Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Chronic Kidney Disease Prediction using Hybrid Optimization Search Model (HOSM)

S. Mohammed Imran[1], N. Prakash[2], Arif Abdul Rahman[3]

[1]Research scholar, Department of Computer Applications, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India
[2]Research Supervisor, Department of Information Technology, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India
[3]Faculty chairman, Department of Information Technology, Aalim Muhammad Salegh College of Engineering, Avadi, Chennai, India
Email: prakash@crescent.education

Kidney is an important organ of the human body helps to keep balancing the functionality of all other organs. Chronic kidney diseases(CKD)create many impacts in recent days due to changes in life style habits, lack of health awareness and irregular treatments on health issues. Diabetics is one of the largest impacted diseases affect the humans. A chronic diabetic allows slow degradation of organs health when untreated regularly. A chronic disease covers various health problems that is resilient in diagnosis. The proposed research work is aims to develop an comparative analysis model that deep dig the attributes of the chronic kidney dataset collected from Kaggle.com. the standard dataset act as the benchmark for initially creating the machine learning models. The novel framework acted upon with tuning the hyper parameters using hybrid optimization approach. Machine learning models such as Adaptive boosting model (ADAB), CatBoostAlgorithm (CA), K-Nearest neighbor algorithm (KNN), Probability boosted regression (PBR) and Wide – Deep decision tree model (WADET) etc. the comparative analysis on CKD is validated through output parameters in the form of prediction accuracy, system precision, Recall and F1 score. The novel system relies on hybrid optimization enabled search model represented as HOSM to find out the best adaptive approach on developed systems. The robust framework is iteratively performing the validation process thus provides N-Cross validation mechanism that adopt dynamic inputs. The presented system using HOSM-WADET achieved 98.77% cumulative accuracy comparing the state of art approaches.

Keywords: Kidney disease, Data analytics, Hybrid algorithm, Regression models, Diabetics.

1. Introduction

One of the body's most important organs is the kidney. In the healthcare industry, chronic kidney prediction has a significant scope for research. A person with a chronic disease requires

ongoing medical care and treatment throughout their lifetime. Without kidneys, people cannot live [1]. A reliable and accurate prediction system is highly sought after. Many medical records are now collected from patients and stored in the cloud thanks to the internet of things. Big Data Analytics offers significant contribution in systematic and predictive selection of various treatments and diagnostic options for patients [2]. The computerized patient data, which is based on physiological data, can be used to analyze various chronic diseases are health records. Due to the technology used to divide this data into a variety of attributes, big data analytics is in high demand [3]. In today's world, accurate prediction of chronic kidney diseases is critical to preventing life-threatening conditions. Diseases of diabetes Joint pain, colds, diabetes, and other problems that come up frequently all fall under the category of chronic diseases [4]. Long-term chronic diseases result from persistent medical illness. People frequently avoid diagnostic treatment for their fundamental issues. Based on the department's current data collector, predictive data analysis is thought to be a cost-effective method for determining future outcomes. Supporting the decision-making process for healthcare devices is made easier with the help of algorithms [5]. In order to conduct predictive analyses of future outcomes, healthcare departments require patient monitoring and treatment validation. Health Care Facilities frequently place an emphasis on developing applications that are better optimized in order to guarantee the health of their patients and address any upcoming issues immediately. These intelligent systems are also helpful for the patience required to provide immediate medical assistance [6].Without knowing the impact of chronic diseases, people underlie with the commonly showing symptoms as hypertension, renovascular disease etc.[7].

Utilizing models based on machine learning [8]and artificial intelligence to accurately diagnose CKD. To identify disease affected patient and normal patient, the model categorizes clinical data and health records. During the prediction phase the unsupervised technique of learning algorithms [9] are used to select features for the CKD health record. K-means clustering based segmentation of kidney images are discussed with the presented system in which based on the region affected the presence of chronic kidney disease is classified[10].

An additional phase of kidney disease evaluation deals with family history. If someone in the family had kidney problems, required dialysis [11], or had to have their kidneys transferred, the same thing would have happened to the next person. The kidneys respond to various high-dose medications in different ways. The renal tissues were accurately developed by the medicine's high impact. Kidney disorders[12] can also be caused by obesity and high blood pressure.

Kidney disease is classified into different levels based on the impact as No kidney disease or normal (NKD), Long term kidney infection (CKD)[13], End stage of kidney disease as (ESKD) etc. are presented using statistical method of cross entropy enabled deep belief network[14](DBN). In the presented article, the system achieved the cumulative score above 90% is achieved by comparing various existing methods.

Deep convolutional_neural_networks (CNNs)[15] are widely utilized in image processing applications. The convolution neural network model is applied in kidneydisease classificationwith kernel-based support vector analysis. The presented system achieved the accuracy of 98.04%.

- The proposed model initiated with hybrid optimization mechanism using transfer learning and adaptive weight management. The performance of the proposed predictive model is considered by the optimization loop where the input features are controlled. The input features are two different data types. The numerical one and the object type. The categorical information is processed separately where the numerical or Boolean data is processed separately.

- The proposed model considers iterative learning algorithms namelyK-Nearest neighbor (KNN) technique, AdaBoost Regression (ADAB), Cat Boost Regression(CB), Wide and Deep Ensemble Tree Classifier(WADET) and Probability Boosting Regression Model(PBRM) etc.

- The feature data are optimized at every instance before fetching into the analysis model. The weight of the features are reflected into the performance result using accuracy, precision, Recall and F1Score. The adaptive loop structure continuously optimize the input features towards number of iterative learning.

- The directed procedure reduces the processing time and complexity in handling unstructured dataset. the dataset is tuned in the initial stage itself. The scaling and dimensioning of data enable the hybrid model to optimize the result and provide early detection.

The rest of the paper of written as detailed discussion on dataset collection, Exploratory data analysis in Section 2. Followed by Background study considering various existing papers in Section 3. Detailed system architecture and step by step procedure of the proposed model in Section 4. Explained over here. Section 5. Discuss the interpretation of results obtained from the proposed HOSM on CKD detection.

2. Materials and Methods

2.1 Dataset

Blood glucose is a crucial factor in determining whether a patient has chronic kidney disease or not. Both patience and high blood levels of red platelet and hemoglobin are associated with a relatively equal number of illnesses. The figure depicts infected individuals with fewer than five red platelets and a volume of 400 filled cells.3 The ratio of red blood cells to volume is also important when determining the difference. There is a strong correlation between red platelet and hemoglobin in people with chronic kidney[16] disease who have less than 5 platelets and a 12.5 hemoglobin level. According to the provided data set, potassium levels of 39 and 49 have an impact on chronic kidney disease and should be monitored closely. The sodium level of 14.5 is also taken into account. The data set's missing values are rounded to the nearest 20. Numerous junk values exist in these data that are irrelevant to the current investigation. The K-nearest to neighbor method is utilized for the analysis and normalization of the missing values. Data reduction is the process of scaling the date's dimension prior to analysis in data mining techniques[17]. These data must be cleaned in order to establish a connection and calculate the relativity score. The process of converting data from one format to zero score normalization, in which the converted data are used for analysis, is known as

data transformation. Decision tree, Naive Bayes, Kernel controlled Support vector Machines(K-SVM), and Adaptive boosting technique are compared to the boosting method in the proposed strategy. Various health care analysis are achieved in recent days using predictive analysis models[18].

2.2 Exploratory Data analysis

The initial analyses of non-linear data necessitate exploratory data analysis. In order to identify the relevant parameters that appear repeatedly in each analysis, it is necessary to discover the pattern in the dataset. The creation of graphical representations of the input data that are provided with a summary of statistical measures[19] is the primary focus of exploratory data analysis. For the prediction process, non-linear data must be structured into normalized data. For the purpose of analyzing the non-linear input with various statistical scores, the proposed system is developed using the Python Tool that is integrated with Google Collaborator. Python libraries are integrated into Google Collab; Consequently, it serves as an online Python code execution tool that is open source. NumePy, Sci-Kit Learn, SciPy, MatplotLib, and other Python libraries were utilized in the proposed design. Accessing the fundamental system features like input out access and other core modules is made easier with these libraries.

