A Gradient Boosting Classifier for Predicting Chronic Kidney Disease Stages

J. P. Scoralick[0000−0001−8116−3124], G. C. Iwashima[0000−0002−7610−7299], F. A. B. Colugnati[0000−0002−8288−203X], P. V. Z. Capriles[0000−0001−9780−4328], and L. Goliatt[0000−0002−2844−9470]

Federal University of Juiz de Fora, Brazil {jpscoralick, gabriele.cesar.iwashima, capriles, goliatt}@ice.ufjf.br
fernando.colugnati@ufjf.edu.br

Abstract. Chronic Kidney Disease (CKD) is a global public health issue and one of the most neglected chronic diseases worldwide. CKD impacts worldwide morbidity and mortality by other conditions such as diabetes and hypertension, and the treatment can be extremely costly. However, CKD could be prevented or delayed by inexpensive interventions. Once the CKD prediction is successful, the quality control in the diagnostic and treatment of chronic kidney disease can be improved. This paper proposes six different classification algorithms to predict chronic kidney disease stages without considering serum creatinine as a predictive value. By testing these algorithms on the database considered in this paper, our findings show that the Gradient Boosting (XGB) algorithm provides higher accuracy on classification and prediction performance for determining the severity stage in chronic kidney disease, reaching similar precision level in comparison with another approach that considers serum creatinine in the classification model.

Keywords: kidney disease · gradient boosting· computational intelligence.

1 Introduction

When the kidneys' functional unities, known as nephrons, lose their blood filtration capacity over a long time, the kidney function shows slowly and progressive loss. These irreversible damages make the kidneys inefficient to supply patient necessities [11]. This scenario of loss in kidney function is defined as chronic kidney disease (CKD) [14], which is a global public health issue [13], known to be one of the most neglected chronic diseases worldwide [16]. According to the Brazilian Society of Nephrology (SBN), the chronic dialysis program had 126.583 patients with CKD in 2017 [22]. Besides, The Global Burden of Disease (GBD) study, 2017 notified that this disease was in charge of 33·7% of the global deaths [21].

Indirectly, CKD impacts global morbidity and mortality by increasing the risks associated with cardiovascular diseases, diabetes, hypertension, kidney disease progression, acute kidney injury, anemia mineral deficiency, bone disorders,

and fractures. Furthermore, kidney disease, whether acute, chronic, or end-stage, can be extremely costly, but this could be prevented or delayed by inexpensive interventions, preventing adverse outcomes [16,10].

As CKD, the severity of chronic diseases can be mainly characterized by stages of progression. CKD is classified based on measured albuminuria and estimated Glomerular Filtration Rate (GFR) categories, helping to risk-stratify patients (see Fig. 1).

Four personal and clinical values are required to calculate the GFR of a patient: age, gender, race, and serum creatinine [7,6]. Although only age and serum creatinine can vary over time, the latter value can suffer significant variation in a short period according to the patient's clinical condition. Hence, the frequent calculation of serum creatinine value is essential for pre-CKD or CKD patients.

Therefore it is essential to understand the clinical and individual reasons why patients move through these stages and how it happens [15]. Early detection is also crucial for identifying disease severity stages as it helps to decrease the costs, enabling the application of other more efficient treatment methods.

CKD Stages according to GFR and Albuminuria levels

				Albuminuria categories		
				A1	**A2**	**A3**
				Normal to mildly increased	Moderately increased	Severely increased
				<30 mg/g <3 mg/mmol	30-299 mg/g 3-29 mg/mmol	≥300 mg/g ≥30 mg/mmol
GFR Stages	**G1**	Normal or high	≥90			
	G2	Mildly decreased	60-90			
	G3a	Mildly to moderately decreased	45-59			
	G3b	Moderately to severely decreased	30-44			
	G4	Severely decreased	15-29			
	G5	Kidney failure	<15			

Fig. 1. Relationship between GFR values and albuminuria levels for stratification of CKD stages. Colors represents the risk for progression, morbidity, and mortality from the best to worst. Green: Low risk for CKD; Yellow: Moderately Increased risk; Orange: High risk; Red: Very high risk; Deep Red: Highest risk. Adapted from NKF, 2020 [6].

