Chronic Renal Disease Prediction using Clinical Data and Different Machine Learning Techniques

Md. Mohsin Sarker Raihan[1], Eshtiak Ahmed[2], Asif Karim[3], Sami Azam*[4],
M. Raihan[5], Laboni Akter[6] and Md. Mehedi Hassan[7]
Department of Biomedical Engineering, Khulna University of Engineering & Technology, Khulna-9203, Bangladesh[1,6]
Faculty of Infomration Technology and Communication Sciences, Tampere University, Tampere, Finland[2]
College of Engineering, IT and Environment, Charles Darwin University, Darwin, NT, Australia[3,4]
Department of Computer Science and Engineering, North Western University, Khulna, Bangladesh[5,7]
Emails: msr.raihan@gmail.com[1] (ORCID: 0000-0002-0401-312X)[1] eshtiak.ahmed@tuni.fi[2], asif.karim@cdu.edu.au[3],
sami.azam@cdu.edu.au[4], raihanbme@gmail.com[5] (ORCID: 0000-0003-1072-3555)[5],
laboni.kuet.bme@gmail.com[6] (ORCID: 0000-0002-0063-3188)[6],
mehedihassan@ieee.org[7] (ORCID: 0000-0002-9890-0968)[7]

Abstract*—Chronic Renal Disease (CRD) or Chronic Kidney Disease (CKD) is defined as the continuous loss of kidney function. It's a long-term condition in which the kidney or renal doesn't work properly, gets damaged and can't filter blood on a regular basis. Diabetes, high blood pressure, swollen feet, ankles or hands and other disorders can cause chronic renal disease. By gradual progression and lack of treatment, it can lead to kidney failure. A prior prognosis of CKD can nourish the quality of life to a higher range in such circumstances and can enhance the attribute of life to a larger province. Now a days, bioscience is playing a significant role in the aspect of diagnosing and detecting numerous health conditions. Machine Learning (ML) as well as Data Mining (DM) methods are playing the leading role in the realm of biosciences. Our objective is to predict and diagnose (CKD) with some machine learning algorithms. In this study, an attempt to diagnose chronic renal disease has been taken with four ML algorithms named XGBoost, Adaboost, Logistic Regression (LR) as well as Random Forest (RF). By using decision tree-based classifiers and analyzing the dataset with comparing their performance, we attempted to diagnose CKD in this study. The results of the model in this study showed prosperous indications of a better prognosis for the diagnosis of kidney diseases. Considering and contemplating the performance analysis, it is accomplished that Random Forest ensemble learning algorithm provides better classification performance than other classification methods.

***Index Terms*—Chronic Kidney Disease, CKD, XGBoost, Adaboost, Logis-tic Regression, Random Forest.**

I. Introduction

The term CKD means Chronic Kidney Disease which can be defined as the inability of kidneys to perform their spontaneous blood functions. The word chronic in this context recounts the gradual deterioration of the kidney cells. This is the disease which can lead to a massive kidney failure; where kidneys lack blood straining. This results in dreadful enhancement of existing potassium as well as calcium salts in the body. Subsistence of lofty equality of these salts cause multiple types of illnesses. The elementary task of kidneys is to filter surplus water as well as wastes from blood. For balancing the existing salts and minerals in our body, this filtering process is very important. A high coagulation of calcium results in different types of cystic ovaries and bone diseases in women. CKD can also cause various types of problems like sudden illness or allergy to specific medicines. As a result, it can also cause increase in blood pressure, leading to a heart disease. In many cases, CKD can lead to kidney transplants or permanent dialysis. A history of kidney disease in the family can also result in a high possibility of CKD. In literature, it is shown that one out of three persons are diagnosed with diabetes which contain CKD [1]. At lower stage, a high remainder of urine and wastes aggregate in the body. The process is held when the glomerular filtration rate (GFR) arrives at a subordinate phase. An onward regression occurs in GFR while initiating a period of irrecoverable sclerosis of surviving nephrons. This initiation occurs at the time of deprivation of nephrons. Degrading changes in kidney tissue can cause chronic inflammation in a short period of time which causes many complications. It may lead to loss of weight, innutrition, muscle problems, dropsy, fatigue, immunity problem, heart diseases, heart failure, increased cardiovascular mortality, several types of skin diseases, cardiovascular diseases, pericarditis and appetite problems [2]. The inauguration of CKD dialysis generally dominates some morbidities like sitophobia, weariness, cognitional damages, depressions, erythema, sleep inconvenience, pulmonary dropsy, pericarditis, neurological dis-tempers, and metabolic anomaly [3]. More than a total of two million people get kidney transplantation or dialysis treatment for staying healthy and alive each year. Due to the extended financial burden; in 112 lower-income countries, more than one million people die from kidney failure annually [4]. However, it is possible to medicate CKD in the prior stage. To predict as well as prognose the causation of CKD in a prior stage, prediction algorithms in machine learning (ML) can be employed.

