International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7211
***Article Received**: 29 April 2023 **Revised:** 17 June 2023 **Accepted:** 29 June 2023*

Ensemble of Homogenous and Heterogeneous Classifiers using K-Fold Cross Validation with Reduced Entropy

R.Venkatarathinam[1], R.Sivakami[2], Prasanna Ranjith Christodoss[3], Mahesh T R[4], E. Mohan[5], Vinoth Kumar V[6]
[1]Professor, Department of Information Technology
Sri Balaji Chockalingam Engineering College, India
e-mail: hodinfotechsbcec@gmail.com
[2]Associate Professor, Department of Computer Science and Engineering
Sona College of Technology, Salem -636005
e-mail: shivasona07@gmail.com
[3]University of Technology & Applied Sciences, Shinas, Oman
e-mail: prasanna.christodoss@shct.edu.om
[4]Department of Computer Science and Engineering
Faculty of Engineering and Technology, JAIN (Deemed-to-be University), Bangalore, India
trmahesh.1978@gmail.com
[5]Professor, Department of ECE, Saveetha School of Engineering, SIMATS, Chennai, Tamilnadu, India
e-mail: emohan1971@gmail.com
[6]School of Information Technology and Engineering
Vellore Institute of Technology University, Vellore, India
e-mail: pvkumar243@gmail.com

Abstract— Chronic kidney disease (CKD) affects millions of people worldwide, greatly reducing their quality of life and creating serious economic, social, and medical problems. Some automated diagnosis methods can detect chronic renal disease. In-depth studies on data mining techniques have recently focused on accuracy in the diagnosis of chronic renal illnesses, either by taking advantage of the disease's simplicity or doing feature selection in addition to pre-processing. In order to handle the unbalanced dataset in this work, Synthetic Minority Over Sampling Technique (SMOTE) is used during pre-processing. For this investigation, 400 data from the publicly accessible UCI machine learning (ML) repository are used. For the implementation, both homogeneous and heterogeneous ensemble classifiers which combine two separate classifiers have been used. Different machine learning (ML) techniques, such as the Classification and Regression Tree (CART), Adaboost classifier, Decision Tree (DT), Reduced Error Pruning Tree, Alternating Decision Tree, and Random Forests Algorithm and their ensembles with a significant reduction in entropy, are used to perform the classification. With a 99.12% accuracy rate and a 99.10% f1 score, the homogeneous classifier Adaboost-Random Forest outperforms other models in the prediction of CKD.

Keywords- machine learning; entropy; Chronic Kidney Disease; ensemble methods; precision.

I. INTRODUCTION

Chronic kidney disease (CKD), a widespread disorder affecting people of all races, claims millions of lives each year due to inadequate timely and accurate diagnosis, particularly in underdeveloped countries where there are appallingly few medical experts and facilities for the treatment. Both kidneys are affected by the illness known as chronic kidney disease (CKD), which makes it harder for them to filter blood wastes excreted in urine and regulate physiological fluids. According to the Centers for Disease Control and Prevention [1], over thirty million people globally have CKD as of 2017, and 96% of those who have renal issues aren't even aware of it. In order to help medical professionals, make decisions regarding the prognosis of CKD, it was necessary to develop a reliable, accurate, and efficient predictive model using data mining, which may be used to discover hidden patterns from data. Data mining [2], has recently been the topic of substantial research in the extraction of hidden patterns from the medical profession for the diagnosis of various diseases such as CKD. The illness known as chronic kidney disease (CKD) impairs the kidney's capacity to operate [11]. The prevalence of CKD is now estimated to be higher in those over 65 (38%) than in those between the ages of 45 and 64 (12%) and 18 to 44 (6%). Compared to men, women have a slightly greater CKD prevalence (14%).

Feature selection (FS), a component of the machine

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7211
***Article Received**: 29 April 2023 **Revised:** 17 June 2023 **Accepted:** 29 June 2023*

learning (ML) process, is a fundamental pre-processing phase that chooses the most pertinent attributes from a dataset. Models can be made simpler and more accurate by removing unneeded and redundant attributes. Relief-F [4] as well as chi-squared [5] feature selection methods are employed in this research as 2 feature selection. To predict CKD, some research studies have used ML approaches. For instance, K-nearest neighbours (KNN), support vector machine (SVM), logistic regression (LR), and decision tree (DT) are four machine learning (ML) techniques that have been utilized by Charleonnan [6] et al. to predict CKD. Several studies have employed feature selection techniques along with hybrid ML algorithms for the prediction of CKD. Methods for reducing the amount of features and choosing the best feature subsets from the dataset have been employed. For instance, in [7], authors chose the crucial features from the database using chi-square and the chosen features were subjected to the same applications of ANN. Big data platforms like Apache Spark [8], a large-scale data processing engine with a unified analytics engine, have recently been used by researchers. For running workloads on massive clusters, Spark is actulaly100 times faster than that of Hadoop. High-level APIs for Java, Scala, Python, and R are included, along with a powerful engine that can handle big execution graphs. A variety of higher-level tools are also included, including Spark for processing.

II. RELATED WORK

Polat, H. et al. [9] used SVM and efficient feature selection techniques to diagnose chronic renal disease. In their research, they implemented SVM, which had an accuracy of 97.75%. In order to forecast cardiac disease, Bashir, S. et al. [10] suggested an ensemble classifier that uses vote-based approach on majority. The ensemble model was built using five heterogeneous classifiers. Following testing using a technique called stratified cross-validation, they found that their framework had an accuracy of 88.5%, 90.83% specificity, 86.96% sensitivity as well as 88.85% F1-Measure. They then compared this accuracy to the base classifiers and found that this increased the ensemble model's average accuracy. A medical decision making support framework was proposed by Bashir, S. [11]. They outperform the competing prediction models with their HMV ensemble framework.