Improving or not your clinical condition, a patient can have your stage changed during treatment according to the GFR value. Hence, this paper aims to analyze and evaluate clinical and individual data related to patient's displacement through all CKD stages. The approach proposed in this paper can predict

the last recorded stage (LRS) of each patient in the dataset. Once this prediction process is successful, healthcare decision-making can be performed in intelligent ways by managers and healthcare professionals, improving the quality control in the diagnostic and treating chronic kidney disease.

Machine learning algorithms have been used to predict and classify in the healthcare field. Support Vector Machine (SVM) Algorithm has been used to classify and predict diabetes and pre-diabetes patients [23]. The results show that SVM is useful in categorizing patients with common diseases. SVM has also been used to classify Alzheimer's disease [17] based on whole-brain anatomical magnetic resonance imaging (MRI). In this study, for a set of patients, and the results show that SVM is a promising approach for Alzheimer's disease early detection. In [5] heart disease prediction using the Probabilistic Neural Network Algorithm, Decision Tree Algorithm, Naive Bayes Algorithm, and PRNN provides the best results compared with other algorithms for heart disease prediction. The study [2] performed the prediction of HBV-induced liver cirrhosis using the Multilayer Perceptron (MLP) Algorithm. The results show that the MLP classifier gives satisfactory liver disease predictions, mostly in HBV-related liver cirrhosis patients. A recent study [19] applied four data mining algorithms on a clinical/laboratory dataset consisting of 361 patients. The addressed algorithms' results have been compared to define the most accurate algorithm results in classifying CKD's severity stage. This study recommends that the Probabilistic Neural Networks algorithm be the best algorithm that physicians can use to eliminate diagnostic and treatment errors.

This paper proposes and compares six different classification algorithms to predict chronic kidney disease stages considering a scenario with the most 25 frequent clinical and laboratory patient data. Our findings show that the Gradient Boosting (XGB) algorithm provides higher accuracy on classification and prediction performance for determining the severity stage in chronic kidney disease.

2 Material and Methods

2.1 Dataset

The dataset used in this paper is composed of data collected between the years of 2010 and 2014 in the IMEPEN Foundation alongside with HIPERDIA Program from the Ministry of Health from Brazil. All the information in the dataset comprises clinical, socioeconomic, and personal data from 7266 patients diagnosed with chronic kidney disease or any comorbidity that would affect kidney function - mainly high blood pressure (HBP) and diabetes mellitus (DM). The data was collected from the Brazilian city of Juiz de Fora and other 36 nearby towns. Furthermore, a project for using this dataset was submitted to the Federal University of Juiz de Fora (UFJF) and got approved under protocol number 36345514.1.0000.5139.

Besides the already know 7266 patients distributed in rows and summing up all data information, the dataset has 255 different variables (numerical and

categorical) distributed in columns. Each row represents a single record which comprises all data related to a patient: personal information such as identification number, gender, age, race, weight, and so on; socioeconomic information like familiar income, education level, place of birth, and so forth; and clinical data as blood and blood pressure tests, numerous medicines, up to eight serum creatinine values (two per year from 2011 to 2014) and their respective GFR and stages values, and several other clinical data. An excerpt from the data used for this paper is described in table 1, and all details of the data and further information will be available by the authors upon request.

Not all the 7266 patients in the raw dataset have at least one serum creatinine value. Hence, for them, there is not even one stage record. A data filtering process was applied in the database to remove all patients without any serum creatinine record, totaling 5689 patients.

Also was applied to the dataset a reorganization process concerning the clinical tests dates from each patient. This process aimed to generate a temporal understanding of the patient's clinical condition throughout the CKD stages that they went through. Consequently, the total number of rows in the dataset changed from 7266 to 40100. In addition, all missing values were replaced by zeros. All analyses described from now on in this paper consider the resulting dataset from applying the steps described previously.