Our objective is to ascertain the accuracy of four well known algorithms called XGBoost, Adaboost, Logistic Regression (LR) and Random Forest (RF). In this study, a dataset containing 455 instances and 25 attributes or features including outcome has been explored. The goal is to utilize the different statistical methods to prognose the risk of kidney disease.

Preprint
IEEE International Conference on Informatics and Software Engineering, 2021, Turkey

Additionally, 10-fold cross validation applied for overcome over fitting under fitting issue and gained promising accuracy. With this approach in the future we will be able to predict CKD disease in advance.

The manuscript structure us arranged as follows: related works have been discussed in sections 2; in section 3, the methodology has been expounded. In section 4, the experiment has been analyzed and discusses. Lastly, section 5 concludes the study.

II. Background

In [5], a system was proposed to use an associated method of three prominent selection techniques to narrate the foremost significant properties related with the subsistence of CKD. The three combined methods were Tree classifier, Recursion property Elimination founded on LR and Univariate Statistics (US). For the final optimization, the XGBoost model was tuned which satisfied three feature selection techniques. In that research, the optimized CKD XGBoost model was used to predict CKD on some test sets and 24 CKD patients were correctly diagnosed. Among those 24 patients, 11 patients were properly determined as CKD-free patients. An XGBoost feature was used for contributing and picking package to supervise the significance of this model. The model had several advanced properties with an accuracy of 97.58% and the proposed model was able to reduce the time and it was quite cost effective. In [6], ML technique was used to narrate the affinity between kidney disease including eleven chronic diseases. There was an illustration of 4384 examples in that study. For diagnosing CKD, several types of ML techniques were used. The AdaBoost model returned the most desired outcome and it showed accuracies of 98.87% and 88.66%. In [7], an execution of three models was appraised. The performance of three diagrams were executed ported on K-NN, RF and NN to predict CKD. For K-NN model as well as Neural Network, the missing data was alternated with the values from IBK algorithm. For better performance the Random Forest model was used. The Random Forest model gave a better performance than two other models. The F1-score was 99.35% and Root Mean Square Error (RMSE) was 1.84%. This model was designed to execute an efficient and easier performance to diagnose CKD. In [8], for CKD prediction, logistic regression analysis and for the apparent outcome of data; the Linear regression analysis was resolved. The Regression coefficient of simple LR analysis of CKD was found 1.118 and for multiple LR analysis of CKD, the regression coefficient was found 0.974. To access the relationship between exposure Linear and logistic regression, analyses were performed in that study.

III. Methodology

Our objective is to ascertain the accuracy of four well known algorithms called XGBoost, Adaboost, Logistic Regression (LR) and Random Forest (RF). In this study, a dataset containing 455 instances and 25 attributes or features including outcome has been explored. The goal is to utilize the different statistical methods to prognose the risk of kidney disease. Additionally, 10-fold cross validation applied for overcome over fitting under fitting issue and gained promising accuracy. Furthermore, we tested our train model using real datasets obtained from hospital, and the model's performance was satisfactory. With this approach in the future we will be able to predict CKD disease in advance. In this section, instances and dataset features, dataset preprocessing i.e., missing data handling, import dataset, feature scaling, dataset splitting, and application of the machine learning algorithms are explained. The Fig. 1 shows the working flow of this study.