By combining DT, SVM, Nave Bays, as well as C4.5, the authors [12] used a classification algorithm Breast Cancer Recurrence prediction. They used efficient feature selection techniques to increase each classifier's accuracy. SVM achieved 75.75% accuracy after being deployed on the Weka tool. The data set, devoid of feature selection, is presented here. Following rigorous feature selection, SVM increased by 1.52%, Naive Bayes(NB) by 9.09% as well as C4.5 by 2.52%.

U. N. Dulhare et al. [13], in order to distinguish between patients with and without CKD, classification models were built, feature selection was utilized to extract an action rule. These techniques combine the best first search with Nave Bays with a wrapper subset evaluator. NB classifier achieved a accuracy of 97.5% after being deployed on the Weka tool. The table 1 depicts the related work for CKD prediction on publicly available CKD dataset.

The authors of [14] chose the crucial features from the database using the feature selection namely, chi-square, CFS, as well as Lasso. They utilized full functionality and chosen features for different ML algorithms. According to the findings, LSVM had the maximum accuracy of 98.86%. The authors in this [15], employed the classifiers K-star, NB, SVM, as well as J48 for the prediction of CKD. Using WEKA software, performance comparisons were conducted. The authors claimed that J48 technique outperformed with 99% accuracy. In [16], the most significant features from the database were chosen using the backward selection, RF feature selection, forward exhaustive selection, forward selection as well as backward exhaustive approaches. To predict CKD, four ML algorithms SVM, RF, SVM, LR, and NB have been implemented. According to the results, RF with RF feature selection outperformed with 98.8% accuracy.

Table 1. Summary of Related work for CKD prediction

Study and Year	Models	Feature selection method
Chittora P et al., (2021) [14]	RF,ANN,KNN, LSVM, LR and C5.0	CFS, Wrapper method and Lasso
Avci E et al., (2019) [15]	SVM , J48 NB, K-Star,	No
Abdullah A et al., (2020) [16]	LR, RF, NB, SVM	backward selection, RF feature selection
Jena L et al., (2021) [17]	DT, NB, J48	Genetic search algorithm
Jongbo O et al., (2020) [18]	Bagging, random subspace, DT, KNN and NB	No
Wibawa M et al.,(2017) [19]	SVM,NB,KNN	CFS
Almansour N et al., (2019) [20]	SVM and ANN	Correlation coefficients

The key features from the CKD dataset were chosen using the genetic search algorithm in [17]. Both the whole set of features and the chosen features have been subjected to DT

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7211
***Article Received**: 29 April 2023 **Revised:** 17 June 2023 **Accepted:** 29 June 2023*

Table, J48, MLP and NB analyses. Using a genetic search technique improved the results. The MLP classifier outperformed with the best performance. In [18], the authors employed three base-learners KNN, DT and NB along with two ensemble techniques, Bagging and Random Subspace methods—to predict CKD. By using a KNN classifier, the random subspace outperformed Bagging. The authors [19], used a feature based correlation selection method to choose the number of significant characteristics (CFS). The detection of CKD has been carried out using AdaBoost, KNN, NB, and SVM. The best result was obtained by the suggested CFS with AdaBoost, which had 98.1% accuracy. The actual findings of this study [20] showed that ANN outperformed SVM in the experiments, with 98.75% and 97.75% accuracy respectively.

Earlier studies did not employ heterogeneous classifiers; instead, they applied ML approaches to explore and interpret. This inspires us to suggest this study and examine CKD data, including homogeneous and heterogeneous classifiers.

III. METHODOLOGY

The proposed study in this paper is divided into two steps. The technique, SMOTE is being utilized in the first step to decrease the effect of data imbalance in the dataset. The second step involves classifying CKD data using the different DT algorithms (ADTree, CART, REPTree, and Random Forest) and their ensembles with Adaboost classifiers. Then the heterogeneous classifiers are used to predict CKD and their performance is compared with that of homogeneous Adaboost classifiers. Finally, the ML model with the best accuracy is recommended. Figure 1 depicts the Proposed Methodology.

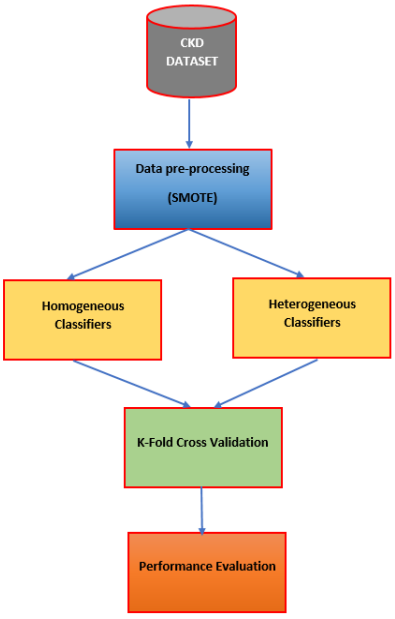

Figure 1. Proposed Methodology

A. Dataset

The publicly available CKD dataset available in [21], contains 400 records.

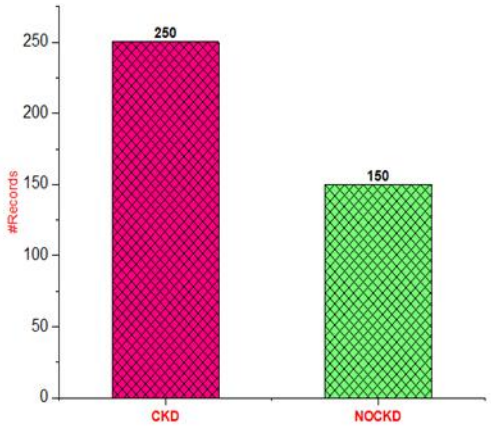

Figure 2. Count of “CKD” and “NOCKD” classes

The dataset also contains the class characteristics "ckd" and "notckd," as well as twenty-four features broken down into thirteen category features and eleven numeric features. Some of the characteristics include hypertension, pedal edoema, coronary artery disease, haemoglobin, specific gravity, red blood cell count, age, blood pressure, red blood cells, albumin, pus cells, sodium, potassium, sugar, diabetes mellitus, appetite, packed cell volume, white blood cell count, and anaemia. There are two diagnostic classes: Ckd and Notckd [22]. Figure 3 depicts Count of “CKD” and “NOCKD” instances in the dataset.

