IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 15, 2021, accepted December 6, 2021, date of publication December 7, 2021,
date of current version December 22, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3133700

A Machine Learning Analysis of Health Records of Patients With Chronic Kidney Disease at Risk of Cardiovascular Disease

DAVIDE CHICCO[1], CHRISTOPHER A. LOVEJOY[2,3], AND LUCA ONETO[4,5]
[1]Institute of Health Policy Management and Evaluation, University of Toronto, Toronto ON M5T 3M7, Canada
[2]Computer Science Department, University College London, London WC1E 6BT, U.K.
[3]Department of Medicine, University College London Hospital, London NW1 2BU, U.K.
[4]DIBRIS, Università di Genova, 16146 Genoa, Italy
[5]ZenaByte S.r.l., 16121 Genoa, Italy

Corresponding author: Davide Chicco (davidechicco@davidechicco.it)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was obtained by the original dataset curators [28] and granted to them by Tawam Hospital and the United Arab Emirates University Research and Ethics Board under Application No. IRR536/17.

ABSTRACT Chronic kidney disease (CKD) describes a long-term decline in kidney function and has many causes. It affects hundreds of millions of people worldwide every year. It can have a strong negative impact on patients, especially when combined with cardiovascular disease (CVD): patients with both conditions have lower survival chances. In this context, computational intelligence applied to electronic health records can provide insights to physicians that can help them make better decisions about prognoses or therapies. In this study we applied machine learning to medical records of patients with CKD and CVD. First, we predicted if patients develop severe CKD, both including and excluding information about the year it occurred or date of the last visit. Our methods achieved top mean Matthews correlation coefficient (MCC) of +0.499 in the former case and a mean MCC of +0.469 in the latter case. Then, we performed a feature ranking analysis to understand which clinical factors are most important: age, eGFR, and creatinine when the temporal component is absent; hypertension, smoking, and diabetes when the year is present. We then compared our results with the current scientific literature, and discussed the different results obtained when the time feature is excluded or included. Our results show that our computational intelligence approach can provide insights about diagnosis and relative important of different clinical variables that otherwise would be impossible to observe.

INDEX TERMS Machine learning, computational intelligence, feature ranking, electronic health records, chronic kidney disease, CKD, cardiovascular diseases, CVD.

I. INTRODUCTION

Chronic kidney disease (CKD) kills around 1.2 million people and affects more than 700 million people worldwide every year [1]. CKD is commonly caused by diabetes and high blood pressure, and are more likely to be developed in subjects with a family history of CKD.

Individuals with chronic kidney disease are at higher risk of cardiovascular disease (such as myocardial infarction, stroke, heart failure) [2], and patients with both diseases are more likely to have worse prognoses [3].

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

In this context, computational intelligence methods applied to electronic medical records of patients can provide interesting and useful information to doctors and physicians, helping them to more precisely predict the trend of the condition and consequently to make decisions on the therapies. Several studies involving analyses done with machine learning applied to clinical records of patients with CKD have appeared in the biomedical literature in the recent past [4]–[26].

Among the studies found, a large number involves applications of machine learning methods to the Chronic Kidney Disease dataset of the University of California Irvine Machine Learning Repository [27].

On this dataset, Shawan *et al.* [16] and Abrar *et al.* [18] employed several data mining methods for patient classification in their PhD theses. Wibawa *et al.* [8] applied a correlation-based feature selection methods and AdaBoost to this dataset, while Al Imran *et al.* [13] employed deep learning techniques to the same end.

Rashed-al-Mahfuz *et al.* [24] also employed a number of machine learning methods for patient classification and described the dataset precisely. Ali *et al.* [21] applied several machine learning methods to the same dataset to determine a global threshold to discriminate between useful clinical factors and irrelevant ones.

Salekin and Stankovic [6] used Lasso for feature selection, while Belina *et al.* [15] applied a hybrid wrapper and filter based feature selection for the same scope.

Tazin *et al.* [5] employed several data mining methods for patient classification. Ogunleye and Wang [11] used an enhanced XGBoost method for patient classification. Satukumati and Satla [17] used several techniques for feature extraction. Elhoseny *et al.* [19] developed a method called Density based Feature Selection (DFS) with Ant Colony based Optimization (D-ACO) algorithm for the classification of patients with CKD. Polat *et al.* [7] showed an application of a Support Vector Machine variant for patient classification to the same dataset. Chittora *et al.* [22] applied numerous machine learning classifiers and their variants for patient classification. Zeynu and Patil [12] published a survey on computational intelligence methods for binary classification and feature selection applied on the same dataset. Charleonnan *et al.* [4] applied numerous machine learning classifiers and their variants for patient classification. Subasi *et al.* [9] focused on Random Forests for patient classification and feature ranking. Zeynu and Patil [10] applied numerous machine learning classifiers for patient classification and clinical feature selection. All these studies were focused more on the improvement and enhancement of computational intelligence methods, rather than on clinical implications of the results.

Few studies published recently employed datasets different from the UC Irvine ML Repository one. Ventrella *et al.* [23] applied several machine learning methods to an original dataset of EHRs collected at the hospital of Vimercate (Italy) for assessing Chronic Kidney Disease progression. This study indicated creatinine level, urea, red blood cells count, eGFR trend among the most relevant clinical factors for CKD advancement, highlighting that eGFR did not resulted being the top most important one.

Ravizza *et al.* [20] employed machine learning methods on a dataset of patients with diabetes from the IBM Explores database to predict if they will develop CKD. This study states that the usage of diabetes-related data can generate better predictions on data of patients with CKD.

To the best of our knowledge, no study published before involves the usage of machine learning methods to investigate a dataset of patients with both CKD and CVD.

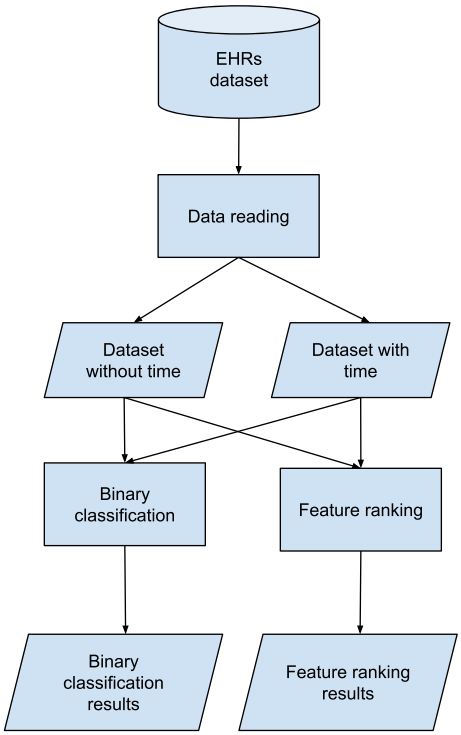

FIGURE 1. Flowchart of the computational pipeline of this study. Cylinder shape: dataset. Rectangular shape: process. Parallelogram shape: input/output.

In this manuscript, we analyzed a dataset of 491 patients from United Arab Emirates, released by Al-Shamsi *et al.* [28] in 2018 (section II). In their original study, the authors employed multivariable Cox's proportional hazards to identify the independent risk factors causing CKD at stages 3-5. Although this analysis was interesting, it did not involve a data mining step, which instead could retrieve additional information or unseen patterns in these data.

To fill this gap, we perform here two analyses: first, we apply machine learning methods to binary classify the serious CKD development, and then to rank the clinical features by importance. Additionally to what Al-Shamsi *et al.* [28] did, we also performed the same analysis excluding the year when the disease happened to each patient (Figure 1).

As major results, we show that computational intelligence is capable of predicting a serious CKD development with or without the time information, and that the most important clinical features change if the temporal component is considered or not.

We organize the rest of the paper as follows. After this Introduction, we describe the dataset we analyzed (section II) and the methods we employed (section III). We then report the binary classification and feature ranking results (section IV) and discuss them afterwards (section V). Finally, we recap the main points of this study and mention limitations and future developments (section VI).

