A Comprehensive Survey on Detecting Chronic Kidney Disease by using a Voting Classifier

Aradala Vamsi Krishna[1], Ganta Srinivasa rao[2], Podapati Naveen[3], Deepali Bhagat[4]
[1,2,3] *Computer science engineering, Lovely professional University*
[4]*Asst Professor, Computer science engineering, Lovely professional University*

***Abstract*— chronic kidney disease (CKD) is a global health issue that causes a high incidence of morbidity and death, as well as the onset of additional illnesses. Because there are no clear symptoms in the early stages of CKD, people frequently miss it. Early identification of CKD allows patients to obtain prompt therapy to slow the disease's development. Due of their rapid and precise identification capabilities, machine learning models can successfully assist doctors in achieving this aim. We present a machine learning framework for diagnosing CKD in this paper. The CKD data set was taken from the machine learning repository at the University of California, Irvine (UCI), which has a substantial number of missing values. To fill up the missing values, KNN imputation was employed, which picks many full samples with the most comparable measurements and processes the missing data for each incomplete sample. For a variety of reasons, patients may miss some measurements, resulting in missing statistics in real-world medical settings. Six machine learning approaches (logistic regression, random forest, support vector machine, k-nearest neighbour, naive Bayes classifier, and feed forward neural network) were employed to create models once the missing data set was filled in appropriately. Random forest outperformed the other machine learning models, with a diagnostic accuracy of 99.75 percent. After examining the misjudgments produced by the current models, we proposed an integrated model that incorporates logistic regression, random forest, and perceptron, which could achieve an average accuracy of 99.83 percent after ten simulations. As a result, we hypothesized that this technology may be used to diagnose diseases using more complex clinical data.**

***Index Terms:* chronic kidney disease (CKD), Data mining, Bagging, Random subspace Naive Bayes, KNN, Decision tree, Accuracy, ROC, Kappa.**

I.INTRODUCTION

Chronic Kidney Disease (CKD) is a worldwide clinical concern. The most frequent kind of kidney illness is chronic kidney disease (CKD). Three-month incurable inflammation and degenerative changes in renal tissue are symptoms of the condition. In the early stages of CKD, patients can be readily treated. Furthermore, in individuals with mild renal impairment, the kidneys can meet the body's demands and there are no clinical or biochemical abnormalities associated with kidney disease.

The majority of CKDs, on the other hand, worsen over time, and the number of nephrons diminishes. CKD generates biochemical and clinical signs after a period of time. With these poor kidney functions, the patient can survive for a long period if the condition is detected early on and the appropriate therapies are started. However, it is common to see that the patient's health deteriorates with time, and the number of nephrons falls rapidly. As a result, increased hazardous materials in the blood endangers the patient's life. The final stage of renal failure is referred to as terminal kidney failure.

Chronic kidney disease (CKD) is a worldwide public health issue that affects around 10% of the world's population.

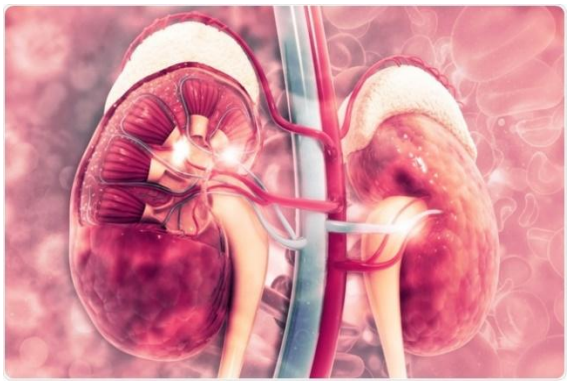

Figure 1: Chronic Kidney Disease Diagnosis

CKD, which eventually leads to renal failure, can create a variety of issues in the society, including economic, social, and medical issues. The high cost

of hemodialysis might be used as an illustration of economic issues. Patients have emotional and social issues as a result of their continued use of medicines and consequences. Furthermore, the medications utilized can interact with many organs and systems, resulting in a considerable reduction in lifespan.

In China, 10.8% of people have chronic kidney disease, whereas in the United States, the frequency ranges from 10% to 15%. According to another survey, this ratio has risen to 14.7 percent of the adult population in Mexico. This condition is marked by a gradual decrease of renal function that finally leads to total renal failure. In the early stages of CKD, there are no visible symptoms. As a result, the illness may not be recognized until the kidneys have lost around 25% of their function. CKD also has a significant morbidity and death rate, as well as a worldwide influence on the human body. It has the potential to cause cardiovascular disease.

In the literature, several classification and feature selection approaches for detecting CKDs are discussed. The most common classifiers include k-NN, SVM, decision trees (J48, BF Tree, REP Tree), Gauss algorithm-based classifiers, and Nave Bayes classifiers. Baby et al., for example, found that the K Star, Random Forest, and J48 algorithms are the most accurate in terms of classification accuracy. The first two algorithms' accuracy values are calculated to be 100 percent, whereas the third algorithm's accuracy value is calculated to be 98.55 percent. Kaladharetal studied the ROC curve using J48, Random Forest, and SVM algorithms in their kidney stone investigations. For the J48 and Random Forest methods, and the SVM algorithm, they attained accuracy values of 93 percent and 91.98 percent, respectively.

