Supervised machine learning for the assessment of Chronic Kidney Disease advancement

Piervincenzo Ventrella [a, *], ***Giovanni Delgrossi*** [b], ***Gianmichele Ferrario*** [b], ***Marco Righetti*** [b], ***Marco Masseroli*** [a]

a. *DEIB, Politecnico di Milano, Piazza L. Da Vinci 32, 20133 Milan (MI), Italy*

b. ASST Vimercate, Via Santi Cosma e Damiano, 10, 20871 Vimercate (MB), Italy

ABSTRACT

Background and objective

Chronic Kidney Disease (CKD) is a condition characterized by a progressive loss of kidney function over time caused by many diseases. The most effective weapons against CKD are early diagnosis and treatment, which in most of the cases can only postpone the onset of complete kidney failure. The CKD grading system is classified based on the estimated Glomerular Filtration Rate (eGFR), and it helps to stratify patients for risk, follow up and management planning. This study aims to effectively predict how soon a CKD patient will need to be dialyzed, thus allowing personalized care and strategic planning of treatment.

Methods

To accurately predict the time frame within which a CKD patient will necessarily have to be dialyzed, a computational model based on a supervised machine learning approach is developed. Many techniques, regarding both information extraction and model training phases, are compared in order to understand which approaches are most effective. The different models compared are trained on the data extracted from the Electronic Medical Records of the Vimercate Hospital.

Results

As final model, we propose a set of Extremely Randomized Trees classifiers considering 27 features, including creatinine level, urea, red blood cells count, eGFR trend (which is not even the most important), age and associated comorbidities. In predicting the occurrence of complete renal failure within the next year rather than later, it obtains a test accuracy of 94%, specificity of 91% and sensitivity of 96%. More and shorter time-frame intervals, up to 6 months of granularity, can be specified without relevantly worsening the model performance.

Conclusions

The developed computational model provides nephrologists with a great support in predicting the patient's clinical pathway. The model promising results, coupled with the knowledge and experience of the clinicians, can effectively lead to better personalized care and strategic planning of both patient's needs and hospital resources.

KEYWORDS

Chronic Kidney Disease, supervised machine learning, predicting renal failure, personalized care, chronicity management.

1 Introduction

The Chronic Kidney Disease (CKD) is a generic condition for several diseases that affect the kidneys, and it generally means permanent and progressive damage to kidneys, until to end-stage renal disease [1]. It affects 12%-14% of people worldwide and its related care costs represent an important percentage of the total health expenditure. According to the Center for Disease Control and Prevention (CDC), Chronic Kidney Disease affects approximately 1 in 7 adults, or an estimated 30 million Americans, consisting in annual national care cost over 32 billion of dollars [2]. Several are the possible causes of onset and rapid evolution of CKD, including diabetes, high blood pressure, or previous episodes in the family history [3]. Prevention and early detection of CKD allow appropriate treatment and are the main factors against the disease, which however, in most of the cases, can only postpone the onset of complete kidney failure.

To highlight the presence of CKD and assess the kidney functionalities, the estimated Glomerular Filtration Rate (eGFR) is computed. It forms the basis of CKD staging, helps stratifying patients and is useful for planning their follow-up and management [4] (see Figure 1 for details about CKD staging). The eGFR is a mathematically derived score based on patient's serum creatinine level, age, sex and race; it is usually computed by the laboratory that analyzes the patient's blood sample and reported with the serum creatinine result. Several recognized and well validated formulas have been used for this purpose, including the Modification of Diet in Renal Disease (MDRD) and Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equations [5].

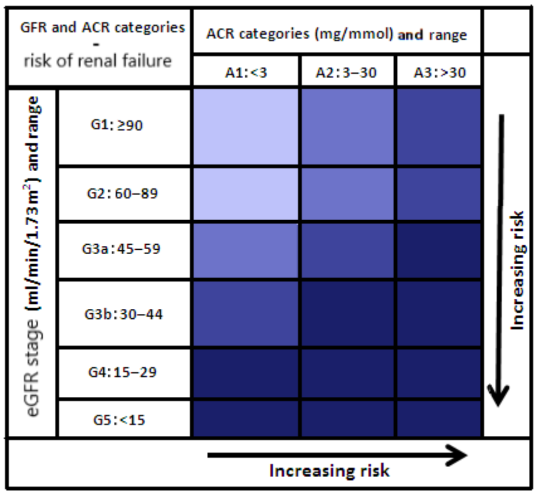

GFR and ACR categories - risk of renal failure
ACR categories (mg/mmol) and range
A1:<3
A2:3–30
A3:>30
eGFR stage (ml/min/1.73m^2) and range
G1: ≥90
G2: 60–89
G3a: 45–59
G3b: 30–44
G4: 15–29
G5: <15
Increasing risk
Increasing risk

Figure 1: Chronic Kidney Disease (CKD) classification based on Glomerular Filtration Rate (GFR) and Albumin-Creatinine Ratio (ACR).

However, it is important to bear in mind several pitfalls and cautions when interpreting the eGFR. First of all, it is only an estimate of kidney function and a relevant error is possible. Therefore, when assessing the CKD advancement, clinicians should be careful in interpreting the results and they should consider also other patient-related factors, relying also on eGFR multiple evaluations and overall trend, rather than

[*] *Corresponding* AUTHOR. Tel.: +39 3887296136.

E-mail addresses: piervincenzov@gmail.com (P. Ventrella), giovanni.delgrossi@asst-vimercate.it (G. Delgrossi), gianmichele.ferrario@asst-vimercate.it (G. Ferrario), marco.righetti@asst-vimercate.it (M. Righetti), marco.masseroli@polimi.it (M. Masseroli)

on the single observation of the latter one [6]. Furthermore, from a clinical point of view, it is currently very complex, if not impossible, to estimate adequately and sufficiently in advance when a patient with CKD will necessarily have to be dialyzed; whereas, it would be very useful for clinicians and patients to have a reasonable indication that such an event will happen within the next 6 or 12 months, or more.

