Panamerican Mathematical Journal
ISSN: 1064-9735
Vol 35 No. 2s (2025)

A Novel Approach Using Artificial Intelligence for Early Prognosis of Chronic Kidney Disease in Asian Population

Anindita Khade[1], Amarsinh Vidhate[2], Siddhant Jaiswal[3], Dr. Sunil V. Prayagi[4], V. Preethi[5], Mal Hari Prasad[6]

[1]Department of Computer Engineering, School of Technology Management and Engineering, SVKM'S NMIMS Deemed to be University, Navi Mumbai, aninditaac1987@gmail.com

[2] Department of Computer Engineering, RAIT, DY Patil Deemed to be University, Navi Mumbai.

amar.vidhate@rait.ac.in

[3]School of Computer Science and Engineering, Ramdeobaba University (RBU), Nagpur, India.
siddhantjaiswal5@gmail.com

[4]Department of Mechanical Engineering, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India. sunil_prayagi@yahoo.com

[5]Asst.Professor , Dept.of ECE, ADITYA UNIVERSITY, SURAMPALEM, INDIA.

joeljason1987@gmail.com

[6]Assistant Professor, Department of Artificial Intelligence, Anurag University, Venkatapur, Ghatkesar

hari.mhp1106@gmail.com

Article History:

Received: *21-09-2024*

Revised: *30-10-2024*

Accepted: *14-11-2024*

Abstract:

Purpose: The goal of this research project is to address the increasing prevalence of chronic kidney disease (CKD) and the challenges associated with its early detection. The study proposes a new approach that combines Linear Discriminant Analysis (LDA) to reduce the number of features and Artificial Neural Networks (ANN) for accurate and early identification of CKD. The primary objective is to create a smart decision-making system that can help nephrologists in India diagnose CKD in the early stages.

Method: The study is based on a dataset collected from DY Patil Hospital in Navi Mumbai. It consists of around 500 records with 21 attributes. To improve accuracy and reduce prediction time, the proposed model combines LDA and ANN. Feature selection is carried out using Recursive Feature Elimination (RFE), which identifies Creatinine, BUN, and Urea as crucial factors. The methodology includes preprocessing, hyperparameter tuning, and classification. Statistical analyses, such as hyperparameter values, Friedman's Test, and parallel computing evaluation, are used in the comprehensive methodology.

Results: The proposed hybrid model, referred to as a Hybridized LDA with ANN (HLDANN), outperforms traditional classifiers like SVM, LR, RF, DT, KNN, and even a standalone ANN. The model achieves an accuracy of 93.22% on a real-time dataset, surpassing other algorithms. Precision, recall, and F1 score metrics further validate the effectiveness of HLDANN. Parallel computing analysis demonstrates reduced prediction time with an increase in worker nodes.

Conclusion: Based on the study, the hybrid model is an effective method for detecting CKD at an early stage. It provides significant accuracy improvements compared to existing methodologies. The key features identified in the study align

with known biomarkers, confirming the model's reliability. To further enhance the model's robustness, future research could explore additional biomarkers, diverse data sets, and non-invasive techniques. The proposed algorithm shows promise as a valuable tool for the medical community that could contribute to early CKD diagnosis and improved patient outcomes.

Keywords: Chronic Kidney Disease, Glomerular Filtration Rate, Artificial Neural Networks, Linear Discriminant Analysis, Recursive Feature Elimination.

1. Introduction

CKD is a global public health issue with an increasing prevalence. CKD leads to difficulties in extracting waste products from the body and in case this worsens, these wastes get mixed with blood and may develop multiple complications like heart diseases, fragile bones, diabetes, etc. Loss of the kidney function leads to some devastating symptoms like abdominal pain, diarrhea, nausea, etc. In the general population, the prevalence is estimated to be between 11 and 13 percent, and it is linked to an elevated risk of cardiovascular disease and mortality when compared to the non-CKD population. CKD, on the other hand, can be asymptomatic in its early stages and is frequently misdiagnosed. The prevalence of this kind of disease is rising in developed as well as developing countries, which is something that needs to be addressed [1]. The hospitalization rate for such a disease increases every year by 6.23 percent, but the mortality rate has not changed [2]. As a result, even in rich nations like India, CKD unawareness is common. Due to this growing unawareness, it is extremely imperative to overcome this disease through early detection, monitoring, and handling of the same. Figure 1 describes the prevalence of CKD in India in the year 2021.

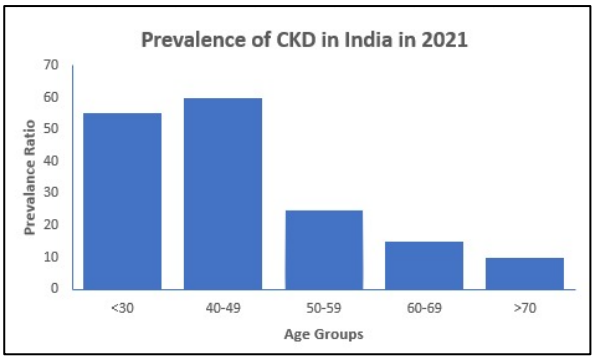

Fig. 1 Prevalence of CKD in India in 2021

Intelligent decision-making systems for early and automatic diagnosis are especially preferred to be used against such kinds of diseases due to the dearth of personnel and infrastructure required for manual sickness diagnosis [3]. Data mining methods can assist in the prediction of the most relevant risk factors that are connected to CKD by using their previous medical records. Data mining algorithms can play a big role in finding out hidden information from huge databases and may help figure out precise treatment plans. Support Vector Machine (SVM), Artificial Neural Networks (ANN), Logistic Regression (LR), Decision Tree (DT), Extreme Gradient Boosting (XGBoost), Naive Bayes (NB), and fuzzy

set theory are all important AI methods that when combined with medical experience can help detect breast cancer,tuberculosis, heart disease, diabetes, etc. CKD is one of the pivot areas of this AI-based medical diagnostics development. In CKD, the kidneys' ability to filter blood and eliminate metabolic waste gradually deteriorates. A blood test is used to detect how well the kidneys are working by looking for biomarkers like urea, creatinine, and other waste products in the blood. Glomerular Filtration Rate (GFR) is a metric that determines how much damage has been done to the kidneys [4]. This is referred to as the estimated Glomerular Filtration Rate (eGFR). The phases of CKD detected based on GFR areshown in Table 1.

Table 1 Stages of CKD

Stage	Description	GFR	Percent Kidney Function
1	Normal functioning kidney	≥ 90ml/min	≥ 90 %
2	Slight decline in kidney function	60-89ml/min	60-89 %
3A	Moderate decline in kidney function	45-59 ml/min	45-59 %
3B	Moderate decline in kidney function	30-44 mL/min	30-44 %
4		15-29 ml/min	15-29 %
5	Severe decline in kidney function Kidney Failure	≤ 15ml/min	≤ 15 %

To design an AI-based treatment for CKD, data must be collected, and evidence needs to be studied. While diagnosing any kind of disease, the clinical decision is mainly dependent on the patient's symptoms and the experience of physicians [5]. The major problem in the prediction of CKD is that the patients do not get any symptoms up to stage 3. Only when a patient tries to get tested for some other disease like hypertension, blood pressure, or diabetes, does he accidentally get to know that his kidneys are also deteriorating.
To streamline the diagnosis of CKD amidst symptoms resembling those of other diseases, it's crucial to prioritize key symptoms (features). This approach minimizes the need for extensive testing and requires efficient feature selection and reduction in data mining to pinpoint the most significant indicators [6]. Until stage 3, there are no signs or symptoms of CKD [7]. As a result, approaches that can detect this form of sickness in its early stages are required. The need for a classification system facilitates the perfect and timely diagnosis of any disease. Consequently, this improves the speed and efficiency of decision-making processes [8]. The objective of this study is to introduce a novel technique designed to swiftly identify the presence of CKD during its initial stages. The work aims to develop a project that will be beneficial to the medical community in India.

