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

Pedro A. Moreno-Sanchez (pedro.morenosanchez@tuni.fi)
Tampere University

Research Article

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

Posted Date: May 16th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1628347/v1

Abstract

Chronic Kidney Disease (CKD), which implies premature mortality if diagnosed late, is currently experiencing a globally increasing incidence and high cost to health systems. Artificial Intelligence (AI) allows discovering subtle patterns in CKD indicators to contribute to an early diagnosis. In addition, eXplainable AI (XAI) meets the clinicians' requirement of understanding AI models' output when patients' life is eventually affected by the AI algorithms' decision. 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 optimal combination, in terms of classification performance, of ensemble trees algorithms with features selected. The main contribution of the paper involves an explainability-driven approach that allows selecting the best prediction model maintaining a balance between accuracy and explainability that provide quantitative information about the effect of the features selected on the probability of having CKD. Therefore, the most balanced explainable prediction model implements an extreme gradient boosting classifier using 3 features (hemoglobin, specific gravity, and hypertension) that achieves an accuracy of 99.2% and 97.5% with a 5-fold cross-validation and with new unseen data respectively. In addition, an analysis of the model´s explainability shows that hemoglobin is the most relevant feature that influences the prediction results of the model, followed by specific gravity and hypertension. This small number of features selected results in a reduced cost of the early diagnosis of CKD implying a promising solution for developing countries.

Introduction

Chronic kidney disease (CKD) has become 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 [2][3]. Typically, CKD has no early symptoms, and when detected the kidney has already lost 25 percent of its capacity and is under progressive damage that, if not slowed by controlling underlying risk factors (hypertension, obesity, heart disease, age) [4], the hemodialysis or even kidney transplantation are crucial for patient survival. [5]–[7]. Consequently, early diagnosis of CKD based on risk factors allows initiating treatments that slow the progression of kidney damage and prolong patients' life.

In the medical domain, Artificial intelligence (AI) has become a promising instrument to build computer-aided diagnosis (CAD) [8], [9]. Thus, AI can be employed to discover latent correlations between CKD and its indicators enabling an early discovery of patients at risk. Related works show the use of a public common CKD dataset from the University of California Irvine-Machine Learning (UCI-ML) [10] to build their prediction models allowing the reproducibility of results as well as benchmarking between other models' implementation. Table 1 shows the most recent and accurate works (accuracy above 98%) that employed the CKD dataset of the UCI-ML repository and applied feature selection in the model development.

When CAD systems' decisions affect patients' life eventually, explanations about the AI models' outputs are essential to support clinicians in their diagnosis and treatments. Thus, 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 [11]. Global model-specific solutions of XAI have been developed for the last decade in different clinical fields, namely: urology [12], toxicology [12], endocrinology [13], neurology [14], cardiology [15], cancer (e.g. breast cancer or prostate cancer) [16], [17], and chronic diseases (e.g. diabetes or Alzheimer's disease) [18], [19]. However, an inherent trade-off arises between predictive accuracy, which provides the reliability of the model, and the explainability requested by clinical experts. This tension must be addressed when developing the AI model because most accurate models are usually less transparent and vice versa. Concerning the application of XAI approaches to CKD prediction models, to the best of our knowledge, any XAI analysis further than applying feature selection has not been found in the literature.

This paper aims at describing the development and assessment of an explainable prediction model of CKD 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 to find an appropriate balance between accuracy and explainability. The remainder of the article adopts the following structure: (1) description of the dataset, the machine learning algorithms, evaluation metrics, and explainability techniques employed in this research; (2) presentation of the pipeline employed, the evaluation results in terms of classification and explainability, and the explainability analysis; (3) the discussion of the results as well as obtained conclusions.

Material And Methods

Chronic kidney disease dataset

To promote the reproducibility of this research, the UCI-ML dataset was employed. Table 2 describes the dataset collected from the Apollo Hospital, Karaikudi, India during a nearly 2-month period in 2015 that includes 400 patients where some presented missing values in their features. Each instance of the dataset is composed of 11 numeric, 10 nominal, 3 ordinal features, and 1 target feature (notCKD/CKD).

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

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

Article	Acc	Sen	Spe	F1	Pre	#F	Machine Learning Classifier
Ekanayake[20]	100	100	-	100	100	7	*DT, RF*, XGB, Ada, ET (*)
Alaoui [21]	100	-	-	-	-	23	XGB Lin, Lin SVM, DT, RF
Ogunleye [22]	100	100	100	-	100	12	XGB (*)
Zeynu [23]	99.5	99.5	-	99.5	99.5	8	KNN, DT, ANN, NB, SVM.
Raju [24]	99.3	99	-	99	100	5	XGB, RF, LR, SVM, NB(*)
Khan [25]	99.1	99.7	-	99.3	98.7	23	NB, LR, SVM, DT, RF
Hasan [26]	99	-	-	99	-	13	Ada, RF, GB, ET(*)
Abdullah [27]	98.8	98.0	100	98.8	98.0	10	RF, SVM, NB, LR
Alaiad [28]	98.5	99.6	96.8	-	98	12	NB, DT, SVM, KNN, Jrip
Kadhum [29]	98.1	98	-	98	98	10	SVM, ELM

classification tasks by weighting various decision trees and combining them to reach a final model that improves each base model [30]. In addition, ensemble methods are used to mitigate challenges like class imbalance or the curse of dimensionality. The classifiers used in this research are: random forest and extra trees [30] 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; as well as adaptative boosting [30] and extreme gradient boosting [31] that apply the technique of boosting focused on instances, in a

Table 2
Attributes description of CKD dataset (*Num*: numerical, *Ord*: ordinal, *Nom*: nominal)