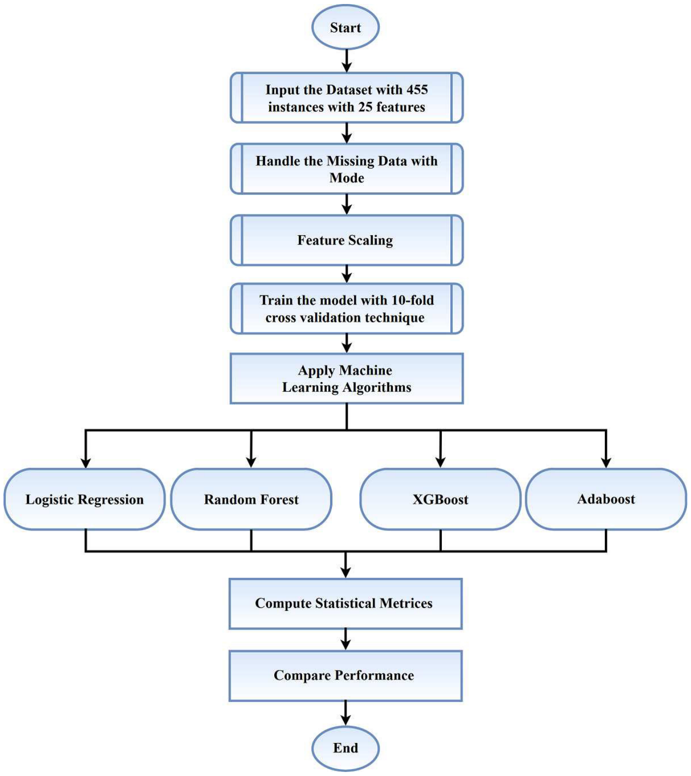
Start
Input the Dataset with 455 instances with 25 features
Handle the Missing Data with Mode
Feature Scaling
Train the model with 10-fold cross validation technique
Apply Machine Learning Algorithms
Logistic Regression
Random Forest
XGBoost
Adaboost
Compute Statistical Metrices
Compare Performance
End

Fig. 1. Work-flow of the study

A. *Instances and Dataset*

In this work, a CKD dataset was procured from the UCI Machine Learning Repository. The dataset contains 455 instances and 25 attributes or features including outcome. The dataset contained two different classes and the percentages of those two classes were shown in Fig. 2. The features were age, blood pressure (bp), specific gravity (sg), albumin (al), sugar (su), red blood cell (rbc), pus cell (pc), bacteria (ba), pus cell clumps (pcc), blood urea (bu), blood glucose random (bgr), serum creatinine (sc), potassium (pot), sodium (sod), hemoglobin (hemo), white blood cell count (wbcc), packed cell volume (pcv), red blood cell count (rbcc), diabetes mellitus (dm), hypertension (htn), coronary artery dis-ease (cad), pedal edema (pe), appetite (appet), anemia (ane) and outcome or class. Out of these 25 features, only 11 features including outcomes were nominal and rest of the features were numerical. The nominal features were pus cell, red blood cell, bacteria, pus cell clumps, diabetes mellitus, hypertension, anemia, coronary artery disease, pedal edema, and appetite.

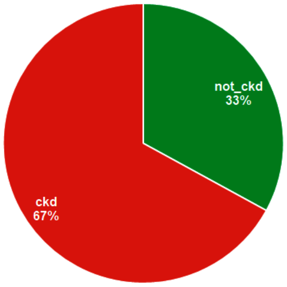

Fig. 2. Pie chart of CKD and Non-CKD Patients

B. Missing Data Handling

The machine learning algorithms cannot deal with missing values; thus, they must be replaced using different statistical techniques. In this dataset, we used different methods depending on attributes properties. Some missing data were replaced by using **MODE**. **MULT** method and some missing data was replaced by MEAN method.

