Journal of Mines, Metals and Fuels, 71(5): 632-639; 2023. DOI: 10.18311/jmmf/2023/34162

Print ISSN : 0022-2755

Journal of Mines, Metals and Fuels

Contents available at: www.informaticsjournals.com/index.php/jmmf

Kidney Disease Detection using Machine Learning

Dibakar Dasgupta*, Subhojit Mukherjee, Amit Chakraborty and Mou Majhi

Department of Computer Science & Engineering Swami Vivekananda University Kolkata, India.
E-mail: dibakar.dasgupta1@gmail.com / subhojit613@gmail.com / Amit13.ons@gmail.com / moumajhi21@gmail.com

Abstract

Different organs present in the human body which performed the relevant work. Kidney is one of the important body components that purifies the blood by removing toxic elements from the body. This is the reason, the kidney plays one of the important roles in the human body. Thus, the kidney needs to be safe in order to keep the body healthy. Different reasons are therefore which kidney is affected by a different disease. It can be seen that the reason behind kidney disease may differ for different persons. In this research, machine learning has been applied to the kidney disease dataset (collected from Kaggle) for the identification of kidney disease. Primarily, the analysis of the data has been conducted followed by the selection of essential attributes of the data so that the symptoms can be identified. In this scenario, the correlation mechanism has been employed to select essential attributes of the data. Using those essential attributes., the data has been finalized based upon which the detection of kidney disease has been commenced with the implication of classifiers of machine learning. To determine the best-performing model, the effectivenesses of the applied models have been done from where it has been identified that Gradient Boosting has performed the best with an accuracy rate of 99.62%.

Keywords: *Machine Learning, Kidney Disease Detection, Feature selection, Classification, Gradient Boosting*

1.0 Introduction

The kidney is one of the important organs in the human body. It helps filtrate the waste materials of the body and keeps the blood purified from any unnecessary waste elements. Thus, the kidney plays an important role in keeping human body hygiene[1]. So, any inconsistency in the operations of the kidney may cause severe effects on human health and the waste materials may not be drained from the body[2]. It raises the risk of toxicity in the human blood and thereby affects health[3]. There may not be any particular or specific reason for which all kidney diseases in humans will be caused. So, the disease is dependent upon the symptoms which need to be identified to detect the disease type correctly[3].

In this research, kidney disease will be detected by emphasizing the important symptoms of the patients[4]. So, from the data, the important symptoms will be found using the features selection process which facilities identifying the important symptoms of the disease in the kidney. After the selection of the symptoms and the details of the patients, the detection of kidney disease commenced based on those symptoms. This will be the concerning facts of the research part of the project[5]. In the next phase of the project commencement, the effectiveness of the applied classifiers will be done so that the best-performing model of machine learning can be identified that can detect the disease in the kidney with the highest effectiveness[6]. In this overall scenario, the effectiveness will be judged on the classification metrics such as the model accuracy, preciseness and recalls of the classifiers and finally the f1-scores.

**Author for correspondence*

2.0 Exiting Works

A. Existing Approaches for Disease Detection

The application of machine learning models has emerged in the healthcare and medical fields for the precise detection of diseases[7]. The researchers have employed RF, KNN, and SVM for the detection of diseases such as liver disease, cardiac disease and diabetes disease. To detect the diseases, they have primarily emphasized the determination of important symptoms and applied those to prepare the final datasets for disease detection. With the implication of those classifiers, they have acquired 90.16% accuracy for cardiac disease detection, 73% accuracy for liver disease detection and 93.23% accuracy for the detection of diabetes disease.

The authors of the research[8], diabetes is a chronic disorder prevalent in people of all age groups caused by poor insulin generation leading to increased blood sugar levels. Being neglected leads to a number of severe health complications like Diabetic Retinopathy (DR), an asymptomatic diseased condition leading to damaged retinal vessels. The advancement of DL in medical imaging has resulted in more accurate outcomes due to its prioritization of automated feature derivation. CNN is the most suitable DL approach utilized for the assessment of the medical image. A number of DL-based DR disease detection and classification methodologies can be assessed and appraised of gaining better insight.