TABLE 1. Meaning, measurement unit, and possible values of each feature of the dataset. ACEI: Angiotensin-converting enzyme inhibitors. ARB: Angiotensin II receptor blockers. mmHg: millimetre of mercury. kg: kilogram. mmol: millimoles.

feature	explanation	measurement unit	values
ACEIARB	if the patient has taken ACEI or ARB	boolean	[0, 1]
AgeBaseline	age of the patient	integer	[23, 24, . . . , 80, 89]
BMIBaseline	body–mass index of the patient	kg/m^2	[13, 16, 17, . . . , 53, 57]
CholesterolBaseline	level of cholesterol	mmol/L	[2.23, 2.40, . . . , 8.20, 9.30]
CreatinineBaseline	level of creatinine in the blood	mol/l	[6, 27, . . . , 113, 123]
dBPBaseline	diastolic blood pressure	mmHg	[41, 45, . . . , 110, 112[
DLDmeds	if the patient has taken dyslipidemia medications	boolean	[0, 1]
DMmeds	if the patient has taken diabetes medications	boolean	[0, 1]
eGFRBaseline	estimated glomerular filtration rate	ml/min/1.73m^2	[60, 60.4, . . . , 242.6]
HistoryCHD	patient history of coronary heart disease	boolean	[0, 1]
HistoryDiabetes	patient history of diabetes	boolean	[0, 1]
HistoryDLD	patient history of dyslipidemia	boolean	[0, 1]
HistoryHTN	patient history of hypertension	boolean	[0, 1]
HistoryObesity	patient history of obesity	boolean	[0, 1]
HistorySmoking	patient history of smoking	boolean	[0, 1]
HistoryVascular	patient history of vascular diseases	boolean	[0, 1]
HTNmeds	if the patient has taken hypertension medications	boolean	[0, 1]
sBPBaseline	systolic blood pressure	mmHg	[92, 95, . . . , 177, 180]
Sex	if the patient is a woman (0) or a man (1)	binary	[0, 1]
time year	year from follow-up start to severe CKD event or last visit	integer	[0, 1, ..., 9, 10]
[target] EventCKD35	if the patient had moderate–extreme CKD	boolean	[0, 1]

II. DATASET

In this study, we examine a dataset of electronic medical records of 491 patients collected at the Tawam Hospital in Al-Ain city (Abu Dhabi, United Arab Emirates), between 1st January and 31st December 2008 [28]. The patients included 241 women and 250 men, with an average age of 53.2 years (Table 2 and Table 3).

Each patient has a chart of 13 clinical variables, expressing her/his values of laboratory tests and exams or data about her/his medical history (Table 1). Each patient included in this study had cardiovascular disease or was at risk of cardiovascular disease, according to the standards of Tawam Hospital [28].

Several features regard the personal history of the patient: diabetes history, dyslipidemia history, hypertension history, obesity history, smoking history, and vascular disease history (Table 2) state if the patient biography had those specific diseases or conditions. Dyslipidemia indicates excessive presence of lipids in the blood. Two variables refer to the blood pressure (diastolic blood pressure and systolic blood pressure), and other variables refer to blood levels obtained through laboratory tests (cholesterol, creatinine). Few features state if the patients have taken specific-disease medicines (dyslipidemia medications, diabetes medications, and hypertension medications) or inhibitors (angiotensin-converting-enzyme inhibitors, or angiotensin II receptor blockers) which are known to be effective against cardiovascular diseases [29] and hypertension [30]. The remaining factors describe the physical conditions of each patient: age, body–mass index, biological sex (Table 2).

Among the clinical features available for this dataset, the EventCKD35 binary variable states if the patient had chronic kidney disease at high stage (3rd, 4th, or 5th stage). According to the Kidney Disease Improving Global Outcomes (KDIGO) organization [31], CKD's can be grouped into 5 stages:

- Stage 1: normal kidney function, no CKD;
- Stage 2: mildly decreased function of kidney, mild CKD;
- Stage 3: moderate decrease of kidney function, moderate CKD;

TABLE 2. Binary features quantitative characteristics. All the binary features have meaning *true* for the value 1 and *false* for the value 0, except sex (0 = female and 1 = male). The dataset contains medical records of 491 patients.

feature	value	#	%
ACEIARB	0	272	55.397
ACEIARB	1	219	44.603
DLDmeds	0	220	44.807
DLDmeds	1	271	55.193
DMmeds	0	330	67.210
DMmeds	1	161	32.790
HistoryCHD	0	446	90.835
HistoryCHD	1	45	9.165
HistoryDiabetes	0	276	56.212
HistoryDiabetes	1	215	43.788
HistoryDLD	0	174	35.438
HistoryDLD	1	317	64.562
HistoryHTN	0	156	31.772
HistoryHTN	1	335	68.228
HistoryObesity	0	243	49.491
HistoryObesity	1	248	50.509
HistorySmoking	0	416	84.725
HistorySmoking	1	75	15.275
HistoryVascular	0	462	94.094
HistoryVascular	1	29	5.906
HTNmeds	0	188	38.289
HTNmeds	1	303	61.711
Sex	0	241	49.084
Sex	1	250	50.916
[target] EventCKD35	0	435	88.595
[target] EventCKD35	1	56	11.405
total		491	100

TABLE 3. Numeric feature quantitative characteristics. σ: standard deviation.

feature	median	mean	range	σ
AgeBaseline	54	53.204	[23, 89]	13.821
BMIBaseline	30	30.183	[13, 57]	6.237
CholesterolBaseline	5	4.979	[2.23, 9.3]	1.097
CreatinineBaseline	66	67.857	[6, 123]	17.919
dBPBaseline	77	76.872	[41, 112]	10.711
eGFRBaseline	98.1	98.116	[60, 242.6]	18.503
sBPBaseline	131	131.375	[92, 180]	15.693
time year	8	7.371	[0, 10]	2.175

- Stage 4: severe decrease of kidney function, severe CKD;
- Stage 5: extreme CKD and kidney failure.

When the EventCKD35 variable has value 0, the patient's kidney condition is at stage 1 or 2. Instead, when EventCKD35 equals to 1, the patient's kidney is at stage 3, 4, or 5 (Table 1).

Even if the value of eGFR has a role to the definition of the CKD stages in the KDIGO guidelines [31], we found weak correlation between the eGFRBaseline variable and the target variable EventCKD35 in this dataset. The two variables have Pearson correlation coefficient equal to -0.36 and Kendall distance of -0.3, both in the $[-1, +1]$ interval where -1 indicates perfectly opposite correlation, 0 indicates no correlation, and $+1$ indicates perfect correlation,

The time year derived factor indicates in which year the patient had a serious chronic kidney disease, or the year when he/she had his/her last outpatient visit, whichever occurred first (Supplementary information),in the follow-up period.

All the dataset features refer to the first visits had by the patients in January 2008, except the EventCKD35 and the time year variables that refer to the end of the follow-up period, in June 2017.

More information about this dataset can be found in the original article [28].

III. METHODS

The problem described earlier (section I) can be addressed as conventional binary classification framework, where the goal is to predict EventCKD35, using the data described earlier (section II). This target feature indicates if the patient has the chronic kidney disease in the stage 3 to 5, which represents an advanced stage.

In binary classification, the problem is to identify the unknown relation $\mathfrak{R}$ between the input space $\mathcal{X}$ (in our case: the features described in Section II) and an output space $\mathcal{Y} \subseteq \{0, 1\}$ (in our case: the EventCKD35 target) [32]. Once a relation is established, one can find a way to discover what the most influencing factors are in the input space for predicting the associated element in the output space, namely to determine the feature importance [33].

Note that, $\mathcal{X}$ can be composed by categorical features (the values of the features belong to a finite unsorted set) and numerical–valued features (the values of the features belong to a possibly infinite sorted set). In case of categorical features, *one-hot encoding* [34] can map them in a series of numerical features. The consequent resulting feature space is $\mathcal{X} \subseteq \mathbb{R}^d$.

A set of data $\mathcal{D}_n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, with $\boldsymbol{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, is available in a binary classification framework. Moreover, some values of $\boldsymbol{x}_i$ might be missing [35]. In this case, if the missing value is categorical, we introduce an additional category for missing values for the specific feature. Instead, if the missing value is associated with a numerical feature, we replace the missing value with the mean value of the specific feature, and we introduce an additional logical feature to indicate if the value of the feature is missing for a particular sample [35].

Our goal is to identify a model $\mathfrak{M} : \mathcal{X} \rightarrow \mathcal{Y}$, which best approximates $\mathfrak{R}$, through an algorithm $\mathscr{A}_{\mathcal{H}}$ characterized by its set of hyper-parameters $\mathcal{H}$. The accuracy of the model $\mathfrak{M}$ to represent the unknown relation $\mathfrak{R}$ is measured using different indices of performance (Supplementary information).

