iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 11 **Issue:** XI **Month of publication:** November 2023

**DOI:** https://doi.org/10.22214/ijraset.2023.56526

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 11 Issue XI Nov 2023- Available at www.ijraset.com*

# An Ensemble Approach to Chronic Kidney Disease's Diagnosis

Uday Jain[1], Daksh Jain[2], Aditya Raj Varshney[3]
[1]*Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi*
[2]*Computer Science and Engineering, Bhagwan Parshuram Institute of Technology,*
[3]*Computer Science and Engineering, Amity University Noida*

***Abstract:*** ***The most essential priority right now is healthcare which incorporates the identification, treatment, prevention, and management of illness, injury, or sickness. Chronic diseases, the most harmful type of diseases, are more prevalent in senior individuals and are often treatable but incur a significant financial cost, adding to the challenges the patient and the patient's family already confront. These have great impacts on the kidneys, damaging the waste-filtering mechanism of the body. As a result, technologies like artificial intelligence (machine learning-ML as well as deep learning-DL) are now being used to forecast and enhance the health of human systems in an efficient, inexpensive and reliable way. In this study, a stacking model with SVC, Adaboost and Random forest is being proposed which is trained on a dataset (n=400) collected in india over a time span of two months which includes 25 features (including red blood cell (RBC) count, white blood cell (WBC) count, etc). The data went through Exploratory Data Analysis (EDA), followed by feature extraction using Adaboost. This data was then used for model training using different classifiers, including the proposed model. The stacking model gave best accuracy (100%), precision (100%) and recall (100%) in comparison to SVC (Support Vector Classifier), Random forest and Adaboost models used individually.***

***Keywords: Kidney disease, CKD, ensemble method, machine learning, prediction, adaboost, SVC, SMOTE, random forest, feature selection, medical data, diagnosis.***

## I. INTRODUCTION

Chronic diseases and disorders are becoming increasingly prevalent all over the world. Age factors and changes in cultural lifestyle are contributing to an increase in these frequently occurring and expensive long-term health complications. Chronic diseases are anticipated to kill almost 41 million people worldwide each year, accounting for seven out of ten fatalities and responsible for 71% of all deaths worldwide. Premature deaths account for around 17 million of these deaths, with deaths occurring far younger than expected on average.

Chronic kidney disease (CKD) is the world's 16th major source of lost years of life around the world [1]. Also known as Chronic Renal Disease, It is distinguished by a continuous decline in kidney health and activity over months or years. Glomerular filtration rate or GFR less than 60 mL/min/1.73 m2, albuminuria of at least 30 mg per 24 hours, or evidence of renal injury (eg, hematuria or structural abnormalities including polycystic or dysplastic kidneys) that linger for more than 3 months are indicators of CKD. [2]. This disease is associated with quite a few yet severe clinical outcomes, including kidney failure necessitating renal transplantation treatment, mortality, and overall poor quality of life in survivors [3].

The four priority Non-communicable Diseases (NCDs) included in the WHO global action plan for the prevention and control of NCDs excludes CKD [4]. However, it is well evident that CKD is closely related to diabetes, hypertension, and cardiovascular disease (CVD). CKD is a significant risk factor for many diseases. Moreover, Diabetes as well as hypertension are among the major risk factors for CKD itself, other factors being obesity, infectors, Acute Kidney Injury (AKI), Maternal, reproductive, pediatric health, kidney stones, and Nephrotoxins [5]. Furthermore, kidney failure increases the considerable burden and public health consequences of both infectious and non-infectious disorders [6, 7].

Chronic illnesses place a significant cost on citizens and governments both. The prominence of CKD is higher in nations with low or moderate incomes than in high-income ones [8]. Expensive treatments such as hemodialysis, peritoneal dialysis, and kidney transplantation can all help restore health. However, early detection of CKD by modern technologies including Machine Learning will be critical in both the prevention and treatment and will aid in the reduction of medical costs [9].

The primary goal of this research is to propose a stacking classifier based on several ML methods to create a powerful model that

© IJRASET: All Rights are Reserved | 

gives better accuracy than already existing ML approaches to detect CKD.

Thus, the proposed methodology in this study ensures a cost-effective, efficient and reliable technique which is an ensemble of various powerful ML algorithms and includes techniques such as SMOTE and K-fold Cross validation . This study comprises 4 sections. Section 2 includes a literature review of various existing studies related to ML-assisted prediction of CKD in a tabular form. Section 3 includes the proposed methodologies followed by Section 4 which incorporates the experimental results and analysis. Section 5 terminates the study with a summary of results, limitations and future scope.

## II. LITERATURE REVIEW

In this section, a review of already existing studies primarily focused on detection of CKD using ML techniques is being done. Many studies implemented support vector machine (SVM), K- nearest neaighbours (KNN), gradient boosting, decision trees and random forests (RF) individually to detect CKD and the best accuracy achieved by them is 99.75%. A detailed review of each of them is as follows:

To overcome the difficulty of predicting chronic kidney disease (CKD), Zewei Chen et al. utilized a system that used a variety of physiological indicators as well as machine learning (ML) techniques. K-nearest neighbors (KNN) and support vector machine (SVM) were utilized. Soft independent modeling of class analogy (SIMCA) combined with SVM performed the best with a 99.0% accuracy [10].

Researchers, including Nusrat Tazin and colleagues, also employed SVM, Decision Trees (DT), Naive Bayes (NB), and K-Nearest Neighbors (KNN) algorithms. and examined the models on the basis of accuracy, Root Mean Squared Error, etc. The DT model had the greatest accuracy of 99% [11].Classifiers such as KNN, NB, and SVM have also been used in research by S.B. Akben. This was done in conjunction with the k-means approach, which served as both a feature extractor and a classifier. KNN performed the best, with a 97.8% accuracy [12].