The remainder of the paper is laid out as follows. Literature reviews based on various categories are discussed in Section 2. Section 3 will discuss in detail the dimensionality

reduction techniques. Section 4 will discuss the hybridized methodology implemented in this work with the proposed model and its salient features. We discuss the results obtained and our subsequent analysis in Section 5. Analysis concerning prediction time using multiple processors is discussed in Section 6. Section 7 provides the statistical analysis results. We provide our conclusion for this work in Section 8 followed by references.

2. Related Works

Numerous intelligent methods have been employed to detect CKD in its early stages. Over the years, AI techniques have demonstrated significant utility in the medical field. The following section will outline several methodologies adopted by different researchers, categorizing them into supervised Machine Learning (ML) and Deep Learning (DL) techniques. This will culminate in a recommendation regarding the preferred AI techniques for prediction. Notably, the datasets utilized in these studies are predominantly sourced from the UCI machine learning repository, which comprises 400 records, including 250 non-CKD and 150 CKD instances, with 25 attributes. Originating from Apollo Hospitals, Tamil Nadu, India, this dataset was released in 2005. Furthermore, the literature review suggests that existing algorithms generally exhibit satisfactory accuracy levels in prediction tasks.

The authors in [9] proposed three classifiers SVM, RF, and DT which were able to diagnose CKD.KNN has been used for data assertion techniques. Amongst these classifiers, RF gives the best accuracy of 97.2%. The authors in [10] in their research implemented Support Vector Machine (SVM) and ANN for their analysis. Their findings indicated that ANN exhibited superior accuracy compared to SVM. However, their analysis also revealed that employing ANN necessitated more time in comparison to SVM. The researchers in [11] in their research used DT. The project had a near-perfect accuracy of 93%. The authors concluded that the analysis had to be based on recent evidence. The authors in [12] devised a technique for doctors to use in predicting the development of Chronic Renal Failure (CRF) in patients. On the CRF data set, the authors used Naive Bayes, KNN, tree-based classification, and random subspace algorithms. The KNN classifier gave a 94 percent accuracy rate in the random subspace according to the authors. The researchers in [13] developed a gradient-boosting-based prediction model to diagnose CKD using patients' EHR and billing data. It provided an accuracy of about 87%. To predict CKD in diabetic patients, the authors in [14] created several artificial intelligence models like Convolutional Neural Networks (CNN), ANN, Light Gradient Boosting Model (GBM), etc. Age, gout, diabetes mellitus, sulphonamide use, and angiotensin's were chosen as the most critical factors for CKD prediction by the Light GBM model. When compared to other models, CNN had the best performance of about 97%. The authors in [15] developed a novel hybrid technique for diagnosing CKD by combining the effects of SVM with Random Forest. It gave a prediction accuracy of about 92.5 percent. They explored the WEKA tool for finding accuracy. The researchers in [16] suggested an artificial neural network model for CKD diagnosis. The diagnostic sensitivity values in this study varied from 91.1 to 92.5 percent, whereas the diagnostic specificity values ranged from 90.9 to 92.2 percent. The authors used Ant Colony Optimization (ACO) for feature selection. They concluded that when they combined the

effects of ANN with ACO, they got satisfying results. The authors in [17] have used Naive Bayes, DT, and KNN for this task of classification. The findings were that Naive Bayes has a lesser precision and KNN took a longer time than other algorithms for prediction. They concluded that of all the three algorithms, DT gave the best performance. The researchers [18] in their research used the Autoencoder network for prediction. They tested it on only 100 patients and got an excellent accuracy of 100%. They suggested that it should be explored more for real-time patients. The authors [19] in their research for the detection of CKD implemented Probabilistic Neural Networks. The authors implemented the same on 361 patients' records received from a hospital and looked after stage-wise classification of patients also. The accuracy achieved was near 93%. The researchers in [20] implemented a hybrid of CNN + SVM. They worked on ultrasonography photographs of patients for their prediction. They also concluded that the hybrid model provided a sound accuracy of 88%.

From the above literature survey, we found out that different algorithms and techniques have been applied to date to detect CKD at its early stages. Few of these algorithms have given exemplary results. Genetic algorithms and fuzzy logic techniques have also been implemented to study their effect on the early detection of CKD. The deep learning algorithms gave better performance than the traditional machine learning algorithms. For our initial analysis based on a comparison of Machine Learning, Deep Learning, and Fuzzy Logic techniques, we implemented six Machine Learning and three Deep Learning algorithms on the mentioned UCI Dataset [21]. The performance of these models has been described in Fig.2. Based on the above review and analysis, we found out that Deep Learning models provide better accuracy (approx. 83%) as compared to Machine Learning models. This paper is based on further enhancements in our Deep Learning model with an intuition to increase accuracy and reduce prediction time.

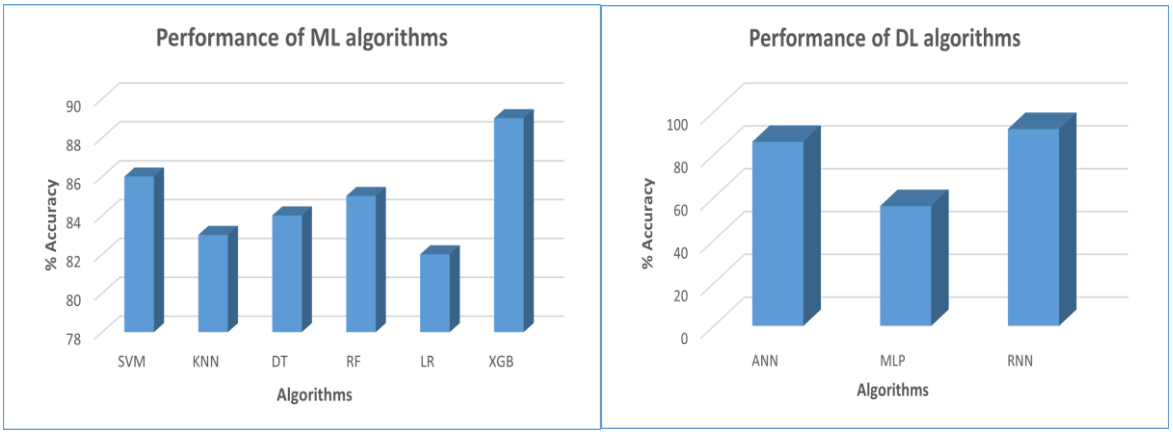

Fig. 2 Performance of standard algorithms.