The ability to develop mathematical models to categorize items based on the statistical analysis of data has made machine learning (ML) obtaining considerable success in medical diagnosis, with a strong growth of diagnostic questions for which ML algorithms are designed [7]. This has certainly been assisted by the digital revolution of the last few years, which made possible to record and archive various types of clinical data, made easily available and usable in digital format [8]. Examples of ML models in medical diagnosis can be found in the work of Lee *et al*. [9], where Conditional Inference Survival Forests and Random Survival Forests are employed for risk stratification of patients with type 2 diabetes for acute myocardial infarction and sudden cardiac death, or in the study of JoonNyoung *et al*. [10], where they show that ML algorithms can improve the prediction of long-term outcomes in ischemic stroke patients. Other examples are the studies of Senders *et al*. [11] for the prediction of survival in glioblastoma patients, or of Tse *et al*. [12] for the risk stratification in heart failure. Finally, Lee *et al*. [13] show that ML techniques can significantly improve overall risk stratification performance.

Regarding nephrological diseases, ML applications are currently being tested to improve the life of patients with CKD and to reduce treatment costs [14]. The main objectives of these applications are early detection and study of the evolution of CKD. Focusing on the second task, Norouzi *et al*. [15] proposed an adaptive neurofuzzy inference system that can accurately predict the GFR variations. Features such as weight, diastolic blood pressure, diabetes mellitus as underlying disease, and current eGFR showed significant correlation with GFR variations. With a different approach, Agarwal and Shah [16] tried to characterize the CKD progression patterns using clustering techniques. They demonstrated how two sub-groups of patients that display distinct patterns of disease progression may be compared on clinical attributes that correspond to the maximum difference in progression patterns. Recently, in the study of Makino *et al*. [17] big data machine learning techniques are used to predict the progression of diabetic kidney disease. Specifically, the authors apply both deep learning approaches to extract time-series data patterns from an Electronic Medical Record (EMR) and natural language processing techniques over textual data to extract patient's diagnosis and treatment information. These techniques were used to develop a binary classifier predicting aggravation or not of the GFR in a 6-month time spam with an overall accuracy of 70%. These studies set the basis for the analysis of the progression of CKD through machine learning approaches, but they fail in answering the important question that clinicians are most concerned about: "*given a patient affected by CKD, how soon will he/she necessarily have to be dialyzed?*"

In the study here discussed, we directly define, as a target of our computational prediction, the number of months within which the beginning of the alternative dialysis treatment would be required for a CKD patient. Then, using a supervised machine learning approach, a computational model, trained on the available data stored in the EMR of a hospital, is developed to accurately predict this interval of time. Thanks to it, clinicians can plan more effectively the next clinical encounter, scheduling it within a shorter or longer period of time, paying more attention to patients at higher risk. The resources used by the hospital (in terms of staff, department crowding, exam prescription, etc.) and the time and energy of the patient undergoing the clinical encounters can be remarkably optimized. The beginning of the dialysis treatment itself can be planned in advance with precision, allowing both clinicians and patients to organize themselves in the most appropriate manner. Additionally, the study allows better understanding of what are the clinical and physiological characteristics of the patient that most determine the speed of the CKD progression. The work has been developed in strict collaboration with the Vimercate Hospital's nephrologists, who contributed to all its phases, from the initial formulation of the diagnostic problem to be addressed, to the discussion of the results in order to verify their consistency.

2 Material and methods

In this Section we describe the steps followed to effectively develop our computational model, able to predict the time frame within which the beginning of the alternative dialysis treatment will be required for a CKD patient. First, the target variable for the supervised learning approach is defined. Then, the feature extraction and engineering, done over the considered information in order to build the dataset used for training and comparison of the supervised algorithms evaluated, is described. Finally, we briefly discuss how missing values are handled, how the feature selection is performed, and which machine learning approaches are employed and compared, including both classification and regression ones, as well as the different algorithms tested.

2.1 Addressed medical question and patient selection

We focus our study on the evolution of CKD between the G4 stage (or almost G4) and the beginning of the dialysis treatment, thus disregarding the previous staging cases G1, G2, G3a and partially G3b (Figure 1). This choice was taken following the discussions with the clinicians: the time interval between the first stages G1, G2 and G3 and the final stage is generally too wide, and it is complicated or even useless to start planning possible dialysis sessions since the first CKD stages. Furthermore, a patient in stage G4 is typically under clinical observation from an interval of time that allows having access to relevant data such as the trend of eGFR in the previous months or years, most recent laboratory test results and their trend, general state of health of the patient, etc.; thus, it is possible to better profile him/her. Therefore, the medical question we want to answer can be formulated in this way: "***given a patient affected by CKD in stage G4 or almost (i.e., ending of G3b), how soon will he/she necessarily be dialyzed?***"

For this study we consider all patients in the Vimercate Hospital EMR in stage G4 or close (with an eGFR test value below 35 ml / min / 1.73 m^2) and who were subsequently dialyzed. The target variable (from now on named '*months until dialysis*') is defined as the elapsed number of months between a patient's clinical encounter and the beginning of his/her dialysis treatment. Chronic Kidney Disease and Acute Kidney Disease cases are differentiated, discarding the latter ones because they are not useful for the study purpose. Eventually, 906 distinct patients have been selected; we retrieved their data regarding 4,266 patient's eGFR measurements (with four or more measurements per patient) from different clinical encounters, with the associated date (defined from now on '*last observation date*'), and the corresponding beginning of the dialysis session, with the relative date.

2.2 Considered data and their extraction

To answer the defined medical question, the data collected in the Vimercate Hospital EMR system since the early 2000s is considered. Data from different databases of the hospital infrastructure is integrated and information from both structured data and unstructured textual medical reports is extracted.

2.2.1 Structured information from EMR querying

Several patient's structured data are extracted from the hospital EMR, including age (calculated as the difference in years between the patient's date of birth and the last observation date) and eGFR values. For the analysis of the patient's eGFR trend, the most recent value and the mean and standard deviation of the eGFR values are considered, both over the last 4 months and the last year of observation. Other clinically relevant blood, urine and general laboratory tests are considered as well, and for each of them the most recent value preceding or referring to the last observation date and the mean and standard deviation of the values in the last 4 months of observation are retrieved.

2.2.2 Unstructured information from medical report text mining

For each considered patient, the most recent textual medical report preceding or referring to the last observation date is retrieved, and from it a lot of relevant information is extracted, such as whether:

- the patient is diabetic, anemic, obese, or he/she has hypertension episodes or other associated pathologies,
- in the patient's family there have been cases of diabetes, cardiopathy, or hypertension episodes,
- the patient has kidney stones, or a solitary kidney,
- the patient is a smoker, or he/she used to smoke,
- the patient had a renal transplant.