The emergence of heart disease is among the top reasons behind the largest number of hospitalizations and demises all around the globe[9]. Technological advancements and computer engineering facilitate easy identification followed by fast and efficient treatment. The evolvement of ML in disease prediction encourages the implementation of seven such ML algorithms with the aim of improving the efficiency of the poor-performing algorithms utilizing Adaptive Boost and the voting association method. The Linear Discriminate Analysis is suitable with Mean Value and Mean Absolute errors of 0.847 and 0.185 respectively with an accuracy of 80% which is lower than logistic regression.

According to the authors of the research[10], the possible prevalence of heart disease in humans can be predicted using a majority voting associational approach facilitated by simple clinical tests. The study aims toward providing efficiency and confidence to experts' diagnoses due to the model training on practical data of healthy and diseased individuals. The patients can be classified as per the voting of the number of ML models for facilitating apt solutions, and finally present with an accuracy of 90% as per the hard-voting classification model.

B. Existing Approaches for Kidney Disease Detection

The authors of the research[11] have identified chronic kidney disease with the application of classifiers of machine learning. As reported by the researchers, the possibility of kidney disease lies in certain factors like symptoms and lifestyle issues. Those accelerate the person to be affected by chronic kidney disease. Additionally, demographic factors such as the place to live and gender etc. are also important parameters that influence the occurrence of chronic kidney disease. To detect chronic kidney disease, the researchers have applied SVM followed by the determination of important factors and features from the data and gained 93% accuracy in detecting chronic kidney disease.

The authors of the research[12] have detected kidney disease with the implication of machine learning classifiers. CKD is also known as Renal Disturbance and requires a number of tests for its diagnosis. Progresses in the fields of ML-oriented disease classification encourage researchers to examine health data with the primary focus of the article on automatic CKD detection utilizing clinical data and deploying a DL model. Moreover, it also aims toward achieving a strong and adjustable model for CKD identification with required clinical efficiency. Data consisting of 400 instances and 23 attributes are subjected to a Deep Learning model for classifying CKD diagnosis with an accuracy achievement of 99%.

The authors of the research[13], CKD is an alarming global health condition, and the situation gets worsened due to the high susceptibility of HIV-infected individuals getting the disease, but early diagnosis restricts the disease progression thereby limiting its severity. Implementing ML in this approach for disease identification and classification has turned quite common, depending on the glomerular filtration estimate, the CKD stages can also be assessed with the DNN model's performance surpassing the other methods in terms of classifying HIV patients with CKD with an accuracy of 99%.

According to the authors of the research[14], kidney disease can be efficiently detected via ML approach classification considering the processing of an individual's ECG signal. Recent studies indicate that people with CKD develop cardiac complications named Cardio-Renal Syndrome often leading to certain cardiac attacks and due to the prevailing interconnection, individuals suffering from the cardiovascular disorder can be tested for determining the presence of CKD or not. Early CKD diagnosis provides some time by lowering the progression using appropriate medical needs. For the study, digital ECG data can be acquired from free sources like PTD and Fantasia via the Physio net database. The validation steps present satisfactory outcomes as the model was able to segment the users from healthy or kidney patients.

The authors of the research[2] have predicted chronic kidney with the help of machine learning algorithms. Infections in the heart are an alarming health condition demanding early diagnosis for facilitating happy and healthy life. Kidney functionality deteriorates with the rising conditions of Hemoglobin, BP, diabetes, Obesity, etc. CKD advances in severity via the number of phases and a summarized study can be presented reflecting the bio-clinical phenotype of instances with CKD susceptibility and repeatedly predicts sufferers over a long span and facilitating CKD staging and presenting levered suggestions for prohibiting the advancement of CKD.

The authors of the paper[15] have researched kidney detection data using machine learning algorithms to determine the patient with a normal case of kidney disease. In this research, they have selected the dataset from UCI with 400 observations and applied classifiers such as Decision tree, support vector machine, XGBoost and K-neighbour. After the application of the model, they obtained 98% perfectness in detecting kidney disease using the Boost classifier.

