Predicting the Kidney Diseases by Using Machine Learning Techniques

N. Sreenivasa[*1], Sudesh Pawaar[2], Shaurya Sparsh[3], P. Ramesh Naidu[4]

[1,2,3,4]*Department of Computer Science and Engineering, Nitte Meenakshi Institute of Technology, Bangalore-560064, Karnataka, India*

sreenivasa.n@nmit.ac.in, Isudesh008@gmail.com , shauryasparsh912@gmail.com, ramesh.naidu@nmit.ac.in

Abstract. CKD (Chronic Kidney Diseases) is a persistent medical state categorized by the kidney damage that hinders their ability to effectively filter blood. Over time, this progressive disease can result in kidney failure. This project compares the performance of the Support Vectos Machines (SVM), logistic regression and Decision Tree algorithms for predicting the risk of CKD. In this project, the dataset utilized comprises a total of 25 attributes, consisting of 11 numerical features and 14 nominal features. In the training of machine learning algorithms for prediction, all 400 instances from the dataset are utilized. Among these instances, 250 are labeled as CKD cases, indicating the presence of chronic kidney disease, while the remaining 150 instances are categorized as non-CKD cases, denoting the absence of the condition. We utilized the UCI dataset, which underwent preprocessing to handle missing data. Using Python, we trained and built Support Vectors Machines (SVM), Logistic Regression, and Decision Tree models. The accuracy achieved with SVM was 97.3%, Logistic Regression was 93.8%, and Decision Tree yielded 95%, which are notable results.

Keywords— Machine *learning, Chronic Kidney Disease, Support Vectors Machines, Decision Tree, Logistic regression.*

1 INTRODUCTION

CKD is an advancing ailment characterized by a gradual decline in the kidney's capacity to eliminate waste substances and surplus fluids from the circulatory system. Consequently, harmful substances accumulate within the body, giving rise to various health Complications such as hypertension, anemia, skeletal abnormalities, and

* Corresponding author: *sreenivasa.n@nmit.ac.in*

cardiovascular disorders. CKD has a substantial impact on public health and affects numerous individuals worldwide. With the high cost of Renal Replacement Therapy (RRT) and limited availability of treatment facilities in developing countries, timely detection of CKD is crucial for reducing economic burdens and optimizing treatment outcomes. In this context, machine learning techniques offer valuable tools for early CKD detection, enabling prompt interventions. By facilitating early diagnosis and management, these techniques contribute to reducing mortality rates and enhancing the quality of life for individuals affected by CKD (Figure 1). The most effective way to measure kidney's function and determine the phase of chronic kidney disease is through a test called Glomerular Filtration Rate (GFR).

Stage	**Description**	**GFR (Ml/min)**
-	At increased risk for CKD	>=90 with risk factors
1	Kidney damage with normal or increased GFR	>=90
2	Mild decrease in GFR	60-89
3	Moderate decrease in GFR	30-59
4	Severe decrease in GFR	15-29
5	Kidney Failure	<15 or dialysis

Figure 1: Chronic kidney disease (CKD) categories

There are several factors that can be used to calculate GFR, including the level of serum creatinine and/or cystatin C, along with demographic factors like age, race, and gender. Detecting the disease early provides a better opportunity to prevent its progression. CKD is categorized) into five stages, which are determined by evaluating the patient's estimated glomerular filtration rate (eGFR) and the extent of proteinuria. Stage 1: In this stage, there is kidney damage, but the kidneys are still functioning norally. The GFR (Glomerular Filtration Rate) is above 90 ml/min/1.74m², indicating good kidney function. Stage 2: Kidney function is mildly decreased in this stage. The GFR varies from 60-89 ml/min/1.73m². Although there is some impairment, the kidneys can still adequately filter waste and fluids from the bloodstream. Stage 3: Moderate decrease in kidney function characterizes this stage. This stage is further classified into two sub-stages: 3a and 3b. In the stage 3a, the GFR is in the range of 45-59 ml/min/1.73m², while in the stage 3b, the GFR ranges from 30-45 ml/min/1.73m². At this point, kidney damage becomes more noticeable, and treatment strategies are crucial to slow down disease progression. Stage 4: Severe decrease in kidney function is observed in the Stage 4. The GFR falls between 15-29 ml/min/1.73m². Significant kidney damage is present, and symptoms may become more pronounced. Medical intervention is required to manage complications and prepare for potential kidney replacement therapy. Stage 5: This is the final stage. Kidney function is severely impaired, with GFR less than 15 ml/min/1.74m². At this point, the kidney will lose their ability to effectively filter waste and excess fluids from the body, necessitating kidney replacement therapy, such as dialysis or transplantation, for survival [16].

