Wireless Networks (2022) 28:1269–1277
https://doi.org/10.1007/s11276-018-01905-4

A data mining approach to classify serum creatinine values in patients undergoing continuous ambulatory peritoneal dialysis

Claúdia Brito[1] · Marisa Esteves[2] · Hugo Peixoto[2] · António Abelha[2] · José Machado[2]

Published online: 25 January 2019

Abstract
Continuous ambulatory peritoneal dialysis (CAPD) is a treatment used by patients in the end-stage of chronic kidney diseases. Those patients need to be monitored using blood tests and those tests can present some patterns or correlations. It could be meaningful to apply data mining (DM) to the data collected from those tests. To discover patterns from meaningless data, it becomes crucial to use DM techniques. DM is an emerging field that is currently being used in machine learning to train machines to later aid health professionals in their decision-making process. The classification process can found patterns useful to understand the patients' health development and to medically act according to such results. Thus, this study focuses on testing a set of DM algorithms that may help in classifying the values of serum creatinine in patients undergoing CAPD procedures. Therefore, it is intended to classify the values of serum creatinine according to assigned quartiles. The better results obtained were highly satisfactory, reaching accuracy rate values of approximately 95%, and low relative absolute error values.

Keywords Data mining · Knowledge extraction · Chronic kidney diseases · Continuous ambulatory peritoneal dialysis · Serum creatinine · Clinical decision support systems · Weka · Classification algorithms

1 Introduction

Chronic diseases are currently a huge obstacle in several procedures. No matter how long the procedure lasts, if it is a chronic disease, it is extremely probable that the patient may require an organ transplant or, otherwise, he/she will stay ill. In chronic kidney diseases (CKD), such setbacks are astoundingly current [1].

Although there are presently many procedures and treatments available, the patients do not present great odds to have their disease cured. One of those procedures is the continuous ambulatory peritoneal dialysis (CAPD). This procedure is mostly used for patients in their end-stage of CKD and it can be performed in their home settings, which highly simplifies patients' life [1]. Since the CAPD procedure is not a cure, patients need to be observed from time to time, which also includes them to be monitored using blood tests. Those blood tests are used to see patients' evolution and monitor therapeutic outcomes and other prognosis [2].

One of the indexes to renal function is the creatinine levels since this waste product presents the potential to calculate the values of glomerular filtration rate and, therefore, it can be used to examine how functional is the kidney [3]. The creatinine levels in end-stage renal disease patients can be classified according to their values [4].

After the study of influential substances found in blood samples of patients undergoing the CAPD procedure, it is

✉ Marisa Esteves
marisa@di.uminho.pt

Claúdia Brito
cvmb.mat@gmail.com

Hugo Peixoto
hpeixoto@di.uminho.pt

António Abelha
abelha@di.uminho.pt

José Machado
jmac@di.uminho.pt

[1] Department of Informatics, University of Minho, Braga, Portugal

[2] Algoritmi Research Center, University of Minho, Braga, Portugal

essential to understand that those samples may introduce new patterns in this field of work. Considering this, it is becoming more and more crucial to introduce data mining (DM) technologies in this medical area as methods of knowledge extraction (KE). In short, DM is a discovering field since almost daily a new algorithm is created to find new patterns or to classify information, and even to create clusters.

Therefore, the main aim of this research project is to find the best way to predict or classify the patients' samples, through a classification method previously used in other studies, but also to understand the results expected and the results obtained. To accomplish such tasks, the Weka software was used in which DM algorithms are already implemented.

We can affirm that the results obtained in this paper present a solid foundation for further investigation, reaching a high performance in classifying the instances. The tested algorithms displayed good accuracy rates, reaching values of approximately 95%, and low relative absolute error values, which proved that the features chosen were the right choice. It also exemplifies that they are some important characteristics to succeed. In the future, the results could be implemented in a clinical decision support system (CDSS) to help health professionals in their decision-making process.

Regarding the structure of this document, in Sect. 2, the state of the art ("Background") related to the research area of this project is described. Thereafter, Sect. 3—"Implementation"—presents a description of the implementation of the study conducted, which includes the data processing and the algorithms. Then, Sect. 4 presents the results achieved, and they are subsequently discussed in the same section. In Sect. 5, the conclusion and future work conclude briefly this paper.