Bagging, adaboost, and random subspace are among the individual and ensemble classifiers utilized in this study to diagnose CKD.In ensemble learning techniques, support vector machines and decision tree-based classifiers like J48, REPTree, and BFTree are used as foundation learners. In addition, the information gain ratio attribute evaluator feature selection approach is used to find the optimal attributes.

II.REVIEW OF LITERATURE

In this [1] the authors JIONGMING QIN, LIN CHEN, YUHUA LIU, CHUANJUN LIU, CHANGHAO FENG and BIN CHEN proposed a technique, chronic kidney disease (CKD) is a global health issue that causes a high incidence of morbidity and death, as well as the onset of additional illnesses. Because there are no clear symptoms in the early stages of CKD, people frequently miss it. Early identification of CKD allows patients to obtain prompt therapy to slow the disease's development. Due of their rapid and precise identification capabilities, machine learning models can successfully assist doctors in achieving this aim. We present a machine learning framework for diagnosing CKD in this paper. The CKD data set was taken from the machine learning repository at the University of California, Irvine (UCI), which has a substantial number of missing values. The planned work's contributions are noted below.

1) To fill in the missing values in the data set, we employed KNN imputation, which could be used to data sets when the diagnostic categories are unknown.

2) Using logistic regression (LOG), RF, SVM, KNN, naive Bayes classifier (NB), and feed forward neural network, the whole CKD data set was used to develop CKD diagnostic models (FNN).Misjudgment analysis was performed on the models that performed better.

3) A perceptron-based integrated model that integrates LOG and RF was developed, and it enhanced the component models' performance in CKD diagnosis after missing values were filled using KNN imputation.

The following machine learning models for diagnosing CKD were created by applying the relevant subset of characteristics or predictions on the entire CKD data sets.

1) Regression-based model: LOG
2) Tree-based model: RF
3) Decision plane-based model: SVM
4) Distance-based model: KNN
5) Probability-based model: NB
6) Neural network: FNN

In terms of data imputation and sample diagnosis, the suggested CKD diagnostic approach is practical. After unsupervised KNN imputation of missing values in the data set, the integrated model was able

to achieve sufficient accuracy. As a result, we believe that adopting this technique to the practical diagnosis of CKD would have a positive outcome. Furthermore, this technology might be used to clinical data from various disorders in real-world medical diagnosis. As a result, the model's generalization performance may be restricted. Furthermore, because the data set contains only two types of data samples (ckd and notckd), the model is unable to determine the severity of CKD.

A significant amount of more complicated and representative data will be collected in the future to train the model to increase generalization performance while also allowing it to recognize illness severity. We anticipate that as the amount and quality of the data grows, our model will become even more perfect.

In this [2] the authors MerveDogruyolBasar, Aydin Akan proposed a technique chronic kidney disease is a severe health issue that affects millions of individuals worldwide and creates serious economic, social, and medical issues. Several automated diagnosis techniques can detect chronic renal disease. The Adaboost, Bagging, and Random Subspaces ensemble learning algorithms are used in this study to identify chronic kidney disease. In the decision step, decision tree-based classifiers are utilized. The accuracy and kappa criteria are used to evaluate the classification performance. According on the results of the suggested systems' performance studies, ensemble learning classifiers outperform individual classifiers in classification.In the literature, several classification and feature selection approaches for detecting CKDs are discussed. k-NN, SVM, decision trees (J48, BFTree, REPTree), Gauss algorithm-based classifiers, and Nave Bayes classifiers are all commonly employed. Baby et al., for example, found that the K Star, Random Forest, and J48 algorithms are the most accurate in terms of classification accuracy. The first two algorithms' accuracy values are calculated to be 100 percent, whereas the third algorithm's accuracy value is calculated to be 98.55 percent. Kaladharetal studied the ROC curve using J48, Random Forest, and SVM algorithms in their kidney stone investigations. For the J48 and Random Forest methods, and the SVM algorithm, they attained accuracy values of 93 percent and 91.98 percent, respectively.

Patient data on chronic renal disease was used in this investigation. The "UC Irvine Machine Learning Repository “database is used to gather data that have or do not contain illness symptoms. This database considers 24 characteristics (11 numerical and 13 nominal) of 400 persons (150 of whom are healthy), which are thought to be the primary contributors to the condition, as well as comparisons and classifications with kidney tissue samples. The database provides classification information that may be used to identify whether or not a person has renal disease.

In the second phase, the most suitable 10 features in terms of data processing are extracted by using Gain Ratio Attribute Evaluator Algorithm and the new data sets are processed. Parameters of the CKD database are given in Table I. The existence and/or the values of these parameters are closely related to the kidney disease formation [14].

“Waikato Environment for Knowledge Analysis (WEKA)” was used to analyze and compare the data in the second step, the Gain Ratio Attribute Evaluator Algorithm is used to find the most acceptable 10 data processing characteristics, and the new data sets are processed. The parameters of the CKD database are listed in Table I. These markers' existence and/or amounts have been associated to the development of renal illness. The "Waikato Environment for Knowledge Analysis (WEKA)" tool was used to evaluate and compare the data.Another goal of this research is to see how the in-formation gain attribute evaluator feature selection approach affects the diagnosis of chronic diseases. Thus, using the information gain attributeevaluator approach, 10 best features with the best information values are picked, and four unique classifiers are applied to these 10 features. With the BFTree classifier, the best accuracy is determined to be 99.25 percent. Utilizing the RSM-BFTree combination, however, 100 percent classification accuracy is attained using ensemble classifiers. Both the ensemble learning algorithm and the suggested feature selection approach are useful tools for classifying chronic renal disease, according to these findings.