Features (units) [*legend*]	**Type of feature (% of non-null values) [classes in ordinal or nominal features]**	**Average (std) for numerical features / number of values for ordinal or nominal features**
Age (year) [*age*]	Num (97,75%)	51.48 (17.17)
Blood pressure (mm/Hg) [*bp*]	Num (97%)	76.46 (13.68)
Specific gravity [*sg*]	Ord (88,25%) [1.005,1.010,1.015, 1.020, 1.025]	7, 84, 75, 106, 81
Albumin [*al*]	Ord (88,5%) [0,1,2,3,4,5]	199,44,43,43,24,1
Sugar [*su*]	Ord (87,75%) [0,1,2,3,4,5]	290,13,18,14,13,3
Red blood cells [*rbc*]	Nom (62%) [normal/abnormal]	47 abnormal
Pus cell [*pc*]	Nom (83,75%) [normal/abnormal]	76 abnormal
Pus cell clumps [*pcc*]	Nom (99%) [not present/ present]	42 present
Bacteria [*ba*]	Nom (99%) [not present/ present]	22 present
Blood glucose random (mgs/dl) [*bgr*]	Num (89%)	148.04 (79.28)
Blood urea (mgs/dl) [*bu*]	Num (95,25%)	57.43 (50.50)
Serum creatinine (mgs/dl) [*sc*]	Num (95,75%)	3.07 (5.74)
Sodium (mEq/l) [*sod*]	Num (78,25%)	137.53 (10.41)
Potassium (mEq/l) [*pot*]	Num (78%)	4.63 (3.19)
Hemoglobin (gms) [*hemo*]	Num (87%)	12.53 (2.91)
Packed cell volume [*pcv*]	Num (82,50%)	38.88 (8.99)
White blood cell count (cells/cumm) [*wc*]	Num (73,75%)	8406.12 (2944.47)
Red blood cell count (cells/ cumm) [*rc*]	Num (67,5%)	4.71 (1.03)

Features (units) [*legend*]	Type of feature (% of non-null values) [classes in ordinal or nominal features]	Average (std) for numerical features / number of values for ordinal or nominal features
Hypertension [*htn*]	Nom (99,5%) [no/yes]	147 yes
Diabetes mellitus [*dm*]	Nom (99,5%) [no/yes]	137 yes
Coronary artery disease [*cad*]	Nom (99,5%) [no/yes]	34 yes
Appetite [*appet*]	Nom (99,75%) [good/poor]	82 poor
Pedal edema [*pe*]	Nom (99,75%) [no/yes]	76 yes
Anemia [*ane*]	Nom (99,75%) [no/yes]	60 yes
Target class	notCKD/CKD	250 CKD

sequential way, that have been previously misclassified when training a new base decision tree.

Explainability techniques for machine learning

In domains (e.g healthcare) where predictions results must be interpretable, maintaining the predictive performance of ensemble trees, known by a black-box behavior, balanced with explainable capabilities is crucial. Therefore, post-hoc XAI techniques aimed at providing understandable information about how an already developed model produces its predictions [32] are required. In this work, the following explainable post-hoc techniques have been used: *permutation feature importance (PFI)*, 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 [33]; *partial dependence plot (PDP)*, 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[34]; and *SHapley Additive exPlanations (SHAP)* that computes by applying coalitional game theory, an additive importance score for each feature in every individual prediction, known as shapley value, with local accuracy and consistency which are then aggregated to give a global explainability of the model [35] [36].

Besides these model-agnostic explainability techniques, it's worth mentioning that feature selection procedures can remove unimportant features with non-relevant information to the classification, hence enhancing models' explainability [37]. This research addresses feature selection by applying filter methods, where intrinsic properties of data measured with ANOVA, Chi-squared or mutual information test justify the inclusion of a subset of features; or by wrappers methods, like Recursive Feature Elimination (RFE), where a classification algorithm (eg. Logistic regression) is utilized to select important features [38].

Classification performance and explainability evaluation metrics

Since the dataset employed presents an imbalance in its target feature (250 CKD/150 notCKD), other metrics than accuracy are needed: sensitivity, specificity, precision, and F1-score [12]. Considering ensemble trees as the classifiers employed, the explainability metrics proposed by Tagaris et.al [39] are used: *Interpretability*, defined as the ratio of those masked features that do not bring information to the final classification result and the total number of features of the dataset; *Fidelity*, measures the accuracy relation between the evaluated model and its equivalent full-interpretable model that is built with a decision tree on the same data input; and *Fidelity-Interpretability Index (FII)* that allow comparing explainability performance between different models. The formulas of these metrics are shown in Table 3.

Table 3
Metrics of classification performance and explainability evaluation

Metric	Equation
Accuracy (Acc)	$Acc = \frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity/Recall (Sen)	$Sen = \frac{TP}{TP+FN}$
Specificity (Spe)	$Spe = \frac{TN}{TN+FP}$
Precision (Pre)	$Pre = \frac{TP}{TP+FP}$
F1-Score (F1)	$F1 = 2 * \frac{Pre * Sen}{Pre+Sen}$
Interpretability (I)	$I = \frac{Maskedfeatures}{Totalfeatures}$
Fidelity (F)	$F = \frac{Acc_{equivalentinterpretablemodel}}{Acc_{originalmodel}}$
Fidelity-Interpretability Index (FII)	$FII = F * I$

Results

Automated pipeline for best-model selection

In this work, the automated pipeline named SCI-XAI (feature Selection and Classification for Improving eXplainable AI) and published in [40] has been employed for developing the explainable CKD prediction model[2]. SCI-XAI pipeline, shown in Fig. 1, is implemented with python scikit-learn package [41] and allows through a brute force algorithm finding the optimal combination of ensemble trees classifier, the number of features selected, the feature selection, and data missing imputation methods. As a first step, a stratified split of the dataset allocates 280 and 120 instances respectively into training and test set (ratio 70/30). Next, the data preparation phase entails data missing, scaling/encoding, and feature

selection for numerical, nominal, and ordinal features separately, to be merged in a 5-fold cross-validation training phase. The initial split with target feature stratification allows evaluating the model's performance over unseen new data from the test set by applying the specific parameters selected by the pipeline in the preprocessing and training phase. Finally, the best model selected is also assessed in terms of explainability with the interpretability, fidelity, and FII metrics.

Feature selection

Table 4 shows the best combination of features selected obtained by the SCI-XAI pipeline for each ensemble trees algorithm. Thus, the feature selection step denotes that at least 50% of the original features are non-relevant (worst case Adaboost with 12 out of 24), being XGBoost the algorithm with the biggest features reduction, leaving 3 out of 24 features, at its best classification results.