In another study by Marwa Almasoud et al., several statistical tests were performed to remove some less important features. Logistic regression (LR), SVM, random forest (RF), and gradient boosting (GB) algorithms were used along with 10-fold cross-validation, resulting in the GB model having the highest accuracy of 99.1% [13].Six algorithms were used in the research performed by Jiongming quin et al, including LR, RF, SVM, KNN, NB classifiers, and feedforward neural networks, with the FR technique yielding the greatest accuracy of 99.75% [14].

Researchers Huseyin Polat et al. tested the wrapper technique and filter approach as two feature selection approaches using two subset evaluators each, namely the greedy stepwise approach and Best First search on the SVM model. In the instance of filtered feature selection using the Best First search engine being used, an accuracy of 98.5 % was achieved [15].

Sarah A. Ebiaredoh-Mienye et al. They implemented a strategy that consisted of both feature learning and classification phases.that incorporates an improved Sparse Autoencoder (SAE) and Softmax regression to cope with the imbalance dataset and achieved a 98% accuracy [16].

Mirza MuntasirNishat et al. applied algorithms such as KNN, LR, DT, RF, SVM, NB, Multi-Layer Perceptron (MLP), and Quadratic Discriminant Analysis (QDA). For hyperparameter tuning, Random search cross-validation was also performed. The research discovered that the RF algorithm without tuning provided the best accuracy of 99.75% [17].

Gazi Mohammed Ifraz et al. applied LR, DT, and KNN classification methods on the UCI dataset, with the LR classification approach producing the greatest accuracy of around 97% [18].

Njoud Abdullah Almansour et al. used Artificial Neural Network (ANN) and SVM approaches on a dataset of 400 patients with missing values, filled using the mean of relevant characteristics. It discovered that the ANN model outperformed the SVM model with an accuracy of 99.75% [19].

Minhaz Uddin Emon and colleagues utilized a variety of machine learning algorithms including Random Forest (RF), Naive Bayes (NB), Logistic Regression (LR), Multi-Layer Perceptron (MLP), Stochastic Gradient Descent (SGD), Adaptive Boosting (AdaBoost), Bagging, Decision Trees (DT), along with 10-fold cross-validation to assess the performance of CKD prediction., with the RF model achieving the greatest accuracy of 99% [20].

Mohan Vijayarani explored ANN, KNN, DT, NB, SVM, and fuzzy classification models and discovered that the fuzzy model was the best strategy for identifying chronic renal disease after parameter assessment with almost 90% accuracy [21].

R. Subhashini and M.K. Jeyakumar examined the accuracy and execution time of SVM and ANN for predicting renal illness and discovered that ANN outperforms SVM with an accuracy of 87.70% [22].

SatriaWibawa et al. used three distinct ways to experiment with NB, KNN, and SVM classifiers. The first was just the Base classifier, the second method employed a classifier along with Correlation-based Feature Selection (CFS), and the third method was

a base classifier with CFS and AdaBoost. CFS and AdaBoost were found to succeed in improving the performance of the base classifiers. KNN along with CFS and AdaBoost generated an accuracy of 98.1% [23].

On the CKD Dataset, Ajay Kumar S et al. used KNN, LR, DT, RF, NB, SVM, and MLP algorithms and determined that the RF model performed the best with 95% accuracy [24].

Yedilkhan Amirgaliyev et al. examined the CKD dataset using SVM for accuracy, sensitivity, and specificity; experimental findings demonstrated over 93% accuracy [25]. Reshma S et al. used the Ant Colony Optimization (ACO) approach for feature selection along with SVM to determine the presence of CKD or not using the fewest features feasible and reached a 96% accuracy [26].

Researchers Junaid Rashid et al. implemented ANN and compared it to RF, LR, SVM, KNN, NB, DT, and Deep Learning approaches with epoch 10.0, which resulted in high accuracy of 99.67%. The study also revealed that the time required for ANN classification was significantly less than that of other classification algorithms [27].

Zixian Wang et al. applied methods such as Apriori Association, ZeroR, OneR, Naive Bayes, J48, and IBk. It was discovered that IBk, in conjunction with the Apriori algorithm, produced the greatest accuracy of around 99%. Seven important risk factors were identified [28]. Researcher Ramesh Revathy used DT, SVM, and RF to predict CKD in patients and found Random Forest to be the highest performer with an accuracy of 99.16% [29]. On the dataset, Yen-Ling Chiu et al. used stochastic gradient boosting (SGB), multivariate adaptive regression splines (MARS), eXtreme gradient, logistic regression (LR), and C5.0 decision tree (C5.0), with C5.0 decision tree having the greatestaccuracy of 82.31% [30].

The literature view of research done utilizing the UCI CKD Dataset is summarized in table 1.

Table 1. Comparative analysis of existing studies

| Authors | Year | Keywords | Methodology | Results | Source |
|---|---|---|---|---|---|
| Zewei Chen, Xin Zhang and Zhuoyong Zhang | 2016 | Chronic kidney disease (CKD), Risk assessment, Multivariate models, Clinical screening | KNN, SVM and soft independent modeling of class analogy (SIMCA) | Accuracy:99% | [10] |
| NusratTazin etal. | 2016 | Chronic Kidney disease (CKD), SVM, Decisiontree, Naïve Bayes, Receiver Operating Characteristic curve (ROC), KNN,WEKA | SVM, DT, NB, and KNN. | Accuracy:99% | [11] |
| S.B. Akben | 2018 | Chronic Kidney Disease (CKD), Kidneydisease diagnosis, Data mining, Machine learning | KNN, SVM, and Naïve Bayes | Accuracy:97.8% | [12] |
| MarwaAlmasoud and Tomas E Ward | 2013 | chronic kidney disease (CKD), Random forest(RF), Gradient boosting (GB), Logistic Regression (LR), SVM | LR, SV, RF, GB | Accuracy:99.1% | [13] |
| Jiongming quinet al. | 2019 | Chronic kidney disease, machine learning, KNN imputation, integrated model. | LR, RF, SVM, KNN, NB, and feedforward neural network | Accuracy:99.75% | [14] |
| HuseyinPolat etal. | 2017 | Feature selection,Support vector machine, Chronic kidney disease, Machine learning | Support Vector Machine (SVM) | Accuracy:98.5% | [15] |
| Sarah A. Ebiaredoh-Mienye, Ebenezer Esenogho, and Theo. G. Swart | 2020 | sparse autoencoder, unsupervised learning, Softmax regression, medical diagnosis, machine learning, ANN, e-health | Softmax Regression (SR), and Enhanced sparse autoencoder (SAE) | Accuracy:98% | [16] |
| Mirza MuntasirNishatet al. | 2021 | Chronic Kidney Disease, ML Algorithms, UCI Dataset, Accuracy, Precision, Sensitivity,F1 | KNN, LR, DT, RF, SVM, NB, Multi-Layer Perceptron (MLP), Quadratic Discriminant Analysis (QDA) | Accuracy:99.75% | [17] |