C. Feature Scaling

Feature scaling in ML is one of the most basic tasks during the pre-processing of information before making an AI model. Scaling can have any kind of effect between a feeble Machine Learning model and a superior one [9]. The most widely recognized procedures of feature scaling are Normalization and Standardization. The popular python library **StandardScaler** was used to feature scaling in this study.

D. Dataset Training

In this research study, a 10-Fold Cross-Validation method has been utilized. It's a re-testing method for assessing prescient models by dividing the first example into a preparation set for preparing the model, as well as a test set for assessing it [10]. The system has a single boundary which is considered as K that implies the calculation number and a provided data test is to be part into. It rearranges the dataset randomly by parting it into 10 gatherings as well as at last abbreviates the ability of the model by using the appraisal scores [11].

E. Application of the Machine Learning Algorithms

Logistic Regression, Random Forest, XGBoost, and Adaboost algorithms have been used in this study which described in the following section.

1) Logistic Regression (LR): LR is a notable framework for gathering data into two random and exhaustive orders, for example, purchaser, non-purchaser and responder, non-responder [12]. It forecasts by propagating and establishing the logit work and by fitting information of an occasion [12]. For Logistic Regression theses parameter have been used: penalty = l2, random state = 0, solver = liblinear.

2) Random Forest (RF): Random Forest generally is an inquired learning calculation exploited for both arrangements. It's fundamentally applied for classification issues in any cases. Basically, a forest is the aggregation of trees, and more trees signifies more plants. Apart from this, random forest calculation makes decision trees on dataset tests. It also chooses the best arrangement methods by casting a ballot. It's an outfit technique which is better than a secluded decision tree as it decreases the overfitting by affecting the outcome [10]. For Random Forest, these parameters have been used: n estimators = 300, criterion = entropy, random state = 0.

3) XGBoost: XGBoost is also known as Extreme Gradient Boosting algorithm. XGBoost is a decision tree-based outfit Machine Learning calculation that utilizes a gradient boosting structure. XGBoost is an optimized dispersed inclination boosting library contemplated to be exceptionally proficient, congenial and multi purposed. It actualizes Machine Learning calculations within the Gradient notifying system. XGBoost provides an equal tree boosting that tackles numerous data science issues in a fast and efficient way [13]. For XGBoost, the parameters are objective = binary: logistic, n estimators = 100, random state = 0, importance type = gain.

4) AdaBoost: AdaBoost is the short form of Adaptive Boosting. Fundamentally, Ada Boosting is a truly effective first boosting prediction created for twofold order. Likewise, it is the superior starting stage for intellect boosting. Also, present day boosting techniques has prolonged on AdaBoost, mostly on quiet stochastic gradient machines. Basically, AdaBoost is utilized with short decision trees. Further, the primary tree is made, the exhibition of the tree on each preparation case is utilized. Likewise, we have used it to weight consideration the following tree. In this way, it is made to focus on each preparation occasion. Consequently, preparing information that is difficult to foresee is given more weight. However, easy to classify the data are given less weight [14]-[22]. For AdaBoost, these parameters have been used: n estimators = 400, learning rate = 1.

F. Simulation Environment and Libraries

- Python 3.6.5
- Anaconda Package Manager
- scikit-lean 0.19.1
- numpy 1.14.3
- matplotlib 2.2.2
- pandas 0.23.0
- seaborn 0.8.1
- xgboost 1.1.1

IV. Experimented Analysis and Discussions

Fig. 3 shows the correlation matrix of each feature. The range of the correlation matrix value between -1 to +1. The value of +1 means strong correlation where the -1 represents poor correlation of the features. From the correlation matrix, the packed call volume has the highest correlation (0.69) with the class. Specific Gravity, The Hemoglobin, Red Blood Cell

Count have also a strong correlation with the class where the value was 0.65, 0.68 and 0.65 separately.

