

Received 4 March 2023, accepted 23 March 2023, date of publication 3 April 2023, date of current version 21 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3264270

## RESEARCH ARTICLE

# Data-Driven Early Diagnosis of Chronic Kidney Disease: Development and Evaluation of an Explainable AI Model

PEDRO A. MORENO-SÁNCHEZ 

Faculty of Medicine and Health Technology, Tampere University, 60320 Seinäjoki, Finland

e-mail: pedro.morenosanchez@tuni.fi

**ABSTRACT** Chronic Kidney Disease (CKD) is currently experiencing a growing worldwide incidence and can lead to premature mortality if diagnosed late, resulting in rising costs to healthcare systems. Artificial Intelligence (AI) and Machine Learning (ML) offer the possibility of an early diagnosis of CKD that could revert further kidney damage. However, clinicians may be hesitant to adopt AI models if the reasoning behind the predictions is not understandable. Since eXplainable AI (XAI) addresses the clinicians' requirement of understanding AI models' output, this work presents the development and evaluation of an explainable CKD prediction model that provides information about how different patient's clinical features contribute to CKD early diagnosis. The model was developed using an optimization framework that balances classification accuracy and explainability. The main contribution of the paper lies in an explainable data-driven approach to offer quantitative information about the contribution of certain clinical features in the early diagnosis of CKD. As a result, the optimal explainable prediction model implements an extreme gradient boosting classifier using 3 features (hemoglobin, specific gravity, and hypertension) with an accuracy of 99.2% (standard deviation 0.8) and 97.5% with a 5-fold cross-validation and with new unseen data respectively. In addition, an explainability analysis shows that hemoglobin is the most relevant feature that influences the prediction, 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.


**INDEX TERMS** Clinical prediction model, early diagnosis, chronic kidney disease, feature selection, medical explainable AI.

## I. INTRODUCTION

Chronic kidney disease (CKD) has become a worldwide public health problem with increasing incidence (more than 800 million individuals in 2017) and prevalence (13.4% globally) which can lead to premature mortality for many patients (1.2 million people died from CKD in 2017) [1]. CKD is one of a small number of non-communicable diseases that have shown an increase in associated deaths over the past 2 decades, producing a significant burden to healthcare systems, especially in low-middle income countries where lack of appropriate renal replacement therapy results in a high mortality rate [2], [3], [4]. CKD, usually caused by diabetes

and hypertension, is a non-communicable chronic disease with comorbidities associated, and cardiovascular diseases are the major cause of early morbidity and mortality sustained by patients with CKD [5].

Typically, CKD has no early symptoms [5], and when detected through laboratory testing, which quantifies the estimated glomerular filtration rate (eGFR), the kidney has already lost 25 percent of its capacity and is under irreversible and progressive damage toward the so-called end-stage kidney disease. At this point, symptoms may appear such as leg swelling, extreme fatigue, generalized weakness, shortness of breath, loss of appetite, or confusion [6]. If this irreversible deterioration is not slowed by controlling underlying risk factors (hypertension, obesity, heart disease, age) [7], hemodialysis or even kidney transplantation becomes crucial

The associate editor coordinating the review of this manuscript and approving it for publication was Inês Domingues .

for the patient to avoid an exponential rise of risk of death. [4], [5], [6]. Consequently, early diagnosis of CKD based on risk factors, along with their monitoring, allows for initiating preventive treatments and therapeutic measures that slow the progression of kidney damage and prolong patients' life. [4]. In addition, the early identification of groups at high risk of CKD has become an important focus in kidney disease management strategies [8].

In the medical domain, Artificial intelligence (AI) and machine learning (ML) have become promising tools for building computer-aided diagnosis (CAD) systems. These systems use algorithms that learn to classify individuals with specific symptoms either as sick or healthy [8], [9]. AI/ML can be employed to discover latent patterns and correlations between CKD and its risk factors, enabling an early discovery of patients at risk in an effective, convenient, and low-cost manner [9], [10]. Feature selection (FS) is a crucial step in the ML process, which removes unnecessary and unimportant attributes to result in less complicated and more accurate and interpretable models [11]. This features selection step becomes a relevant aspect when dealing with medical datasets due to their high dimensionality resulting from the combination of different variables and multiple measurement techniques when registering patient information.

When CAD systems' decisions affect patients' lives, the transparency and explanations of the AI models' outputs are essential to support clinicians in their diagnosis and treatments. Thus, eXplainable Artificial Intelligence (XAI), which can be defined as a class of systems that provides insight into how an AI system makes decisions and predictions by giving details or reasons to make its functioning clear or easy to understand, allows healthcare experts to make reasonable and data-driven decisions that would enhance the clinical adoption of AI models and their acceptance [12]. XAI is a research area in AI that is acquiring recently an emergent relevance [13], and different solutions have been developed over the last decade in several clinical fields, namely: urology [14], toxicology [14], endocrinology [15], neurology [16], cardiology [17], cancer (e.g. breast cancer or prostate cancer) [18], and chronic diseases (e.g. diabetes or Alzheimer's disease) [19], [20]; and it. Developing explainable AI models in the medical domain involves an inherent trade-off between predictive accuracy, which provides the reliability of the model, and the explainability requested by clinical experts. This tension must be addressed properly by engaging the clinical experts in the development process because the most accurate models, which might be the most interesting to them, are usually less transparent and vice versa. Concerning the application of XAI approaches to CKD prediction models, to the best of our knowledge, no XAI analysis beyond applying feature selection has been found in the literature.

The aim of this paper is to describe the development and assessment of an explainable prediction model of CKD.<sup>1</sup>

<sup>1</sup>Source code of this research can be found at: [https://github.com/petmoreno/Chronic\\_Kidney\\_Disease\\_Predictor](https://github.com/petmoreno/Chronic_Kidney_Disease_Predictor)

This model is developed using an automated optimization framework, named SCI-XAI (feature Selection and Classification for Improving eXplainable AI), that calculates the best combination of different ensemble tree algorithms and feature selection techniques in terms of accuracy performance and minimum number of selected features. Additionally, an explainability analysis is conducted to determine the relevance of the different features selected by the optimal model, as well as to assess the model using explainability metrics to find an appropriate balance between classification performance and explainability. To adhere this research to standardized reporting practices for prediction models in medicine, the TRIPOD statement guideline [21] has been adopted, and its 22-item checklist is reported as supplementary material. This work contributes to the state-of-art by extending the feature selection approach and providing an explainability analysis that quantifies the relevance of different risk factors for the early diagnosis of CKD. Furthermore, the results obtained in terms of feature selection improve upon those reported by related works, making this work the one that selects the least number of features while maintaining a reasonably good classification performance.

The remainder of the article adopts the following structure: (1) an overview of the different related works identified that aim to develop a CKD prediction model considering feature selection, (2) a description of the dataset, the optimization framework employed to develop our model, the machine learning algorithms, evaluation metrics, and explainability techniques employed in this research; (3) the evaluation results in terms of classification and explainability, and the explainability analysis; (4) the discussion of the results as well as the conclusions obtained.

