A generalized linear model for cardiovascular complications prediction in PD patients

Carlos Fernandez-Lozano
Computer Science Department, University of A Coruna. Faculty of Computer Science, A Coruña, Spain
carlos.fernandez@udc.es

Rafael Alonso Valente
Nephrology Service. Complexo Hospitalario Universitario de Santiago, Santiago de Compostela, Spain
Rafael.Alonso.Valente@sergas.es

Manuel Fidalgo Díaz
Nephrology Service. Complexo Hospitalario Universitario de Santiago, Santiago de Compostela, Spain
Manuel.Fidalgo.Diaz@sergas.es

Alejandro Pazos
Computer Science Department, University of A Coruna. Faculty of Computer Science, A Coruña, Spain
apazos@udc.es

ABSTRACT
This study was conducted using machine learning models to identify patient non-invasive information for cardiovascular complications prediction in peritoneal dialysis patients. Nowadays is well known that cardiovascular diseases are the key to mortality in patients undergoing peritoneal dialysis as the risk of cardiovascular disease increases with the progression of renal failure. Primary aim is to establish variables most associated with cardiovascular complications. To achieve this goal four different machine learning techniques were used. We found that the best classification algorithm was a Generalized Linear Model, which achieved AUC values above 96% using a small subset of the original variables following a feature selection approach. Our approach allows us to increase the interpretability of the combinations of traditional factors, advanced chronic kidney disease factors and peritoneal dialysis factors all related with cardiovascular risk profile. The final model is based primarily in the traditional factors.

CCS CONCEPTS
Computing methodologies; Artificial intelligence; Machine learning; Modeling and simulation

KEYWORDS
Peritoneal Dialysis, Machine Learning, Cardiovascular Risk Prediction, Feature Selection, glmnet

1 INTRODUCTION

The prevalence of Chronic Kidney Disease is increasing in the world nowadays. Moreover, it is known that there is a clear association between this disease and heart disease, resulting in an increased risk of the latter as renal failure progresses.

In fact, cardiovascular disease is one of the key causes of morbidity and mortality in those patients with any kind of renal therapy, along with other infectious complications not directly associated with the renal process. Nowadays, peritoneal dialysis is a well- known and a standard modality of renal replacement therapy in patients with end-stage renal disease [8, 9].

The main objective of this study was to identify the key non- invasive clinical variables most frequently associated with cardio- vascular disease in peritoneal dialysis patients. Other objectives of this proposal are:

- prediction of cardiovascular risk in patients on peritoneal dialysis using Artificial Intelligence techniques
- establishing non-invasive clinical variables most associated with cardiovascular disease in patients on peritoneal dialysis
- evaluating cardiovascular risk and prevalence in patients on peritoneal dialysis

2 MATERIALS AND METHODS

For the present study, we used a cohort of more than 114 patients from the Peritoneal Dialysis Unit at the Complexo Hospitalario Universitario de Santiago de Compostela (CHUS) to predict cardio- vascular disease as the main cause of morbidity and mortality in patients with Chronic Kidney Disease.

For each patient we had traditional factors, advanced chronic kidney disease factors and peritoneal dialysis factors all related with cardiovascular risk profile. One of the aims of this research is to understand which of the three are better factors for cardiovascular disease prediction.

Different computational approaches were used that resort to artificial intelligence to predict cardiovascular disease as the main cause of morbidity and mortality in patients with Chronic Kidney Disease. Specifically, the following machine learning techniques were used: Random Forest (RF) [2], Support Vector Machines (SVM) [3, 13], Regularization paths for Generalized Linear Models via Coordinate descent (GLMNET) [6] and K-Nearest Neighbor (KNN) [7].

One of the main problems of those who suffer medical data and can confuse machine learning algorithms is noise and redundant information that can be found between different variables even more in the era of big data. Feature Selection (FS) have become a necessity in bioinformatics [5], were most of the algorithms were not designed to deal with large amounts of variables or actually, correlated variables. The main aims of a FS process is to avoid possible overfitting and improve the performance of the algorithm and to reduce the computational efforts of the experiments.

One of the key points of FS is that the variables remain unaltered and also the process selects the best subset of them. There are, in the context of binary classification problems, three different categories: filter, wrapper and embedded depending on how the researcher combine the search with the algorithms. In this work, we used a filter approach (faster, simpler and not dependent of the particular characteristics of the algorithms) for the search of the best subset of features. [10].

GLMNET via penalized maximum likelihood [6] is a regularized statistical model whose response variable is a Bernoulli indicator used for classification. It is based on the least absolute shrinkage and selection operator (LASSO) [12]. This algorithm performs an internal feature selection and regularization processes (shrinkage) automatically as the LASSO. Thus, this particular algorithm is able to find groups of variables highly correlated and important for the problem. The implementation goes from lasso to ridge penalties combining l1 and l2 penalties for regularization according with an alpha hyperparameter. In our experiments we used values 0, 0.15, 0.25, 0.35, 0.5, 0.65, 0.75, 0.85, and 1. The other hyperparameter of this algorithm is the one related with the degree of regularization, we considered values of 0.0001, 0.001, 0.01, 0.1, and 1.