To extract this information, we use simple text mining techniques based on keywords and regular expression matching. When checking if in the text there is a specific condition, first a vocabulary is defined, consisting in a set of keywords composed by the name of the condition (e.g., 'diabetes') and a list of synonyms (e.g., 'high blood sugar', 'high glycosuria', 'hyperglycemia', ...). Then stemming, a common procedure in text mining that consists in extracting the base or root form of a word, is applied to these keywords in order to improve their recall. Finally, through pattern matching, the presence in the text of any of the stemmed keywords is checked, paying attention that they are not negated (it is common to find sentences like "the patient does not have diabetes", "hypertension: absent/not present", etc.). Moreover, always through pattern matching, the cases when the condition refers to the patient himself/herself or to a member of the family are differentiated. For the cases where only one or more synonyms, but not the name of the condition, are automatically found, the medical report is manually checked to confirm or discard the condition (e.g., diabetes) as associated with the patient.

2.3 Extracted data preprocessing

The extracted dataset is split into training set (with about 75% of the patient's clinical encounters) and test set (regarding the remaining 25% of encounters) in a stratified manner (i.e., keeping a similar distribution of both target variable and input features between training and test set), using the Scikit-learn Python library [18].

From the training set, the features with more than 40% of missing values and the clinical encounters with at least half of the features missing (less than 9% of the cases in the training set) are discarded. Then, the remaining missing values are imputed simply with the mean of the values of their feature in the training set; we consider this adequate since no evident relation exists between the missing value distribution of an input feature and the target variable, or the other input variables. The exact same imputation procedure is applied also to the test set, but with the mean of the values in the training set.

2.4 Feature selection

To reduce the probability of overfitting and improve the generalization capabilities of predictive algorithms, a feature selection over the initial set of input variables is advised. Reducing the number of variables used for training has also the advantage of lowering the amount of information needed by an algorithm to perform the prediction, thus making the obtained computational models more usable.

We perform backward feature elimination, a recursive feature selection that, starting from the original set of variables, eliminates one feature at the time until lowering the performance of the considered prediction model. The selected feature subset is retrieved by performing the mentioned procedure over the initial best performing model. Therefore, in selecting the features to eliminate first, the feature ranking of the initial best performing model and the correlation ranking of these features are taken into account, starting to drop out the variables that show small importance for the model and/or low correlation with the target variable.

2.5 Approaches and algorithms compared

To predict when the alternative dialysis treatment will be required, we approach the problem with two different machine learning supervised techniques: classification and regression.

Using a classification approach, a priori we define default intervals of months and then the different trained algorithms are compared through common classification performance metrics, including overall accuracy, precision, recall and F1-score, obtained by 10-fold cross-validation. Eventually, also confusion matrices are used to analyze false positive, false negative, true positive, and true negative classifications. These metrics allow understanding how well a model discriminates between cases belonging to one class or another, but in the case of misprediction of an instance it is difficult to say if the model misclassified it because it was difficult to be classified or because the model was completely wrong.

On the other hand, addressing the problem as a regression and trying to predict the exact number of months within which the patient will need to be dialyzed allows analyzing the trained models also with other metrics, such as the Mean Square Error (MSE) or Root Mean Square Error (RMSE), which provide further information on the model ability to predict the cases.

To compare classification models with regression ones and understand which approach works better for our problem, we assign the output of a regression model to a class by performing different discretizations depending on the number of classes considered, such as:

$$class \; = \begin{cases} '\text{within 1 year}', & if \text{ predicted months} < 12 \\ 'after\ 1\ year', & if\ predicted\ months \geq 12 \end{cases}$$

for two classes, or defining more intervals of months (more classes):

$$class \; = \begin{cases} 1st, & if\ predicted\ months \; < \; 6 \\ 2nd, & if\ 6 \leq \; predicted\ months \; < \; 18 \\ 3rd, & if\ predicted\ months \; \geq \; 18 \end{cases}$$

for three classes, or for four classes as follows:

$$class \; = \begin{cases} 1st, & if\ predicted\ months \; < \; 6 \\ 2nd, & if\ 6 \; \leq predicted\ months \; < \; 14 \\ 3rd, & if\ 14 \; \leq \; predicted\ months \; < \; 24 \\ 4th, & if\ predicted\ months \; \geq \; 24 \end{cases}$$

Regarding the compared algorithms, using the open source Scikit-Learn Python library for machine learning [18], we focus on:

- *Logistic Regression*, a simple but effective binary classifier providing a useful baseline,
- *Decision Trees* (DT), an intuitive but powerful algorithm that effectively manages highly correlated features,

- *Random Forest*, *Extremely Randomized Trees* and *Gradient Tree Boosting*, DT ensemble techniques that aim to improve the performance of a single DT,
- (Fully connected) *Neural Networks*, known for achieving the best performances in a variety of different problems, if enough training data are provided.

All these algorithms can be used both for regression and classification. Parameter tuning has been applied through cross-validation for selecting the hyperparameters of the different algorithms (i.e., max_depth of the Decision Trees, number of trees in the used ensembles, number of layers in the Neural Networks). Binary trees (order = 2) have been used for the Decision Trees algorithm and its ensembles, while the Gini index has been used as splitting criteria to evaluate the best feature to be used in each node of the tree. The hyperparameter max_depth is used as stopping criteria for the Decision Trees algorithm; for the algorithms of ensemble of decision trees instead, no stopping criteria is specified.

3 Results

3.1 The dataset

The considered patients' data extracted from the Vimercate Hospital EMR regard 58 different clinical features (not considering the ones discarded during data preprocessing) listed in Table 1; they are used for the development of our computational model as input variables, together with the target variable '*months until dialysis*', which indicates the number of months up to the occurrence of the complete renal failure with consequent beginning of the dialysis treatment.

Table 1: List of the 58 clinical features considered; the 27 selected for model development are in bold.

