A Random Forest Classifier combined with Missing Data Strategies 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 global 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 the treatment of chronic kidney disease can be improved. In this paper, we propose a Random Forest (RF) classifier combined with the imputation of missing data to predict chronic kidney disease stages. We tested four different scenarios, and our findings show that the RF algorithm provides higher accuracy on classification and prediction performance for determining the severity stage in chronic kidney disease.

Keywords: kidney disease · random forest· 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 [7]. This scenario of loss in kidney function is defined as chronic kidney disease (CKD) [9], which is a global public health issue [8], known to be one of the most neglected chronic diseases worldwide [11]. According to the Brazilian Society of Nephrology (SBN), the chronic dialysis program had 126.583 patients with CKD in 2017[17]. Besides, in The Global Burden of Disease (GBD) study, 2017 notified that this disease was in charge of 33·7% of the global deaths [16].

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 [6,11].

As CKD, the severity of chronic diseases can be mainly characterized by stages of progression or regression. CKD is classified based on measured albuminuria and estimated Glomerular Filtration Rate (GFR) categories, helping to risk-stratify patients (see Fig. 1). Therefore it is essential to understand the clinical and individual reasons why patients move through these stages and how it happens [10]. 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.

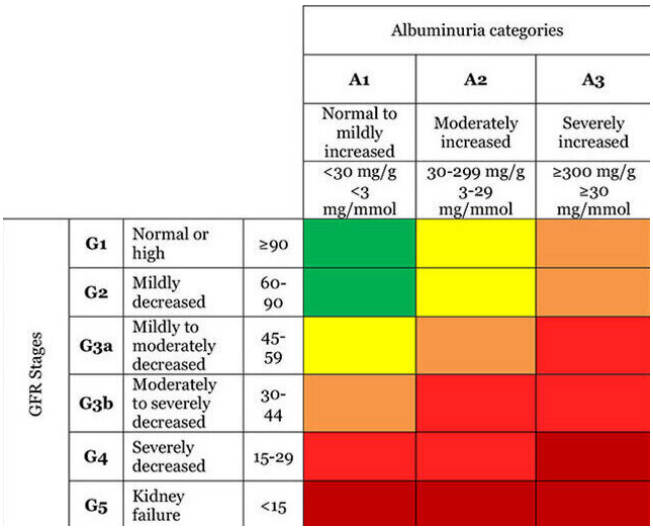

Fig. 1. Colors represents the risk for progression, morbidity, and mortality from the best to worst. Green: Low risk for CKD; Yellow: Moderately Increased risk; Orange: High risk; Red: Very high risk; Deep Red: Highest risk. Adapted from NKF, 2020[4].

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, health care 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 prediabetes patients [18]. The results show that SVM is useful in classifying patients with common diseases. SVM has also been used to classify Alzheimer's disease [12] 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 [3] heart disease prediction using the Probabilistic Neural Network Algorithm, Decision tree Algorithm, Naïve Bayes Algorithm, and PRNN provides the best results compared with other algorithms for heart disease prediction. The study [1] performed the prediction of HBV-induced liver cirrhosis using the Multilayered 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 [15] four data mining algorithms on a clinical/laboratory dataset consisting of 361 CKD 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.

In this paper, we propose a Random Forest (RF) classifier combined with the imputation of missing data to predict chronic kidney disease stages. Besides, missing data were not extensively been studied in the literature. Four different scenarios are tested, and our findings show that the GB 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 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 CKD 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, sex, 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.

2.2 Data Analysis and Filtering

Not all the 7266 CKD 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 CKD patients.

Also was applied to the dataset a reorganization process concerning the exams 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.

All the scenarios and analyses described from now on in this paper consider the resulting dataset from applying the steps described previously.

Five different data input methods were applied for handling missing data: (1) replacement by zeros, (2) by mean and by (3) median; (4) replacement by the K-nearest neighbor algorithm (KNN) [2], with two neighbors; and (5) replacement by Multiple Imputation by Chained Equations method (MICE), since the dataset type can be categorized as Missing at Random (MAR), according to the approach proposed by Pedersen et al. [13]

2.3 Random Forest Classifier

Random Forest (RF) is an ensemble method that produces a prediction model using a collection of decision tree models [5]. 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 considers additive models of the following form:

$$F(x) = \sum_{m=1}^{N} \gamma_m h_m(x) \tag{1}$$

where $h_m(x)$ are the tree models, γ_m is the step length of each tree and N is the number of trees.

In each iteration, it first randomly selects a set of samples from the training set. To reproduce a decision tree from this subset, the RF randomly chooses a subset of features as each node's candidate features. Then, RF constructs the additive model in mth stage as:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x). \tag{2}$$

The initial model F_0 is usually the average of the label values.

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.

3 Computational Experiments

3.1 Scenarios

Four prediction scenarios were proposed to predict the last recorded stage in the dataset for each of the 5689 CKD patients. In all cases, the target variable was LRS, which is why only the first recorded creatinine value was considered in scenarios that encompass a serum creatinine value.

The first scenario only has the four main values in the calculation of GFR: sex, age, race, and serum creatinine. Following the approach proposed by Rady et al. [15], the second scenario considers the 25 most frequent exams and data, that is, those with the largest number of patients. In addition to the first recorded creatinine value, for the third scenario was considered the total number of times a patient went through each of the three ambulatories, per semester, during the 2011-2014 quadrennium: systemic arterial hypertension (SAH), diabetes mellitus (DM) and chronic kidney disease (CKD). Finally, the fourth scenario consists of combining scenarios 2 and 3: the 25 most frequent exams and data, the first recorded creatinine value, and the total number of times of a patient in each ambulatory.