Since the hyper-parameters $\mathcal{H}$ influence the ability of $\mathscr{A}_{\mathcal{H}}$ to estimate $\mathfrak{R}$, we need to adopt a proper Model Selection (MS) procedure [36]. In this work, we exploited the Complete Cross Validation (CCV) procedure [36]. CCV relies on a simple idea: we resample the original dataset $\mathcal{D}_n$ many ($n_r = 500$) times without replacement to build a training set of size l $\mathcal{L}_l^r$ while the remaining samples are kept in the validation set $\mathcal{V}_v^r$, with $r \in \{1, \cdots, n_r\}$. In order to perform the MS phase, to select the best combination of the hyper-parameters $\mathcal{H}$ in the set of possible ones $\mathfrak{H} = \{\mathcal{H}_1, \mathcal{H}_2, \cdots\}$ using the algorithm $\mathscr{A}_{\mathcal{H}}$, the hyper-parameters which minimize the average performance of the model, trained on the training set, and evaluated on the validation set, should be selected. Since the data in $\mathcal{L}_l^r$ are independent from the ones in $\mathcal{V}_v^r$, the idea is that $\mathcal{H}^*$ should be the set of hyper-parameters which allows to achieve a small error on a data set that is independent from the training set.

Finally, we need to estimate the error (EE) of the optimal model with a separate set of data $\mathcal{T}_m = \{(\boldsymbol{x}_1^t, y_1^t), \cdots, (\boldsymbol{x}_m^t, y_m^t)\}$ since the error that our model commits over $\mathcal{D}_n$ would be optimistically biased since $\mathcal{D}_n$ has been used to find $\mathfrak{M}$.

Additionally, another aspect to consider in this analysis is that data available in health informatics are often unbalanced [37]–[39], and most learning algorithms do not work well with imbalanced datasets and tend to poorly perform on the minority class. For these reasons, several techniques have been developed in order to address this issue [40]. Currently the most practical and effective method involves the resampling of the data in order to synthesize a balanced dataset [40]. For this purpose, we can under-sample or over-sample the dataset. Under-sampling balances the dataset by reducing the size of the abundant class. By keeping all samples in the rare class and randomly selecting an equal number of samples in the abundant class, a new balanced dataset can be retrieved for further modeling. Note that this method wastes a lot of information (many

samples might be discarded). For this reason, scientists take advantage of the over-sampling strategy more often. Over-sample tries to balance the dataset by increasing the size of rare samples. Rather than removing abundant samples, new rare samples are generated (for example by repetition, by bootstrapping, or by synthetic minority). The latter method is the one that we employed in this study: synthetic minority oversampling [41], [42].

Another important property of $\mathfrak{M}$ is its interpretability, namely the possibility to understand how it behaves. There are two options to investigate this property. The first one is to learn a $\mathfrak{M}$ such that its functional form is, by construction, interpretable [43] (for example, Decision Trees and Rule based models); this solution, however, usually results in poor generalization performances. The second one, used when the functional form of $\mathfrak{M}$ is not interpretable by construction [43] (for example, Kernel Methods or Neural Network), is to derive its interpretability a posteriori. A classical method for reaching this goal is to perform a feature ranking procedure [33], [44] which gives an hint to the users of $\mathfrak{M}$ about the most important features which influence its results.

A. BINARY CLASSIFICATION ALGORITHMS

In this paper, for the $\mathscr{A}$, we will exploit different state-of-the-art models. In particular we will exploit Random Forests [45], Support Vector Machines (linear and kernelized with the Gaussian Kernel) [46], [47], Neural Network [48], Decision Tree [49], XGBoost [50], and One Rule [51].

We tried a number of different hyper-parameter configurations for the machine learning methods employed in this study.

For Random Forests, we set the number of trees to 1000 and we searched number of variables randomly sampled as candidates at each split in $\{1, 2, 4, 8, 16\}$, the minimum size of samples in the terminal nodes of the trees in $\{1, 2, 4, 8\}$, the percentage samples (sampled with bootstrap) during the creation of each tree in $\{60, 80, 100, 120\}$ [52]–[55]. For the linear and kernelized Support Vector Machines [46], we searched the regularization hyper-parameters in $\{10^{-6.0,-5.8,\cdots,4}\}$ and, for the kernelized Support Vector Machines, we used the Gaussian Kernel [47] and we searched the kernel hyper-parameters in $\{10^{-6.0,-5.8,\cdots,4}\}$. For the Neural Network we used a single hidden layer network (hyperbolic tangent as activation function in the hidden layer) with dropout (`mlpKerasDropout` in the caret [56] R package), we train it with adaptive subgradient methods (batch size equal to 32), and we tuned the following hyper-parameters: the number of neurons in the hidden layer in $\{10, 20, 40, 80, 160, 320, 640, 1280\}$, the dropout rate of the hidden layer in $\{0.001, 0.002, 0.004, 0.008\}$, the learning rate in $\{0.001, 0.002, 0.005, 0.01, 0.02.0.05\}$, the fraction of gradient to keep at each step in $\{0.01, 0.05, 0.1, 0.5\}$, and the learning rate decay in $\{0.01, 0.05, 0.1, 0.5\}$. For Decision Tree we searched the max depth of the trees in $\{4, 8, 16, 24, 32\}$ (`rpart2` in the caret [56] R package). For XGBoost we set tree gradient boosting and we searched the Booster Parameters in $\{0.001, 0.002, 0.004, 0.008, 0.01, 0.02, 0.04, 0.08\}$ the number of trees in $\{100, 500, 1000\}$, the minimum loss reduction to make a split in $\{0, 0.001, 0.005, 0.01\}$, the fraction of samples in $\{1, 0.9, 0.7\}$ and features $\{1, 0.5, 0.2, 0.1\}$ used train the trees and the maxim number of leaves in $\{1, 2, 4, 8, 16\}$, and the regularization hyper-parameters in $\{10^{-6.0,-5.8,\cdots,4}\}$ [50]. For One Rule we did not have to tune hyper-parameters (`OneR` in the caret [56] R package).

Note that these methods have shown to be a set of the simplest yet best performing methods available in scientific literature [57], [58]. The difference between the methods is just the functional form of the model which tries to better approximate a learning principle.

For example, Random Forests and XGBoost try to implement the wisdom of the crowd principles, Support Vector Machines are robust maximum margin classifiers, and Decision Tree and One Rule represent very easy to interpret models. In this paper we tested multiple algorithms since the no-free-lunch theorem [59] assures us that, for a specific application, it is not possible to know, a-priori, what algorithm will better perform on a specific task. Then we tested the ones which, in the past, have shown to perform well on many tasks and identified the best one for our application.

B. FEATURE RANKING

Feature rankings methods based on Random Forests are among the most effective techniques [60], [61], particularly in the context of bioinformatics [62], [63] and health informatics [64]. Since Random Forests obtained the top prediction scores for binary classification, we focus on this method for feature ranking.

Several measures are available for feature importance in Random Forests. A powerful approach is the one based on the Permutation Importance or Mean Decrease in Accuracy (MDA), where the importance is assessed for each feature by removing the association between that feature and the target. This effect is achieved by randomly permuting [65] the values of the feature and measuring the resulting increase in error. The influence of the correlated features is also removed.

In details, for every tree, the method computes two quantities: the first one is the error on the out-of-bag samples as they are used during prediction, while the second one is the error on the out-of-bag samples after a random permutation of the values of a variable. These two values are then subtracted and the average of the result over all the trees in the ensemble is the raw importance score for the variable under exam.

Despite the effectiveness of MDA, when the number of samples is small these methods might result being unstable [66]–[68]. For this reason, in this work, instead of running the Feature Ranking (FR) procedure just once, analogously to what we have done for MS and EE, we subsample the original dataset and we repeat the procedure many

TABLE 4. CKD development binary classification results. Linear SVM: Support Vector Machine with linear kernel. Gaussian SVM: Support Vector Machine with Gaussian kernel. MCC: Matthews correlation coefficient (worst value = −1 and best value = +1). TP rate: true positive rate, sensitivity, recall. TN rate: true negative rate, specificity. PR: precision-recall curve. PPV: positive predictive value, precision. NPV: negative predictive value. ROC: receiver operating characteristic curve. AUC: area under the curve. F_1 score, accuracy, TP rate, TN rate, PPV, NPV, PR AUC, ROC AUC: worst value = 0 and best value = +1. Confusion matrix threshold for TP rate, TN rate, PPV, and NPV: 0.5. We highlighted in blue and with an asterisk * the top results for each score. We report the formulas of these rates in the Supplementary Information.

method	MCC	F_1 score	accuracy	TP rate	TN rate
Random Forests	***+0.501** ± 0.035	*0.550 ± 0.034	*0.843 ± 0.012	*0.793 ± 0.038	0.852 ± 0.012
Gaussian SVM	**+0.319** ± 0.065	0.387 ± 0.063	0.873 ± 0.018	0.353 ± 0.082	*0.940 ± 0.009
Neural Network	**+0.302** ± 0.075	0.353 ± 0.065	0.840 ± 0.028	0.436 ± 0.173	0.882 ± 0.048
Linear SVM	**+0.266** ± 0.113	0.340 ± 0.077	0.767 ± 0.032	0.600 ± 0.193	0.785 ± 0.032
Decision Tree	**+0.253** ± 0.085	0.345 ± 0.078	0.747 ± 0.036	0.588 ± 0.079	0.767 ± 0.036
XGBoost	**+0.160** ± 0.033	0.286 ± 0.033	0.767 ± 0.035	0.368 ± 0.056	0.824 ± 0.044
One Rule	**+0.145** ± 0.063	0.267 ± 0.044	0.707 ± 0.034	0.471 ± 0.094	0.737 ± 0.035