3 RESULTS

We used four different machine learning algorithms RF, SVM, GLM- NET and KNN. It is necessary to perform a hyperparameter tuning of the algorithms to achieve the better possible performance of the algorithms. For fair comparisons we used a nested resampling approach [1, 11] involving two different levels. Externally we ran 10 independent repetitions of a 10-fold cross-validation with stratification of the patients in the validation set (same proportion of positive and negative cases). Thus, we partitioned the dataset in 10 equal sized subsets. We retained one for validation and used the nine remaining subsets for training.

We repeated this process 10 times, using each one of the partitions exactly once as the validation set. In each iteration, the corresponding training set is used in an inner loop, to find the optimal values of the hyperparameters of the algorithms. A holdout strategy was implemented to this aim, where the given training set from the external loop is partitioned again in 2/3rd of the set for training and 1/3rd of the set for validation. Hyperparameter search was ran with a grid search approach. The main goal of this resampling strategy in two levels is to better estimate the fit of the models. We decided to use the Area Under the ROC Curve (AUC) as the measure of fit because it is independent of the threshold of binarization and considers both Type I and Type II errors [4].

Initially we considered the original set of features and subsequently, we performed a filter feature selection approach (T-test) and we considered different subsets from the original, sizes 4, 16 and 32. As shown in Figure 1, the best subset of non-invasive clinical variables that we found is 16 and therefore we were able to reduce noisy and useless features.

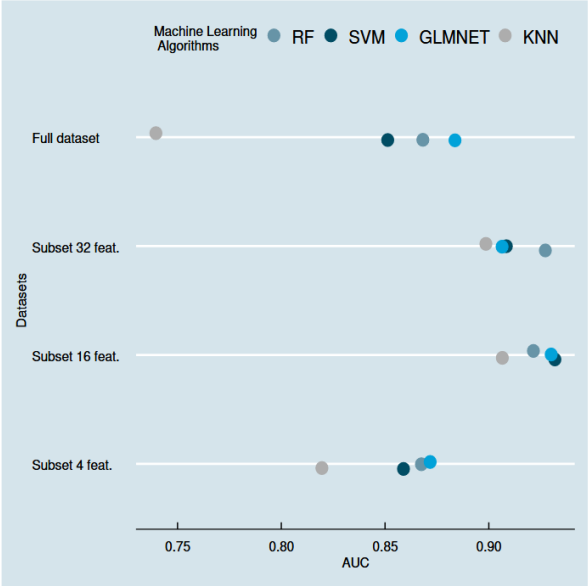

Figure 1: Aggregated results of the nested resampling strategy. Each of the classifiers was evaluated using AUC.

There is a clear trend with all the classifiers that seems to indicate that this number of features is probably the best as, at the end of the experimental design [2,5], a predictive value higher than 0.93 of AUC was achieved through more than one technique. The stability of the algorithms through the 10 independent runs is shown in Figure 2.

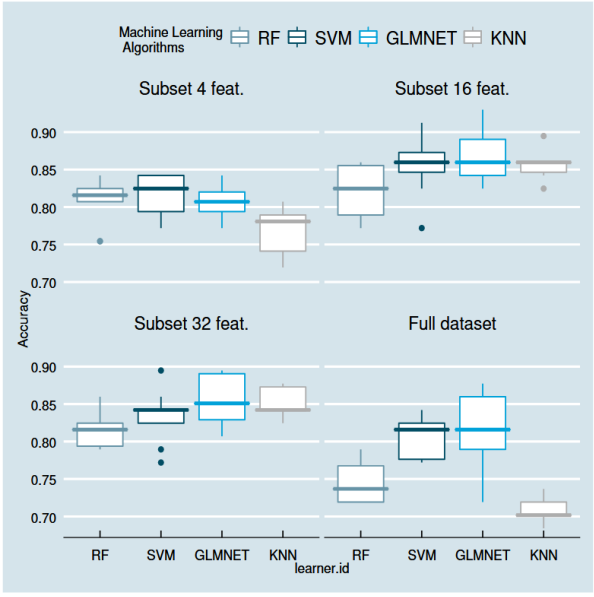

Figure 2: Violin plot results of the nested resampling strategy. Each of the classifiers was evaluated using AUC. Observed performance is very stable in general terms.

Finally, the importance of each one of the non-invasive variables aggregated through all the GLMNET models is shown in Figure 3.