2 Background

2.1 Creatinine

Creatinine is a waste chemical molecule which is generated from muscle metabolism. This substance is of major importance when dealing with patients undergoing continuous ambulatory peritoneal dialysis (CAPD) treatment and, consequently, it was an essential pivotal point for this study [5].

Patients undertaking CAPD may present altered blood samples. It is known that the values of serum creatinine are highly affected by kidneys malfunction since those organs should filter out most of the creatinine and, subsequently, dispose it in the urine [5]. Therefore, when the values of creatinine increase from its normal value of < 1.0 to 10 mg/dL in the blood, it is undoubtedly a warning sign for the malfunction of kidneys since it reflects their waste-excretion function [6].

On the same hand, it is important to denote the role of creatinine in the calculations of the estimated glomerular filtration rate (eGFR). Even though GFR—the best indicator to kidney function—can be measured directly by other blood tests, serum creatinine and urea are still being used to calculate eGFR nowadays. There is an association of low serum creatinine corresponding to eGFR values near the 60 mL/min/1.73 m^2 threshold with the identification of kidney diseases or even end-stage chronic kidney diseases (CKD) [7]. This paper will focus on serum creatinine values since those values can be used to further reveal more information about the disease progression and, as also above-mentioned, once its role to calculate posteriorly the eGFR.

2.2 Data mining and machine learning

The process of predicting or finding specific patterns in a large amount of data is known as data mining (DM). In recent years, this process has been continuously in evolution and it is increasingly more integrated in new software.

DM and machine learning (ML) could be classified based on the output into three different types, i.e. classes, namely: [10, 11]

1. Supervised Learning: it uses the label in the training process and, consequently, in the testing process. If the label is categorical, it is called a classification problem, but if it is continuous numbers, it corresponds to a regression problem;
2. Unsupervised Learning: it trains without using a label, such as clustering algorithms;
3. Semi-supervised Learning: it uses some label values in the learning process.

The most important purpose of DM is to create models and, thereafter, improving the way users can read data and correctly interpret it.

The set of three classes previously mentioned (Supervised Learning, Unsupervised Learning, and Semi-supervised Learning) demonstrates that there are many possible ways to create a model. However, the major and most difficult step in a DM process is to create a good model, but also to identify exactly which techniques to use [10].

Software As stated previously, data mining (DM) is currently an emerging technology. Therefore, the software presently available are continuously improving and increasing in number. Considering this, there is in the market an extensive list of open-source software which can be used to perform DM. Nonetheless, the most important

features of any software are their outcomes, as well as the number of algorithms available in their library. There are several free open-source tools that can help to perform a better DM process rather than to use directly the R language. Weka, Orange or even RapidMiner are some of the technologies that have been consistently updated to provide better performance results to their users [12].

Thus, this project aims to use a software that encompasses a comprehensive collection of algorithms and processing tools focusing on machine learning (ML). Weka allows visual interaction, but also an easiness to users to try out new algorithms and access their results. Vis-à-vis already existing algorithms, Weka presents an ample library with algorithms for regression, classification, clustering, association rules, and attribute selection [13]. Therefore, our choice regarding which software to use was evident.

Weka The Weka software was developed by the University of Waikato in Hamilton, New Zealand. Initially, this platform was just a project with the intention to allow researchers an easier access to several techniques in machine learning (ML). Later, the project felt the need of developing a framework and, consequently, Weka went from being a simple toolbox to become the complete and complex software that it is today. Nowadays, this platform is recognized worldwide and it is widely used to perform DM and ML. Consequently, its documentation concerning DM is cited frequently by scientific researchers [13].

Furthermore, nowadays, machine learning is a very common topic in the informatics, electronics, and health fields. Its main goal is to find patterns and then performing tasks in generalized datasets. ML uses learning and knowledge systems as basis, being able to perform diverse tasks and analysis in structured and unstructured data [8].

On the other hand, ML implementation is been considered a low cost system when compared to manual programming [9].

Nonetheless, one of the principal problems associated with the use of ML are the few philosophical questions that it arises. One of the most pertinent questions rely on the big question of "What is Learning?". In this case, learning is considered the sum of representation, evaluation, and optimization. The set of those three factors is very important since it depicts data representation, the evaluation of performances, and the hypothesis of optimization of the results [8, 9].