	PPV	NPV	PR AUC	ROC AUC
Random Forests	0.422 ± 0.034	*0.971 ± 0.006	*0.475 ± 0.046	*0.885 ± 0.015
Gaussian SVM	*0.429 ± 0.068	0.919 ± 0.021	0.275 ± 0.032	0.646 ± 0.040
Neural Network	0.313 ± 0.046	0.912 ± 0.052	0.257 ± 0.032	0.669 ± 0.066
Linear SVM	0.237 ± 0.051	0.946 ± 0.023	0.280 ± 0.018	0.693 ± 0.095
Decision Tree	0.244 ± 0.067	0.936 ± 0.015	0.274 ± 0.013	0.678 ± 0.050
XGBoost	0.233 ± 0.036	0.900 ± 0.012	0.267 ± 0.017	0.596 ± 0.019
One Rule	0.186 ± 0.035	0.916 ± 0.021	0.179 ± 0.017	0.604 ± 0.051

times. The final rank of a feature will be the aggregation of the different ranking using the Borda's method [69].

C. BIOSTATISTICS UNIVARIATE TESTS

Before employing machine learning algorithms, we applied traditional univariate biostatistics techniques to evaluate the relationship between the EventCKD35 target and each feature.

We made use of the Mann–Whitney U test (also known as Wilcoxon rank–sum test) [70] for the numerical features and of the chi–square test [71] for the binary features. The p-values of both these tests range between 0 and 1: a low p-value of this test means that the analyzed variable strongly relates to the target feature, while a high p-value means the no evident relation. These tests are also useful to detect the importance of each feature with respect to the target: the lower the p-value of a feature, the stronger its association with the target. Following the recent advice of Benjamin *et al.* [72], we use 0.005 as threshold of significance for the p-values, that is 5×10^{-3}. If the p-value of a test applied to a variable and the target results being lower than 0.005, we consider significant the association between the variable and the target.

D. PREDICTION AND FEATURE RANKING INCLUDING TEMPORAL FEATURE

In the second analysis we performed for chronic kidney disease prediction, we decided to include the temporal component expressing in which year the disease occurred for the CKD patients or which year they had their last outpatient visit (Supplementary information).

We applied a Stratified Logistic Regression [73], [74] to this complete dataset, including all the original clinical features and the derived year feature, both for supervised binary classification and feature ranking. We measured the prediction with the typical confusion matrix rates (MCC, F_1 score, and others), and the importance for each variable as the logistic regression model coefficient. This method has no significant hyper-parameters so we did not perform any optimization (`glm` method of the stats R package).

IV. RESULTS

In this section, we report the results for the prediction of the chronic kidney disease (subsection IV-A) and its feature ranking (subsection IV-B).

A. CHRONIC KIDNEY DISEASE PREDICTION RESULTS

1) CKD PREDICTION

We report the results obtained for the static prediction of the CKD measured with traditional confusion matrix indicators in Table 4. We rank our results by the Matthews correlation coefficient (MCC) because it is the only confusion matrix rate that generates a high score if the classifier was able to correctly predict most of the data instances and correctly make most of the predictions, both on the positive class and the negative class [75]–[78].

Random Forests outperformed all the other methods for MCC, F_1 score, accuracy, sensitivity, negative predictive value, precision recall AUC, and receiver operating characteristic AUC (Table 4), while the support vector machine with Gaussian kernel achieved the top specificity and precision.

Because of the imbalance of the dataset (section II), all the classifiers attained better results among the negative data instances (specificity and NPV) than among the positive elements (sensitivity and precision). This consequence

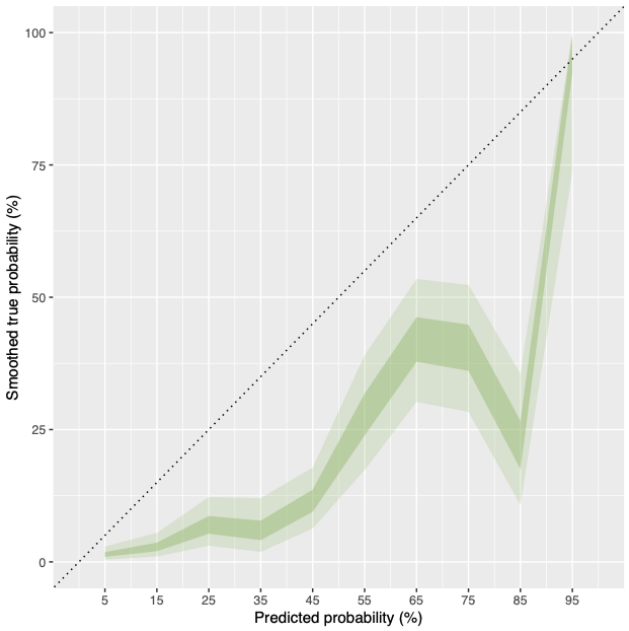

FIGURE 2. Calibration curve and plots for the results obtained by Random Forests predictions applied on the dataset excluding the temporal component (Table 4).

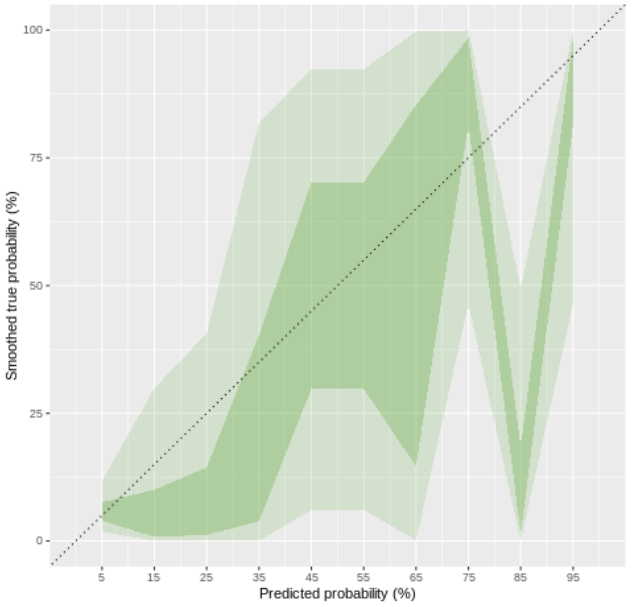

FIGURE 3. Calibration plot for the Stratified Logistic Regression predictions applied on the dataset including the temporal component (Table 5).

happens because each classifier can observe and learn to recognize more individuals without CKD during training, and therefore are more capable of recognizing them than recognizing patients with CKD during testing.

XGBoost and One Rule obtained Matthews correlation coefficients close to 0, meaning that their performance was similar to random guessing. Random Forests, linear SVM, and Decision Tree were the only methods able to correctly classify most of the true positives (TP rate = 0.792, 0.6, and 0.588, respectively). No technique was capable of correctly making most of the positive predictions: all PPVs are below 0.5 Table 4.

Regarding positives, SVM with Gaussian kernel obtained an almost perfect specificity (0.940), while Random Forests achieved an almost perfect NPV of 0.968 Table 4.

These results show that the machine learning classifiers Random Forests and SVM with Gaussian kernel can efficiently predict patients with CKD and patients without CKD from their electronic health records, with high prediction scores, in few minutes.

Since Random Forests resulted being the best performing classifier, we also included the calibration curve plot [79] of its predictions (Figure 2), for the sake of completeness. The curve follows the trend of the $x = y$ perfect line translated on the x axis between approximately 5% and approximately 65%, indicating well calibrated predictions in this interval.

2) CKD PREDICTION EXCLUDING TEMPORAL COMPONENT

To show a scenario where no previous disease history of a patient is available, we did not include any temporal component providing information about the progress of the disease in the previous analysis. We then decided to performed a stratified prediction including a time feature indicating the year when the patient developed the chronic kidney disease, or the last visit for non-CKD patients (Supplementary information). After having included the year information in the dataset, we applied a Stratified Logistic Regression [74], [80], as described earlier (section III).

The presence of the temporal feature actually improved the prediction, allowing the regression to obtain a MCC of +0.469, better than all the MCC's achieved by the classifiers applied to the static dataset version except Random Forests (Table 5). Also in this case, sensitivity and precision result being much higher than sensitivity and NPV, because of the imbalance of the dataset.