It's essential to note the descriptions are common and can vary depending on distinct circumstances. Healthcare professionals closely monitor patients with CKD to provide personalized care based on their specific condition and needs. In this project, we aim to

develop a predictive model for CKD using ML techniques. We have a dataset of 400 CKD patients with 25 attributes. We will use four ML algorithms, Support Vectors Machines, Decision Tree and logistic regression to build a predictive model for CKD risk The SVMs are highly effective machine learning algorithms utilized for classification and regression purposes. Their objective is to classify data points into distinct classes and identify the optimal hyper-plane that increases the border between these classes. SVMs are capable of handling linear and non-linear problems through the utilization of different kernel functions. They exhibit resilience against over fitting and are efficient in processing high-dimensional data. SVMs find wide application across various domains, known for their versatility and accuracy Decision trees, on the other hand, are supervised machine learning algorithms commonly employed for grouping and regression jobs. They function by recursively subdividing the data into minor subsets based on the values of input variables. Each division is determined by specific criteria, such as maximizing information or minimizing impurities, thereby creating a tree-like structure. Internal nodes of the tree represent decision rules based on input variables, while the leaf nodes represent the predicted classes or values. Decision trees are intuitive and easily interpretable and can handle both categorical and numerical data. Due to their simplicity and ability to discern complex relationships within data, decision trees find extensive application across various fields. Logistic regression, a statistical model frequently employed in machine learning, is particularly useful for binary classification problems. It aims to estimate the probability of an input belonging to one of two possible classes. By modeling the relationship between input variables and binary outcomes, logistic regression enables the prediction of class probabilities. Logistic regression belongs to the generalized linear model family and utilizes the logistic function (sigmoid function). This transformation allows the linear model's output to be mapped to probability values ranging from 0 to 1. Through the application of the logistic function, logistic regression effectively captures the association between input variables and the probabilities of binary outcomes.

2 RELATED WORK

Researchers have discovered the use of various classification algorithms to predict chronic kidney disease (CKD) and have obtained promising results.1]. Investigated different stages of CKD according to its gravity, and their analysis exposed that the RBF algorithm out-performed the other classifiers with an accuracy of 85.3%. In the research [2], they evaluated the performance of 12 classification algorithms on a datasets containing 410 records and 26 attributes. Their findings revealed that the decision tree techniques achieved remarkable results, with an accurateness of 98.8%. Additionally, the decision trees demonstrated a sensitivityof 0.9820, precision of 1, and specificity of 1In the research [3] addressed the issue of missing values in a dataset of CKD by performing a re-calculation procedure on CKD stages and replacingmissing values with recalculated values. Asif Salekin and John Stankovic [4] successfully employed machine learning algorithms to detect CKD. Their approach achieved accurate results on a dataset of400 accounts and 26 attributes. They utilized k-nearest neighbor, the random forest, and neural networks algorithm, along with the feature reduction wrapper methods. Pinar Yildirim [5] investigated the consequences of classimbalance during the training of a neural network algorithm for decision-making in CKD. The study revealed that employing sampling algorithms has the potential to improve the performance. It was observed that the rate of