B. Data-preprocessing

In the dataset, there are missing values in all the features except diagnostic class. The KNN Imputer method was used in this investigation to fill in the missing information. The dataset is imbalanced since there are 150 examples of "notckd" (37.5%) and 250 cases of the "ckd" (62.5%) class as shown in figure 2. SMOTE is an oversampling method of balancing class distribution in the dataset. The minority examples that are near the feature space are chosen. Then, a new sample is drawn at a location along the line that is drawn between the examples in the features space [23]. SMOTE technique uses K Closest Neighbors to choose a random neighbour and a random example from the minority class. Between two instances in the feature space, the synthetic example is produced. The algorithm 1 depicts the pseudocode for SMOTE [24,25].

The analysis of the features before and after the application of the SMOTE technique is shown in table 2.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7211
***Article Received**: 29 April 2023 **Revised:** 17 June 2023 **Accepted:** 29 June 2023*

Table 2. SMOTE analysis

Components		Before SMOTE	After SMOTE	(Mean±SD)
		Overall (N=400)	Overall (N=750)	
Age (Mean±SD)		37.42 ± 15.40	38.40 ± 16.53	-
Gender, [IN (%)]	Male	280 (70)	375 (50)	37.26 ± 15.52
	Female	120 (30)	375 (50)	
CKD		250 (62.5)	300 (40)	37.20± 16.52
Non-CKD		150 (37.5)	300 (40)	
Testing Samples		-	150 (20)	37.29 ± 16.45

C. *Feature Selection*

The Pearson correlation coefficient can be used to describe a pair of properties' linear correlation [26]. The heat map is depicted in Figure 3.

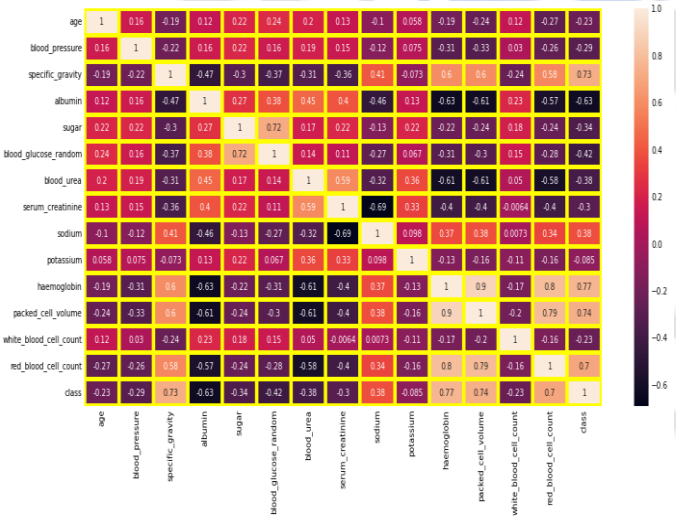

Figure 3. Heat map

Apart from class label, we selected 14 pertinent features among the 24 dependent characteristics present in the dataset to build the prediction model.

D. *Classification Techniques*

In this section, different machine learning techniques namely Adaboost classifier, CART, Reduced Error Pruning Tree, Alternating Decision Tree, Random Forests Algorithm along with homogeneous Adaboost classifiers and heterogeneous classifiers are discussed.

Decision Tree (DT) : A decision tree is a type of flowchart that, in its most basic form, depicts a clear path to a choice. It is a kind of algorithm used in data analytics that classifies data using conditional "control" expressions [27]. A decision tree begins at a single node (or "node") and branches out in two or more ways from there. Each branch presents various potential results, combining a range of choices and unforeseen circumstances until a decisive result is reached. They resemble trees when displayed graphically, hence the name. Since they divide complex data into more digestible pieces, decision trees are very helpful for data analytics and machine learning. For data classification, regression, and prediction analysis, they are frequently employed in these domains.

CART: Any classification or regression tree's study aims to establish a series of if-then scenarios that permit precise case prediction or categorization. Based on the collection of if-else conditions, classification as well as regression trees offer precise predictions [28-30]. CART makes it possible to quickly classify fresh observations. This is due to the fact that scoring each group using complicated nonlinear equations is far more difficult than just evaluating one or two logical requirements. It frequently leads to a more straightforward model that explains why the observations are categorized or forecasted in a particular way. Analytics includes feature selection and variable screening as key components. Here, feature selection is carried out automatically.

Reduced Error Pruning Tree (REPT) :The REPT method is based on the idea that variance-induced error can be reduced by computing information gain using entropy as well as backfitting. Entropy is a metric used in data science to assess how "mixed" a column is. Entropy is specifically used to quantify disorder. The (im)purity of any random set of instances is characterized by the entropy, which is extremely frequent in information theory. It is computed using the equation (1).

$$Entropy = -\sum p(X)logp(X) \quad (1)$$

where p(X) represents the percentage of examples in a given class.

Information Gain is the anticipated decrease in entropy brought on by dividing the instances into groups based on a specific attribute. By comparing the entropy of the dataset before and after a transformation, information gain is computed. Mutual information, also known as information gain when used to variable selection, determines the statistical dependence between two variables. It is computed using the equation (2).

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7211
***Article Received**: 29 April 2023 **Revised:** 17 June 2023 **Accepted:** 29 June 2023*

ALTERNATING DECISION TREE (ADT) :

A unique class of categorization models is called an Alternating DT. It is an extension of Voted Decision Trees, Voted Decision Stumps, and Classic Decision Trees. It enables any boosting implementation to extract the ADT model from the data as a learning method. A contemporary computational statistical approach called "boosting" can be used to enhance classification performance overall. In the context of the DT, ADT is a compelling extension of boosting. It enables the adoption of different boosting strategies to create an ADT model with distinctive properties to handle a good amount of applications.

