IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Selecting Important Features For CKD Diagnosis And Prediction By Using LightGBM.

Prof. Sayaram Shingote[1],Pratiksha Davkhar[2], Shweta Lokhande[3], Vaishali Parande[4]

[1](Computer Engineering, SGOI's COE, Belhe / SPPU,Pune India)
[2](Computer Engineering, SGOI's COE, Belhe /SPPU, Pune India)
[3](Computer Engineering, SGOI's COE, Belhe / SPPU, Pune India)
[4](Computer Engineering, SGOI's COE, Belhe / SPPU, Pune India)

Abstract: *Chronic Kidney Disease (CKD) is an exceeding health problem which is affecting 10% of world's population and 17% of the Indian population. So it is more important to diagnose it correctly. As we know that the machine learning (ML) techniques have unique advantages and that's why they are widely used for analysis and prediction of diseases. Researchers have used the CKD dataset provided by UCI repository to develop predictive models for CKD. In this paper we have compared and reviewed Extreme Gradient boosting and LightGBM and the LightGBM is chosen as a further predictive model for its high performance of speeding up the training time and less memory consumption. The dataset has 24 attributes and requires18 plus tests to be conducted to conclude these attributes which are costly. That is why we have also reviewed some of the feature selection methods and tried to use only the most important predictive attributes to save the training time of model and the most important tried to reduce the cost required to do these testsa. Also we have analyzed GFR calculation méthods and used mostly suggested method CKD-EPI tofind from which stage of CKD the patient is suffering.*

Keywords - *Chronic Kidney Disease, Extreme Gradient Boosting, feature selection, GFR, LightGBM, Machine learning, Medical diagnosis.*

1. Introduction

Chronic kidney disease is in short CKD is a disease which affects functionality of human kidneys and if it lasts for long time, it results into kidney failure. Kidneys fuction is to filter out the blood and CKD used to brake this. When kidney disease reaches an advanced stage, it means if the damage is very heartache then the kidneys may stop working.If the kidneys will failed, we will be needing the dialysis or the kidney transplantation which is very costly and basically impossible for the peoples who belongs to middle class or poor families. If the disease is detected as early as possible, the further progression of chronic kidney disease will be prevented.Thus the lifeline of the patient will be increased. So keeping it as our prior objective we are using machine learnig techniques for early detection of CKD.

1.1 Related Work

Adeola Ogunleye and Qing-Guo Wang[1] developed Extreme Gradient Boosting (XGBoost) based Chronic Kidney disease diagnosis system.They reviewed existing classification algorithmsand have used (XGBoost) model by combining three features selection techniques for rapid and accurate diagnosis of CKD.

C. S. Lee and M. H. Wang [2]designed a fuzzy expert system for diabetes prediction. They developed a unique five layer fuzzy ontology to vesion the area understanding with uncertainty and volume the bushy ontology to the diabetes area. Additionally they worked on a semantic decision support agent[SDSA] for semantic decision making in diagnosis of diabetes.

Deepa Gupta , Sangita Khare and Ashish Aggarwal [3] proposed a method for the prediction of diagnostic codes for chronic diseases by applying machine learning algorithms. They made a specialty of scientific records, claims records for analysing 11 continual disorder which includes kidney disorder, osteoporosis, arthritis etc. the use of the declared records.

V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra [4] proposed a lung cancer diagnosis system with the help of data mining type techniques.They compared different machine learning classification techniques such as Naïve Bayes, Decision Tree and Neural Network. In that Naïve Bayes performed better than others as it picks out allo of the considerable predictors.

Younghwan Shin, Hyun Soo Chung, Sangdo Kim ,Sang Gil Han, Jong Moon Chung, Junho Cho [5] proposed a unique Blood sample-based Emergency department (ED) Return (BER) systemwhich is usedfor the prediction of the ED return. Here they have used LightGBM algorithm against XGBoost algorithm. They in this comparison got effective outcomes as compared to xgboost. LightGBM effectively predicted the ED return probability of hospitals.

Guolin Ke , Qi Meng , Thomas Finley , Taifeng Wang , Wei Chen , Weidong Ma , QiweiYe Tie-Yan Liu[6], developed a fast, an accurate and highly efficient algorithm called LightGBM.They have included GOSS and EFB sampling methods in LightGBM for boosting up the performance of LightGBM. LightGBM appreciably and effectively outperformed than XGBoost and SGB in the contrast of training speed and consumption of memory.

Fatimah Alzamzam, Mohmad Hoda and Abdul motalel EL soddik[7], usedLight gradient boosting machine (LightGBM) for sentiment analysis and classification based on short texts. Their LGBM model was trained to classify tweets sentiments in: positive,negative or neutral categories. Result showed that, LightGBM based LGBM sentiment classifier outperformedthan the other classification algorithms in case of accuracy and F-scores.