learning, as a crucial parameter, had a significant effect on the multilayer perceptron model. Gunarathnes W.H.S.D et al. [6] conducted a comparative analysis of ML techniques and determined that the Multi class Decision Forest algorithms exhibited superior accurateness compared to other models. The Multiclass Decision Forest achieved an accurateness of approximately 99.3% when applied to a reduced datasets consisting of 15 attributes. Charleonnan et al. [7] compared predictive models (KNN, SVM, LR, DT) on an Indian CKD dataset. They found that SVM exhibited the highest accuracy (98.3%) and sensitivity (0.99), making it the top-performing classifier.Salekin and Stankovic [8] evaluated classifiers like RF, K- NN, ANNs on a datasets consisting of 410 instances using feature selection and constructing models with 5 features. RF had the highest accuracy (98%) and RMSE of 0.12. Tekale et al. [9] utilized a preprocessed dataset consisting of 400 instances and 14 features to predict CKD. SVM and Decision tree algorithms were employed for classification. The results indicated that the SVM model achieved accuracy of 96.75%, indicating its effectiveness in CKD prediction. Xiao et al. [10] proposed a CKD progression prediction model using multiple ML algorithms. Results of Logistic regression where best with an AUC of 0.873, sensitivity of 0.83, and specificity of 0.82 on a dataset of 551 patients with 18 features. Mohammed and Beshah [11] developed the self-learning data-based system for diagnosing and treating early stages of CKD using machine learning. They achieved anoverall accuracy of 91% using a decision tree algorithm on a small dataset, creating a prototype for patient advice.

Priyanka et al. [12] predicted CKD using the naive Bayes algorithm and tested other algorithms such as KNN, SVM, decision tree, and ANN. They found that naive Bayes had the best accurateness of 94.7% compared to the other algorithm. A study was done by Almasoud and Ward [13] to determine how well machine learning algorithms canforetell chronic renal disease. To find these features, they used feature selection methods like ANOVA, Pearson correlations, and Cramer's V tests. For modelling, their study used the SVM, LR, GB, and RF algorithms. With an F- measure of 99.2, their research revealed that Gradient Boosting had the highest accuracy. Yashfi [14] proposed using machine learning algorithm to predict the risk of CKD by analyzing data from CKD patients. Artificial Neural Network and Random Forest models were employed, utilizing 20 out of 25 features. The study found that Random Forest achieved the maximum accuracy, reaching 97.12%.Rady and Anwar [15] compared the performance of four algorithms (MLP, PNN, SVM, RBF) for predicting kidneys disease stages. The dataset used in their investigation was small and contained limited features. The findings of the studyrevealed that the Probabilistic Neural Network (PNNs) algorithm achieved the maximum overall, reaching 96.7%.

3 METHODOLOGY

The Figure 2 depicts the architecture flow as it applies various classifications of the model. To raise accuracy and select the best classifiers, above model is used. The prediction datasets for CKDs is used by the model.