Classification and explainability metrics results

Table 5 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 continues with unseen data, obtaining in all classifiers considered an accuracy result of more than 97.5%, and above 95% in the rest of the classification metrics.

The evaluation of explainability is shown in Table 6 considering the relevant features selected for each classifier. As FII gives a balanced measure between interpretability and fidelity to be used to compare different algorithms, XGBoost (FII = 0.85) achieved the most balanced model. Therefore, the XGBoost and its group of selected features: hemoglobin (hemo), hypertension(htn), and specific gravity (sg) are used to conduct an explainability analysis of its predictions.

Explainability analysis of the prediction model

Being XGBoost the most balanced model in terms of explainability and accuracy, the relevance of hemo, htn, and sg features is analyzed with different post-hoc explainability techniques to show their influence on the model's outputs.

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
	#	Feats [sel method]	#	Feats [sel method]	#	Feats [sel method]	
Random Forest	1	hemo [ANOVA]	5	htn, dm, appet, rbc, pc [RFE]	1	sg [mut-inf]	7
Extra Trees	4	hemo, pcv, rc, sc [mut-inf]	3	htn, dm, appet [chi2]	1	sg [mut-inf]	8
AdaBoost	7	hemo, pcv, rc, sc, sod, pot, wc [mut-inf]	4	htn, dm, appet, pe [mut-inf]	1	sg [mut-inf]	12
XGBoost	1	hemo [mut-inf]	1	htn [mut-inf]	1	sg [mut-inf]	3

Table 5
Classification metrics results (in %). Cross-validation training results expressed with mean (standard deviation).

Classifier	Training set with Cross-Val					Test set (new unseen data)				
	Acc.	Sens.	Spec.	F1	Prec.	Acc.	Sens.	Spec.	F1	Prec.
Random Forest	100 (0,0)	100 (0,0)	100 (0,0)	100 (0,0)	100 (0,0)	97.5	96	100	98	100
Extra Trees	100 (0,0)	100 (0,0)	100 (0,0)	100 (0,0)	100 (0,0)	98.3	97.3	100.0	98.6	100.0
AdaBoost	100 (0,0)	100 (0,0)	100 (0,0)	100 (0,0)	100 (0,0)	98.3	97.3	100.0	98.6	100.0
XGBoost	99.2 (0,8)	100 (0,0)	98.1 (2,3)	99.4 (0,6)	98.8 (1,3)	97.5	98.7	95.6	98	97.4

Table 6
Explainability metrics results

Classifier	Interpretability	Fidelity	FII
Random Forest	71%	100%	0,71
Extra Trees	67%	99%	0,66
AdaBoost	50%	99%	0,50
XGBoost	88%	97%	0,85

Figure 2 shows those features' importance obtained with PFI that allows visualizing the global explainability of each feature without informing about the direction of the contribution, i.e. if they

increase or decrease the probability of having CKD. Thus, hemo is denoted as the most relevant feature followed by sg and htn (in descending order of importance).

PDP plots, as shown in Fig. 3, offer information about the marginal effect of the features selected's values (x-axis) on the probability of a positive CKD prediction (y-axis). Thus, for hemo values between 12.3 gms and 13.5 gms the contribution to predicting CKD drops from 0.98 to 0.53 in several steps decrease with values of 0.94, 0.57, 0.55, and 0.53; being monotonic at the latter for the rest of hemo values above 13.5 gms. In addition, patients with hypertension (htn = 1) have an increase of 0.33 (from 0.6 to 0.93) in the probability of suffering CKD. In the case of sg, values of 1.020 and 1.025 reduce the probability of predicting CKD by a 0.4 (from 0.98 to 0.58).

SHAP technique also allows explaining the general contribution to the model's probability of every feature concerning its values. Similar to trends shown when using PDP, Fig. 4 depicts that hemo feature has the greatest attribution to the CKD probability, decreasing it at high hemo values (red/magenta color) and vice versa. Similarly, high values of sg feature contribute by reducing the probability of CKD. In addition, the presence of hypertension (htn equals 1) increases the CKD probability (red color). Besides, SHAP offers explanations when concerning predictions of individual cases (shown in Fig. 5), by depicting the attribution of each feature value not only specifying the direction force towards the final shapley value (red: positive contribution, blue: negative contribution) but also the feature's weight (length of the bar). Thus, Fig. 5.a and Fig. 5.b show a true negative case (y = 0, the patient does not have CKD) and a true positive case (y = 1, the patient does have CKD), where both cases the prediction starts from a base of 1.58 which means the average shapley value of the model output over the training set. In the case of the true negative with a final shapley value of -4.87, hemo equals 17 gms is denoted as the most relevant feature in the prediction with a shapley value attribution of -3.2, meanwhile, sg and hth, with values 1.025 and 0 respectively, have negative shapley values attributions (-1.92 and − 1.44). As regards the true positive case (shapley value equals 5.76), the values of hemo = 11.4, sg = 1.015, and htn = 1 contribute to a positive prediction of CKD with a nearly similar additive shapley values (+ 1.7 + 1.77, + 1.27 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.

Footnote:

[2] Source code of SCI-XAI is available at: https://github.com/petmoreno/SCI-XAI-Pipeline

Discussion

Due to the current increase in the global incidence of CKD, the classification of patients at risk becomes a relevant tool for doctors to achieve a disease 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 to the kidney.

This paper describes a CKD prediction model developed to tackle the early diagnosis not only seeking high accuracy but also analyzing the explainability of its results. Thus, this research contributes to enlarging the works dedicated to AI for CKD diagnosis from a novel perspective to the best of our knowledge, that focuses on the model's explainability. By using post-hoc explainability techniques, this work aims to "open" the black-box paradigm of the ensemble trees classifiers when predicting CKD.

The development of the explainable CKD prediction model is based on a data management pipeline that 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 deployment in a real clinical environment, however, the model's performance might differ since actual medical records are not usually as curated as the dataset employed.