This result comes with no surprise: it makes complete sense that the inclusion of a temporal feature describing the trend of a disease could improve the prediction quality.

To better understand the prediction obtained by the Stratified Logistic Regression, we plotted a calibration curve [79] of its predictions (Figure 3). As one can notice, the Stratified Logistic Regression returns well calibrated predictions, as it trends follows the $x = y$ line which represents the perfect calibration from approximately 5% to approximately 75% of the probabilities. This calibration curve confirms that the Stratified Logistic Regression made a good prediction.

B. FEATURE RANKING RESULTS

1) CKD PREDICTIVE FEATURE RANKING

After verifying that computational intelligence is able to predict CKD developments among patients, we applied

TABLE 5. CKD prediction results including the temporal feature. The dataset analyzed for these tests contains the time year feature indicating in which year after the baseline visits the patient developed the CKD. All the abbreviations have the same meaning described in the caption of Table 4.

method	MCC	F_1 score	accuracy	TP rate	TN rate
Stratified Logistic Regression	**+0.469** $\pm$ 0.141	0.507 $\pm$ 0.130	0.903 $\pm$ 0.031	0.604 $\pm$ 0.177	0.933 $\pm$ 0.025
	PPV	**NPV**	**PR AUC**	**ROC AUC**	
Stratified Logistic Regression	0.458 $\pm$ 0.141	0.960 $\pm$ 0.024	0.345 $\pm$ 0.122	0.768 $\pm$ 0.089	

TABLE 6. Feature ranking through biostatistics univariate tests. We employed the Mann–Whitney *U* test [70] for the numerical features and the chi–square test [71] for the binary features. We reported in blue and with an asterisk * the only feature having a *p*-value lower than the 0.005 threshold, that is 5×10^{-03}.

position	feature	Mann-Whitney U test p-value
1	*AgeBaseline	0
2	*CreatinineBaseline	0
3	*eGFRBaseline	0
4	*CholesterolBaseline	9.490×10^{-04}
5	*sBPBaseline	4.379×10^{-03}
6	dBPBaseline	1.083×10^{-01}
7	BMIBaseline	9.134×10^{-01}
position	**feature**	**chi–squared p-value**
1	*HistoryDiabetes	5×10^{-04}
2	*HistoryCHD	5×10^{-04}
3	*HistoryHTN	5×10^{-04}
4	*DLDmeds	5×10^{-04}
5	*DMmeds	5×10^{-04}
6	*ACEIARB	5×10^{-04}
7	*HistoryDLD	1.999×10^{-03}
8	*HTNmeds	1.999×10^{-03}
9	HistoryVascular	3.698×10^{-02}
10	Sex	4.398×10^{-02}
11	HistorySmoking	5.397×10^{-02}
12	HistoryObesity	4.948×10^{-01}

a feature ranking approach to detect the most predictive features in the clinical records. We employed two techniques: one based on traditional univariate biostatistics tests, and one based on machine learning.

Regarding the biostatistics phase, applied the Mann–Whitney test and of chi-squared test to each variable in relationship with the CKD target (subsection III-C), and ranked the features by *p*-value (Table 6).

The application of these biostatistics univariate tests, although useful, show a huge number of relevant variables: 13 variable of out 19 result being significant, having a *p*-value smaller than 0.005 (Table 6). Since the biostatistics tests affirm that 68.42% of clinical factors are important, this information does not help us to detect the relevance of the features with enough precision. For this reason, we decided to calculate the feature ranking with machine learning, by employing Random Forests, which is the method that achieved the top performance results in the binary classification earlier (subsection IV-A).

We therefore applied the Random Forests feature ranking, and ranked the results by mean accuracy decrease position (Table 7 and Figure 4).

TABLE 7. Feature ranking generated by Random Forests. MDA average position: average position obtained by each feature through the accuracy decrease feature ranking of Random Forests.

position	MDA average position	feature
1	1.2	AgeBaseline
2	1.8	eGFRBaseline
3	3.3	DMmeds
4	3.7	dBPBaseline
5	5.2	CholesterolBaseline
6	6.0	HistoryVascular
7	7.0	HistoryCHD
8	8.3	sBPBaseline
9	8.7	CreatinineBaseline
10	11.4	HistoryHTN
11	11.6	HistorySmoking
12	11.9	DLDmeds
13	12.1	Sex
14	13.4	HTNmeds
15	14.6	HistoryObesity
16	15.9	HistoryDLD
17	17.4	ACEIARB
18	17.7	BMIBaseline
19	18.8	HistoryDiabetes

The two rankings show some common aspects, both listing AgeBaseline and eGFRBaseline in top positions, but show also some significant differences. The biostatistics standing, for example, lists dBPBaseline as unrelevant predictive feature (Table 6), while Random Forests puts it on the 4^{th} position out of 19 (Table 7). Also, the biostatistics tests stated that HistoryDiabetes is one of the most significant factors, with *p*-value of 0.0005 (Table 6), while the machine learning approach put the same feature on the last position of its ranking.

The two rankings contain other minor differences that we consider unimportant.

2) CKD PREDICTIVE FEATURE RANKING CONSIDERING THE TEMPORAL COMPONENT

As we did early for the CKD prediction, we decided to re-run the feature ranking procedure by including the temporal component regarding the year when the patient developed chronic kidney disease or the year of the last visit. Again, we employed Stratified Logistic Regression.

The ranking generated considering the time component (Table 8) showed several differences with respect to the previously described ranking generated without it (Table 7). The most relevant differences in ranking positions are the following:

- HTNmeds is at the 1^{st} position in this ranking, while it is 14^{th} without considering time;

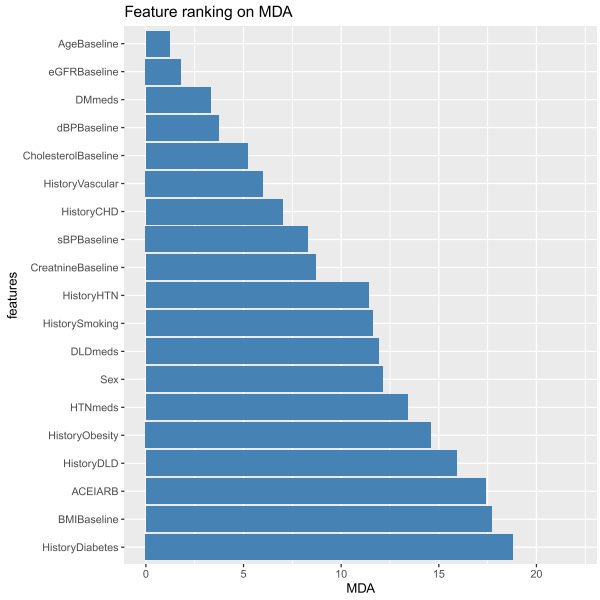

FIGURE 4. Barplot of the Random Forests feature ranking. MDA average position: average position obtained by each feature through the accuracy decrease feature ranking of Random Forests.

- HistoryHTN is at the 3^{rd} position in this ranking, while it is 10^{th} without considering time;
- ACEIARB is at the 4^{th} position in this ranking, while it is 17^{th} without considering time;
- AgeBaseline is at the last position in this ranking, while it is 1^{st} without considering time;
- CreatinineBaseline is at the 18^{th} position in this ranking, while it is 9^{th} without considering time.

We also decided to measure the difference between these two rankings through two traditional metrics such as Spearman's rank correlation coefficient and Kendall distance [81]–[83]. Both these metrics range between –1.0 and +1, with –1 meaning opposite rank orders, 0.0 meaning no correlation between lists, and +1.0 meaning identical ranking.

The comparison between ranking without time (Table 7) and ranking considering time (Table 8) generated Spearman's $\rho = -0.209$ and Kendall $\tau = -0.146$.

V. DISCUSSION

A. CKD PREDICTION

Our results show that machine learning methods are capable of predicting chronic kidney disease from medical records of patients at risk of cardiovascular disease, both including the temporal information about the year when the patient has developed the CKD and without it. These findings can have an immediate impact in the clinical settings: physicians, in fact, can take advantage of our methods to forecast the likelihood of a patient having chronic kidney disease, in a few minutes, and then use this information to establish the urgency of the case. Our techniques, of course, do not replace laboratory exams and tests, that will still be needed to further verify and understand the prognosis of the disease. However, if used efficiently, our methods will provide quick, reliable, fast information to physicians to help them with medical decision making.