| | | | | | |
|---|---|---|---|---|---|
| | | score, ROC. | | | |
| Gazi Mohammed Ifraz et al. | 2021 | CKD, Intelligent Machine Learning Methods, KNN, LR,DT | LR, DT, and KNN | Accuracy:97% | [18] |
| Njoud Abdullah Almansour et al. | 2019 | Machine Learning,Artificial Neural Network (ANN), Support Vector Machine (SVM), Chronic Kidney Disease (CKD) | Artificial Neural Network(ANN), and SVM | Accuracy:99.75% | [19] |
| Minhaz UddinEmon et al. | 2021 | Chronic Kidney Disease, Machine Learning, Prediction,PCA, Co-relation Metrics, Random Forest. | RF, NB, LR, MLP, Stochastic Gradient Descent (SGD), Adaptive Boosting (AdaBoost),Bagging, DT, with 10-fold cross-validation | Accuracy:99% | [20] |
| R. Subhashini<br><br>and, M.K. Jeyakumar | 2017 | Chronic kidney disease, classification, artificial neural network, k-nearest neighbors,decision tree, naive Bayes. | ANN, KNN, DT, NB, SVM, and Fuzzy Logic | Accuracy:90% | [21] |
| Mohan Vijayarani | 2015 | Data Mining, Datamining techniques,Kidney disease, Support Vector Machine, ArtificialNeural Network. | SVM and ANN | Accuracy:87.70% | [22] |
| Made SatriaWibawa etal. | 2017 | chronic kidney disease,CFS, KNN, Naive Bayes, SVM, AdaBoost, feature selection | NB, KNN, SVM<br><br>With CFS and AdaBoost | Accuracy:98.1% | [23] |
| Ajay Kumar S etal. | 2020 | Chronic Kidney Disease (CKD), Chronic Kidney Disease Prediction System (CKDPS), MLalgorithms, Random Forest Algorithm, Userinput, System Output. | SVM, RF, DT, NB, KNN, Kbest, Recursive Feature Elimination (RFE), and Principal ComponentAnalysis (PCA) | Accuracy:95% | [24] |
| YedilkhanAmirg aliyev et al. | 2018 | Chronic kidney disease, Support vector machine, Biomedical engineering. | SVM with 10-foldcross-validation | Accuracy:93% | [25] |
| Reshma S et al. | 2020 | Chronic kidney, SVM,Ant colony optimization | SVM with Ant ColonyOptimization (ACO) | Accuracy:<br><br>96% | [26] |
| Junaid Rashid etal. | 2022 | medical diagnosis, feature selection, chronic diseases, artificial neural network (ANN), prediction | ANN, RF, DL, LR, SVM, KNN,NB, DT<br><br>10-fold cross validation, Particle Swarm Optimization (PSO) | Accuracy:99.67% | [27] |
| Zixian Wang etal. | 2018 | Machine Learning, Classification Technique, PredictionSystem | Apriori association, ZeroR,OneR, NB, J48, and IBk | Accuracy:99.17% | [28] |
| Ramesh Revathy | 2019 | Chronic Kidney Disease, Decision Tree, | DT, SVM, and RF | Accuracy: | [29] |
| | | Machine Learning,Random Forest, Support Vector. | | 99.16% | |

| | | | | | |
|---|---|---|---|---|---|
| Yen-Ling Chiuet al. | 2021 | chronic kidney disease, health screening, machine learning algorithms, risk indicators assessment, education. | Stochastic Gradient Boosting (SGB), Multivariate Adaptive Regression Splines (MARS), eXtreme gradient, LR, and C5.0Decision Tree (C5.0) | Accuracy:82.31% | [30] |

## III. PROPOSED METHODOLOGIES

The proposed methodology As depicted in Figure 1 alongside... with the raw input data. An exploratory data analysis is done on the data followed by data pre-processing. The AdaBoost algorithm is then applied on pre-processed data for feature extraction. Then, several different classifiers including AdaBoost, SVC (Support vector classifier), random forests and the stacking model consisting of all the three algorithms are trained on this data. K-fold cross validation is done to tune the hyperparameters. The description about the main techniques and algorithms used in this study along with the quality or performance indicators is given in brief in the following subsections.