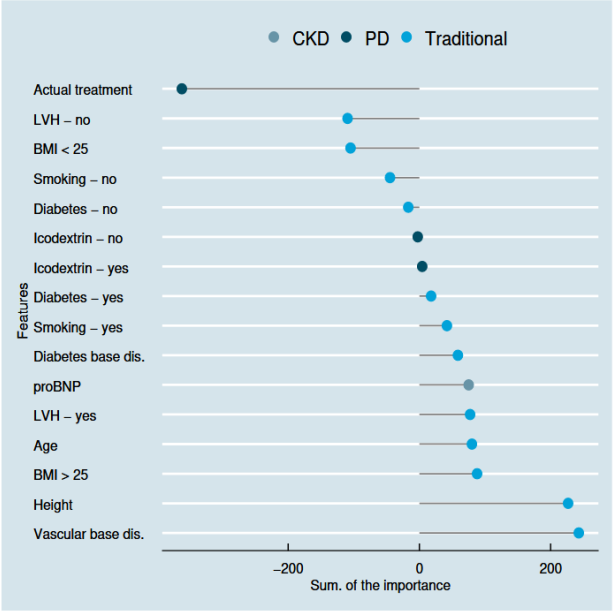

Figure 3: Importance of the non-invasive clinical variables (GLMNET). Results are separated according with the belonging of the variable to one of the three groups: traditional factors, advanced chronic kidney disease factors and peritoneal dialysis factors all related with cardiovascular risk profile.

4 CONCLUSIONS

This study has demonstrated the ability of machine learning for predicting cardiovascular disease in peritoneal dialysis patients using non-invasive, clinical variables. The results obtained in this work support those obtained in the literature [7]. After analyzing the importance of the non-invasive clinical variables, the model considers that the following are critical: diabetes mellitus and vascular base diseases, age, presence of diabetes mellitus, height, mass index body over 25, smoking, age, icodextrin, proBNP (brain natriuretic peptide), actual treatment and occurrence of left ventricular hypertrophy.

Because machine learning models did not work at any time under biological assumptions, a more comprehensive approach is provided than in the case of biological approaches. Although some of those factors cannot be treated by physicians, doctors must pay attention to disorders in those values to reduce as much as possible the risk of cardiovascular disease in this population subset. Furthermore, a reduction in both morbidity and mortality rates is possible and will hopefully improve the long-term outcomes.

ACKNOWLEDGMENTS

This work is supported by “Collaborative Project in Genomic Data Integration (CICLOGEN)” PI17/01826 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013âĂŞ2016 and the European Regional Development Funds (FEDER) - “A way to build Europe”. This project was also supported by the General Directorate of Culture, Education and University Management of Xunta de Galicia (Ref. ED431G/01, ED431D 2017/16) and the “Galician Network for Colorectal Can- cer Research” (Ref. ED431D 2017/23), and finally by the Spanish Ministry of Economy and Competitiveness for its support with the funding of the unique installation BIOCAI (UNLC08-1E-002, UNLC13-13-3503) and the European Regional Development Funds (FEDER) by the European Union and the “Juan de la Cierva” fellow- ship program supported by the Spanish Ministry of Economy and Competitiveness (Carlos Fernandez-Lozano, Ref. FJCI- 2015-26071).

REFERENCES

[1] Bernd Bischl, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Zachary Jones, and Giuseppe Casalicchio. 2016. mlr: Machine Learning in R . https://CRAN. R-project.org/package=mlr R package version 2.9.

[2] Leo Breiman. 2001. Random Forests. In Machine Learning . 5–32.

[3] Nello Cristianini and John Shawe-Taylor. 2000. An Introduction to Support Vector Machines and other kernel-base learning methods . Cambridge University Press.

[4] Tom Fawcett. 2006. An Introduction to ROC Analysis. Pattern Recogn. Lett. 27, 8 (2006), 861–874.

[5] Carlos Fernandez-Lozano, José Antonio Seoane Fernández, Marcos Gestal, Tom R. Gaunt, Julian Dorado, and Colin Campbell. 2015. Texture classification using feature selection and kernel-based techniques. Soft Comput. 19, 9 (2015), 2469– 2480.

[6] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33, 1 (2010), 1–22. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/

[7] K. Hechenbichler and K. Schliep. 2004. Weighted k-Nearest-Neighbor Techniques and Ordinal Classification. (2004). http://nbn-resolving.de/urn/resolver.pl?urn= nbn:de:bvb:19-epub-1769-9

[8] Arsh K. Jain, Peter Blake, Peter Cordy, and Amit X. Garg. 2012. Global Trends in Rates of Peritoneal Dialysis. Journal of the American Society of Nephrology 23, 3 (2012), 533–544. https://doi.org/10.1681/ASN.2011060607 arXiv:https://jasn.asnjournals.org/content/23/3/533.full.pdf

[9] Zhi-Hong Liu. 2013. Nephrology in China. Nature Reviews Nephrology 9 (2013), 523–528.

[10] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. 2007. A Review of Feature Selection Techniques in Bioinformatics. Bioinformatics 23, 19 (Sept. 2007), 2507–2517. https://doi.org/10.1093/bioinformatics/btm344

[11] Richard Simon. 2007. Resampling Strategies for Model Assessment and Selection . Springer US, Boston, MA, 173–186.

[12] Robert Tibshirani. 1994. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society, Series B 58 (1994), 267–288.

[13] Vladimir N Vapnik. 1998. Statistical learning theory . John Wiley & Sons, New York. A Wiley-Interscience Publication