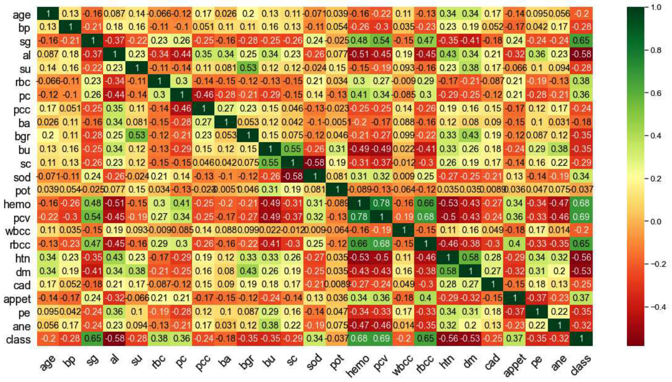

Fig. 3. Correlation Matrix of the Features

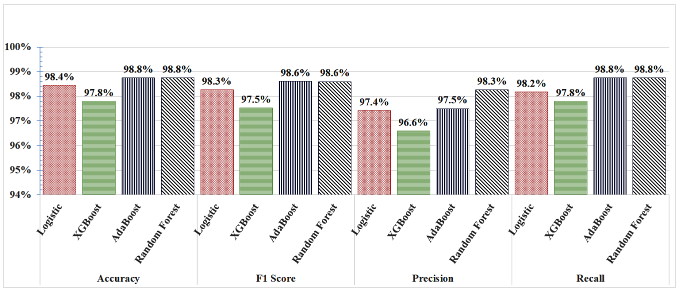

Fig. 4. Bar graph of different statistical parameters of algorithms

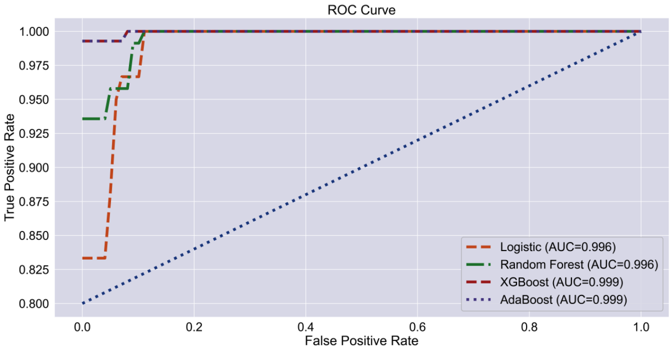

Fig. 5. Receiver Operating Characteristic (ROC) Curve

The applied machine learning algorithms results are shown in Fig. 3. The accuracy, recall, precision and F1 score were found 98.4%, 98.2%, 97.4% and 98.3% respectively for Logistic Regression algorithm. Random Forest algorithm has provided the 98.8% accuracy, 98.2% recall, 98.3% precision and 98.6% f1 score. The statistical score of XGBoost algorithm were 97.8% accuracy, 97.2% recall, 96.6% precision and 97.5% f1 score. The accuracy, recall, precision and F1 score of AdaBoost algorithm were 98.8%, 99.1%, 97.5% and 98.6% respectively. Fig. 4 shows the AUC score from the ROC curve. XGBoost and AdaBoost algorithms have obtained maximum AUC score which was 99.9%. The similar AUC score was found in both Logistic and Random Forest algorithms where the value was 99.6% in both of cases.

From all the statistical results, we have found that the highest accuracy was 98.8% of Random Forest algorithm and the highest precision (98.3%) was also found using Random Forest algorithm. Similarly, the maximum F1 score (98.6%) was found Random Forest algorithm where Logistic algorithm presented the least performance score. Though Adaboost also performed with same accuracy of 98.8% and the highest AUC score was found in XGBoost and AdaBoost algorithms, both algorithms failed to reach the maximum accuracy, recall, precision and F1 scores and according to precision score. Furthermore, according to the ROC curve result, Random Forest gave the best performance admittedly. The main objectives of this study were early-stage prediction of CKD using different machine learning algorithms. Finally, this study confirmed that the Random Forest is the suitable algorithm to predict the early stage of CKD compared to other algorithms.

