JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# Chronic Kidney Disease Prediction System using Machine Learning Approach

**Mrs. V. D. Jadhav[1], Mr. Somnath Shinde[2], Mr. Ganesh Kaldhone, Mr. Atharv Deshpande[4], Mr. Jeevan Koli[5], Mr. Shreyas Surwase[6]**

[1]Assistant Professor,[2]Student, [3]Student, [4]Student,[5]Student, [6]Student
[123456]Computer Science & Engineering,
[123456]Sveris College Of Engineering, Pandharpur, Solapur, Maharashtra, India

**Abstract :**Chronic Kidney Disease (CKD) has become a major health concern worldwide, affecting millions of people and often leading to fatal outcomes. Early detection and accurate prediction of CKD are crucial for timely intervention and effective management. In this study, we propose a machine learning-based prediction system for CKD, leveraging clinical and demographic data to predict the risk of CKD onset. The system utilizes a dataset comprising comprehensive clinical features, including age, gender, blood pressure, serum creatinine, and other relevant parameters.

We explore various machine learning algorithms, such as Random Forest, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN), and Logistic Regression, to develop and compare the performance of our prediction model. The dataset was preprocessed, including handling missing values, normalization, and feature selection, to ensure optimal model performance. Evaluation metrics, including accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC), were employed to assess the models' performance.

Our results indicate that the Random Forest algorithm outperforms other models with an accuracy of over 95% and an AUC-ROC of 0.98. The developed prediction system demonstrates promising results in identifying individuals at risk of developing CKD. The proposed system could be integrated into clinical practice to assist healthcare professionals in making informed decisions, leading to early detection and timely intervention, thereby improving patient outcomes and reducing the burden on healthcare systems.

**Keywords –** Chronic Kidney Disease (CKD), Machine Learning, Prediction System, Random Forest, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, Clinical Data, Early Detection, Healthcare, Risk Prediction, Feature Selection, Model Evaluation, Accuracy, Sensitivity, Specificity, Receiver Operating Characteristic Curve (ROC), Healthcare Management.

## 1. INTRODUCTION

Chronic Kidney Disease (CKD) is a global health issue that is growing at a steady rate. A person having CKD can only be saved by providing kidney transplantation or dialysis but as it involves huge medical expenditure it remains unaffordable to low and middle-income countries. Due to unaffordability in developing countries, more than 1 million people die annually from CKD as the larger population is not able to receive the required treatment on time. Therefore, early-stage diagnosis is critical to contain CKD at the initial stage which in turn can reduce the mortality rate and bring down the cost of treatment significantly. Recently, Machine learning algorithms have shown promising results in detecting CKD using patients' Electronic Health Records (EHRs). Apart from that researchers have explored different feature selection approaches such as filter, wrapper, embedded, and hybrid for extracting critical features for the prediction of CKD. But sometimes the existing feature selection methods in literature generally ignore the most relevant features resulting which in turn reduces the accuracy of the model.

In this project, we propose a two-stage feature selection method for the detection of CKD using a machine learning approach. In this method, we have used majority voting of filter-based and embedded feature selection technique for selecting the most significant features and hence named Majority Vote Feature Selection (MV-FS). We have also used Random Forest, Extreme Gradient Boosting (XGBoost), and Adaboost machine learning algorithms for classifying CKD patients.

In this project, we have used the chronic kidney disease dataset available on the UCI Machine Learning repository. For evaluating the machine learning models we have used accuracy, precision, sensitivity, F1-score, and area under the ROC curve (AUC) as our evaluation metrics.

At last, we have also interpreted the proposed model by analyzing the most critical features using Shapley Additive exPlanations (SHAP) feature importance plot, summary plot, dependence plot, and Partial dependence plots. With detailed model interpretation, our proposed MV-FS method can assist nephrologists and medical practitioners in diagnosing early-

stage CKD which can reduce the diagnostic cost and can shorten the diagnosis time significantly.

Further, we build an end-to-end Python-based web application using Flask microframework in which we can train the machine learning models, and predict CKD by selecting the best ML algorithm either by filling in medical details of the patient or uploading an excel file comprising details of multiple patients. The web app can assist doctors as it gives a model explanation for each of the predictions by plotting the Shapley plot and gives a warning indicator in case the patient is suffering from CKD.

The web app's front end involves HTML, CSS, and Javascript, and the back end is developed using Python and MySQL to save training data, model predictions, and model artifacts. The framework used in the web app is Flask.