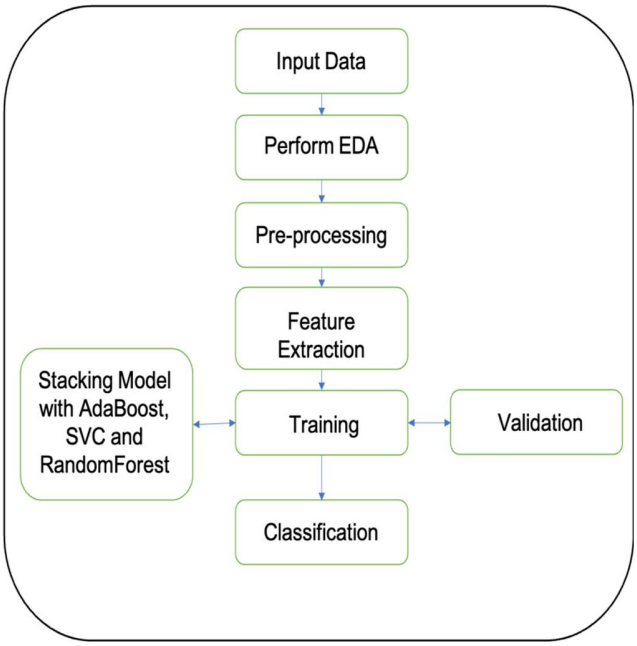



Fig. 1. Proposed flowchart

### *A. Quality Indicators*

In this experiment, accuracy, weighted mean recall, and weighted mean precision have been used as the three key error indicators to gauge the classification's quality. Precision and recall are two metrics that are frequently employed in data mining issues because they more clearly demonstrate the recognition quality as they are independent (explicitly) from the count of items in the test set. Accuracy represents the proportion of correctly predicted outcomes, or the relative count of accurately classified instances. Mathematically, it can be denoted as follows:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

They are determined from the confusion matrix.

$$precision = \frac{TP}{TP+FP} \times 100 \quad (2)$$

The percentage of correctly classified answers that are obtained from the classifier is known as precision. The number of erroneous hits decreases as precision increases. In all identified items, it calculates the proportion of actual detections. The weighted mean of all measurements for per- class precision is known as weighted mean precision. It is determined using class precisions for certain classes. The precision metric does not fully inform us about the accuracy of all classifier findings. There is one more statistic for this purpose, called recall:

$$recall = \frac{TP}{TP+FN} \times 100 \quad (3)$$

Recall is a metric that measures a classifier's capacity to "predict" the highest number of affirmative responses from the expected number. It gauges the proportion of real recognition among all the objects we are considering. Note that this statistic is unaffected by false-positive replies. The weighted mean of all per-class recall metrics is referred to as weighted mean recall. It is determined using class recall data for certain courses.

*B. Synthetic Minority Oversampling Technique (SMOTE)*

By giving "fake" instances as opposed to oversampling with replacement in the oversampling technique, the minority class would be eliminated. SMOTE reduces application- specificity by offering synthetic examples in "feature space" as opposed to "data space."

The method to address the minority class involves selecting samples from the minority class and generating synthetic data points along line segments connecting each of the k nearest minority class neighbors. The selection of neighbors from the k-nearest neighbors is done randomly, taking into account the required oversampling level. If a 200 percent oversampling is needed, only two out of the five nearest neighbors are chosen, resulting in the creation of one new sample in each direction.

Initially, the total number of sample events, denoted as N, is determined. Typically, the goal is to ensure an equal distribution between the two classes. However, based on the situation, this might be decreased. A true positive instance is chosen at random to start the iteration. The next step is to obtain the KNNs (by default, 5) for that instance. The foundation for the generation of additional synthetic instances is chosen as N of the K groups [14]. To do this, the value difference between the vector and its neighbor is calculated using any distance measure.

Now, this disparity is incorporated into the previously generated feature vector, along with any random number between (0, 1). Fig 2 represents the pictorial representation of SMOTE.

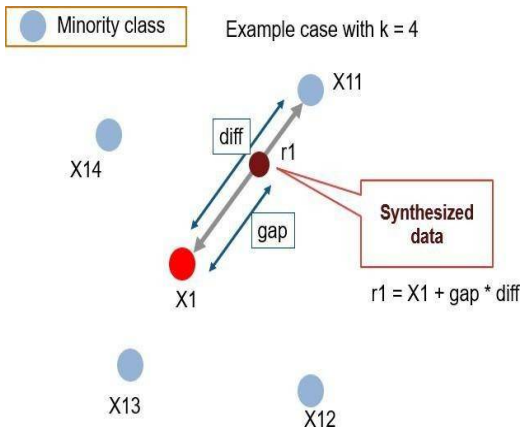



**Fig 2. Pictorial representation of SMOTE**

*C. K-Fold Cross-Validation*

To perform evaluation of machine learning models on a small data sample, a re-sampling approach known as cross-validation is employed. The procedure includes a single parameter, k, which specifies how many groups should be formed from a given data sample. As a result, the procedure is commonly known as k-fold cross-validation. When a specific value for k is specified, it can be substituted in the model's description, such as k=10 indicating 10-fold cross-validation. . Fig 3 shows the mean f1 score for different algorithms when k=10. Cross-validation is typically used to assess how well a machine learning model performs on untrained data. That is, a small sample size will be used to evaluate how the model will perform in general when used to make predictions on data that was not used during the model's training. As a general rule, the following happens:

*1)* Shuffle the dataset at random.
*2)* Splitting the dataset into k groups
*3)* For each separate group:
*a)* Employ the group as a test or holdout dataset.
*b)* Utilize the remaining data from the groups as a training set.
*c)* Train a model on the training data, then assess its performance on the test data.
*d)* Retain the evaluation results but discard the model.

*4)* Summarize the model's competence using a sample of model assessment ratings.

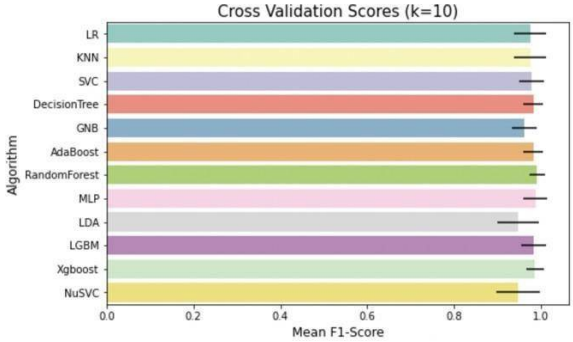



Fig. 3. Cross Validation Scores (k=10).