TABLE 8. Clinical feature ranking generated by the Stratified Logistic Regression, depending on the temporal component (the year when the CKD happened or of patient's last visit). Importance: average coefficient of the trained logistic regression model out of 100 executions.

position	importance	clinical feature
1	2.56	HTNmeds
2	2.43	dBPBaseline
3	2.32	HistoryHTN
4	1.85	ACEIARB
5	1.68	HistorySmoking
6	1.52	HistoryDiabetes
7	1.42	sBPBaseline
8	1.31	BMIBaseline
9	0.88	CholesterolBaseline
10	0.87	HistoryCHD
11	0.80	eGFRBaseline
12	0.45	DMmeds
13	0.35	Sex
14	0.19	HistoryVascular
15	0.18	HistoryObesity
16	0.16	HistoryDLD
17	0.14	DLDmeds
18	0.01	CreatinineBaseline
19	0.00	AgeBaseline

B. FEATURE RANKING

As mentioned earlier (subsection IV-B), some significant differences emerge between the feature ranking obtained without the time component and generated through Random Forests (Table 7) and the feature ranking obtained considering the year when the patient had the serious CKD development and generated through Stratified Logistic Regression (Table 8).

The features HTNmeds, ACEIARB, and HistoryDiabetes had an increase of 13 positions in the year standing (Table 8), compared to their original position in the static ranking (Table 7). Also, the feature BMIBaseline had an increase, of 10 positions. The AgeBaseline variable, instead, had the biggest position drop possible: it moved from the most important feature in the static standing (Table 7) to the less relevant position in the year standing (Table 8). The other variables in the year standing did not show so high position changes.

These results show that taking medication for hypertension, taking ACE inhibitors, having a personal history of diabetes, and body–mass index have an important role in predicting if a patient will have serious CKD, when the information about the disease event is included. The age of the patient is very important when the CKD year is unknown, but becomes irrelevant here.

C. DIFFERENCE BETWEEN TEMPORAL FEATURE RANKING AND NON-TEMPORAL FEATURE RANKING

The significant differences that emerge suggest strong overlap between the information contained within the time variable with certain variables in the previous model. It is

plausible that some predictors encode a 'baseline' level of risk of developing CKD, which is negated if the model knows in which year the CKD developed.

The variables which reduce most significantly between the models are age, eGFR and creatinine, which are all clinical indicators of an individual's baseline risk of CKD. Inspection of variables which maintain or increase their position when the year feature is added identifies hypertension, smoking and diabetes as key predictive factors in the model (subsection IV-B). These are all known to play a central role in the pathogenesis of micro- and macrovascular disease, including of the kidney. While the former variables may encode baseline risk, the latter are stronger indicators for rate of progression.

It is also worth noting that without the temporal information, the model is tasked with predicting whether the individual will develop CKD within the next 10 years. Here, the baseline is highly relevant as it indicates how much further the renal function needs to deteriorate. However, when the configuration is altered to include the year in which year the CKD developed, the relative importance of risk factors may be expected to increase – and indeed, we observed this in our models.

D. COMPARISON WITH RESULTS OF THE ORIGINAL STUDY

The original study of Al-Shamsi *et al.* [28] included a feature ranking phase generated through a multivariable Cox's proportional hazards analysis, which included the temporal component [84]. Their ranking listed older age (AgeBaseline), personal history of coronary heart disease (HistoryCHD), personal history of diabetes mellitus (HistoryDLD), and personal history of smoking (HistorySmoking) as most important factors for risk of CKD serious event.

In contrast to their findings, AgeBaseline was ranked in the last position in our Stratified Logistic Regression standing, while HistoryCHD and HistoryDLD were at unimportant positions: 10^{th} and 16^{th} ranks out of 19 variables, respectively. Smoking history, instead, occupied a high rank both in our standing and in the original study standing: our approach, in fact, listed it as 5^{th} out of 19.

E. COMPARISON WITH RESULTS OF OTHER STUDIES

Several published studies include a feature ranking phase to detect the most relevant variables to predict chronic kidney disease from electronic medical records. Most of them, however, use feature ranking to reduce the number of variables for the binary classification, without reporting a final standing of clinical factors ranked by importance [10], [12], [21].

Only the article of Salekin and Stankovic [6] reports the most relevant variables found in their study: specific gravity, albumin, diabetes, hypertension, hemoglobin, serum creatinine, red blood cells count, packed cell volume, appetite, and sodium resulted being at top positions. Even if the clinical features present in our datasets mainly differ from theirs, we can notice the difference in the ranking positions between the two studies.

Hypertension resulted being the 4^{th} most important factor in Salekin's study [6], confirming the importance of the HistoryHTN variable which is ranked at the 3^{rd} position in our Stratified Logistic Regression ranking (Table 8). Also diabetes history has high ranking in both the standings: 3^{rd} position in the ranking of Salekin's study [6], and 6^{th} of importance in our Stratified Logistic Regression ranking, as HistoryDiabetes (Table 8).

VI. CONCLUSION

Chronic kidney disease affects more than 700 millions people in the world annually, and kills approximately 1.2 million of them. Computational intelligence can be an effective means to quickly analyze electronic health records of patients affected by this disease, providing information about how likely they will develop severe stages of this disease, or stating which clinical variables are the most important for diagnosis.

In this article, we analyzed a medical record dataset of 491 patients from UAE with CKD and at risk of cardiovascular disease, and developed machine learning methods able to predict the likelihood they will develop CKD at stages 3-5, with high accuracy. Afterwards, we employed machine learning to detect the most important variables contained in the dataset, first excluding the temporal component indicating the year when the CKD happened or the patient's last visit, and then including it. Our results confirmed the effectiveness of our approach.

Regarding limitations, we have to report that we performed our analysis only on a single dataset. We looked for alternative public datasets to use as validation cohorts, but unfortunately we could not find any that have the same clinical features.

In the future, we plan to further investigate the probability of diagnosis prediction in this dataset through classifier calibration and calibration plots [85], and to perform the feature ranking with a different feature ranking method such as SHapley Additive exPlanations (SHAP) [86]. Moreover, we also plan to study chronic kidney disease by applying our methods to CKD datasets of other types, such as microarray gene expression [87], [88] and ultrasonography images [89].

LIST OF ABBREVIATIONS

AUC: area under the curve. BP: blood pressure. CHD: coronary hearth disease. CKD: chronic kidney disease. CVD: cardiovascular disease. DLD: dyslipidemia. EE: error estimation. FR: feature ranking. KDIGO: Kidney Disease Improving Global Outcomes. HTN: hypertension. MCC: Matthews correlation coefficient. MDA: Model Decrease in Accuracy. MS: model selection. NPV: negative predictive value. *p*-value: probability value. PPV: positive predictive value. PR: precision–recall. ROC: receiver operating characteristic. SHAP: SHapley Additive exPlanations. SVM: Support Vector Machine. TN rate: true negative rate. TP rate: true positive rate. UAE: United Arab Emirates.

COMPETING INTERESTS

The authors declare they have no competing interest.

ACKNOWLEDGMENT

The authors thank Saif Al-Shamsi (United Arab Emirates University) for having provided additional information about the dataset.

DATA AND SOFTWARE AVAILABILITY

The dataset used in this study is publicly available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license at: https://figshare.com/articles/dataset/Chronic_kidney_disease_in_patients_at_high_risk_of_cardiovascular_disease_in_the_United_Arab_Emirates_A_population-based_study/6711155?file=12242270

Our software code is publicly available under GNU General Public License v3.0 at: https://github.com/davidechicco/chronic_kidney_disease_and_cardiovascular_disease

