Logistic Regression	96.08
3	Support Vector Machines	96
4	Linear SVC	94
5	Naive Bayes	93.46
6	Decision Tree	92

Table 2. Analysis of the proposed model with existing classification techniques on CKD data set.

This section highlights the primary feature identified by the Recursive Feature Elimination (RFE) algorithm, emphasizing their significance in classifying Chronic Kidney Disease (CKD). The figure visually represents these chosen features and their respective importance levels. Noteworthy risk factors include Hemoglobin, Serum Creatinine, Specific Gravity, Packed Cell Volume, Red Blood Cell Count, Hypertension, and Albumin, detailed in Table 3. Nephrologists are urged to prioritize these factors when diagnosing CKD patients. Additionally, Figure 7 provides a clear illustration of the RFE-selected features and their associated importance, aiding in understanding their relevance in CKD classification. In the literature, some hybrid model, cross validation, and feature selection techniques were used, which are missing in our study. A greater memory and longer compilation time are required for such complicated techniques [20]. To reduce compilation time of a model, we avoid a cross-validation technique or a hybrid model.

Risk Factor Name
Hemoglobin
Serum
Creatinine
Red Blood Cell Count
Packed Cell Volume
Albumin
Specific Gravity
Hypertension

Table 3. The most critical risk factors from CKD data

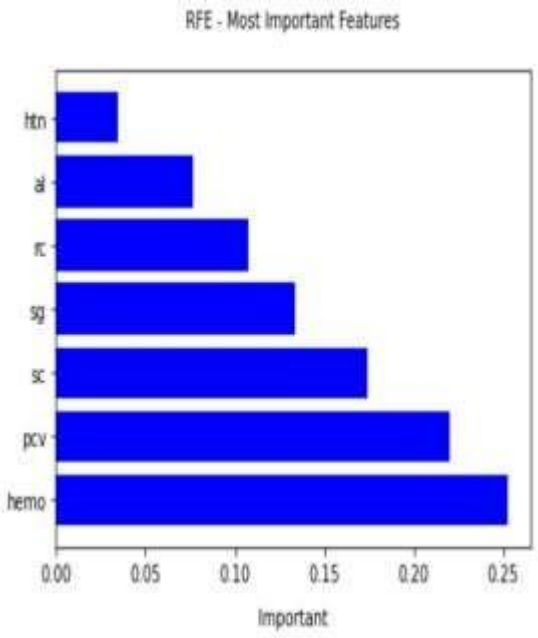

Fig 7: Important features selected by RFE.

Instead of cross-validation techniques, we proposed our model with some balanced dense layers, and kernel regulations are added to some neurons. Then, we disable a few parameters in order to balance our data. Issues with overfitting are handled via kernel regulations and dropouts. Finally, we achieve a balanced performance from all the different optimizers, and our model generates the findings in a very short time.

5. CONCLUSION

A neural-network-based method for detecting chronic kidney disease (CKD) has been successfully developed here. This prediction technique has a high level of accuracy and may be used by physicians as an alternative approach. It can also be used by ordinary people to study, by filling in the most likely values for all missing variables, the raw dataset is transformed into a highly pre- processed dataset. All the learning algorithms, optimized and traditional, were evaluated on the same dataset and on the heavily pre-processed data.

The most essential CKD features are packed red blood cell count, albumin, cell volume, serum creatinine, specific gravity, hemoglobin, and hypertension. Classification algorithms are fed with a set of features. Different metrics, including classification accuracy, recall, precision, and f-measure, are used to estimate the comparative analysis. The proposed deep neural model outperformed the other five classifiers (Support Vector Machine (SVM), Logistic regression, Linear SVC, Random Forest, and Naive Bayes classifier) by achieving 100% accuracy. The accuracy of Linear SVC, SVM, Naïve Bayes, Decision tree, Random Forest, logistic regression is 94%, 96%, 93.46%, 92%, and
96.08%, respectively.

The performance of the proposed model compared with several recent scholarly studies, such as Ant Colony-based Optimization Classifier by Elhoseny et al. [19], Neural network by Vasquez-Morales et al. [27], KNN by M Senan et al. [37], Convolutional Neural Networks by Krishnamurthy et al. [38], SVM by Polat, H. et al. [45], and SAE and Soft max Regression proposed by Sarah A. et al. [46]. The exiting works obtained the accuracy from 85% to 98.5%, while the proposed model has obtained an accuracy of 95%. The proposed approach could be a useful tool for nephrologists in detecting CKD.

The limitation of the proposed model was that it had been tested on small data sets. To improve the model performance, significant volumes of increasingly sophisticated and representative CKD data will be collected in the future to detect disease severity. The clinical data to be collected from pathologist's experts. The performance of the proposed model will be evaluated on a large clinical data set based on acid-base parameters, hyperparathyroidism, inorganic phosphorus concentration, and night urination in the future. Additionally, new features will be applied to get a broader perspective

on the informative parameters related to CKD disease to test the prediction accuracy.