Knowledge extraction Machine learning is important in the process of knowledge extraction (KE) since it uses data mining to find patterns and rules, but also to learn from such data. In this way, KE became a very important aspect when dealing with Big Data—large set of structured and unstructured data—by helping in the process of transforming "nonsense" data into meaningful and predictive data.

Therefore, DM should process data collected from continuous ambulatory peritoneal dialysis (CAPD) procedures to classify patients' samples/instances since they generate Big Data in terms of big volume, high velocity, and different variety. The classification and prediction of samples are designed using DM and ML algorithms. In this study, the algorithms were implemented using the well-known Weka tool.

2.3 Clinical decision support systems

Due to the increasing quantity of data generated in healthcare settings, health professionals in such organisations do not have frequently enough time to analyse and act quickly upon any discrepancy in patients' blood test results. Therefore, these institutions need to uncover a solution to summarize, simplify, and improve health professionals' work.

In order to solve this situation, clinical decision support systems (CDSS) are currently being developed and implemented worldwide in healthcare settings [14, 15]. However, those systems must become intelligent by also starting to use Artificial Intelligence (AI) and processes such as data mining (DM) to achieve better results, as well as to deal more effectively with the usual high demand for health services [16–18].

This paper presents also a unique opportunity to involve CKD indicators in clinical decision support systems as it may help the caregivers to disclose the patients' conditions and give the proper treatment timely.

2.4 Related work

In recent years, a lot of research projects regarding the prediction and classification of several indicators have been conducted in healthcare. Currently, there is a significant number of studies that merge data mining (DM) techniques and chronic kidney diseases (CKD) [1, 19, 20]. Nonetheless, even though there is undoubtedly a lack of information in this field merely focused on serum creatinine, diverse research projects already undertaken can be explored and brought to this issue.

Bala and Kumar [21] conducted a review paper focused on DM classification techniques in kidney diseases prediction, going from kidney stone to CKD, the authors studied several DM techniques used to predict kidney diseases in order to find the best techniques available. It was concluded that no single classifier can produce the best results for every dataset which proves that the DM

techniques must be adapted to the data information available [21].

One of the most interesting studies is described in the scientific article "Process Mining Routinely Collected Electronic Health Records to Define Real-life Clinical Pathways during Chemotherapy" [22]. The authors endorse that it can be possible to find patterns of care and, subsequently, to use the information gathered to monitor how patients are improving. The principal aim of the study was to focus in creating a method that could be used to process Electronic Health Records (EHR) and, thereafter, to use the outputs from the process in order to define and classify the patients' improvements or deteriorations during chemotherapy [22].

Data mining can also be used to prevent certain events in healthcare, such as demonstrated in the scientific manuscript "Prevent Patient Cardiac Arrhythmias by Using Data Mining Techniques" [16]. This study promotes a way to address real-time monitoring in clinical decision support systems (CDSS) to help the healthcare provider to be more efficient when he/she is taking care of patients. Their results are very remarkable since a value of 95% was reached for sensitivity by using a DM algorithm. In our opinion, this work was very ambitious and it could became the foundation of several other similar research projects [16].

A study conducted by Aqlan et al. [19] is focused in several DM techniques in order to predict CKD. The authors tested different predictive analytics techniques, going from Decision Trees to Artificial Neural Networks to create a decision support tool which might help in the diagnosis of CKD. Using the blood samples of patients, this paper exhibits a greater performance when using Random Trees above all the other five techniques used. Although the results only focused in predicting the presence of CKD or not CKD, the authors emphasize that posterior work will focus in the five stages of CKD and their consequent prediction.

3 Implementation

Through the realization of this study, it was mainly intended to find data mining (DM) algorithms and a correlation between several blood indicators in order to predict the serum creatinine values in blood tests. Thus, via knowledge extraction (KE), and more precisely DM, it was expected to correctly classify or predict those patients' blood indicators values.

Table 1 presents a small sample of the dataset used before its pre-processing, which includes the following eight blood indicators: gender, age, total calcium, chlorides, creatinine, ferritin, iron, and urea. The dataset was retrieved from the Health Information Systems (HIS) of a Portuguese health institution—*Centro Hospitalar do Porto* (CHP) and, therefore, the information used corresponds to current real data of patients. The dataset used had 2489 rows (instances), i.e. information regarding 2489 medical examinations of different patients.