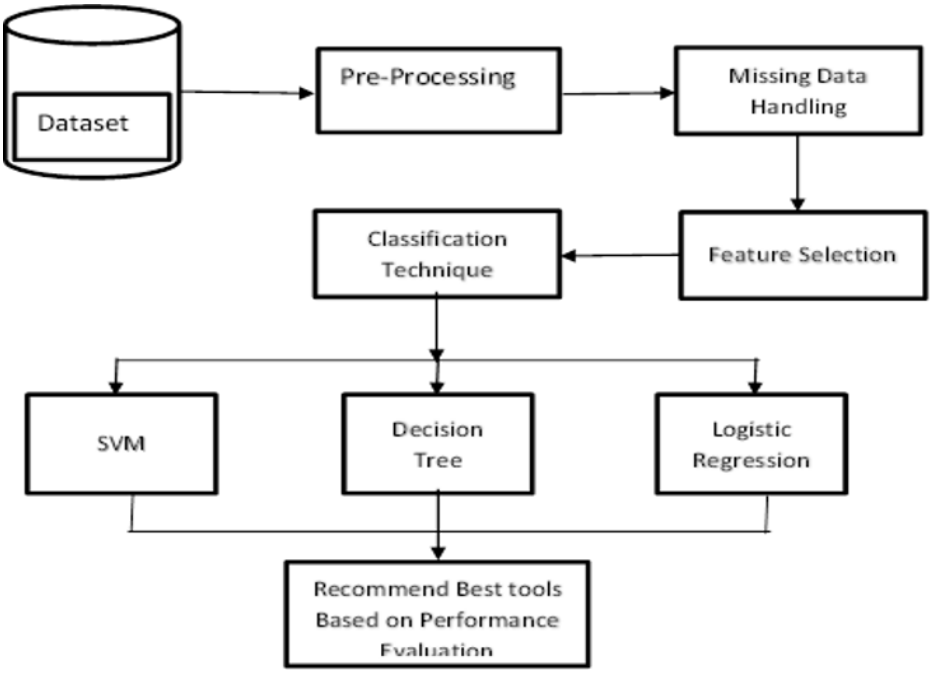

Figure 2. Methodology

Following features selection and pre-processing, the SVM (Support Vector Machine), DT (Decision Tree Naive Bayes), and RT (Random Forest), algorithms are useful. For the performance of the model assessment, initially, find TP, FP, TN, and FN as true positive, false positive, true negative, and improper positive, respectively. The true positive mentions to the total number of cases positively forecasted as required, untrue positive deals with the total number of instances incorrectly predicted as necessary, and so on.The following are the four dimensions that can be attained: precision, recall, F1 measure and accuracy. Precision=True Positive/(True Positive + False Positive)

4 RESULTS & ANALYSIS

Dataset: This method makes use of the CKD datasets from the UCI ML Repository, which incorporates 24 functions and 1 target variables. The target variable is binary, classified as both "sure" or "no" for chronic kidney ailment. The dataset carries 25 attributes, along with 11 numerical and 14 nominal features. The whole dataset of four hundred instances is applied to train system learning algorithms for making predictions. Of the 400 sample, 250 are categorized as having CKD and a 150 are classified as non-CKD. diverse attributes are covered within the dataset, such as age, blood stress, serum creatinine, and haemoglobin, among others. those functions are applied to teach and verify the performance of diverse system studying algorithms in predicting the likelihood of chronic kidney diseases.

The attributes present inside the dataset are indexed in Table 1.

Table 1 Descriptions of each feature in the main CKD datasets.

CKD dataset	**CKD dataset with PCA**	**Attributes meaning**	**Category**	**Scale**	**Missing**
age	-	age	Numerical	Years	9
bp	-	Blood Pressure	Numerical	Mm/Hg	12
sg	sg	Specific gravity	Nominal	1.005 to 1.025	47
al	al	Albumin	Nominal	0 to 5	46
su	su	Sugar	Nominal	0 to 5	49
rbc	-	Red blood cells	Nominal	Abnormal/Normal	152
Pc	-	Pus Cell	Nominal	Abnormal/Normal	65
pcc	-	Pus cell clumps	Nominal	Not Present	4
ba	-	Bacteria	Nominal	Present	4
bgr	bgr	Blood glucose random	Numerical	Mgs/dl	44
bu	-	Blood Urea	Numerical	Mgs/dl	19
sc	sc	Serum creatinine	Numerical	Mgs/dl	17
sod	-	Sodium	Numerical	mEq/L	87
pot	pot	Potasium	Numerical	mEq/L	88
Classification	Classification	Class	Nominal	Not CKD,CKD	0

The Data Sets has medical related variables which can be related to the CKD presence. Few variables can be perhaps more significant for this model, and after scrutiny few of them are correlated, so it's suggested to examine the datasets and choose the best approach's based on distinct needs. One attributes is the binary result which tells about patient actually developing the condition or not. Hence these data sets help to train supervised algorithm. Data set has a lesser percentage of missing values, which people can decide to fill out with either median values, or leave them blank to allow the algorithm learn without noise data.