## II. RELATED WORKS

In order to promote timely identification of patients at high risk of kidney function deterioration, researchers have developed several disease prediction models. Although these models performed well at internal validation, they present an uncertain generalization capability due in part to the use of non-public datasets, such as medical images or clinical data from EHR [22], which hinders benchmarking of the models [8].

This research paper advocates for an open-science approach and describes CKD early-diagnosis model developed using a public open dataset from UCI-ML repository [23]. This allows other researchers to benchmark the generalization performance of their models. Table 1 shows the most recent and accurate works (with accuracy above 98%) that use the CKD dataset from the UCI-ML repository and implement feature selection as a preprocessing step in their CKD data analysis pipeline.

Although the related works consider the reduction of the original number of features, there has not been a proper tackling, to the best of our knowledge, of the explainability of their results. Thus, our research provides a novel approach by analyzing the explainability of the prediction model

**TABLE 1. Classification results (in % and descending order) and their machine learning classifiers (best ones in italic underlined) of related works.**

| Article               | Acc   | Sen   | Spe   | F1    | Pre   | #F | Machine Learning Classifiers used in the paper           |
|-----------------------|-------|-------|-------|-------|-------|----|--|
| Ekanayake [24]        | 100   | 100   | -     | 100   | 100   | 7  | <i>DT, RF, XGB, Ada, ET</i> (*)                          |
| Alaoui [25]           | 100   | -     | -     | -     | -     | 23 | <i>XGB Lin, Lin SVM, DT, RF</i>                          |
| Ogunleye [26]         | 100   | 100   | 100   | -     | 100   | 12 | <i>XGB</i> (*)   |
| Abdel-Fattah [11]     | 100   | 100   | 100   | 100   | 100   | 12 | SVM, RF, DT, <i>GBT</i> , LR, NB(*)                      |
| Ebiaredoh-Mienye [10] | 99.9  | 100   | 99.8  | -     | -     | 18 | LR, DT, XGB, RF, SVM, <i>Ada</i>                         |
| Zeynu [27]            | 99.5  | 99.5  | -     | 99.5  | 99.5  | 8  | KNN, DT, <i>ANY</i> , NB, SVM.                           |
| Raju [28]             | 99.3  | 99    | -     | 99    | 100   | 5  | XGB, <i>RF</i> , LR, SVM, NB(*)                          |
| Imran Ali[29]         | 99,1  | 100   | 97.5  | 99.4  | 98.8  | 6  | NB, <i>LG</i> , ANN, DT, RF, GBT, SVM                    |
| Khan [30]             | 99.1  | 99.7  | -     | 99.3  | 98.7  | 23 | <i>NB</i> , LR, SVM, DT, RF                              |
| Hasan [31]            | 99    | -     | -     | 99    | -     | 13 | <i>Ada</i> , RF, GB, ET(*)                               |
| Antony [32]           | 99    | 100   | -     | 99.2  | 98.4  | 10 | <i>KMeans</i> , DBScan, Autoencoder, IForest             |
| Chaudhuri [33]        | 99    | 96    | 100   | -     | -     | 13 | LR, NB, SVM, DT, RF, <i>EDT</i> (*)                      |
| Abdullah [34]         | 98.8  | 98.0  | 100   | 98.8  | 98.0  | 10 | <i>RF</i> , SVM, NB, LR                                  |
| Poonia [35]           | 98.75 | 98    | -     | 99    | 100   | 14 | <i>LR</i> , NB, SVM, KNN, ANN                            |
| Siddhartha [36]       | 98.75 | 100   | 96.67 | 99    | 98.03 | 5  | <i>RF</i> , Ada, XGB                                     |
| Alaiad [37]           | 98.5  | 99.6  | 96.8  | -     | 98    | 12 | NB, DT, SVM, <i>KNN</i> , Jrip                           |
| Kadhun [38]           | 98.1  | 98    | -     | 98    | 98    | 10 | SVM, <i>ELM</i>  |
| Akter [39]            | 97    | 98    | -     | 96    | 97    | 10 | <i>ANN</i> , LSTM, Bi LSTM, GRU, Bi GRU, Simple RNN, MLP |
| Theerthagiri [40]     | 96    | 97    | 99    | 94.9  | 95.8  | 6  | LR, SVM, KNN, NB, <i>RF</i>                              |
| Ali [41]              | 91.25 | 91.89 | 97.37 | 94.81 | 97.81 | 5  | NB, LG, DL, <i>ANN</i> , RF, GBT, SVM                    |

\*: Studies that perform the best classifier with unseen new data; *Acc*: accuracy; *Classification metrics legend*: *Sen*: sensitivity; *Spe*: specificity; *F1*:f1-score; *Pre*: Precision; *#F*: number of features selected; *Machine learning classifier legend*: *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, *KMeans*: K-means clustering, *DBScan*: Density-Based Spatial Clustering of Applications with Noise; *IForest*: Isolation Forest, *LSTM*: Long Short-Term Memory, *Bi LSTM*: Bidirectional LSTM, *GRU*: Gated Recurrent Units, *Bi GRU*: Bidirectional GRU, *RNN*: Recurrent Neural Network, *MLP*: Multi-Layer Perceptron

developed offering information on the influence of the selected features in the classification of CKD.

### III. MATERIAL AND METHODS

#### A. CHRONIC KIDNEY DISEASE DATASET

To promote the reproducibility of this research as well as to benchmark to the existing related works, the CKD dataset from UCI-ML 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).

The features contained in the dataset represent the following information [legend in brackets]: age in years [age], diastolic blood pressure in mm/Hg [bp], specific gravity to compare the density of urine to the density of water [sg], presence of albumin in urine[al], level of sugar is present in urine[su], red blood cells present in urine[rbc], pus cells present in the urine, indicating major or minor infection [pc], pus cell clumps indicating if the infection is present in the urine [pcc], if the growth of bacteria is evident in urine [ba], sugar level in blood in mgs/dl [bgr], level of urea nitrogen in blood in mgs/dl [bu], level of creatinine in blood in mgs/dl [sc], level of sodium in blood in mEq/L [sod], level of potassium in blood in mEq/L [pot], protein in red blood cells in Gms [hemo], percentage of cells in blood [pcv],

amount of white blood cells present in the blood (cells/cumm) [wc], amount of red blood cells present in the blood (millions/cmm) [rc], whether the patient has higher level of blood pressure [htn], presence of diabetes [dm], whether the patient is suffering from coronary artery disease [cad], loss of appetite [appet], level of leg swelling [pedal], whether the patient is suffering from anemia [ane], and whether the patient has CKD or not [target class].