## 2. LITERATURE SURVEY:

The identification of chronic kidney disease prediction using machine learning techniques and algorithms has gained significant attention in recent years due to its potential to streamline and improve the medical process. Several studies and research papers have explored various aspects of kidney disease prediction in different contexts. Here is a literature survey highlighting some key research in this field:

1. "Chronic Kidney Disease Prediction Using Machine Learning Algorithms: A Comparative Study" by Kong, D., Yu, H., and Zhang, X. (2021): This paper presents a comprehensive framework for improving crop production and soil health through the synergistic integration of IoT and ML technologies. The proposed system offers practical solutions to address the challenges faced by farmers and agricultural stakeholders.

2 "Chronic Kidney Disease Prediction using Random Forest with Grid Search Algorithm." Purwitasari, D., Hartama, D., and Handayani, P. W. (2020): This study focused on the use of the Random Forest algorithm with the Grid Search algorithm to predict Chronic Kidney Disease. The research demonstrated the effectiveness of this approach in accurately predicting CKD, highlighting its potential for implementation in clinical settings..

3 "Chronic Kidney Disease Prediction Using Random Forest Algorithm." by Kumar, P., and Vinothina, V. (2020 This research utilized the Random Forest algorithm to predict Chronic Kidney Disease. The study focused on the application of machine learning in healthcare, demonstrating the effectiveness of Random Forest in accurately predicting CKD.

4 Performance Analysis of Machine Learning Techniques in Chronic Kidney Disease Prediction." By Sivaranjani, K., and Rajalakshmi, P. (2020): The research compared the effectiveness of Random Forest, Support Vector Machine (SVM), and Decision Tree algorithms, providing insights into the most suitable algorithm for CKD prediction.

## 3. Requirement:

The Entire Development Process Has Been Subdivided Into Two: the Front End Development and the Backend Development. The Front End Comprises of the Visually Visible Parts Such as the Main Page, Admin login page, User Login Page, User Sign in Page and Home page. The Back End Contains the Database and Its Interaction With the Front-end.

1) Front End Development:

For front end HTML, CSS, JavaScript and Bootstrap is used. Our aim is to develop a system with responsive design and is user friendly. The design developed is simple and clean to understand by any user. HTML defined a structure of the website and for styling CSS is used. To add the functionality in the website JavaScript is used. Bootstrap enhanced the development process by minimizing the time required to develop a design.

2) Backend Development:

Backend is developed using JavaScript and Python. Machine learning is used for image processing and give the kidney disease predictions. Along with this patient recommendations, disease recommendation is provided to the patient on successful analysis of medical reports. GPU resources are utilized for faster training, and the LIME technique is applied to interpret model predictions.

## 4. RESEARCH METHODOLOGY:

This paper presents a comprehensive overview of the chronic kidney disease prediction using machine learning technologies to optimize chronic kidney disease. The application comprises various modules, each designed to address specific medical challenges, disease detection. The following sections provide detailed insights into the design, implementation, and evaluation of each module:

### A. The Application

1. **Data Preprocessing**: Data pre-processing is an important step of data preparation. In the healthcare domain, raw data often contains missing values, inaccurate data, noisy data, and this is due to measurement error, data entry error, or any other human error. To build a useful predictive model in the healthcare domain, handling data quality issues is a challenging task. The data pre-processing step consists of handling missing values, removing inconsistent data or noise from the data, handling extreme values, i.e., outliers and selecting optimal features before building a machine learning model. Therefore, data pre-processing involves five steps handling missing values, data transformation, outlier removal, data normalization or data scaling and feature selection. In this study, we have not removed the outliers from the dataset as the dataset is small and removing outliers would further reduce the records resulting in loss of information. Therefore, we have employed four main steps of data preprocessing: handling missing values, data transformation, and data normalization which are described in the below sub-sections. In our study, we have proposed a novel feature selection method for selecting most optimal feature subset which is a part of data pre-processing step but we have covered it in proposed methodology section.