6. REFERENCES

1. 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.
2. Debal, Dibaba Adeba, and Tilahun Melak Sitote. "Chronic kidney disease prediction using machine learning techniques." *Journal of Big Data* 9, no. 1 (2022): 109.
3. Pal, Saurabh. "Chronic kidney disease prediction using machine learning techniques." *Biomedical Materials & Devices* 1, no. 1 (2023): 534-540.
4. Kaur, Chamandeep, M. Sunil Kumar, Afsana Anjum, M. B. Binda, Maheswara Reddy Mallu, and Mohammed Saleh Al Ansari. "Chronic kidney disease prediction using machine learning." *Journal of Advances in Information Technology* 14, no. 2 (2023): 384-391.
5. Farjana, Afia, Fatema Tabassum Liza, Partha Pratim Pandit, Madhab Chandra Das, Mahadi Hasan, Fariha Tabassum, and Md Helal Hossen. "Predicting chronic kidney disease using machine learning algorithms." *In 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 1267-1271. IEEE, 2023.
6. Dritsas, Elias, and Maria Trigka. "Machine learning techniques for chronic kidney disease risk prediction." *Big Data and Cognitive Computing* 6, no. 3 (2022): 98.
7. Saif, Dina, Amany M. Sarhan, and Nada M. Elshennawy. "Deep-kidney: an effective deep learning framework for chronic kidney disease prediction." *Health Information Science and Systems* 12, no. 1 (2023): 3.
8. Bai, Qiong, Chunyan Su, Wen Tang, and Yike Li. "Machine learning to predict end stage kidney disease in chronic kidney disease." *Scientific reports* 12, no. 1 (2022): 8377.
9. Mondol, Chaity, FM Javed Mehedi Shamrat, Md Robiul Hasan, Saidul Alam, Pronab Ghosh, Zarrin Tasnim, Kawsar Ahmed, Francis M. Bui, and Sobhy M. Ibrahim. "Early prediction of chronic kidney disease: A comprehensive performance analysis of deep learning models." *Algorithms* 15, no. 9 (2022): 308.
10. Hassan, Md Mehedi, Md Mahedi Hassan, Swarnali Mollick, Md Asif Rakib Khan, Farhana Yasmin, Anupam Kumar Bairagi, M. Raihan, Shibbir Ahmed Arif, and Amrina Rahman. "A Comparative Study, Prediction and Development of Chronic Kidney Disease Using Machine Learning on Patients Clinical Records." *Human-Centric Intelligent Systems* 3, no. 2 (2023): 92-104.
11. Swain, Debabrata, Utsav Mehta, Ayush Bhatt, Hardeep Patel, Kevin Patel, Devanshu Mehta, Biswaranjan Acharya, Vassilis C. Gerogiannis, Andreas Kanavos, and Stella Manika. "A robust chronic kidney disease classifier using machine learning." *Electronics* 12, no. 1 (2023): 212.
12. Aswathy, R. H., P. Suresh, Mohamed Yacin Sikkandar, S. Abdel-Khalek, Hesham Alhumyani, Rashid A. Saeed, and Romany F. Mansour. "Optimized tuned deep learning model for chronic kidney disease classification." *Comput. Mater. Contin* 70 (2022): 2097-2111.
13. Sawhney, Rahul, Aabha Malik, Shilpi Sharma, and Vipul Narayan. "A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease." *Decision Analytics Journal* 6 (2023): 100169.
14. Alsekait, Deema Mohammed, Hager Saleh, Lubna Abdelkareim Gabralla, Khaled Alnowaiser, Shaker El-Sappagh, Radhya Sahal, and Nora El-Rashidy. "Toward Comprehensive Chronic Kidney Disease Prediction Based on Ensemble Deep Learning Models." *Applied Sciences* 13, no. 6 (2023): 3937.
15. Baidya, Deepanita, Umme Umaima, Md Nazrul Islam, FM Javed Mehedi Shamrat, Anik Pramanik, and Md Sadekur Rahman. "A deep prediction of chronic kidney disease by employing machine learning method." *In 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 1305-1310. IEEE, 2022.
16. Sanmarchi, Francesco, Claudio Fanconi, Davide Golinelli, Davide Gori, Tina Hernandez-Boussard, and Angelo Capodici. "Predict, diagnose, and treat chronic kidney disease with machine learning: a systematic literature review." *Journal of nephrology* 36, no. 4

(2023): 1101-1117.

17. Ebiaredoh-Mienye, Sarah A., Theo G. Swart, Ebenezer Esenogho, and Ibomoiye Domor Mienye. "A machine learning method with filter-based feature selection for improved prediction of chronic kidney disease." *Bioengineering* 9, no. 8 (2022): 350.
18. Holmstrom, Lauri, Matthew Christensen, Neal Yuan,
J. Weston Hughes, John Theurer, Melvin Jujjavarapu, Pedram Fatehi et al. "Deep learning-based electrocardiographic screening for chronic kidney disease." *Communications Medicine* 3, no. 1 (2023):
73.
19. Iftikhar, Hasnain, Murad Khan, Zardad Khan, Faridoon Khan, Huda M. Alshanbari, and Zubair Ahmad. "A comparative analysis of machine learning models: a case study in predicting chronic kidney disease." *Sustainability* 15, no. 3 (2023): 2754.
20. Schena, Francesco Paolo, Vito Walter Anelli, Daniela Isabel Abbrescia, and Tommaso Di Noia. "Prediction of chronic kidney disease and its progression by artificial intelligence algorithms." *Journal of Nephrology* 35, no. 8 (2022): 1953-1971.