3.1 Data processing

First, the data should be normalized. Then, the labels for classification should be determined by dividing creatinine values into groups for classification. The subjects' dataset used contains the results of blood tests of six different substances as presented above, which includes the values of serum creatinine, represented by "Creatinine (label)".

A study that classifies creatinine by its values was found—"Significance of Serum Creatinine Values in New End-stage Renal Disease Patients" [4]. Nonetheless, it is relevant to note that it uses factors in its classification such as gender, race, and age as factors that influence the increase or decrease of serum creatinine values [4]. On the other hand, this classification has a range of values from [4.6 $\pm$ 2.7 to 16.3 $\pm$ 0.2] mg/dL, and the dataset used in this study presents a range that goes from [1.82 to 126] mg/dL. Since these two ranges have a huge discrepancy, the dataset could not be treated straightforwardly in its data processing. Therefore, our process for the classification of creatinine values involved adding a new range called "Outliers" since those values were not initially classified by the study, i.e. mean creatinine values greater than 16.5 mg/dL.

The representation of the serum creatinine values by quintile are presented in Table 2.

It is important to refer that this pre-processing of data was performed once the initial values were not positively exploitable through only regression and normalization.

After the first tests, it was observed that one more pre-processing was needed on the dataset, namely oversampling, which was thereafter also executed. As so, the results presented below do use the oversampling method.

Oversampling The method of oversampling in data analysis implies adjusting the class distribution of a dataset (i.e. the ratio between the different classes) in order to equalize the number of the classes instances. This method is used when there is a need for classification [23].

3.2 Algorithms

Focusing in achieving a good classification of a dataset, it is crucial to find an algorithm (or more than one algorithm) that have a reliable performance for the dataset in question.

Table 1 A representation of a small sample of the dataset before its pre-processing

Gender	Age	Total Calcium	Chlorides	Urea	Ferritin	Iron	Creatinine (label)
F	40	2.17	87	132	11	15	126
M	37	1.91	102	761	299	116	88
F	33	2.28	104	274	321	134	147
…	…	…	…	…	…	…	…

Table 2 Representation of the serum creatinine values by quintile (Adapted from [4])

	Creatinine Quintiles					
	Lowest	Second	Third	Fourth	Highest	Outliers
Mean Creatinine (mg/dL)	4.6 ± 2.7	6.6 ± 1.4	8.3 ± 1.4	10.1 ± 2.1	16.3 ± 0.2	> 16.5

Nonetheless, it is impossible to generalize and affirm that one specific algorithm will have a better performance than the others since each dataset has its own algorithm that reaches better performance values.

However, a large set of algorithms to classify is available. In this particular case study, the *IBK*, *KStar*, *REPTree*, *RandomTree*, and *RandomForest* algorithms were chosen.

3.2.1 Lazy algorithms

The lazy algorithms are used as K-nearest Neighbours Algorithms. They store the training test until it is needed to classify the testing test or else when a query is made to the system [24].

This group of algorithms includes the next two algorithms described, namely:

- ***IBK* Algorithm** The *IBK* algorithm belongs to the set of Instance-based Classification Methods and it is, as a Lazy Algorithm, a K-nearest Neighbours Classifier. This classifier uses the same distance metric yet the number of nearest neighbours is a parameter that can be specified [24];
- ***KStar* Algorithm** The *KStar* algorithm follows the same methodology as *IBK* yet the distance metric used in *KStar* uses the concept of entropy. The classification of *KStar* is done by the sum of the probabilities "from the new instance to all of the members of a category" [25]. In order to fully achieve the results, this sum must be executed in all the instances.

3.2.2 Decision tree algorithms

The decision tree algorithms are used as predictive models that "use a set of binary rules to calculate a target value" [26]. This type of algorithms can be divided into two types, namely Classification Trees and Regression Trees. The first is applied to a dataset to create categorical datasets, the other one is used to build continuous datasets [26].