2. **Missing Value Treatment**: In this step, we mainly handle missing values, inconsistent data, and noisy data. It is the most important step in building a machine learning pipeline because for building an accurate predictive model we need to input quality data. In this study, the dataset used consists of large number of missing values in all the input variables. Apart from that, the dataset is small so dropping missing values from the dataset results in loss of information. Therefore, in this study, we have used Predictive Mean Matching (PMM) based imputation for handling missing values, which is available under Multiple Imputations via Chained equations (MICE) package in R language. PMM method is based on the concept of multiple imputations. Multiple imputations are used in the case of missing at random (MAR), i.e., the probability of a

missing value is dependent on some of the observed data. PMM is a more effective method for handling missing values in comparison to commonly used mean imputation. The mean imputation has three significant disadvantages. Firstly, it assumes that missing values are missing completely at random (MCAR), i.e., the probability of a missing value is independent of complete observations . However, this is not true in every dataset, especially in healthcare data, where some observations are dependent on other observations. Secondly, mean imputation does not use the information of other observations, which may improve imputation accuracy. Lastly, it is biased as it imputes constant value in place of missing values.

3.**Data Transformation:** Data transformation is used to transform original data into the required format before building machine learning model. The machine learning model only accepts numerical values so data transformation needed to transform categorical values into appropriate numerical values using label encoding technique.

4.**Data Normalization**: Data normalization is the part of data preparation step of machine learning pipeline. The major objective of data normalization is to scale the numerical features into uniform scale so that numerical features having greater values does not dominate the smaller numeric value features. In this way, data normalization helps minimize the bias of higher numeric features in classifying different classes. In this study we have used min-max normalization, which transforms data into a specified range and preserves the original relationship among data values . In our study, min-max normalization transform the numeric value features into the range of (0, 1). The primary reason for using min-max normalization is that more than 50% of the features in the data have binary variables, i.e., (0, 1).

5. **Modelling Techniques**: In this project, we have used five machine learning algorithms Logistic Regression, Random forest, XG Boost, Support Vector Machines and AdaBoost.

**i) Logistic Regression:** Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means there would be only two possible classes 0 for failure and 1 for success.

**ii) Random Forest:** Random forest was first proposed by Breiman. It works on the principle of bootstrap aggregation also known as bagging. The bagging concept involves building multiple predictors to combine their predictions and generate one final aggregated predictor. The random forest is a more refined version of bagging as it selects only a subset of features at random for splitting each node in a tree instead of all the features. The random forest has also shown significant results in the healthcare domain.

**iii) Extreme Gradient Boosting:** XGBoost is a new variant of gradient boosting family of ensemble learning.It is a supervised machine learning algorithm and can be used for both regression and classification tasks. XGboost is the distributed, portable, and scalable boosting framework in machine learning. It has given state-of-the-art results in different problem domains, machine learning competitions in Kaggle and KDD Cup 2015. It is an enhanced version of the gradient boosting decision tree (GBDT). XGboost's objective function has two parts: a second-order loss function, which tells about the model's predictiveness in terms of training data, and a regularization term, which balances the model's complexity.

**iv) Support Vector Machine:** Support vector machines are a set of supervised learning methods used for classification, regression, and outliers detection. All of these are common tasks in machine learning. SVMs are different from other classification algorithms because of the way they choose the decision boundary that maximizes the distance from the nearest data points of all the classes. The decision boundary created by SVMs is called the maximum margin classifier or the maximum margin hyper plane.

**v) Adaptive Boosting (Adaboost):** The Adaboost algorithm works by adjusting the set of weights over the training set. Initially, equal weights are assigned to the instances of the training set. However, as the training progresses, the weights are increased for misclassified instances so that weak learners can better focus on wrong predictions in future iterations. It is relatively faster in execution in comparison to other algorithms and uses to give better results with no parameter tuning except the number of iterations.

6.**Evaluation Metrics**: In this study we will be using Accuracy, Sensitivity, Specificity, Precision, F1-score,
and AUC i.e., Area Under the Receiver Operating Characteristics (ROC) curve. The description of seven performance metrics is given below:

1. **Accuracy:** It is the ratio of total correct predictions made by the model to the total number of observations.
2.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

1. **Precision:** Precision is defined as the ratio of actual positive observations to the total number of positive predictions made by the model.

$$Precision = \frac{TP}{TP + FP}$$

2. **Sensitivity**: It is defined as the ratio of positive predictions made by the model to the actual number of positive observations.

$$Sensitivity = \frac{TP}{TP + FN}$$

3. **F1-score**: It is defined as the harmonic mean of precision and sensitivity or recall. It measures the balance between precision and recall. The ideal value of the F1-score is 1 whereas the worst value is 0.

$$F1 - score = 2 * \frac{(Precision * Recall)}{Precision + Recall}$$

4. **AUC:** AUC is the area under the ROC curve and is numerically defined as follows:

$$AUC = \frac{True\ Positive\ Rate + True\ Negative\ Rate}{2}$$

Where,

True Positive Rate = TP / (TP+FN),

True Negative Rate = TN / (TN+FP)

7.**Model Interpretation:** In this study, we will explain the prediction of best performing machine learning classifier using model interpretation techniques: SHapley Additive exPlanations (SHAP).