Table 1. Dataset analysis

Attribute	Description
Age_	Age of the patient declared as numerical
bp.	Systolic blood pressure measured in terms of mm/Hg
Sg.	Specific_gravity utilized as [1.005,1.010,1.015,1.020,1.025]
Al.	Albumin_level utilized as (nominal) al - (0,1,2,3,4,5)
Su.	Human body Sugar utilized as (nominal) su - (0,1,2,3,4,5)
Rbc.	Quantity of red blood utilized as (nominal) rbc - (normal,abnormal)
Pc.	Pus_Cells in the term pc - (normal,abnormal)
Pcc.	Pus_Cells as clumps utilized as (nominal) pcc - (present,notpresent)
Ba.	Bacterial_infection utilized as (nominal) ba - (present,notpresent)
Bgr.	Blood_Glucose_level as Random utilized as (numerical) bgr in mgs/dl
Bu.	Blood_Urea_level utilized as (numerical) bu in mgs/dl
Sc.	Serum_Creatinine_level utilized as (numerical) sc in mgs/dl
Sod.	Sodium_level utilized as (numerical) sod in mEq/L
Pot.	Blood potassium_level utilized as (numerical) pot in mEq/L
Hemo.	Blood density in the form of hemo utilized as (numerical) hemo in gms
Pcv.	Amount of Packed Cells utilized as (numerical)
Wc.	Number of white blood cells utilized as (numerical) wc in cells/cumm
Rc.	Blood cell volume utilized as (numerical) rc in millions/cmm
Htn.	Mental stress utilized as (nominal) htn - (yes,no)
Dm.	Diabetes setback utilized as (nominal) dm - (yes,no)

Cad.	Heart problem utilized as (nominal) with class cad - (yes,no)
Appet.	Appetite utilized as (nominal) appet - (good,poor)
Pe.	Pedal Edema utilized as (nominal) pe - (yes,no)
Ane.	Anemia(nominal) ane - (yes,no)
Class.	Class (nominal) class - (ckd,notckd)

Table 1. shows the UCI repository collected dataset on chronic kidney disease. The dataset attributes are displayed over here.

The unique vfralues present in the list of attributes are given below

- Red_blood_cellsis the attribute name contains the output value nan, 'normal' , abnormal' etc.
- pus_cellis the attribute name contains the output value['normal'/ 'abnormal'/ nan] values
- pus_cell_clumpsis the attribute name contains the output value['notpresent' /'present' /nan] values
- bacteria is the attribute name contains the output value['notpresent' 'present' nan] values
- hypertension is the attribute name contains the output value['yes' /'no'/ nan] values
- diabetes_mellitusis the attribute name contains the output value['yes'/ 'no'/ ' yes'/ '\tno' /'\tyes' nan] values
- coronary_artery_diseaseis the attribute name contains the output value['no'/ 'yes'/ '\tno' nan] values
- appetite is the attribute name contains the output value['good'/ 'poor'/ nan] values
- peda_edemais the attribute name contains the output value['no'/ 'yes' /nan] values
- aanemiais the attribute name contains the output value['no'/ 'yes' /nan] values
- class is the attribute name contains the output value['ckd'/ 'ckd\t'/ 'notckd'] values. On the other hand Still some invalid values are present in the columns are removed.

Clinical trials are the standard data collected live from the patients and opted for chronic syndromes. Based on the results obtained after the specific treatments these data are validated and posted in public forum for research on disease interpretation also to enhance the pattern of treatment [20]. The machine learning algorithms are commonly used in existing frameworks, although predictive analysis is executed in specific cases, development of ensemble approach is recommended due to the complexity in dynamic data handling [21].

3. Background study

Recently, anticipating the possibility of infection by utilizing the enormous information

method in conjunction with expanding recurrence. Specialists have conducted numerous calculations and compartments of tools and concentrated them. These have demonstrated the vast capabilities of this research field. The study of existing implementations arediscussed here. The comprehensive study discusses the advantages and disadvantages of each techniques that relate with the proposed research work.

Wu et al., (2022) the author presented a system where continuous real time monitoring data is formulated with (AI) Artificial intelligence enabled platform. The complete health care platform is discussed here. The chronic diseases, health monitoring, post-surgical health monitoring, emergency alert towards API control is provided[22].

Kanda et al.(2022) Long term kidney disease, Cardiac arrest arecommonlyhitting diseases in recent days. Diabetics are becoming more common health impact. The author presented a system with CKD and Heart failure(HF) dataset to analyze the abnormality. Chronic diseases creates various impact on the patient health status. Accumulation of slow impact creates many changes within the human body. Over 16822 patients data are collected and formulated the Kaplan-Meier curves analysis[23].

Yang GMet al.(2023)Presenteda system that considersclinical reports and records as databases collected from hospitals essential for recalling a specific objective in order to produce productive and accommodating confirmation. In order to refresh the idea of information, preprocessing was appropriate. The changed dataset was used by text analysis through Naïve Bayes model. The presented article achieved precision of 72.3% [24].

Samrat Kumar Dey et al. (2022) the author presented a system where hybrid (Chi2) Chi-squared test is utilized to evaluate the pro-longed kidney disease detection system. Presented system considers unique features from input data collected from chronic kidney patients and formulated the CKD classification with accuracy of 98%. The result of bagging performance out of various machines learning model is performed well [25].

P. Chittora et al.(2021) the author presented a system in which proposed solutions framed using neural networks on long way kidney abnormalities. In the clinical consideration section, Maker reveals the demonstration of such insightful assessment, focusing on electronic prosperity records, the widespread use of sparks, trademark language handling used in decision systems. Using statistical methods like chi-square test, least absolute shrinkage, correlation coefficient the identification of chronic kidney disease detection is modelled [26].

Aqueel Ahmed et al.(2022) discussed the assumption of coronary disease. Finding useful information is a crucial but muddled task that must be carried out clearly and profitably. The skilled data analysis techniques are known to find unique information from large clinical data. The clinical benefits frameworks provide access to a wealth of data. To find canvassed associations and models in data, a variety of plausible examination tools are available. Data processing, feature modelling are utilized in numerous applications. One of the applications is disease detection, in which data processing devices demonstrate practical results. The presented system considers learning algorithms such as Support Vector Machine (SVM), Genetic Algorithm (GA), Reinforced learning, and artificial neural networks were all proposed in this evaluation paper to treat gut disorders[27].

Bai et al. (2022) the author presented a machine learning based approach for End stage kidney

disease detection system (EKDD). Gradual loss of kidney health due to chronic impact of diabetics becomes common in recent days. The comparative study of Naïve Bayes algorithm, Logistic Regression, and Random forestalgorithm to find the sensitivity of kidney failure with risk assessment. The lasting stage kidney disease (ESKD) detection using prognostic approach is evaluated and achieved the accuracy improvement of 9.4% comparatively through 70+ patient data[28].

Durairaj et althe author presented neural network based system to make inferences based on useful data. The Universal markers are the name given to neural networks. Diabetes mellitus, or DM for short, is a condition caused by an increase in blood glucose levels. Diabetes can be diagnosed using a variety of standard methods, including physical and designed tests. The hypertension risk estimation framework based on unreal neural networks (ANNs) can be used in an acceptable manner[29].

Akchurin et al.(2021) the author discussed the outcome of chronic kidney disease and its impact to the human body through certain diet. By considering the dietary measures required to control the blood pressure, metabolic activity chronic diseases are identified. Based on the impact created by the diets the implantation process is initiated to the kidneys[30].

Vestergaard SVet al.(2021), the authors presented a system in which 99000 unique individual dataset are collected from real time patients. Various clinical data are considered to analyze the pattern of abnormality present in it. Over the period of 1999 to 2008 chronic kidney infection affected database is formulated[31].

Hossain et al., (2022) The author presented a comparative study on chronic kidney disease detection through ensemble approach. Features extraction from CKD dataset is modeled using linear discriminant analysis (LDA). Feature optimization is included in the design to enable the feature scaling process. 10-fold cross validation is implemented here to make the chronic kidney disease presence in a significant way[32].