This group of algorithms contains the next three algorithms defined, i.e.:

- **REPTree Algorithm** Reduced Error Pruning (REP) Tree is a classifier and a Fast Decision Tree Learning Algorithm. It is built under the assumptions that adding entropy to information and minimizing the error that arises from variance culminates to the achievement of better results. It acts upon the dataset creating several decision trees and, in the end, the outcome is the best tree of them all. This algorithm, as the name indicates, prunes the tree with reduced-error, which means that the tree is trimmed and altered in order to get better performance results [27];
- ***RandomForest* Algorithm** The *RandomForest* algorithm acts on the dataset to create several and different decision trees and, thereafter, it sets the object to be classified in each of the decision trees. In the end, the algorithm calculates the results (total) by the evaluation of the results of each tree [28]. This algorithm is considered easier to use and it has a forgiving threshold greater than the other algorithms [26];
- ***RandomTree* Algorithm** The *RandomTree* algorithm is an algorithm that was created by combining the Single Tree Model and the *RandomForest* algorithm. It is a classifier that can generate many individual learners. The general idea consists in taking the input vector, classifying it with every tree in the forest, and, subsequently, the output includes the class label that contains the majority of the elements [29].

4 Results and discussion

After going through a challenging process in the search for the right and better data mining (DM) algorithm, it was possible to reduce the best results achieved to three

algorithms. Those results were obtained with the dataset already pre-processed.

However, as stated previously, it is important to note that after analysing the initial results, it was concluded that the results were not good enough. Since this study involved classification, it was necessary to perform an oversampling on the dataset in order to reach feasible results. More specifically, the tests were performed using a tenfold cross-validation method, which allowed to establish an average accuracy for the classifier.

Cross validation The cross-validation technique uses the dataset repeatedly to train the algorithm N times and tests 1/N of the training examples. It reproduces the use of training and test sets without the use of such methodology [30].

Table 3 displays the results achieved with the DM algorithms after the complete pre-processing of the dataset. A few algorithms tested were initially put aside since the results were not satisfactory enough to be considered. Therefore, those algorithms are not mentioned in this study.

As it was already referred previously in this paper, it does not exist a specific universal algorithm that always presents the better results for every dataset it is performed on. Depending on the dataset, the results may differ, and a different algorithm can reach better performance values.

After analysing the results obtained, it was possible to realise that the *REPTree* does not present a good outcome since the percentage of instances correctly identified does not reach the standard value of approximately 95% as the other algorithms accomplish, namely the *IBK* (94.89%), *KStar* (94.97%), *RandomForest* (95.86%), and *RandomTree* (94.89%) algorithms. This algorithm also exhibits a relative absolute error value of more than 50% (54.59%), which is highly disappointing.

Regarding another Decision Tree Algorithm, namely the *RandomForest* algorithm, as stated above, it displays a good accuracy value of 95.86%. However, its relative absolute error value is greater than the other algorithms—approximately 29.31% in comparison with 6.5% (*IBK*), 8.02% (*KStar*), and 6.37% (*RandomTree*). Although its result regarding the correctly classified instances is greater than any other algorithm, its relative absolute error value is undoubtedly an issue.

Concerning the *RandomTree* algorithm, another Decision Tree Algorithm, the results achieved were the better achieved with this study since its accuracy is of 94.89% and its relative absolute error value of 6.37%. Those values are better than the obtained from the lazy algorithms.

Both lazy algorithms, namely *IBK* and *KStar*, had similar outcomes, since they reached values of approximately 95% for correctly classified instances. Nevertheless, regarding the relative absolute error values, *IBK* achieved a value of 6.5% and *KStar* of 8.02%. Therefore, those values demonstrate that these results are also very satisfactory even though they were slightly better for the *RandomTree* algorithm.

For a better visualization of the outcomes achieved, two charts were built in order to compare the algorithms used to conduct this study. The algorithms were grouped by their type, i.e. lazy algorithms or decision tree algorithms.

Figure 1 represents both lazy algorithms, namely *IBK* and *KStar*.

Figure 2 represents the comparison between the three decision tree algorithms, specifically *REPTree*, *RandomForest*, and *Random Tree*.

5 Conclusion and future work

Finally, this study had the aim to analyse, understand, and prove how an initially large amount of insignificant clinical data to an Engineer can become extremely useful and approachable to be used in a clinical decision support system (CDSS) to assist health professionals in their decision-making process.