2.2 Classification Algorithms

For the classification process, the performance of Gradient Boosting (XGB) on the proposed data scenario is compared with five methods: Random Forest (RF), Support Vector Classification (SVC), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP) and Extreme Learning Machine (ELM).

Gradient Boosting is an ensemble method for combining several minor problems, called weak learners, to generate a complete problem, a strong learner [3]. In the case of the XGB, weak learners are regularized decision trees [4]. The XGB prediction for a instance i is

$$\hat{y_i} = F_M(x_i) = \sum_{m=1}^{M} h_m(x_i)$$

where `M` is the number of estimators, and h_m are the weak learners built based on parameters of `max_depth` and a minimum loss factor for partitioning a new tree leaf (Γ_{tree}). Since it is a integrative boosting method, it is built in a greedy fashion of the form

$$F_m(x) = F_{m-1}(x) + \eta h_m(x)$$

where η is a learning rate and the newly added tree $h_m(x)$ is fitted in order to minimize a sum of losses L_m and is given by Eq. (1).

$$h_m = \arg\min_h L_m = \arg\min_h \sum_{i=1}^{N} l(y_i, F_{m-1}(x_i) + h(x_i)) + \Omega(F_{m-1}) \qquad (1)$$

In Eq. (1), we take $l(y_i, F_m(x)) = [y_i - F_m(x_i)]^2$, where N is the number of samples and

$$\Omega(F_m) = \alpha_{reg} T + \frac{1}{2}\lambda_{reg}||\mathbf{w}||^2$$

as a regularization term, where T is the number of leaves, $\mathbf{w}$ is the leaf weights, and α_{reg} and λ_{reg} are a L_1 and L_2 regularization term on weights, respectively.

Random Forest Classifier is an ensemble method that produces a prediction model using a collection of decision tree models [8]. A decision tree is a classification model in the form of a tree structure splitting a dataset into increasingly smaller subsets while an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. The RF uses decision trees of fixed size as tree models. In this way, each decision tree is built thought the ensemble, employing random independent subsets of both features and samples. The prediction of a new sample class is performed as follows, each classifier votes, and the most voted class is elected. The minimum number of samples in newly created leaves is the parameter of this method.

Support Vector Classification and Support Vector Machine (SVM) are methods for the classification of both linear and nonlinear data [19] and can be described as follows: given two classes and a set of points that belong to those classes, a SVM determines the hyperplane that separates the points in order to place the largest number of the same class on the same side, while maximizing the distance of each class to that hyperplane [12]. The distance of a class to a hyperplane is the shortest distance between it and the points of that class and is called the separation margin. The hyperplane generated by SVM is determined by a subset of the points of the two classes, called support vectors [12]. SVMs and SVCs can be used for prediction and classification in applications such as Alzheimer's disease early detection, handwritten digit recognition, object recognition as well as benchmark time-series prediction tests [19].

K-Nearest Neighbors is a supervised learning algorithm algorithm that can be used for regression and classification purposes. For the prediction of each data class, KNN checks the labels of data points surrounding a target data point [1,18]. Also, KNN is a non-parametric algorithm, meaning that the model is exclusively created from the data. An important feature of KNN is that the algorithm does not separate the dataset into training and test groups. Consequently, the entire training set is also used for making predictions. Finally, K-Nearest Neighbors is one of the most popular classification and regression algorithms used in general machine learning [18].

Multilayer Perceptron is a class of feedforward artificial neural network with one or more hidden layers with an undetermined number of neurons linked together by synapses with weights. The hidden layer has this name because it is

not possible to predict the desired output in the intermediate layers. Learning in this type of network is usually done through a supervised learning algorithm called backpropagation [20].