4. System Model

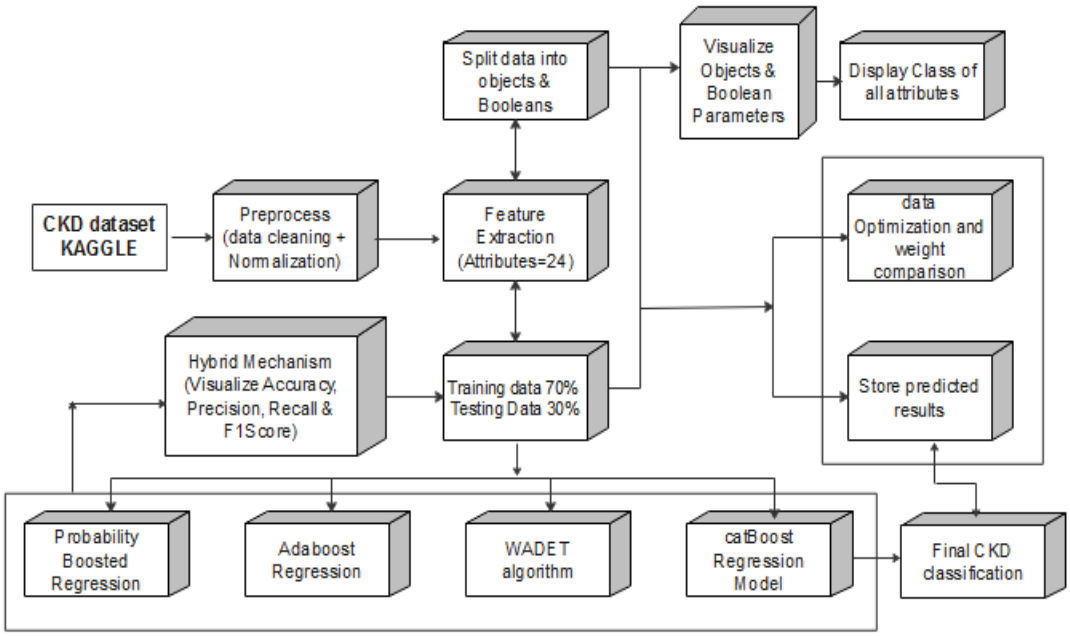

Fig. 1 System architecture of Hybrid optimization enabled Search Model (HOSM)

Fig. 1 shows the System architecture of proposed Hybrid mechanism on optimized search process. The optimization with various machine learning algorithms together is highly helpful to provide adjustable decision process with unique features. The existing frameworks on chronic kidney disease detection utilized with machine learning algorithm with training and testing of dataset.

Dataset Information/Dataset Review

The dataset collected from CKD dataset in the form of .CSV from Kaggle.com. The initial data processing is nothing but connecting the cloud platform with the data storage server using Google drive. The mounted drive fetch the files automatically and the updatins are happening in the cloud itself.

Exploratory Data Analysis

The initial analyses of non-linear data necessitate exploratory data analysis. It is expected to find the example of the dataset to recognize the significant boundaries that more than once happen at each investigation. The creation of graphical representations of the input data that are provided with a summary of statistical measures is the primary focus of exploratory data analysis. For the prediction process, non-linear data must be structured into normalized data. For the purpose of analyzing the non-linear input with various statistical scores, the proposed system is developed using the Python Tool that is integrated with Google Collaborator. Python libraries are integrated into Google Collab; Consequently, it serves as an online Python code execution tool that is open source. NumePy, Sci-Kit Learn, SciPy, MatplotLib, and other

Python libraries were utilized in the proposed design. Accessing the fundamental system features like input-out access and other core modules is made easier with these libraries.

Feature Encoding and Hybrid Optimization

Buiding the Model a. KNN Algorithm(KNNA) b. AdaBoost Regression(ADAB) c. Cat Boost Regression(CB) d. Wide and Deep Ensemble Tree Classifier(WADET) e. Probability Boosting Regression Model(PBRM) with optimized features are the highlighted steps in the proposed HOSm model.

K-Nearest Neighbor algorithm (KNN)

K-nearest neighbor algorithm classifies the data with respect to the label provided to the input dataset. the system considers the label as hint for the classification process. Based on the training data and group data the KNN model classifies the most relevant data from the pattern of inputs.

Adaptive Boosting (ADAB)

The adaptive boosting algorithm enhance the input features to make the correlation factor better. classifier. In order to boost detection accuracy, the adaptive boosting algorithm rotates the existing classifier at each hydration and expands the analysis system. The performance of the probability weighted AdaBoost algorithm is compared to that of the standard boost algorithm using the proposed strategy. At each iteration, the weight's probabilities are evaluated. For these rates, the standard weights are added to the adjusted iterations; It may alter the proposed value analysis as a whole. The overhead issue in the standard AdaBoost algorithm is incorporated into the probability weighted Ada boost algorithm. In order to achieve improved classification accuracy and further reduce the amount of time required to complete the analysis, the proposed paper takes into account a probability-based ReBoost strategy. The proposed method uses both positive and negative classification to adaptively alter the boosting algorithm's weight. The weight refreshing of the likelihood supporter versatile calculation in light of misleading positive and bogus negative rate is examined underneath.

CatBoost regression (CB)

CatBoost regression algorithm considers the object class type to eventually divide the data into object values such as labels, hyper tension presence, pus cells presence etc. in the presented approach categorical based approach play an important role. The relativity between the physiological data such as blood glucose, potassium level, blood pressure and pus cell count, hyper tension are significantly analyzed in the presented method. CatBoost is one of the fastest learning model. The categorical data calls features thatreduces GPU computation instead of numerical features. CatBoost uses a lot less GPU memory than other algorithms when compared to their performance. If the data set has more memory available for current access, the GPU operation will also slow down. Multiple histograms are used to model how well the algorithm works with GPU operation.

The CatBoost algorithm generates a histogram with two statistics and eight features per group. The fact that CatBoosttechniqueutilized here includes shared memory and temporary memory rather than permanent memory is the best feature of the algorithm. By making use of the

parallelism of instructions to get the best computation time, enrolling loops enable the method to achieve high performance. The CatBoost algorithm's ability to identify hash values and store them in temporary memory was tested for the first time. During the subsequent titration, these hash values can be used to identify, make pattern identification, and identifies the similar pattern using the trained data. CatBoost computes the learned data and the new data separately in each group of information. The challenge in CB includes, the data size impacts the processing time hence, the CatBoost algorithm's overall performance will improve. The CB model adopt the size of the data available towards the learning model.

Wide-Deep Decision tree model (WADET)

In the linear gradient regression model, the goal is to get the best fit to the regression line. In this model, the given data value of x is compared to the database value of Y and trained. Certain statistics are used to determine whether the predictions are true or false, such as the root mean square error. The correlation would be higher the lower the cost function. The model must be randomly trained with recursive data updates between higher iterations in order to minimize the cost function. The prediction score ought to be one in order to carry out the final update of the hypothesis equation for the linear data on the nonlinear data. Using the following formula

$$J = \frac{1}{n}\sum_{i=1}^{n}(xi - yi)^2 \quad (1)$$

$$Min\frac{1}{n}\sum_{i=1}^{n}(xi - yi)^2 \quad (2)$$

Using the Nelder–Mead algorithm, WADET pseudocode Input: Obtain the dataset and set convergence thresholds. W optimized weights for each column Mm, d1 is the number of rows, W is a null vector, and length d is the length of the null vector Mm S is zero; At first,

while (Conv0)do S=s+1;

For j=1: d1, calculate X[j] as metric(y, P+z(:,j)/s);

Finalize Max(j) = argmax(x(j));

W(jmax) = wjmax+1; P = p+x(:,jmax);

Stop the Return process w/s;

Probability boosted Regression model (PBRM)

Boosting algorithms are utilized to tune the data and enhance the prediction pattern to meet the goal of accuracy. Adaptive boosting model is capable of improving the data features and its weights with respect to that of pattern correlations. Adaptive boosting technique suffers from data overhead issue. To reduce the large data handling issues, probability boosted regression model is utilized. The probability boosted model reweight the feature quality and improves the accuracy without overhead issues.