Therefore, several data mining (DM) algorithms were performed on a subjects' dataset that contained the results of blood tests of six substances, which included the values of serum creatinine. After the results were obtained, it was possible to analyse them and, thereafter, to conclude that

Table 3 Results from the classification of the dataset using the Weka software

Algorithms	Correctly classified instances	TP rate	FP rate	Precision	F-measure	ROC area	Relative absolute error
IBK	0.9489	0.949	0.013	0.949	0.949	0.944	0.065
KStar	0.9497	0.95	0.013	0.95	0.95	0.996	0.0802
REPTree	0.647	0.647	0.1	0.642	0.644	0.901	0.5459
RandomForest	0.9586	0.959	0.021	0.959	0.959	0.995	0.2931
RandomTree	0.9489	0.949	0.087	0.949	0.949	0.968	0.0637

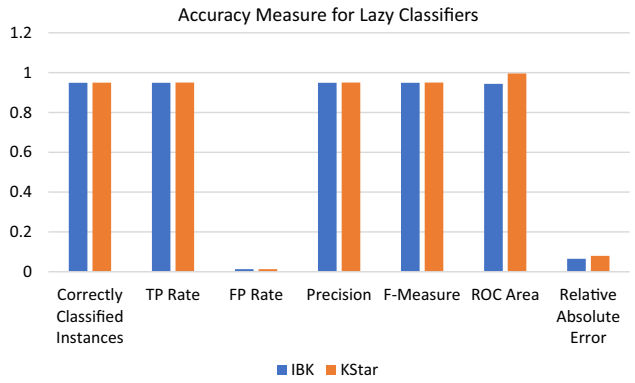


Fig. 1 Accuracy measure for Lazy Classifiers (*IBK* and *KStar*)

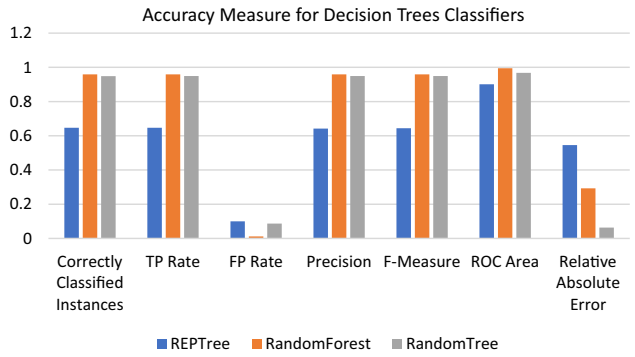


Fig. 2 Accuracy measure for Decision Tree Classifiers (*REPTree*, *RandomForest*, and *RandomTree*)

the algorithms chosen had a reliable performance except for the *REPTree* and *RandomForest* algorithms.

Reaching accuracy values that go from 94.89 to 94.97%, the other three algorithms used, i.e. *IBK*, *KStar*, and *Random Tree*, exhibited a prominent level of accuracy, and low values of relative absolute error, and it is therefore conceivable to conclude that those algorithms could be applied in a larger dataset with the same features in order to be used in the future in a CDSS where efficiency and scalability should also be tested. Applying machine learning (ML) to this type of information adds value to the system and promotes a better efficiency between health professionals.

Thus, it was possible to accomplish the proposed main goals and to understand how data can be handled to reach results in which it is possible to classify the values of serum creatinine.

As future work, the dataset could be handled differently. For instance, it could be used for regression rather than classification, which could enable the prediction of the true values of serum creatinine of patients and, subsequently, inferring about their improvements more directly. Additionally, it could also include the combination of the better algorithms obtained rather than using them separately in order to potentially obtain better results. Furthermore, this potential new study could provide some insights regarding their comparison. Finally, the results obtained could be potentially added and used in a real-life CDSS in healthcare settings in order to aid healthcare providers in their tasks.

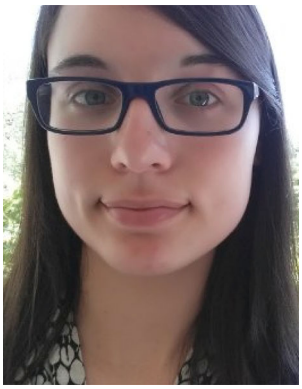
Acknowledgements This work has been supported by Compete POCI-01-0145—FEDER-007043 and FCT—*Fundação para a Ciência e Tecnologia* within the Project Scope UID/CEC/00319/2013.