In this [3] the authors Mohamed Elhoseny, K. Shankar & J. Uthayakumar proposed a technique In today's world, new capabilities like as machine learning (ML), data mining, and artificial intelligence are being added to healthcare systems in order to

provide humans with more intelligent and expert healthcare services. This study for chronic kidney illness introduces the Density based Feature Selection (DFS) with Ant Colony based Optimization (D-ACO) approach, which is an intelligent prediction and classification system for healthcare (CKD).Prior to the ACO-based classifier development, the suggested intelligent system uses DFS to reduce unnecessary or duplicate characteristics. Various strategies for accurately predicting CKD using medical data from patients have been offered. For the identification of CKD17, a Cuckoo Search trained neural network (NN-CS) technique is given. Initially, the proposed methodology is intended to address problems with local search-based learning methods. The CS method aids in the best selection of the NN's input weight vector in order to correctly train data. The suggested algorithm's classifier results revealed that it achieves superior performance. To tackle the problem of local optima in the NN-CS method, a modified version of the algorithm (NN-MCS)18 is devised.The FS process plays a significant role in data classification, as it is used to extract a smaller set of rules from a training dataset with defined goals. For FS, many strategies like as AI techniques and bio-inspired algorithms are applied. In22, the hybridization of GA with support vector machine (SVM) is offered as a wrapper technique named GA-SVM method to appropriately pick the feature subset. The suggested method's elimination of duplicate characteristics enhances classification performance, which is confirmed using four different illness datasets.The DFS with ACO method, commonly known as the D-ACO algorithm, is proposed for the classification of the CKD dataset in this research.The suggested D-ACO framework, on the other hand, combines FS and ACO-based learning and eliminates extraneous characteristics. The D-ACO algorithm's efficiency is assessed using a benchmark CKD dataset, and a comparison with existing approaches is done. When compared to current approaches, the suggested D-ACO algorithm exceeded them in terms of classification performance in a variety of ways. Overall, the suggested D-ACO method is found to be a suitable classifier for CKD identification.

In this [4] the authors Nikhila G, Meghashree A.C proposed a technique chronic kidney disease (CKD) is a worldwide problem that affects around 10% of the world's adult population. Early detection of CKD is difficult for the majority of persons. As a result, it's critical to employ today's computer-assisted supported tactics to enable the traditional CKD detection framework become more successful and exact. Six current machine learning approaches were utilised on the chronic renal disease dataset from the UCI Repository: Multilayer Perceptron Neural Network, Support Vector Machine, Nave Bayes, K-Nearest Neighbor, Decision Tree, and Logistic regression. Ensemble Algorithms such as ADABoost, Gradient Boosting, Random Forest, Majority Voting, Bagging, and Weighted Average were employed to increase the model's performance. The model was fine-tuned to get the optimal hyper parameters for training. Accuracy, Precision, Recall, F1-score, Mathew's Correlation Coefficient, and the ROC-AUC curve were used to evaluate the model's performance. The experiment was carried out on individual classifiers first, and subsequently on Ensemble classifiers. Ensemble classifiers like Random Forest and ADABoost outperformed individual classifiers with 100% accuracy, precision, and recall when compared to Decision Tree Algorithm-generated classifiers with 99.16 percent accuracy, 98.8 percent precision, and 100 percent recall.A few drugs, notably certain pain relievers (analgesics), might induce CKD if used over an extended period of time. Often, experts are stumped as to what caused the problem. There are usually no signs or symptoms in the beginning, but later on, side effects such as swelling ankles, weariness, regurgitating, lack of desire, or disarray may occur. The therapy for CKD is primarily focused on slowing the progression of kidney damage, which is usually accomplished by addressing the root causes. Early detection and treatment can prevent CKD from worsening. When the findings reach the final stage, regular dialysis or a kidney transplant are required to maintain a normal life. On the chronic kidney disease dataset, the experiment was carried out. Support Vector Machine (SVM), Logistic Regression (LR), Nave Bayes (NB), K-Nearest Neighbor (K-NN), Decision Tree (DT), and Multilayer Perceptron were used to evaluate the dataset (MLP). The dataset was then put through its paces with six ensemble classifiers: Majority Voting, Weighted Average, Bagging, ADABoost, Gradient Boosting, and Random Forest. Accuracy, precision, recall, Area Under Curve (AUC), and Mathew's Correlation Coefficient were used to assess

performance in both scenarios. Decision tree fared the best in the base classifier, with an accuracy of 99.16 percent, precision of 98.8%, recall of 100 percent, MCC of 98.07 percent, and AUC of 98.6 percent. ADABoost and Random Forest were the best performers in the Ensemble classifier, with accuracy, precision, recall, and precision.