Dingling Ge ,shunyu chang [8] proposeda LightGBM based Credit Card Fraud Detection system.They compared various algorithms such as SVM (Support Vector Machine), Random Forest and XGBoost. The results shown that LightGBM is an effective modelling algorithm than the above.

Jian Ping Li, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan [9], presented heart disease diagnosis using machine learning techniques. They reviewed various classification algorithms includes Support vector machine, Logistic regression, Artificial neural network, K-nearest neighbor, Naïve bays, and Decision tree.Also they took a look on some of the standardized feature selection methods such as Minimal redundancy maximal relevance, Least absolute shrinkage selection operator and Local learning to remove irrelevant features and save the training time of model. They also proposed a unique and fast conditional mutual information feature selection algorithm.

Jović , K. Brkić and N. Bogunović [10],presented feature selection and has provided an overview of the available feature selection techniques to handleseveral different typesof problems. Alsoanalyzed application of feature selection domain and reviewed comparative studies on feature selection to analyzed which method is best suited for the particular class of problem.

1.2 Our Contribution

We proposed a machine learning based diagnosis method for identification of chronic kidney disease in this paper. We reviewed XGBoost algorithm and compared with LightGBM which is lighter version version of the Gradient Boosting Decision Tree. The LightGBM is chosen as a further predictive model for its high performance of speeding up the training time and less memory consumption.

- Firstly we reviewed lightGBM model,then used for chronic kidney classification problem.LightGBM is an open-source framework for gradient boosted machines.
- It requires very less training time as compared to all machine learning algorithms and its also faster than the other gradient boosting algorithms such as Xgboost,Catboost etc.
- Secondly, we are applying some of the feature selection methods from filter, wrapper, and embedded approach.By combining all results we are selecting 5 attributes which are common in the results of all methods.
- Lastly, we are applying CKD-EPI equation to find out the stage of ckd patient.Serum creatine which is considered as a important indicator of CKD is used to compute the stages of CKD by putting it into the CKD-EPI equation.

2. Existing System

Existing system was based on the XGBoost algorithm:

- We have to manually create dummy variable/label encoding for categorical feature before feeding them into the model.
- Slow,less efficient.
- Requires more memory space.