Not selected features	Selected features
aspartate aminotransferase delta	**age**
cardiopathic family	**anemic**
cataract	**aspartate aminotransferase**
chlorine delta	**cardiopathic**
cirrhosis	**chlorine**
corpuscular hemoglobin concentration	**creatinine**
corpuscular hemoglobin concentration delta	**creatinine delta**
diabetic family	**diabetic**
ex smoker	**erythrocytes**
glucose	**erythrocytes delta**
glucose delta	**GFR delta last 4 months**
hematocrit delta	**GFR delta last year**
hemoglobin delta	**GFR standard deviation last 4 months**
hypercholesterolemia	**GFR standard deviation last year**
Hypertension episodes family	**hematocrit**
hyperthyroid	**hemoglobin**
kidney stones	**hypertension episodes**
leukocytes	**last GFR**
leukocytes average	**male**
leukocytes delta	**mean corpuscular hemoglobin**
mean corpuscular hemoglobin delta	**mean corpuscular volume**
mean corpuscular volume delta	**potassium**
obese	**sodium**
potassium delta	**specific gravity standard deviation**
smoker	**urate**
sodium delta	**urea**
solitary kidney	**urea delta**
specific gravity	
specific gravity delta	
transplanted	
urate delta	

After data preprocessing, the training set is composed of 2,911 clinical encounters; out of these encounters, 74% are related to patients with hypertension episodes, 68% regard male patients, 58% refer to the beginning of the dialysis treatment within 1 year, 41% are associated with diabetic patients, 30% with anemic patients, 25% refer to cardiopathic patients, 17% to patients with kidney stones, 8% are associated with obese patients, 7% with patients who previously had a kidney transplant, 3% with patients with a solitary kidney and 3% with patients suffering of hypercholesterolemia. The average patients' age is 68 years. Thanks to the stratified sampling used to extract the test set, it presents a distribution very similar to the training set one. The number and percentage of samples for each class are reported in Table 2 both for the training set and the test set.

Table 2: Number and percentage of samples of each defined class.

N. of classes	Class	Training set samples		Test set samples	
		Count	Percentage	Count	Percentage
2 classes	1st	1,717	58.98%	619	58.01%
	2nd	1,194	41.02%	448	41.99%
3 classes	1st	1,164	39.99%	435	40.76%
	2nd	934	32.09%	340	31.87%
	3rd	813	27.92%	292	27.37%
4 classes	1st	1,164	39.99%	435	40.77%
	2nd	707	24.29%	255	23.90%
	3rd	454	15.60%	139	13.03%
	4th	586	20.12%	238	22.31%

Figure 2 shows the distribution of the target variable in the training set. It can be noticed that it is skewed on the left, with 95% of the considered encounters occurring no more than 47 months (about 4 years) before dialysis. Hence, we focus the analysis and predictions on the first 3-4 years following the considered medical check.

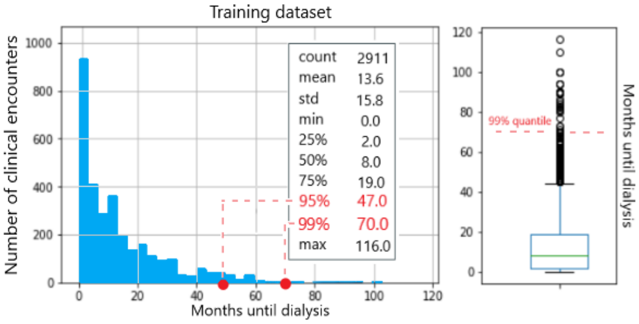


Figure 2: Target variable distribution in the training data set.

3.2 Evaluation of approaches and algorithms

We evaluated the considered algorithms using both classification and regression approaches, as well as different feature sets. Table 3 reports the cross-validation performances in predicting the time to dialysis using binary (within one year or later) classification algorithms trained with all 58 input variables, or with only the features selected through backward feature elimination on the initial best performing algorithm, i.e., the Extremely Randomized Trees classifier (27 in total, in bold in Table 1) and when not considering the aggregated features (mean and standard deviation of laboratory test results), or when disregarding information extracted from textual medical reports. Similarly, Table 4 reports the cross-validation performances when predicting directly the time to dialysis using regression algorithms instead.

Comparing Table 3 and Table 4 it can be noticed that using a classification approach to the problem leads to better results than with a regression approach (also when defining more than 2 classes, see Table 5 and Table 6). This is probably due to the high difficulty of

Table 3: Comparison of cross-validation performances of binary classification algorithms using different sets of features.

Classifier	Feature set	Cross-validation			
		Accuracy	Precision	Recall	F1-score
Decision Trees	All features	0.84 ± 0.01	0.85	0.85	0.85
	Selected features	0.84 ± 0.01	0.85	0.85	0.85
	No aggregated features	0.81 ± 0.02	0.78	0.78	0.78
	No textual reports	0.82 ± 0.02	0.80	0.80	0.80
Random Forest	All features	0.91 ± 0.01	0.93	0.93	0.93
	Selected features	0.91 ± 0.02	0.94	0.94	0.94
	No aggregated features	0.88 ± 0.01	0.86	0.86	0.86
	No textual reports	0.90 ± 0.01	0.91	0.90	0.90
Extremely Randomized Trees	All features	0.94 ± 0.01	0.93	0.93	0.93
	Selected features	0.94 ± 0.01	0.95	0.95	0.95
	No aggregated features	0.90 ± 0.02	0.92	0.92	0.92
	No textual reports	0.91 ± 0.01	0.92	0.92	0.92
Gradient Tree Boosting	All features	0.93 ± 0.01	0.92	0.92	0.92
	Selected features	0.93 ± 0.01	0.94	0.94	0.94
	No aggregated features	0.89 ± 0.02	0.89	0.89	0.88
	No textual reports	0.91 ± 0.02	0.89	0.89	0.89
Neural Networks	All features	0.90 ± 0.01	0.90	0.90	0.90
	Selected features	0.88 ± 0.02	0.89	0.89	0.89
	No aggregated features	0.81 ± 0.02	0.85	0.85	0.85
	No textual reports	0.82 ± 0.02	0.84	0.84	0.84
Logistic Regression	All features	0.73 ± 0.02	0.75	0.75	0.75
	Selected features	0.74 ± 0.02	0.74	0.72	0.73
	No aggregated features	0.71 ± 0.02	0.66	0.66	0.66
	No textual reports	0.71 ± 0.02	0.71	0.71	0.71

directly predicting the exact number of months until the dialysis; in fact, even the best regression model has a cross-validation RMSE of about 7 months. Therefore, we focus on the classification approach.