Adaboost : AdaBoost, often known as adaptive boosting, is a boosting algorithm. The strategy to correct its predecessor is applied by this algorithm. It focuses more on training instances where the prior model's fit was inadequate. As a result, the complicated situations receive more attention than the others for each new predictor. A series of weak learners are fitted to various weighted training data. It begins by forecasting the initial batch of data and gives each observation the same weight. When the first learner makes an erroneous prediction, the observation that was incorrectly forecasted is given more weight. Due to the iterative nature of the process, it keeps adding learners until the accuracy or quantity of models can no longer be increased. AdaBoost mostly employs decision stamps. But, if a machine learning algorithm accepts weights from a training data set, we can use that algorithm as our base learner. In machine learning, we can utilize the AdaBoost algorithm for both classification and regression issues.

Homogeneous and heterogeneous Classifiers: The Adaboost classifier is combined with other single classifiers to form homogeneous classifiers. This results in four combinations namely, Adaboost-REPT, Adaboost-RF, Adaboost-CART and Adaboost-ADT. The performance of all these are then evaluated against different metrics. The heterogeneous combinations include REPT-ADT, REPT-RF, REPT-CART, ADT-RF, ADT-CART, RF-CART resulting in six different combinations. First, the performance of these heterogeneous classifiers are compared, then they are compared with that of single classifiers as well as homogeneous classifiers to figure out the best ML model for the prediction of CKD.

IV. RESULTS AND DISCUSSIONS

The effectiveness of the base classifiers and the ensemble of homogeneous and heterogeneous classifiers in diagnosing and predicting CKD is evaluated in this section. Any classification algorithm's effectiveness is evaluated in terms of accuracy. Yet, depending just on classification accuracy, particularly for a medical dataset that is imbalanced, can occasionally be deceptive. As a result, in addition to accuracy, the performance of the classifier models is evaluated using measures including f1 measure, precision, sensitivity, Matthew's Correlation coefficient (MCC) and specificity. Using the parameters derived from the confusion matrix, namely True Positive (TPs), which denotes CKD predicted as CKD, True Negative (TNs), False Positive (FPs), which denotes Normal predicted as CKD, and False Negative (FNs), which denotes CKD predicted as Normal, the effectiveness of the classifier is experimentally assessed.

Accuracy is defined as the metrics that identify the proportion of samples in the testing database that are correctly classified into a given class. The accuracy is mathematically computed using the equation (3).

$$Information\ Gain = entropy(parent) - [averahe\ entropy\{children)] \quad (2)$$

$$MCC = \frac{TPs * TNs * FPs * FNs}{\sqrt{(TPs + FPs)(TPs + FNs)(TNs + FPs)(TNs + FNs)}} \quad (8)$$

$$Accuracy = \frac{TNs + TPs}{TNs + TPs + FPs + FNs} \quad (3)$$

The sensitivity and specificity are computed using the equations (4) and (5) as shown.

$$Sensitivity = \frac{TPs}{TPs + FNs} \quad (4)$$

$$Specificity = \frac{TNs}{TNs + FPs} \quad (5)$$

Using equations (6) and (7), precision and F1 measures are generated to assess the effectiveness of the algorithm.

$$Precision = \frac{TPs}{TPs + FPs} \quad (6)$$

$$F1\ measure = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivit} \quad (7)$$

Matthew's Correlation Coefficient (MCC), which is calculated using equation (8), is a metric used to assess the effectiveness of a binary classifier for identifying CKD in a patient.

The amount of time needed to finish training or modelling a dataset is referred to as Time Took to Build the Model (TTBM). It is computed in seconds.

Mean absolute error (MAE) as well as root-mean-squared error (RMSE) are two common metrics used in model

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7211
***Article Received**: 29 April 2023 **Revised:** 17 June 2023 **Accepted:** 29 June 2023*

evaluation. Equations (9) and (10) are utilized to determine the MAE and RMSE for a sample of n observations. yi denotes the realized value , $\hat{y}$ denotes the predicted value and $\underline{y}$ denotes the mean of the realized values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \quad (9)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (10)$$

The square root of the squared errors of a predictive model normalized by the squared errors of a simple model is known as the Root Relative Squared Error (RRSE) and is computed using (11).

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \underline{y_i})^2}} \quad (11)$$

$$RAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|y_i - \underline{y_i}|} \quad (12)$$

The Relative Absolute Error (RAE) measures how well a model predicts actual data as opposed to just the average and it is computed using the equation (12). The model performs better than the simple model if the RAE is less than one. A perfect model has a relative absolute error of zero.

The models are optimized and the hyperparameters are tuned using stratified K-Fold cross-validation. Grid search is the approach used most frequently for hyper parameter optimization. First, we establish a collection of values for each hyper parameter. The hyperparameters are then evaluated for each conceivable value, and the model selects the hyper parameter with the best performance. The CKD dataset is partitioned into 10-folds of same and equal size for 10-Fold cross-validation. The classifiers are tested in the remaining time after the k-1 group training has been completed. For each k, the classifiers' effectiveness is also evaluated. Lastly, the evaluation classifier is developed based on average performance. The table 3 records the performance of single classifiers against different performance metrics.