Probability based random distribution formulate the feature selection with sampled phase.

Hence the processing time also reduced. The input data is split into training data of 70% and testing data of 30%. Once the PBRM is created, then the test data is fetched into the model to make analysis. the weights of the system are adjusted to make strong classifier decision. The proposed system classifies the data into CKD and NON-CKD.

The implementation steps are explained below.

Probability error(PE) is utilized in control with False positive data and False negative rate etc.

$PE_{(FP)}$ = False Positive Rate at every probability process

$PE_{(FN)}$ = False Negative Rate at every probability process

The PBR model estimate the regression ratio using the formula below

$$R_{ratio} = \log(E_t/ (1- E_t)) * ((1-PE_{(FP)})/FP) \quad (3)$$

Where,

RP_{ratio}= Positive Regression ratio,

With respect to false negative rate, the probability of error rate is defined by the regression value as shown in the expression below.

$$RN_{ratio} = \log(E_t/ (1- E_t)) * ((1-PE_{(FN)})/FN) \quad (4)$$

Where,

RN_{ratio} = NegativeRegression ratio etc.

Implementation summary of Hybrid optimized search model(HOSM)

- The novel system considers the CKD dataset collected from UCI repository [33] as source location utilized in Kaggle.com which is publicly available to download.
- The dataset consists of 26 attributes including the age of the patient, blood pressure status, blood glucose level, hypertension even more related to the kidney related chronic symptoms.
- From the 26 attributes, 15 numerical attributes are considered for linear analysis, 11 categorical attributes are considered for supporting data. The implementation is developed in python- google collab platform.
- The discussed machine learning algorithms are separately developed to meet the goal of chronic kidney disease classification with different accuracy obtained and tabulated in the Table.
- The novel approach collaborated here is the implementation of Hybrid optimization enabled Search model (HOSM) is incorporated with the feature tuning process.
- The raw data is initially considered for analysis. The overall performance of the system is tuned by scaling the features extracted. The HOSM model tune the feature values before fetching it into analysis model.

- The optimization model adjust the feature values to reduce the processing time, improve the sensitivity, improving the accuracy of prediction etc.
- GridSearchCVmodel is implemented here which acts as a cross validation process for the presented algorithms. The cross validation model evaluate the performance at every iteration of the Hybrid machine learning algorithms.
- The proposed system also have the responsibility to select the best performing model from the overall process.
- The performance is validated through accuracy, precision, Recall, F1Score etc.

Various results obtained from the simulation is presented in the results and discussions section.

Performance metrics

True positive rateis calculated based on the number of correctly classified results occurs same as than that of expected results.

True Negative rate is calculated based on the amount of predicted values on incorrect data that occurs as positive. The result of true negative rate is the confirmation of negative values only.

False positive rateis calculated as per the overall process, the truly predicted results are pretend to be positive, but the actual level is negative.

False negative rateis calculated based on the truly negative prediction occurs to be false in certain cases. The result impact the accuracy a lot.

From the given test parameters, TPR, TNR, FPR, FNRare evaluated from confusion matrix. The relevant confusion matrix obtained from each machine learning model are explored in results and discussion section 5. Below are the performance measuring formulas.

$$\text{Accuracy(acc)} = \frac{\text{TNr+TPr}}{\text{TPr+TNr+FPr+FNr}} \quad (5)$$

$$\text{Precision(P)} = \frac{\text{TPr}}{\text{TPr+FPr}} \quad (6)$$

$$\text{Recall(R)} = \frac{\text{TPr}}{\text{TPr+FNr}} \quad (7)$$

$$\text{F1Score(F)} = 2 * \frac{\text{Precison*Recall}}{\text{Precision+Recall}} \quad (8)$$

5. Results and Discussions

Exploratory data analysis

Fig 2. Exploratory data analysis of CKD dataset

Fig 2. Shows the exploratory data analysis of CKD dataset in which the raw parameters are visualized. Different parameters are impacted towards the CKD presence. Hence EDA is helpful to analyze it separately.

Heat-Map visualization

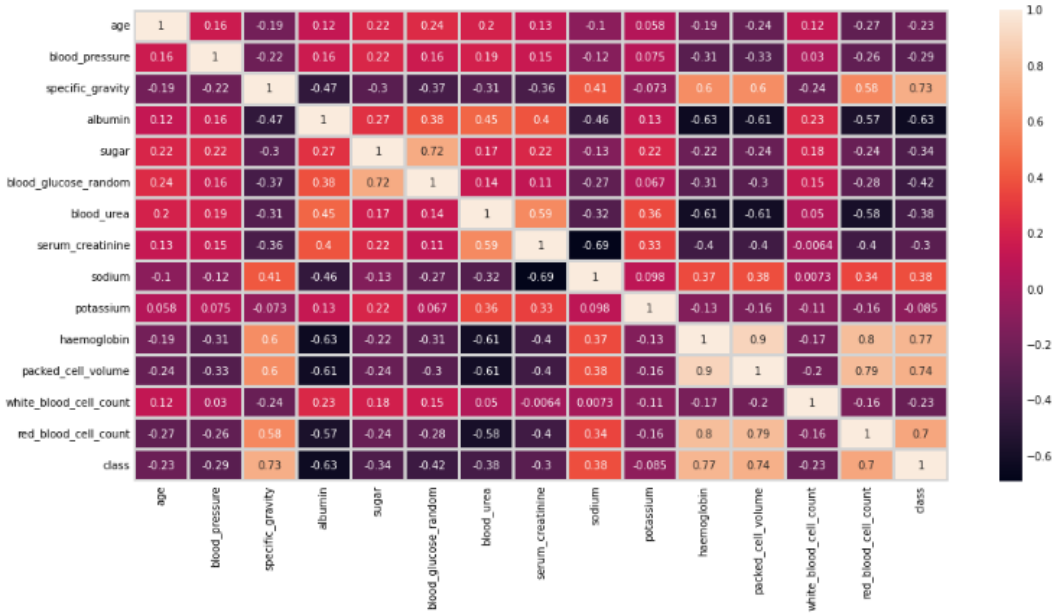

Fig 3. Heat Map visualization

Fig 3. Shows the Heat map visualization of 15 numerical attributes. The relativity towards the presence of CKD and non-CKD falls on the impacted value of numerical data of 15 attributes such as age factor, blood pressure, specific gravity, albumin, Sugar, blood glucose random evaluation, blood urea, serum creatine, sodium, potassium, hemoglobin, packed cell volume, white blood cell volume, red blood cell volume, and related classification falls on it.

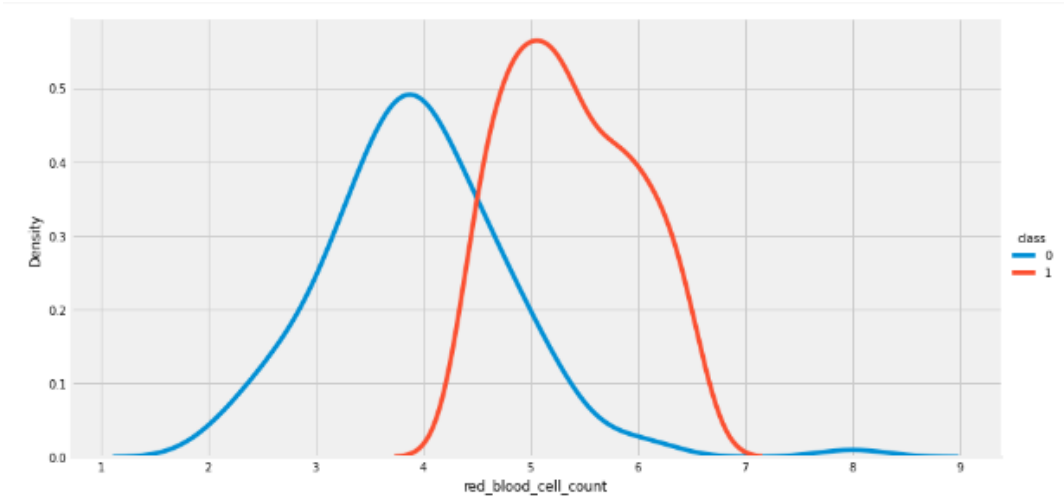

Fig 4. Redblood cell count vs. Density

Fig 4. Shows the red blood cell count(RBC) vs. density of the blood etc. class 1 falls on the

density in the range of 0.6, class 0 falls on the range of 0.5 etc. with the help of exploratory data analysis(EDA) process the deeper insight on relativity between the parameters are explored. From the above graph the RBC count is more as the density of the blood increases.

