



SEINÄJOEN AMMATTIKORKEAKOULU
SEINÄJOKI UNIVERSITY OF APPLIED SCIENCES

Pedro A. Moreno-Sanchez

**Development and evaluation of an explainable prediction
model for chronic kidney disease patients
based on ensemble trees**

2022

ISBN 978-952-7317-69-3



Development and evaluation of an Explainable Prediction Model for Chronic Kidney Disease Patients based on Ensemble Trees

Pedro A. Moreno-Sanchez ¹[0000-0003-3987-1789]

¹ School of Health Care and Social Work, Seinäjoki University of Applied Sciences Seinäjoki,
Finland.

`pedro.morenosanchez@seamk.fi`

Phone: +358 403752546

Total words of manuscript: 4349 words

Total words of abstract: 207 words

Number of figures: 4 figures

Number of tables: 6 tables

Development and evaluation of an Explainable Prediction Model for Chronic Kidney Disease Patients based on Ensemble Trees

Pedro A. Moreno-Sanchez ¹[0000-0003-3987-1789]

¹ School of Health Care and Social Work, Seinäjoki University of Applied Sciences Seinäjoki, Finland.

pedro.morenosanchez@seamk.fi

Abstract. Chronic Kidney Disease (CKD), where delayed recognition implies premature mortality, is currently experiencing a globally increasing incidence and high cost to health systems. Data mining allows discovering subtle patterns in CKD indicators to contribute to an early diagnosis. This work presents the development and evaluation of an explainable prediction model that would support clinicians in the early diagnosis of CKD patients. The model development is based on a data management pipeline that detects the best combination of ensemble trees algorithms and features selected concerning classification performance. Furthermore, the main contribution of the paper involves an explainability-driven approach that allows selecting the best predictive model maintaining a balance between accuracy and explainability. Therefore, the most balanced explainable predictive model implements an extreme gradient boosting classifier over 3 features (packed cell value, specific gravity, and hypertension), achieving an accuracy of 99.2% and 97.5% with cross-validation technique and with new unseen data respectively. In addition, an analysis of the model's explainability shows that the packed cell value is the most relevant feature that influences the prediction results of the model, followed by specific gravity and hypertension. This small number of feature selected results in a reduced cost of the early diagnosis of CKD implying a promising solution for developing countries.

Keywords: medical XAI; clinical prediction model; Chronic Kidney Disease; feature selection; explainability.

1 Introduction

Chronic kidney disease (CKD) is manifested as a worldwide public health problem with increasing incidence and prevalence that leads an ample number of patients to premature mortality [1], and implies high cost to healthcare systems especially in developing countries where lack of appropriate treatment results in a high mortality rate within their population [2][3]. Typically, CKD is a disease with no early symptoms that when detected the kidney has already lost 25 per cent of its capacity and is under progressive damage that, if not slowed by controlling underlying risk factors

(hypertension, obesity, heart disease, age, diabetes, drug abuse, family history of kidney disease, race/ethnicity) [4], the hemodialysis or even kidney transplantation are crucial for patient survival. [5–7]. Therefore, an early diagnosis of CKD based on those indicators allows initiating treatments to slow the progression of kidney damage and prolong patients' life even with an insufficient kidney function.

In the field of medicine, Artificial intelligence (AI) have become, in recent years, a promising instrument to build computer-aided diagnosis (CAD) [8], [9], that could be employed to discover latent correlations between CKD onset and its indicators enabling an early discover of those patients at risk. In the literature, several AI approaches for CKD detection employed their own data sources as medical images or tabular datasets of clinical indicators extracted from EHR [10]. Other researchers used a public common CKD dataset from the University of California Irvine-Machine Learning [11] to build their prediction models allowing reproducibility of results as well as benchmarking between other models implementation. For the sake of these latter aspects, Table 1 shows the most recent and accurate works (accuracy above 98%) that employed the CKD dataset of the UCI-ML repository [11] and applied feature selection approaches in their data preparation stages to build the prediction models.

When CAD systems' decisions affect patients' life eventually, explanations about the output's logic of the AI models are crucial to support clinicians in their diagnosis and treatments. Thus in the medical domain, eXplainable Artificial Intelligence (XAI) would allow healthcare experts to make reasonable and data-driven decisions as well as improve the clinical adoption of AI models [12]. Global model-specific solutions of XAI have been developed for the last decade in different clinical fields, namely: urology [13], toxicology [13], endocrinology [14], neurology [15], cardiology [16], cancer (e.g. breast cancer or prostate cancer) [17], [18], and chronic diseases (e.g. diabetes or Alzheimer's disease) [19, 20]. Concerning XAI application to prediction models, an inherent trade-off must be addressed between AI models' predictive accuracy and their explainability, because most accurate models are usually less transparent and vice versa.

This paper aims at describing the development and assessment of an explainable prediction model of CKD that address the balance between accuracy and explainability through an automated data pipeline that implements different ensemble trees algorithms and feature selection techniques to achieve the best accuracy. In addition, an explainability analysis is conducted in terms of feature relevance and explainability metrics. The outline of the article adopts the following structure: Section 2 gives an overview of the dataset as well as the machine learning algorithms, evaluation metrics and explainability techniques employed in this research. Section 3 presents the pipeline employed to build the prediction model, the evaluation results in terms of classification and explainability, and the explainability analysis. The discussion of the results, the comparison to the related works, as well as the potential impacts of the prediction model on the healthcare domain are described in Section 4. Finally, Section 5 includes the conclusions drawn from the work.

Table 1. Classification results (in %) of related works and machine learning classifiers (best ones in italic). (*Acc*: accuracy; *Sen*: sensitivity; *Spe*: specificity; *F1*:f1-score; *Pre*: Precision; *#F*: number of features; *: Studies that perform the best classifier with unseen new data).

Article	Acc	Sen	Spe	F1	Pre	#F	Machine Learning Classifier
Ekanayake [21]	100	100	-	100	100	7	<i>DT, RF, XGB, Ada, ET</i> (*)
Alaoui [22]	100	-	-	-	-	23	<i>XGB Lin, Lin SVM, DT, RF</i>
Ogunleye [23]	100	100	100	-	100	12	<i>XGB</i> (*)
Zeynu [24]	99.5	99.5	-	99.5	99.5	8	<i>KNN, DT, ANN, NB, SVM.</i>
Raju [25]	99.3	99	-	99	100	5	<i>XGB RF, LR, SVM, NB</i> (*)
Khan [26]	99.1	99.7	-	99.3	98.7	23	<i>NB, LR, SVM, DT, RF</i>
Hasan [27]	99	-	-	99	-	13	<i>Ada, RF, GB, ET</i> (*)
Abdullah [28]	98.8	98.0	100	98.8	98.0	10	<i>RF, SVM, NB, LR</i>
Alaiad [29]	98.5	99.6	96.8	-	98	12	<i>NB, DT, SVM, KNN, Jrip</i>
Kadhun [30]	98.1	98	-	98	98	10	<i>SVM, ELM</i>

2 Material and methods

2.1 Chronic Kidney Disease dataset

To promote the reproducibility of this research, the UCI-ML dataset was employed. Table 2 shows the structure of the dataset that includes 400 samples collected from the Apollo Hospitals, Karaikudi, India during a nearly 2-month period in 2015. Each instance of the dataset is composed of 11 numeric, 10 nominal, 3 ordinal features and 1 target feature (notckd/ckd).

2.2 Ensemble Trees machine learning techniques

Ensemble trees have become one of the most popular machine learning classifiers due to their stability and robustness when dealing with datasets of any size, as well as to a reasonably good predictive performance. Ensemble trees perform classification tasks by weighting various decision trees and combining them to reach a final model that improves each base model [31]. In addition, ensemble methods are used to mitigate challenges like class imbalance or the curse of dimensionality. Apart from decision trees, the classifiers used in this research are: random forest or extra trees [31] that follow the bagging technique where each base decision tree is trained using a sample with the same number of instances taken with replacement from the original dataset; and adaptive boosting [31] or extreme gradient boosting [32] that employ the technique of boosting focused on instances, in a sequential way, that have been previously misclassified when training a new base decision tree.

Table 2. Attributes description of ckd dataset

Features (units) [<i>legend</i>]	Type of feature (classes)	Average (std) / number of values
Age (year) [<i>age</i>]	Numerical	51.48 (17.17)
Blood pressure (mm/Hg) [<i>bp</i>]	Numerical	76.46 (13.68)
Specific gravity [<i>sg</i>]	Ordinal (1.005,1.010,1.015, 1.020, 1.025)	7, 84, 75, 106, 81
Albumin [<i>al</i>]	Ordinal (0,1,2,3,4,5)	199,44,43,43,24,1
Sugar [<i>su</i>]	Ordinal (0,1,2,3,4,5)	290,13,18,14,13,3
Red blood cells [<i>rbc</i>]	Nominal (normal/abnormal)	47 abnormal
Pus cell [<i>pc</i>]	Nominal (normal/abnormal)	76 abnormal
Pus cell clumps [<i>pcc</i>]	Nominal (not present/ present)	42 present
Bacteria [<i>ba</i>]	Nominal (not present/ present)	22 present
Blood glucose random (mgs/dl) [<i>bgr</i>]	Numerical	148.04 (79.28)
Blood urea (mgs/dl) [<i>bu</i>]	Numerical	57.43 (50.50)
Serum creatinine (mgs/dl) [<i>sc</i>]	Numerical	3.07 (5.74)
Sodium (mEq/l) [<i>sod</i>]	Numerical	137.53 (10.41)
Potassium (mEq/l) [<i>pot</i>]	Numerical	4.63 (3.19)
Hemoglobin (gms) [<i>hemo</i>]	Numerical	12.53 (2.91)
Packed cell volume [<i>pcv</i>]	Numerical	38.88 (8.99)
White blood cell count (cells/cumm) [<i>wc</i>]	Numerical	8406.12 (2944.47)
Red blood cell count (cells/cumm) [<i>rc</i>]	Numerical	4.71 (1.03)
Hypertension [<i>htn</i>]	Nominal (no/yes)	147 yes
Diabetes mellitus [<i>dm</i>]	Nominal (no/yes)	137 yes
Coronary artery disease [<i>cad</i>]	Nominal (no/yes)	34 yes
Appetite [<i>appet</i>]	Nominal (good/poor)	82 poor
Pedal edema [<i>pe</i>]	Nominal (no/yes)	76 yes
Anemia [<i>ane</i>]	Nominal (no/yes)	60 yes
Target class	notckd/ckd	250 ckd

2.3 Explainability techniques for ML

Concerning explainability, in domains (e.g healthcare) where predictions results must be interpretable, classifiers like decision trees are preferred due to their transparency. However, albeit ensemble trees usually present a better performance in classification tasks than transparent models, they require post-hoc explainability techniques to interpret their results because of their black-box behaviour. Post-hoc explainability techniques aim at providing understandable information about how an already developed model produces its predictions [33]. In this work, the following techniques have been used: *implicit feature importance*, which measures the mean decrease impurity, known as Gini, for each feature when obtaining the prediction; *feature permutation importance*, which quantifies the prediction error increase of the model after permuting a specific feature's values, being the most important features those that provoke an error increase [34]; *partial dependence plot*, that shows visually the marginal effect in terms of the probability that a given feature has on the predicted outcome over a

range of different observed values, while all other features are kept constant [35]; and *SHapley Additive exPlanations (SHAP)* that computes by applying coalitional game theory, an additive importance score for each feature in every individual prediction with local accuracy and consistency which are aggregated to give a global explainability of the model [36] [37].

Besides these explainability techniques, it's worth mentioning that feature selection procedures can be applied to remove unimportant features that bring non-relevant information to the classification, hence enhancing models' explainability [38]. This research has embraced feature selection by applying filter methods, where intrinsic properties of data, measured with ANOVA, Chi-squared or mutual information test, justify the inclusion of an attribute or a subset of attributes; or through wrappers methods like Recursive Feature Elimination (RFE) where a classification algorithm is utilized to select important features.

2.4 Classification performance and explainability metrics

Since the dataset employed presents an imbalance in its target feature (250 CKD/150 non-CKD) other metrics than accuracy, that measure the rate of true predictions over the total, are needed, namely: sensitivity (fraction of positive instances predicted correctly); specificity (fraction of negative instances predicted correctly); precision (fraction of true positive data given all true predicted data), and F1-score (harmonic mean from precision and sensitivity) [13]. The formulas of these metrics are shown in Table 3.

Moreover, considering ensemble trees as the classifiers employed, the explainability metrics proposed by Tagaris et.al [39] are used: *Interpretability*, defined as the percentage of those masked features that do not bring information to the final classification result and the total number of features of the dataset; *Fidelity*, that measures the accuracy relation of the equivalent full-interpretable model (i.e. decision trees) and its un-interpretable model counterpart; and *Fidelity-Interpretability Ratio (FIR)* that shows how much of the model's interpretability is sacrificed for performance, being 0.5 the optimal score.

Table 3. Metrics' formula (TN: true negative, FN: false negative, FP: false positive, TP: true positive).

Metrics	Formula
<i>Accuracy (Acc)</i>	$(TP+TN) / (TP+TN+FP+FN)$ (1)
<i>Sensitivity/Recall (Sen)</i>	$TP / (TP+FN)$ (2)
<i>Specificity (Spe)</i>	$TN / (FP+TN)$ (3)
<i>Precision (Pre)</i>	$TP / (TP+FP)$ (4)
<i>F1-Score (F1)</i>	$2*((Precision*Recall) / (Precision+Recall))$ (5)
<i>Interpretability (I)</i>	$I = \text{masked features} / \text{total input features}$ (6)
<i>Fidelity (F)</i>	$F = \text{Acc. Fully Interpretable} / \text{Acc. Un-Interpretable}$ (7)
<i>Fidelity-Interpretability Ratio (FIR)</i>	$FIR = F / (F+I)$ (8)

3 Results

3.1 Automated data workflow pipeline

In this work, the SCI-XAI automated data workflow pipeline published in [40] has been employed for developing the explainable CKD prediction model. Figure 1 shows the SCI-XAI pipeline that is based on the python scikit-learn package [41] and allows through a brute force algorithm finding the specific combination of ensemble trees classifier, the number of features selected, the feature selection and data missing imputation methods that output the best classification performance of the model. As a first step, the original dataset is split, with a target feature stratification, into training and test set with 280 and 120 instances respectively (ratio 70/30). This initial split allows building the model exclusively with training set's instances, and hence, evaluating its performance over unseen new data from the test set, when the above-mentioned preprocessing and classification discovered parameters are applied. Next, the data preparation phase takes place separately for numerical, nominal, and ordinal features to be merged in the modelling or training phase, where through a 5-fold cross-validation approach different ensemble trees classifiers are applied.

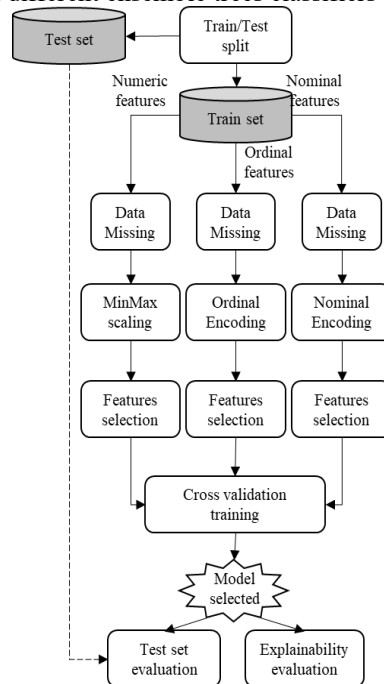


Figure 1. SCI-XAI automated data workflow pipeline.

3.2 Feature selection

The best combination of features selected found by SCI-XAI pipeline for each ensemble trees algorithms is shown in Table 5. The number of features, the techniques employed for selection (ANOVA, Chi-squared, mutual information or Recursive Feature Elimination), and the name of those selected features are detailed for numerical, nominal and ordinal features of the dataset. Thus, the feature selection step denotes at least 50% of the original features a non-relevant for the classification output. XGBoost is the algorithm with the biggest feature reduction in its best classification results, remaining with 3 out of 24 features.

3.3 Classification and explainability metrics results

Table shows the classification performance of the different ensemble trees algorithms after the training cross-validation module as well as the evaluation with the test set. The results show a solid classification performance in the training phase with a range of 98.1 to 100% in all metrics considered. This robust performance is maintained with unseen data from the test, obtaining in all classifiers considered an accuracy of more than 97.5%, and above 95% in the rest of the classification metrics.

Once the number of relevant features selected is known, the evaluation of explainability can be performed (shown in Table). For each ensemble trees algorithm and its group of selected features, a specific decision tree was implemented to obtain the equivalent fully interpretable model to calculate the Fidelity. As FIR gives a balanced measure between interpretability and fidelity with 0.5 as the optimal point, XGBoost (FIR=0.53) achieved the most balanced model. Therefore, the XGBoost and its group of selected features are used to conduct an explainability analysis of its predictions.

3.4 Explainability analysis of the prediction model

Since the CKD prediction model based on XGBoost is selected as the most balanced model in terms of explainability and accuracy, in this subsection, the relevance of the following features pcv (packed cell volume), htn (hypertension), sg (specific gravity) is analysed with different post-hoc explainability techniques to show their influence in model's outputs.

The implicit feature importance and feature permutation importance techniques allow visualizing the global explainability of each model's feature without informing about the direction of the contribution, i.e. increase or decrease of CKD probability. Figure 2.a shows the ranking of the features selected's importance based on mean impurity decrease (Gini) by applying feature implicit importance. In addition, Figure 2.b shows those features' importance, using the feature permutation importance technique, calculated as the decrease in accuracy when a specific feature's values are permuted. Both techniques denote pcv as the most relevant feature followed by sg and htn (in descending order of importance).

When employing PDP, information about the contribution that different features' values have to the probability when predicting a positive case of CKD can be extract-

ed. Figure 3 shows the PDP plots of the three selected features where the x-axis represents the values distribution, and the y-axis the contribution in probability to the prediction of CKD. Thus, for pcv values above 39 the contribution to predicting CKD decrease from 1 to 0.6, being monotonic at 0.6 for the rest of values except at 41 where a small peak can be appreciated. In addition, patients with hypertension (htn=1) have an increase of 0.4 (from 0.6 to 1.0) in the probability of suffering CKD. In the case of sg, for values of 1.020 and 1.025, the feature contributes to reducing the probability of predicting CKD in a 0.4 (from 1.0 to 0.6). It could be appreciated that approximately 0.6 represents a base value in the partial dependence plots which correspond with the proportion of total CKD positive (250) and negative (150) cases in the dataset.

Table 4: Feature selection results (#: number of features selected; Feats: name of features selected; mut-inf: mutual information, RFE: Recursive Feature Elimination)

Classifier	Numerical features		Nominal features		Ordinal features		Total of feats.
	#	Feats (selection method)	#	Feats (selection method)	#	Feats (selection method)	
<i>Decision Trees</i>	1	hemo (ANOVA)	3	htn, dm, appet (chi2)	1	sg (mut-inf)	5
<i>Random Forest</i>	1	hemo (ANOVA)	5	(RFE)	1	sg (mut-inf)	7
<i>Extra Trees</i>	4	pcv, hemo, rc, sc (mut-inf)	3	htn, dm, appet (chi2)	1	sg (mut-inf)	8
<i>AdaBoost</i>	7	pcv, hemo, rc, sc, pot, sod, bu (mut-inf)	4	htn, dm, appet, pe (mut-inf)	1	sg (mut-inf)	12
<i>XGBoost</i>	1	pcv (mut-inf)	1	htn (mut-inf)	1	sg (mut-inf)	3

Table 5. Classification metrics results (in %)

Classifier	Training set (Cross-val approach)					Test set (new unseen data)				
	Acc.	Sens.	Spec.	F1	Prec.	Acc.	Sens.	Spec.	F1	Prec.
<i>Decision Trees</i>	99.6	99.4	100	99.7	100	97.5	96	100	98	100
<i>Random Forest</i>	100	100	100	100	100	97.5	96	100	98	100
<i>Extra Trees</i>	100	100	100	100	100	98.3	97.3	100.0	98.6	100.0
<i>AdaBoost</i>	100	100	100	100	100	98.3	97.3	100.0	98.6	100.0
<i>XGBoost</i>	99.2	100	98.1	99.4	98.8	97.5	98.7	95.6	98	97.4

Table 6. Explainability metrics results

Classifier	Interpretability	Fidelity	FIR
<i>Decision Trees</i>	79 %	100 %	0.56
<i>Random Forest</i>	71 %	100 %	0.59
<i>Extra Trees</i>	67 %	99 %	0.60
<i>AdaBoost</i>	50 %	99 %	0.66
<i>XGBoost</i>	88 %	100 %	0.53

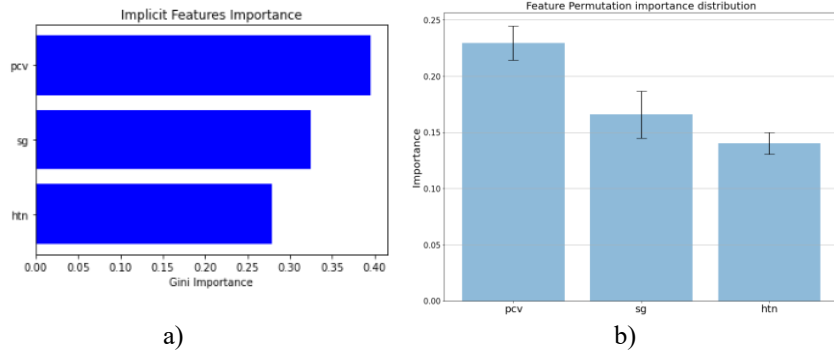


Figure 2: Global explainability of prediction model (a. Implicit feature importance, b. Feature permutation importance)

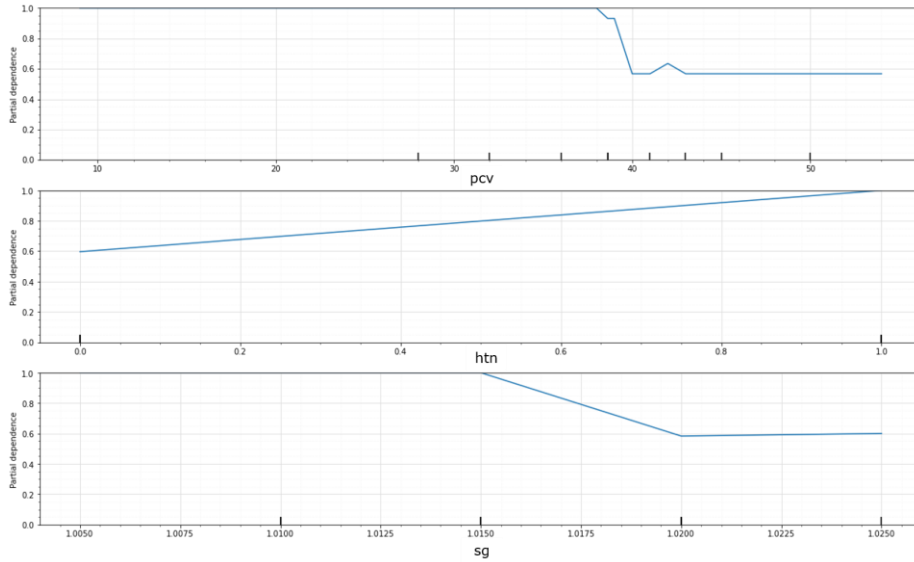


Figure 3: PDP plots for CKD probability contribution of each model's feature.

As an example of post-hoc explainability techniques when concerning predictions of individual cases, the SHAP technique depicts through its “waterfall plots”, the attribution of each feature value not only specifying the direction force towards the final score (red: positive contribution, blue: negative contribution) but also the feature's weight (length of the bar). Thus, Figure 4.a and Figure 4.b show the features' attributions when predicting respectively a true negative case ($y=0$, the patient does not have CKD) and a true positive case ($y=1$, the patient does have CKD). For both cases, the prediction starts from a base SHAP value (1.66) equal to the initial distribution of CKD/nonCKD cases (250/150). In the case of the true negative, pcv, with a value of 47, reveals as the most relevant feature in the prediction with an attribution

of -3.09, meanwhile sg and htn, with values 1.025 and 0 respectively, has negative attributions towards the non-CKD prediction (-1.92 and -1.44). As regards the true positive case, the values of pcv=35, sg=1.015 and htn=1 contribute to a positive prediction of CKD with a nearly similar additive attribution (+1.43, +1.38, +1.22 respectively). It is worth noting that the contributions shown for the features' values in these individual cases agree with the insights gained with the PDP plots.

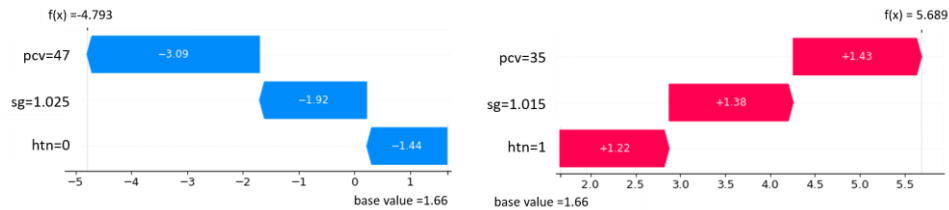


Figure 4. Local Explainability through SHAP (a. True Negative case; b. True Positive case)

4 Discussion

Due to the current increase of global incidence of CKD, the classification of those patients who might suffer the disease using prediction models could become a relevant tool for doctors to achieve an early diagnosis. In addition to that, XAI could imply an improvement to those prediction models by meeting the healthcare professionals' demands about understanding the decisions made by the models. Having more explainable CKD prediction models, doctors could make more data-driven decisions and focus on controlling those underlying features or indicators to slow the progressive damage of the kidney.

This paper describes a CKD prediction model developed to tackle the early diagnosis not only seeking high accuracy of the predictions but also analysing the explainability of its results. Thus, this research contributes to enlarging the works dedicated to CKD diagnosis through AI from a novelty perspective, to the best of our knowledge, that focus on the model's explainability. By using post-hoc explainability techniques, this work aims at "opening" the black-box paradigm of the ensemble trees classifiers employed to build the CKD prediction model.

The development of the explainable CKD prediction model is based on a data management pipeline, previously developed in other author's work, that automates the data preparation, modelling and evaluation phases. The pipeline allows inferring automatically different parameters like the appropriate ensemble tree algorithm, the relevant features selected, the feature selection method, and data imputation techniques to obtain the best classification performance of the prediction model. Moreover, the pipeline allows evaluating the model's performance over new unseen data (30% of the original dataset), which could emulate a real deployment, however, the model's performance in a real clinical environment might differ from the results because the medical records are not usually as curated as the dataset employed.

Concerning the classification performance, different ensemble tree algorithms have been considered (i.e. random forest, extra trees, AdaBoost, and XGBoost) along with different combinations of features selected to build the prediction model and consequently to conduct its evaluation. Considering the classification results achieved, this work achieves the state-of-art of CKD prediction models found in the literature. Thus, the application of the SCI-XAI pipeline shows a fairly good performance, especially when compared to other related works.

The pipeline's feature selection step has also proven to be valuable due to the substantial reduction of the original number of features, leaving 3 out of 24 when using the XGBoost classifier, which is the best CKD prediction model in the literature in terms of minimum features considered. Furthermore, 4 out of 5 ensemble learning algorithms considered obtaining their best classification results with only 33% of the original features showing the capability of the pipeline to detect relevant features when building the prediction model.

To the best of our knowledge, this paper is the first to entail an explainability analysis of a CKD prediction model. By considering the FIR value of 0.5 as the optimal point in terms of a trade-off between accuracy and explainability, the model that uses the XGBoost is selected as the most balanced. By doing so, this selected model does not achieve the best classification performance, showing an example of the tension occurring between accuracy and explainability when concerning prediction models aimed at being used in specific domains where understanding the results is crucial.

Regarding the analysis of the features' importance in the prediction model, the pcv (packed cell volume) feature seems to be the most relevant in all post-hoc analysis techniques considered, followed by the sg (specific gravity) and then htn (hypertension). It is worth highlighting the utility of the PDP plots to identify thresholds on which a certain feature modifies the probability prediction. For instance, this work establishes thresholds in 37 and 1.015 for pcv and sg respectively from the probability starts to decrease, implying that doctors could set up a treatment for the patient to be above these thresholds and reduce the probability of CKD disease. Moreover, the local explainability results exemplify how explainability techniques contribute to the promotion of precision or personalized medicine by showing the relevance of the different features for an individual prediction case.

With the results described in this work, the added value of explainability to a clinical prediction model is exhibited. Moreover, the feature selection approach is valuable not only for improving the explainability of clinical prediction models but also for reducing the cost of the diagnosis having fewer clinical indicators to extract. In particular, since this explainable CKD prediction model embraces the processing of 3 features (pcv, sg, and htn), the cost associated to extract them, by following the price list defined by Salekin et al [42], is 1.62 USD for pcv in a hematocrit test, free for sg, and free for htn. Therefore, the cost associated with an early diagnosis of CKD by using this explainable prediction model would be around 1.6 USD, which would have an important impact on developing countries where medical access is more difficult [43].

5 Conclusions

The development and evaluation of an explainable CKD prediction model have been presented in this work with the aim of showing the importance of considering XAI in early diagnosis systems in the healthcare field, maintaining a balance between the classification performance of the model and its explainability. The prediction model is built by using a data management pipeline that allows the detection of the most adequate ensemble tree algorithm as well as the number of features selected. In order to detect the best-balanced model in terms of accuracy and explainability, different evaluations are carried out by applying classification and explainability metrics. Therefore, the best explainable prediction model implements an XGBoost classifier over the following 3 features: packed cell values (pcv), specific gravity (sg) and hypertension (htn). After an explainability analysis by employing different techniques, the features' relevance in descendent order is: pcv, sg and htn. The prediction model developed equals the classification performance of the best CKD prediction models identified in the literature. In addition, the novelty presented by this work is the explainability approach adopted in the model's performance which revolves around giving healthcare professionals an easier understanding and interpretability of the outcomes generated by the model. Thus, not only would clinicians achieve an early diagnosis with a reduced group of indicators, but they could also focus on tackling relevant features to avoid the CKD onset or even to revert its progress. Future works would also be aimed at testing the prediction model developed in a clinical setting to test the accuracy robustness of the model with new patients' data as well as to gather insights of healthcare professionals about the explainability of the results provided.

6 References

1. Chen, Z., Zhang, X., & Zhang, Z. (2016). Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. *International Urology and Nephrology*, 48(12), 2069–2075. <https://doi.org/10.1007/s11255-016-1346-4>
2. Lysaght, M. J. (2002). Maintenance Dialysis Population Dynamics: Current Trends and Long-Term Implications. *Journal of the American Society of Nephrology*, 13(suppl 1), S37–S40. https://doi.org/10.1681/ASN.V13suppl_1s37
3. Gupta, R., Koli, N., Mahor, N., & Tejashri, N. (2020). Performance Analysis of Machine Learning Classifier for Predicting Chronic Kidney Disease. In *2020 International Conference for Emerging Technology (INCET)* (pp. 1–4). Presented at the 2020 International Conference for Emerging Technology (INCET). <https://doi.org/10.1109/INCET49848.2020.9154147>
4. Jeevantha, R. A., Halgamuge, M. N., Mohammad, A., & Ekici, G. (2017). Classification Performance Analysis in Medical Science: Using Kidney Disease Data. In *Proceedings of the 2017 International Conference on Big Data Research* (pp. 1–6). Osaka, Japan: Association for Computing Machinery. <https://doi.org/10.1145/3152723.3152724>
5. Keith, D. S., Nichols, G. A., Gullion, C. M., Brown, J. B., & Smith, D. H. (2004). Longitudinal follow-up and outcomes among a population with chronic kidney disease in a large managed care organization. *Archives of Internal Medicine*, 164(6), 659–663. <https://doi.org/10.1001/archinte.164.6.659>
6. Levin, A., Bakris, G. L., Molitch, M., Smulders, M., Tian, J., Williams, L. A., & Andress, D. L. (2007). Prevalence of abnormal serum vitamin D, PTH, calcium, and phosphorus in pa-

- tients with chronic kidney disease: results of the study to evaluate early kidney disease. *Kidney International*, 71(1), 31–38. <https://doi.org/10.1038/sj.ki.5002009>
7. Polat, H., Danaei Mehr, H., & Cetin, A. (2017). Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods. *Journal of Medical Systems*, 41(4), 55. <https://doi.org/10.1007/s10916-017-0703-x>
 8. Baby, P. S., & Vital, T. P. (2015). Statistical analysis and predicting kidney diseases using machine learning algorithms. *International Journal of Engineering Research and Technology*, 4(7).
 9. Lakshmi, K., Nagesh, Y., & Krishna, M. V. (2014). Performance comparison of three data mining techniques for predicting kidney dialysis survivability. *International Journal of Advances in Engineering & Technology*, 7(1), 242.
 10. Ray, A., & Chaudhuri, A. K. (2021). Smart healthcare disease diagnosis and patient management: Innovation, improvement and skill development. *Machine Learning with Applications*, 3, 100011. <https://doi.org/10.1016/j.mlwa.2020.100011>
 11. Dua, D., & Graff, C. (2017). *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences. Retrieved from <http://archive.ics.uci.edu/ml>
 12. Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5). <https://doi.org/10.1002/widm.1379>
 13. Zhang, H., Ren, J.-X., Ma, J.-X., & Ding, L. (2019). Development of an in silico prediction model for chemical-induced urinary tract toxicity by using naïve Bayes classifier. *Molecular Diversity*, 23(2), 381–392. <https://doi.org/10.1007/s11030-018-9882-8>
 14. Sossi Alaoui, S., Aksasse, B., & Farhaoui, Y. (2020). Data Mining and Machine Learning Approaches and Technologies for Diagnosing Diabetes in Women. In Y. Farhaoui (Ed.), *Big Data and Networks Technologies* (pp. 59–72). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-23672-4_6
 15. Zhang, Y., & Ma, Y. (2019). Application of supervised machine learning algorithms in the classification of sagittal gait patterns of cerebral palsy children with spastic diplegia. *Computers in Biology and Medicine*, 106, 33–39. <https://doi.org/10.1016/j.compbiomed.2019.01.009>
 16. Feeny, A. K., Rickard, J., Patel, D., Toro, S., Trulock, K. M., Park, C. J., ... Chung, M. K. (2019). Machine Learning Prediction of Response to Cardiac Resynchronization Therapy: Improvement Versus Current Guidelines. *Circulation. Arrhythmia and Electrophysiology*, 12(7), e007316. <https://doi.org/10.1161/CIRCEP.119.007316>
 17. Aro, T. O., Akande, H. B., Jibrin, M. B., & Jauro, U. A. (2019). Homogenous Ensembles on Data Mining Techniques for Breast Cancer Diagnosis. *Daffodil international university journal of science and technology*, 14(1).
 18. Seker, H., Odetayo, M. O., Petrovic, D., Naguib, R., & Hamdy, F. (2000). A Soft Measurement Technique for Searching Significant Subsets of Prostate Cancer Prognostic Markers. In P. Sinčák, J. Vaščák, V. Kvasnička, & R. Mesiar (Eds.), *The State of the Art in Computational Intelligence* (pp. 325–328). Heidelberg: Physica-Verlag HD. https://doi.org/10.1007/978-3-7908-1844-4_52
 19. Karun, S., Raj, A., & Attigeri, G. (2019). Comparative Analysis of Prediction Algorithms for Diabetes. In S. K. Bhatia, S. Tiwari, K. K. Mishra, & M. C. Trivedi (Eds.), *Advances in Computer Communication and Computational Sciences* (pp. 177–187). Singapore: Springer. https://doi.org/10.1007/978-981-13-0341-8_16
 20. Bucholc, M., Ding, X., Wang, H., Glass, D. H., Wang, H., Prasad, G., ... Wong-Lin, K. (2019). A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. *Expert Systems with Applications*, 130, 157–171. <https://doi.org/10.1016/j.eswa.2019.04.022>

21. Ekanayake, I. U., & Herath, D. (2020). Chronic Kidney Disease Prediction Using Machine Learning Methods. In *2020 Moratuwa Engineering Research Conference (MERCon)* (pp. 260–265). Presented at the 2020 Moratuwa Engineering Research Conference (MERCon). <https://doi.org/10.1109/MERCon50084.2020.9185249>
22. Sossi Alaoui, S., Aksasse, B., & Farhaoui, Y. (2019). Statistical and Predictive Analytics of Chronic Kidney Disease. In M. Ezziyyani (Ed.), *Advanced Intelligent Systems for Sustainable Development (AI2SD'2018)* (pp. 27–38). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-11884-6_3
23. Ogunleye, A., & Wang, Q.-G. (2020). XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(6), 2131–2140. Presented at the IEEE/ACM Transactions on Computational Biology and Bioinformatics. <https://doi.org/10.1109/TCBB.2019.2911071>
24. Zeynu, S. (2018). Prediction of Chronic Kidney Disease Using Data Mining Feature Selection and Ensemble Method. *WSEAS TRANSACTIONS ON INFORMATION SCIENCE AND APPLICATIONS*.
25. Raju, N. V. G., Lakshmi, K. P., Praharshitha, K. G., & Likhitha, C. (2019). Prediction of chronic kidney disease (CKD) using Data Science. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* (pp. 642–647). Presented at the 2019 International Conference on Intelligent Computing and Control Systems (ICCS). <https://doi.org/10.1109/ICCS45141.2019.9065309>
26. Khan, B., Naseem, R., Muhammad, F., Abbas, G., & Kim, S. (2020). An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy. *IEEE Access*, 8, 55012–55022. Presented at the IEEE Access. <https://doi.org/10.1109/ACCESS.2020.2981689>
27. Zubair Hasan, K. M., & Zahid Hasan, Md. (2019). Performance Evaluation of Ensemble-Based Machine Learning Techniques for Prediction of Chronic Kidney Disease. In N. R. Shetty, L. M. Patnaik, H. C. Nagaraj, P. N. Hamsavath, & N. Nalini (Eds.), *Emerging Research in Computing, Information, Communication and Applications* (pp. 415–426). Singapore: Springer. https://doi.org/10.1007/978-981-13-5953-8_34
28. Abdullah, A. A., Hafidz, S. A., & Khairunizam, W. (2020). Performance Comparison of Machine Learning Algorithms for Classification of Chronic Kidney Disease (CKD). *Journal of Physics: Conference Series*, 1529, 052077. <https://doi.org/10.1088/1742-6596/1529/5/052077>
29. Alaiad, A., Najadat, H., Mohsen, B., & Balhaf, K. (2020). Classification and Association Rule Mining Technique for Predicting Chronic Kidney Disease. *Journal of Information & Knowledge Management*, 19(01), 2040015. <https://doi.org/10.1142/S0219649220400158>
30. Kadhum, M., Manaseer, S., & Dalhoum, A. L. A. (2021). Evaluation Feature Selection Technique on Classification by Using Evolutionary ELM Wrapper Method with Features Priorities. *Journal of Advances in Information Technology*, 12(1), 21–28. <https://doi.org/10.12720/jait.12.1.21-28>
31. Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4), e1249. <https://doi.org/10.1002/widm.1249>
32. Sagi, O., & Rokach, L. (2020). Explainable decision forest: Transforming a decision forest into an interpretable tree. *Information Fusion*, 61, 124–138. <https://doi.org/10.1016/j.inffus.2020.03.013>
33. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
34. Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *arXiv:1801.01489 [stat]*. Retrieved from <http://arxiv.org/abs/1801.01489>

35. Friedman, J. H. (2000). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29, 1189–1232.
36. Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs, stat]*. Retrieved from <http://arxiv.org/abs/1705.07874>
37. Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., ... Lee, S.-I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10), 749–760. <https://doi.org/10.1038/s41551-018-0304-0>
38. Shilaskar, S., & Ghatol, A. (2013). Feature selection for medical diagnosis : Evaluation for cardiovascular diseases. *Expert Systems with Applications*, 40(10), 4146–4153. <https://doi.org/10.1016/j.eswa.2013.01.032>
39. Tagaris, T., & Stafylopatis, A. (2020). Hide-and-Seek: A Template for Explainable AI. *arXiv:2005.00130 [cs, stat]*. Retrieved from <http://arxiv.org/abs/2005.00130>
40. Moreno-Sanchez, P. A. (2021). An automated feature selection and classification pipeline to improve explainability of clinical prediction models. In *2021 IEEE 9th International Conference on Healthcare Informatics* (pp. 527–534). Presented at the 1st Workshop on Machine Learning in Healthcare Data for Precision Medicine.
41. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Cournapeau, D. (n.d.). Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*, 6.
42. Salekin, A., & Stankovic, J. (2016). Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 262–270). Presented at the 2016 IEEE International Conference on Healthcare Informatics (ICHI). <https://doi.org/10.1109/ICHI.2016.36>
43. Sobrinho, A., Queiroz, A. C. M. D. S., Dias Da Silva, L., De Barros Costa, E., Eliete Pinheiro, M., & Perkusich, A. (2020). Computer-Aided Diagnosis of Chronic Kidney Disease in Developing Countries: A Comparative Analysis of Machine Learning Techniques. *IEEE Access*, 8, 25407–25419. Presented at the IEEE Access. <https://doi.org/10.1109/ACCESS.2020.2971208>

Glossary of Terms

<ul style="list-style-type: none"> • DT: Decision Trees • RF: Random Forest • XGB: eXtreme Gradient Boosting • Ada: Adaptive Boosting • ET: Extra Trees • XGB lin: XGB linear • Lin SVM: Linear Support Vector Machine • KNN: K-Nearest Neighbors • ANN: Artificial Neural Network • NB: Gaussian Naïve Bayes • LR: Logistic Regression • GB: Gradient Boosting • Jrip: Jrip associated rule • ELM: Extreme Marchine Learning • Acc: Accuracy • Sen: Sensivity 	<ul style="list-style-type: none"> • Spe: Specificity • F1: f1-score • Pre: Precision pcv: Packed cell volume hemo: Hemoglobin (gms) rc: Red blood cell count (cells/ cumm) sc: Serum creatinine (mgs/dl) pot: Potassium (mEq/l) sod: Sodium (mEq/l) bu: Blood urea (mgs/dl) htn: Hypertension dm: Diabetes mellitus appet: Appetite rbc: Red blood cells pc: Pus cell pe: Pedal edema sg: Specific gravity
--	--

Author's Biography



Pedro A. Moreno-Sanchez currently works as Researcher at Seinäjoki University of Applied Sciences, Finland. He received the Ph. D. on 2014 by the Polytechnical University of Madrid, Spain. His main research interest are eXplainable Artificial Intelligence applied to healthcare.