3. Dimensionality Reduction Techniques

To design any efficient machine learning algorithm, the most difficult part is identifying the most important attributes in any given data set. Feature selection and Feature extraction methods extract the most significant features and eliminate unrelated features from the dataset for enhancing classification performance [21]. These types of methods in turn help

lower computational costs and improve results. Machine learning, data mining, bioinformatics, biometrics, and information retrieval all have benefitted from dimensionality reduction techniques [22]. Dimensionality reduction solutions are designed to minimize dimensions by eliminating duplicate and dependent elements and transferring them from a higher dimensional environment, which could lead to a curse of dimensionality, to a space with less dimensions. This research work focuses on using LDA for dimensionality reduction. It captures attribute interaction as well as local interaction between features. This approach can also handle multiclass problems and is resistant to noisy and partial data. The LDA approach is used to convert the features into a shallow-dimensional space that enhances the proportion of between-class disparity to within-class variance, ensuring utmost class separation [22]. LDA can be categorized into two types. In class-dependent LDA, each class possesses its unique shallow-dimensional space for projecting its data, whereas in class-independent LDA, each class is considered distinct from the others. In this variant, all classes project their data onto a single shallow-dimensional space. The LDA methodology operates by converting the original data matrix into a lower-dimensional space [22]. The LDA approach works by projecting the original data matrix into a shallow dimensional environment. The entire process of dimensionality reduction has been explained in Figure 3.

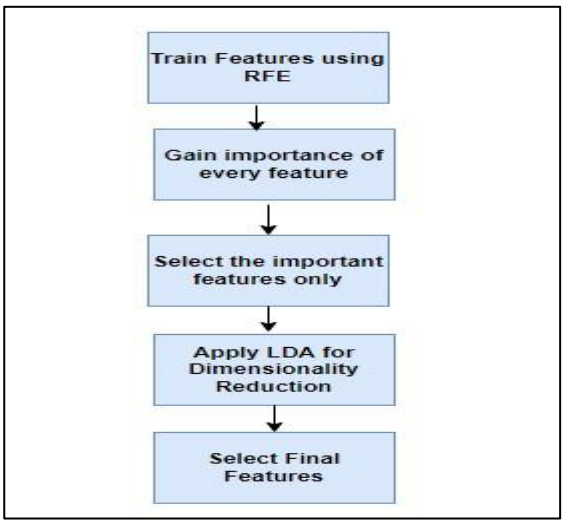

Fig. 3 Dimensionality Reduction Technique

4. Proposed Model Hybridized LDA-Based ANN(HLDANN)

In this paper, we propose a hybrid model which has been built to predict CKD in its early stages. The main idea of the hybrid model is to combine the features of LDA and ANN which will create a more robust architecture. The benefits of dimensionality reduction along with the aid of ANN will help in giving more accurate and faster results. The test results show that the proposed model has a sound accuracy and takes less time in prediction as compared to the other algorithms discussed.

Salient Features of the proposed model.

1. Less complicated design.
2. Better Accuracy
3. Faster prediction due to implementation of dimensionality reduction techniques with ANN.
4. Focus on important features only by using Feature Extraction Techniques.

4.1 Selection of LDA over PCA

PCA aims to uncover the primary directions of variance within a dataset as part of its unsupervised dimensionality reduction process. The goal is to identify a more manageable subset of variables or traits that best represent the salient patterns in the data. Preprocessing data for machine learning algorithms is a common application of PCA. The goal of LDA, a supervised technique for dimensionality reduction, is to identify the linear feature combination that best divides a dataset's classes. The goal is to make the data less dimensional while keeping the information that matters most for class differentiation. While LDA is widely used for feature selection and classification, PCA is frequently used for exploratory data analysis and data preprocessing for machine learning algorithms. Due to the above reasons, the above research uses LDA over PCA for feature extraction.

4.2 Study Design and Dataset

This is an analytical study that has been conducted on the patients of DY Patil Hospital Navi Mumbai, for 2.5 years. We collected approximately 500 records from the patients who came to be tested for CKD. These had 250 CKD and 250 Non-CKD records. It consisted of around 21 attributes that are usually used by hospitals/clinicians to check whether a patient has CKD or not. A list of such parameters and their ranges has been provided in Table 2.

Table 2: Details of CKD Dataset

Features	Units	Ranges
age	-	25-60
gender	-	M-Male F-
vol	ml	Female
sg	mg/dl	2-5 ml
freq	ml	0-1.25
sod	mg/dl	0-5
pot	mEq/L	0-163
chlo	mEq/L	0-47
phos	mEq/L	0-76
prot	gms	0-83
alb	gms	0-9
glob	gms	0-5
urea	mgs/dl	0-7
creatinine	mgs/dl	4.75-183.3

bun	mgs/dl	0.03-11.4
uricacid	mEq/L	0-165
rbc	millions/cmm	0-391
wbc	cells/cumm	5.0-9.0
pcv	cells/cumm	6.0-11.0
pe	-	0-54
ane	-	yes,no
classification	-	yes,no ckd/notckd

A line graph was plotted to understand the amount of these biochemical parameters in the blood stream of the CKD as well as Non-CKD patients. This graph helped us in identifying major parameters for decision making. Out of the 21 major parameters, we found out that the major 3 parameters for predicting CKD were Urea, BUN, and Creatinine. This has been evidently presented in Fig 4 ,5, 6 and 7.

Since usually patients having diabetes and hypertension tend to have more pressure on their kidneys, we targeted their records for our analysis. Keeping this thing in mind, we aim to predict CKD at least 6 months before a person enters stage 3 of CKD.