For the efficient performance of data mining method this is vulnerable to over fitting. Techniques like pruning and ensemble approaches like random forests are frequently employed to overcome this. Overall, decision trees provide a flexible and effective method to machine learning that may be used in a variety of applications high- quality data is essential. In the case of the CKD dataset, missing variables need to be filled in and continuous characteristics may need to be converted into discrete ones. The dataset contains noisy and missing values, which requires data preprocessing to improve its behavior.

Pre-processing: Adjusting the raw data to make compatible for the utilization in a machine learning model. The data preprocessing steps is shown in Figure 3, which highlights a considerable number of lost values in the datasets. The Table 1 provides detailed information on each variable. NaNs are present in the data of the three datasets and numerical features need to be converted to floats. All rows containing NaNs are removed, without a specific threshold, to comply with the given directive.

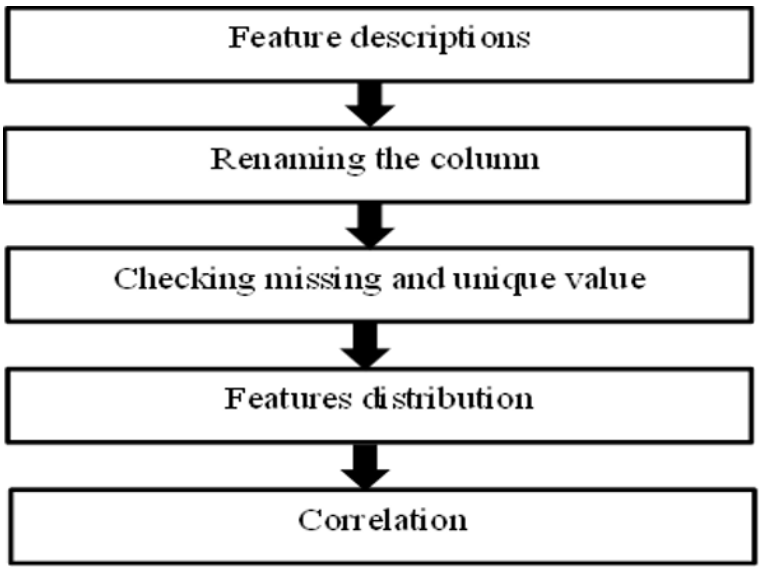

Figure 3. Establishing and Evaluating Individual Models

Decision Tree: A popular and simple machine learning technique used for both regression and classification applications is the decision tree algorithm (Figure 4). It creates a tree-like hierarchical structure, with internal nodes standing in for feature tests and leaf nodes standing in for class labels or predicted values. The algorithm makes decisions by following a path in the root nodes to specific leaf nodes depending on the expected values of the input data. Decision trees' interpretability, which makes it simple to visualize and comprehend the acquired principles, is one of their main advantages. Decision trees can recognize complex associations between variables since they can handle both categorical and numerical characteristics. Additionally, they can effectively handle missing values and noisy data.