3.0 Proposed Method

A. Proposed Approach of Research

The proposed method for the research is shown in Fig.1:

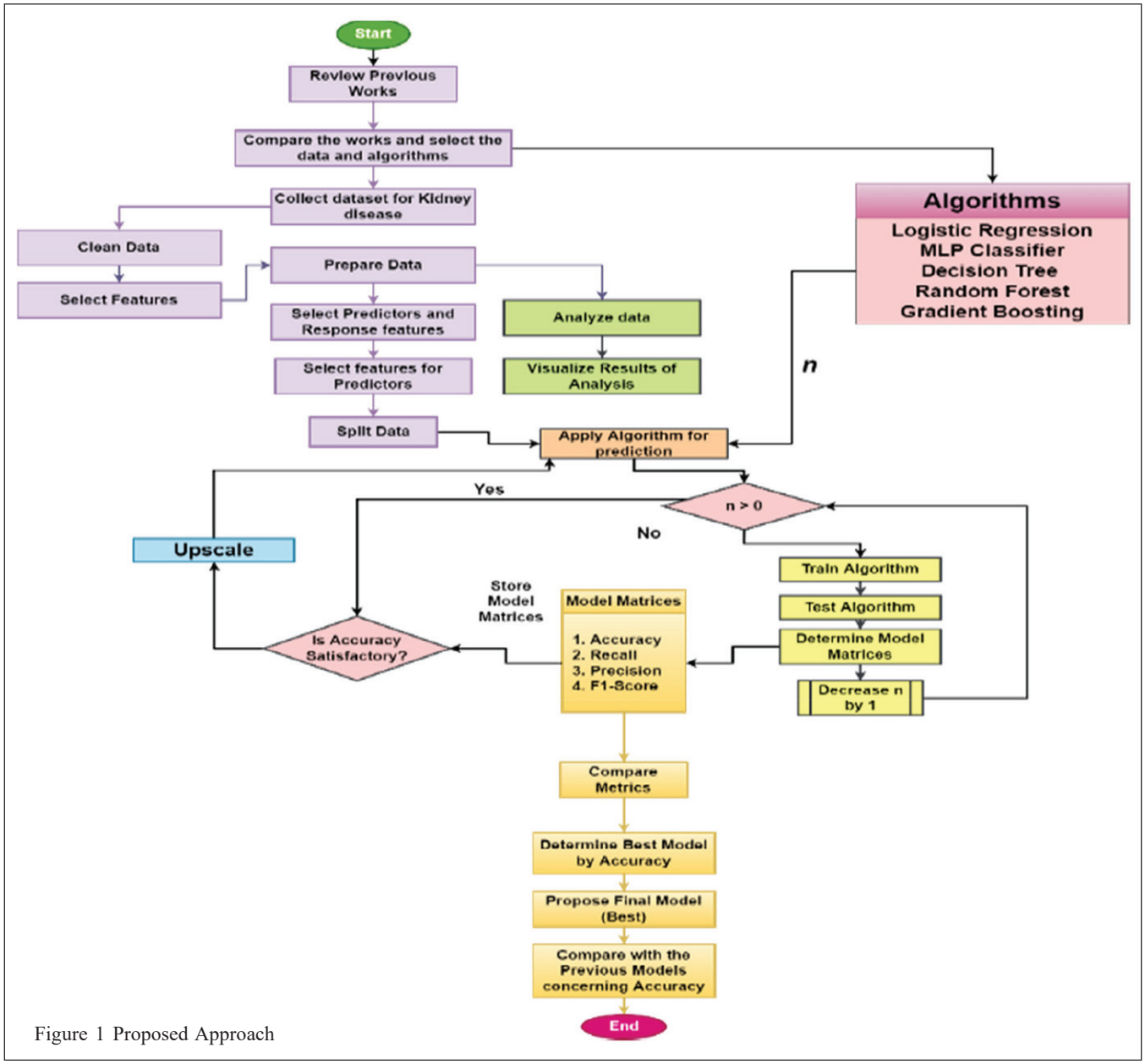


Figure 1 Proposed Approach

In this proposed approach, the first step that will be taken is the review of the existing works. This step has been done already in the previous section[2]. the main motivation behind the reviews is to gather the required knowledge for the detection of disease by focusing on the identification of kidney disease[4].