In this [5] the author Khaled Mohamad Almustafaproposed a technique chronic kidney disease (CKD) is a prevalent kidney function issue that causes renal perfusion to deteriorate. renal failure as a result of poor performance. A way for determining kidney functioning that may be used early in the diagnostic process is in many circumstances, necessary and exceedingly significant. Different classifiers were used in this study to classify the data. fiction of a chronic kidney disease dataset Random tree, decision table (DT), and K-nearest neighbor algorithms were used. A prediction model was created using K-NN, J48, SGD, and Nave Bayes classifiers, as well as a prediction model. A methodology based on feature selection has been proposed for accurately predicting CKD patients. The J48 was proven to be effective, as was the decision. With 99 percent accuracy, 0.999 and 0.992 ROCs, and 0.999 and 0.992 MAEs,table classifiers beat the other classifiers. of 0.0225 and 0.1815, respectively, with RMSEs of 0.0807 and 0.2507.A sensitivity study was done on selected classifiers to examine how well they performed when their parameters were altered. With accuracy of 99 percent and RMSEs of 0.0807 and 0.2507, respectively, the J48 and decision table classifiers surpassed all other classifiers. Furthermore, K-NN (K = 1) was shown to improve classification performance.When feature selection methods were used, Nave Bayes and decision table classification improved to 99.75 percent, 98.25 percent, and 99.25 percent, respectively, and only a few features were used for classification of the CKD dataset, in which such an enhancement can add value and support healthcare provided to identify certain CKD cases at an early stage using the presented selected features, in which such an enhancement can add value and support healthcare provided to identify certain CKD cases at an early stage using the presented selected features, in. A prediction model, such as the one provided in this research, that monitors only specific aspects of the needed tests for the patients would be a fantastic and efficient addition to the already existing diagnostic prediction tools. Many academics are interested in studying CKD datasets in order to build feasible and efficient models to aid healthcare practitioners in the monitoring of probable CKD patients. Machine learning (ML) algorithms have been widely used in various healthcare-related conditions, resulting in a variety of trustworthy and productive models to assist healthcare professionals. In this study, a variety of machine learning (ML) classifiers were employed to classify a CKD dataset, including naïve Bays, stochastic gradient descent (SGD), random tree, J48, K-nearest neighbor (K-NN), and decision table. Classification subset evaluators were used to develop prediction models for the CKD dataset and to assess classifier performance after applying feature selection procedures for different classification algorithms.

In this [6] the authors Olayinka Ayodele Jongbo, Adebayo Olusola Adetunmbi, Roseline BosedeOgunrinde, Bukola Badeji-Ajisafeproposed a techniquechronic kidney disease is a serious worldwide health concern that kills millions of people each year as a result of poor lifestyle choices and inherited factors. The adoption of data mining techniques was prompted by the requirement for quick and precise diagnosis. In recent years, data mining approaches have been extensively researched in the diagnosis of chronic renal disease, with a focus on accuracy, either through the simplicity of the illness by feature selection in addition to preprocessing or not before classification. For increasing the classification performance of the models, this study uses two ensemble approaches: Bagging and Random Subspace techniques on three base-learners: k Nearest Neighbors, Nave Bayes, and Decision Tree. Data preprocessing and data scaling were performed before to classification to manage missing values and normalize the range of independent variables. The accuracy, specificity, sensitivity, kappa, and ROC criteria were used to assess the model's performance. In majority of the examined situations, empirical findings on the chronic kidney disease dataset available from the UCI machine learning repository reveal that assemble approaches outperform individual base learner performances, with random subspace outperforming bagging. Using a random subspace ensemble on a KNN classifier, 100 percent prediction accuracy is achieved. As a result, the model is appropriate for

accurate detection of chronic renal disease. The global frequency of chronic kidney disease is alarming, and it is considered a serious hazard to human life. Three base learners and two ensemble strategies are presented in this study to increase classification accuracy for efficient detection of chronic renal disease. Because human lives are at stake, the accuracy of such a model is critical.

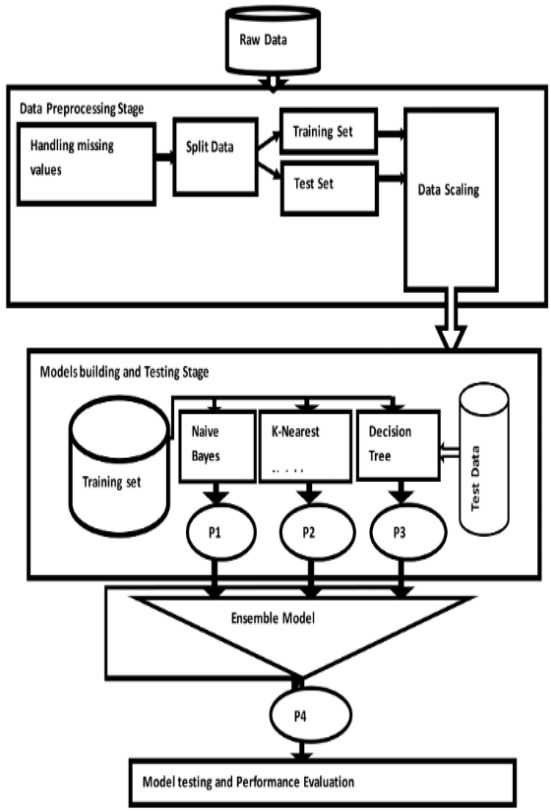

Figure 2: Architecture of the proposed system

In this study, One of the measures by which the classification accuracy of these computer aided diagnostic models could be improved is the use of an ensemble learning approach to aggregate the output of various data mining models (Nave Bayes, KNN, and Decision tree classifier) over a chronic kidney disease dataset. The suggested approach was implemented using the Python programming language. The use of an ensemble learning approach to aggregate the output of various data mining models (Nave Bayes, KNN, and Decision tree classifier) over a chronic kidney disease dataset is demonstrated in this study as one of the ways to improve the classification accuracy of these computer aided diagnostic models. The suggested approach was implemented using the Python programming language.The suggested model's performance was assessed, and comparisons were performed between individual learner and ensemble models. Individual classification algorithms were outperformed by the developed ensemble approach to chronic kidney disease diagnosis, with the best classification accuracy of 1.00 accuracy, 1.00 sensitivity, 1.00 specificity, 1.00 kappa value, and 1.00 ROC value obtained from Random subspace with KNN classifier.The study's findings demonstrated that the ensemble technique predicts better than individual base classifiers, making it acceptable for chronic kidney disease prediction.