**7.1 SHapley Additive exPlanations (SHAP)**

SHAP was first introduced by (Lundberg and Lee, 2017). The main objective of SHAP is to explain the model's prediction by showing the contribution of each feature in the prediction. It works on the principle of game theory where the success of a team is determined by the contribution of each player in the game. SHAP values are similar to feature importances of the linear machine learning models with multicollinearity (Lundberg and Lee, 2017).

In this method model retraining happened on all the subset of features S ⊆ F, where F denotes the subset of all the features in the dataset. This method assigns a feature importance value to each of the features which shows effects on the model output. In this method, a model is trained with each of the features present iteratively and a corresponding model is trained without including that feature for comparing the prediction output of the model. The exclusion of a feature depends on the presence of other features in the dataset, so the preceding differences are calculated for all possible combinations of feature subsets S ⊆ F \ {i}.

After computing the preceding differences of feature subset, SHAP values are calculated and used as feature attributions. The weighted average of all the possible differences is given by following equation 1.

$$\phi_i = \sum_{S \subseteq F \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (1)$$

Here f(S) represents output of the model to be explained using a set S of features, and F is the complete set of all features. The SHAP value of feature i (ϕi) is computed as the average of its contributions across all possible combination of a feature set.

8.**Proposed Methods**: Feature selection (FS) is the most crucial step in the machine learning pipeline. It filters irrelevant features from the dataset, which helps machine learning algorithms reduce training and execution time, build simple models, interpret features, and improve the data quality and data understanding. In this project, we propose a All Vote feature selection method to classify CKD patients using machine learning algorithms.

The method comprises three filters-based FS methods: Pearson Correlation Coefficient, Chi-square and, Mutual Information (MI), and one embedded logistic regression FS method.

## 5. RESULTS AND DISCUSSION:

In this section, we have demonstrated the results of our experiment. In this study, we have performed 10-fold stratified cross-validation to estimate the performance of machine learning models. In 10-fold cross-validation, entire training data randomly partitioned into 10 equal subsamples in which one sample was used for validation and the rest 9 sample used to train the model.

The process repeated 10 times i.e., the number of folds, with every 10 subsamples used once for validating the model. The results from 10 folds averaged to estimate the performance of the model. The 10-fold cross-validation accuracy of machine learning models with and without feature selection is shown in Table 3. From the table, we can observe that with the proposed feature selection method the average 10-fold accuracy of the Random Forest model increased to 97.81% from 96.87% whereas the accuracy of Adaboost, SVC, and XGBoost reduced post-feature selection.

**Table 3: Comparison of 10-fold cross-validation accuracy**

| Machine Learning Algorithm | Without Feature Selection | With Feature Selection |
|---|---|---|
| | Accuracy (%) | Accuracy (%) |
| Logistic regression | 96.25 | 96.56 |
| Random forest | 96.87 | 97.81 |
| XGBoost | 97.50 | 96.25 |
| SVC | 97.81 | 97.50 |
| Adaboost | 98.43 | 97.50 |

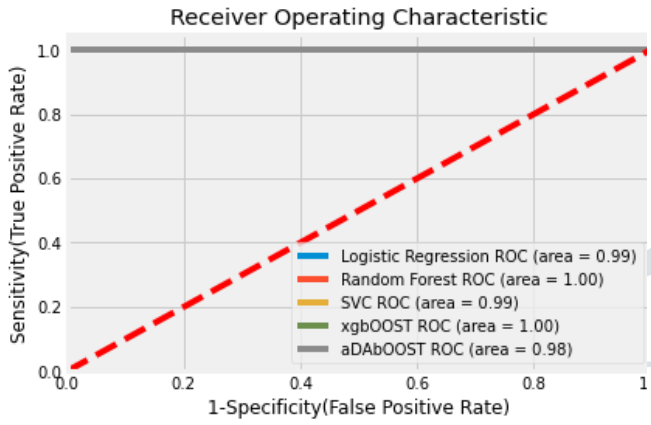
Table 4 shows the comparison of machine learning models post feature selection. After the proposed feature selection, 10 optimal features are selected i.e., Hemoglobin, specific gravity, red blood cell count, diabetes Mellitus, hypertension, appetite, pedal edema, packed cell volume, red blood cells, and albumin. As per the results, Random forest and XGBoost attained highest accuracy, precision, sensitivity, and F1-score of 100%.