REFERENCES

[1] V. A. Luyckx, M. Tonelli, and J. W. Stanifer, "The global burden of kidney disease and the sustainable development goals," *Bull. World Health Org.*, vol. 96, no. 6, p. 414, 2018.
[2] S. Said and G. T. Hernandez, "The link between chronic kidney disease and cardiovascular disease," *J. Nephropathol.*, vol. 3, no. 3, p. 99, 2014.
[3] K. Damman, M. A. E. Valente, A. A. Voors, C. M. O'Connor, D. J. van Veldhuisen, and H. L. Hillege, "Renal impairment, worsening renal function, and outcome in patients with heart failure: An updated meta-analysis," *Eur. Heart J.*, vol. 35, no. 7, pp. 455–469, Feb. 2014.
[4] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," in *Proc. Manage. Innov. Technol. Int. Conf. (MITicon)*, Bang-Saen, Thailand, Oct. 2016, pp. 80–83.
[5] N. Tazin, S. A. Sabab, and M. T. Chowdhury, "Diagnosis of chronic kidney disease using effective classification and feature selection technique," in *Proc. Int. Conf. Med. Eng., Health Informat. Technol. (MediTec)*, Dhaka, Bangladesh, Dec. 2016, pp. 1–6.
[6] A. Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Chicago, IL, USA, Oct. 2016, pp. 262–270.
[7] 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, no. 4, p. 55, 2017.
[8] M. S. Wibawa, I. M. D. Maysanjaya, and I. M. A. W. Putra, "Boosted classifier and features selection for enhancing chronic kidney disease diagnose," in *Proc. 5th Int. Conf. Cyber IT Service Manage. (CITSM)*, Denpasar, Indonesia, Aug. 2017, pp. 1–6.
[9] A. Subasi, E. Alickovic, and J. Kevric, "Diagnosis of chronic kidney disease by using random forest," in *Proc. Int. Conf. Med. Biol. Eng. (CMBEBIH)*. Singapore: Springer, 2017, pp. 589–594.
[10] S. Zeynu and S. Patil, "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.
[11] A. Ogunleye and Q.-G. Wang, "Enhanced XGBoost-based automatic diagnosis system for chronic kidney disease," in *Proc. IEEE 14th Int. Conf. Control Autom. (ICCA)*, Anchorage, AK, USA, Jun. 2018, pp. 805–810.
[12] S. Zeynu and S. Patil, "Survey on prediction of chronic kidney disease using data mining classification techniques and feature selection," *Int. J. Pure Appl. Math.*, vol. 118, no. 8, pp. 149–156, 2018.
[13] A. A. Imran, M. N. Amin, and F. T. Johora, "Classification of chronic kidney disease using logistic regression, feedforward neural network and wide & deep learning," in *Proc. Int. Conf. Innov. Eng. Technol. (ICIET)*, Osaka, Japan, Dec. 2018, pp. 1–6.
[14] A. Shrivas, S. K. Sahu, and H. Hota, "Classification of chronic kidney disease with proposed union based feature selection technique," in *Proc. 3rd Int. Conf. Internet Things Connected Technol.*, Jaipur, India, 2018, pp. 26–27.
[15] S. Belina V. J. Sara and K. Kalaiselvi, "Ensemble swarm behaviour based feature selection and support vector machine classifier for chronic kidney disease prediction," *Int. J. Eng. Technol.*, vol. 7, no. 2, p. 190, May 2018.
[16] N. R. Shawan, S. S. A. Mehrab, F. Ahmed, and A. S. Hasmi, "Chronic kidney disease detection using ensemble classifiers and feature set reduction," Ph.D. dissertation, Dept. Comput. Sci. Eng., BRAC Univ., Dhaka, Bangladesh, 2019.
[17] S. B. Satukumati and R. K. S. Satla, "Feature extraction techniques for chronic kidney disease identification," *Kidney*, vol. 24, no. 1, p. 29, 2019.
[18] T. Abrar, S. Tasnim, and M. Hossain, "Early detection of chronic kidney disease using machine learning," Ph.D. dissertation, Dept. Comput. Sci. Eng., BRAC Univ., Dhaka, Bangladesh, 2019.
[19] M. Elhoseny, K. Shankar, and J. Uthayakumar, "Intelligent diagnostic prediction and classification system for chronic kidney disease," *Sci. Rep.*, vol. 9, no. 1, pp. 1–14, Dec. 2019.
[20] S. Ravizza, T. Huschto, A. Adamov, L. Böhm, A. Büsser, F. F. Flöther, R. Hinzmann, H. König, S. M. McAhren, D. H. Robertson, T. Schleyer, B. Schneidinger, and W. Petrich, "Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data," *Nature Med.*, vol. 25, no. 1, pp. 57–59, Jan. 2019.
[21] S. I. 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.
[22] P. Chittora, S. Chaurasia, P. Chakrabarti, G. Kumawat, T. Chakrabarti, Z. Leonowicz, M. Jasiński, Ł. Jasiński, R. Gono, E. Jasińska, and V. Bolshev, "Prediction of chronic kidney disease—A machine learning perspective," *IEEE Access*, vol. 9, pp. 17312–17334, 2021.
[23] P. Ventrella, G. Delgrossi, G. Ferrario, M. Righetti, and M. Masseroli, "Supervised machine learning for the assessment of chronic kidney disease advancement," *Comput. Methods Programs Biomed.*, vol. 209, Sep. 2021, Art. no. 106329.
[24] M. Rashed-Al-Mahfuz, A. Haque, A. Azad, S. A. Alyami, J. M. W. Quinn, and M. A. Moni, "Clinically applicable machine learning approaches to identify attributes of chronic kidney disease (CKD) for use in low-cost diagnostic screening," *IEEE J. Transl. Eng. Health Med.*, vol. 9, pp. 1–11, 2021.
[25] S. Krishnamurthy, K. Ks, E. Dovgan, M. Luštrek, B. G. Piletič, K. Srinivasan, Y.-C.-J. Li, A. Gradišek, and S. Syed-Abdul, "Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan," *Healthcare*, vol. 9, no. 5, p. 546, May 2021.
[26] M. Gupta and P. Gupta, "Predicting chronic kidney disease using machine learning," in *Emerging Technologies for Healthcare: Internet of Things and Deep Learning Models*. Hoboken, NJ, USA: Wiley, 2021, pp. 251–277.
[27] University of California Irvine Machine Learning Repository. (Oct. 4, 2021). *Chronic Kidney Disease Data Set*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease
[28] S. Al-Shamsi, D. Regmi, and R. D. Govender, "Chronic kidney disease in patients at high risk of cardiovascular disease in the United Arab Emirates: A population-based study," *PLoS ONE*, vol. 13, no. 6, Jun. 2018, Art. no. e0199920.
[29] G. S. Francis, "ACE inhibition in cardiovascular disease," *New England J. Med.*, vol. 342, no. 3, pp. 201–202, Jan. 2000.
[30] J. Agata, D. Nagahara, S. Kinoshita, Y. Takagawa, N. Moniwa, D. Yoshida, N. Ura, and K. Shimamoto, "Angiotensin II receptor blocker prevents increased arterial stiffness in patients with essential hypertension," *Circulat. J.*, vol. 68, no. 12, pp. 1194–1198, 2004.
[31] Kidney Disease: Improving Global Outcomes (KDIGO) Transplant Work Group, "KDIGO clinical practice guideline for the care of kidney transplant recipients," *Amer. J. Transplantation*, vol. 9, p. S1, Nov. 2009.
[32] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.
[33] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.
[34] M. A. Hardy, *Regression With Dummy Variables*. Newbury Park, CA, USA: Sage, 1993.
[35] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, "Review: A gentle introduction to imputation of missing values," *J. Clin. Epidemiol.*, vol. 59, no. 10, pp. 1087–1091, Oct. 2006.
[36] L. Oneto, *Model Selection and Error Estimation in a Nutshell*. Berlin, Germany: Springer, 2020.