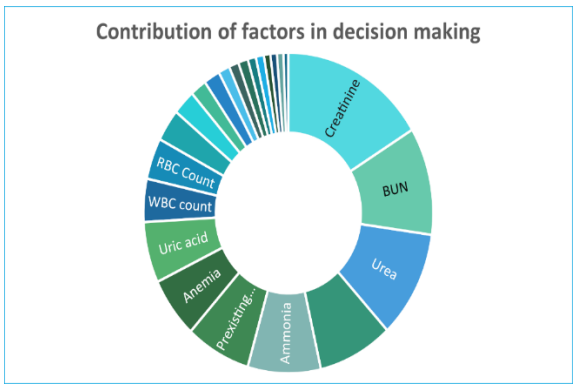

Fig. 4 Contributing Factors in Decision Making

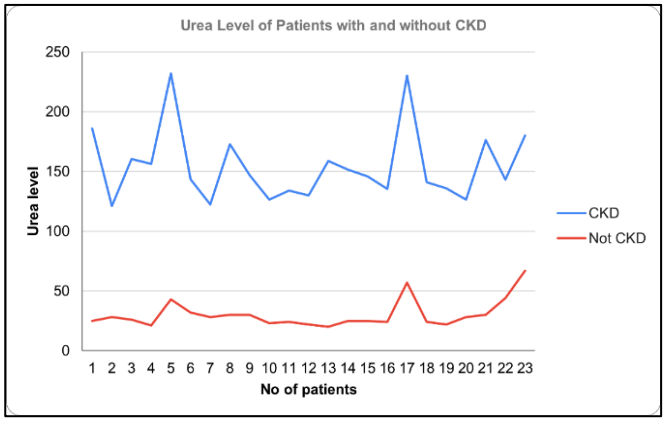

Fig. 5 Urea Levels of Patients with and without CKD

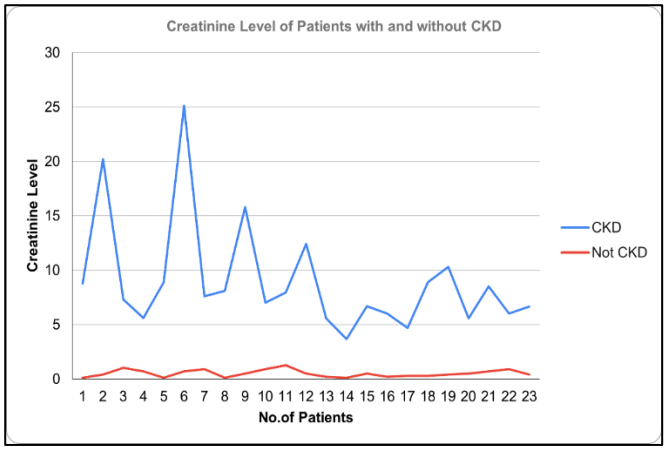

Fig. 6 Creatinine Levels of Patients with and without CKD

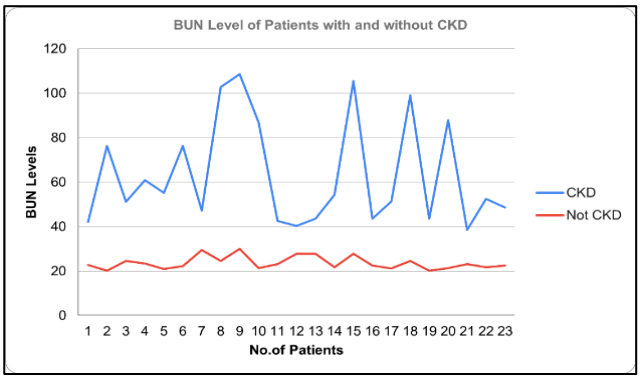

Fig. 7 BUN Levels of Patients with and without CKD

4.3 Data Pre-processing

The preparation phases included the estimation of missing values, the normalization of data, and the elimination of noise. Some data received from the hospital was incomplete, so by applying imputation methods, we tried to fill them.

4.3.1 Handling Missing Values

124 records were found to have some missing attributes. The easiest method for dealing with missing values is to ignore records; however, this is not a good solution for tiny data sets, such as ours. During the data preparation process, the data set is inspected tosee whether any of the attribute values are not present. The missing tuple values were scaled using statistical methods of median imputation. This was possible only for numerical data. According to us, the data received was very skewed, hence we decided to use the median approach for replacing missing values.

4.3.2 Categorical Data Encoding

Categorical values must be encoded into number values because this will be helpful for the ML algorithms for their analysis. The binary numbers “0” and “1” are used to indicate the features of categories like “no” and “yes,” respectively.

4.3.3 Data Transformation

The technique of altering values on the same scale so that normalization can be maintained throughout data is called data transformation [23]. Regardless of the unit of weight, machine learning algorithms interpret higher values as greater and lower values as lesser. Data transformations change the values in a data set to allow for additional processing. This study uses a data normalization strategy to increase the accuracy of machine learning models [24]. This process standardizes the data by rescaling it to a range of -1 to +1 with a mean of 0 and standard deviation of 1.

4.3.4 Outlier Detection

Outliers are anomalous observations that stand out from the remaining data. An outlier could be the result of the error or measurement variability. An outlier can mislead the machine learning algorithm's learning process. It results in longer training times, lower model accuracy, and, ultimately, poorer results [25]. This study uses the Interquartile Range (IQR) technique to remove outliers before feeding data into the learning algorithm.

4.3.5 Feature Selection Process

Feature nomination means selecting only those sub-features for training that may be helpful in the prediction of CKD. Usually, training a model on a few selected features for prediction is always better than training on all features. The benefits of Feature selection are as follows:

1. The machine learning algorithm can train more quickly as a result.
2. A model becomes less complicated and is simpler to interpret as a result.
3. If the proper attribute gets selected, a model’s accuracy increases.
4. Overfitting is decreased.

In this work, we tried to implement the Recursive Feature Elimination Technique. As a result, we explored three possible methods:

1. Forward Selection: Here, we start with no model being added to the process. With every iteration, we add one feature at a time and keep on analyzing till the performance improves.

2. Backward Elimination: In this method, the model starts with all features and tries to remove the least significant feature during each iteration. This process continues up to the time there is no change in performance metrics even after the removal of any feature [26].

3. Recursive Feature Elimination (RFE): In this, it repeatedly finds the best and the worst performing subsets after each iteration. It follows the greedy approach to find an optimal solution. It creates the feature sets every time based on left-out features. It then selects the best features out of the given ones. We selected RFE for our project since the method was very easy to configure, and we found it to be computationally less expensive as compared to the other two wrapper methods.

Regarding the given research, the RFE algorithm has proved Blood Urea Nitrogen (BUN), Serum Creatinine, and Urea to be the most important features. These selected features are

depicted in Figure 8.

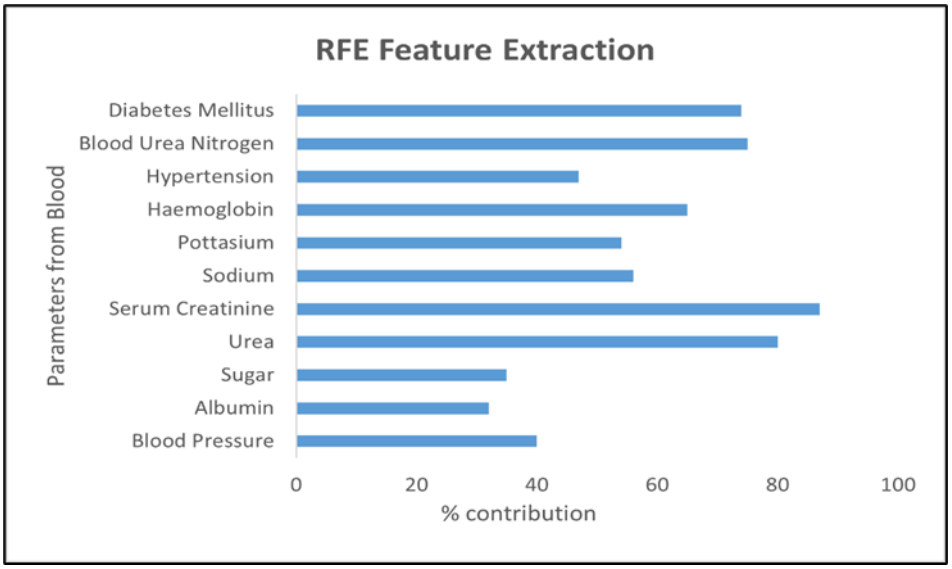

Fig. 8 Important Features

4.4 Methodology

The suggested model has been illustrated in Figure 9.