*D. Stacking Classifier – Ensemble*

Stacked Generalization, often known as Stacking, is a machine learning ensemble algorithm. The architecture of a stacking model consists of two or more base models, also known as level-0 models, and a meta-model, or level-1 model, which integrates the predictions of the base models. Level-0 Models (Base-Models) are models that have been fitted to training data and whose predictions have been collected. Level 1 model (meta-model) that learns the best approach to integrate predictions from base models. This is pictorially represented in fig 4.

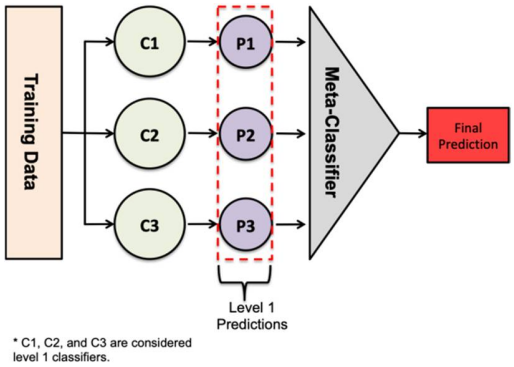



Fig. 4. Internal architecture of stacking classifier

The meta-model is trained using extrapolations made by the base models from data not present in the training sample.. In essence, the base models are presented with data that was not utilized during their training. They subsequently generate predictions based on this unseen data, yielding anticipated outcomes.which then serve as the input and output pairs of the training dataset that the meta-model is fitted to. The most prevalent method for creating the training dataset for the meta-modelinvolves employing k-fold cross-validation on the base models. In this process, the out-of-fold predictions made by the base models serve as the fundamental input data for training the meta-model.Once the training dataset for the meta-model has been produced, The meta-model can be exclusively trained on this dataset, while the base models have the option to be trained on the entire original training dataset.Stacking is appropriate when many machine learning models demonstrate competency on a dataset in different ways. Basically, the forecasts or mistakes in predictions of the models are uncorrelated or have a low correlation. Base-models tend to be complicated and diverse. As a result, employing a number of models that make widely disparate assumptions about how to solve the predictive modeling issue, such as linear models, decision trees, support vector machines, neural networks, and others, is often a wise approach. Other ensemble techniques, such as random forests, can also be utilized as foundation models.

## IV. EXPERIMENTAL ANALYSIS

Different techniques at different stages of developing the model were used and the results were analyzed and interpreted using graphs and tables. The detailed analysis of each step given in the proposed methodology and the results obtained at each phase are given in the coming subsections.

© IJRASET: All Rights are Reserved | 

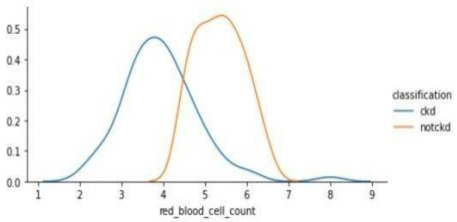
### *A. Dataset Collection*

The data was gathered in India during a span of two months. and included 25 features (such RBC count, WBC count, etc). The 'classification' serves as the target variable, taking on values of either 'ckd' or 'notckd' (where "ckd" indicates chronic kidney disease). The dataset was sourced from the UCI Machine Learning Repository.and was used to identify CDK patients. This study was carried out in order to "downstage" the illness (raise the proportion of CDK discovered at an early stage) to phases more susceptible to curative therapy. There are 400 entries in total. The information is available in a public repository [17].

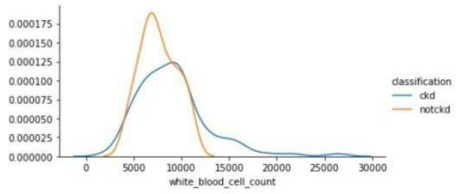
This dataset comprises 400 rows and 14 columns. The value of the output column "class" is either "1" or "0." The value "0" indicates that the patient does not have CKD, whereas the value "1" indicates that the patient does have CKD. There are a total of 250 instances of CKD data, and 150 instances of non-CKD data in this dataset.

### *B. Performing EDA*

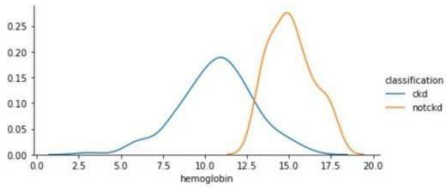
Exploratory Data Analysis is applied for discovering the most essential characteristics of a dataset. This dataset is utilized for comprehending data and placing it into context. identify components and associations, and offer suggestions to help in the creation of forecasting models. It allows for a more thorough examination of the entire dataset as well as a summary of important features such as class and size distribution. Fig 5 depicts the red blood cells, white blood cells, hemoglobin, and specific gravity relative to the disease.



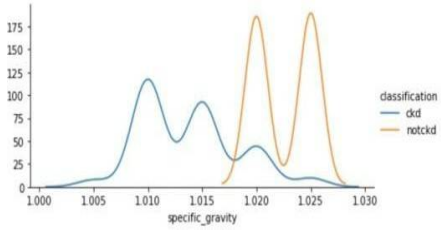

(a)




(b)




(c)




(d)

Fig 5. Type concerning RBCs, WBCs, hemoglobin, and specificgravity

*C. Feature Extraction*

The feature significance offered by AdaBoost is derived from the feature importance provided by its base classifier. If you use a Decision Tree as your basis classifier, the AdaBoost feature importance is determined by the average feature importance supplied by each Decision Tree.