Table 3. Performance of Individual Classifiers

Performance metrics	REPT	CART	ADT	RF
TTBM(sec)	34.76	23.55	12.67	9.56
MAE	0.38	0.41	0.45	0.32
RMSE	0.41	0.39	0.42	0.43
RAE	84.12	72.12	67.72	68.82
RRSE	87.33	88.45	96.23	97.12
MCC	0.86	0.90	0.88	0.92
Accuracy (%)	87.50	90.88	89.90	91.04
Precision(%)	87.00	90.00	88.90	91.02
Sensitivity(%)	89.10	89.99	88.78	91.00
Specificity(%)	88.10	88.19	87.27	90.80
F1 measure(%)	88.03	89.99	88.78	91.01

Table 3 depicts that RF classifier has taken the least time (9.56 seconds) for building the model. Figure 4 shows the different error measures of the individual classifiers. It reveals that RF has a very minimum of MAE and RMSE of 0.32

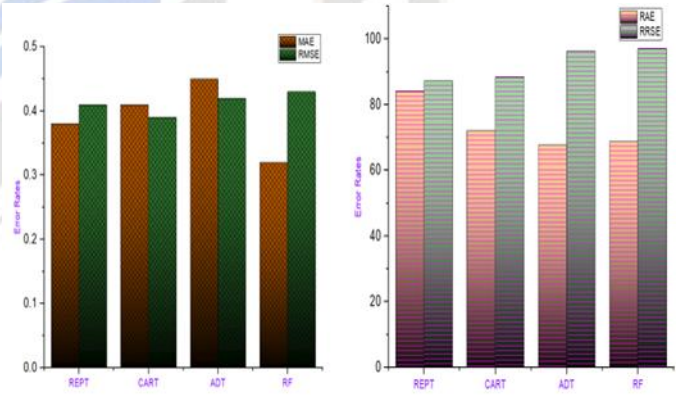

Figure 4. Error rates of Single Classifiers

RF outperforms with a precision of 91.02%, sensitivity of 91.00%, specificity of 90.80% and f1 score of 91.01%. REPT has shown the least performance with a f1 score of 88.03%. Figure 5 depicts the performance of single classifiers against precision, specificity, sensitivity, and f1 measure.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7211
***Article Received**: 29 April 2023 **Revised:** 17 June 2023 **Accepted:** 29 June 2023*

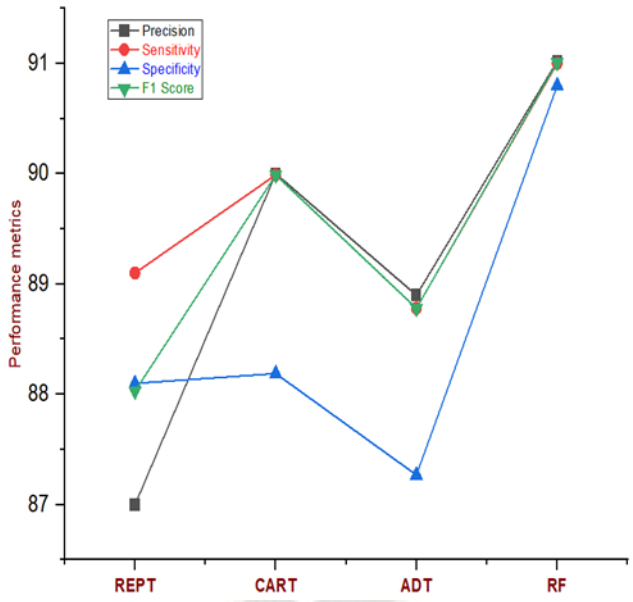

Figure 5. Performance of single classifiers

Compared to all the individual classifiers, RF outperforms with an accuracy of 91.04% and REPT has the least accuracy of 87.50% as shown in figure 6.

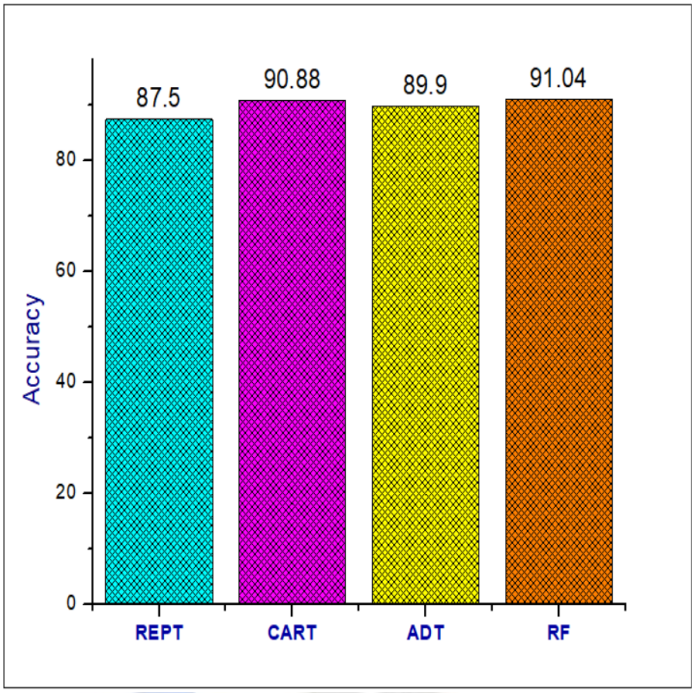

Figure 6. Accuracy of Single classifiers

The table 4 shows the performance of heterogeneous classifiers against different performance metrics. CART-RF has taken 40.66 seconds to build the model, however CART-ADT has taken 100.24 seconds to build the model.

Table 4. Performance of Heterogeneous classifiers

Performance metrics	REPT-CART	REPT-ADT	REPT-RF	CART-ADT	CART-RF	ADT-RF
TTBM (sec)	52.24	45.65	78.88	100.24	40.66	88.90
MAE	0.89	0.78	0.66	0.88	0.91	0.92
RMSE	0.56	0.67	0.78	0.79	0.89	0.90
RAE	86.10	67.87	77.90	69.87	82.12	88.18
RRSE	66.89	67.89	76.89	77.88	81.12	80.12
MCC	0.81	0.78	0.67	0.66	0.59	0.58
Accuracy (%)	88.12	89.24	90.18	89.14	92.12	91.90
Precision(%)	87.12	88.98	89.89	89.02	91.12	90.87
Sensitivity(%)	87.10	88.56	89.12	88.70	90.88	90.94
Specificity(%)	86.12	88.54	88.90	87.78	89.78	90.04
F1 measure(%)	87.11	88.77	89.50	88.86	90.99	90.50