References

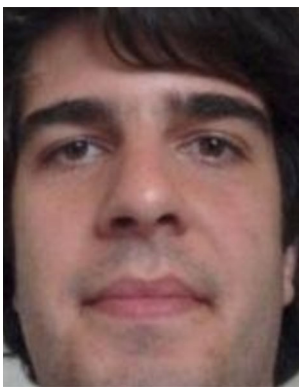
1. Rodrigues, M., Peixoto, H., Esteves, M., Machado, J., & Abelha, A. (2017). Understanding stroke in dialysis and chronic kidney disease. *Procedia Computer Science, 113,* 591–596.
2. Venkatapathy, R., Govindarajan, V., Oza, N., Parameswaran, S., Pennagaram Dhanasekaran, B., & Prashad, K. V. (2014). Salivary creatinine estimation as an alternative to serum creatinine in chronic kidney disease patients. *International Journal of Nephrology, 2014,* 1–6.
3. Guyton, A. C., & Hall, J. E. (2006). *Guyton and hall textbook of medical physiology*. Amsterdam: Elsevier.
4. Fink, J. C., Burdick, R. A., Kurth, S. J., Blahut, S. A., Armistead, N. C., Turner, M. S., et al. (1999). Significance of serum creatinine values in new end-stage renal disease patients. *The American Journal of Kidney Diseases, 34,* 694–701.
5. Davis, C. P., & Shield Jr., W. C. (2018). *Creatinine (low, high, blood test results explained).* https://www.medicinenet.com/creatinine_blood_test/article.htm#what_is_creatinine. Accessed 21 Jan 2019.
6. Mildred Lam, M. (2018). *Kidney failure—Understanding end stage renal disease (ESRD).* http://www.netwellness.org/healthtopics/kidney/kidney2.cfm. Accessed 21 Jan 2019.
7. Peake, M., & Whiting, M. (2006). Measurement of serum creatinine—Current status and future goals. *The Clinical Biochemist Reviews, 27,* 173–184.
8. Oliveira, P., Portela, F., Santos, M. F., Machado, J., Abelha, A., Silva, Á., & Rua, F. (2016). Optimization techniques to detect early ventilation extubation in intensive care units. In *Advances in Intelligent Systems and Computing (AISC)* (pp. 599–608). Cham: Springer.
9. Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55,* 78–87.
10. Abernethy, M. (2010). Data mining with WEKA, Part 2: Classification and clustering. https://www.ibm.com/developerworks/library/os-weka2/. Accessed 21 Jan 2019.
11. Veloso, R., Portela, F., Santos, M. F., Machado, J., da Silva Abelha, A., Rua, F., et al. (2017). Categorize readmitted patients in intensive medicine by means of clustering data mining. *International Journal of E-Health and Medical Communications, 8,* 22–37.
12. Naik, A., & Samant, L. (2016). Correlation review of classification algorithm using data mining tool: WEKA. *Procedia Computer Science, 85,* 662–668.
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter, 11,* 10–18.
14. Blobel, B. (2002). Analysis, design and implementation of secure and interoperable distributed health information systems. *Studies in Health Technology and Informatics, 89,* 1–352.