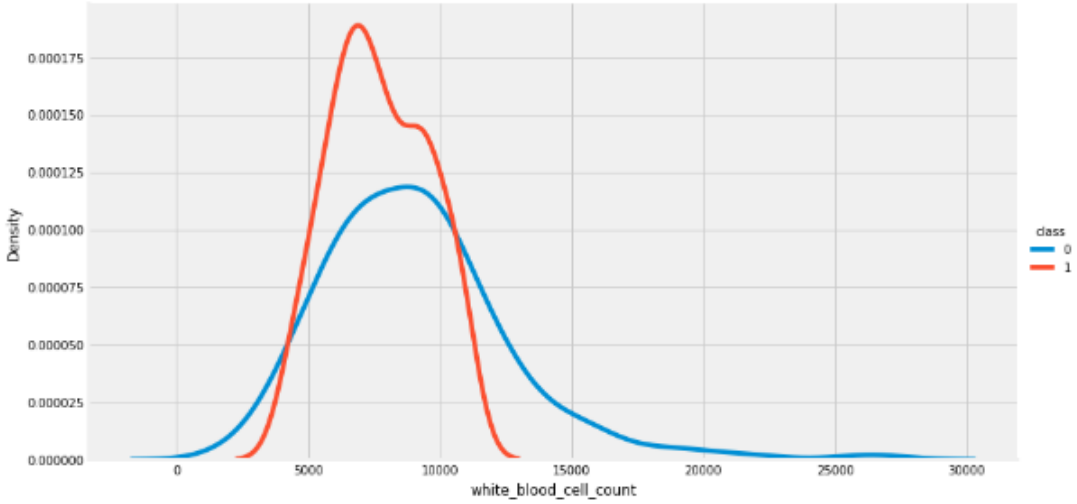

Fig 5. White blood cell count vs Density

Fig 5. Shows the white blood cell count with respect to density. class 1 falls on the density in the range of 0.000175, class 0 falls on the range of 0.000100 etc. In the similar way as RBC count, white blood cell count are responsible for chronic impacts in kidney. As the density is not higher comparing with the RBC counts, it shows clear raise in disease existence as shown in Class 1.

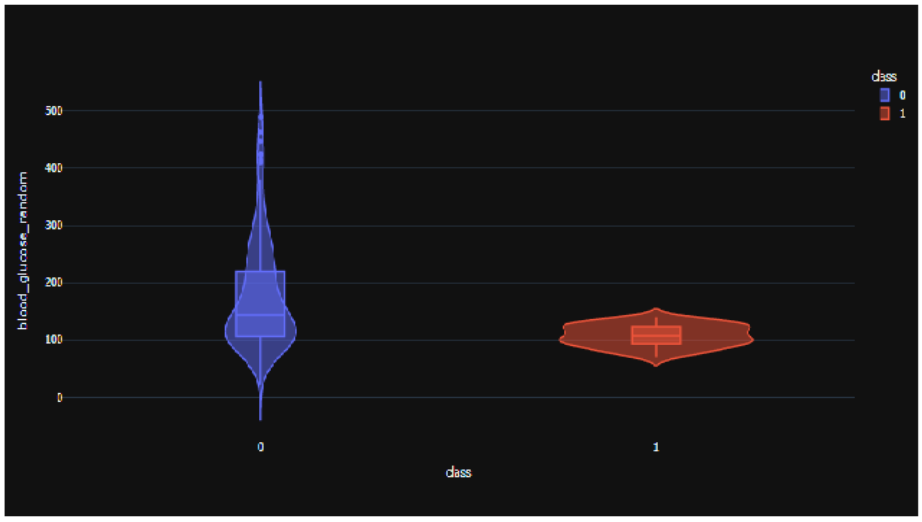

Fig 6. Blood glucose level vs. CKD class

Fig 6. Shows the EDA graph comparing the blood glucose level with respect to CKD

classification etc. the random glucose level raise high in class 1 as the glucose level is low in class 1. Most of the chronic kidney impacted patients are diabetics. The graph predicts the existence of such scenario.

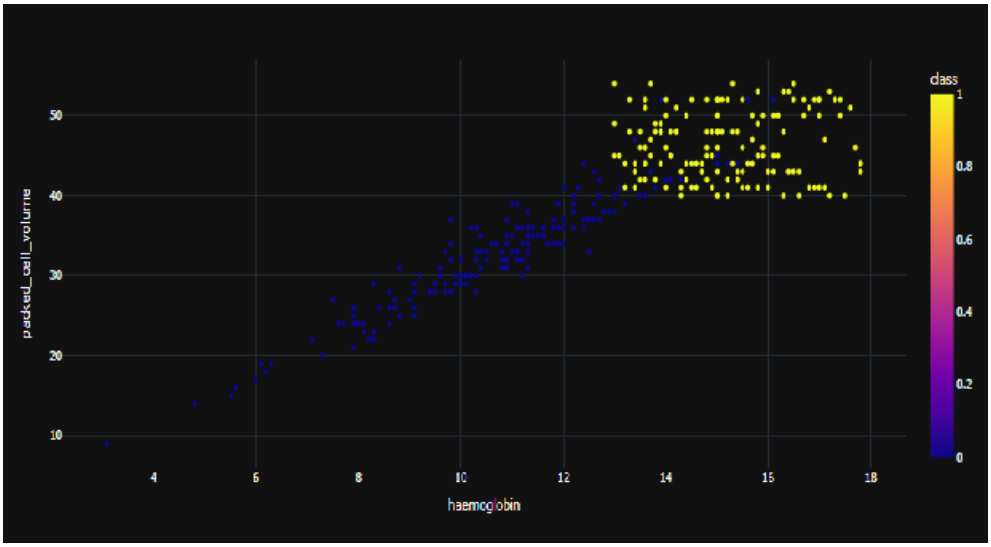

Fig 7. Hemoglobin vs. Packed cell volume

Fig 7. Shows the EDA graph comparing the Hemoglobin with respect to packed cell volume etc. the convergence of both the parameters falls near the class 1. Both the parameters impact more in class 1 chronic disease category.