TABLE I
COMPARISONS WITH OTHER EXISTING SYSTEMS

Reference No.	Features	Algorithm	Accuracy (%)
[5]	25	XGboost	97.58
[6]	11	Adaboost	88.66
[15]	-	Random Forest	88.7
[16]	-	XGBoost	83
		Logistic Regression	82
Our Proposed System	25	Logistic	98.4
		Random Forest	98.8
		XGBoost	97.8
		Adaboost	98.8

The Table I shows some previous study results compared to our methods. From all of the statistical results and comparisons with other existing system it is found that among several used algorithms, Random Forest showed the best accuracy with 99.8%. Whereas XGBoost, Adaboost, Logistic regression analysis and Linear regression analyses also showed satisfactory range of accuracy like 88%-98% range. Our four proposed algorithms were Logistic regression, XG-Boost, Random Forest and Adaboost. Among those algorithms, Random Forest showed the paramount performance with high accuracy of 98.8%. Logistic, XG-Boost and Adaboost also showed quite favorable performance with the accuracy. Snegha et.al applied two data mining algorithms for prognosing the chronic kidney disease prediction (CKD). The output of the analysis of the Random Forest Algorithm obtains certainty of 88.7% percent and an area under the curve for receiver working characteristics 99.2% of the time [15]. Arulanthu et.al launched a system for the prediction of chronic kidney disease (CKD), an online patient decision support system (OMDSS). For the prediction of CKD, the presented model included several steps, including data collection, preprocessing, and classification of medical data. For classification, the logistic regression (LR) model was used to divide the data into CKD and non-CKD cases. Adaptive Moment Estimation (Adam) and an adaptive learning rate optimization algorithm were both used to tune the parameters of LR. XGBoost and Logistic Regression showed an accuracy of 83% and 82% [16].

V. Conclusion

This research article exhibited several prediction algorithms to predict and diagnose chronic kidney disease or chronic renal disease at a prior stage. The dataset showed various collected input parameters and the models were created and trained for the provided input parameters. By appraising the algorithms with an attribute set; the prediction of CKD or CRD was evaluated with the higher accuracy. To conclude, outcomes of this study define novel aspects to be used by classifiers for detecting CKD more precisely at an early stage. XGBoost, Adaboost, Logistic Regression (LR) and Random Forest (RF) were constructed and used for classification to predict and diagnose CKD. Though all the classifiers performed satisfactory with similar accuracy, considering the precision score and in accordance with the ROC curve; the Random Forest (RF) model constructed for CRD performed beyond expectation. Among all used algorithms, Random Forest showed better and promising performance in diagnosing CKD or CRD at an early stage.

References

[1] S. Revathy, B. Bharathi, P. Jeyanthi and M. Ramesh "Chronic Kidney Disease Prediction using Machine Learning Models", *International Journal of Engineering and Advanced Technology Regular Issue*, vol. 9, no. 1, pp. 6364-6367, 2019.

[2] M. D. Basar and A. Akan, "Chronic Kidney Disease Prediction with Reduced Individual Classifiers", *Electrica*, vol. 18, no. 2, pp. 249–255, 2018.

[3] V. J. Cabrera, J. Hansson, A. S. Kliger, and F. O. Finkelstein, "Symptom Management of the Patient with CKD: The Role of Dialysis", *Clinical Journal of the American Society of Nephrology*, vol. 12, no. 4, pp. 687-693, 2017.

[4] E. H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms", *Informatics in Medicine Unlocked*, vol. 15, p. 1-7, 2019.

[5] A. Ogunleye and Q.-G. Wang, "Enhanced XGBoost-Based Automatic Diagnosis System for Chronic Kidney Disease", *2018 IEEE 14th International Conference on Control and Automation (ICCA)*, 2018.

[6] D. Gupta, S. Khare, and A. Aggarwal, "A method to predict diagnostic codes for chronic diseases using machine learning techniques", *2016 International Conference on Computing, Communication and Automation (ICCCA)*, 2016.

[7] A. Salekin and J. Stankovic, "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes", *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, 2016.