It can be also noticed that considering only the selected 27 features (rather than using all features) does not worsen the overall performance of the classification algorithms, except for the Neural Networks, which however have worse performances than any of the ensemble classifiers considered. Conversely, when omitting aggregated features or information from textual medical reports, the algorithms perform remarkably worse. This proves the effectiveness of the feature engineering and feature extraction techniques that we employed. Using the text mining algorithm adopted for extracting structured information from textual medical reports resulted in a simple, but effective, approach. We also considered the procedure of deducing the patient's comorbidities from the structured pharmacotherapy information by using the Anatomical Therapeutic Chemical (ATC) Classification System [19]. ATC identifies the main anatomical group and the main therapeutic group which the patient administered drug refers to. However, using the ATC code it is not possible to uniquely trace back the patient's clinical conditions (a specific drug can be administered for different pathologies). Conversely, using text mining techniques it is possible to remarkably improve precision and recall in retrieving these pathologies. Moreover, other information such as whether the patient is a smoker, or he/she has cirrhosis, solitary kidney, or kidney stones and the clinical family history could not be retrieved without text mining or more sophisticated, but difficult to implement, natural language processing techniques.

Table 4: Comparison of cross-validation performances of regression algorithms using different sets of features.

Regressor	Feature set	Cross-validation				
		RMSE	Accuracy	Precision	Recall	F1-score
Decision Trees	All features	12.03 ± 1.46	0.81 ± 0.02	0.82	0.82	0.82
	Selected features	11.32 ± 1.31	0.78 ± 0.03	0.79	0.79	0.79
	No aggregated features	13.52 ± 0.97	0.79 ± 0.01	0.80	0.80	0.80
	No textual reports	12.46 ± 1.05	0.80 ± 0.02	0.81	0.81	0.81
Random Forest	All features	8.05 ± 0.77	0.86 ± 0.01	0.86	0.86	0.86
	Selected features	8.12 ± 0.88	0.86 ± 0.01	0.85	0.85	0.85
	No aggregated features	9.03 ± 0.99	0.83 ± 0.01	0.83	0.83	0.83
	No textual reports	8.75 ± 0.64	0.85 ± 0.02	0.83	0.83	0.83
Extremely Randomized Trees	All features	6.99 ± 0.87	0.90 ± 0.02	0.91	0.91	0.91
	Selected features	7.28 ± 0.89	0.89 ± 0.02	0.89	0.89	0.89
	No aggregated features	9.03 ± 0.99	0.83 ± 0.01	0.83	0.83	0.83
	No textual reports	8.05 ± 0.94	0.87 ± 0.02	0.87	0.87	0.87
Gradient Tree Boosting	All features	7.42 ± 0.86	0.85 ± 0.01	0.83	0.83	0.83
	Selected features	7.54 ± 1.27	0.85 ± 0.02	0.86	0.86	0.86
	No aggregated features	8.53 ± 0.59	0.82 ± 0.02	0.79	0.79	0.79
	No textual reports	7.86 ± 0.69	0.85 ± 0.02	0.84	0.84	0.84
Neural Networks	All features	11.57 ± 5.99	0.85 ± 0.02	0.86	0.86	0.86
	Selected features	11.96 ± 5.99	0.80 ± 0.02	0.84	0.84	0.84
	No aggregated features	13.62 ± 4.26	0.73 ± 0.03	0.81	0.81	0.81
	No textual reports	13.47 ± 4.26	0.74 ± 0.03	0.81	0.81	0.81

Table 5 shows the cross-validation performances in the time to dialysis prediction of the considered classification algorithms when using the 27 features selected and considering the defined 3 or 4 classes, i.e., month intervals. Notice that the algorithm performances do not decrease considerably with respect to the binary classification ones, suggesting the possibility to predict the beginning of the dialysis session with more granularity. Clearly, the shorter the specified intervals, the more error-prone the algorithms are, but the right trade-off can be found. Finally, Table 3 and Table 5 show that, independently on the number of specified classes, the best performing algorithms are the ones using an ensemble of decision trees. Particularly, the Extremely Randomized Trees (shortly ExtraTrees) classifier is the best performing one for all number of classes considered, both in terms of accuracy and F1-score. Thus, we select the ExtraTrees classification algorithm for the time to dialysis prediction. In the next Section, we report its performances computed on the unseen test set. As for the number of trees used as hyperparameter in the ExtraTrees classifier, cross-validation shows that the best performance is obtained with 180 trees. Similarly, the optimal number of trees for the Random Forest classifier is 250. For the Decision Trees algorithm instead, the best performance is obtained with max_depth equal to 15 and the generated tree has about 320 leaves.

Table 5: Comparison of cross-validation performances of classification algorithms using the 27 features selected and the 3 or 4 classes defined.

Classifier	N. of Classes	Accuracy	Precision	Recall	F1-score
Decision Trees	3	0.76 ± 0.03	0.79	0.79	0.79
	4	0.75 ± 0.03	0.76	0.76	0.76
Random Forest	3	0.87 ± 0.01	0.89	0.89	0.89
	4	0.85 ± 0.03	0.89	0.89	0.89
Extremely Randomized Trees	3	0.90 ± 0.02	0.92	0.92	0.92
	4	0.89 ± 0.02	0.90	0.90	0.90
Gradient Tree Boosting	3	0.86 ± 0.01	0.88	0.88	0.88
	4	0.81 ± 0.02	0.88	0.88	0.88
Neural Networks	3	0.84 ± 0.02	0.87	0.87	0.87
	4	0.82 ± 0.02	0.86	0.86	0.86

Regarding Table 6, indeed the RMSE remains the one in Table 4; what changes are the classification metrics, according to the discretization formula used to map the continuous predictions to the discrete output (i.e., the corresponding classes).

Table 6: Comparison of cross-validation performances of regression algorithms using the 27 features selected and the mapping to the 3 or 4 classes defined.