**Table 4: Comparison of machine learning models post feature selection**

| Algorithm | Accuracy (%) | Precision (%) | Sensitiv (%) |
|---|---|---|---|
| Logistic Regression | 98.75 | 100 | 98 |
| Random Forest | 100 | 100 | 100 |
| Adaboost | 97.50 | 100 | 96 |
| SVC | 98.75 | 100 | 98 |
| XGBoost | 100 | 100 | 100 |

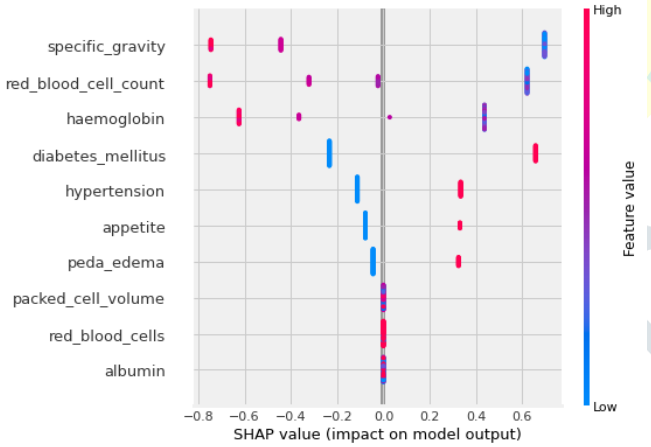
The below figure demonstrated the Receiver Operating Characteristics (ROC) curves corresponding to five machine

learning classifiers i.e., Logistic regression, Random Forest, Adaboost, Support vector machine, and XGBoost. The closer the curve follows the path in the direction of the top-left region, the more generalized the modeling method. The ROC curve of the Random Forest and XGBoost model followed the extreme top-left path and achieved the highest area under the ROC (AUC) value of 1.00 which outperformed other machine learning classifiers.




As we found that random forest and XGboost are the best models based on different performance metrics. So we have selected XGboost as our main model and further, we have interpreted the model prediction using Shapley force plot and summary plot.

**1.SHAP Summary plot:**

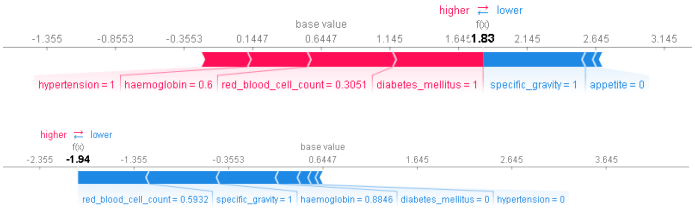



SHAP summary plot of a 10 features XGBoost CKD prediction. The higher the SHAP value of a feature, the higher is the log odds of heart disease in this CKD prediction model. Every patient in the dataset is run through the model and a dot is created for each feature attribution value, so one patient gets one dot on each feature's line.

Dot's are colored by the feature's value for that patient and pile up vertically to show density. In above plot we see that **specific_gravity** is the most important risk factor for CKD patients. The lower values of specific_gravity leads to CKD, whereas for normal patients its mostly of higher lower values.

Higher values of **diabetes_mellitus** increases the risk of CKD whereas its lower values decreases the chances of CKD. Similarly, higher values of **hypertension**, **appetite** and **peda_edema** increases the risk of CKD whereas lower values signify normality.

**2.SHAP Force plot**




The above graph is generated when we applied SHAP algorithm on instance number 17 and 9 from our test set. In first plot, we predicted 1.83, whereas the base_value is 0.6447 which means model predicted that patient is suffering from CKD.

Important to note that, feature values causing increased predictions are in **pink**, and their visual size shows the magnitude of the feature's effect. Feature values decreasing the prediction are in **blue**. The biggest impact comes **diabetes_mellitus** =1, while **specific_gravity**=1 value has the effect of decreasing the prediction.