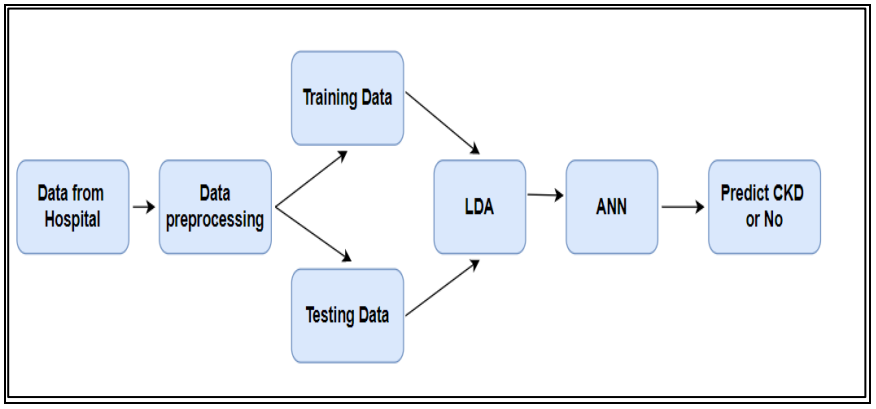

Fig. 9 Proposed Model

With an intent to receive better accuracy for our model, we tried enhancing the model by applying Dimensionality Reduction and Feature Elimination techniques. In this approach, we have tried to implement LDA and RFE over ANN. Pre-processing, model training and hyper-tuning, and label classification are the three stages of the proposed model. The CKD data set is split into training and testing data sets after applying fundamental pre-processing techniques. The percentages are set as 80% and 20%, respectively. Only a few of the study's 21 features were selected via RFE. The processing complexity of the approach is decreased by the RFE algorithm, which values each feature according to its significance. Finally, attributes that are redundant or unimportant are eliminated. The learning model is fed the most crucial traits. LDA has been added on top of ANN By addition of this layer, the total number of features used for decision-making has now been reduced to 3 instead of 11. A list of such features has been specified in Table 3. In essence, this is done so that we can advance

toward intra-class low variance and inter-class high variance without penalizing the misclassifications [27]

Table 3: Number of features selected by RFE and RFE+ LDA

Sr. No	RFE	RFE+LDA
1	dm	urea
2	sod	creatinine
3	pot	bun
4	alb	
5	urea	
6	creatinine	
7	bun	
8	ht	
9	hgb	
10	su	
11	bp	

The ANN receives the results of this layer. The suggested model has 5 layers: 1 input layer, 2 hidden layers, 1 dropout layer, and 1 output layer. The dropout layer is typically added to neural networks to lessen the impact of underfitting. The two concealed levels are followed by the introduction of this layer. Dropout rates for the drop layer are 0.2. Since LDA and RFE were already in place when ANN was introduced, there were now just 3 neurons in the input layer of ANN as opposed to 23 neurons previously. The hyperparameters must be used to optimize any ANN. Our network uses the "adam" optimizer [28]. When it comes to offering an ideal gradient descent, the Adam optimizer performs far better than many other algorithms. This is so because the effects of the two separate gradient descent approaches can be combined. Because of its simple and efficient classification approach, the Rectified Linear Unit (ReLU) stands out as the most widely recognized activation function in deep learning models [30]. The trials' findings show that the representations made using the suggested approach are discriminative and improve classification accuracy. When implementing real-time data prediction using shallow ANN (ANN with a single hidden layer), the prediction received was near 87.32 percent. As discussed above,we implemented a heterogeneous model by implementing the features of RFE and LDA above ANN. As a result, the ANN was trained and subsequently tested only for the selected features. Due to this, we received a slightly high accuracy of approximately 91 percent. Since the dataset received was quite unbalanced, we also planned to check the precision, recall, and F1 score as our performance metrics. The entire process has been explained using the below algorithm.

Algorithm: Hybridized LDA with ANNInput: A dataset consisting of 200 real-time records

Output: Classification: Whether CKD or Not CKD

Algorithm:

1. Provide the dataset to the Feature extraction module.

2. Out of the 21 features provided, the module returns the most important 11 features.

2.1 for Each subset Pi, i=1 to S

Keep the Pi most important VariablesPreprocess the data

Train the model on the training set using Pi predictorsCalculate the model's performance

Recalculate the rankings of each predictorEnd

2.2 Calculate the performance profile and determine the number of predictors

3. Feed these predictors to the LDA module for dimensionality reduction

3.1 The mean (m) value of each input (x) for each class (z) can be estimated as:

$$\boldsymbol{m = 1 \,/\, nz * \Sigma \cdot sum(x)} \quad \textit{(4.4.1)}$$

where nz is the total number of instances in class k.

3.2 Using Bayes' theorem, we can show that the base probability (Pk) of each class(z) canbe calculated as

$$\boldsymbol{Pk = nz/n} \quad \textit{(4.4.2)}$$

3.3 Based on the above data, the discriminant function can be considered as:

$\boldsymbol{Dk(x) = x * (m/var) - (m\text{^}2/(2 * var)) + ln(Pk)}$ *(4.4.3)Where Dk(x) is the discriminant function and var is the variance respectively.*

4. The output which consists of only 3 important features now is fed to the ANN

We take the input variables and the Neural Network equation of

$$Z = W0 + W1X1 + W2X2 + \ldots + WnXn$$

to compute the final output or the predicted Y values, called the Ypred.W0 is called thebias

We apply ***the ReLu*** *activation function and* ***adam*** *optimizer*

5. *Ypred* gives us the final prediction of whether a person is having CKD or No.

5. Results and Discussions

The proposed model has been tested for accuracy, precision, recall, and prediction time. Figure 10 represents the heatmap labelling the correlations between the attributes.