Regressor	N. of Classes	Accuracy	Precision	Recall	F1-score
Decision Trees	3	0.71 ± 0.02	0.71	0.71	0.71
	4	0.68 ± 0.03	0.70	0.69	0.69
Random Forest	3	0.71 ± 0.03	0.76	0.71	0.71
	4	0.66 ± 0.02	0.76	0.65	0.66
Extremely Randomized Trees	3	0.79 ± 0.03	0.80	0.75	0.76
	4	0.75 ± 0.03	0.78	0.75	0.75
Gradient Tree Boosting	3	0.72 ± 0.02	0.76	0.71	0.72
	4	0.68 ± 0.02	0.69	0.66	0.66
Neural Networks	3	0.72 ± 0.02	0.74	0.74	0.74
	4	0.67 ± 0.03	0.68	0.65	0.66

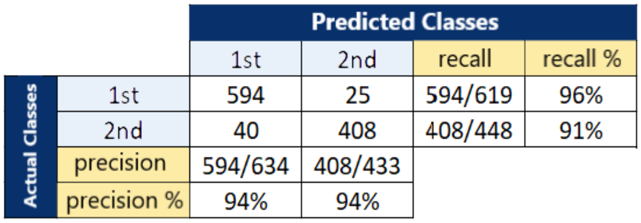
3.3 Extremely Randomized Trees classifier models

The performances of the ExtraTrees classifier using the 27 selected features when applied on the test set are reported in Table 7; they are equivalent to the ones obtained for the same classifier during cross-validation, showing that the chosen models are not overfitted.

Table 7: Performances of the proposed models (ExtraTrees classifiers) obtained on the unseen test set using the 27 selected features.

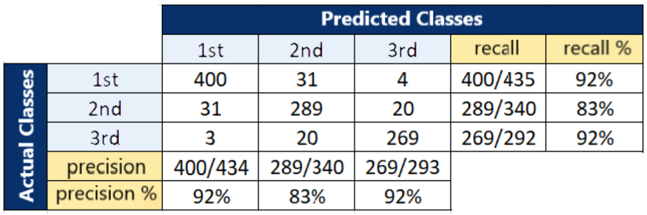
N. of Classes	Accuracy	Precision	Recall	F1-score	Specificity
2	0.94	0.96	0.96	0.93	0.91
3	0.91	0.93	0.93	0.93	0.91
4	0.87	0.90	0.89	0.89	0.87

The confusion matrices of the proposed models are shown in Figure 3, Figure 4 and Figure 5, where the precision and recall in predicting each specific class are also reported. These matrices show that the models manage to correctly classify the great majority of the test set new cases, with greater precision for the classes at the extremes of the overall time interval considered for the prediction (i.e., the ones closest or most distant in time) and with slightly less precision for the intermediate classes, where some cases are classified in the adjacent classes; this is well adequate for the intended clinical purposes of personalized care.



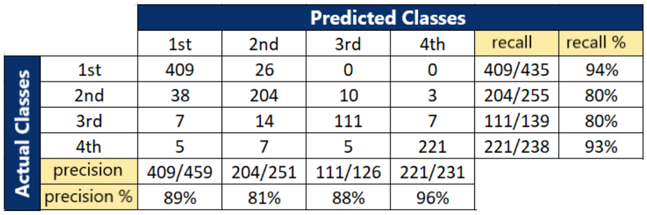
		Predicted Classes			
		1st	2nd	recall	recall %
Actual Classes	1st	594	25	594/619	96%
	2nd	40	408	408/448	91%
	precision	594/634	408/433		
	precision %	94%	94%		

Figure 3: Confusion matrix of the ExtraTrees binary model.



		Predicted Classes				
		1st	2nd	3rd	recall	recall %
Actual Classes	1st	400	31	4	400/435	92%
	2nd	31	289	20	289/340	83%
	3rd	3	20	269	269/292	92%
	precision	400/434	289/340	269/293		
	precision %	92%	83%	92%		

Figure 4: Confusion matrix of the ExtraTrees model with 3 classes.



		Predicted Classes					
		1st	2nd	3rd	4th	recall	recall %
Actual Classes	1st	409	26	0	0	409/435	94%
	2nd	38	204	10	3	204/255	80%
	3rd	7	14	111	7	111/139	80%
	4th	5	7	5	221	221/238	93%
	precision	409/459	204/251	111/126	221/231		
	precision %	89%	81%	88%	96%		

Figure 5: Confusion matrix of the ExtraTrees model with 4 classes.

3.4 Implementation and dataset availability

At [20], the collected dataset used for the analysis, together with the Python code for computing an overview of the dataset and for training and testing the proposed computational models, and the developed computational models for their direct use are publicly available.

3.5 Relevant factors for the dialysis onset

To identify the selected features most relevant in predicting the time to dialysis of CKD patients, we compute both the feature importance ranking, based on the feature coefficients of the ExtraTrees binary classifier (Figure 6), and the feature correlation ranking with the target variable '*months until dialysis*' (Figure 7). The former one helps identifying which features are more relevant for the prediction according to the proposed binary model. Instead, the latter one is based on the value of the Pearson's correlation coefficients computed over the training data set; in Figure 7 we report them, with in dark red the features resulting inversely correlated with the target variable (i.e., representing factors associated with a rapid onset of renal failure) and in light blue the features directly correlated with the target variable.

Figure 7 shows that the last observed values of ***creatinine*** and ***urea*** are the features most correlated, and inversely, with the target variable; thus, they are associated with a rapid onset of complete renal failure.

Creatinine is a chemical waste produced in the muscles and filtered by the kidneys. If the kidneys are not functioning properly, the amount of creatinine in the blood increases. Thus, as our study shows, the level of creatinine in the blood is a reliable indicator of kidney's functionality. Similar considerations can be done for urea, as well as '***creatinine delta***' and '***urea delta***', both the latter ones computed as the difference between their last test result observed and the mean of their previous test results over the last 4 months of observation. The increase of these delta values means that the values of creatinine and urea increased from their last clinical check, highlighting a rapid degradation of the renal function; the opposite indicates their decrease. The same reasoning applies to '***GFR delta last 4 months***' and, even if less correlated, to '***GFR delta last year***', both computed in the opposite way, i.e., as the difference between the mean of the previous results and the last observed test result; thus, a big positive delta value means that the eGFR remarkably decreased in the last period, the opposite for a negative delta value. The last value of GFR (***last GFR***) is an important indicator and, as expected, it results directly correlated to a wider time span before the necessary start of the dialysis treatment. These features, overall, occupy a high ranking position also in Figure 6.

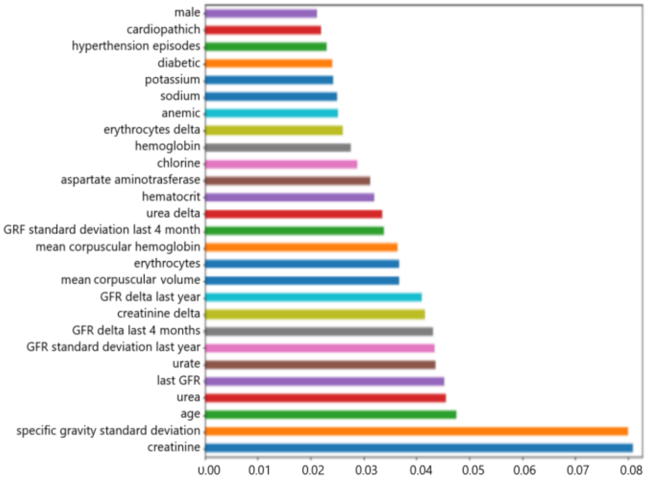


Figure 6: Feature importance ranking of the ExtraTrees binary model.

Other variables associated with slower advancement of the disease are the last observations of ***erythrocytes***, ***hematocrit*** and ***hemoglobin***, which are indicators of the quantity and percentage of red blood cells in the body and are linked to the general well-being of the patient. Both the feature ranking and the correlation analysis show that another important factor is the ***standard deviation of specific gravity*** (a urine test that compares the density of urine with the density of water). Interestingly, the binary features '***diabetic***', '***cardiopathic***', and '***hypertension episodes***' occupy a higher position in the correlation ranking with respect to their position in the feature importance ranking. Their low position in the feature ranking can be explained by the intrinsic nature of these variables: Decision Trees algorithms intrinsically perform feature selection by selecting the appropriate branch split points through the information gain, or Gini index criteria; their basic idea is that the more often a feature is used in the split points of a tree, the more important that feature is. Therefore, continuous variables that can be discretized and used for splitting several times are better ranked than binary variables that can be used for splitting only once. As known, the presence of such comorbidities worsens the general well-being of the patient; from Figure 7, where these features are shown in red, we can deduce that people affected by these pathologies are more prone to rapid worsening of kidney functionalities. These results are confirmed by previous medical investigations showing that the CKD incidence increases in people affected by diabetes, high blood pressure, or other comorbidities [3, 4, 15, 21]. Our results confirm also that an advanced ***age*** is related to a rapid worsening of renal conditions [3, 4, 21]. Finally, the ***potassium***, ***sodium,*** and ***anemic*** features do not seem to have a strong impact on the prediction of the dialysis onset. The scarce relevance of the binary variable 'anemic' can be explained by the fact that the presence or absence of anemia per se is less important than the actual degree of anemia, expressed by the laboratory results of the hematocrit and erythrocyte levels. Our results are also in line with the medical literature and the opinion of the Nephrologists who supervised the study.

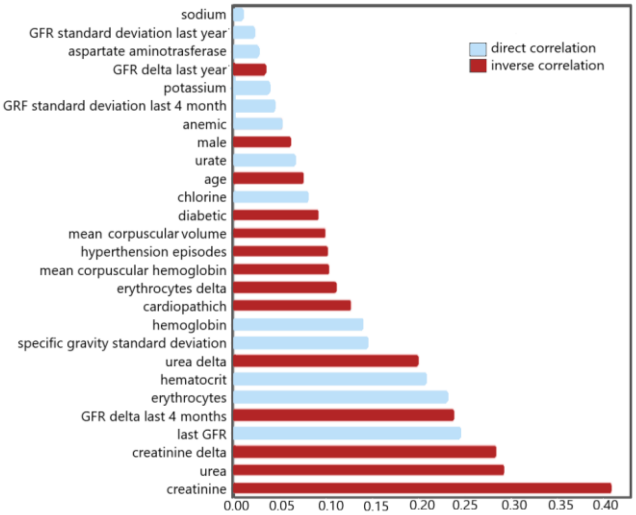


Figure 7: Correlation ranking of the selected features with the target variable '*months until dialysis*'.

4 Discussion

The conducted study shows that several factors influence the CKD progression rate. The ones resulted more important for predicting the dialysis onset include, in addition to the patient's age and main comorbidities, recent observations on the trends of the amount of creatinine, urea and red cells in the blood, urine specific gravity and eGFR. Interestingly, the latter one does not even appear to be the most important.

The developed computational method is meant to enable clinicians to reliably estimate in advance by when the dialysis treatment will be necessary for a patient; additionally, it allows specifying different levels of granularity for the target time interval estimation, corresponding to different reliability levels. This is achieved by training the ExtraTrees classification algorithm multiple times with an increasing number of predefined classes, eventually providing different classification models to be used for different numbers of classes; the appropriate model is then used according to the granularity specified by the clinician.

The availability of our method allows best planning of the next clinical check or of the beginning of the dialysis treatment, prioritizing the controls of patients at risk and allowing clinicians and end-staging patients to organize themselves in the most appropriate manner. Furthermore, it provides more information on the CKD progression of a specific patient, by analyzing how the computational predictions change from a clinical check to the subsequent ones with respect to the administered treatment, the lifestyle, or the diet of the patient.

Regarding the different machine learning approaches evaluated, tackling our problem with a classification approach leads to better results than when using a regression approach. Focusing on classification algorithms, ensemble techniques based on decision trees are the key to achieve good results in the addressed problem. First, given the possible limited size of the data set, it is important to avoid overfitting, which ensemble methods can help eluding. Second, since the interpretation of results is a critical aspect in the clinical context, decision trees are intuitive and explainable algorithms, therefore preferable to more complex computational architectures.

Regarding the feature engineering phase, the obtained results show that the last observed values and their simple aggregation statistics, such as mean, trend and standard deviation, allow describing reliably the clinical status of a patient with a limited number of attributes; these allow generating much simpler and more understandable models, avoiding the use of more complex computational architectures such as recurrent neural networks. The latter ones are commonly used for effective pattern analysis of time series, but they require much more data, often not available in a clinical context.

The obtained results also show that even simple text mining techniques, based primarily on keyword search in the patient's textual medical history, allow extracting relevant patient information that would otherwise be impossible to retrieve, because not present in a structured form within a hospital database.

We are aware of the limitations of the developed predictive models, since the study is carried out in a single Center and is not reevaluated yet using data sets from EMRs of other Institutions. For example, the population we consider for the analysis is composed only of people of Caucasian ethnicity, but the medical literature suggests that the ethnicity could be another important factor in the CKD development. Unfortunately, it is difficult to find public datasets adequate for our purpose, since the analysis of the CKD evolution still needs to be deepened. Furthermore, the overall accuracy of the defined model may be improved considering other relevant information, such as patient's level of proteinuria, diet, body mass index (BMI), drug therapy, or the assessment scales compiled by physicians or nurses, which were not available or had an excessive percentage of missing entries for the conducted study.

We have integrated the developed computational models in the Vimercate Hospital informatics infrastructure to further evaluate them; the goal is a long-term prospective study aimed at assessing the improved diagnostic accuracy of the clinicians when coupling their experience and knowledge to the use of such models.

5 Conclusions

The current assessment scales of CKD progression, mainly based on eGFR and already proven to be ineffective for predicting in advance when a CKD patient will necessarily have to be dialyzed, can be improved by considering multiple factors through machine learning techniques, which allow a reliable prediction of a CKD patient at dialysis risk in a short or longer time. The promising results obtained show that machine learning techniques allow developing a computational model that, coupled with the knowledge and experience of the clinicians, can effectively lead to better personalized care and strategic planning of both patient's needs and hospital resources.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgement

We thank the Information Systems team of the Vimercate Hospital for helping in the analysis of the raw data available in the hospital infrastructure and for the integration of the proposed computational models within the Clinical Decision Support System of the hospital.

REFERENCES

[1] Lora CM, Daviglus ML, Kusek JW, Porter A, Ricardo AC, Go AS *et al.* Chronic Kidney Disease in United States Hispanics: A Growing Public Health Problem. *Ethnicity & Disease*. 2009; 19(4): 466–472.

[2] Center for Disease Control and Prevention. Chronic Kidney Disease in the United States, 2019. https://www.cdc.gov/kidneydisease/publications-resources/2019-national-facts.html . Accessed on May 27th, 2021.

[3] Jha V, Garcia-Garcia G, Iseki K, Li Z, Naicker S, Plattner B *et al.* Chronic kidney disease: global dimension and perspectives. *ScienceDirect*. 2013; 382(9888): 260-272.

[4] Johnson CA, Levey AS, Coresh J, Levin A, Lau J, Eknoyan G. Clinical Practice Guidelines for Chronic Kidney Disease in Adults: Part 1. Definition, Disease Stages, Evaluation, Treatment, and Risk Factors. *American Family Physician*. 2004; 70(5): 869-876.

[5] Levey AS, Stevens LA. Estimating GFR Using the CKD Epidemiology Collaboration (CKD-EPI) Creatinine Equation: More Accurate GFR Estimates, Lower CKD Prevalence Estimates, and Better Risk Predictions. *American Journal of Kidney Diseases*. 2010; 55(4): 622-627.

[6] Michels WM, Grooterdorst DC, Verdujin M, Elliott EG, Dekker FW, Krediet RT. Performance of the Cockcroft-Gault, MDRD, and New CKD-EPI Formulas in Relation to GFR, Age, and Body Size. *Clinical Journal of American Society of Nephrology*. 2010; 5(6): 1003-1009.

[7] Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *JAMA*. 2018; 319(13): 1317-1318.

[8] Edmund LSC, Ramaiah CK, Gulla SP. Electronic Medical Records Management Systems: An Overview. *DESIDOC Journal of Library & Information Technology*. 2009; 29(6): 3-12.

[9] Lee S, Zhou J, Guo LC, Wong WT, Liu T, Wong ICK *et al.* Predictive scores for identifying patients with type 2 diabetes mellitus at risk of acute myocardial infarction and sudden cardiac death. *Endocrinology, Diabetes & Metabolism*. 2021; 00: e00240.

[10] JooNyung H, Jihoon GY, Hyungjong P, Young DK, Hyo SN, Ji HH. Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke. *Stoke*. 2019; 50(5):1263-1265.

[11] Senders JT, Staples P, Mehrtash A, Cote DJ, Taphoorn MJB, Reardon DA *et al.* An Online Calculator for the Prediction of Survival in Glioblastoma Patients Using Classical Statistics and Machine Learning. *Neurosurgery*. 2020; 86(2): E184-E192.

[12] Tse G, Zhou J, Woo SWD, Ko CH, Lai RWC, Liu T *et al.* Multi-modality machine learning approach for risk stratification in heart failure with left ventricular ejection fraction ≤ 45%. *ESC Hearth Failure*. 2020; 7(6): 3716-3725.

[13] Lee S, Zhou J, Li KHC, Leung KSK, Lakhani I, Liu T *et al.* Territory-wide cohort study of Brugada syndrome in Hong Kong: predictors of long-term outcomes using random survival forests and non-negative matrix factorisation. *Open Heart*. 2021; 8(1): e1505.

[14] Sennaar K. AI Applications for Managing Chronic Kidney Disease. https://emerj.com/ai-sector-overviews/ai-managing-chronic-kidney-disease/ . Accessed on May 27th, 2021.

[15] Norouzi J, Yadollahpour A, Mirbagheri SA, Mazdeh MM, Hosseini SA. Predicting Renal Failure Progression in Chronic Kidney Disease Using Integrated Intelligent Fuzzy Expert System. *Computational and Mathematical Methods in Medicine*. 2016; 2016; 6080814.

[16] Agarwal V, Shah NH. Learning attributes of disease progression from trajectories of sparse lab values. *Pacific Symposium on Biocomputing*. 2017. 2017; 22:184-194.

[17] Makino M, Yoshimoto R, Ono M, Itoko T, Katsuki T, Koseki A *et al.* Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. *Scientific Reports*. 2019; 9(1): 11862.

[18] Pedrosa F, Varoquaux G, Gramfort A, Michel V, Thirison B *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011; 12: 2825-2830.

[19] Wikipedia. Anatomical Therapeutic Chemical Classification. https://en.wikipedia.org/wiki/Anatomical_Therapeutic_Chemical_Classification_System . Accessed on May 27th, 2021.

[20] Ventrella P, Delgrossi G, Ferrario G, Righetti M, Masseroli M. Dataset and Python code for CKD advancement assessment. https://github.com/DEIB-GECO/CKD_advancement_assessment . Accessed on May 27th, 2021.

[21] Kazancioglu R. Risk factors for chronic kidney disease: an update. *Kidney International Supplement (2011)*. 2013; 3(4): 368-371.