#### B. AUTOMATED FRAMEWORK FOR MODEL SELECTION OPTIMIZATION

In this work, the automated framework named SCI-XAI (feature Selection and Classification for Improving eXplainable AI) and published in [42] is employed to develop the explainable CKD prediction model (Figure 1). SCI-XAI, implemented with the Python scikit-learn package [43], allows obtaining a balanced prediction model in terms of classification performance (accuracy) and explainability (number of features selected) by considering different kind of parameters. Through a brute force optimization algorithm implemented with GridSearchCV method of scikit-learn, SCI-XAI finds the optimal combination regarding ensemble trees classifier, the number of features selected, and the feature selection method, which provides the best accurate classification.

Initially, the dataset is split with target feature stratification allocating 280 and 120 instances respectively into training and held-out test sets (ratio 70/30). Thus, the model's

**TABLE 2.** Description of statistical information of the features included in the dataset.

| Features (units) [legend]                | Type of feature [classes in ordinal or nominal features] | % of non-null values | 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 [1.005,1.010,1.015, 1.020, 1.025]                    | 88,25                | 7, 84, 75, 106, 81  |
| Albumin [al]                             | Ord [0,1,2,3,4,5]  | 88,5                 | 199,44,43,43,24,1   |
| Sugar [su]                               | Ord [0,1,2,3,4,5]  | 87,75                | 290,13,18,14,13,3   |
| Red blood cells [rbc]                    | Nom [normal/abnormal]                                    | 62                   | 201/47 normal/abnormal  |
| Pus cell [pc]                            | Nom [normal/abnormal]                                    | 83,75                | 259/76 normal/abnormal  |
| Pus cell clumps [pcc]                    | Nom [not present/ present]                               | 99                   | 354/42 not present/present  |
| Bacteria [ba]                            | Nom [not present/ present]                               | 99                   | 374/22 not present/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)   |
| Hypertension [htn]                       | Nom [no/yes]   | 99,5                 | 251/147 no/yes  |
| Diabetes mellitus [dm]                   | Nom [no/yes]   | 99,5                 | 261/137 no/yes  |
| Coronary artery disease [cad]            | Nom [no/yes]   | 99,5                 | 364/34 no/yes   |
| Appetite [appet]                         | Nom [good/poor]  | 99,75                | 317/82 good/poor  |
| Pedal edema [pe]                         | Nom [no/yes]   | 99,75                | 323/76 no/yes   |
| Anemia [ane]                             | Nom [no/yes]   | 99,75                | 339/60 no/yes   |
| Target class                             | Nom [notCKD/CKD]   | 100                  | 150/250 not CKD/CKD   |

The information included is the type of feature (Num= numerical, Ord= ordinal, Nom= nominal), % of non-null values, mean and standard deviation (for numerical features), categories, and the number of instances per each category (for ordinal and nominal features)

performance is evaluated over unseen new data from the held-out test set that is applied to the optimal parameters selected by the framework in the preprocessing and training phases. The data preprocessing phase is performed in three separate threads respectively for numerical, nominal, and ordinal features; and entails missing data imputation, scaling/encoding, and feature selection. Next, the preprocessing data thread merged in a 5-fold cross-validation training phase. Due to the small sample size (400 instances), the choice of 5-folds allows reducing overfitting and improve the generalizability of the classification. Finally, the optimal model selected is also assessed in terms of explainability with the interpretability, fidelity, and FII metrics. The methods employed in the different phases of the framework are described in the next subsections.

### C. DATA PREPROCESSING

The SCI-XAI framework tackles the preprocessing of the data in three phases: missing data handling, data encoding, and feature selection. Regarding data missing, the strategy of imputation is implemented depending on the data type of features. Mean value imputation is used for numerical features, while the mode (or most frequent value) is used for ordinal and nominal features. In the case of the encoding phase, a minimum-maximum scaling process is applied to numerical features, while the categories of the ordinal and nominal features are encoded into numbers, i.e. 0-5 with 1 step unit for ordinal, and 0 or 1 for nominal features.

These two steps of missing data handling and encoding are not considered parameters for the optimization algorithm.

In addition to the explainability approaches, it's worth mentioning that feature selection procedures can remove features with non-relevant information from the classification, thereby enhancing models' explainability [44]. This research addresses feature selection by applying filter methods, where intrinsic properties between the dataset's features and the target class are measured with methods like ANOVA, Chi-squared, or mutual information. These methods determine the univariate statistical mutual dependence or significance that justifies the inclusion or withdrawal of a subset of features. Moreover, wrapper feature selection methods are also used, like Recursive Feature Elimination (RFE), where a classification algorithm (e.g. logistic regression) is utilized to find the most significant features by finding a high correlation between the target feature and the rest of the features [45].

### D. ENSEMBLE TREES CLASSIFIER

Thanks to their stability and robustness with datasets of different sizes, as well as a reasonably good predictive performance, ensemble trees have become one of the most popular ML classifiers nowadays. Ensemble trees perform classification tasks by weighting various decision trees and combining them to reach a final model that improves each base estimator [46]. In addition, these classifiers also offer a good performance to mitigate class imbalance situations.

The classifiers used in this research belong to the family of ensemble trees and their different approaches of bagging and boosting, namely: Random Forest and Extra Trees [46], which 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. Additionally, adaptative boosting (AdaBoost) [46] and extreme gradient boosting (XGBoost) [47] apply the technique of boosting, which is focused on instances that have been previously misclassified when training a new base decision tree.

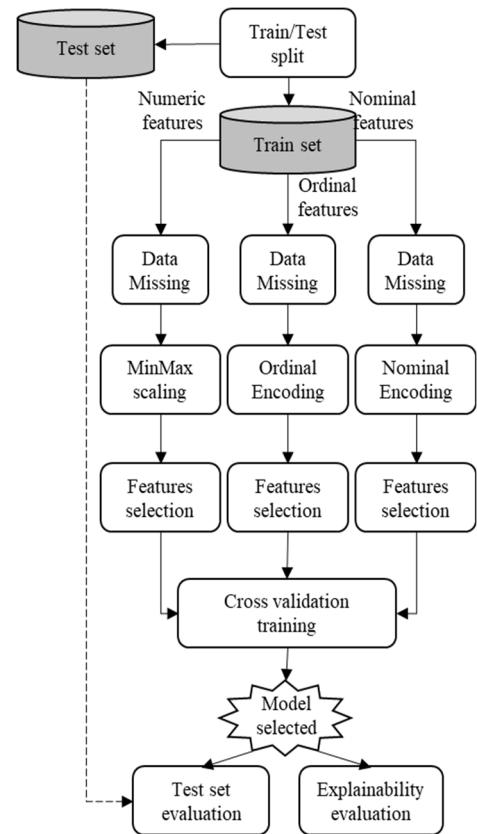
**E. EXPLAINABILITY TECHNIQUES FOR AI**

In domains like healthcare where the model’s predictions should be as interpretable as possible, maintaining the predictive performance of the classifier balanced with explainable capabilities is crucial. Ensemble trees are considered black-box classifiers in terms of explainability due to the poor transparency offered when dealing with multiple base estimators to estimate their predictions. Therefore, post-hoc XAI techniques are required to facilitate providing understandable information about how an already developed model produces its predictions [48].