**6.Diagram:**

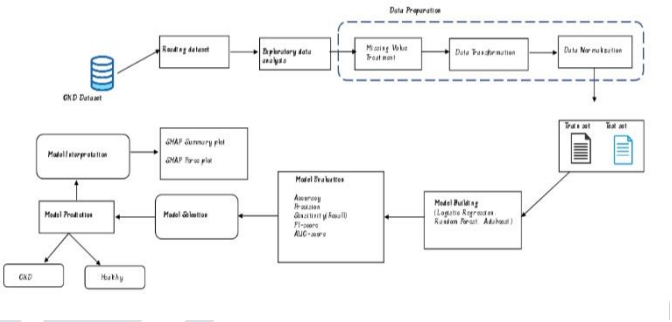


**Fig.6.1. Control Flow Diagram**

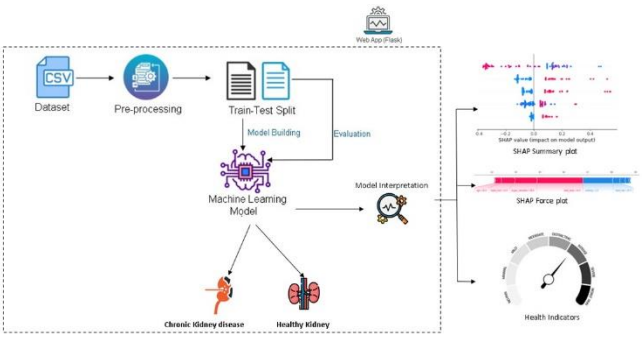



**Fig.6.2. System Architecture**

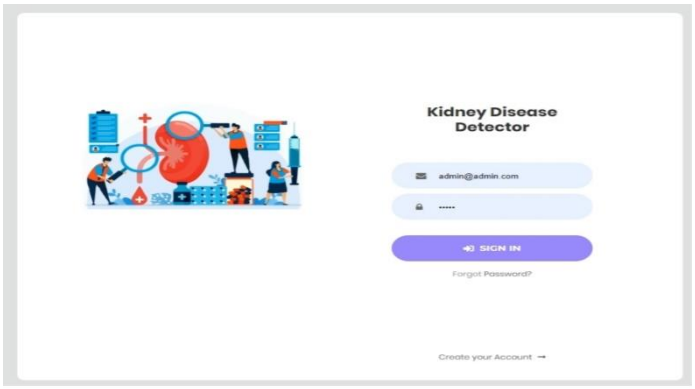



**Fig6.3. Login Page**

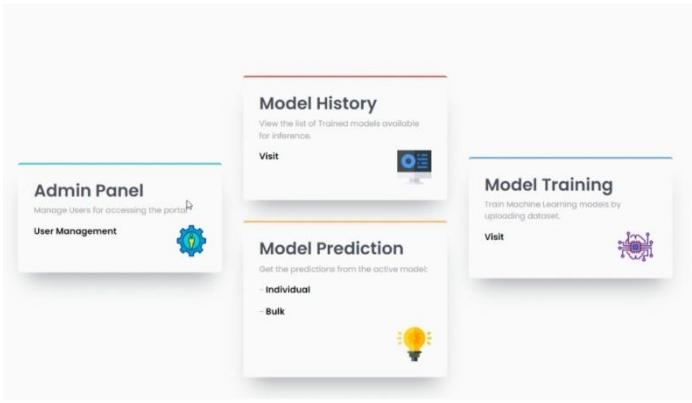



**Fig6.4. Admin Panel**

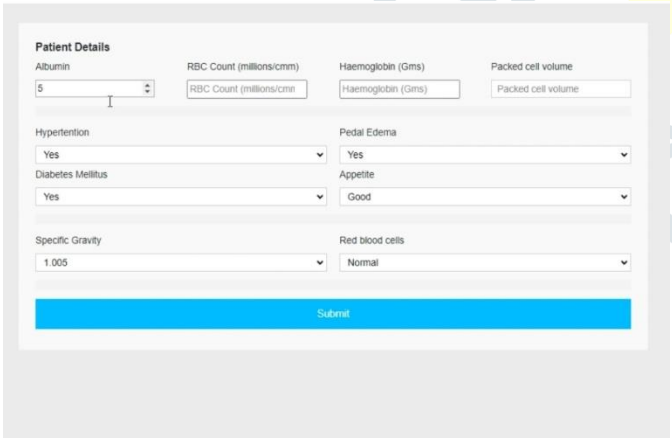



**Fig6.5. Patient Details Page**

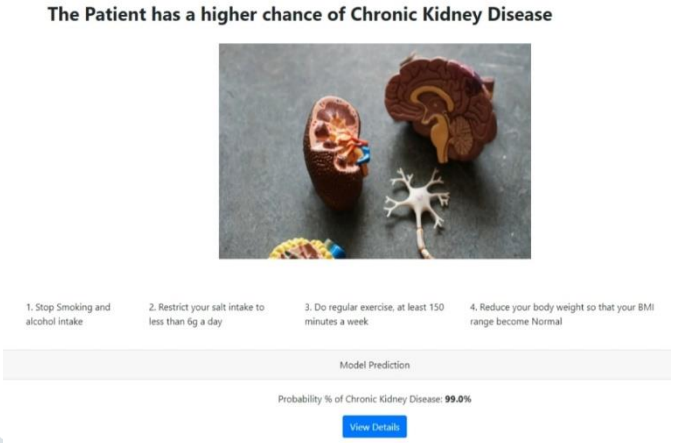



**Fig6.6. Output/Prediction Page**

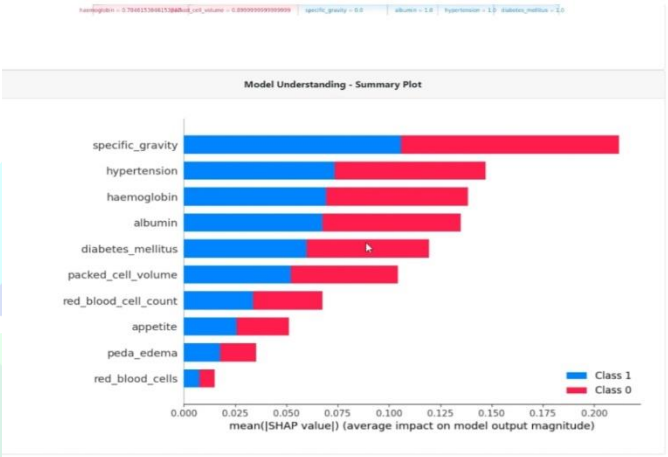



**Fig6.7. Summary Plotting Page**

## 7.Conclusion:

This project demonstrated a all vote feature selection method AV-FS for the prediction of CKD using five machine learning algorithms Logistic Regression, Random Forest, Adaboost, Support vector Machine, and XGBoost. The dataset used in this study was taken from UCI Machine Learning Repository. The proposed MV-GWO method selected 10 significant features i.e., Hemoglobin, specific gravity, red blood cell count, diabetes Mellitus, hypertension, appetite, pedal edema, packed cell volume, red blood cells, and albumin. The machine learning classifiers Random Forest and XGBoost resulted in higher performance with the proposed feature selection method whereas Adaboost has shown better performance with all the features. Further, we have also used the SHAP force plot and summary plot to analyze the effect of the top 10 critical features and explained how these features are contributing to the prediction of CKD.