3. Proposed System

3.1 Flowchart for proposed system:

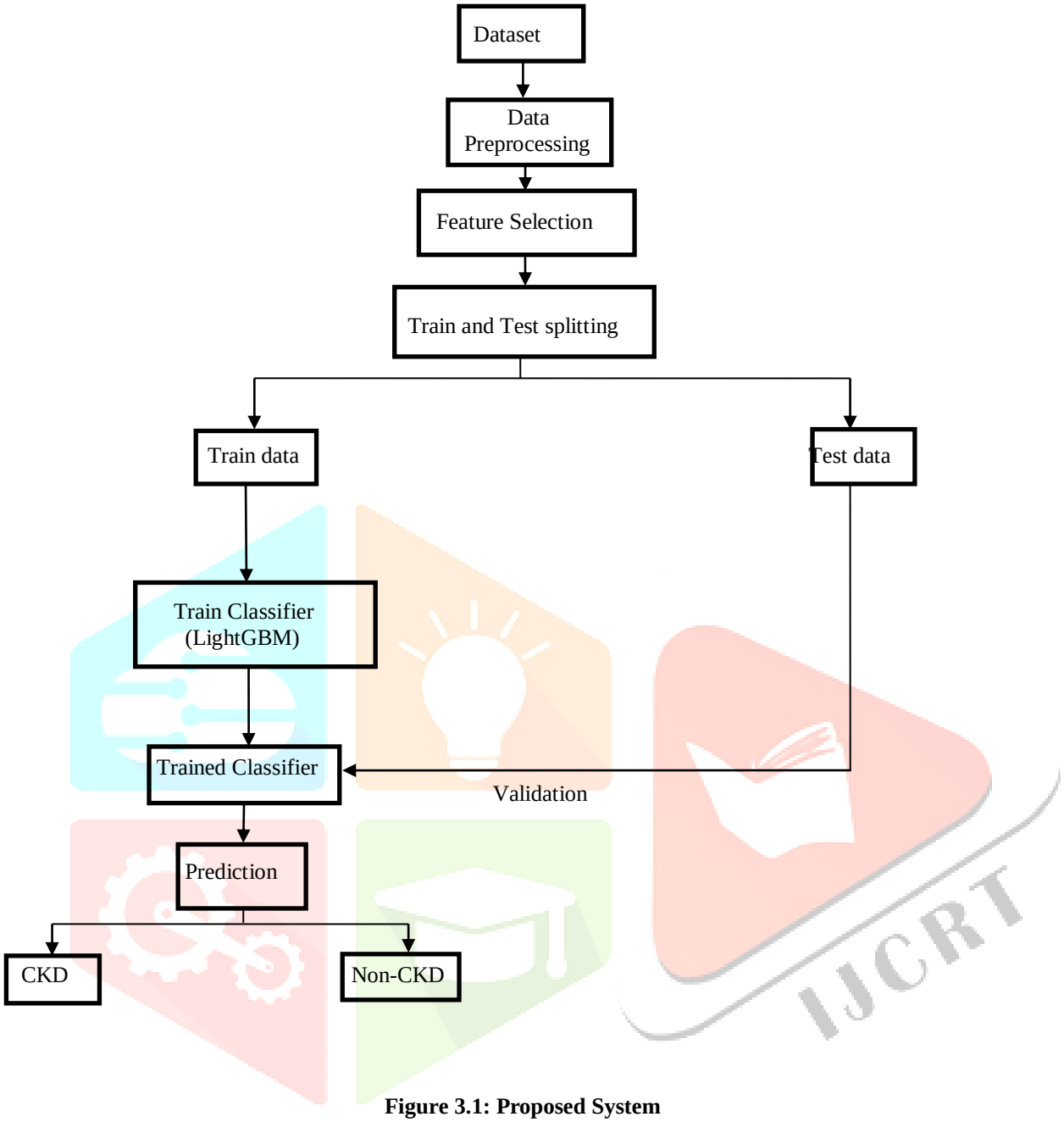

Figure 3.1: Proposed System

3.2 Dataset

The CKD dataset used to build the model in our paper was taken from University of California, Irvine (UCI) learning repository[11].

The dataset contains twenty-six attributes one of them is id which is un-necessary. The remaining twenty-four attributes are listed in the below table and one attribute denotes the output referred as class attribute which contains categorical value i.e. (ckd or notckd) which indicates the class of patient. There are some categorical (Nominal) attributes in the dataset and some are numerical.We have used abbreviation of these attributes in the dataset which are also given in the below table.

TABLE I

Attribute information and Abbreviation

Attribute Number	CKD Attributes	Data type	Abbreviation
1	Age	Number (in years)	age
2	Blood Pressure	Number (in mm/hg)	bp
3	Specific Gravity	Nominal	sg
4	Albumin	Nominal(1-5)	al
5	Sugar	Nominal(1-5)	su
6	Red Blood Cells	Nominal (normal,abnormal)	rbc
7	Pus Cells	Nominal (normal, abnormal)	pc
8	Pus Cell Clumps	Nominal (present, notpresent)	pcc
9	Bacteria	Nominal (present, notpresent)	ba
10	Blood Glucose Random	Number (in mgs/dl)	bgr
11	Blood Urea	Number (in mgs/dl)	bu
12	Serum Creatinine	Number (in mgs/dl)	sc
13	Sodium	Number (in mEq/L)	sod
14	Pottasium	Number (in mEq/L)	pot
15	Hemoglobin	Number (in gms)	hemo
16	Packed Cell Volume	Nominal	pcv
17	White Blood Cell Count	Number (in cells/cumm)	wc
18	Red Blood Cell Count	Number (millions/cmm)	rc
19	Hypertension	Nominal (yes, no)	htn
20	Diabetes Mellitus	Nominal (yes, no)	dm
21	Coronary Artery Disease	Nominal (yes, no)	cad
22	Appetite	Nominal (good, poor)	appet
23	Pedal Edema	Nominal (yes, no)	pe
24	Anemia	Nominal (yes, no)	ane

3.3 Preprocessing

The data pre-processingis required for good representation as well as for best results. Techniques of pre-processing such as removing attribute missing values or doing missing value imputation using mean, modeand median. Our dataset has some categorical attributes so to normalize it we have to apply anyone of the normalization techniques such as Standard Scalar (SS) or Min-Max Scalar. Here we have applied Standard scalar normalizing method to our dataset.

3.4 Feature selection

After data pre-processing, we are applying feature selection process.We have used three feature selection techniques from Filter method, one from wrapper and one from embedded method,

- From filter method we have used SelectKBest and chi2statistic test.
- From wrapper method we have RFE(Recursive Feature Elimination).
- In embedded we have used SelectFromModel for LighGBM.

4. LightGBM based CKD prediction model

Light Gradient Boosting Machine referred as LightGBM is a gradient boosting framework.It is a decision treebased learning algorithm, mostly used for machine learning classification problems. It can also be appliedon machine learning based regression tasks. It takes very less computational time for training of data as compared to other gradient boosting algorithms as the trees in LightGBM algorithm grows leaf-wise while in other algorithms it grows level-wise.