Extreme Learning Machine is a single-layer feedforward artificial neural networks where the input to hidden weights are randomly generated [9] . The output function of the ELM is given by $\hat{y}(\mathbf{x}) = \sum_{i=1}^{L} \beta_i G(\mathbf{w}_i, b_i, \mathbf{x})$ where $\hat{y}$ is the ELM prediction associated to the input vector $\mathbf{x}$, $\mathbf{w}_i$ is the weight vector of the ith hidden layer, b_i are the biases of the neurons in the hidden layer, β_i are output weights, $G(\cdot)$ is the nonlinear activation function and L is the the number of neurons in the hidden layer. The parameters $(\mathbf{w}, b)$ are randomly generated (normally distributed with zero mean and standard deviation equals to one), and weights β_i of the output layer are determined analytically using least squares. The ReLU $G(\mathbf{w}, b, \mathbf{x}) = \max_i (0, (\mathbf{w} \cdot \mathbf{x} + b))$ is used as activation function.

3 Computational Experiment

3.1 Classification scenario

Following the approach proposed by Rady *et al.* [19], a classification scenario was proposed, comprising the 25 most frequent clinical and laboratory data, that is, those with the largest number of patients. To predict the last recorded stage in the dataset for each of the 5689 patients, LRS was set as the target variable. All the 26 variables are described in table 1.

It is important to note that although the GFR value depends on four values: age, gender, race, and serum creatinine [7], only the first three were considered for the testing scenario. Since creatinine is the most determining variable in the calculation of GFR, its value was disregarded for the classification. Therefore, our approach considers 22 clinical tests added to the three personal data cited in order to predict the last recorded stage of a patient in the database.

3.2 Experimental setup

In order to evaluate the performance of the methods we used the accuracy metric $\frac{1}{N} \sum_{i=1}^{N} I(f(x_i) = y_i)$ that measures the percentage of correct classes by comparing the predicted classes with the actual ones, where $f(x_i)$ is the predicted class of a sample, y_i is the true class of this sample, $I(true) = 1$ and $I(false) = 0$.

In the experiments conducted here, 70% of the dataset was allocated for training and the remaining 30% for validation. And all missing values were replaced by zeros. A computational experiment, consisting of 100 iterations of the classification algorithm, was implemented for each of them to obtain the average accuracy in predicting the target variable: the last stage recorded for each patient. All codes and data will be available by the authors upon request.

Table 1. Description of the variables used in the analysis.

	Value	Variables	Class	Type	Missing rows
1	-	Gender	Predictor	Categorical	0
2	-	Age	Predictor	Numerical	0
3	-	Race	Predictor	Categorical	0
4	Initial	Systemic_Blood_Pressure	Predictor	Numerical	21
5	Final	Systemic_Blood_Pressure	Predictor	Numerical	21
6	Initial	Diastolic_Blood_Pressure	Predictor	Numerical	21
7	Final	Diastolic_Blood_Pressure	Predictor	Numerical	21
8	Initial	Weight	Predictor	Numerical	50
9	Final	Weight	Predictor	Numerical	50
10	Initial	Hemoglobin	Predictor	Numerical	280
11	Initial	Total_Cholesterol	Predictor	Numerical	285
12	Initial	Fasting_Glucose	Predictor	Numerical	328
13	Initial	Triglycerides	Predictor	Numerical	335
14	Initial	Potassium	Predictor	Numerical	819
15	Initial	HDL_Cholesterol	Predictor	Numerical	389
16	Initial	Urea	Predictor	Numerical	1132
17	Initial	TSH	Predictor	Numerical	1229
18	Initial	Uric_Acid	Predictor	Numerical	1280
19	Initial	Glycated_Hemoglobin	Predictor	Numerical	1338
20	Initial	ALT	Predictor	Numerical	1405
21	Final	Fasting_Glucose	Predictor	Numerical	1453
22	Final	Total_Cholesterol	Predictor	Numerical	1615
23	Final	Triglycerides	Predictor	Numerical	1713
24	Final	HDL_Cholesterol	Predictor	Numerical	1758
25	Final	Hemoglobin	Predictor	Numerical	1770
26	-	Last_Recorded_Stage	Target	Categorical	0

3.3 Results and Discussion

Table 2 shows the average accuracies obtained from the computational experiments for all of the classification algorithms.