15. Portela, F., Santos, M. F., Machado, J., Abelha, A., Rua, F., & Silva, Á. (2015). Real-time decision support using data mining to predict blood pressure critical events in intensive medicine patients. In *Lecture Notes in Computer Science (LNCS)* (pp. 77–90). New York: Springer.
16. Portela, F., Filipe Santos, M., Silva, A., Rua, F., Abelha, A., & Machado, J. (2014). Preventing patient cardiac arrhythmias by using data mining techniques. In *2014 IEEE Conference on Biomedical Engineering and Sciences (IECBES). IEEE (2014)* (pp. 165–170).
17. Pereira, S., Portela, F., Santos, M., Machado, J., & Abelha, A. (2016). Predicting pre-triage waiting time in a maternity emergency room through data mining. In *Lecture Notes in Computer Science (LNCS)—Smart Health.* New York: Springer.
18. Oliveira, S., Portela, F., Santos, M. F., Machado, J., & Abelha, A. (2014). Predictive models for hospital bed management using data mining techniques. In *Advances in Intelligent Systems and Computing (AISC)* (pp. 407–416). New York: Springer.
19. Aqlan, F., Markle, R., & Shamsan, A. (2017). Data mining for chronic kidney disease prediction. In *Industrial and Systems Engineering Research Conference (ISERC).*
20. Sharma, S., Sharma, V., & Sharma, A. (2016). Performance based evaluation of various machine learning classification techniques for chronic kidney disease diagnosis. *International Journal of Modern Computer Scienc, 4,* 11–16.
21. Bala, S., & Kumar, K. (2014). A literature review on kidney disease prediction using data mining classification technique. *International Journal of Computer Science and Mobile Computing, 37,* 960–967.
22. Baker, K., Dunwoodie, E., Jones, R. G., Newsham, A., Johnson, O., Price, C. P., et al. (2017). Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy. *International Journal of Medical Informatics, 103,* 32–41.
23. Chawla, N. V. (2005). *Data mining and knowledge discovery handbook*. New York: Springer.
24. Vijayarani, S., & Muthulakshmi, M. (2013). Comparative analysis of bayes and lazy classification algorithms. *International Journal of Advanced Research in Computer and Communication Engineering, 2,* 3118–3124.
25. Tejera Hernández, D. C. (2015). An experimental study of K* algorithm. *International Journal of Information Engineering and Electronic Business, 7,* 14–19.
26. Horning, N. (2010). Random forests: An algorithm for image classification and generation of continuous fields data sets. In *The International Conference on GeoInformatics for Spatial-Infrastructure Development in Earth & Allied Sciences 2010* (pp. 1–6).
27. Devasena, L. (2014). Comparative analysis of random forest, REP tree and J48 classifiers for credit risk prediction. In *IJCA Proceedings on International Conference on Communication, Computing and Information Technology* (pp. 30–36).
28. Breiman, L., & Cutler, A. (2018). Random forests—Classification description. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm. Accessed 21 Jan 2019.
29. Kalmegh, S. (2015). Analysis of WEKA data mining algorithm REPTree, SimpleCart and RandomTree for classification of indian news. *International Journal of Innovative Science Engineering and Technology, 2,* 438–446.
30. Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research, 5,* 1089–1105.

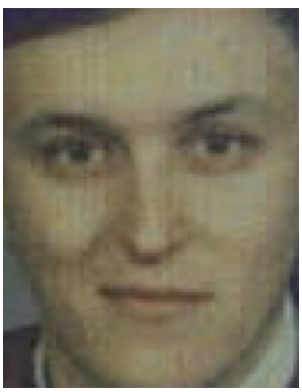
Claúdia Brito is currently a PhD Student and a Researcher at HASLab, as well as an Associate Laborator at INESC Technology and Science, at the University of Minho, in Portugal. Her areas of expertise are in Biomedical Engineering and Medical Informatics.



Marisa Esteves is currently a PhD Student and a Researcher at Algoritmi Research Center, as well as a Guest Lecturer at the University of Minho, in Portugal. She is a member of the Knowledge Engineering Group (KEG) at Algoritmi Research Center. Her areas of expertise are in Biomedical Engineering and Medical Informatics.



Hugo Peixoto is currently an Information Technology professional at Centro Hospitalar entre Douro e Vouga, as well as a Guest Lecturer at the University of Minho, in Portugal. He has a PhD title in Biomedical Engineering. His areas of expertise are in Biomedical Engineering and Medical Informatics.



António Abelha is currently an Assistant Professor at the University of Minho, as well as an Information Technology professional in several Portuguese hospitals, in Portugal. He is a member of the research line Computer Science and Technology (CST) and the Knowledge Engineering Group (KEG) at the Algoritmi Research Center. His areas of expertise are in Informatics.

José Machado is currently an Associate Professor with Habilitation at the University of Minho, as well as an Information Technology professional in several Portuguese hospitals, in Portugal. He is the director of Algoritmi Research Center since 2018, as well as a member of the research line Computer Science and Technology (CST) and the Knowledge Engineering Group (KEG). His areas of expertise are in Informatics.