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 [49]. *Partial dependence plot (PDP)* 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 [50]. *SHapley Additive exPlanations (SHAP)* compute an additive importance score, known as Shapley value, for each feature in every individual prediction by applying coalitional game theory, which are then aggregated to give a global explainability of the model [51], [52].

**F. 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 recommended to use to have a more comprehensive view of the performance, namely: sensitivity, specificity, precision, and F1-score [14]. Additionally, to assess the explainability performance of the classifiers employed, the explainability metrics proposed by Tagaris et al [53] are used: *Interpretability*, defined as the ratio of those masked features that do not bring relevant 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 the *Fidelity-Interpretability Index (FII)* that allow comparing explainability performance between different models.



**FIGURE 1. SCI-XAI automated model selection framework.**

**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)                  | $I = \frac{Masked\ features}{Total\ features}$                             |
| Fidelity (F)                          | $F = \frac{Acc.\ equivalent\ interpretable\ model}{Acc.\ original\ model}$ |
| Fidelity-Interpretability Index (FII) | $FII = F * I$  |
| Fidelity-Accuracy Index (FAI)         | $FAI = F * Acc$  |

With the aim to benchmark the classification and explainability balance of our research results with the ones identified in the related works, we propose a new metric:

**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] |       |
| <i>Random Forest</i> | 1                  | hemo [ANOVA]                              | 5                | htn, dm, appet, rbc, pc [RFE] | 1                | sg [mut-inf]       | 7     |
| <i>Extra Trees</i>   | 4                  | hemo, pcv, rc, sc [mut-inf]               | 3                | htn, dm, appet [chi2]         | 1                | sg [mut-inf]       | 8     |
| <i>AdaBoost</i>      | 7                  | hemo, pcv, rc, sc, sod, pot, wc [mut-inf] | 4                | htn, dm, appet, pc [mut-inf]  | 1                | sg [mut-inf]       | 12    |
| <i>XGBoost</i>       | 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 (held-out -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  |

Fidelity-Accuracy Index that relates the number of features selected and the accuracy performance. The formulas of these metrics are shown in Table 3.

**IV. RESULTS**

**A. FEATURE SELECTION**

Table 4 shows the best combination of features selected for each ensemble trees algorithm obtained by the SCI-XAI framework. The features selected are denoted by their type (numerical, nominal, and ordinal) as well as the selection method (i.e. ANOVA, Chi-squared, Mutual Information, or Recursive Feature Elimination). Considering all the ML classifiers, the framework achieves a reduction of at least 50% of the original features. The major reduction, leaving 3 out of 24 features, is achieved with XGBoost implying that 21 features are non-relevant for the classification. Thus, using the mutual information technique, the features selected are hemo, htn, and sg. Random Forest and Extra Trees (both bagging ensemble trees) achieved a similar general reduction with 7 and 8 features left, respectively. However, there is a substantial difference concerning the numerical features (1 selected by Random Forest and 4 by Extra trees). The worst case is performed by AdaBoost, where 12 features are selected, with a relatively high number of numerical features (7) selected compared to the others ML classifiers.

**B. CLASSIFICATION PERFORMANCE RESULTS**

Table 5 shows the classification performance of the different ensemble trees algorithms considered after the training cross-validation module, as well as the evaluation with the held-out test set (unseen data). The results show a solid classification performance in the training phase, where the classification results for all the metrics considered i.e. accuracy, sensitivity, specificity, f1-score, and precision are 100% in the cases of Random Forest, Extra Trees, and AdaBoost. Concerning XGBoost, the performance decreases slightly although it maintains fairly good results, with the highest value of 100% in sensitivity and the lowest value of 98.1% in specificity

**TABLE 6. Explainability metrics results.**

| Classifier           | Interpretability | Fidelity | FII  |
|----------------------|------------------|----------|------|
| <i>Random Forest</i> | 71 %             | 100 %    | 0,71 |
| <i>Extra Trees</i>   | 67 %             | 99 %     | 0,66 |
| <i>AdaBoost</i>      | 50 %             | 99 %     | 0,50 |
| <i>XGBoost</i>       | 88 %             | 97 %     | 0,85 |

meaning that XGBoost would generate a reduced number of false positives (around 2%).

When evaluating the model with the held-out test set, all classifiers considered obtain an accuracy higher than 97.5%, and values above 95% in the rest of the classification metrics. This implies a robust classification performance even with new and unseen data. Similar to training data, the lowest value is obtained for specificity with XGBoost classifier, indicating a less strong performance in classifying true negative cases.

**C. EXPLAINABILITY METRICS RESULTS**

For the evaluation of explainability, the metrics employed are Fidelity, Interpretability, and Fidelity Interpretability Index. Table 6 shows the results of these three metrics considering the features selected by SCI-XAI for each of the classifiers. Concerning Interpretability, the different models achieve values between 50 to 88%, denoting that at a minimum, half of the initial features are removed. The highest value is for XGBoost, which achieves 88% due to the selection of only three features (hemo, htn, and sg) in the resultant optimal model when using that ML classifier algorithm. In terms of Fidelity, all models achieved high values close to 100%, which indicates that a decision tree built over the same selected features as input would have almost the same an accuracy than the original ML classifier. Moreover, FII gives a balanced measure between interpretability and fidelity to compare different algorithms. Thus, XGBoost (FII = 0.85) achieved the most balanced model in comparison to the other models (Random Forest 0.71, Extra Trees 0.66, and AdaBoost 0.50).

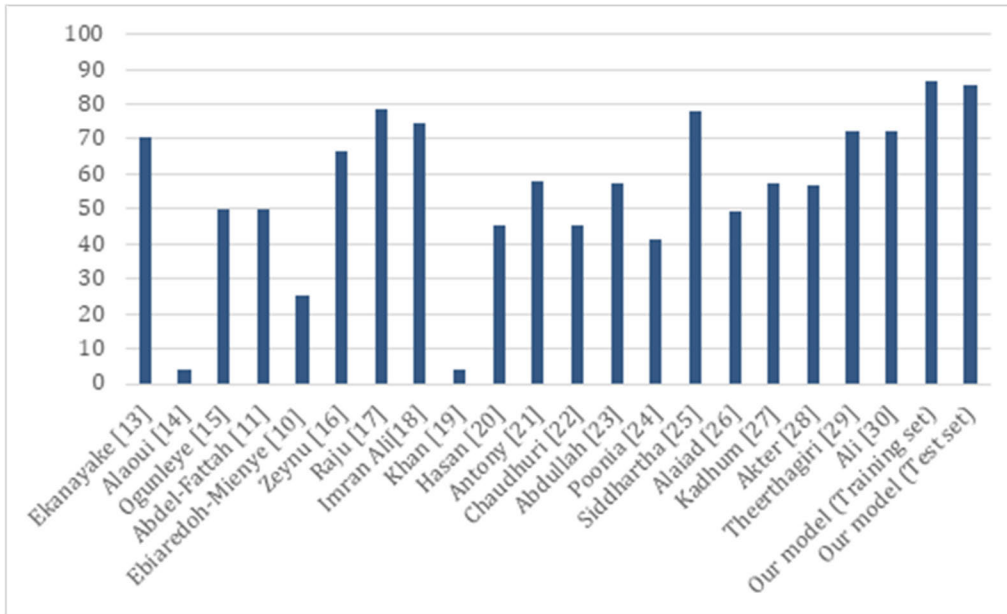


FIGURE 2. Comparison of fidelity accuracy index of the different CKD prediction models from the related compared to our prediction model.