In this [7] the authors Ebrahime Mohammed Senan, Fawaz WaselallahAlsaade, MoslehHmoud Al-Adhaileh, Theyazn H. H. Aldhyani, Ahmed Abdullah Alqarni, Nizar Alsharif, Ahmed H. Alahmadi, Mukti E Jadhav, M. Irfan Uddin and Mohammed Y. Alzahraniproposed a techniquechronic kidney disease (CKD) is one of the top 20 causes of mortality in the world, affecting around 10% of the adult population. CKD is a condition in which normal kidney function is disrupted. Effective prediction tools for the early detection of CKD are essential due to the growing number of patients with the disease. The study's originality is in the development of a diagnostic method for chronic renal disease detection. This research aids specialists in the investigation of CKD prevention approaches through early detection utilizing machine learning techniques. The study looked at a dataset with 24 attributes that was obtained from 400 patients. The missing numerical and nominal data were replaced using the mean and mode statistical analysis methods. Recursive Feature Elimination was used to choose the most important properties (RFE).The four classification algorithms utilized in this study were support vector machine (SVM), k-nearest neighbors (KNN), decision tree, and random forest. All of the categorization techniques did a fantastic job. The random forest technique outperformed all other available algorithms on every criterion, obtaining 100 percent accuracy, precision, recall, and F1-score. CKD is a life-threatening disease with high rates of morbidity and mortality. As a result, artificial intelligence technologies for early detection of CKD are crucial.

These procedures aid specialists and doctors in making early diagnoses in order to avert renal failure.

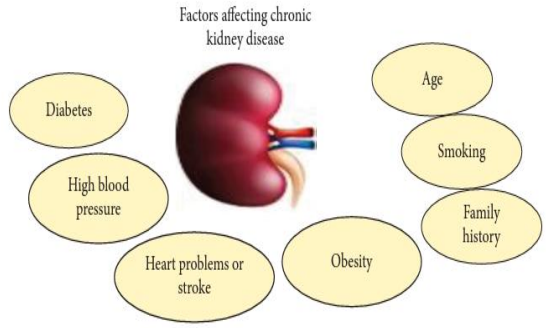

Figure 3: Factors affecting chronic kidney disease.

This study shed light on how to diagnose CKD patients so that they can address their illness and obtain therapy at an early stage. A total of 400 patients were used to create a dataset with 24 characteristics. The dataset was split into two parts: training and testing and validation. Using mean and mode statistical measures, the dataset was processed to eliminate outliers and replace missing numerical and nominal values. The most strongly representative characteristics of CKD were chosen using the RFE method. SVM, KNN, decision tree, and random forest were used to feed selected characteristics into classification algorithms. All classifiers' parameters were fine-tuned to get the best classification results, and all methods produced promising results. For all metrics, the random forest approach surpassed all other algorithms, obtaining 100% accuracy, precision, recall, and F1-score. With respect to the accuracy measure, the empirical findings of SVM, KNN, and decision tree algorithms discovered significant values of 96.67 percent, 98.33 percent, and 99.17 percent, respectively.

In this [8] the authors Hamida Ilyas, Sajid Ali, MahvishPonum, Osman Hasan, Muhammad Tahir Mahmood,Mehwish Iftikhar, and Mubasher Hussain Malikproposed a techniqueChronic Kidney Condition (CKD) is a life-threatening disease characterized by a slow decline in renal function over months to years with no obvious symptoms. According to the severity degree, it advances through six phases. It is divided into phases depending on the Glomerular Filtration Rate (GFR), which is calculated using multiple factors such as age, gender, race, and serum creatinine. The Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI), a linear model, has been proven to be highly efficient among the several available models for predicting GFR value since it permits detection of all CKD stages.Early identification and treatment of CKD is very desired since it can lead to the avoidance of negative outcomes. Recently, machine learning technologies have been widely recommended for early identification of symptoms and diagnosis of a variety of illnesses. With the same motive, the goal of this study is to use machine learning classification algorithms on a dataset derived from afflicted people's medical records to predict the various phases of CKD. The Random Forest and J48 algorithms were used to develop a model that is both sustainable and practical for diagnosing various stages of CKD with high medical accuracy.We developed and compared two algorithms, J48 and random forest, to predict the various phases of CKD in this study. J48 has an 85.5 percent properly categorised instance ratio, whereas Random Forest has a 78.25 percent correctly classified instance ratio. J48, on the other hand, takes 0.03 seconds, whereas Random Forest takes 0.28 seconds. As a consequence, J48 may be stated to be accurate and efficient in terms of execution time, because when compared to Random Forest, it produces more accurate results in less time.

J48 beats Random Forest because it takes into account both categorical and continuous data, whereas Random Forest favours categorical values. Random forest creates many decision trees and then combines them to create a reliable prediction model. However, because of this method, the algorithm is sluggish and useless for real-time prediction. J48 is simple to construct, while Random Forest is difficult because to the vast number of trees. As a consequence of our findings, we propose that clinicians use j48 to assist them in developing an automated decision support system for diagnosing CKD.