## 8. Acknowledgment:

I would Like to acknowledge Hod of CSE Department (Dr. Swati. P. Pawar) and our guide Mrs. V. D. Jadhav ma'am of Sveri's College of Engineering, Pandharpur that provided me with the idea of project. All team members and staff also helped me a lotto give important information regarding how the system work and can help peoples.

**9.References:**

[1] Kong, D., Yu, H., & Zhang, X. (2021). Chronic Kidney Disease Prediction Using Machine Learning Algorithms: A Comparative Study. Journal of Healthcare Engineering.

[2] Purwitasari, D., Hartama, D., & Handayani, P. W. (2020). Chronic Kidney Disease Prediction using Random Forest with Grid Search Algorithm. Journal of Physics: Conference Series.

[3] El Hajji, S., Amane, M., & Elachqar, A. (2019). Comparative Study of Different Machine Learning Algorithms for Chronic Kidney Disease Prediction. International Journal of Computer Applications.

[4] Sagheer, A., Waseem, M., & Mehmood, Z. (2020). A Comparative Analysis of Machine Learning Techniques for Chronic Kidney Disease Prediction. Journal of Healthcare Engineering.

[5] Kumar, P., & Vinothina, V. (2020). Chronic Kidney Disease Prediction Using Random Forest Algorithm. 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA).