A machine learning analysis of health records of patients with chronic kidney disease at risk of cardiovascular disease

DAVIDE CHICCO*1, CHRISTOPHER A. LOVEJOY2,3, and LUCA ONETO4,5

1University of Toronto, Toronto, Ontario, Canada
2University College London, London, United Kingdom
3University College London Hospital, London, United Kingdom
4Università di Genova, Genoa, Italy
5ZenaByte Srl, Genoa, Italy

*Corresponding author: Davide Chicco (e-mail: davidechicco@davidechicco.it).

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 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.

1. 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 [2]. 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].

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 [16] and Abrah [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 and colleagues [24] also employed
a number of machine learning methods for patient classification and described the dataset precisely. Syed Imran Ali and co-authors [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 and colleagues [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. Ogungbey and Wang [11] used an enhanced XGBoost method for patient classification. Satukumati and coauthors [17] used several techniques for feature extraction. Elhoseny and colleagues [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. Chitter and colleagues [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 and coauthors [4] applied numerous machine learning classifiers and their variants for patient classification. Subas et al. [9] focused on Random Forests for patient classification and feature ranking. Zeynu and colleagues [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 and coauthors [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 result being the top most important one.

Ravizza and colleagues [20] employed machine learning methods on a dataset of patients with diabetes from the IBM Explorys 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.

In this manuscript, we analyzed a dataset of 491 patients from United Arab Emirates, released by Al-Shamsi and colleagues [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 and colleagues [23] 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).

**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 3) 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 variables 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 3).

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;
- 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 $\mathcal{R}$ 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 unordered 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 $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, with $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, is available in a binary classification framework. Moreover, some values of $x_i$ 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 $\mathcal{R}$, through an algorithm $\mathfrak{A}_{\mathcal{H}}$ characterized by its set of hyper-parameters $\mathcal{H}$. The accuracy of the model $\mathfrak{M}$ to represent the unknown relation $\mathcal{R}$ is measured using different indices of performance (Supplementary information).

Since the hyper-parameters $\mathcal{H}$ influence the ability of $\mathfrak{A}_{\mathcal{H}}$ to estimate $\mathcal{R}$, we need to adopt a proper Model Selection (MS) procedure [36]. In this work, we exploit the Complete Cross Validation (CCV) procedure [36]. CCV relies on a simple idea: we resample the original dataset $D_n$ many ($n_s = 500$) times without replacement to build a training set of size $l \mathcal{C}_r$ while the remaining samples are kept in the validation set $\mathcal{V}_v$, with $v \in \{1, \ldots, n_s\}$. In order to perform the MS phase, to select the best combination of the hyper-parameters $\mathcal{H}$ in the set of possible ones $\mathcal{S} = \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}$ using the algorithm $\mathfrak{A}_{\mathcal{H}}$, the hyper-parameters which minimize the average performance of the model, trained on the
In this study: synthetic minority oversampling [41], [42]. removing abundant samples, new rare samples are generated on the dataset by increasing the size of rare samples. Rather than ples in the abundant class, a new balanced dataset dataset [40]. For this purpose, we can under-sample or over- the resampling of the data in order to synthesize a balanced on the minority class. For these reasons, several techniques well with imbalanced datasets and tend to poorly perform 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 \( \mathcal{M} \) 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 \( \mathcal{M} \) 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 \( \mathcal{M} \) 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 \( \mathcal{M} \) about the most important features which influence its results.

### A. BINARY CLASSIFICATION ALGORITHMS

In this paper, for the \( \mathcal{A} \), 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\} [2], [52–54]. For the linear and kernelized Support Vector Machines [46], we searched the regularization hyper-parameters in \{10^{−6.0}, 5.8, \ldots, 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, \ldots, 4\}. For the Neural Network we used a single hidden layer network (hyperbolic tangent as activation function in the hidden

| Feature | explanation | measurement unit | values |
|---------|-------------|-----------------|--------|
| ACEI:ARB | 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/m² | [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 | mmol/L | [6, 27, ..., 113, 123] |
| dBPBaseline | diastolic blood pressure | mmHg | [41, 45, ..., 110, 112] |
| DMdmeds | if the patient has taken diabetes medications | boolean | [0, 1] |
| DMdmeds | if the patient has taken diabetes medications | boolean | [0, 1] |
| eGFRBaseline | estimated glomerular filtration rate | ml/min/1.73m² | [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] |

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3133700, IEEE Access

---

Enzyme inhibitors. ARB: Angiotensin II receptor blockers. mmHg: millimetre of mercury. kg: kilogram. mmol: millimoles.

---

**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.
The dataset contains medical records of 491 patients.

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.

TABLE 3: Numeric feature quantitative characteristics. \(\sigma\): standard deviation.

| Feature            | median | mean [range]       | \(\sigma\) |
|--------------------|--------|--------------------|-----------|
| 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     |

Feature rankings methods based on Random Forests are among the most effective techniques [59], [60], particularly in the context of bioinformatics [61], [62] and health informatics [63]. 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 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 [65]–[67]. 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 sub-sample the original dataset and we repeat the procedure many times. The final rank of a feature will be the aggregation of the different ranking using the Borda’s method [68].

B. FEATURE RANKING

Feature rankings methods based on Random Forests are among the most effective techniques [59], [60], particularly in the context of bioinformatics [61], [62] and health informatics [63]. 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 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 [65]–[67]. 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 sub-sample the original dataset and we repeat the procedure many times. The final rank of a feature will be the aggregation of the different ranking using the Borda’s method [68].
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) and of the chi-square test 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 and colleagues, we use 0.005 as threshold of significance for the p-values, that is $5 \times 10^{-3}$. 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 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, F1 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

CKD prediction. We report the results obtained for the static prediction of the CKD measured with traditional confusion matrix indicators in Table 3. 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.

Random Forests outperformed all the other methods for MCC, F1 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 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 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.

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 perform 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, 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 of its predictions (Figure 3). As one can notice, the Stratified
B. FEATURE RANKING RESULTS

CKD predictive feature ranking. After verifying that computational intelligence is able to predict CKD developments among patients, we applied 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, we 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).

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 irrelevant predictive feature (Table 6), while Random Forests puts it on the 4th 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.

CKD predictive feature ranking considering 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.

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.
poral 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 1st position in this ranking, while it

| position | feature       | Mann-Whitney U test p-value |
|----------|---------------|-----------------------------|
| 1        | *AgeBaseline  | 0                           |
| 2        | *CreatinineBaseline | 0                          |
| 3        | *eGFRBaseline | 0                           |
| 4        | *CholesterolBaseline | $9.490 \times 10^{-04}$   |
| 5        | *sBPBaseline  | $4.379 \times 10^{-03}$    |
| 6        | dBPPBaseine   | $1.083 \times 10^{-01}$    |
| 7        | BMIBaseline   | $9.134 \times 10^{-01}$    |

| position | feature       | chi-squared p-value |
|----------|---------------|---------------------|
| 1        | *HistoryDiabetes | $5 \times 10^{-04}$ |
| 2        | *HistoryCHD    | $5 \times 10^{-04}$  |
| 3        | *HistoryHTN    | $5 \times 10^{-04}$  |
| 4        | *DLmeds        | $5 \times 10^{-04}$  |
| 5        | *DMmeds        | $5 \times 10^{-04}$  |
| 6        | *ACEIARB       | $5 \times 10^{-04}$  |
| 7        | *HistoryDLD    | $1.999 \times 10^{-03}$ |
| 8        | *HTNmeds       | $1.999 \times 10^{-03}$ |
| 9        | HistoryVascular| $3.698 \times 10^{-02}$ |
| 10       | Sex            | $4.398 \times 10^{-02}$ |
| 11       | HistorySmoking | $5.397 \times 10^{-02}$ |
| 12       | HistoryObesity | $4.948 \times 10^{-01}$ |

| position | MDA average position feature |
|----------|-----------------------------|
| 1        | 1.2 AgeBaseline              |
| 2        | 1.8 eGFRBaseline             |
| 3        | 3.3 DLmeds                   |
| 4        | 3.7 dBPPBaseline             |
| 5        | 5.2 CholesterolBaseline      |
| 6        | 6.0 HistoryVascular          |
| 7        | 7.0 HistoryCHD               |
| 8        | 8.3 dBPPBaseline             |
| 9        | 8.7 CreatinineBaseline       |
| 10       | 11.4 HistoryHTN              |
| 11       | 11.6 HistorySmoking          |
| 12       | 11.9 DLmeds                  |
| 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         |
is 14\textsuperscript{th} without considering time;
- HistoryHTN is at the 3\textsuperscript{rd} position in this ranking, while it is 10\textsuperscript{th} without considering time;
- ACEIARB is at the 4\textsuperscript{th} position in this ranking, while it is 17\textsuperscript{th} without considering time;
- AgeBaseline is at the last position in this ranking, while it is 1\textsuperscript{st} without considering time;
- CreatinineBaseline is at the 18\textsuperscript{th} position in this ranking, while it is 9\textsuperscript{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 [80]–[82]. Both these metrics range between −1.0 and +1.0, 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

#### 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.

#### 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 happened 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.

#### 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 (section 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.

Comparison with results of the original study. The original study of Al-Shamsi and colleagues [28] included a feature ranking phase generated through a multivariable Cox’s proportional hazards analysis, which included the temporal component [33]. 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 ranking, while HistoryCHD and HistoryDLD were at unimportant positions: 10th and 16th ranks out of 19 variables, respectively.

Smoking history, instead, occupied a high rank both in our ranking of Salekin’s study [6], confirming the importance of the HistoryHTN variable which is ranked at the 3rd position in our Stratified Logistic Regression ranking (Table 8). Also diabetes history has high ranking in both the standings: 3rd position in the ranking of Salekin’s study [6], and 6th of importance in our Stratified Logistic Regression ranking, as HistoryDiabetes (Table 8).

VI. CONCLUSIONS

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 [84], and to perform the feature ranking with a different feature ranking method such as SHapley Additive exPlanations (SHAP) [85]. Moreover, we also plan to study chronic kidney disease by applying our methods to CKD datasets of other types, such as microarray gene expression [86], [87] and ultrasonography images [88].

LIST OF ABBREVIATIONS

AUC: area under the curve. BP: blood pressure. CHD: coronary heart 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.

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

DATA AND SOFTWARE AVAILABLE
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/Chronic_kidney_disease_in_patients_at_high_risk_of_cardiovascular_disease_in_the_United_Arab_Emirates_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] Valerie A Luyckx, Marcello Tonelli, and John W Stanifer. The global burden of kidney disease and the sustainable development goals. Bulletin of the World Health Organization, 96(6):414, 2018.
[2] Sarmad Said and German T Hernandez. The link between chronic kidney disease and cardiovascular disease. Journal of Nephropathology, 3(3):99, 2014.
[3] Kevin Damman, Mattia A Valente, Adriaan A Voors, Christopher M O’Connor, Dirk J van Veldhuisen, and Hans L Hillegé. Renal impairment, worsening renal function, and outcome in patients with heart failure: an updated meta-analysis. European Heart Journal, 35(7):455–469, 2014.
[4] Anusorn Charleonnan, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueyputtananon, Satthi Swannawatch, and Nitat Pinchawee. Predictive analytics for chronic kidney disease using machine learning techniques. In Proceedings of 2016 MITicon – the 2016 Management and Innovation Technology International Conference, pages 80–83, Bang Saen, Chonburi, Thailand, 2016. IEEE.
[5] Nusrat Tazin, Shahed Anzarus Sabab, and Muhammed Tawfiq Chowdhury. Diagnosis of chronic kidney disease using effective classification and feature selection technique. In Proceedings of MedTec 2016 – the 2016 International Conference on Medical Engineering, Health Informatics and Technology, pages 1–6, Dhaka, Bangladesh, 2016. IEEE.
[6] Asif Salekin and John Stankovic. Detection of chronic kidney disease and selecting important predictive attributes. In Proceedings of IEEE ICHI 2016 – the 2016 IEEE International Conference on Healthcare Informatics, pages 262–270, Chicago, Illinois, USA, 2016. IEEE.
[7] Huseyin Polat, Homay Danaei Mehr, and Aydin Cetin. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. Journal of Medical Systems, 41(4):55, 2017.
[8] Made Satria Wibawa, I Made Dendi Maysanjaya, and I Made Agus Wirahadi Putra. Boosted classifier and features selection for enhancing chronic kidney disease diagnosis. In Proceedings of CITSM 2017 – the 5th International Conference on Cyber and IT Service Management, pages 1–6, Denpasar, Bali, Indonesia, 2017. IEEE.
[9] Abdulhamit Subasi, Emina Alicikovic, and Jasmin Kevric. Diagnosis of chronic kidney disease by using random forest. In Proceedings of CMBEBIH 2017 – the 2017 International Conference on Medical and Biological Engineering, pages 589–594. Springer, 2017.
[10] Siraje Zeynu and Shrutti Patil. Prediction of chronic kidney disease using data mining feature selection and ensemble method. International Journal of Data Mining in Genomics & Proteomics, 9(1):1–9, 2018.
[11] Adrola Ogureleye and Qing Gao Wang. Enhanced XGBoost-based automatic diagnosis system for chronic kidney disease. In Proceedings of IEEE ICCA 2018 – the 14th IEEE International Conference on Control and Automation, pages 805–810, Anchorage, Alaska, USA, 2018. IEEE.
[12] Siraje Zeynu and Shruthi Patil. Survey on prediction of chronic kidney disease using data mining classification techniques and feature selection. International Journal of Pure and Applied Mathematics, 118(8):149–156, 2018.
[13] Abdulkarim Al Imran, MD Nur Amin, and Fatema Tuj Johora. Classification of chronic kidney disease using logistic regression, feedforward neural network and wide & deep learning. In Proceedings of ICICIET 2018 – the 2018 International Conference on Innovation in Engineering and Technology, pages 1–6, Osaka, Japan, 2018. IEEE.
[14] AK Shrivas, Sanat Kumar Sahu, and HS Hota. Classification of chronic kidney disease with proposed union based feature selection technique. In Proceedings of ICIoTCCT 2018 – the 3rd International Conference on Internet of Things and Connected Technologies, pages 26–27, Jaipur, India, 2018.
[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. International Journal of Engineering & Technology, 7(2.31):190–195, 2018.
[16] Naveed Rahman Shawn, Syed Samiul Alam Mehrab, Fardeen Ahmed, and Mohammad Shararat Hasmi. Chronic kidney disease detection using ensemble classifiers and feature set reduction. PhD thesis, BRAC University, 2019.
[17] Suresh Babu Satukumati and Raghu Kogila Shivaprasad Satla. Feature extraction techniques for chronic kidney disease identification. Kidney, 15:29, 2019.
[18] Tahmid Abrar, Samia Tasnim, and MD Hossain. Early detection of chronic kidney disease using machine learning. PhD thesis, BRAC University, 2019.
[19] Mohamed Elhoseny, K Shankar, and J Uthayakumar. Intelligent diagnostic prediction and classification system for chronic kidney disease. Scientific Reports, 9(1):1–14, 2019.
[20] Stefan Ravizza, Tony Huschto, Anja Adamow, Lars Bohm, Alexander Büscher, Frederik F Flöther, Rolf Hinzmann, Helena König, Scott M Mcahren, Daniel H Robertson, Titus Schleyer, Bernd Schneidering, and Wolfgang Petrich. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. Nature Medicine, 25(1):57–59, 2019.
[21] Syed Imran Ali, Gwag Hoon Park, and Sungyoung Lee. Cost-sensitive ensemble feature ranking and automatic threshold selection for chronic kidney disease diagnosis. Preprints, 202005045S, 2020.
[22] Pankaj Chittora, Sandeep Chaurasia, Prasun Chakrabarti, Gaurav Kumawat, Tulika Chakrabarti, Zbigniew Leonowicz, Michal Jasinski, Lukasz Jasinski, Radomir Gono, Elżbieta Jasińska, and Vadim Bolshev. Prediction of chronic kidney disease-a machine learning perspective. IEEE Access, 9:17312–17334, 2021.
[23] Piervincenzo Ventrella, Giovanni Delgrossi, Giannichele Ferrario, Marco Righetti, and Marco Masseroli. Supervised machine learning for the assessment of chronic kidney disease advancement. Computer Methods and Programs in Biomedicine, 209:106329, 2021.
[24] Md Rashed-Al-Mahfuz, Abdul Haque, Akaz Azad, Salem A Alyami, Julian MW Quinn, and Mohammad Ali Moni. Clinically applicable machine learning approaches to identify attributes of chronic kidney disease (CKD) for use in low-cost diagnostic screening. IEEE Journal of Translational Engineering in Health and Medicine, 9:1–11, 2021.
[25] Surya Krishnamurthy, Kapelesh KS, Erik Dovgan, Mitja Luštrek, Barbara Gradšek Piletič, Kathiravan Srinivasan, Yu-Chuan Jack Li, Anton Gradšek, and Shabbir Syed-Abdul. Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan. In Healthcare, volume 9, page 546. Multidisciplinary Digital Publishing Institute, 2021.
[26] Monika Gupta and Parul Gupta. Predicting chronic kidney disease using machine learning. Emerging Technologies for Healthcare: Internet of Things and Deep Learning Models, pages 251–277, 2021.
[27] University of California Irvine Machine Learning Repository. Chronic kidney disease data set. https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease URL visited on 4th October, 2021.
[28] Saif Al-Shamsi, Dybesh Regim, and Romona D Govender. Chronic kidney disease in patients at high risk of cardiovascular disease in the United Arab Emirates: a population-based study. PLos One, 13(6), 2018.
[29] Gary S Francis. ACE inhibition in cardiovascular disease. New England Journal of Medicine, 370:2021–2032, 2019.
[30] Jun Agata, Daigo Nagahara, Shuichi Kinoshita, Yoshitoki Takagawa, Nohito Moniwa, Daiksu Yoshiha, Nobuyuki Ura, and Kazuaki Shimamoto. Angiotensin II receptor blocker prevents increased arterial stiffness in
patients with essential hypertension. Circulation Journal, 68(12):1194–1198, 2004.

[31] Kidney Disease: Improving Global Outcomes (KDIGO) Transplant Work Group. KDIGO clinical practice guideline for the care of kidney transplant recipients. American Journal of Transplantation, 9:51, 2009.

[32] Shai Shalev-Shwartz and Shai Ben-David. Understanding machine learning: from theory to algorithms. Cambridge University Press, 2014.

[33] André Altman, Laura Tolosi, Olivier Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10):1340–1347, 2010.

[34] Melissa A Hardy. Regression with Dummy Variables. Sage, 1993.

[35] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT Press, Cambridge, Massachusetts, USA, 2016.

[36] Christoph Molnar. Interpretable machine learning. https://christophm.github.io/book/, 2018.

[37] Kathleen F Kerr. Comments on the analysis of unbalanced microarray data. Bioinformatics, 25(16):2035–2041, 2009.

[38] Rosalia Laza, Reyes Pavón, Miguel Rebrojo-Jato, and Florentino Fdez-Riverola. Evaluating the effect of unbalanced data in biomedical document classification. Journal of Integrative Bioinformatics, 8(3):105–117, 2011.

[39] Kyanghee Han, Kyee Zu Kim, Jung Mi Oh, In Wha Kim, Kyungjin Kim, and Taesung Park. Unbalanced sample size effect on the genome-wide population differentiation studies. International Journal of Data Mining and Bioinformatics, 6(5):490–504, 2012.

[40] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: review of methods and applications. Data Systems with Applications, 73:220–229, 2017.

[41] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357, 2002.

[42] Tuanfei Zhu, Yaping Lin, and Yonghe Liu. Synthetic minority oversampling technique for multiclass imbalance problems. Pattern Recognition, 72:327–340, 2017.

[43] Thomas W MacFarland and Jan M Yates. Mann–Whitney U test. In Tom F Dastic, editor, Encyclopedia of Human Biology, volume 4, pages 577–585. Academic Press, 1996.

[44] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. Journal of Machine Learning Research, 3(Mar):1157–1182, 2003.

[45] Leo Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[46] John Shawe-Taylor and Nello Cristianini. Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge, England, United Kingdom, 2004.

[47] S Sathiya Keerthi and Chih-Jen Lin. Asymptotic behaviors of support vector machines. Journal of Machine Learning Research, 3(Mar):1157–1182, 2003.

[48] Autumn Heskes and Peter Sollich. A bayesian approach for assessing hyperparameter importance. In International Conference on Neural Information Processing, pages 171–180, Reykjavik, Iceland, 2009.

[49] David Breiman. Bagging predictors. Machine Learning, 24(2):123–140, 1996.

[50] David H Wolpert. The lack of a priori distinctions between learning algorithms. Neural Computation, 8(7):1341–1390, 1996.

[51] Yvan Saey, Thomas Abeel, and Yves Van de Peer. Robust feature selection using ensemble feature selection techniques. In Proceedings of ECML PKDD 2008 — the 2008 Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Antwerp, Belgium, 2008.

[52] Jesse Genuer, Jean-Michel Poggi, and Christine Tuleau-Malet. Variable selection using random forests. Pattern Recognition Letters, 31(14):2225–2236, 2010.

[53] Ruiyan Qi. Random forest for bioinformatics. In Ensemble Machine Learning, Boston, Massachusetts, USA, 2012.

[54] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7(1):3, 2006.

[55] Davide Chicco and Cristina Rovelli. Computational prediction of diagnosis and feature selection on mesothelioma patient health records. PLoS One, 14(1):e0208737, 2019.

[56] Thomas W MacFarland and Jan M Yates. Mann–Whitney U test. In Tom F Dastic, editor, Encyclopedia of Human Biology, volume 4, pages 577–585. Academic Press, 1996.

[57] Daniel J Benjamin, Jamie O Berger, Magnus Johannesson, Brian A Nosek, E J Wagenmakers, Richard Berk, Kenneth A Bollen, Bjorn Brembs, Lawrence Brown, Colin Camerer, David Cesaroni, Christopher D Chambers, Merlise Clyde, Thomas D Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P Field, Malcolm Forster, Edward I George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P Green, Anthony G Greenwald, Jarrod D Hadfield, Larry V Hedges, Leonhard Held, Tiek Hua Ho, Herbert Hoogeveen, David J Hruschka, Kosuke Imai, Guido Imbens, John P A Ioannidis, Mininjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E Maxwell, Michael McCarthy, Don A Moore, Stephen L Morgan, Marcus Munafò, Shinichi Nakagawa, Brendan Nyhan, Timothy H Parker, Luis Pericchi, Marco Perugini, Jeff Ruder, Judith Rousseau, Victoria Squezi, Felix D Schönbrodt, Thomas Selkie, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Sinemine Vazire, Duncan J Watts, Christopher Winship, Robert L Wolpert, Yu Xie, Cristobal Young, Jonathan Zimman, and Valen E Johnson. Redefine statistical significance. Nature Human Behaviour, 1:6–10, 2018.

[58] Cyrus R Mehta and Ninit R Patel. Exact logistic regression: theory and examples. Statistics in Medicine, 14(19):2143–2160, 1995.

[59] Davide Chicco and Giuseppe Jurman. The Matthews correlation coefficient (MCC) is more informative than Cohen’s Kappa and Brier score in binary classification assessment. IEEE Access, 9:78368–78381, 2021.
bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining, 14, 2021.

[77] Davide Chicco, Valery Starovoitov, and Giuseppe Jurman. The benefits of the Matthews correlation coefficient (MCC) over the diagnostic odds ratio (DOR) in binary classification assessment. IEEE Access, 9:47112–47124, 2021.

[78] Peter C Austin and Ewout W Steyerberg. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Statistics in Medicine, 33(3):517–535, 2014.

[79] Norman E Breslow, Lue P Zhao, Thomas R Fears, and Charles C Brown. Logistic regression for stratified case-control studies. Biometrics, 44(3):891–899, 1988.

[80] Jerrold H Zar. Spearman rank correlation. In Encyclopedia of Biostatistics, volume 7, Hoboken, New Jersey, USA, 2005. Wiley Online Library.

[81] Franz J Brandenburg, Andreas Gleißner, and Andreas Hofmeier. Comparing and aggregating partial orders with Kendall tau distances. In Proceedings of WALCOM 2012 – the 6th International Workshop on Algorithms and Computation, pages 88–99, Dhaka, Bangladesh, 2012. Springer.

[82] Davide Chicco, Eleonora Ciceri, and Marco Masseroli. Extended Spearman and Kendall coefficients for gene annotation list correlation. In Proceedings of CIIEB 2014 – the 11th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, volume 9623 of Lecture Notes in Computer Science, pages 19–32, Cambridge, England, United Kingdom, 2015. Springer.

[83] David Clayton and Jack Cuzick. Multivariate generalizations of the proportional hazards model. Journal of the Royal Statistical Society: Series A (General), 148(2):82–108, 1985.

[84] Peter A Flach. Classifier calibration. In Encyclopedia of Machine Learning and Data Mining. Springer, Berlin, Germany, 2016.

[85] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Proceedings of NIPS 2017 – the 31st International Conference on Neural Information Processing Systems, pages 4768–4777, 2017.

[86] Le-Ting Zhou, Shen Qiu, Lin-Li Lv, Zuo-Lin Li, Hong Liu, Ri-Ning Tang, Kun-Ling Ma, and Bi-Cheng Liu. Integrative bioinformatics analysis provides insight into the molecular mechanisms of chronic kidney disease. Kidney and Blood Pressure Research, 43(2):568–581, 2018.

[87] Zhi Zuo, Jian-Xiao Shen, Yan Pan, Juan Pu, Yong-Gang Li, Xing-hua Shao, and Wan-Peng Wang. Weighted gene correlation network analysis (WGCNA) detected loss of MAGI2 promotes chronic kidney disease (CKD) by podocyte damage. Cellular Physiology and Biochemistry, 51(1):244–261, 2018.

[88] Chih-Yin Ho, Tun-Wen Pai, Yuan-Chi Peng, Chien-Hung Lee, Yung-Chih Chen, Yang-Ting Chen, and Kuo-Su Chen. Ultrasonography image analysis for detection and classification of chronic kidney disease. In Proceedings of CISIS 2012 – the 6th International Conference on Complex, Intelligent, and Software Intensive Systems, pages 624–629, Palermo, Italy, 2012. IEEE.
SUPPLEMENTARY INFORMATION

DATA ENGINEERING

We derived the \textit{time year} feature from the \textit{TimeToEvent-Months} variable present in the original dataset. We associated to the time year 1 all the patients who had \textit{TimeToEvent-Months} between 0 and 12 months, time year 2 all the patients who had \textit{TimeToEvent-Months} between 13 and 24 months, time year 3 all the patients who had \textit{TimeToEvent-Months} between 25 and 36 months, and so on. If a patient has \textit{time year} = \textit{x} (where \( x \in \mathbb{N} \)), it means that the CKD development occurred in the \textit{x}\textsuperscript{th} year of the follow-up (since 2008) for each patient who developed stages 3-5 CKD, or that the subject last visit happened in the \textit{x}\textsuperscript{th} year, for healthy controls.

BINARY STATISTICAL RATES

List of statistical rates to evaluate confusion matrices and their formulas:

\[
\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}
\]

(worst value = −1; best value = +1)

\[
\text{F}_1 \text{ score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}
\]

(worst value = 0; best value = 1)

\[
\text{accuracy} = \frac{TP + TN}{TP + FN + TN + FP}
\]

(worst value = 0; best value = 1)

\[
\text{true positive rate}, \text{ recall}, \text{ sensitivity} = \frac{TP}{TP + FN}
\]

(worst value = 0; best value = 1)

\[
\text{true negative rate}, \text{ specificity} = \frac{TN}{TN + FP}
\]

(worst value = 0; best value = 1)

\[
\text{positive predictive value}, \text{ precision} = \frac{TP}{TP + FP}
\]

(worst value = 0; best value = 1)

\[
\text{negative predictive value} = \frac{TN}{TN + FN}
\]

(worst value = 0; best value = 1)

\[
\text{Precision-Recall (PR) curve} = \begin{cases} 
\text{true positive rate} & \text{on the x axis} \\
\text{precision} & \text{on the y axis}
\end{cases}
\]

(worst value = 0; best value = 1)

\[
\text{ROC curve} = \begin{cases} 
\text{false positive rate} & \text{on the x axis} \\
\text{true positive rate} & \text{on the y axis}
\end{cases}
\]

(worst value = 0; best value = 1)