Now, to detect kidney disease, the data will be selected. In this context, the kidney disease data will be selected from a public source (Kaggle). After collecting the data, analytics has been done to identify the data statistics[16]. Next, the feature selection will be done to identify the important symptoms from the data with help of the pairwise correlation process[17]. Finally, the classifiers which have been chosen have been applied for the identification of disease in kidneys and the effectiveness has been queued. Those performances will be compared to determine the best model with higher accuracy[14].

B. Data Selection

The data for chronic kidney disease has been taken from a Kaggle source which comprises the historical patient records (400 patients) who have been tested for kidney disease[18]. Out of the 400 records, 289 patients have been recorded with health status and the rest 111 patients have been detected with chronic kidney disease. The feature namely "classification" will be used as the target or dependent feature which will be put on the identification of disease in the kidney with the implication of the predictor features. The details of the predictor features are shown in Table 1:

Table 1: Predictor Features

Feature	Type	Feature	Type	Feature	Type
age	Num	pcc	Cat	hemo	Num
bp	Num	ba	Cat	pcv	Num
sg	Num	bgr	Num	wc	Num
al	Num	bu	Num	rc	Num
su	Num	sc	Num	htn	Cat
rbc	Cat	sod	Num	dm	Cat
pc	Cat	pot	Num	cad	Cat
appet	Cat	pe	Cat	ane	Cat

**Num: Numerical, Cat: Categorical

C. Data Analysis

The analyses have been done on Haemoglobin and Glucose. It has been observed that the level of haemoglobin is lower for CKD patients but the level of glucose is higher as shown in Fig.2:

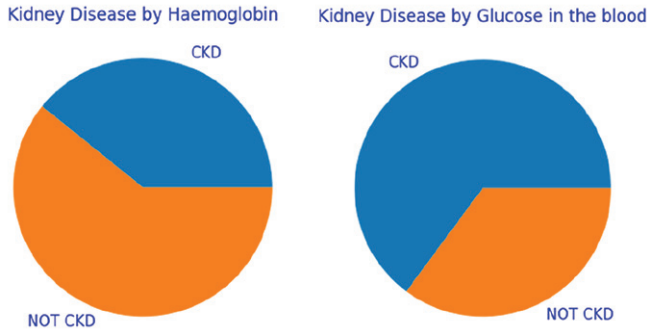


Figure 2: Analysis by Haemoglobin and Glucose

The analyses below are showing that the level of Serum Creatinine and Blood urea are higher for CKD patients (Fig.3).

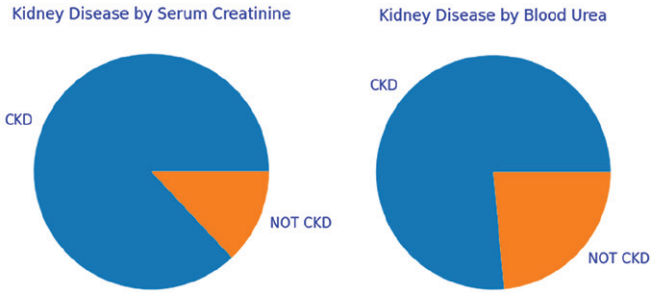


Figure 3: Analysis of Creatinine and Blood Urea

D. Selection of Feature

To select the essential features from the data, Pairwise correlation has been employed[19]. With the application of correlation, the heatmap has been generated which is shown in Fig.4:

In the heatmap, the entire correlation result has been shown. By applying the correlation filter (-0.42>feature_correaltion>-0.4), the final features have been selected and shown in Fig.5:

E. Data Split and Preparation

The classifiers of machine learning require sufficient data to be trained so that the classifiers can detect unseen data (test data)[15]. However, the collected data has a lower number of instances (400 rows) which may not train the model properly. This is the reason, the data has been upscaled by multiplying the present row number 10 times and the final data has been set. After that, the data splitting has been done to create the training and testing set[13]. Here, the split ratio has been taken by 0.8. After splitting the data, the support for the data label "CKD" becomes 237 and the support for the data label "notckd" becomes 563.