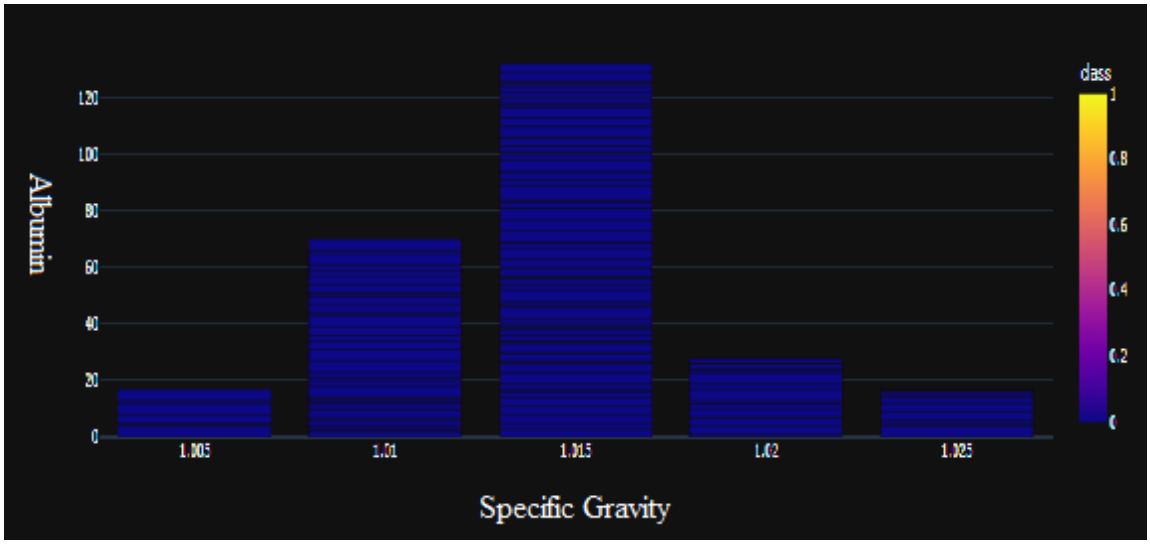

Fig 8. Specific gravity vs. Albumin

Fig 8. Shows the EDA graph comparing the Specific gravity with respect to Albumin etc. the graph depicts the specific gravity of 1.015raise on class 1 category towards the maximum albumin level 1200.

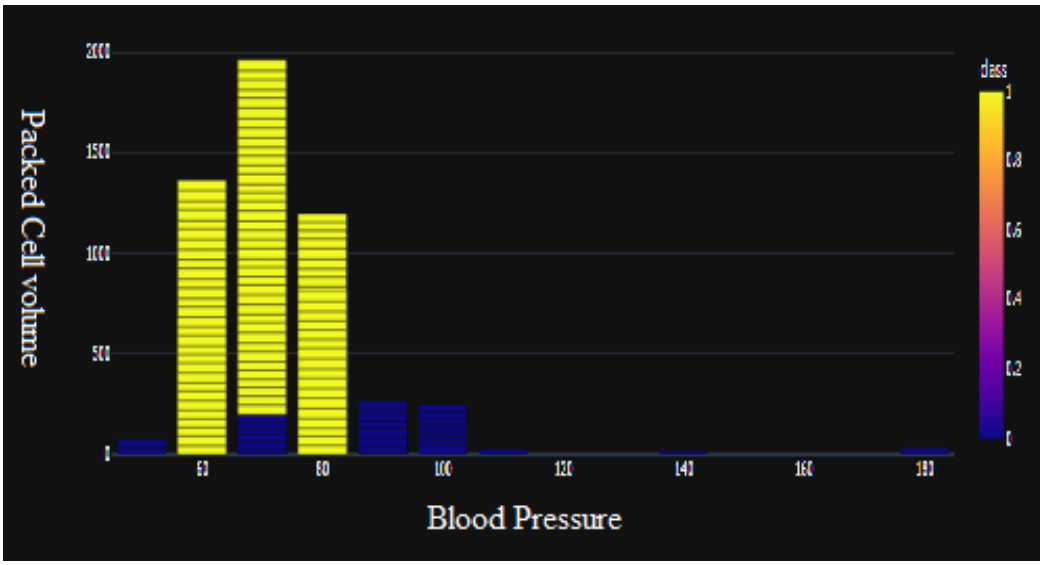

Fig 9. Blood Pressure vs. Packed cell volume

Fig 9. Shows the Blood pressure vs. Packed cell volume on chronic kidney disease dataset. Blood pressure (BP) is relatively low in class 1 category as the packed cell volume raise to the maximum for class 0 category.

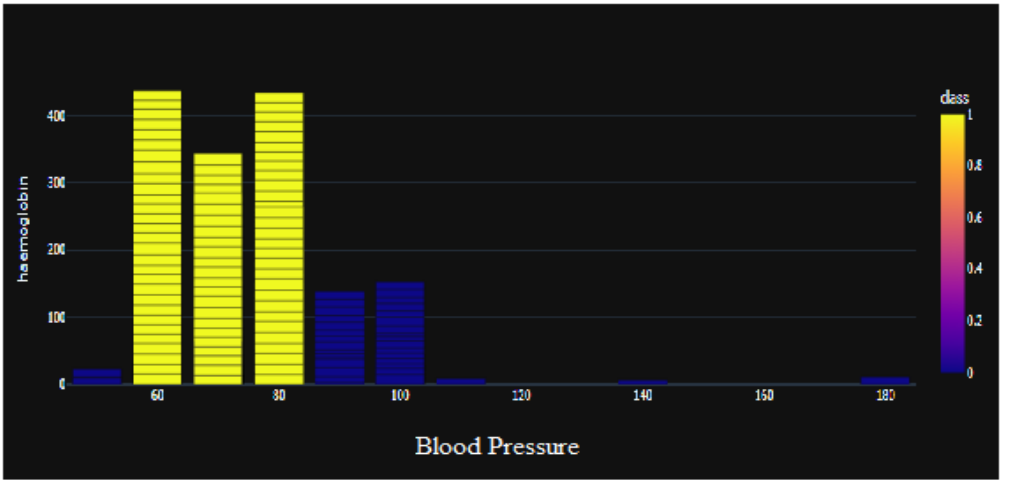

Fig 10. Blood Pressure vs. Hemoglobin

Fig 10. Shows the Blood pressure vs. Hemoglobin on chronic kidney disease dataset. The BP level falls normal as comparing the relativity the eventual raise occurs on hemoglobin too. People with chronic impacts have the both the parameters falls under the range depicted above.

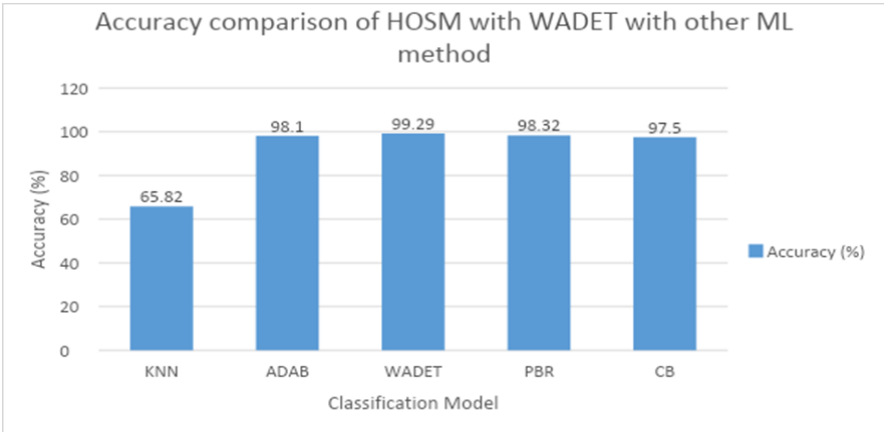

Fig 11. Comparison of analysis models

Fig 11. Shows the accuracy comparison of KNN, AdaBoost, WADET, PBR and CB model with respect to various levels of test iterations.The comparison is made with respect to the accuracy score as KNN achieved 65.8% , ADAB model 98.1%, WADET model 99.2%, PBR model 98.3%, CB model achieved 97.5% accuracy respectively.

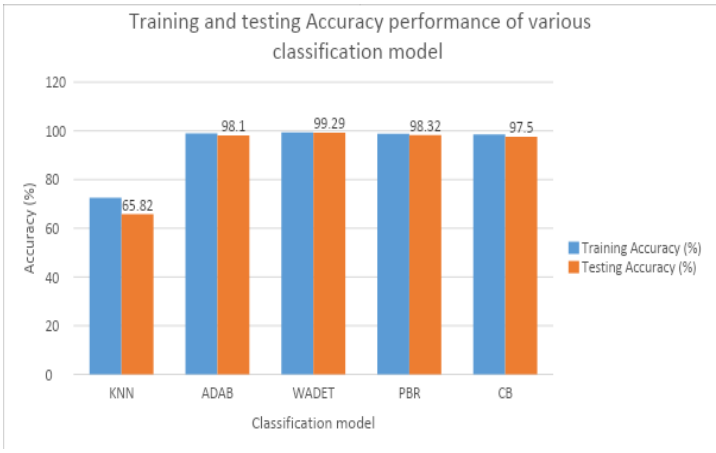

Fig 12. Comparison of Training accuracy vs. Testing Accuracy

Fig 12. Shows the comparison graph of training accuracy with testing accuracy on various state of art(SOA) approaches developed here.