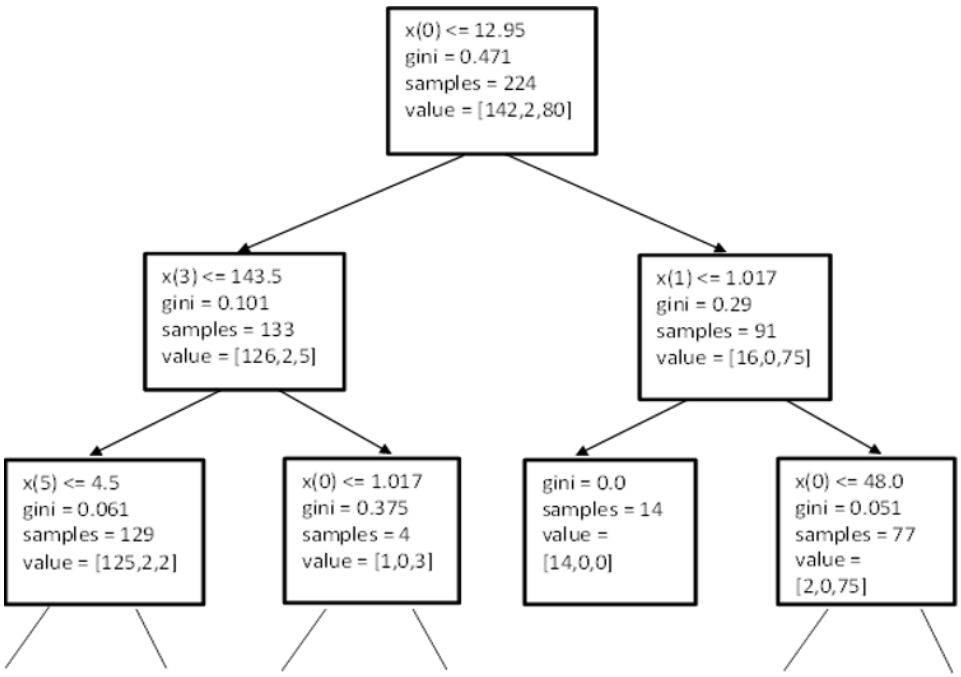

Figure 4. Simplified Structure of Decision Tree.

Support Vector Machine (SVM): A common and most effective machine learning method, the Support Vector Machine (SVM) algorithm excels in classification problems. It looks for the best hyperplane possible within the feature space to efficiently divide classes. SVM creates a strong decision boundary that adapts well to new data by maximizing the space between this hyper plane and the adjacent data points from each class. Theuse of kernel functions allows SVM to handle both linearly separable and non-linearly separable data (Figure 5), demonstrating versatility in handling high-dimensional feature spaces. Itssuccess has been attributed to its capacity to handle complex data distributions and strong theoretical underpinnings in avariety of fields, including image recognition, text classification, and bioinformatics.

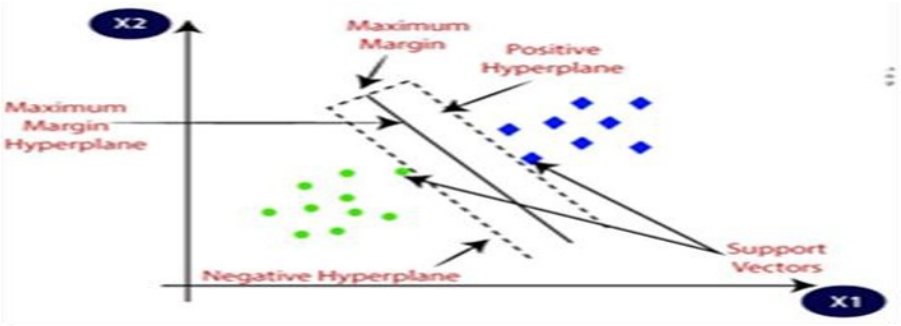

Figure 5. linearly separable and non-linearly separable data

Logistic Regression : In statistics and computer learning, the classification algorithm known as logistic regression is frequently utilized. Although it is typically used for binary classification tasks, this supervised learning technique can also be expanded to solve multi-class issues. The goal of logistic regression is to estimate the likelihood that a given instance will belong to agiven class. It achieves this by utilizing a logistic function, often known as sigmoid functions, to represent the relation between the input attributes and probability. The coefficients of the input features are estimated via logistic regression through a procedure called maximum likelihood estimation. These coefficients are modified during training to reduce the discrepancy between the anticipated probability and the actual class labels. The effectiveness, interpretability, and simplicity of logistic regression are valued qualities. It can handle both numerical and categorical information, and because its results can be translated into probabilities, it is helpful for applications requiring decision-making. The performance of logistic regression(Figure 6) may be constrained in increasingly complicated situations since it presupposes a linear relation between the input data and the log-odds of the target class.