Since the aim of this research is to achieve the most balanced CKD prediction model in terms of classification performance and explainability, we propose the model built with XGBoost and its group of selected features: hemoglobin (hemo), hypertension(htn), and specific gravity (sg) to conduct an explainability analysis of its predictions.

Finally, to benchmark this balance accuracy-interpretability capability of our model with the ones identified in related works, we propose a comparison by using the Fidelity-Accuracy-Index for each of the prediction models. Figure 5 shows the graphical comparison where our model achieves the best result, above 80%, either considering the performance with the training or the test set. Since some related works express their classification performance using cross-validation and other related works use the test set, the FAI of this work is calculated and depicted with the training set, which uses cross-validation, and the test set.

**D. EXPLAINABILITY ANALYSIS OF THE PREDICTION MODEL**

Since XGBoost is the most balanced model in terms of explainability and accuracy, the relevance of hemo, htn, and sg features is analyzed using different post-hoc explainability techniques to demonstrate their influence on the model’s outputs.

Figure 3 shows the features’ importance obtained with PFI, which allows visualizing the global explainability of each feature without indicating the direction (positive or negative) of the contribution to the probability of CKD. According to the Figure 3, hemo is the most relevant feature followed by sg and htn, (in descending order of importance.

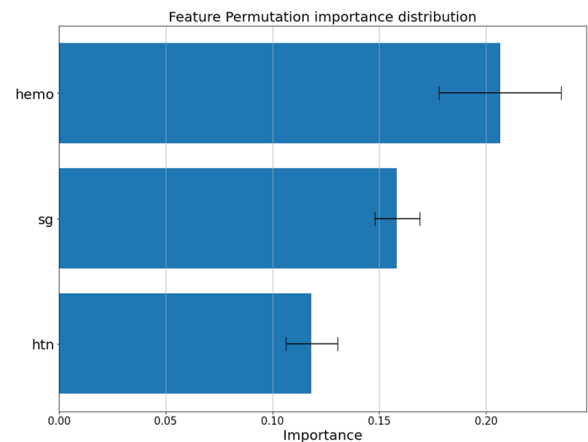


FIGURE 3. Global explainability obtained with Permutation Feature Importance technique.

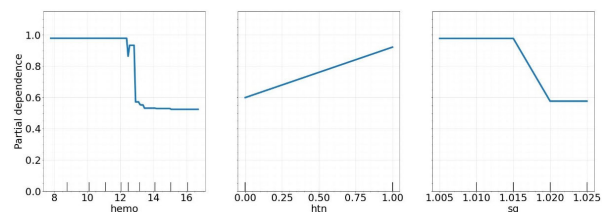


FIGURE 4. PDP plots of CKD probability contribution for each model’s feature.

The PDP plots presented in Figure 4 provide information about the marginal effect of the selected features values (x-axis) on the probability of a positive CKD prediction (y-axis). Thus, the marginal contribution of hemo values

between 12.3 gms and 13.5 gms to predict CKD decreases from 0.98 to 0.53 in several steps with values of 0.94, 0.57, 0.55, and 0.53, and remains monotonic for the rest of hemo values above 13.5 gms. Moreover, patients with hypertension (htn = 1) have a marginal increase of 0.33 (from 0.6 to 0.93) in the probability of developing CKD. In the case of sg, values of 1.020 and 1.025 decrease the marginal probability of predicting CKD by 0.4 (from 0.98 to 0.58).

SHAP technique also allows for explaining the general contribution of every feature to the model’s probability concerning its values. Similar to the trends shown when using PDP, Figure 5 depicts that the 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 the sg feature contribute by reducing the probability of CKD. Additionally, the presence of hypertension (htn equals 1) increases the CKD probability (red color). Besides, SHAP offers explanations concerning predictions of individual cases (shown in Figure 6), 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). As an example of this individual explainability, Figure 5.a and Figure 5.b show the explanations of a predicted true negative case ( $y = 0$ , the patient does not have CKD, and the features’ values for that specific case are hemo = 17.1, sg = 1.025, htn = 0) and a predicted true positive case ( $y = 1$ , the patient does have CKD, and the features’ values for that specific case are hemo = 11.4, sg = 1.015, htn = 0). In both cases, the prediction contribution 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 17.1 gms is denoted as the most relevant feature in the prediction with a Shapley value attribution of  $-3.2$ , meanwhile, sg and htn, with values 1.025 and 0 respectively, have negative Shapley values attributions ( $-1.92$  and  $-1.44$ ). Regarding 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 nearly similar additive Shapley values ( $+1.7 + 1.77, +1.27$  respectively). It is worth noting that the contributions shown for the feature values in these individual cases agree with the findings obtained with the PDP plots.

V. DISCUSSION

Due to the current increase in the global incidence of CKD, timely detection of patients at risk becomes a relevant tool for doctors to achieve a disease early diagnosis. Besides the advent of ML algorithms to develop prediction models that support CKD early diagnosis, XAI could additionally improve these models by meeting the healthcare professionals’ demands for a clearer understanding of models’ decisions. With more explainable CKD early diagnosis models, doctors could make more data-driven decisions and focus



FIGURE 5. General explainability of CKD probability contribution by using the SHAP technique.

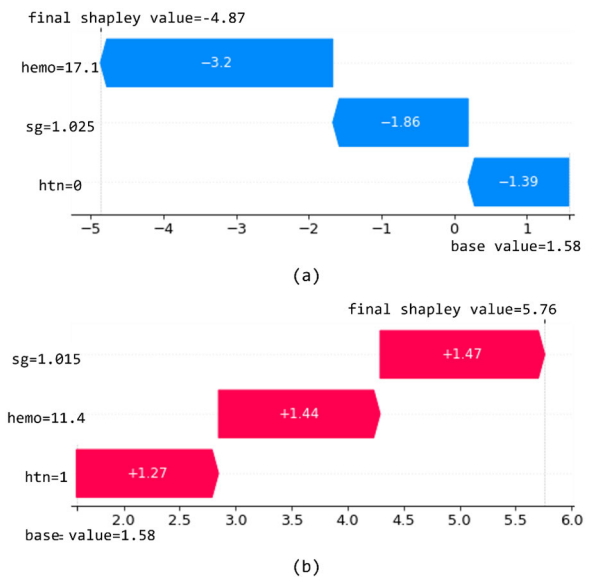


FIGURE 6. Local explainability through SHAP (a. True negative case; b. True positive case).

on controlling underlying features or indicators to slow the progressive damage of the kidneys.