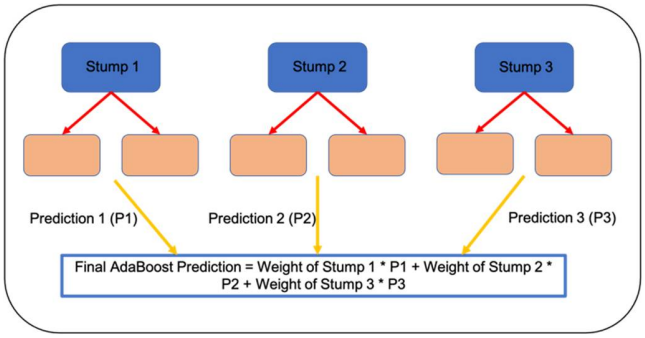



Fig. 6. Visualization of how AdaBoost's final predictions are made

It takes use of the fact that features towards the top of the tree influences the final prediction choice of a greater percentage of input samples. Using this expected fraction, one may assess the relative value of a feature. The difference between AdaBoost and, say, a Random Forest (forest of trees) that may influence feature importance determination is in how they produce variants of the base classifier. The former generates variants with a greater emphasis on "difficult" examples, while the latter generates variants by introducing randomness into the tree-building process.

AdaBoost feature importance represents that features such as serum creatinine and specific gravity have the highest relative importance among all, followed by hemoglobin and age in fig 7. The evaluation is done using various performance parameters which are accuracy, precision, recall, and F-measure which are obtained with help of confusion matrix which comprises of:

1) *True Negatives (TN)* – These are the values. that had a negative response in their actual class values and their predicted value also came out to be a negative response. For example, if a person does not have CKD and the model also predicts that she does not have CKD, then this instance comes under True negative.
2) *True Positives (TP)* - The values that had positive response in their actual class values and their predicted value also came out to be a positive response. For instance, if a person has CKD and the model also predicts that she has CKD, then this instance comes under True positive.
3) *False Negatives (FN)* – The values that had positive response in their actual class values and their predicted value came out to be a negative response. For example, if a person has CKD and the model predicts that she does not have CKD, then this instance comes under false negative.
4) *False Positives (FP)* – The values that had negative response in their actual class values and their predicted value came out to be a positive response. For example, if a person does not have CKD and the model predicts that she has CKD, then this instance comes under false positive.

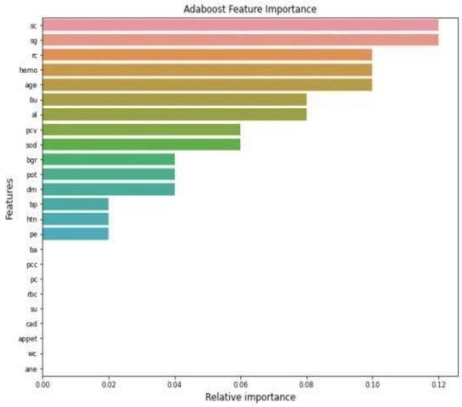


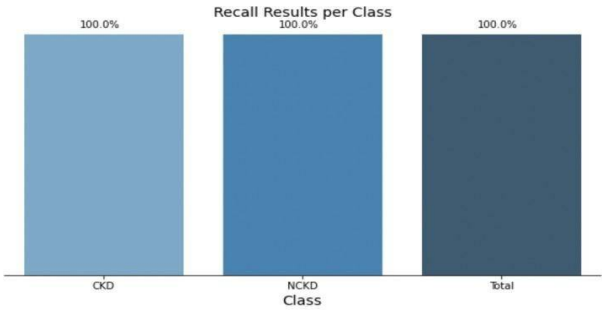

Fig. 7. AdaBoost feature importance

© IJRASET: All Rights are Reserved | 

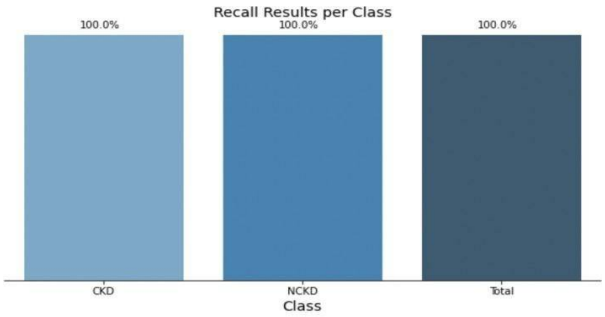
Table 2. Comparison of results

| Results | SVC | AdaBoost | Random Forest | Stacking Model with AdaBoost, SVC and Random Forest (Proposed) |
|---|---|---|---|---|
| Recall | 0.9 | 0.9921 | 0.9865 | 1 |
| Precision | 0.887 | 0.9868 | 0.9922 | 1 |
| F1 Score | 0.8987 | 0.9893 | 0.9892 | 1 |
| Accuracy | 0.8998 | 0.986842 | 0.9920 | 1 |

After training the stacking model, the accuracy graph for the test dataset (figure 6) and the comparative analysis of thedifferent models (in table 2) was obtained.




(a)




(b)




(c)

Fig. 6. Accuracy graphs

A confusion matrix is a tabular representation used to visualize the performance of a classification technique or classifier on a testing dataset where the actual values are known. It is particularly useful for supervised learning methods, typically displayinga matching matrix. In this matrix:

- Each row corresponds to instances from an actual class.
- Each column corresponds to instances from a predicted or forecasted class, or vice versa.

Figure 7 depicts the actual and predicted labels for features classified as malicious and non-malicious, typically represented as 0 and 1. The diagonal of the matrix represents True Positives (TP) and True Negatives (TN) for each class, while the off-diagonal elements represent False Positives (FP) and False Negatives (FN) in the predicted class labels, along with their respective scores.