Only two algorithms showed an average accuracy greater than 50%. And yet, such algorithms showed very different values. The Random Forest algorithm correctly classified the LRS of just over three in every five patients. On the other hand, the Gradient Boosting algorithm could predict the LRS of the vast majority of the 5689 patients from the database.

Table 2. The average accuracy for each of classification algorithm.

AVERAGE ACCURACY (%)						
Algorithm	**RF**	**XGB**	**SVC**	**KNN**	**MLP**	**ELM**
Accuracy	68	96	25	48	35	36

In comparison to results obtained by Rady *et al.* [19], our approach presented similar accuracy values. By using the Probabilistic Neural Network (PNN) algorithm, the approach proposed by Rady *et al.* [19] reached 96,7% in the overall accuracy, which is essentially the same average accuracy obtained in this paper with XGB algorithm. Although the accuracy values are the same and also the total number of variables, there are essential differences between both approaches.

The first one is the use of serum creatinine for the classification. As we discussed earlier in this paper, serum creatinine is the most influential variable in the calculation of GFR. Hence, its use in the prediction of a patient's LRS is highly recommended, as Rady *et al.* [19] pointed creatinine having 100% of importance as a predictive variable. On the other hand, our approach does not use serum creatinine as a predictive variable, as we described in table 2.

The second difference is that we used a database with 5689 patients with pre-CKD or CKD clinical situations in this paper. Rady *et al.* [19] approach considered a database with only 361 CKD patients. Therefore, our database is more than 15 times larger than the other. Even though our database has many more patients and a significant amount of missing values, with XGB algorithm, we obtained the same accuracy value by Rady *et al.* [19] with PNN algorithm.

4 Conclusion

The continuous and frequent growth of chronic kidney disease (CKD) as a worldwide health problem, especially among patients diagnosed with diabetes mellitus or high blood pressure, fosters novel approaches for its early detection.

For the stratification and classification of CKD, the Glomerular Filtration Rate is the reference value, which can be easily estimated using three personal data from a patient: age, gender, and race, in addition to serum creatinine. Once the GFR is known, it can be used to categorize patients into six predetermined CKD stages. It is essential to understanding how and why a patient has his stage modified during the treatment, taking into account his personal and socioeconomic data and several of his clinical and laboratory data.

In the addressed scenario, comprised of the 25 most frequent clinical and laboratory patient data, the classification evinced significant stage prediction results. Although we used six different classification algorithms, only two showed relevant results: Random Forest algorithm and Gradient Boosting algorithm. However, and by far, XGB presented higher predictive precision, reaching 96% of average accuracy.

This result is the same obtained by Rady *et al.* [19] with the use of PNN algorithm. However, their approach considered a database with only 361 patients and serum creatinine as a predictive variable. Differently, the method presented in this paper does not assess creatinine for the classification task. This is a novel approach not found in related literature.

The proposed method can be seen as an alternative way to obtain Glomerular Filtration Rate values without using serum creatinine values. Hence, it can also

be used in several future works to understand and predict the different cycles that a pre-CKD or CKD patient goes through during his treatment period.