[8] G. Tripepi, K. Jager, F. Dekker, and C. Zoccali, "Linear and logistic regression analysis", *Kidney International*, vol. 73, no. 7, pp. 806–810, 2008.

[9] "6.3. Preprocessing data — scikit-learn 0.23.2 documentation", *Scikit-learn.org*, 2020. [Online]. Available: https://scikit-learn.org/stable/modules/preprocessing.html. [Accessed: 06- Apr-2020].

[10] M. M. Hassan, M. A. Mamun Billah, M. M. Rahman, S. Zaman, M. M. Hasan Shakil and J. H. Angonn, "Early Predictive Analytics in Healthcare for Diabetes Prediction Using Machine Learning Approach ", *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2021, pp. 01-05, doi: 10.1109/ICCCNT51525.2021.9579799.

[11] M. T. Islam, M. Raihan, F. Farzana, M. G. M. Raju and M. B. Hossain, "An Empirical Study on Diabetes Mellitus Prediction for Typical and Non-Typical Cases using Machine Learning Approaches", *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kanpur, India, 2019, pp. 1-7, doi: 10.1109/ICCCNT45670.2019.8944528.

[12] B. Ratner, *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data*, 2nd ed. CRC Press, 2011, pp.97-98

[13] A. Kadiyala and A. Kumar, "Applications of python to evaluate the performance of decision tree-based boosting algorithms", *Environmental Progress & Sustainable Energy*, vol. 37, no. 2, pp. 618-623, 2018. Available: 10.1002/ep.12888.

[14] T. Hastie, S. Rosset, J. Zhu and H. Zou, "Multi-class AdaBoost", *Statistics and Its Interface*, vol. 2, no. 3, pp. 349-360, 2009. Available: 10.4310/sii.2009.v2.n3.a8.

[15] J. Snegha, V. Tharani, S. D. Preetha, R. Charanya, and S. Bhavani, "Chronic Kidney Disease Prediction Using Data Mining", *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 2020.

[16] P. Arulanthu and E. Perumal, "An Intelligent IoT With Cloud Centric Medical Decision Support System For Chronic Kidney Disease Prediction", *International Journal of Imaging Systems and Technology*, vol. 30, no. 3, pp. 815–827, 2020.

[17] M. Ghosh, M. M. S. Raihan, M. Raihan, L. Akter, A. K. Bairagi, S. S. Alshamrani, M. Masud, "A Comparative Analysis of Machine Learning Algorithms to Predict Liver Disease", *Intelligent Automation & Soft Computing*, vol. 30, no. 3, pp. 917-928, 2021. Available: 10.32604/iasc.2021.017989

[18] S. Kabiraj, L. Akter, M. Raihan, N. J. Diba, E. Podder and M. M. Hassan, "Prediction of Recurrence and Non-recurrence Events of Breast Cancer using Bagging Algorithm", *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*,2020, pp. 1-5, doi: 10.1109/ICCCNT49239.2020.9225440.

[19] P. Ghosh, A. Karim, S. T. Atik, S. Afrin, and M. Saifuzzaman, "Expert cancer model using supervised algorithms with a lasso selection approach", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 3, p. 2631, 2021.

[20] C. Liang, B. Shanmugam, S. Azam, M. Jonkman, F. D. Boer and G. Narayansamy, "Intrusion Detection System for Internet of Things based on a Machine Learning approach", *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*, 2019, pp. 1-6, doi: 10.1109/ViTECoN.2019.8899448

[21] P. Ghosh, S. Azam, A. Karim, M. Hassan, K. Roy, and M. Jonkman, "A comparative study of different machine learning tools in detecting diabetes", *Procedia Computer Science*, vol. 192, pp. 467–477, 2021.

[22] Z. Tasnim, S. Chakraborty, F. M. Shamrat, A. N. Chowdhury, H. A. Nuha, A. Karim, S. B. Zahir, and M. M. Billah, "Deep learning predictive model for colon cancer patient using CNN-based classification", *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, 2021.