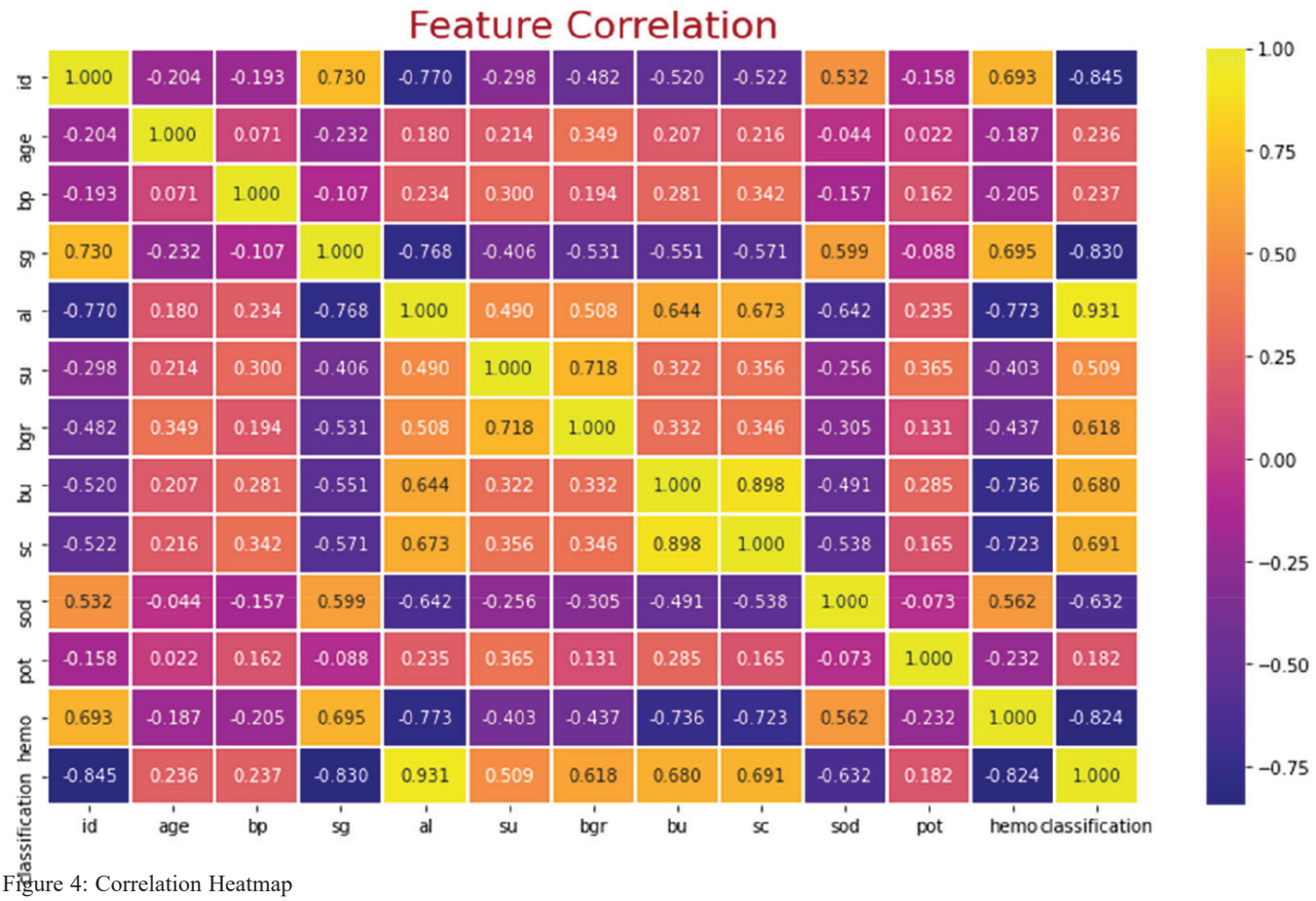


Figure 4: Correlation Heatmap

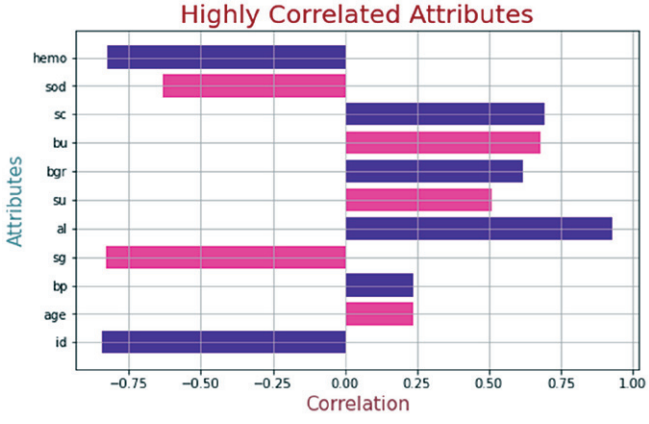


Figure 5: Selected Features

F. Model Tuning and Preparation

To detect kidney disease, the following algorithms in machine learning have been selected:

- Logistic Regression
- MLP Classifier
- Decision Tree
- Random Forest
- Gradient Boosting

Each of the classifiers has been tuned and prepared and applied to the data for the detection of kidney disease. The overall process of tuning and applying the classifiers is as Fig.6:

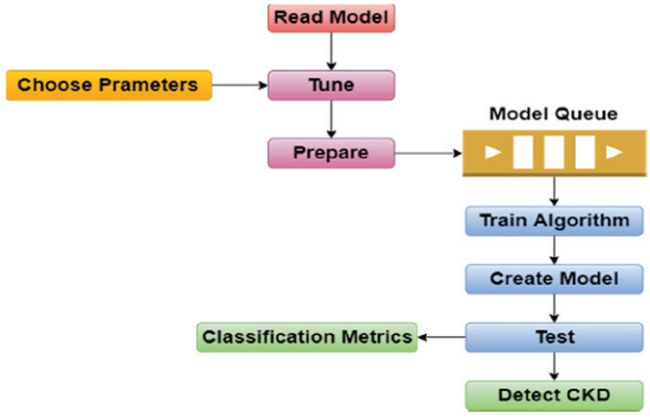


Figure 6: Algorithmic Process

After applying each of the algorithms through the model queue, the performance has been evaluated through the following classification metrics:

- Accuracy: This is used to measure the effectiveness of the model by emphasising the correct detection using the following equation:

$$Acc = \frac{Correct\ Detection}{Total\ Test\ Observations} \quad \ldots (1)$$

- Precision: This is used to measure the preciseness of detecting by classes using the following equation:

$$\frac{Correct\ Detection\ by\ Class}{Total\ Test\ Observations\ in\ that\ Class} \quad \ldots (2)$$

- Recall: This is used to measure the sensitivity of classification using the following equation:

$$Recall = \frac{Correct\ Detection\ by\ Class}{Total\ Test\ Observations\ Predicted\ by\ class} \quad \ldots (3)$$

- 'F1-Score: This is the combined measure of precision and recall and can be computed using the following equation:

$$F1 = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \quad (4)$$

4.0 Result of Kidney Disease Detection

The metrics for classification have been achieved for the algorithms which have been used for the detection of kidney disease. Those will be presented in this section.

A. Result of Logistic Regression

The result of logistic regression concerning the classification (detection) of kidney disease has been presented in Table 2:

Table 2: Result of Logistic Regression

Logistic Regression		Predicted	
		CKD	NOTCKD
True	CKD	184	53
	NOTCKD	0	563

This model has detected 184 CKD instances out of 237 actual instances and 563 NOTCKD instances out of 563 actual instances correctly with an overall 93.38% accuracy, 94% precision, 93% recall and 93% f1-score.

B. Result of MLP Classifier

The result of the MLP classifier has been presented in Table 3:

Table 3: Result of MLP

MLP Classifier		Predicted	
		CKD	NOTCKD
True	CKD	199	38
	NOTCKD	0	563

This model has detected 234 CKD instances out of 237 actual instances and 563 NOTCKD instances out of 563 actual instances correctly with an overall 95.25% accuracy, 95% precision, 95% recall and 95% f1-score.

C. Result of Decision Tree

The result of decision tree concerning the classification (detection) of kidney disease has been presented in Table 4:

Table 4: Result of Decision Tree

Decision Tree		Predicted	
		CKD	NOTCKD
True	CKD	233	4
	NOTCKD	0	563

This model has detected 233 CKD instances out of 237 actual instances and 563 NOTCKD instances out of 563 actual instances correctly with an overall 99.6% accuracy, 99% precision, 99% recall and 99% f1-score.

D. Result of Random Forest

The result of the Random Forest classifier has been presented in Table 5:

Table 5: Result of Random Forest

Random Forest		Predicted	
		CKD	NOTCKD
True	CKD	200	37
	NOTCKD	0	557

This model has detected 200 CKD instances out of 237 actual instances and NOTCKD instances out of 563 actual instances correctly with an overall 94.62% accuracy, 95% precision, 95% recall and 95% f1-score.

E. Result of Gradient Boosting

The result of the Gradient Boosting classifier has been presented in Table 6:

Table 6: Result of Gradient Boosting

Gradient Boosting		Predicted	
		CKD	NOTCKD
True	CKD	234	3
	NOTCKD	0	563

This model has detected 234 CKD instances out of 237 actual instances and 563 NOTCKD instances out of 563 actual instances correctly with an overall 99.62% accuracy, 99% precision, 99% recall and 99% f1-score.

F. Effectiveness Comparison

The classification metrics have been used to determine the performance of the employed classifiers of machine learning. The comparison of the effectivenesses has been presented in Fig.7:

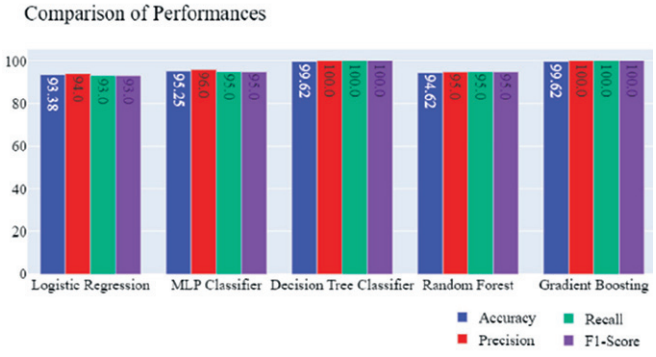


Figure 7: Comparison of Performance

From the overall comparison, it can be observed that the Gradient Boosting classifier has detected kidney disease with the highest accuracy by 99.62% with misclassification of 3. So, Gradient Boosting will be taken as the most efficient model to detect kidney disease.

5.0 Future Scopes

The models of machine learning have been applied to detect kidney disease in this research. According to the result, the accuracy has been achieved satisfactorily. However, some future aspects are there that can be implied in the further progress of research. Some of those are the application of deep learning models or hybrid machine learning models are important.

6.0 Conclusion

The identification of kidney disease has been commenced by employing classifiers of machine learning which have been implied on the selected dataset (feature selected). To conduct the research and experiment, the data has been collected from Kggale and the analysis of the features has been done. The necessary features which can also be denoted by the essential symptoms have been examined and detected using the application of the feature selection process. Hence, the correlation mechanism has been applied to select essential symptoms. With the selection of essential attributes or symptoms, the final data has been prepared. Now, this data has been used for the identiifcation of disease in the kidney with the employment of classifiers of machine learning. By comparing the effectivenesses of the applied models, it has been observed that 99.62% accuracy has been attained by Gradient Boosting which is the highest among all the applied models. So, finally, Gradient Boosting is the proposed model to detect kidney disease with the highest effectiveness.

7.0 Dedication

The research and the paper is dedicated to the loving memory of Late Prof. Dr. Samir Dasgupta (Emeritus Professor, University of Kalyani)

8.0 References

1. S. Bookheimer, M. Strojwas, M. Cohen, A. Saunders, M. Pericak-Vance and J. Mazziotta, (2000): "Patterns of brain activation in people at risk of Alzheimer's disease," N Engl J Med, vol. 6, pp. 343-450.
2. C. Suresh, B. Chakra Pani and C. Swatisri, (2020): "A Neural Network based Model for Predicting Chronic Kidney Diseases," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 157-162.
3. M. Zarour, M. T. J. Ansari, M. Alenezi, A. K. Sarkar, M. Faizan, A. Agrawal, R. Kumar and R. A. Khan, (2020): "Evaluating the Impact of Blockchain Models for Secure and Trustworthy Electronic Healthcare Records," IEEE Access, pp. 157959 - 157973.
4. M. Wickramasinghe, D. Perera and K. Kahandawaarachchi, (2017): "Dietary prediction for patients with Chronic Kidney Disease (CKD) by considering blood potassium level using machine learning algorithms," IEEE Life Sciences Conference (LSC), pp. 300-303.
5. A. Burlacu, A. Iftene, D. Jugrin, I. Popa, P. Lupu and C. A. C. A. Vlad, (2020): "Using artificial intelligence resources in dialysis and kidney transplant patients: a literature review.," BioMed research international, pp. 1-12.
6. M. D. Basar and A. Akan, (2017): "Detection of chronic

kidney disease by using ensemble classifiers," 2017 10th International Conference on Electrical and Electronics Engineering (ELECO), pp.544-547.
7. S. Gupta, U. Namdev and V. Gupta, "Data-driven Preprocessing Techniques for Early Diagnosis of Diabetes, Heart and Liver Diseases,," 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1-8, 2021.
8. S. Valarmathi and R. Vijayabhanu, (2021): "A Survey on Diabetic Retinopathy Disease Detection and Classification using Deep Learning Techniques," 2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII), pp. 1-4.
9. A. Kumari and A. Kumar Mehta, (2021): "A Novel Approach for Prediction of Heart Disease using Machine Learning Algorithms," 2021 Asian Conference on Innovation in Technology (ASIANCON), pp. 1-5.
10. R. Atallah and a. Al-Mousa, (2019): "Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method,," 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), pp. 1-6.
11. Amirgaliyev, S. Shamiluulu and A. Serek, (2018): "Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods," 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), pp. 1-4.
12. F. Arafat, T. Khan and A. D. Bapon, (2021): "A Deep Learning Approach to Predict Chronic Kidney Disease in Human," 2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 1010-1015.
13. A. Y. Darveshwala, D. K. Singh and Y. Farooqi, (2021): "Chronic Kidney Disease Stage Identification in HIV Infected Patients using Machine Learning," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1509-1514.
14. T. M. Rahman, S. Siddiqua and S. E. Rabby, (2019): "Early Detection of Kidney Disease Using ECG Signals Through Machine Learning Based Modelling," 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp. 319-323.
15. S. Patel, R. Patel and N. Ganatra, (2022): "An Experimental Study and Performance Analysis of Supervised Machine Learning Algorithms for Prognosis of Chronic Kidney Disease,," 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), pp.1-6.
16. Z. Xiong, H. Zhang and Y. Chen, (2019): "Deep Ensemble Learning Network for Kidney Lesion Detection," 2019 Chinese Automation Congress (CAC), pp. 3841-3846.
17. S. R. Tanuku, A. K. Addike and S. R. Somaraju, (2022): "Liver Disease Prediction Using Ensemble Technique," 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 1522-1525.
18. M. Iqbal, (2017): "Chronic KIdney Disease dataset,". [Online]. Available: https://www.kaggle.com/mansoordaku/ckdisease.
19. Z. Seidighi, H. Ebrahinpour-Komleh and S. J. Mousavirad, (2015): "Featue selection effects on kidney desease analysis," 2015 International Congress on Technology, Communication and Knowledge (ICTCK), pp. 455-459.
20. H. Alaskar, S. Alharkan and W. Alharkan, (2017): "Detection of kidney disease using various intelligent classifiers," 2017 3rd International Conference on Science in Information Technology (ICSITech), pp. 681-684.
21. N. Abdulhadi and A. Al-Mousa, (2021): "Diabetes Detection Using Machine Learning Classification Methods," 2021 International Conference on Information Technology (ICIT), pp. 350-354.
22. N. Basha, A. Kumar P.S. and G. Krishna C., (2019): "Early Detection of Heart Syndrome Using Machine Learning Technique," 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT),, pp. 387-391.
23. S. Kaur, J. Singla and L. Nkenyereye, (2020): "Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives," *IEEE Access*, vol. 8, pp. 228049-228069.
24. R. Katarya and P. Srinivas, (2020): "Predicting Heart Disease at Early Stages using Machine Learning: A Survey,," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 302-305.
25. G. Nandhini and J. Aravinth, (2021): "Chronic kidney disease prediction using machine learning techniques," 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), pp. 227-232.