In this [9] the authors Marwa Almasoud, Tomas E Wardproposed a techniqueBecause of its rising incidence, chronic kidney disease (CKD) is one of the most serious health issues. In this study, we use the lowest subset of characteristics to investigate the efficacy of machine learning algorithms to predict chronic kidney disease. Several statistical tests, such as the ANOVA test, Pearson's correlation, and Cramer's V test, were used to eliminate duplicate

characteristics. The 10-fold cross-validation method was used to train and evaluate the algorithms of logistic regression, support vector machines, random forest, and gradient boosting. According to the F1-measure from the Gradient Boosting classifier, we attain an accuracy of 99.1.In addition, we discovered that haemoglobin is more important in diagnosing CKD in both random forest and gradient boosting. Finally, when compared to past research, our results are among the best, despite the fact that we have only achieved a small number of characteristics. As a result, we can identify CKD for about $26.65 using three basic tests.Chronic kidney disease is a major public health issue across the world, particularly in low- and middle-income nations. Chronic renal disease occurs when the kidneys do not function properly and are unable to filter blood properly. Around 10% of the world's population has (CKD), and millions die each year due to a lack of inexpensive treatment options, with the number of older people growing. Chronic kidney disease (CKD) has risen as a major cause of mortality globally, according to the International Society of Nephrology's Global Burden Disease 2010 research, with the number of fatalities growing by 82.3 percent in the previous two decades. Furthermore, the number of individuals who develop end-stage renal disease (ESRD) is rising, necessitating kidney transplantation or dialysis to save their lives.This study examines the ability of machine learning algorithms to detect CKD with the fewest number of tests or characteristics feasible. On a small dataset of 400 records, we use four machine learning classifiers to achieve this goal: logistic regression, SVM, random forest, and gradient boosting.

In this [10] the authors Vijendra Singh, Vijayan K. Asari and Rajkumar Rajasekaran, proposed a techniqueDiabetes and high blood pressure are the leading causes of chronic renal disease (CKD).

Researchers all around the globe utilise the Glomerular Filtration Rate (GFR) and kidney damage indicators to diagnose CKD as a disorder that causes declining renal function over time. Chronic kidney disease (CKD) patients have a higher chance of dying young. Doctors face a difficult task in detecting the various diseases connected with CKD at an early stage in order to avoid the illness.This study provides a unique deep learning model for detecting and predicting CKD in its early stages. The goal of this study is to build a deep neural network and compare it to the performance of other modern machine learning approaches.

In testing, the database's missing values were replaced with the average of the corresponding characteristics. After that, the ideal parameters of the neural network were determined by setting the parameters and executing several trials. Recursive Feature Elimination was used to choose the most significant features (RFE). In the RFE, significant characteristics were haemoglobin, specific gravity, serum creatinine, red blood cell count, albumin, packed cell volume, and hypertension. For categorization, selected characteristics were fed into machine learning models.By obtaining 100 percent accuracy, the suggested Deep neural model surpassed the other four classifiers (Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic regression, Random Forest, and Naive Bayes classifier). For nephrologists, the suggested method might be a valuable tool for identifying CKD.This paper presents a deep learning model for early chronic illness diagnosis. The authors of this study used the Recursive Feature Elimination method to determine which traits are most significant for prediction. Packed red blood cell count, albumin, cell volume, serum creatinine, specific gravity, hemoglobin, and hypertension are the most important CKD characteristics. A collection of features is supplied into classification algorithms. The comparison analysis is estimated using a variety of measures, including classification accuracy, recall, precision, and f-measure.

The suggested model's flaw was that it had only been evaluated on tiny data sets.In the future, large amounts of more complex and representative CKD data will be collected to improve the model's effectiveness in detecting disease severity. Clinical information will be gathered from pathologists. In the future, the suggested model's performance will be assessed using a large clinical data set that includes acid-base parameters, hyperparathyroidism, inorganic phosphorus concentration, and nocturnal urination.

III.PROPOSED WORK

Chronic kidney disease (CKD) is a global public health issue that affects around 10% of the global population. In China, 10.8% of people have chronic

kidney disease, while in the United States, the prevalence ranges from 10% to 15%. This figure has already reached 14.7 percent in the Mexican adult general population, according to another study. This condition is marked by a gradual decrease of renal function that finally leads to total renal failure. Currently, the health-care industry offers a number of advantages, including fraud detection in health insurance, the availability of medical facilities to patients at low costs, the identification of smarter treatment methodologies, the development of effective healthcare policies, effective hospital resource management, improved customer relations, improved patient care, and hospital infection control.

Several strategies for predicting CKD using medical data from patients have been proposed. For the detection of CKD17, a Cuckoo Search trained neural network (NN-CS) technique is given. Initially, the proposed methodology is intended to address problems with local search-based learning methods. The CS method aids in the best selection of the NN's input weight vector in order to correctly train data. The suggested algorithm's classifier results revealed that it achieves superior performance. To tackle the problem of local optima in the NN-CS method, a modified version of the algorithm (NN-MCS)18 is devised.