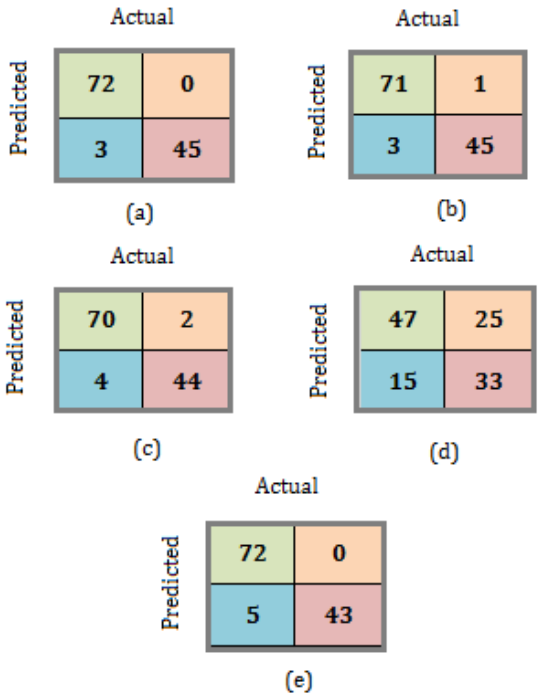

Fig 13. Confusion Matrix of Proposed ML models (a) CatBoost model, (b) Adaboost model ,(c) WADET model, (d) KNN (e) PBR model

Fig 13. Shows the confusion matrix of CatBoost model, Adaboost model ,WADET model, KNN, PBR model. Confusion matrix is responsible for deriving the true positive value, true negative value, false positive value and false negative value etc. the more positive predictions made within the analysis model then the accuracy increases with the true positive rate.

Real-time Prediction

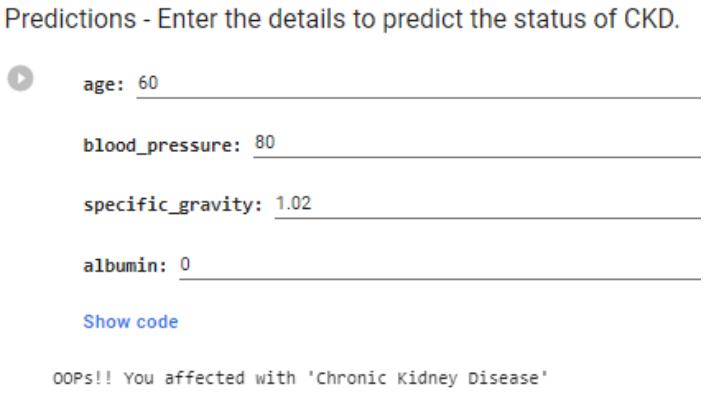

Fig 14.Prediction – User Data

Fig 14. shows the real time prediction of user data collected from real-time analysis. The dynamic data can be tested with the proposed model. The model developed here with the help of input prompt created using the python platform, the method is further extended as android application in future for self screening purpose or clinical screening purposes.

Table 2. Parametric comparisons

Method	Accuracy	Precision	Recall	F1Score
CB	0.9755	0.96	1.00	0.98
		1.00	0.94	0.97
ADAB	0.9810	0.96	0.99	0.97
		0.98	0.94	0.96
WADET	0.9877	0.95	0.97	0.96
		0.96	0.92	0.94
KNN	0.6580	0.76	0.65	0.70
		0.57	0.69	0.62
PBR	0.9830	0.94	1.00	0.97
		1.00	0.90	0.95

Table 2. shows the comparative prediction score of CB, ADAB, WADET, KNN and PBR model using HOSM architecture. The parametric comparisons derived here with existing developments such as CB model achieved accuracy =0.9755, Precision=0.96, Recall=1.00, F1score=0.98, where ADAB achieved accuracy =0.9810, Precision=0.98, Recall=0.99, F1score=0.97, where WADET achieved accuracy =0.9877, Precision=0.96, Recall=0.97, F1score=0.96 as maximum value, where with KNN model achieved accuracy =0.6580, Precision=0.76, Recall=69 F1score=0.70, where PBR achieved accuracy =0.9830, Precision=1.00, Recall=1.00, F1score=0.97 etc.

Table 3. Comparison of proposed HOSM model with existing approaches

S No	Reference	Algorithms Involved	Accuracy	Dataset
1	Wu et al. (2022) [15]	Hybrid ML and DL model	88.46%	AI-TeleCare data
2	E.M. Senan et al. (2021)[16]	Random Forest with SVM model	97.30%	UCI-Repository
5	Proposed model	HOSM-WADET	98.77%	KAGGLE-UCI

Table 3.Shows the Comparison of proposed HOSM model with existing state of art approaches. [15] utilized hybrid idea of machine learning and deep learning model using AI telecare dataset and achieved the accuracy of 88.46%, [16] utilized random forest and SVM algorithm ensemble technique and achieved the accuracy of 97.3%. The proposed HOSM – WADET model achieved 98.77% accuracy.

6. Challenges

The major challenge of the proposed approach is that the division of unstructured data into required class and the time taken to optimize the unstructured data. further the system needs to be improved by evaluating deep learning-based Transfer learning approach for accurate results with degraded sampling time.

More parametric comparison often increases the accuracy of prediction on the other hand the role of few not relative parameters that doesn't make any big impact on the prediction quality need to be omitted in future implementations to reduce the processing delay.

7. Conclusion

One of the most devastating diseases that has a lasting impact on a person's life is chronic kidney disease. Problems that can be fatal if left untreated can arise from chronic kidney diseases. Using machine learning algorithms and a variety of physiological parameters, the proposed research will concentrate on analyzing chronic kidney diseases. They and the terms of their analysis of chronic kidney diseases employing the KAGGLE website-connected data set. The proposed method takes into account a variety of attributes from the CKD data set, as well as the relative parameters incorporated by mining the relativity extraction analysis and the blood glucose level. Between these dataset analyses, various clinical records, diagnostic procedures, and modifications that aid patients and doctors in the early stages of illness. The proposed HOSM model with WADET architecture after the optimization process achieved the accuracy of 98.77%when utilizing optimizing theCatBoostregressionachieved 97.5% accuracy. As a future enhancement Combining multiple machine learning models into ensemble algorithms need to be developed to improve the accuracy and sensitivity.

References

1. Evangelidis, N., Craig, J., Bauman, A., Manera, K., Saglimbene, V., & Tong, A. (2019). Lifestyle behaviour change for preventing the progression of chronic kidney disease: a systematic review. BMJ open, 9(10), e031625.
2. Arulanthu, P., & Perumal, E. (2020). An intelligent IoT with cloud centric medical decision support system for chronic kidney disease prediction. International Journal of Imaging Systems and Technology, 30(3), 815-827.
3. Shang, Ning, Atlas Khan, Fernanda Polubriaginof, Francesca Zanoni, Karla Mehl, David Fasel, Paul E. Drawz et al. "Medical records-based chronic kidney disease phenotype for clinical care and "big data" observational and genetic studies." Npj Digital Medicine 4, no. 1 (2021): 70.
4. Rashid, Junaid, Saba Batool, Jungeun Kim, Muhammad Wasif Nisar, Amir Hussain, Sapna Juneja, and Riti Kushwaha. "An augmented artificial intelligence approach for chronic diseases prediction." Frontiers in Public Health 10 (2022): 860396.
5. Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). A novel diabetes healthcare disease prediction framework using machine learning techniques. Journal of Healthcare Engineering, 2022.
6. Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., & Chen, B. (2019). A machine learning methodology for diagnosing chronic kidney disease. IEEE Access, 8, 20991-21002.
7. Nishat, Mirza Muntasir, et al. "A comprehensive analysis on detecting chronic kidney disease by employing machine learning algorithms." EAI Endorsed Transactions on Pervasive Health and Technology 7.29 (2021): e1-e1.
8. Makino, M., Yoshimoto, R., Ono, M., Itoko, T., Katsuki, T., Koseki, A., ... & Suzuki, A. (2019). Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. Scientific reports, 9(1), 11862.