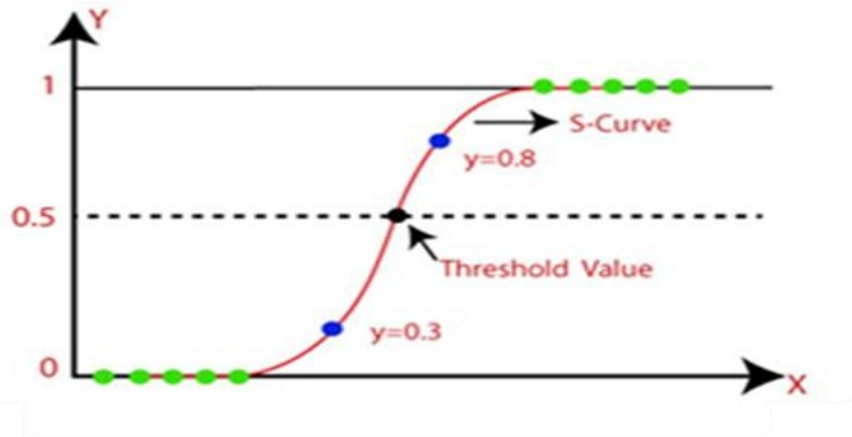

Figure 6. Logistic regression

Developing machine learning algorithms for the early detection of Chronic Kidney Disease (CKD) was the goal of this study. Using data from CKD patients, a variety of methods, including Support Vectors Machine, Logistic Regression and Decision Trees were used to train and verify models. The main performance criterion used to assess the algorithms' performance was precision. The Support Vectors Machine technique beat Decision Trees and Logistic Regression in this study's specific medical situation for predicting CKD, according to the results. Overall, the study shows how machine learning has the potential to improve CKD early diagnosis.

The research involved an Analyses of 24 numerical and nominal characteristics of 400 patients with CKD were performed. There were some patients with unreported test findings, nevertheless, and these were resolved by computational techniques. Missing numerical values were handled using the mean approach, and missing nominal values were handled using the mode method.

Both positive and negative correlations between various features were found by the investigation. For instance, whereas sugar and blood glucose showed positive relationships at random, specific gravity showed positive correlations with packed cells volume, hemoglobin, and red blood cell count. Additionally, serum creatinine and hemoglobin showed positive associations with blood urea. On the other hand serum creatinine showed a negative association with salt, while albumin and blood urea showed negative correlations with hemoglobin, packed cell volume, and the number of red blood cells.Regarding the performance of classification algorithms, Support Vector Machines (SVM) achieved an accuracy of 97.3% with an error rate of 2.88%. The precision, recall, and F-measure for SVM were all 0.97. Logistic Regression attained an accuracy of 93.8% with an error rate of 3.4%. The precision, recall, and F-measure for Logistic Regression were all 0.96. Decision Tree yielded an accuracy of 95% with an error rate of 2.1%. The precision, recall, and F- measure for Decision Tree were all 0.94.

No.	**Model/Method**	**Accuracy**
1	Support Vector Machines (SVM)	97.3%
2	Logistic Regression	93.8%
3	Decision Tree	95%

5 CONCLUSION

Detecting diseases at an early stage plays a vital role in preventing their spread, especially in countries like Bangladesh where medical expenses are high, and people face significant challenges. The cases of Chronic Kidney Disease (CKD) in the Bangladesh is growing rapidly, motivating our objective to develop a systems for predicting the risk of CKD. For our research work, we utilized the UCI dataset, which underwent preprocessing to handle missing data. Using Python, we trained and built

Support Vectors Machines (SVM), Logistic Regression, and Decision Tree models. The accuracy achieved with SVM was 97.3%, Logistic Regression was 93.8%, and Decision Tree yielded 95%, which are notable results. Our proposed method enables early-stage risk prediction of CKD. However, one of the main challenges we encountered was collecting real-time datasets and handling missing values. The success of our system is attributed to our innovative idea. The two newest iterations in renal replacement technology, the implantable BAK and kidney regeneration technology, address the limitations of the AWAK and WAK devices in that both aim to provide full kidney functionality as well as improved patient mobility and autonomy. To enhance accessibility, we plan to develop a mobile application that can be easily used by individuals across different incomelevels since mobile phone usage is prevalent. While our current work involved utilizing the Chi-square test filter method for feature selection, we intend to explore wrapper methods in the future for more accurate feature extraction. Additionally, we aim to collaborate with sources providing real-time datasets to further improve the accuracy of our system.