The FS process plays an important role in data classification, as it is used to extract a smaller set of rules from a training dataset with specific purposes. For FS, many strategies like as AI techniques and bio-inspired algorithms are applied. In22, the hybridization of GA with support vector machine (SVM) is offered as a wrapper technique named GA-SVM method to appropriately pick the feature subset. The suggested strategy enhances classification performance by reducing duplicate features, which is proven using five distinct illness datasets.

IV.METHODOLOGY

A prediction model like the one provided in this study, which simply monitors specific aspects of the needed tests for the patients, would be a fantastic and efficient complement to the already existing prediction tools for such a diagnosis. Many academics are interested in studying CKD datasets in order to build feasible and efficient models to aid healthcare practitioners in the monitoring of probable CKD patients. Machine learning (ML) algorithms have been widely used in various healthcare-related conditions, resulting in a variety of trustworthy and productive models to assist healthcare professionals.In this study, a variety of machine learning (ML) classifiers were employed to classify a CKD dataset, including naïve Bays, stochastic gradient descent (SGD), random tree, J48, K-nearest neighbor (K-NN), and decision table. Classification subset evaluators were used to develop prediction models for the CKD dataset and to assess classifier performance after applying feature selection procedures for different classification algorithms.Feature selection algorithms might aid healthcare practitioners in detecting and predicting CKD patients early on. In all situations where feature selection was used, the performance of these prediction models was evaluated and compared, and positive findings were produced in terms of the accuracy of three distinct classifiers, namely naive Bayes, K-NN (K = 1), and decision table.To our knowledge, no previous work has described the classification and feature section, as well as sensitivity analysis, of a CKD dataset, and this is the main contribution of this work, which we hope adds value and supports the healthcare provided by identifying certain CKD cases at an early stage using the features presented in our feature selection section of this paper.

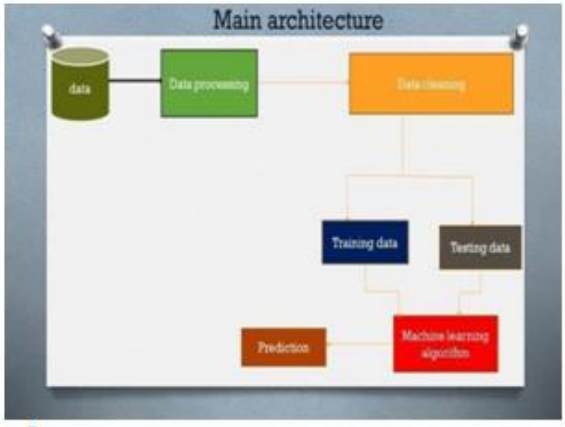

Figure 4: Main Architecture.

However, this study is noteworthy since no prior research has used age, sex, race, or serum creatinine levels to determine the stages of CKD. We use two machine learning algorithms, J48 and Random Forest, to predict the phases of CKD in this study. Our results are more accurate than most other

research, with 85.5 percent accuracy using the J48method in 0.03 seconds and 78.25 percent accuracy using the random forest technique in 0.28 seconds.

V.TABLE

S.no	Parameter	Abbr	Data
1	Age	age	51.5 average
2	Blood Pressure (mm/Hg)	bp	76.5
3	Specific Gravity (1.005, 1.010, 1.015, 1.020, 1.025)	sg	7 84 75 206 81
4	Albumin (0,1,2,3,4,5)	al	199 44 43 43 24 1
5	Sugar Degree (0,1,2,3,4,5)	su	209 13 18 14 13 3
6	Red Blood Cells rbc 47 abnormal (normal, abnormal) Red Blood Cells rbc 47 abnormal (normal, abnormal) Red Blood Cells (Normal, abnormal)	rbc	47 abnormal
7	Pus cell (Normal, abnormal)	pc	76 abnormal
8	Pus cell Clumps (Present, not Present)	pc c	42 present
9	Bacteria (Present, not Present)	ba	22 present
10	Blood Glucose Random (mgs/dl)	bgr	148.04
11	Blood Urea (mgs/dl)	bu	57.43
12	Cerum Creatinine (mgs/dl)	sc	3.07
13	Sodium (mEq/L)	sod	Avg. 137.53
14	Potassium (mEq/L) (mEq/L)	pot	Avg. 4.63
15	Haemoglobin (gms)	hemo	Avg. 12.53
16	Packed Cell Volume	pcv	Avg. 38.88
17	White Blood Count Count (cells/cumm)	wbcc	Avg. 8406.12
18	Red blood cell count (millions/cmm)	rbcc	Avg. 4.71
19	Hypertension (yes/no)	htn	147 Yes
20	Diabetes Mellitus (yes/no)	dm	137 Yes
21	Coronery Artery Disease (yes/no)	cad	34 Yes
22	Appetite (good/poor)	appet	82 Poor
23	Pedal edema (yes/no)	pc	76 Yes
24	Anemia (yes/no)	anc	60 Yes

Table 1:Parameters of Chronic Kidney Disease

VI. CONCLUSIONS

In terms of data imputation and sample diagnosis, the suggested CKD diagnostic approach is practical. The integrated model could obtain adequate accuracy after unsupervised imputation of missing values in the data set using KNN imputation. As a result, we hypothesize that when this technique is applied to the practical diagnosis of CKD, it has a positive result. In Furthermore, this concept may be applied to clinical data from different disorders in the real world. A medical assessment However, due to the constraints of the model, throughout the process of developing it, the settings, as well as the accessible data samples, are limited, with just 400 samples available.