This paper describes a CKD prediction model developed not only to seek high accuracy but also to analyze the explainability of its results, thus contributing 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 framework developed by the authors, which allows for automatic inference of the optimal combination of different parameters such as the appropriate ensemble tree algorithm, relevant features selected, and feature selection method to obtain the best classification performance of the prediction model. Moreover, the framework allows for the evaluation of the model’s performance over new unseen data (by allocating 30% of the original dataset to a held-out test set), which could emulate 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.

The optimization framework considers parameters focused not only on classification but also on the preprocessing stage, where feature selection strategies have proven to be influential in achieving the most accurate classifier. Regarding this preprocessing step, a feature type-driven approach was implemented to process the features into three parallel threads (numerical, ordinal, and nominal) where data imputation, data encoding, and feature selection are applied. Thus, filter and wrapper feature selection methods (ANOVA, chi-squared, mutual information, and recursive feature elimination) were implemented in the optimization framework to remove unimportant features based on the statistical mutual dependence or significance with the target feature.

Concerning the classification, a set of four different ensemble trees ML algorithms (Random Forest, Extra Trees, AdaBoost and XGBoost) were used to obtain the optimal classification model to support CKD early diagnosis since that kind of classifiers are reported in the literature to provide stability and robustness with datasets of different sizes as well as a reasonably good prediction capability. Considering our classification results, this work obtains achieve 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 framework's feature selection step has proven to be valuable by substantially reducing the original number of features, leaving 3 out of 24 when using the XGBoost classifier, which is the best CKD prediction model when compared to other related works in terms of minimum features considered. This insight is supported by a benchmark with the related works, comparing the results regarding the metric Fidelity Accuracy Index. Furthermore, 3 out of 4 ensemble learning algorithms used in the framework obtained their best classification results with only 33% of the original features, showing the capability of the framework 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, providing an example of the tension between accuracy and explainability that occurs in prediction models intended for use in specific domains where understanding the results is crucial, such as 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 at which a particular feature modifies the marginal 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 demonstrating the relevance of the different features for an individual prediction case.

The results described in this work exhibit the added value of explainability to a clinical prediction model. Additionally, 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 three features (hemoglobin, specific gravity, and hypertension), the cost associated with their extraction, following the price list defined by Salekin et al [54], 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 [55].

However, our research has some limitations. First, the present study employs a widely used CKD dataset from a UCI-ML repository, which, although it allows benchmarking with other related research works, lacks an external validation to support objective experimentation. Therefore, to conclusively validate the results, more CKD data would be needed from a different clinical setting from the original.

## VI. CONCLUSION

This research work presents the development and evaluation of an explainable prediction model for CKD early diagnosis. The main goal is to show how XAI contributes to improving prediction models used in the medical field. This research also pursues to exemplifying how to address the existing trade-off between accuracy and explainability when dealing with black box AI models. Therefore, using an automated optimization framework, the best combination of the ensemble tree algorithm and the number of features are selected to provide the best balanced model according to the classification and explainability metrics. The optimal balanced explainable model detected by the framework was an XGBoost classifier that used three features for the CKD prediction: hemoglobin (hemo), specific gravity (sg), and hypertension (htn). After conducting an explainability analysis with different post-hoc techniques, the features' relevance in descending order of importance was found to be hemo, sg, and htn. The prediction model developed in this work achieved the classification performance of the best CKD prediction models identified in the literature with the least number of features selected compared to the other works.

To advance in the line of trustworthiness and transparency of our model, we propose as future works to perform an

external validation with other datasets that contain the same group of features to evaluate the generalization capability of the model in the early CKD diagnosis. Additionally, this external validation could be deployed in a clinical setting with the aim at also gathering insights from clinicians about the explainability results and discussing how it could affect the CKD treatment plans. Therefore, we could confirm that the explainability approach presented in this paper would provide clinicians with an easier understanding and interpretability of how CKD is diagnosed early with a reduced group of indicators. This information would allow them to also focus on tackling relevant features and their values to avoid the CKD onset or even to revert its progress.

## APPENDIX

The TRIPOD statement guideline filled according to the research study characteristics is amended as appendix.