From table 4, it is clear that CART-ADT showed poor performance as it has taken 100.24 seconds to build the model and also, error rates of MAE, RMSE, RAE and RRSE are 0.92,0.90,88.18 and 80.12 respectively as shown in the figure 7.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7211
***Article Received**: 29 April 2023 **Revised:** 17 June 2023 **Accepted:** 29 June 2023*

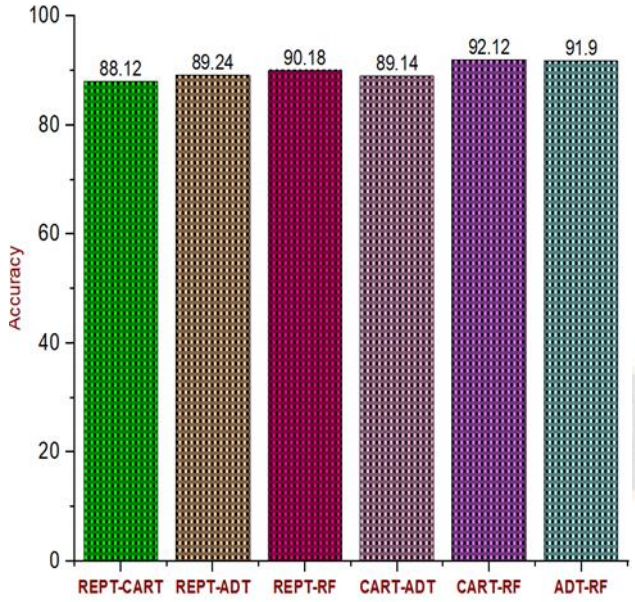

Figure 7. Error Rates of heterogeneous classifiers

Figure 8 shows the performance of different heterogeneous classifiers. It is observed that ADT-RF classifier has recorded the highest precision, sensitivity and specificity of 90.87%, 90.94% and 90.04 respectively. REPT-CART has recorded the least values.

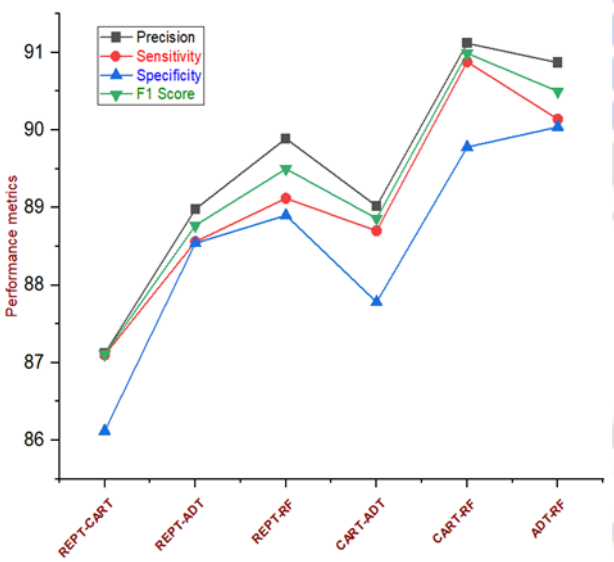

Figure 8. Performance of heterogeneous classifiers

The accuracy of REPT-ADT, REPT-RF, CART-ADT and ADT-RF are 89.24%, 90.18%, 89.14% and 91.90 respectively as shown in figure 10. CART-RF has outperformed with an accuracy of 92.12 %. The accuracy of REPT-CART has exhibited poor performance with an accuracy of 88.12%. However, the performance of heterogeneous classifiers is lesser when compared to that of single classifiers performance.

Table 5 shows the performance of homogeneous Adaboost classifiers against different performance metrics. Overall, the performance of homogeneous Adaboost classifiers is better compared to that of individual classifiers as well as heterogeneous classifiers.

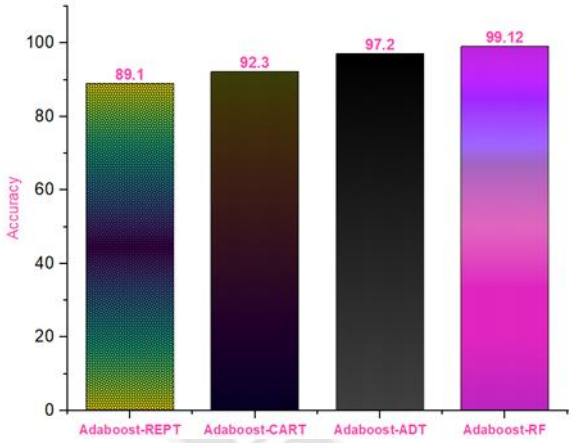

Figure 9. Comparison of Accuracy of Homogeneous classifiers

Table 5. Performance of Homogeneous classifiers

Performance metrics	Adaboost-REPT	Adaboost-CART	Adaboost-ADT	Adaboost-RF
TTBM (sec)	30.12	22.52	10.54	2.12
MAE	0.30	0.27	0.26	0.18
RMSE	0.36	0.35	0.34	0.32
RAE	44.14	33.24	23.43	12.24
RRSE	56.34	54.87	51.12	34.24
MCC	0.89	0.92	0.97	0.99
Accuracy (%)	89.10	92.30	97.20	99.12
Precision(%)	88.90	91.10	96.22	99.20
Sensitivity(%)	89.00	91.22	97.10	99.02
Specificity(%)	88.90	91.00	96.90	98.89
F1 measure(%)	88.94	91.16	96.66	99.11

The results from table 5 indicates that MAE and RMSE error rates reported for homogeneous Adaboost-RF are 0.18 and 0.32 respectively, which are very less compared to other models as shown in figure 10. The MAE, RMSE, RAE and RRSE values of Adaboost-RF are 0.18, 0.32,12.24 and 34.24 respectively. Adaboost has outperformed compared to all other models

Adaboost-RF has taken as least as 2.12 seconds to build the model. It is observed that the Adaboost-RF classifier has recorded the highest precision, sensitivity as well as specificity

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7211
***Article Received**: 29 April 2023 **Revised:** 17 June 2023 **Accepted:** 29 June 2023*

of 99.20%,99.02%,98.89% and 99.11% respectively as shown in figure 11.