As a result, the model's generalization performance may be restricted. Furthermore, because the data set contains only two types of data samples (ckd and notckd), the model is unable to determine the severity of CKD.

A significant amount of more complicated and representative data will be collected in the future to train the model to increase generalization performance while also allowing it to recognize illness severity. We anticipate that as the amount and quality of the data grows, our model will become even more perfect.

REFERENCES

[1] JIONGMING QIN, LIN CHEN, YUHUA LIU, CHUANJUN LIU, CHANGHAO FENG and BIN CHEN, A Machine Learning Methodology for Diagnosing Chronic Kidney Disease, 10.1109/ACCESS.2019.2963053.

[2] MerveDogruyolBasar, Aydin Akan, Detection of Chronic Kidney Disease by Using Ensemble Classifiers, 10th International Conference on Electrical and Electronics Engineering (ELECO).

[3] Mohamed Elhoseny, K. Shankar & J. Uthayakumar, Intelligent Diagnostic prediction and Classification System for Chronic Kidney Disease, (2019) 9:9583 | https://doi.org/10.1038/s41598-019-46074-2.

[4] Nikhila G, Meghashree, Machine Learning Framework to Predict Chronic Kidney Disease using Ensemble Algorithm, ISSN: 2249 – 8958 (Online), Volume-9 Issue-5, June 2020.

[5] Khaled Mohamad Almustafa, Prediction of chronic kidney disease using different

classification algorithms, Volume 24, 2021, 100631.

[6] Olayinka Ayodele Jongbo, Adebayo Olusola Adetunmbi, Roseline BosedeOgunrinde, Bukola Badeji-Ajisafe, Development of an ensemble approach to chronic kidney disease diagnosis, Volume 8, July 2020, e00456.

[7] Ebrahime Mohammed Senan, Fawaz WaselallahAlsaade, MoslehHmoud Al-Adhaileh, Theyazn H. H. Aldhyani, Ahmed Abdullah Alqarni, Nizar Alsharif, Ahmed H. Alahmadi, Mukti E Jadhav, M. Irfan Uddin and Mohammed Y. Alzahrani, Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques, Volume 2021, Article ID 1004767.

[8] Hamida Ilyas, Sajid Ali, MahvishPonum, Osman Hasan, Muhammad Tahir Mahmood, Mehwish Iftikhar, and Mubasher Hussain Malik, Chronic kidney disease diagnosis using decision tree algorithms, 2021 Aug 9. doi: 10.1186/s12882-021-02474-z.

[9] Marwa Almasoud, Tomas E Ward, Detection of Chronic Kidney Disease Using Machine Learning Algorithms with Least Number of Predictors, Vol. XXX, No. XXX, 2013.

[10] Vijendra Singh, Vijayan K. Asari and Rajkumar Rajasekaran, A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease, 2022,12,116. https://doi.org/10.3390/ diagnostics12010116

[11] Nishanth, A. &Thiruvaran, T. Identifying important attributes for early detection of Chronic Kidney Disease. IEEE reviews in biomedical engineering. 11, 208–216 (2017).

[12] Ardhanari, S., Alpert, M. A. & Aggarwal, K. Cardiovascular disease in chronic kidney disease: risk factors, pathogenesis, and prevention. Adv Perit Dial. 30, 40–53 (2014).

[13] Sarnak, M. J. et al. Kidney disease as a risk factor for development of cardiovascular disease: a statement from the American Heart Association Councils on Kidney in Cardiovascular Disease, High Blood Pressure Research, Clinical Cardiology, and Epidemiology and Prevention. Circulation. 108, 2154–2169 (2003).

[14] Walker, R., Marshall, M. R. &Polaschek, N. Improving self-management in chronic kidney disease: a pilot study. Renal Society of Australasia Journal. 9, 116–125 (2013).

[15] Shardlow, M. An analysis of feature selection techniques. The University of Manchester, 1–7 (2016).

[16] Dash, M. & Liu, H. Feature Selection for classification. Intell Data Anal. 1, 131–56 (1997).

[17] Guyon, I., Gunn, S., Nikravesh, M. & Zadeh, L. A. (Eds). Feature extraction: foundations and applications. (Springer 2018).

[18] Ekbal, A. &Saha, S. Joint model for feature selection and parameter optimization coupled with classifier ensemble in chemical mention recognition. Knowledge-Based Systems 85, 37–51 (2015).

[19] Jiang, L., Zhang, H. & Cai, Z. A novel Bayes model: Hidden naive Bayes. IEEE Transactions on knowledge and data engineering. 21, 1361–1371 (2008).

[20] Li, C. & Li, H. One dependence value difference metric. Knowledge-Based Systems 24, 589–594 (2011).

[21] Jensen, R. Combining rough and fuzzy sets for feature selection. (Doctoral dissertation, University of Edinburgh, 2005).

[22] Xue, B., Zhang, M., Browne, W. N. & Yao, X. A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation. 20, 606–626 (2015).

[23] Chandrashekar, G. &Sahin, F. A survey on Feature eS methods. ComputElectr Eng. 40, 16–28 (2014).

[24] Xue, B., Zhang, M. & Browne, W. N. A comprehensive comparison on evolutionary feature selection approaches to classification. International Journal of Computational Intelligence and Applications. 14, 1550008 (2015).

[25] Tsanas, A., Little, M. A. &McSharry, P. E. A simple filter benchmark for feature selection. Journal of Machine Learning Researchm, 1–24 (2010).