## REFERENCES

- [1] C. P. Kovesdy, "Epidemiology of chronic kidney disease: An update 2022," *Kidney Int. Supplements*, vol. 12, no. 1, pp. 7–11, Apr. 2022, doi: [10.1016/j.kisu.2021.11.003](https://doi.org/10.1016/j.kisu.2021.11.003).
- [2] R. Gupta, N. Koli, N. Mahor, and N. Tejashri, "Performance analysis of machine learning classifier for predicting chronic kidney disease," in *Proc. Int. Conf. Emerg. Technol. (INCET)*, Jun. 2020, pp. 1–4, doi: [10.1109/INCET49848.2020.9154147](https://doi.org/10.1109/INCET49848.2020.9154147).
- [3] World Health Organization. (2019). *World Health Statistics 2019: Monitoring Health for the SDGs, Sustainable Development Goals*. Accessed: Feb. 7, 2023. [Online]. Available: <https://apps.who.int/iris/handle/10665/324835>
- [4] P. Cockwell and L.-A. Fisher, "The global burden of chronic kidney disease," *Lancet*, vol. 395, no. 10225, pp. 662–664, Feb. 2020, doi: [10.1016/S0140-6736\(19\)32977-0](https://doi.org/10.1016/S0140-6736(19)32977-0).
- [5] A. C. Webster, E. V. Nagler, R. L. Morton, and P. Masson, "Chronic kidney disease," *Lancet*, vol. 389, no. 10075, pp. 1238–1252, Mar. 2017, doi: [10.1016/S0140-6736\(16\)32064-5](https://doi.org/10.1016/S0140-6736(16)32064-5).
- [6] H. Polat, H. D. Mehr, and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," *J. Med. Syst.*, vol. 41, p. 55, Apr. 2017, doi: [10.1007/s10916-017-0703-x](https://doi.org/10.1007/s10916-017-0703-x).
- [7] R. A. Jeewantha, M. N. Halgamuge, A. Mohammad, and G. Ekici, "Classification performance analysis in medical science: Using kidney disease data," in *Proc. Int. Conf. Big Data Res.*, Osaka, Japan, 2017, pp. 1–6, doi: [10.1145/3152723.3152724](https://doi.org/10.1145/3152723.3152724).
- [8] N. Lei, X. Zhang, M. Wei, B. Lao, X. Xu, M. Zhang, H. Chen, Y. Xu, B. Xia, D. Zhang, C. Dong, L. Fu, F. Tang, and Y. Wu, "Machine learning algorithms' accuracy in predicting kidney disease progression: A systematic review and meta-analysis," *BMC Med. Informat. Decis. Making*, vol. 22, no. 1, p. 205, Aug. 2022, doi: [10.1186/s12911-022-01951-1](https://doi.org/10.1186/s12911-022-01951-1).
- [9] J. Qezelbash-Chamak, S. Badamchizadeh, K. Eshghi, and Y. Asadi, "A survey of machine learning in kidney disease diagnosis," *Mach. Learn. Appl.*, vol. 10, Dec. 2022, Art. no. 100418, doi: [10.1016/j.mlwa.2022.100418](https://doi.org/10.1016/j.mlwa.2022.100418).
- [10] S. A. Ebiaredoh-Mienye, T. G. Swart, E. Esenogho, and I. D. Mienye, "A machine learning method with filter-based feature selection for improved prediction of chronic kidney disease," *Bioengineering*, vol. 9, no. 8, p. 350, Jul. 2022, doi: [10.3390/bioengineering9080350](https://doi.org/10.3390/bioengineering9080350).
- [11] M. A. Abdel-Fattah, N. A. Othman, and N. Goher, "Predicting chronic kidney disease using hybrid machine learning based on Apache Spark," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–12, Feb. 2022, doi: [10.1155/2022/9898831](https://doi.org/10.1155/2022/9898831).
- [12] 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. Discovery*, vol. 10, no. 5, p. e1379, Sep. 2020, doi: [10.1002/widm.1379](https://doi.org/10.1002/widm.1379).
- [13] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 1–21, Nov. 2021, doi: [10.1109/TNNLS.2020.3027314](https://doi.org/10.1109/TNNLS.2020.3027314).
- [14] 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," *Mol. Diversity*, vol. 23, no. 2, pp. 381–392, May 2019, doi: [10.1007/s11030-018-9882-8](https://doi.org/10.1007/s11030-018-9882-8).
- [15] S. S. 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, Switzerland: Springer, 2020, pp. 59–72, doi: [10.1007/978-3-030-23672-4\\_6](https://doi.org/10.1007/978-3-030-23672-4_6).
- [16] 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, Mar. 2019, doi: [10.1016/j.compbiomed.2019.01.009](https://doi.org/10.1016/j.compbiomed.2019.01.009).
- [17] A. K. Feeny, J. Rickard, D. Patel, S. Toro, K. M. Trulock, C. J. Park, M. A. LaBarbera, N. Varma, M. J. Niebauer, S. Sinha, E. Z. Gorodeski, R. A. Grimm, X. Ji, J. Barnard, A. Madabhushi, D. D. Spragg, and M. K. Chung, "Machine learning prediction of response to cardiac resynchronization therapy: Improvement versus current guidelines," *Circulat., Arrhythmia Electrophysiol.*, vol. 12, no. 7, Jul. 2019, Art. no. e007316, doi: [10.1161/CIRCEP.119.007316](https://doi.org/10.1161/CIRCEP.119.007316).
- [18] T. O. Aro, H. B. Akande, M. B. Jibrin, and U. A. Jauro, "Homogenous ensembles on data mining techniques for breast cancer diagnosis," *Daffodil Int. Univ. J. Sci. Technol.*, vol. 14, no. 1, pp. 1–10, 2019.
- [19] S. Karun, A. Raj, and G. Attigeri, "Comparative analysis of prediction algorithms for diabetes," in *Advances in Computer Communication and Computational Sciences*. Singapore: Springer, 2019, pp. 177–187, doi: [10.1007/978-981-13-0341-8\\_16](https://doi.org/10.1007/978-981-13-0341-8_16).
- [20] M. Bucholc, X. Ding, H. Wang, D. H. Glass, H. Wang, G. Prasad, L. P. Maguire, A. J. Bjourson, P. L. McClean, S. Todd, D. P. Finn, and K. Wong-Lin, "A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual," *Expert Syst. Appl.*, vol. 130, pp. 157–171, Sep. 2019, doi: [10.1016/j.eswa.2019.04.022](https://doi.org/10.1016/j.eswa.2019.04.022).
- [21] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement," *Ann. Internal Med.*, vol. 162, no. 1, pp. 55–63, Jan. 2015, doi: [10.7326/M14-0697](https://doi.org/10.7326/M14-0697).
- [22] A. Ray and A. K. Chaudhuri, "Smart healthcare disease diagnosis and patient management: Innovation, improvement and skill development," *Mach. Learn. Appl.*, vol. 3, Mar. 2021, Art. no. 100011, doi: [10.1016/j.mlwa.2020.100011](https://doi.org/10.1016/j.mlwa.2020.100011).
- [23] D. Dua and C. Graff. (2017). *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [24] I. U. Ekanayake and D. Herath, "Chronic kidney disease prediction using machine learning methods," in *Proc. Moratuwa Eng. Res. Conf. (MERCon)*, Jul. 2020, pp. 260–265, doi: [10.1109/MER-CON50084.2020.9185249](https://doi.org/10.1109/MER-CON50084.2020.9185249).
- [25] S. S. Alaoui, B. Aksasse, and Y. Farhaoui, "Statistical and predictive analytics of chronic kidney disease," in *Advanced Intelligent Systems for Sustainable Development*. Cham, Switzerland: Springer, 2019, pp. 27–38, doi: [10.1007/978-3-030-11884-6\\_3](https://doi.org/10.1007/978-3-030-11884-6_3).
- [26] A. Ogunleye and Q.-G. Wang, "XGBoost model for chronic kidney disease diagnosis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 6, pp. 2131–2140, Nov. 2020, doi: [10.1109/TCBB.2019.2911071](https://doi.org/10.1109/TCBB.2019.2911071).
- [27] S. Zeynu, "Prediction of chronic kidney disease using data mining feature selection and ensemble method," *Int. J. Data Mining Genomics Proteomics*, vol. 9, no. 1, pp. 1–9, 2018.
- [28] N. V. G. Raju, K. P. Lakshmi, K. G. Praharsitha, and C. Likhitha, "Prediction of chronic kidney disease (CKD) using data science," in *Proc. Int. Conf. Intell. Comput. Control Syst. (ICCS)*, May 2019, pp. 642–647, doi: [10.1109/ICCS45141.2019.9065309](https://doi.org/10.1109/ICCS45141.2019.9065309).
- [29] S. Imran Ali, B. Ali, J. Hussain, M. Hussain, F. A. Satti, G. H. Park, and S. Lee, "Cost-sensitive ensemble feature ranking and automatic threshold selection for chronic kidney disease diagnosis," *Appl. Sci.*, vol. 10, no. 16, p. 5663, Aug. 2020, doi: [10.3390/app10165663](https://doi.org/10.3390/app10165663).
- [30] 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](https://doi.org/10.1109/ACCESS.2020.2981689).
- [31] K. M. Z. Hasan and M. Z. 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: Springer, 2019, pp. 415–426, doi: [10.1007/978-981-13-5953-8\\_34](https://doi.org/10.1007/978-981-13-5953-8_34).