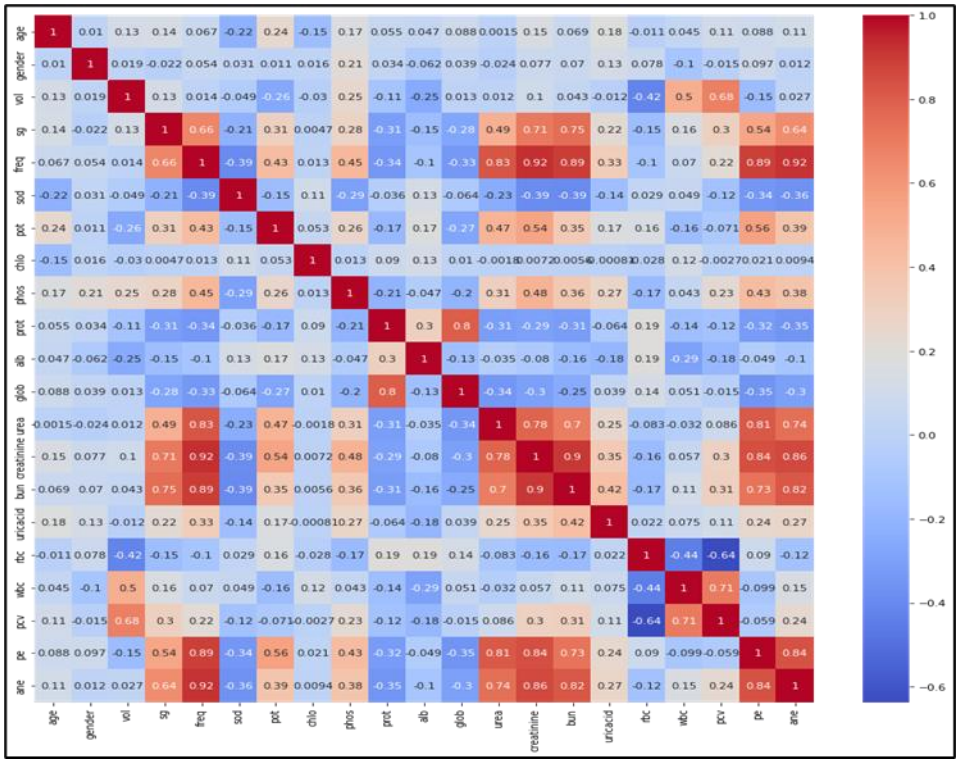

Fig. 10 Heatmap

Hyperparameters are variables that cannot be learned directly during the training process [33]. The adjustment of hyperparameters enables excellent ANN classifier accuracy while simultaneously lowering computing costs. Table 3 depicts the hyperparameter tuning implemented for the proposed model. We implemented all the above-mentioned algorithms and our proposed hybrid model on the dataset described above. The optimal parameter values are discovered using a grid search and 10-fold cross-validation [31]. The results are quite impressive since the proposed model gives a better accuracy as compared to the other models. The comparative analysis of these models is described in Table 4.

Table 3: Hyperparameters

Hyperparameters	Values
Epochs	30
Batch Size	15
Activation function	‘ReLu’
Optimizer	‘adam’

The model is implemented on GPU processors and its performance has been tested first using 2 and then using 4 worker nodes. This is done to understand the intent of parallel processing that can be involved. The proposed model is compared with different classifiers like SVM, LR, DT, RF, KNN, and ANN.

Table 4: Performance Analysis

Algorithms	Accuracy in %
SVM	78.26%
KNN	72.21%
LR	80.23%
DT	83.1 %
RF	85.34%
ANN	87.32%
HLDANN	93.22%

We have also compared the precision, recall, and F1 scores of all the above-mentioned algorithms. All the 3 parameters had sound results for the proposed hybrid model as compared to the other models. Figure 11 describes the same.

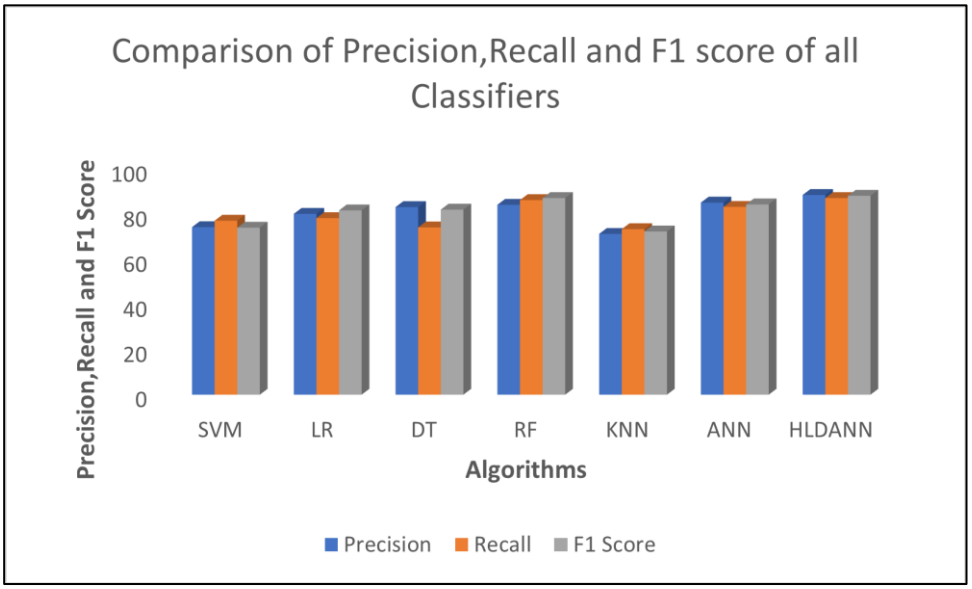

Fig.11 Comparison of precision, recall F1 score

Fig. 12 represents the proposed model's accuracy and loss. This shows that the accuracy of the proposed HLDANN model is very promising. Fig 13 represents the AUC and ROC of the proposed model.

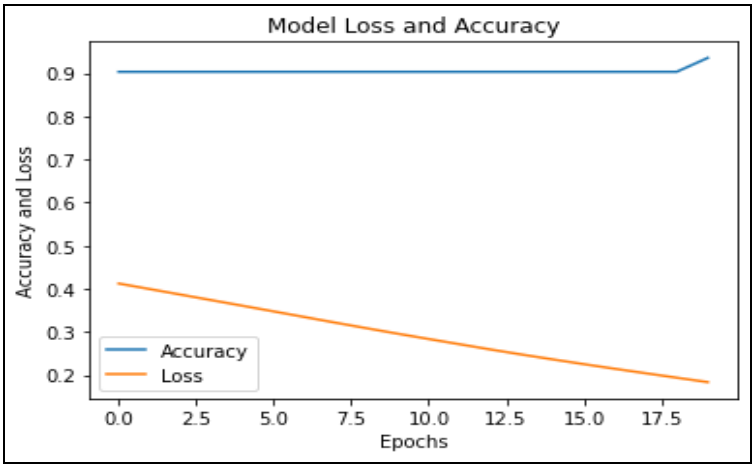

Fig. 12 Model Loss and Accuracy

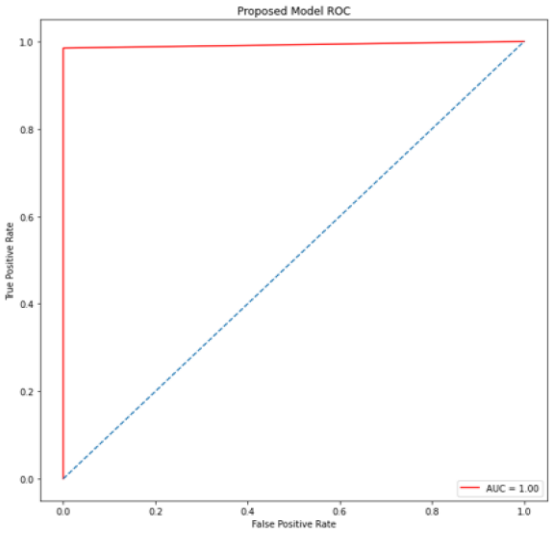

Fig. 13 ROC and AUC curve of proposed model

We also perform a comparative analysis of how the different algorithms used in the paper perform on the online available as well as real time dataset depicted in Table 5.