3.2 Experimental setup

A classification process using the Extremely Randomized Trees Classifier (Extra Trees) was applied to all four scenarios, using the Train-Test Split function from scikit-learn, a free software machine learning library for the Python programming language [14]. Moreover, 70% of the dataset was allocated for training, and the remaining 30% for validation. The computational experiments described here were conducted based on the scikit-learn framework [14]. All codes and data are made available by the authors upon request. To obtain consistent and reliable results, 25 independent runs with different random seeds.

For each scenario, we use each of the five data entry methods only once. And for each of them, a computational experiment, consisting of 100 iterations of the classification algorithm, was implemented to obtain the average accuracy in predicting the target variable: the last stage recorded for each patient.

3.3 Results and Discussion

The table 1 shows the average accuracies obtained from the computational experiments of the four scenarios and the five types of data input applied to them. 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 $I(false) = 0$.

As shown in table 1, considering only the four necessary data in the calculation of GFR, scenario 1, it is possible to predict the last recorded stage of more than 4 out of 5 CKD patients. In scenario 2, the average accuracies show the lower results in this analysis. Although this scenario was established following Rady et al. [15] proposal, which also considers 25 data, the results highlight a relevant loss in the accuracy compared to the other three scenarios. This discrepancy can be explained by the differences between the database used in this paper and the one by Rady et al. [15]. The similarity between these two datasets is way lower than 50%.

Table 1. The average accuracy of the four scenarios according to each data input method.

	AVERAGE ACCURACY (%)			
	Scenario 1	***Scenario 2***	***Scenario 3***	***Scenario 4***
Zero	83.75	68.77	95.17	98.53
Mean	83.76	69.82	95.17	98.82
Median	83.76	69.85	90.58	97.24
KNN	83.76	69.83	95.17	96.42
MICE	83.75	69.76	95.13	98.05

In scenario 3, the combination between the first recorded creatinine value and the total times a patient went through DM, DRC, and SAH ambulatories significantly increased the average accuracy for its second-highest values. Therefore, we concluded that knowing how many times and what ambulatories a patient attended during his treatment is essential to LES prediction and, therefore, understanding how the clinical conditions evolve. Up to now, this approach does not have any known equivalence in related literature.

Scenario 4, combining both 2 and 3, shows the highest average accuracy values, presenting results close to 100%. This approach also does not have any known equivalence in related literature. Hence, the 25 most frequent exams and data alongside initial creatinine value and data related to the previously described ambulatories can predict the last stage recorded for almost all of the 5689 CKD patients considered in this paper.

4 Conclusion

The continuous and frequent growth of chronic kidney disease 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, sex, 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 exams.

The last recorded stage of a patient can be predicted, with different confidence levels, by the random forest classifier proposed in this paper alongside five other imputation methods for handling missing data.

In the four addressed scenarios, the classification evinced significant results in stage prediction. Scenario 1 encompassed only the necessary data for the calculation of GFR. With only three personal data and one clinical exam, the classification was 83% accurate, regardless of the data imputation method.

On the other hand, consisting of the most 25 frequent exams, Scenario 2 showed approximately 69% of accuracy in the prediction. However, this was an approach based on the data selection done by Rady et al. [15].

Scenarios 3 and 4 provide the most accurate classification, being able to predict, in most cases, the LRS for more than 95% of 5689 selected patients from the dataset. The lower accuracy was obtained with the median imputation method in scenario 3: 90.58%. And in scenario 4, 98.82% was the highest accuracy among all others, and it was obtained using the median imputation method.

Consequently, having the information on how many times and what ambulatories a patient attended during his treatment, combined with the first recorded creatinine value and the most frequent exams, proves to be a powerful novel approach in the classification and prediction of the last recorded stage a CKD patient. Therefore, the scenario 4 approach can be used in different future works to understand and predict the different cycles that a patient goes through during his treatment period.

References

1. 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)
2. Cunningham, P., Delany, S.: k-nearest neighbour classifiers. Mult Classif Syst (04 2007)
3. 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)
4. Foundation, N.K.: Estimated glomerular filtration rate (egfr) (October 2020), `https://www.kidney.org/atoz/content/gfr`, access in: 2020-10-11
5. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics **29**(5), 1189–1232 (2001)
6. 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)
7. Junior, R.: Doença renal crônica: Definição, epidemiologia e classificação. J Bras Nefrol. **26** (01 2004)

8. 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)
9. 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)
10. 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)
11. 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)
12. 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)
13. Pedersen, A., Mikkelsen, E., Cronin-Fenton, D., Kristensen, N., Pham, T.M., Pedersen, L., Petersen, I.: Missing data and multiple imputation in clinical epidemiological research. Clinical Epidemiology **Volume 9**, 157–166 (03 2017)
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Édouard Duchesnay: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**(85), 2825–2830 (2011)
15. Rady, E.H.A., Anwar, A.S.: Prediction of kidney disease stages using data mining algorithms. Informatics in Medicine Unlocked **15**, 100178 (2019)
16. 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)
17. 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)
18. 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)