- [32] L. Antony, S. Azam, E. Ignatious, R. Quadir, A. R. Beeravolu, M. Jonkman, and F. De Boer, "A comprehensive unsupervised framework for chronic kidney disease prediction," *IEEE Access*, vol. 9, pp. 126481–126501, 2021, doi: [10.1109/ACCESS.2021.3109168](https://doi.org/10.1109/ACCESS.2021.3109168).
- [33] A. K. Chaudhuri, D. Sinha, D. K. Banerjee, and A. Das, "A novel enhanced decision tree model for detecting chronic kidney disease," *Netw. Model. Anal. Health Informat. Bioinf.*, vol. 10, no. 1, p. 29, Apr. 2021, doi: [10.1007/s13721-021-00302-w](https://doi.org/10.1007/s13721-021-00302-w).
- [34] 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, no. 5, May 2020, Art. no. 052077, doi: [10.1088/1742-6596/1529/5/052077](https://doi.org/10.1088/1742-6596/1529/5/052077).
- [35] R. C. Poonia, M. K. Gupta, I. Abunadi, A. A. Albraikan, F. N. Al-Wesabi, and M. A. Hamza, "Intelligent diagnostic prediction and classification models for detection of kidney disease," *Healthcare*, vol. 10, no. 2, p. 371, Feb. 2022, doi: [10.3390/healthcare10020371](https://doi.org/10.3390/healthcare10020371).
- [36] M. Siddhartha, V. Kumar, and R. Nath, "Early-stage diagnosis of chronic kidney disease using majority vote—Grey wolf optimization (MV-GWO)," *Health Technol.*, vol. 12, no. 1, pp. 117–136, Jan. 2022, doi: [10.1007/s12553-021-00617-8](https://doi.org/10.1007/s12553-021-00617-8).
- [37] A. Alaiad, H. Najadat, B. Mohsen, and K. Balhaf, "Classification and association rule mining technique for predicting chronic kidney disease," *J. Inf. Knowl. Manage.*, vol. 19, no. 1, Mar. 2020, Art. no. 2040015, doi: [10.1142/S0219649220400158](https://doi.org/10.1142/S0219649220400158).
- [38] M. Kadhum, S. Manaseer, and A. L. A. Dalhoum, "Evaluation feature selection technique on classification by using evolutionary ELM wrapper method with features priorities," *J. Adv. Inf. Technol.*, vol. 12, no. 1, pp. 21–28, 2021, doi: [10.12720/jait.12.1.21-28](https://doi.org/10.12720/jait.12.1.21-28).
- [39] S. Akter, A. Habib, M. A. Islam, M. S. Hossen, W. A. Fahim, P. R. Sarkar, and M. Ahmed, "Comprehensive performance assessment of deep learning models in early prediction and risk identification of chronic kidney disease," *IEEE Access*, vol. 9, pp. 165184–165206, 2021, doi: [10.1109/ACCESS.2021.3129491](https://doi.org/10.1109/ACCESS.2021.3129491).
- [40] P. Theerthagiri and A. U. Ruby, "RFFS: Recursive random forest feature selection based ensemble algorithm for chronic kidney disease prediction," *Expert Syst.*, vol. 39, no. 9, pp. 1–12, Nov. 2022, doi: [10.1111/exsy.13048](https://doi.org/10.1111/exsy.13048).
- [41] S. I. Ali, H. S. M. Bilal, M. Hussain, J. Hussain, F. A. Satti, M. Hussain, G. H. Park, T. Chung, and S. Lee, "Ensemble feature ranking for cost-based non-overlapping groups: A case study of chronic kidney disease diagnosis in developing countries," *IEEE Access*, vol. 8, pp. 215623–215648, 2020, doi: [10.1109/ACCESS.2020.3040650](https://doi.org/10.1109/ACCESS.2020.3040650).
- [42] P. A. Moreno-Sánchez, "An automated feature selection and classification pipeline to improve explainability of clinical prediction models," in *Proc. IEEE 9th Int. Conf. Healthcare Informat. (ICHI)*, Aug. 2021, pp. 527–534, doi: [10.1109/ICHI52183.2021.00100](https://doi.org/10.1109/ICHI52183.2021.00100).
- [43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and M. Blonde, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.
- [44] S. Shilaskar and A. Ghatol, "Feature selection for medical diagnosis: Evaluation for cardiovascular diseases," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4146–4153, Aug. 2013, doi: [10.1016/j.eswa.2013.01.032](https://doi.org/10.1016/j.eswa.2013.01.032).
- [45] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," in *Data Classification*. Boca Raton, FL, USA: CRC Press, 2014, pp. 37–64, doi: [10.1201/b17320](https://doi.org/10.1201/b17320).
- [46] O. Sagi and L. Rokach, "Ensemble learning: A survey," *WIREs Data Mining Knowl. Discovery*, vol. 8, no. 4, pp. 1–15, Jul. 2018, doi: [10.1002/widm.1249](https://doi.org/10.1002/widm.1249).
- [47] O. Sagi and L. Rokach, "Explainable decision forest: Transforming a decision forest into an interpretable tree," *Inf. Fusion*, vol. 61, pp. 124–138, Sep. 2020, doi: [10.1016/j.inffus.2020.03.013](https://doi.org/10.1016/j.inffus.2020.03.013).
- [48] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.
- [49] 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," 2018, *arXiv:1801.01489*.
- [50] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
- [51] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, *arXiv:1705.07874*.
- [52] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, and S.-I. Lee, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 749–760, Oct. 2018, doi: [10.1038/s41551-018-0304-0](https://doi.org/10.1038/s41551-018-0304-0).
- [53] T. Tagaris and A. Stafylopatis, "Hide-and-seek: A template for explainable AI," 2020, *arXiv:2005.00130*.
- [54] A. Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Oct. 2016, pp. 262–270, doi: [10.1109/ICHI.2016.36](https://doi.org/10.1109/ICHI.2016.36).
- [55] A. Sobrinho, A. C. M. D. S. Queiroz, L. D. Da Silva, E. De Barros Costa, M. E. 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](https://doi.org/10.1109/ACCESS.2020.2971208).



**PEDRO A. MORENO-SÁNCHEZ** received the B.S. degree in telecommunication engineering, the M.S. degree in telemedicine and bioengineering, and the Ph.D. degree in biomedical engineering from the Technical University of Madrid, in 2007, 2008, and 2014, respectively.

Since 2022, he has been a Postdoctoral Research Fellow with Tampere University, Finland. He was a RDI Expert and a Lecturer with the Seinäjoki University of Applied Sciences, Finland. He has been a Researcher with Getafe University Hospital, Spain, and the Technical University of Madrid, since 2007. His research interests include digital health and in the application of artificial intelligence and machine learning to develop clinical prediction models. Currently, his research area is centered in explainable and trustworthy AI in healthcare. He has also collaborated with Horizon Europe Research Program as an Expert Evaluator.

Dr. Moreno-Sánchez is a member of the Finnish Society of Artificial Intelligence.

• • •