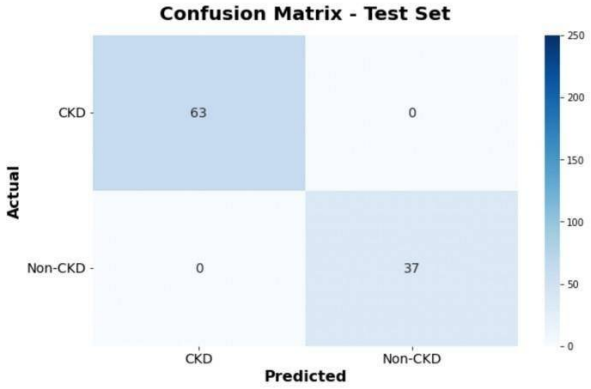



Fig. 7. Confusion matrix of predicted and true labels

## V. CONCLUSION

The dataset goes through an exploratory data analysis followed by data pre-processing. AdaBoost algorithm is applied on pre-processed data for feature extraction. Several different classifiers including AdaBoost, SVC (Support vector classifier), random forests and the stacking model comprising of all the three algorithms are trained on this data. K-fold cross validation is done to tune the hyperparameters. The result obtained showed that the stacking model was superior to the individual models with best quality indicators (accuracy, recall & precision) each having 100% score. Thus, suggested model demonstrates how SMOTE, feature importance with AdaBoost and ensemble stacking classifier with SVC, AdaBoost and random forest models may be merged to create superior models for chronic kidney disease classification and detection tasks.

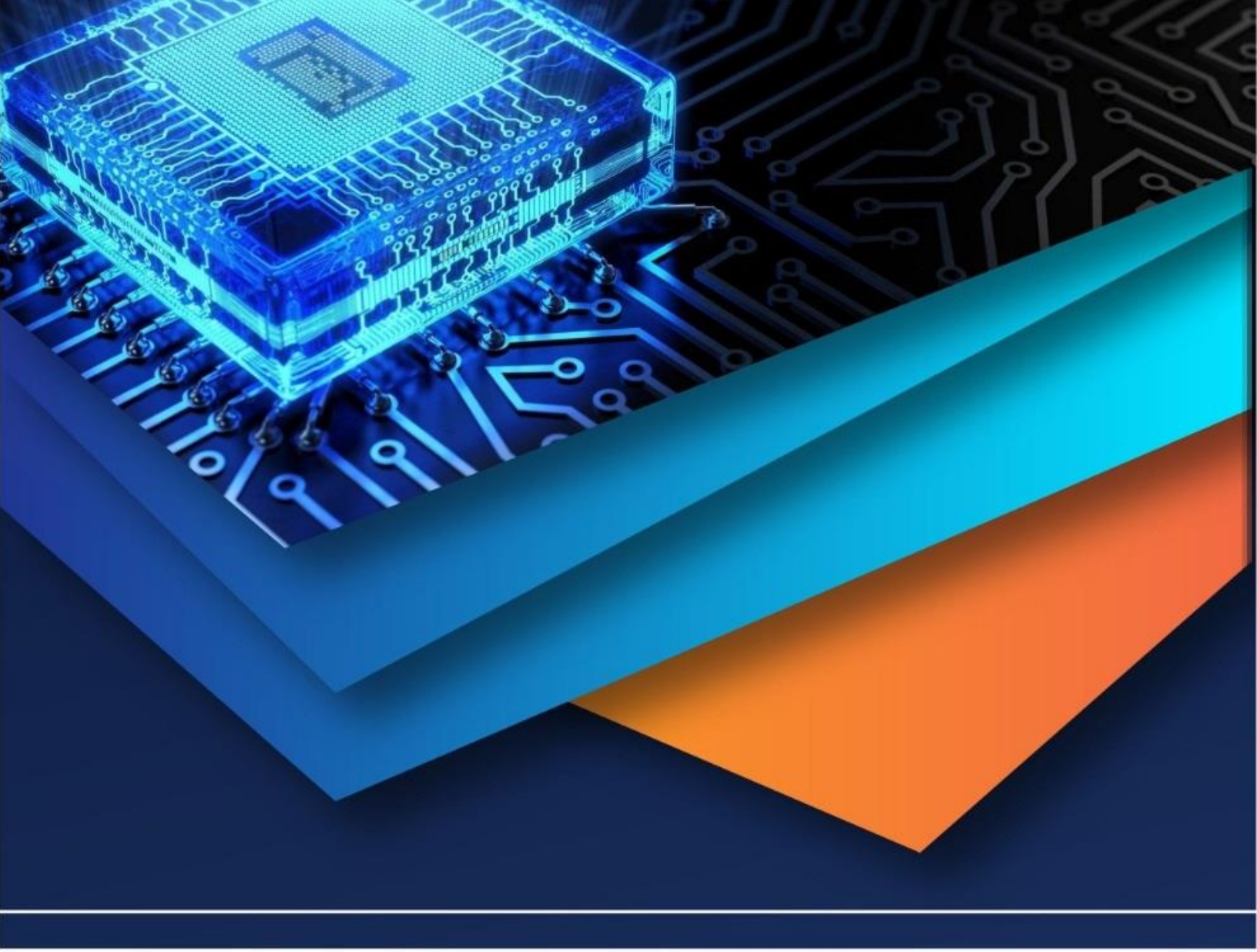
Despite of this approach's exemplary results, there are some limitations to this study. The dataset used is very small in size with only 400 entries, and it lacks diversity. Thus, using this same approach with the same hyperparameters on large and diverse datasets can be an objective for future studies.

## REFERENCES

[1] Chen, Teresa K et al. "Chronic Kidney Disease Diagnosis and Management: A Review." JAMA vol. 322,13 (2019): 1294-1304. doi:10.1001/jama.2019.14745
[2] Levin, Adeera, et al. "Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease." Kidney international supplements 3.1 (2013): 1-150.
[3] Bello, Aminu K et al. "Complications of chronic kidney disease: current state, knowledge gaps, and strategy for action." Kidney international supplements vol. 7,2 (2017): 122-129. doi:10.1016/j.kisu.2017.07.007
[4] Chestnov, O. "World Health Organization global action plan for the prevention and control of noncommunicable diseases." Geneva, Switzerland (2013).
[5] Luyckx, Valerie A et al. "Reducing major risk factors for chronic kidney disease." Kidney international supplements vol. 7,2 (2017): 71-87. doi:10.1016/j.kisu.2017.07.003
[6] Bansal N., Katz R., Robinson-Cohen C. Absolute rates of heart failure, coronary heart disease, and stroke in chronic kidney disease: an analysis of 3 community-based cohort studies. JAMA Cardiol. 2017;2:314–318.
[7] Kent S., Schlackow I., Lozano-Kuhne J. What is the impact of chronic kidney disease stage and cardiovascular disease on the annual costof hospital care in moderate-to-severe kidney disease? BMC Nephrol. 2015;16:65.
[8] Mills, Katherine T., et al. "A systematic analysis of worldwide population-based data on the global burden of chronic kidney disease in 2010." Kidney international 88.5 (2015): 950-957.
[9] AnusornCharleonnan, et al., Predictive analytics for chronic kidney disease using machine learning techniques, in: 2016 Management and Innovation Technology International Conference, MITicon, IEEE, 2016, http://dx.doi.org/10.1109/MITICON.2016.8025242.
[10] Chen, Z.; Zhang, X.; Zhang, Z. Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. Int. Urol. Nephrol. 2016, 48, 2069–2075.