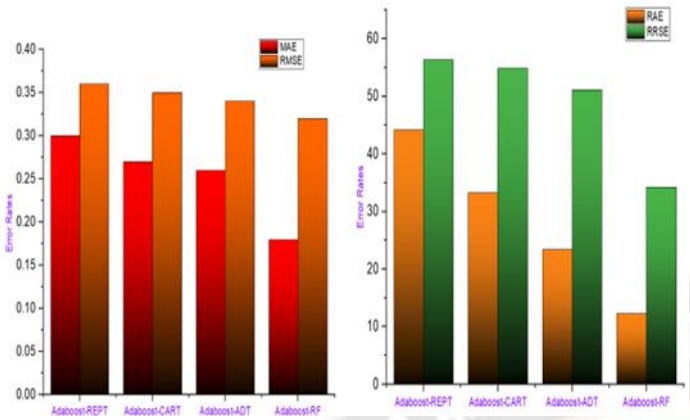

Figure 10. Error Rates of homogeneous Adaboost classifiers.

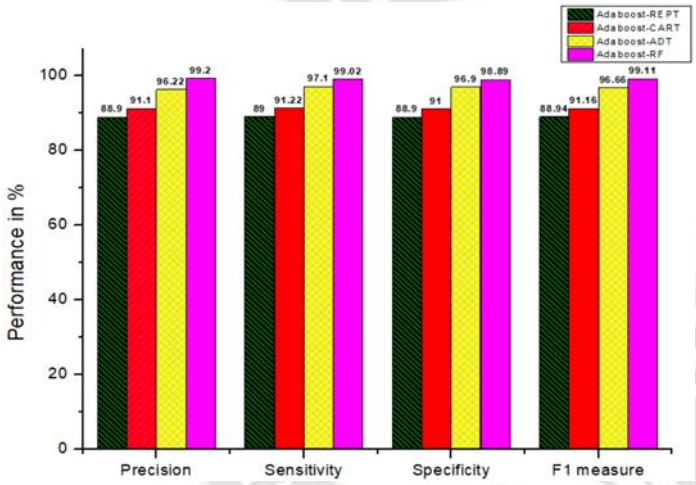

Figure 11. Performance of homogeneous Adaboost classifiers

Adaboost-RF has outperformed with 99.12% accuracy and 99.11% f1 score. Adaboost-REPT has exhibited poor performance as it has taken 30.12 seconds to build the model and achieved an accuracy of 89.10%.

V. CONCLUSION

In this proposed study, both homogeneous and heterogeneous classifiers are suggested to help with accurate patient diagnosis and CKD disease prediction. Using a publicly accessible UCI ML repository, we assessed the performance of the proposed ensemble technique and base classifiers to test the predictive or diagnostic performance of the model to accurately identify CKD data. In order to balance the data on the training dataset during pre-processing, we used the SMOTE approach. We then chose 14 features based on feature relevance. The algorithm's performance is then independently evaluated on the unseen test data after 10-fold cross-validation. Compared to all the classifiers, Adaboost-RF, a homogeneous Adaboost classifier has shown the best performance with 99.12% accuracy and 99.11% f1 score. Adaboost-RF has taken as less as only 2.12 seconds to build the model.

Funding: This research received no external funding.

Data Availability Statement: The dataset used for the findings is included in the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1] Centers for Disease Control and Prevention (2017). Chronic Kidney Disease Surveillance Systems. Accesses March 9, 2017 http://www.cdc.gov/ckd

[2] Ramakrishna, M.T.; Venkatesan, V.K.; Izonin, I.; Havryliuk, M.; Bhat, C.R. Homogeneous Adaboost Ensemble Machine Learning Algorithms with Reduced Entropy on Balanced Data. Entropy 2023, 25, 245. https://doi.org/10.3390/e25020245

[3] Disease K. Improving global outcomes (kdigo) transplant work group. kdigo clinical practice guideline for the care of kidney transplant recipients. American Journal of Transplantation . 2009;9(3):S1–S155. doi: 10.1111/j.1600-6143.2009.02834.x.

[4] Palma-Mendoza R. J., Rodriguez D., Marcos L. D. Distributed relieff-based feature selection in spark. Knowledge and Information Systems . 2018;57(1):1–20. doi: 10.1007/s10115-017-1145-y. [CrossRef] [Google Scholar]

[5] Nassar M., Safa H., Mutawa A. A., Helal A., Gaba I. Chi squared feature selection over Apache spark. Proceedings of the 23rd International Database Applications & Engineering Symposium; June 2019; Athens Greece. pp. 1–5. [CrossRef] [Google Scholar]

[6] Charleonnan A., Fufaung T., Niyomwong T., Chokchueypattanakit W., Suwannawach S., Ninchawee N. Predictive analytics for chronic kidney disease using machine learning techniques. Proceedings of the 2016 management and innovation technology international conference (MITicon); October 2016; Bang-San, Thailand.

[7] Chittora P., Chaurasia S., Chakrabarti P., et al. Prediction of chronic kidney disease-a machine learning perspective. IEEE Access . 2021;9(17312):p. 17334. doi: 10.1109/access.2021.3053763.

[8] T. R. M, Vinoth Kumar V, Lim S-J (2023) UsCoTc: Improved Collaborative Filtering (CFL) recommendation methodology using user confidence, time context with impact factors for performance enhancement. PLoS ONE 18(3): e0282904. https://doi.org/10.1371/journal.pone.0282904

[9] H. Polat, H. D. Mehr, and A. Cetin, Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods, J. Med. Syst., vol. 41, no. 4, p. 55, Apr. 2017.