Table 5: Comparison of accuracy of the models on different datasets

	Accuracy	
Model	**Combined UCI+ Our Dataset**	**Our Dataset**
SVM	0.7870	0.7826
KNN	0.7468	0.7221
LR	0.7802	0.8023
DT	0.8219	0.8310
RF	0.8740	0.8534
ANN	0.8812	0.8732
HLDANN	0.9210	0.9322

The main reason behind implementing RFE was that we came across the most important features. The features that played a major role in the detection of CKD were Creatinine, BUN, and Urea. As per records displayed by the International Society of Nephrology, till now the best-known biomarker for detecting kidney disease is Creatinine. The positive outcomes of the experiments have spurred additional investigation into developing enhanced hybrid methods that leverage combinations of ML algorithms. The proposed methodology can also be applied to the detection of many other diseases.

6. Analysis of results in terms of prediction time

To understand the effect of parallel computing to reduce prediction time, we planned to execute the algorithm in iterations and check the time required using systems with 2 and 4 worker nodes subsequently. Our analysis is depicted in Table 6. We executed the algorithm for 5 iterations on the above-said infrastructure. For this, we used Python's IParallel Package. The analysis shows that the time for prediction was subsequently reduced when we worked with 4 worker nodes instead of 2 worker nodes.

Table 6: Analysis of parallel execution

Iterations	Worker 2(in ms)	Worker 4(in ms)
1	8.02	6.04
2	7.16	6.00
3	7.11	5.44
4	6.22	5.12
5	6.21	5.08

This shows that the algorithm is computationally efficient in predicting CKD in case if executed using parallel computations [32]. Further, we compare the time taken for prediction by all the algorithms mentioned above. This has been depicted in Figure 14.This shows that the proposed model takes less time for prediction.

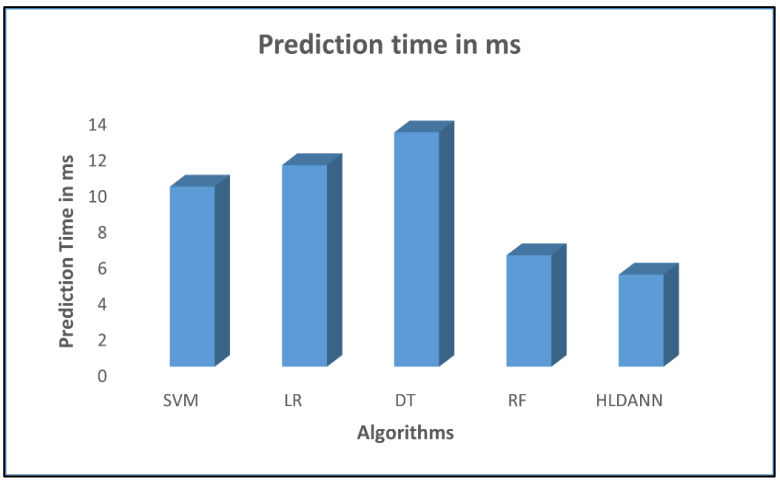

Fig. 14 Prediction time in ms

Table 7 shows a comparative analysis of the proposed model with some of the recent works. For this comparison, this proposed research was tested on the data set available on the UCI machine learning repository for CKD analysis. The proposed model gave an outstanding performance of approximately 93% on the online dataset.

Table 7: Comparative Analysis of Recent Works

Authors	Model	Accuracy (%)
[25]	Ant Colony Optimizer	79.45
[26]		79.31
[27]	Neural Network	83.2
[28]	KNN	82.45
[29]	CNN	87.21
Proposed Model	SVM HLDANN	93.22

7. Statistical Analysis

To compare the performance of the models, we used Friedman's Test [31]. Here the null hypothesis andalternate hypothesis have been formulated as below.

Hnh: No significant difference between all mentioned algorithms.

Hah: Difference between all specified algorithms.

The significance level(α) was 0.05. Since we got the ρ value of the test as 0.25 which was greater than the significant level, we would reject the null hypothesis. Thus, there is a statistically significant difference between all the applied algorithms.

8. Conclusions and Future Scope

We tried to present a hybrid model with a combination of LDA and ANN for the early prediction of CKD.

The subjects analyzed in our study were individuals treated at DY Patil Hospital in Navi Mumbai. The performance metrics that were taken into consideration were accuracy, precision, recall, F1 score, and prediction time. Dimensionality reduction has identified important attributes that have improved accuracy compared to all existing methodologies. The proposed algorithm outperformed all the other algorithms in terms of all the above-mentioned metrics. The identified factors were confirmed by pathologists for confirmation of the effectiveness of the results. The algorithm was validated on real-time records received from the hospital and the results were quite promising when compared to the standard ML algorithms like SVM, LR, RF, DT, KNN, and subsequently shallow ANN. This work provided the affirmation of DL algorithms being able to classify patients into various categories in terms of the decision-making process. Further to that, more and more data is received continuously, which will help us to make the model even stronger to work efficiently with diversified data.

In near future, research tends to move in the direction of testing many new biomarkers, which may be helpful in the detection of CKD in its early stages. More research could be done to see how different combinations of machine learning algorithms predict CKD. An information-driven approach plays a very important role in handling uncertainty. The development of algorithms that apply a probabilistic approach in the determination of CKD may also be helpful in this scenario. The development of a non-invasive technique for detecting such diseases is a need of the time and surely it will help the medical fraternity of our country and outside.

References

[1] Khade, A. A., Vidhate, A. V., & Vidhate, D. A Comparative Analysis of Applied AI Techniques for an Early Prediction of Chronic Kidney Disease. Proceedings - 2nd International Conference on Smart Electronics and Communication, ICOSEC 2021; 1386–1392. https://doi.org/10.1109/ICOSEC51865.2021.9591869

[2] Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., Jasinski,M., Jasinski, L., Gono, R., Jasinska, E., & Bolshev, V. Prediction of Chronic Kidney Disease - A Machine Learning Perspective. IEEE Access, 9(January),2021; 17312–17334. https://doi.org/10.1109/ACCESS.2021.3053763

[3] Dare, A. J., Fu, S. H., Patra, J., Rodriguez, P. S., Thakur, J. S., & Jha, P. Renal failure deathsand their risk factors in India 2001–13: nationally representative estimates from the Million Death Study. The Lancet Global Health, 5(1), 2017; e89–e95. https://doi.org/10.1016/S2214-109X(16)30308-4

[4] Ifraz, G. M., Rashid, M. H., Tazin, T., Bourouis, S., & Khan, M. M. Comparative Analysisfor Prediction of Kidney Disease Using Intelligent Machine Learning Methods. Computational and Mathematical Methods in Medicine, 2021; https://doi.org/10.1155/2021/6141470

[5] Yuan, Q., Zhang, H., Deng, T., Tang, S., Yuan, X., Tang, W., Xie, Y., Ge, H., Wang, X., Zhou, Q.,& Xiao, X.. Role of artificial intelligence in kidney disease. International Journal of Medical Sciences, 17(7),2020; 970–984. https://doi.org/10.7150/ijms.42078

[6] Bhaskar, N., & Suchetha, M. A Computationally Efficient Correlational Neural Network for Automated Prediction of Chronic Kidney Disease. IRBM, 42(4), 2020;268–276. https://doi.org/10.1016/j.irbm.2020.07.002

[7] Muslim, M. A., Kurniawati, I., & Sugiharti, E. Expert system diagnosis chronic kidney disease based on Mamdani fuzzy inference system. Journal of Theoretical and Applied Information Technology,2015; 78(1), 70–75.