LightGBM uses two important sampling methods such as GOSS and EFB,

- With GOSS the model excludes the instances on the basis of gradients, the instances with large gradients which contributes more to the information gainare kept and it excludes instances with small gradients based on some predefined threshold.
- EFB reduces the features to reduce the training complexity which are basically mutually exclusive. Then these exclusive features later on bundled into a single feature called Exclusive feature bundle and treated as a single feature.

4.1 GOSS Calculation

1) Keep all the large gradient instances.
2) Randomly samplethe small gradient instances
3) Use a constant multiplier for small gradient data instancesduring information gain computation in the tree building process.
4) If *a* be the instances with large gradients and *b* is the randomly sampled small gradientinstances, the sampled data will be amplified by (1-a/b).

4.2 EFB

1) Find out the mutually exclusive features which can be bundled together.
2) Merge the features into a single feature,resulted in the first step.

5. GFR calculation

GFR refers to Glomerular filtration rate. It basically measures filtration rate of kidney. The National Kidney Foundation has recommendedto use CKD-EPI Creatinine Equation (2009) for estimation of GFR.
GFR value tells the stage of kidney disease and helps to plan the corresponding treatment.If the GFR value is low, then kidneys are not working as they have to. As the earlier detection of CKD is possible then then the progression of CKD can be stopped.

5.1 CKD-EPI Creatinine Equation (2009)

CKD-EPI equation is expressed as follows:

GFR = $141 * \min(Scr / k, 1)^{\alpha} * \max(Scr / k, 1)^{-1.209} * 0.993^{age} * 1.018$ [if female] * 1.159 [if Black]

5.2Abbreviations / Units

SCr: serum creatinine (mg/dl)
k: 0.7 for females and 0.9 for males
α : -0.329 for femalesand -0.411 males
min: indicates the minimum of SCr/κ or 1
max: indicates the maximum of SCr/κ or 1
age: age of patient in year

6.Conclusion

CKD is a gradual loss in kidney function causes kidney failure, if it is detected as early as possible then by taking preventive measures against it we can completely cure it or we can prevent the progression of CKD in patients.

In this study, we have developed a machine learning based diagnosis system for diagnosis of chronic kidney disease.We have used LightGBM model for this classification. This model gave us 100% accuracy in model training and also took very less training time as compared to other machine learning algorithms.Also we did features selection, we have selected important features to reduce the training time and as the all attributes requires 18 plus test to be conducted which are costly , by selecting important attributes the cost of the test is also reduced.

Acknowledgement

We would like to acknowledge the Department of Computer Engineering for the support extended for the completion of this work.Also we would like to express our special thanks to Prof. Shingote S. N (Guide) as well as Prof. Borhade B. M (H.O.D of Computer Department) for their whole hearted co-operation and valuable suggestions, technical guidance throughout the work and his kind official support given and encouragement. We would also like to thank our parents and friends for their encouragement in completing this work.

References

[1] Adeola Ogunleye and Qing-Guo Wang, "XGBoost Model for Chronic Kidney Disease Diagnosis",DOI 10.1109/TCBB.2019.2911071, IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[2] C.-S. Lee and M.-H. Wang, "A fuzzy expert system for diabetes decision support application." IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics.

[3]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), pp. 281–287, 2016.

[4] V.Krishnaiah , Dr.G.Narsimha, Dr.N.Subhash Chandra Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniquesnternational Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 39 – 45

[5] Younghwan Shin, Hyun Soo Chung, Sangdo Kim ,Sang Gil Han, Jong Moon Chung, Junho Cho, Emergency "Department Return Prediction System using Blood Samples with LightGBM for Smart Health Care Services", DOI 10.1109/MCE.2020.3015439, IEEE Consumer Electronics Magazine.

[6]Guolin Ke , Qi Meng , Thomas Finley , Taifeng Wang, Wei Chen , Weidong Ma , Qiwei Ye Tie-Yan Liu,"LightGBM: A Highly Efficient Gradient Boosting Decision Tree".In Advances in neural information processing system.

[7]Fatimah Alzamzam, Mohmad Hoda and Abdul motalel EL soddik :"Light gradient boosting machine for general sentiment classification on short texts:A comparative evalution"

[8] Dingling Ge ,shunyu chang "Credit card fraud detection system using LightGBM Model",2020 International Conference on E-Commerce and Internet Technology (ECIT).

[9] Jian Ping Li, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan,And Abdus Saboor:"Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare" Digital Object Identifier 10.1109/ACCESS.2020.3001149.

[10]A. Jović , K. Brkić and N. Bogunović "A review of feature selection methods with applications", MIPRO 2015, 25-29 May 2015, Opatija, Croatia 1200.

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