[37] K. F. Kerr, "Comments on the analysis of unbalanced microarray data," *Bioinformatics*, vol. 25, no. 16, pp. 2035–2041, Aug. 2009.
[38] R. Laza, R. Pavón, M. Reboiro-Jato, and F. Fdez-Riverola, "Evaluating the effect of unbalanced data in biomedical document classification," *J. Integrative Bioinf.*, vol. 8, no. 3, pp. 105–117, Dec. 2011.
[39] K. Han, K. Z. Kim, J. M. Oh, I. W. Kim, K. Kim, and T. Park, "Unbalanced sample size effect on the genome-wide population differentiation studies," *Int. J. Data Mining Bioinf.*, vol. 6, no. 5, pp. 490–504, 2012.
[40] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.
[41] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.
[42] T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems," *Pattern Recognit.*, vol. 72, pp. 327–340, Dec. 2017.
[43] C. Molnar. (2018). *Interpretable Machine Learning*. [Online]. Available: https://christophm.github.io/book/
[44] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.
[45] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[46] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[47] S. S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel," *Neural Comput.*, vol. 15, no. 7, pp. 1667–1689, Mar. 2003.
[48] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[49] M. J. Zaki and W. Meira, Jr., *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2019.
[50] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, San Francisco, CA, USA, 2016, pp. 785–794.
[51] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Mach. Learn.*, vol. 11, no. 1, pp. 63–90, Apr. 1993.
[52] I. Orlandi, L. Oneto, and D. Anguita, "Random forests model selection," in *Proc. Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn.*, Bruges, Belgium, 2016, pp. 441–446.
[53] F. Hutter, H. Hoos, and K. Leyton-Brown, "An efficient approach for assessing hyperparameter importance," in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, Beijing, China, 2014, pp. 754–762.
[54] S. Bernard, L. Heutte, and S. Adam, "Influence of hyperparameters on random forest accuracy," in *Proc. Int. Workshop Multiple Classifier Syst.*, Reykjavik, Iceland, 2009, pp. 171–180.
[55] P. Probst, M. Wright, and A.-L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 9, no. 3, p. e1301, 2019.
[56] M. Kuhn, "Building predictive models in R using the caret package," *J. Statist. Softw.*, vol. 28, no. 5, pp. 1–26, 2008.
[57] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3133–3181, 2014.
[58] M. Wainberg, B. Alipanahi, and B. J. Frey, "Are random forests truly the best classifiers?" *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 3837–3841, 2016.
[59] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Comput.*, vol. 8, no. 7, pp. 1341–1390, Oct. 1996.
[60] Y. Saeys, T. Abeel, and Y. V. D. Peer, "Robust feature selection using ensemble feature selection techniques," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML PKDD)*, Antwerp, Belgium, 2008, pp. 313–325.
[61] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognit. Lett.*, vol. 31, no. 14, pp. 2225–2236, Oct. 2010.
[62] Y. Qi, "Random forest for bioinformatics," in *Ensemble Machine Learning*. Boston, MA, USA: Springer, 2012.
[63] R. Díaz-Uriarte and S. A. De Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinf.*, vol. 7, no. 1, p. 3, Dec. 2006.
[64] D. Chicco and C. Rovelli, "Computational prediction of diagnosis and feature selection on mesothelioma patient health records," *PLoS ONE*, vol. 14, no. 1, Jan. 2019, Art. no. e0208737.
[65] P. Good, *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York, NY, USA: Springer, 2013.
[66] M. L. Calle and V. Urrea, "Letter to the editor: Stability of random forest importance measures," *Briefings Bioinf.*, vol. 12, no. 1, pp. 86–89, Jan. 2011.
[67] M. B. Kursa, "Robustness of random forest-based gene selection methods," *BMC Bioinf.*, vol. 15, no. 1, pp. 1–8, Dec. 2014.
[68] H. Wang, F. Yang, and Z. Luo, "An experimental study of the intrinsic stability of random forest variable importance measures," *BMC Bioinf.*, vol. 17, no. 1, p. 60, 2016.
[69] D. Sculley, "Rank aggregation for similar items," in *Proc. SIAM Int. Conf. Data Mining*, Minneapolis, MN, USA, Apr. 2007, pp. 587–592.
[70] T. W. MacFarland and J. M. Yates, "Mann–Whitney U test," in *Introduction to Nonparametric Statistics for the Biological Sciences Using R*. Berlin, Germany: Springer, 2016, pp. 103–132.
[71] P. E. Greenwood and M. S. Nikulin, *A Guide to Chi–Squared Testing*, vol. 280. Hoboken, NJ, USA: Wiley, 1996.
[72] D. J. Benjamin *et al.*, "Redefine statistical significance," *Nature Hum. Behav.*, vol. 2, no. 1, pp. 6–10, 2018.
[73] C. R. Mehta and N. R. Patel, "Exact logistic regression: Theory and examples," *Statist. Med.*, vol. 14, no. 19, pp. 2143–2160, Oct. 1995.
[74] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, p. 16, Dec. 2020.
[75] D. Chicco, "Ten quick tips for machine learning in computational biology," *BioData Mining*, vol. 10, no. 35, pp. 1–17, 2017.
[76] D. Chicco, M. J. Warrens, and G. Jurman, "The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and brier score in binary classification assessment," *IEEE Access*, vol. 9, pp. 78368–78381, 2021.
[77] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Mining*, vol. 14, Feb. 2021, Art. no. 13.
[78] D. Chicco, V. Starovoitov, and G. Jurman, "The benefits of the Matthews correlation coefficient (MCC) over the diagnostic odds ratio (DOR) in binary classification assessment," *IEEE Access*, vol. 9, pp. 47112–47124, 2021.
[79] P. C. Austin and E. W. Steyerberg, "Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers," *Statist. Med.*, vol. 33, no. 3, pp. 517–535, Feb. 2014.
[80] N. E. Breslow, L. P. Zhao, T. R. Fears, and C. C. Brown, "Logistic regression for stratified case–control studies," *Biometrics*, vol. 44, no. 3, pp. 891–899, 1988.
[81] J. H. Zar, "Spearman rank correlation," in *Encyclopedia Biostatistics*, vol. 7. Hoboken, NJ, USA: Wiley, 2005.
[82] F. J. Brandenburg, A. Gleißner, and A. Hofmeier, "Comparing and aggregating partial orders with Kendall tau distances," in *Proc. 6th Int. Workshop Algorithms Comput. (WALCOM)*. Dhaka, Bangladesh: Springer, 2012, pp. 88–99.
[83] D. Chicco, E. Ciceri, and M. Masseroli, "Extended Spearman and Kendall coefficients for gene annotation list correlation," in *Proc. 11th Int. Meeting Comput. Intell. Methods Bioinf. Biostatistics (CIBB)*, in Lecture Notes in Computer Science, vol. 8623. Cambridge, U.K.: Springer, 2015, pp. 19–32.
[84] D. Clayton and J. Cuzick, "Multivariate generalizations of the proportional hazards model," *J. Roy. Stat. Soc., A, General*, vol. 148, no. 2, pp. 82–108, 1985.
[85] P. A. Flach, "Classifier calibration," in *Encyclopedia of Machine Learning and Data Mining*. Berlin, Germany: Springer, 2016.
[86] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 4768–4777.
[87] L.-T. Zhou, S. Qiu, L.-L. Lv, Z.-L. Li, H. Liu, R.-N. Tang, K.-L. Ma, and B.-C. Liu, "Integrative bioinformatics analysis provides insight into the molecular mechanisms of chronic kidney disease," *Kidney Blood Pressure Res.*, vol. 43, no. 2, pp. 568–581, 2018.

[88] Z. Zuo, J.-X. Shen, Y. Pan, J. Pu, Y.-G. Li, X.-H. Shao, and W.-P. Wang, "Weighted gene correlation network analysis (WGCNA) detected loss of MAGI2 promotes chronic kidney disease (CKD) by podocyte damage," *Cellular Physiol. Biochem.*, vol. 51, no. 1, pp. 244–261, 2018.
[89] C.-Y. Ho, T.-W. Pai, Y.-C. Peng, C.-H. Lee, Y.-C. Chen, Y.-T. Chen, and K.-S. Chen, "Ultrasonography image analysis for detection and classification of chronic kidney disease," in *Proc. 6th Int. Conf. Complex, Intell., Softw. Intensive Syst. (CISIS)*, Palermo, Italy, Jul. 2012, pp. 624–629.

DAVIDE CHICCO received the Bachelor of Science and Master of Science degrees in computer science from the Università di Genova, Genoa, Italy, in 2007 and 2010, respectively, and the Ph.D. degree in computer engineering from the Politecnico di Milano University, Milan, Italy, in Spring 2014. He also spent a semester as a Visiting Doctoral Scholar with the University of California Irvine, USA. From September 2014 to September 2018, he was a Postdoctoral Researcher with the Princess Margaret Cancer Centre and a Guest with the University of Toronto. From September 2018 to December 2019, he was a Scientific Associate Researcher with the Peter Munk Cardiac Centre, Toronto, ON, Canada. From January 2020 to January 2021, he was a Scientific Associate Researcher with the Krembil Research Institute, Toronto. In January 2021, he started to work as a Scientific Research Associate with the Institute of Health Policy Management and Evaluation, University of Toronto.

CHRISTOPHER A. LOVEJOY received the bachelor's degree in medicine from the University of Cambridge, U.K., and the master's degree in data science and machine learning from University College London, U.K. He is currently a Medical Doctor with interests in applied machine learning and bioinformatics.

LUCA ONETO received the Bachelor of Science and Master of Science degrees in electronic engineering from the Università di Genova, Italy, in 2008 and 2010, respectively, and the Ph.D. degree from the School of Sciences and Technologies for Knowledge and Information Retrieval, Università di Genova, in 2014, with the thesis entitled Learning Based on Empirical Data. In 2017, he obtained the Italian National Scientific Qualification for the role of an Associate Professor in computer engineering, and in 2018, he obtained the one in computer science. He worked as an Assistant Professor in computer engineering with the Università di Genova, from 2016 to 2019, where he is currently an Associate Professor in computer engineering. In 2018, he was a Co-Funder of ZenaByte s.r.l., spin-off company. In 2019, he obtained the Italian National Scientific Qualification for the role of a Full Professor in computer science and computer engineering. In 2019, he became an Associate Professor in computer science with the Università di Pisa. His first main topic of research is the statistical learning theory with particular focus on the theoretical aspects of the problems of (semi) supervised model selection and error estimation. His second main topic of research is data science with particular reference to the problem of trustworthy AI and the solution of real world problems by exploiting and improving the most recent learning algorithms and theoretical results in the fields of machine learning and data mining. He has been involved in several Horizon 2020 projects (S2RJU, ICT, and DS) and awarded with the Amazon AWS Machine Learning and Somalvico (Best Italian Young AI Researcher) Awards.