[8] Hu, D., Nie, F., & Li, X. Deep linear discriminant analysis hashing. Scientia Sinica Information is 51(2), 2021;279–293. https://doi.org/10.1360/SSI-2019-0175

[9] Singh, V., Asari, V. K., & Rajasekaran, R. A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease. Diagnostics, 2022; 12(1),1–22.https://doi.org/10.3390/diagnostics12010116

[10] Almansour, N. A., Syed, H. F., Khayat, N. R., Altheeb, R. K., Juri, R. E., Alhiyafi, J., Alrashed, S., & Olatunji, S. O. Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. Computers in Biology and Medicine, 2019; 109(April), 101–111. https://doi.org/10.1016/j.compbiomed.2019.04.017

[11] Senan, E. M., Al-Adhaileh, M. H., Alsaade, et al. Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques. Journal of

Healthcare Engineering, 2021;. https://doi.org/10.1155/2021/1004767

[12] Rady, E. H. A., & Anwar, A. S. Prediction of kidney disease stages using data mining algorithms. Informatics in Medicine Unlocked, 15(December 2018), 2019; 100178. https://doi.org/10.1016/j.imu.2019.100178

[13] Khamparia, A., Saini, G., Pandey, B., Tiwari, S., Gupta, D., & Khanna, A. KDSAE: Chronic kidney disease classification with multimedia data learning using deep stacked autoencoder network. Multimedia Tools and Applications, 79(47–48), 2020; 35425–35440. https://doi.org/10.1007/s11042-019-07839-z

[14] Song, X., Waitman, L. R., Yu, A. S. L., et al, Longitudinal risk prediction of chronic kidney disease in diabetic patients using a temporal-enhanced gradient boosting machine: Retrospective cohort study. JMIR Medical Informatics, 2020;8(1), 1–16. https://doi.org/10.2196/15510

[15] Pasadana, I. A., Hartama, D., Zarlis, M., et al. Chronic Kidney Disease Prediction by Using Different Decision Tree Techniques. Journal of Physics: Conference Series,2019; 1255(1). https://doi.org/10.1088/1742-6596/1255/1/012024

[16] Krishnamurthy, S., Kapeleshh, K. S., Dovgan, E., et al. Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan. Healthcare (Switzerland), 2021;9(5), 1–13. https://doi.org/10.3390/healthcare9050546

[17] Neves, J., Martins, M. R., Vilhena, J., Neves, J., Gomes, S., Abelha, A., Machado, J., & Vicente, H. A Soft Computing Approach to Kidney Diseases Evaluation. Journal of MedicalSystems, 2015;39(10). https://doi.org/10.1007/s10916-015-0313-4

[18] Fokas, A. S., Dikaios, N., & Kastis, G. A. Mathematical models and deep learning for predicting the number of individuals reported to be infected with SARS-CoV-2. Journal of the Royal Society Interface,2020; 17(169). https://doi.org/10.1098/rsif.2020.0494

[19] Chen, X., & Jeong, J. C. Enhanced Recursive Feature Elimination. January 2008.2014; https://doi.org/10.1109/ICMLA.2007.35

[20] Shastri, S., Kour, P., Kumar, S., Singh, K., & Mansotra, V. XGBoost: A novel Grading-Ada Boostensemble approach for automatic identification of erythemato-squamous disease. International Journal of Information Technology (Singapore), 2021; 13(3), 959–971. https://doi.org/10.1007/s41870-020-00589-4

[21] Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., & Mohi Ud Din, M. Machine learning-based approaches for detecting COVID-19 using clinical text data. International Journal of Information Technology (Singapore), 2020;12(3), 731–739. https://doi.org/10.1007/s41870-020-00495-9

[22] Dasari, S. K., & Prasad, V. A novel and proposed comprehensive methodology using deep convolutional neural networks for flue-cured tobacco leaves classification. International Journal of Information Technology (Singapore), 2019;11(1), 107–117. https://doi.org/10.1007/s41870-018-0174-4

[23] Khade, A. A. and Vidhate, A. V. Application of Artificial Intelligence Techniques in the Early-Stage Detection of Chronic Kidney Disease, Data Science-Techniques and Intelligent Applications(NewYork), 2022.

[24] H.-C. Lee et al., "Prediction of Acute Kidney Injury after Liver Transplantation: Machine Learning Approaches vs. Logistic Regression Model," Journal of Clinical Medicine, vol. 7, no. 11, p. 428, Nov. 2018, doi: 10.3390/jcm7110428

[25] Elhoseny, M.; Shankar, K.; Uthayakumar, J. Intelligent diagnostic prediction and classification system for chronic kidney disease. Sci. Rep. 2019, 9, 9583. [CrossRef] [PubMed].

[26] Vasquez-Morales, G.R.; Martinez-Monterrubio, S.M.; Moreno-Ger, P.; Recio-Garcia, J.A. Explainable Prediction of Chronic Renal Disease in the Colombian Population Using Neural Networks and Case-Based Reasoning. IEEE Access 2019, 7, 152900–152910.

[27] Senan, E.M.; Al-Adhaileh, M.H.; Alsaade, F.W.; Aldhyani, T.H.H.; Alqarni, A.A.; Alsharif, N.; Uddin, M.I.; Alahmadi, A.H.; Jadhav, M.E.; Alzahrani, M.Y. Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques. J. Healthc. Eng. 2021, 2021, 1004767.

[28] Krishnamurthy, S.; Ks, K.; Dovgan, E.; Luštrek, M.; Pileti˘c, B.G.; Srinivasan, K.; Li, Y.-C.; Gradišek, A.; Syed-Abdul, S. Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan. Healthcare 2021, 9, 546.

[29] Polat, H.; Mehr, H.D.; Cetin, A. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. J. Med. Syst. 2020, 41, 55..

[30] S. K. Agarwal, S. C. Dash, M. Irshad, S. Raju, R. Singh, and R. M. Pandey, "Prevalence of chronic renal failure in adults in Delhi, India," Nephrology Dialysis Transplantation, vol. 20, no. 8, pp. 1638–1642, Apr. 2005, doi: 10.1093/ndt/gfh855.

[31] S. Singh, S. Shreevastava, T. Som, and G. Somani, "A fuzzy similarity-based rough set approach for attribute selection in set valued information systems," Soft Computing, vol. 24, no. 6, pp. 4675–4691, Jul. 2019, doi: 10.1007/s00500-019-04228-4.

[32] Khade, A., Vidhate, A.V. & Vidhate, D. FFN-XGB- design of a hybrid feed forward neural network and extreme gradient boosting model for early prediction of chronic kidney disease. Int J Syst Assur Eng Manag 2023. https://doi.org/10.1007/s13198-023-01993-2

[33] Khade, A.., Vidhate, A. V., & Vidhate, D., Design of an Optimized Self-Acclimation Graded Boolean PSO with Back Propagation Model and Cuckoo Search Heuristics for Automatic Prediction of Chronic Kidney Disease. Journal of Mobile Multimedia, 19(06), 1395–1414. https://doi.org/10.13052/jmm1550-4646.1962