References

1. Altman, N.S.: An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician **46**(3), 175–185 (1992), http://www.jstor.org/stable/2685209
2. Cao, Y., Hu, Z.D., Liu, X.F., Deng, A.M., Hu, C.J.: An mlp classifier for prediction of hbv-induced liver cirrhosis using routinely available clinical parameters. Disease markers **35** (2013)
3. Chen, T., Guestrin, C.: Xgboost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Aug 2016). https://doi.org/10.1145/2939672.2939785, http://dx.doi.org/10.1145/2939672.2939785
4. Chen, T., He, T.: Higgs boson discovery with boosted trees. In: Proceedings of the 2014 International Conference on High-Energy Physics and Machine Learning - Volume 42. p. 69–80. HEPML'14, JMLR.org (2014)
5. Dessai, I.S.F.: Intelligent heart disease prediction system using probabilistic neural network. International Journal on Advanced Computer Theory and Engineering (IJACTE) **2**(3), 2319–2526 (2013)
6. Foundation, N.K.: Estimated glomerular filtration rate (egfr) (October 2020), https://www.kidney.org/atoz/content/gfr, access in: 2020-10-11
7. Foundation, N.K.: Glomerular filtration rate calculator (October 2020), https://www.kidney.org/professionals/KDOQI/gfr_calculator, access in: 2020-10-30
8. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics **29**(5), 1189–1232 (2001)
9. Huang, G., Huang, G.B., Song, S., You, K.: Trends in extreme learning machines: A review. Neural Networks **61**(Supplement C), 32 – 48 (2015)
10. Jha, V., Garcia Garcia, G., Iseki, K., Li, Z., Naicker, S., Plattner, B., Saran, R., Wang, A., Yang, C.W.: Chronic kidney disease: Global dimension and perspectives. Lancet **382** (05 2013)
11. Junior, R.: Doença renal crônica: Definição, epidemiologia e classificação. J Bras Nefrol. **26** (01 2004)
12. K, A.A., Aljahdali, S., Hussain, S.N.: Article: Comparative prediction performance with support vector machine and random forest classification techniques. International Journal of Computer Applications **69**(11), 12–16 (May 2013), full text available
13. Levey, A., Atkins, R., Coresh, J., Cohen, E., Collins, A., Eckardt, K.U., Nahas, M., Jaber, B., Jadoul, M., Levin, A., Powe, N., Rossert, J., Wheeler, D., Lameire, N., Eknoyan, G.: Chronic kidney disease as a global public health problem: Approaches and initiatives - a position statement from kidney disease improving global outcomes. Kidney international **72**, 247–59 (09 2007)
14. Li, L., Astor, B., Lewis, J., Hu, B., Appel, L., Lipkowitz, M., Toto, R., Wang, X., Wright, J., Greene, T.: Longitudinal progression trajectory of gfr among patients with ckd. American journal of kidney diseases : the official journal of the National Kidney Foundation **59**, 504–12 (01 2012)
15. Luo, L., Small, D., Stewart, W., Roy, J.: Methods for estimating kidney disease stage transition probabilities using electronic medical records. eGEMs (Generating Evidence & Methods to improve patient outcomes) **1** (12 2013)

16. Luyckx, V., Tonelli, M., Stanifer, J.: The global burden of kidney disease and the sustainable development goals. Bulletin of the World Health Organization **96**, 414–422D (06 2018)
17. Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., Sarazin, M., Dubois, B., Lehéricy, S., Benali, H.: Support vector machine-based classification of alzheimer's disease from whole-brain anatomical mri. Neuroradiology **51**(2), 73–83 (2009)
18. Nelson, D.: What is a knn (k-nearest neighbors)? (October 2020), `https://www.unite.ai/what-is-k-nearest-neighbors/`, access in: 2020-10-31
19. Rady, E.H.A., Anwar, A.S.: Prediction of kidney disease stages using data mining algorithms. Informatics in Medicine Unlocked **15**, 100178 (2019)
20. Ramchoun, H., Amine, M., Janati Idrissi, M.A., Ghanou, Y., Ettaouil, M.: Multilayer perceptron: Architecture optimization and training. International Journal of Interactive Multimedia and Artificial Inteligence **4**, 26–30 (01 2016). https://doi.org/10.9781/ijimai.2016.415
21. Roth, G.A., Abate, D., Abate, K.H., Abay, S.M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., et al.: Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. The Lancet **392**(10159), 1736–1788 (2018)
22. Thomé, F.S., Sesso, R.C., Lopes, A.A., Lugon, J.R., Martins, C.T.: Brazilian chronic dialysis survey 2017. Brazilian Journal of Nephrology **41**(2), 208–214 (2019)
23. Yu, W., Liu, T., Valdez, R., Gwinn, M., Khoury, M.J.: Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC medical informatics and decision making **10**(1), 16 (2010)