[11] Tazin, N.; Sabab, S.A.; Chowdhury, M.T. Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique. In Proceedings of the 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec), Dhaka, Bangladesh, 17–18 December 2016; pp. 1–6.
[12] Akben, S. Early stage chronic kidney disease diagnosis by applying data mining methods to urinalysis, blood analysis and disease history. IRBM 2018, 39, 353–358.
[13] Almasoud, M.; Ward, T.E. Detection of chronic kidney disease using machine learning algorithms with least number of predictors. Int. J.Soft Comput. Appl. 2019, 10, 89–96.
[14] Qin, J.; Chen, L.; Liu, Y.; Liu, C.; Feng, C.; Chen, B. A Machine Learning Methodology for Diagnosing Chronic Kidney Disease. IEEE Access 2019, 8, 20991–21002.
[15] Polat, H.; Mehr, H.D.; Cetin, A. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. J. Med. Syst. 2017, 41, 55.
[16] Ebiaredoh-Mienye, S.A.; Esenogho, E.; Swart, T.G. Integrating Enhanced Sparse Autoencoder-Based Artificial Neural Network Technique and Softmax Regression for Medical Diagnosis. Electronics 2020, 9, 1963.
[17] Nishat, Mirza & Faisal, Fahim& Dip,Rezuanur&Nasrullah, Sarker& Ahsan, Ragib&Shikder, Md& Asif, Md. Asfi&Hoque, Md. (2021). A Comprehensive Analysis on Detecting Chronic Kidney Disease by Employing Machine Learning Algorithms. EAI Endorsed Transactions on Pervasive Health and Technology. 7. 1-12. 10.4108/eai.13-8-2021.170671.
[18] G. M. Ifraz, M. H. Rashid, T. Tazin, S. Bourouis, and M. M. Khan, "Comparative analysis for prediction of kidney disease using intelligent machine learning methods," Computational and Mathematical Methods in Medicine, vol. 2021, Article ID 6141470, 10 pages,2021.
[19] Almansour, Njoud& Syed, Hajra&Khayat, Nuha&Altheeb, Rawan&Juri, Renad&Alhiyafi, Jamal &Alrashed, Saleh &Olatunji, Sunday. (2019). Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. Computers in Biology and Medicine. 109. 10.1016/j.compbiomed.2019.04.017.
[20] M. U. Emon, A. M. Imran, R. Islam, M. S. Keya, R. Zannat and Ohidujjaman, "Performance Analysis of Chronic Kidney Disease through Machine Learning Approaches," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 713-719, doi: 10.1109/ICICT50816.2021.9358491.
[21] R. Subashini and M.K. Jeyakumar. 2017. Peformance Analysis of Different Classification Techniques for The Prediction of Chronic Kidney Disease. International Journal of Pharmacy and Technology, Volume 9, Issue No. 4, pp. 6563-6582.
[22] Mohan, Vijayarani. (2015). KIDNEY DISEASE PREDICTION USING SVM AND ANN ALGORITHMS.
[23] M. S. Wibawa, I. M. D. Maysanjaya and I. M. A. W. Putra, "Boosted classifier and features selection for enhancing chronic kidney disease diagnose," 2017 5th International Conference on Cyber and IT Service Management (CITSM), 2017, pp. 1-6.
[24] Survey on Chronic Kidney Disease Prediction System with Feature Selection and Feature Extraction using Machine Learning Technique.Ajay kumar S, Karthik Raja R, Jebaz Sherwin P, Revathi M4.
[25] Amirgaliyev, Y., Shamiluulu, S., &Serek, A. (2018). Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods. 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT).doi:10.1109/icaict.2018.8747140.
[26] Reshma S , Salma Shaji , S R Ajina , Vishnu Priya S R, Janisha A, 2020, Chronic Kidney Disease Prediction using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 07 (July 2020).
[27] Rashid J, Batool S, Kim J, WasifNisar M, Hussain A, Juneja S and Kushwaha R (2022) An Augmented Artificial Intelligence Approach for Chronic Diseases Prediction. Front. Public Health 10:860396. doi: 10.3389/fpubh.2022.860396.
[28] Wang, Zixian, Jae Won Chung, Xilin Jiang, Yantong Cui, Muning Wang, &Anqi Zheng. "Machine Learning-Based Prediction System For Chronic Kidney Disease Using Associative Classification Technique." International Journal of Engineering & Technology [Online], 7.4.36 (2018): 1161-1167. Web. 11 Jul. 2022.
[29] Ramesh, Revathy. (2020). Chronic Kidney Disease Prediction using machine Learning Models. 9. 6364. 10.35940/ijeat.A2213.109119.
[30] Chiu YL, Jhou MJ, Lee TS, Lu CJ, Chen MS. Health Data-Driven Machine Learning Algorithms Applied to Risk Indicators Assessment for Chronic Kidney Disease. Risk ManagHealthc Policy. 2021;14:4401-4412. https://doi.org/10.2147/RMHP.S319405.

© IJRASET: All Rights are Reserved | 



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)