Considering our classification results, this work obtains a fairly good performance by achieving the state-of-art of CKD prediction models found in the literature, especially when comparing the number of features selected. Therefore, the SCI-XAI pipeline´s feature selection step has proven to be valuable by reducing substantially the original number of features, leaving 3 out of 24 when using the XGBoost classifier, being the best CKD prediction model in the literature in terms of minimum features considered. Furthermore, 3 out of 4 considered ensemble learning algorithms obtain 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 in the literature to address an explainability analysis of a CKD prediction model selected through an accuracy-explainability trade-off perspective. Thus, albeit not obtaining the best classification performance, XGBoost is selected as the most balanced model, showing an example of the tension between accuracy and explainability dealt by prediction models aimed at being used in specific domains where understanding the results is crucial (e.g., healthcare).

Regarding the analysis of the features' importance in the prediction model, the hemo (hemoglobin) feature is denoted as 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 12.3 gms and 1.015 for hemo and sg respectively where the probability starts to decrease, implying that doctors could set up a treatment for the patient to be above these values and reduce the probability of CKD disease. Moreover, the local explainability results exemplify how XAI could contribute to the promotion of 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. Besides, 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. Thus, since this explainable CKD prediction model implies the processing of 3 features (hemo, sg, and htn), the cost associated to extract them, by following the price list defined by Salekin et al [42], is 1.65 USD for hemo in a hemoglobin test, and no cost for specific gravity (sg) and hypertension (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].

Our research presents some limitations. First, the present study employs a widely used CKD dataset from a UCI-ML repository that although allows benchmarking with other related research works, impedes performing an objective experiment. Since the number of patients is relatively small, a K-fold cross-validation approach has been adopted to foster generalization ability. However, to conclusively validate the results, more CKD data would be needed from a different clinical setting from the original which is planned as future works.

Conclusions

The development and evaluation of an explainable CKD prediction model have been presented in this work with the aim of showing the contribution of XAI and its accuracy and explainability trade-off in the medical field. Our prediction model is built by using an automated model selection pipeline that allows the optimal selection of the ensemble tree algorithm as well as the number of features selected through classification and explainability metrics. Therefore, the best explainable prediction model is an XGBoost classifier over the following 3 features: hemoglobin (hemo), specific gravity (sg), and hypertension (htn). After an explainability analysis by employing different post-hoc techniques, the features' relevance in descendent order is: hemo, sg and htn. The prediction model developed achieves 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 development that would provide healthcare professionals with an easier understanding and interpretability of the output generated. Thus, not only would clinicians achieve an early diagnosis with a reduced group of indicators, but they could also focus on tackling relevant features and its values to avoid the CKD onset or even to revert its progress. For the sake of trustworthiness and transparency of the model, future works would entail the prospective validation of the prediction model developed in a clinical setting to test its classification robustness with new patients' data, gather insights from healthcare professionals about the explainability of the results, and optimize CKD treatment plans.

Declarations

Author contributions

PAMS has carried out all relevant authorship roles concerning Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Supervision; Validation; Visualization; Writing - original draft; and Writing - review & editing

Funding

No funding was received to assist with the preparation of this manuscript.

Ethics approval and consent to participate This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Z. Chen, X. Zhang, and Z. Zhang, “Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models,” Int Urol Nephrol, vol. 48, no. 12, pp. 2069–2075, Dec. 2016, doi: 10.1007/s11255-016-1346-4.
2. M. J. Lysaght, “Maintenance Dialysis Population Dynamics: Current Trends and Long-Term Implications,” JASN, vol. 13, no. suppl 1, pp. S37–S40, Jan. 2002, doi: 10.1681/ASN.V13suppl_1s37.
3. R. Gupta, N. Koli, N. Mahor, and N. Tejashri, “Performance Analysis of Machine Learning Classifier for Predicting Chronic Kidney Disease,” in 2020 *International Conference for Emerging Technology (INCET)*, Jun. 2020, pp. 1–4. doi: 10.1109/INCET49848.2020.9154147.
4. R. A. Jeewantha, M. N. Halgamuge, A. Mohammad, and G. Ekici, “Classification Performance Analysis in Medical Science: Using Kidney Disease Data,” in *Proceedings of the* 2017 *International Conference on Big Data Research*, Osaka, Japan, Oct. 2017, pp. 1–6. doi: 10.1145/3152723.3152724.
5. D. S. Keith, G. A. Nichols, C. M. Gullion, J. B. Brown, and D. H. Smith, “Longitudinal follow-up and outcomes among a population with chronic kidney disease in a large managed care organization,” *Arch. Intern. Med.*, vol. 164, no. 6, pp. 659–663, Mar. 2004, doi: 10.1001/archinte.164.6.659.
6. A. Levin *et al.*, “Prevalence of abnormal serum vitamin D, PTH, calcium, and phosphorus in patients with chronic kidney disease: results of the study to evaluate early kidney disease,” *Kidney Int.*, vol. 71, no. 1, pp. 31–38, Jan. 2007, doi: 10.1038/sj.ki.5002009.
7. H. Polat, H. Danaei Mehr, and A. Cetin, “Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods,” J Med Syst, vol. 41, no. 4, p. 55, Feb. 2017, doi: 10.1007/s10916-017-0703-x.
8. P. S. Baby and T. P. Vital, “Statistical analysis and predicting kidney diseases using machine learning algorithms,” International Journal of Engineering Research and Technology, vol. 4, no. 7, 2015.
9. K. Lakshmi, Y. Nagesh, and M. V. Krishna, “Performance comparison of three data mining techniques for predicting kidney dialysis survivability,” International Journal of Advances in Engineering & Technology, vol. 7, no. 1, p. 242, 2014.
10. D. Dua and C. Graff, *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences, 2017. [Online]. Available: http://archive.ics.uci.edu/ml

11. G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, “Interpretability of machine learning based prediction models in healthcare,” *WIREs Data Mining Knowl Discov*, vol. 10, no. 5, Sep. 2020, doi: 10.1002/widm.1379.
12. H. Zhang, J.-X. Ren, J.-X. Ma, and L. Ding, “Development of an in silico prediction model for chemical-induced urinary tract toxicity by using naïve Bayes classifier,” Molecular Diversity, vol. 23, no. 2, pp. 381–392, May 2019, doi: 10.1007/s11030-018-9882-8.
13. S. Sossi Alaoui, B. Aksasse, and Y. Farhaoui, “Data Mining and Machine Learning Approaches and Technologies for Diagnosing Diabetes in Women,” in *Big Data and Networks Technologies*, Cham, 2020, pp. 59–72. doi: 10.1007/978-3-030-23672-4_6.
14. Y. Zhang and Y. Ma, “Application of supervised machine learning algorithms in the classification of sagittal gait patterns of cerebral palsy children with spastic diplegia,” Comput. Biol. Med., vol. 106, pp. 33–39, 2019, doi: 10.1016/j.compbiomed.2019.01.009.
15. A. K. Feeny *et al.*, “Machine Learning Prediction of Response to Cardiac Resynchronization Therapy: Improvement Versus Current Guidelines,” Circ Arrhythm Electrophysiol, vol. 12, no. 7, p. e007316, 2019, doi: 10.1161/CIRCEP.119.007316.
16. T. O. Aro, H. B. Akande, M. B. Jibrin, and U. A. Jauro, “Homogenous Ensembles on Data Mining Techniques for Breast Cancer Diagnosis,” Daffodil international university journal of science and technology, vol. 14, no. 1, 2019.
17. H. Seker, M. O. Odetayo, D. Petrovic, R. Naguib, and F. Hamdy, “A Soft Measurement Technique for Searching Significant Subsets of Prostate Cancer Prognostic Markers,” in *The State of the Art in Computational Intelligence*, Heidelberg, 2000, pp. 325–328. doi: 10.1007/978-3-7908-1844-4_52.
18. S. Karun, A. Raj, and G. Attigeri, “Comparative Analysis of Prediction Algorithms for Diabetes,” in *Advances in Computer Communication and Computational Sciences*, Singapore, 2019, pp. 177–187. doi: 10.1007/978-981-13-0341-8_16.
19. M. Bucholc *et al.*, “A practical computerized decision support system for predicting the severity of Alzheimer’s disease of an individual,” *Expert Systems with Applications*, vol. 130, pp. 157–171, Sep. 2019, doi: 10.1016/j.eswa.2019.04.022.
20. I. U. Ekanayake and D. Herath, “Chronic Kidney Disease Prediction Using Machine Learning Methods,” in 2020 *Moratuwa Engineering Research Conference (MERCon)*, Jul. 2020, pp. 260–265. doi: 10.1109/MERCon50084.2020.9185249.
21. S. Sossi Alaoui, B. Aksasse, and Y. Farhaoui, “Statistical and Predictive Analytics of Chronic Kidney Disease,” in *Advanced Intelligent Systems for Sustainable Development (AI2SD'*2018), Cham, 2019, pp. 27–38. doi: 10.1007/978-3-030-11884-6_3.
22. A. Ogunleye and Q.-G. Wang, “XGBoost Model for Chronic Kidney Disease Diagnosis,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 6, pp. 2131–2140, Nov. 2020, doi: 10.1109/TCBB.2019.2911071.
23. S. Zeynu, “Prediction of Chronic Kidney Disease Using Data Mining Feature Selection and Ensemble Method,” WSEAS TRANSACTIONS ON INFORMATION SCIENCE AND APPLICATIONS, 2018.

24. N. V. G. Raju, K. P. Lakshmi, K. G. Praharshitha, and C. Likhitha, “Prediction of chronic kidney disease (CKD) using Data Science,” in 2019 *International Conference on Intelligent Computing and Control* Systems *(ICCS)*, May 2019, pp. 642–647. doi: 10.1109/ICCS45141.2019.9065309.
25. B. Khan, R. Naseem, F. Muhammad, G. Abbas, and S. Kim, “An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy,” IEEE Access, vol. 8, pp. 55012–55022, 2020, doi: 10.1109/ACCESS.2020.2981689.
26. K. M. Zubair Hasan and Md. Zahid Hasan, “Performance Evaluation of Ensemble-Based Machine Learning Techniques for Prediction of Chronic Kidney Disease,” in *Emerging Research in Computing, Information, Communication and Applications*, Singapore, 2019, pp. 415–426. doi: 10.1007/978-981-13-5953-8_34.
27. A. A. Abdullah, S. A. Hafidz, and W. Khairunizam, “Performance Comparison of Machine Learning Algorithms for Classification of Chronic Kidney Disease (CKD),” J. Phys.: Conf. Ser., vol. 1529, p. 052077, May 2020, doi: 10.1088/1742-6596/1529/5/052077.
28. A. Alaiad, H. Najadat, B. Mohsen, and K. Balhaf, “Classification and Association Rule Mining Technique for Predicting Chronic Kidney Disease,” J. Info. Know. Mgmt., vol. 19, no. 01, p. 2040015, Mar. 2020, doi: 10.1142/S0219649220400158.
29. M. Kadhum, S. Manaseer, and A. L. A. Dalhoum, “Evaluation Feature Selection Technique on Classification by Using Evolutionary ELM Wrapper Method with Features Priorities,” JAIT, vol. 12, no. 1, pp. 21–28, 2021, doi: 10.12720/jait.12.1.21-28.
30. O. Sagi and L. Rokach, “Ensemble learning: A survey,” *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018, doi: https://doi.org/10.1002/widm.1249.
31. O. Sagi and L. Rokach, “Explainable decision forest: Transforming a decision forest into an interpretable tree,” Information Fusion, vol. 61, pp. 124–138, Sep. 2020, doi: 10.1016/j.inffus.2020.03.013.
32. A. Barredo Arrieta *et al.*, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.
33. A. Fisher, C. Rudin, and F. Dominici, “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]*, Dec. 2019, Accessed: Feb. 11, 2021. [Online]. Available: http://arxiv.org/abs/1801.01489
34. J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” Annals of Statistics, vol. 29, pp. 1189–1232, 2000.
35. S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *arXiv:1705.07874 [cs, stat]*, Nov. 2017, Accessed: Feb. 03, 2021. [Online]. Available: http://arxiv.org/abs/1705.07874
36. S. M. Lundberg *et al.*, “Explainable machine-learning predictions for the prevention of hypoxaemia during surgery,” *Nature Biomedical Engineering*, vol. 2, no. 10, Art. no. 10, Oct. 2018, doi: 10.1038/s41551-018-0304-0.

37. S. Shilaskar and A. Ghatol, “Feature selection for medical diagnosis: Evaluation for cardiovascular diseases,” *Expert Systems with Applications*, vol. 40, no. 10, pp. 4146–4153, Aug. 2013, doi: 10.1016/j.eswa.2013.01.032.
38. J. Tang, S. Alelyani, and H. Liu, “Feature Selection for Classification: A Review,” p. 33.
39. T. Tagaris and A. Stafylopatis, “Hide-and-Seek: A Template for Explainable AI,” *arXiv:2005.00130 [cs, stat]*, Apr. 2020, Accessed: Aug. 21, 2020. [Online]. Available: http://arxiv.org/abs/2005.00130
40. P. A. Moreno-Sanchez, “An automated feature selection and classification pipeline to improve explainability of clinical prediction models,” in 2021 *IEEE 9th International Conference on Healthcare Informatics (ICHI)*, Aug. 2021, pp. 527–534. doi: 10.1109/ICHI52183.2021.00100.
41. F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” MACHINE LEARNING IN PYTHON, p. 6.
42. A. Salekin and J. Stankovic, “Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes,” in 2016 *IEEE International Conference on Healthcare Informatics (ICHI)*, Oct. 2016, pp. 262–270. doi: 10.1109/ICHI.2016.36.
43. A. Sobrinho, A. C. M. D. S. Queiroz, L. Dias Da Silva, E. De Barros Costa, M. Eliete Pinheiro, and A. Perkusich, “Computer-Aided Diagnosis of Chronic Kidney Disease in Developing Countries: A Comparative Analysis of Machine Learning Techniques,” IEEE Access, vol. 8, pp. 25407–25419, 2020, doi: 10.1109/ACCESS.2020.2971208.

Figures

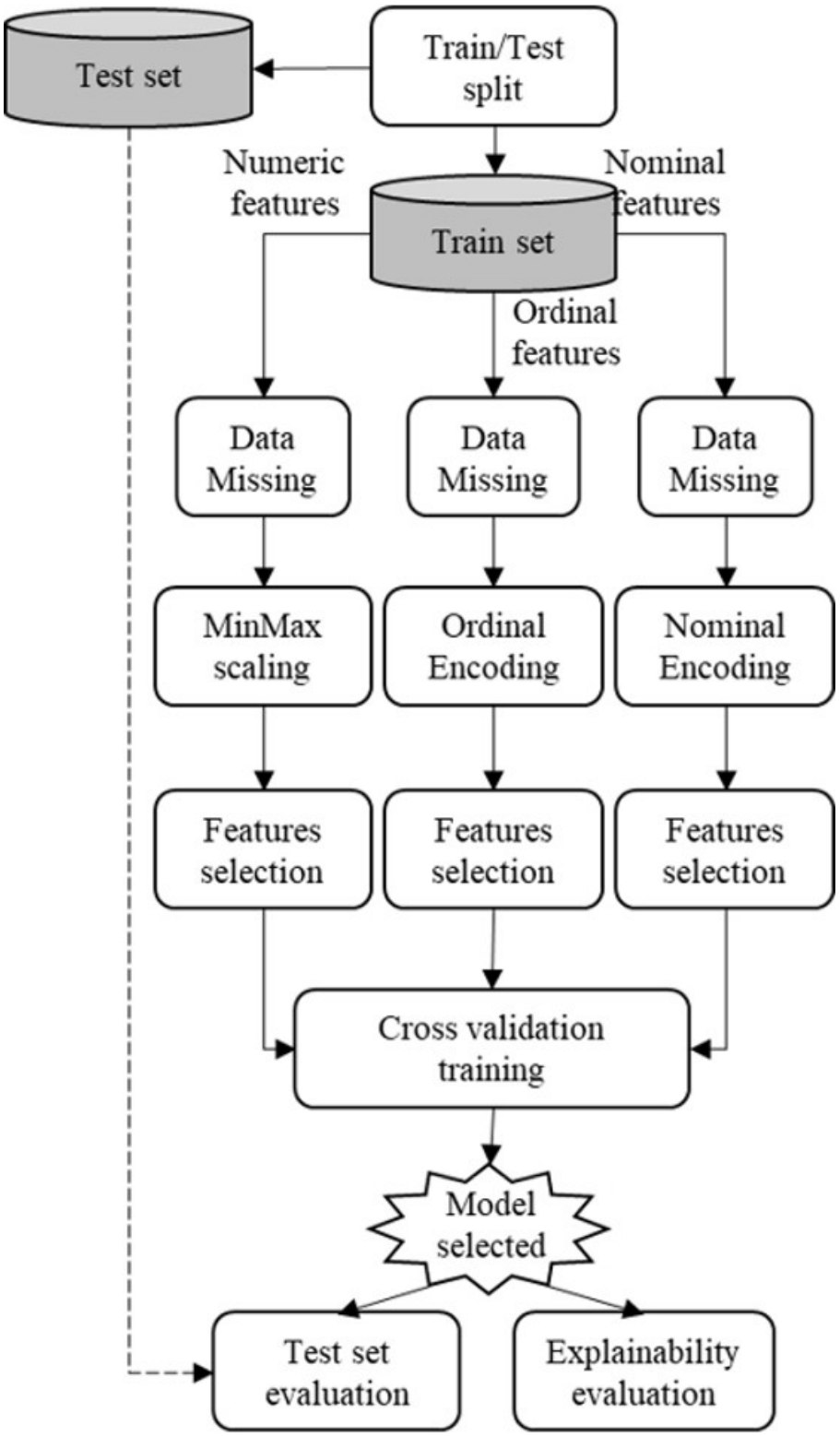

Figure 1

SCI-XAI automated model selection pipeline.

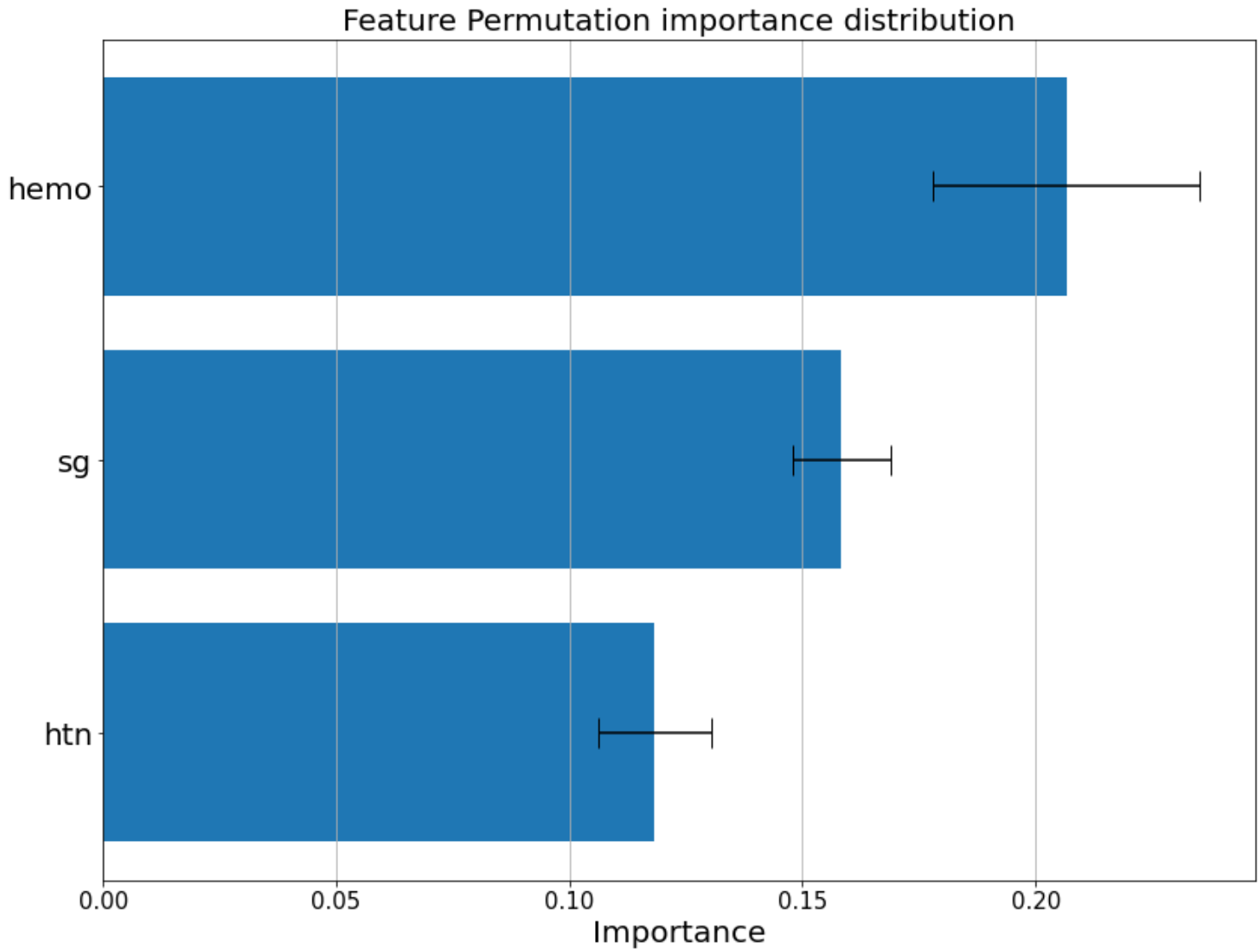

Figure 2

Global explainability obtained with Permutation Feature Importance technique.

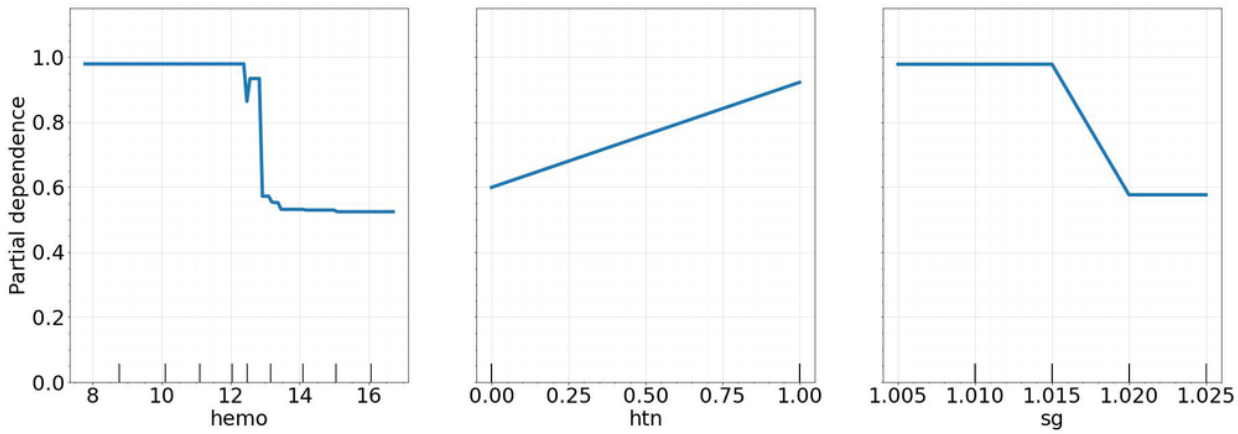

Figure 3

PDP plots of CKD probability contribution for each model's feature

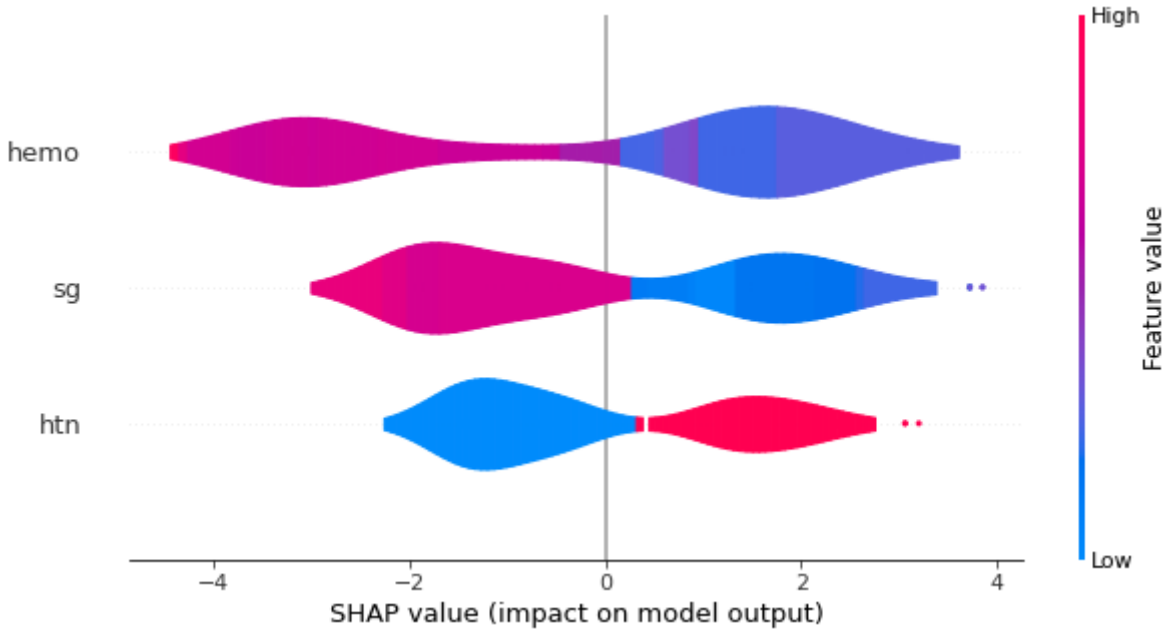

Figure 4

General explainability of CKD probability contribution by using the SHAP technique

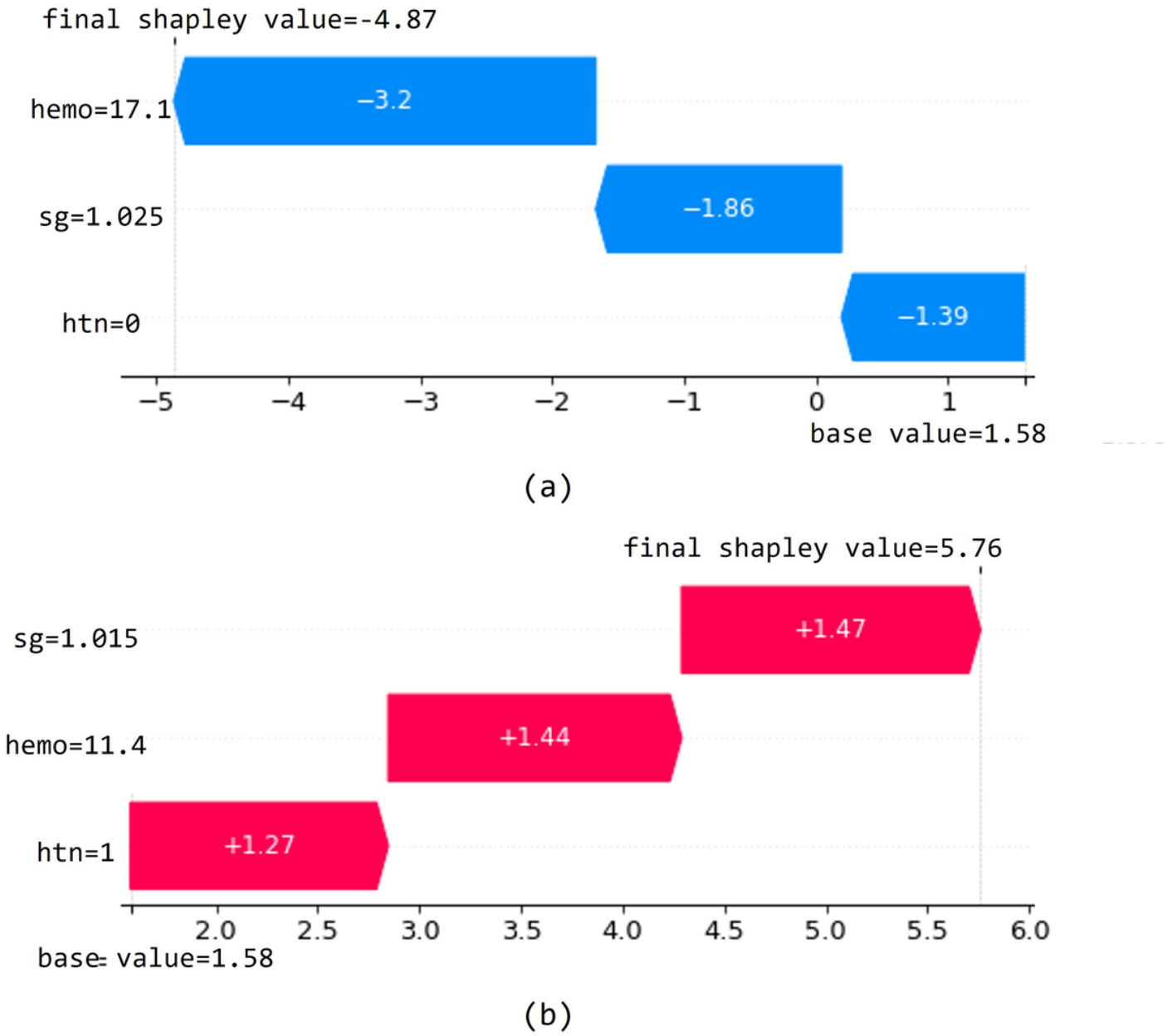

Figure 5

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