[10] S. Bashir, U. Qamar, F. H. Khan, and M. Y. Javed, MV5: A Clinical Decision Support Framework for Heart Disease Prediction Using Majority Vote Based Classifier Ensemble, Arab. J. Sci. Eng., vol. 39, no. 11, pp. 77717783, Nov. 2014.

[11] D. Devarajan D Stalin Alex, T R Mahesh, Rajanikanth Aluvalu, V Vinoth Kumar, Uma Maheswari, S Shitharth, "Cervical Cancer Diagnosis Using Intelligent Living Behaviour of Artificial

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7211
***Article Received**: 29 April 2023 **Revised:** 17 June 2023 **Accepted:** 29 June 2023*

Jellyfish Optimized with Artificial Neural Network," in IEEE Access, 2022, https://doi.org/10.1109/ACCESS.2022.3221451

[12] I. Pritom, M. A. R. Munshi, S. A. Sabab, and S. Shihab, Predicting breast cancer recurrence using effective classification and feature selection technique, in 2016 19th International Conference on Computer and Information Technology (ICCIT), 2016, pp. 310314.

[13] U. N. Dulhare and M. Ayesha, Extraction of action rules for chronic kidney disease using Naive Bayes classifier, in 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2016, pp. 15.

[14] Chittora P., Chaurasia S., Chakrabarti P., et al. Prediction of chronic kidney disease-a machine learning perspective. IEEE Access . 2021;9(17312):p. 17334. doi: 10.1109/access.2021.3053763

[15] Avci E., Karakus S., Ozmen O., Avci D. Performance comparison of some classifiers on chronic kidney disease data. Proceedings of the 2018 6th International Symposium on Digital Forensic and Security (ISDFS); March 2018; Antalya, Turkey. pp. 1–4.

[16] Abdullah A. A., Hafidz S. A., Khairunizam W. Performance comparison of machine learning algorithms for classification of chronic kidney disease (ckd) Journal of Physics: Conference Series . 2020;1529(5) doi: 10.1088/1742-6596/1529/5/052077.052077

[17] Jena L., Patra B., Nayak S., Mishra S., Tripathy S. Intelligent and Cloud Computing . New York, NY, USA: Springer; 2021. Risk prediction of kidney disease using machine learning strategies; pp. 485–494.

[18] Jongbo O. A., Adetunmbi A. O., Ogunrinde R. B., Ajisafe B. B. Development of an ensemble approach to chronic kidney disease diagnosis. Scientific African . 2020;8 doi: 10.1016/j.sciaf.2020.e00456.e00456

[19] Wibawa M. S., Maysanjaya I. M. D., Putra I. M. A. W. Boosted classifier and features selection for enhancing chronic kidney disease diagnose. Proceedings of the 2017 5th international conference on cyber and IT service management (CITSM); August 2017; Denpasar, Indonesia. pp. 1–6.

[20] Venkatesan, V.K.; Ramakrishna, M.T.; Batyuk, A.; Barna, A.; Havrysh, B. High-Performance Artificial Intelligence Recommendation of Quality Research Papers Using Effective Collaborative Approach. Systems 2023, 11, 81. https://doi.org/10.3390/systems11020081

[21] "UCI Machine Learning Repository" [Online] Available:https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease.

[22] Ifraz, G.M.; Rashid, M.H.; Tazin, T.; Bourouis, S.; Khan, M.M. Comparative Analysis for Prediction of Kidney Disease Using Intelligent Machine Learning Methods. Comput. Math. Methods Med. 2021, 2021, 6141470

[23] S. Y. Yashfi et al., "Risk Prediction Of Chronic Kidney Disease Using Machine Learning Algorithms," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-5, doi: 10.1109/ICCCNT49239.2020.9225548.

[24] Venkatesan, V.K.; Ramakrishna, M.T.; Izonin, I.; Tkachenko, R.; Havryliuk, M. Efficient Data Preprocessing with Ensemble Machine Learning Technique for the Early Detection of Chronic Kidney Disease. Appl. Sci. 2023, 13, 2885. https://doi.org/10.3390/app13052885

[25] Kovesdy, C.P.; Matsushita, K.; Sang, Y.; Brunskill, N.J.; Carrero, J.J.; Chodick, G.; Hasegawa, T.; Heerspink, H.L.; Hirayama, A.; Landman, G.W.; et al. Serum potassium and adverse outcomes across the range of kidney function: A CKD Prognosis Consortium meta-analysis. Eur. Heart J. 2018, 39, 1535–1542.

[26] N. K. Baskaran and T. R. Mahesh, "Performance Analysis of Deep Learning based Segmentation of Retinal Lesions in Fundus Images," 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2023, pp. 1306-1313, doi: 10.1109/ICEARS56392.2023.10085616.

[27] M. T R and G. G, "MRI techniques using image processing Brain Tumor Detection," 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), Greater Noida, India, 2023, pp. 477-480, doi: 10.1109/AISC56616.2023.10085289.

[28] T. R. Mahesh, V. Vivek, V. V. Kumar, R. Natarajan, S. Sathya and S. Kanimozhi, "A Comparative Performance Analysis of Machine Learning Approaches for the Early Prediction of Diabetes Disease," 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), 2022, pp. 1-6, doi: 10.1109/ACCAI53970.2022.9752543

[29] P. Shrestha, A. Singh, R. Garg, I. Sarraf, T. R. Mahesh and G. Sindhu Madhuri, "Early Stage Detection of Scoliosis Using Machine Learning Algorithms," 2021 International Conference on Forensics, Analytics, Big Data, Security (FABS), 2021, pp. 1-4, doi: 10.1109/FABS52071.2021.9702699

[30] K. K. Jha, A. K. Jha, K. Rathore and T. R. Mahesh, "Forecasting of Heart Diseases in Early Stages Using Machine Learning Approaches," 2021 International Conference on Forensics, Analytics, Big Data, Security (FABS), 2021, pp. 1-5, doi: 10.1109/FABS52071.2021.9702665.