9. Janani, J., and R. Sathyaraj. "Diagnosing Chronic Kidney Disease Using Hybrid Machine Learning Techniques." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12, no. 13 (2021): 6383-6390.
10. Ilyas, Hamida, Sajid Ali, Mahvish Ponum, Osman Hasan, Muhammad Tahir Mahmood, Mehwish Iftikhar, and Mubasher Hussain Malik. "Chronic kidney disease diagnosis using decision tree algorithms." BMC nephrology 22, no. 1 (2021): 1-11.
11. Shi, S. (2021). A novel hybrid deep learning architecture for predicting acute kidney injury using patient record data and ultrasound kidney images. Applied Artificial Intelligence, 35(15), 1329-1345.
12. Singla, J., Kaur, B., Prashar, D., Jha, S., Joshi, G. P., Park, K., ... & Seo, C. (2020). A novel fuzzy logic-based medical expert system for diagnosis of chronic kidney disease. Mobile Information Systems, 2020.
13. Saha, Anik, Abir Saha, and Tanni Mittra. "Performance measurements of machine learning approaches for prediction and diagnosis of chronic kidney disease (CKD)." In Proceedings of the 7th international conference on computer and communications management, pp. 200-204. 2019.
14. Chen, Guozhen, Chenguang Ding, Yang Li, Xiaojun Hu, Xiao Li, Li Ren, Xiaoming Ding, Puxun Tian, and Wujun Xue. "Prediction of chronic kidney disease using adaptive hybridized deep convolutional neural network on the internet of medical things platform." IEEE Access 8 (2020): 100497-100508.
15. Ma, F., Sun, T., Liu, L., & Jing, H. (2020). Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network. Future Generation Computer Systems, 111, 17-26.
16. Chen, Chi-Jim, Tun-Wen Pai, Hui-Huang Hsu, Chien-Hung Lee, Kuo-Su Chen, and Yung-Chih Chen. "Prediction of chronic kidney disease stages by renal ultrasound imaging." Enterprise Information Systems 14, no. 2 (2020): 178-195.
17. Rady, El-Houssainy A., and Ayman S. Anwar. "Prediction of kidney disease stages using data mining algorithms." Informatics in Medicine Unlocked 15 (2019): 100178.
18. Stiglic, Gregor, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. "Interpretability of machine learning-based prediction models in healthcare." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10, no. 5 (2020): e1379.
19. Nusinovici, Simon, Yih Chung Tham, Marco Yu Chak Yan, Daniel Shu Wei Ting, Jialiang Li, Charumathi Sabanayagam, Tien Yin Wong, and Ching-Yu Cheng. "Logistic regression was as good as machine learning for predicting major chronic diseases." Journal of clinical epidemiology 122 (2020): 56-69.
20. Belur Nagaraj, Sunil, Michelle J. Pena, Wenjun Ju, Hiddo L. Heerspink, and BEAt-DKD Consortium. "Machine-learning–based early prediction of end-stage renal disease in patients with diabetic kidney disease using clinical trials data." Diabetes, Obesity and Metabolism 22, no. 12 (2020): 2479-2486.
21. Jongbo, Olayinka Ayodele, Adebayo Olusola Adetunmbi, Roseline Bosede Ogunrinde, and Bukola Badeji-Ajisafe. "Development of an ensemble approach to chronic kidney disease diagnosis." Scientific African 8 (2020): e00456.
22. Wu, C. T., Wang, S. M., Su, Y. E., Hsieh, T. T., Chen, P. C., Cheng, Y. C., Tseng, T. W., Chang, W. S., Su, C. S., Kuo, L. C., Chien, J. Y., & Lai, F. (2022). A Precision Health Service for Chronic Diseases: Development and Cohort Study Using Wearable Device, Machine Learning, and Deep Learning. IEEE journal of translational engineering in health and medicine, 10, 2700414. https://doi.org/10.1109/JTEHM.2022.3207825.
23. Kanda, E., Suzuki, A., Makino, M. et al. Machine learning models for prediction of HF and CKD development in early-stage type 2 diabetes patients. Sci Rep 12, 20012 (2022). https://doi.org/10.1038/s41598-022-24562-2

24. YangGM,Li J, Zhang Y, Dong PY, Gurunathan S. A comprehensive review on the composition, biogenesis, purification, and multifunctional role of exosome as delivery vehicles for cancer therapy. Biomedicine & Pharmacotherapy. 2023 Sep 1;165:115087.
25. Samrat Kumar Dey, Khandaker Mohammad Mohi Uddin, Hafiz Md. Hasan Babu, Md. Mahbubur Rahman, Arpita Howlader, K.M. Aslam Uddin, Chi2-MI: A hybrid feature selection based machine learning approach in diagnosis of chronic kidney disease, Intelligent Systems with Applications, Volume 16, 2022, 200144, ISSN 2667-3053, https://doi.org/10.1016/j.iswa.2022.200144.
26. P. Chittora et al., "Prediction of Chronic Kidney Disease - A Machine Learning Perspective," in IEEE Access, vol. 9, pp. 17312-17334, 2021, doi: 10.1109/ACCESS.2021.3053763.
27. Ahmed, Aqeel, et al. "Malnutrition Among Pre-Dialysis Patients of Chronic Kidney Disease." Pakistan Journal of Medical & Health Sciences 16.07 (2022): 520-520.
28. Bai, Q., Su, C., Tang, W. et al. Machine learning to predict end stage kidney disease in chronic kidney disease. Sci Rep 12, 8377 (2022). https://doi.org/10.1038/s41598-022-12316-z
29. Durairaj, Sureshkumar, Kaoshik, Kaviya Srinivasan, Khang Wen Goh, Krishna Undela, Vijayakumar Thangavel Mahalingam, ChrismawanArdianto, Long Chiau Ming, and RajanandhMuhasaparur Ganesan. "Effect of L-carnosine in patients with age-related diseases: A systematic review and meta-analysis." Frontiers in Bioscience-Landmark 28, no. 1 (2023): 18.
30. Akchurin OM. Chronic Kidney Disease and Dietary Measures to Improve Outcomes. Pediatr Clin North Am. 2019 Feb;66(1):247-267. doi: 10.1016/j.pcl.2018.09.007. PMID: 30454747; PMCID: PMC6623973.
31. Vestergaard SV, Christiansen CF, Thomsen RW, Birn H, Heide-Jørgensen U. Identification of Patients with CKD in Medical Databases: A Comparison of Different Algorithms. Clin J Am Soc Nephrol. 2021 Apr 7;16(4):543-551. doi: 10.2215/CJN.15691020. Epub 2021 Mar 11. PMID: 33707181; PMCID: PMC8092062.
32. Hossain, M.M. et al. (2022) "Analysis of the performance of feature optimization techniques for the diagnosis of machine learning-based chronic kidney disease," Machine Learning with Applications, 9, p. 100330. Available at: https://doi.org/10.1016/j.mlwa.2022.100330.
33. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
34. Wu, C. T., Wang, S. M., Su, Y. E., Hsieh, T. T., Chen, P. C., Cheng, Y. C., Tseng, T. W., Chang, W. S., Su, C. S., Kuo, L. C., Chien, J. Y., & Lai, F. (2022). A Precision Health Service for Chronic Diseases: Development and Cohort Study Using Wearable Device, Machine Learning, and Deep Learning. IEEE journal of translational engineering in health and medicine, 10, 2700414. https://doi.org/10.1109/JTEHM.2022.3207825.
35. Senan, E. M., Al-Adhaileh, M. H., Alsaade, F. W., Aldhyani, T. H., Alqarni, A. A., Alsharif, N., ... & Alzahrani, M. Y. (2021). Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. Journal of Healthcare Engineering, 2021.