REFERENCES

[1.] Prasanta Kumar Sahoo;Goraknath Kashyap Modali "An Effective Way to Identify Chronic Kidney Disease Using Machine Learning" 2023 International Conference on Emerging Smart Computing and Informatics (ESCI)

[2.] Debabrata Swain; Hardeep Patel, Kevin Patel, Vivek Sakariya, Nishtha Chaudhar "An Intelligent Clinical Support System For The Early Diagnosis Of The Chronic Kidney Disease" 2022 IEEE 2nd International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)

[3.] Prasanta Kumar Sahoo;Goraknath Kashyap Modali "An Effective Way to Identify Chronic Kidney Disease Using Machine Learning" 2023 International Conference on Emerging Smart Computing and Informatics (ESCI)

[4.] Debabrata Swain; Hardeep Patel, Kevin Patel, Vivek Sakariya, Nishtha Chaudhar "An Intelligent Clinical Support System For The Early Diagnosis Of The Chronic Kidney Disease" 2022 IEEE 2nd International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)

[5.] Women More Affected By Kidney Diseases, Women.prothomalo.com,2018.[Online].Available::// women.prothomalo.com/bangladesh/Women more-affected-by-kidney-diseases.[Accessed:20- Sep-2019].

[6.] Sujata Drall, Gurdeep Singh Drall, Sugandha Singh, Bharat Bhushan Naib, "Chronic Kidney Disease Prediction Using Machine Learning: A New Approach", International Journal of Management, Technology And Engineering, vol. 8, no. 5, pp.278- 287, 2018.

[7.] Vivekanand Jha," Chronic Kidney Disease Global Dimension and Perspectives", Lancet, National Library of Medicine, 2013

[8.] Siddeshwar Tekale, "Prediction of Chronic Kidney disease Using Machine Learning, International Journal of Advanced Research in Computer and Communication Engineering, 2018.

[9.] L. Rubini, “Early stage of chronic kidney disease UCI machine learning repository,”2015. [Online].Available: Kidney Disease. http://archive.ics.uci.edu/ml/datasets/Chronic

[10.] P. Yildirim, "Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction," in Computer Software and Applications Conference (COMPSAC), 2017 IEEE 41st Annual, 2017.

[11.] Khamparia, G. Saini, B. Pandey, S. Tiwari, D. Gupta, andA. Khanna, “KDSAE: chronic kidney disease classificationwith multimedia data learning using deep stacked autoen coder network,” Multimedia Tools and Applications, vol. 79, no. 47-48, pp. 35425–35440, 2019.

[12.] Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," in Healthcare Informatics (ICHI), 2016 IEEE International Conference On, 2016,

[13.] M. S. Wibawa, I. M. D. Maysanjaya, and I. M. A. W. Putra, “Boosted classifier and features selection for enhancing chronic kidney disease diagnoses,” in Proceedings of the 2017

[14.] 5th international conference on cyber and IT service management (CITSM), pp. 1–6, IEEE, Denpasar, Indonesia, August 2017.

[15.] Ene-Iordache B, Perico N, Bikbov B, et al. Chronic kidney disease and cardiovascular risk in six regions of the world (ISN-KDDC): a cross-sectional study. Lancet Glob Health 2016; 4: e307–19.

[16.] Wong CW, Wong TY, Cheng CY, Sabanayagam C. Kidney and eye diseases: common risk factors, etiological mechanisms, and pathways. Kidney Int 2014; 85: 1290–302.