2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Improving the compromise between accuracy, interpretability and personalization of rule-based machine learning in medical problems

Francisco Valente[1], Jorge Henriques[1], Simão Paredes[2], Teresa Rocha[2], Paulo de Carvalho[1], João Morais[3]

Abstract— **One of the key challenges when developing a predictive model is the capability to describe the domain knowledge and the cause-effect relationships in a simple way. Decision rules are a useful and important methodology in this context, justifying their application in several areas, particularly in clinical practice. Several machine-learning classifiers have exploited the advantageous properties of decision rules to build intelligent prediction models, namely decision trees and ensembles of trees (ETs). However, such methodologies usually suffer from a trade-off between interpretability and predictive performance. Some procedures consider a simplification of ETs, using heuristic approaches to select an optimal reduced set of decision rules. In this paper, we introduce a novel step to those methodologies. We create a new component to predict if a given rule will be correct or not for a particular patient, which introduces personalization into the procedure. Furthermore, the validation results using three public clinical datasets suggest that it also allows to increase the predictive performance of the selected set of rules, improving the mentioned trade-off.**

I. INTRODUCTION

Physicians usually incorporate clinical decision rules (CDRs) in several domains of their practice, such as bedside diagnostic or therapeutic choices. CDRs appeared as a tool to help the decision-making of the clinical staff, reducing its uncertainty and making it more evidence-based [1]. In the last few decades, several machine learning (ML) methods have been proposed as decision support systems in almost every medical specialties, where they have been shown to achieve high performances [2]. However, the majority of those models are often seen as a "black-box" with a lack of explainability capabilities, which limits their use in medicine. In fact, despite a large investment in the development of novel ML applications in medical areas, its translation to the daily clinical practice is still very limited [3].

Decision trees (DTs) are widely acknowledged as a very interpretable ML approach, and therefore a useful solution when the scenario requires a deep understanding of the generated model [4], such as the medical cases. DTs present several features that make them appealing from the interpretability point of view, such as: 1) they mimic the human reasoning, incorporating a combination of decision rules; 2) the IF-THEN nature of such rules allow to easily extract domain knowledge from the cause-effect relationships; 3) the tree-like visual representation makes the overall classification process easy to understand. Despite these characteristics, DTs typically present a worse predictive performance when compared to more complex ML methodologies.

In order to overcome that disadvantage, ensembles of trees (ET) were proposed. ET are based on the idea that the combination of several weaker classifiers (several DTs) achieves better performance than a single classifier (one DT). Random Forest (RF) [5], a tree ensemble methodology, is one of the most widely applied classifiers as it often assures a high performance [6]. In RF, multiple DTs are built, each one using a randomly selected subset of the whole set of features, and (optionally) also a subset of the whole set of samples. The outputs of the individual trees are then combined to produce the final output. A DT is easy to interpret individually, as long as its dimensions (number of rules and length of each rule) keeps low. However, to study several (hundreds or thousands) DTs simultaneously is unfeasible for the final user (e.g. physicians). Thus, RFs are typically considered as "black-box" models.

Therefore, there is a trade-off between interpretability and accuracy in such rule-based ML methodologies. In order to overcome it, some approaches have been proposed, mainly suggesting techniques to improve the interpretability of the generated ET [4][7][8]. A group of procedures aims at simplifying the ET by decomposing the respective several individual DTs into a set of decision rules. Then, they select a sparser set of the best rules, based on heuristics such as LASSO [9], hill climbing [10] or quadratic programs [11]. These approaches allow to obtain a set of important decision rules directly extracted from the ET, which is simpler to analyze than a group of DTs, improving the interpretability. Even so, those methods often select a final set of several dozens of rules, which is still a high amount of rules for an easy and fast interpretation of the output and the respective extracted knowledge in the clinical practice. Furthermore, the final set of rules is usually applied uniformly to all patients, i.e. the decision rules have the same weight to all samples.

In this study, we propose an approach that predicts the correctness of each one of the decision rules to each patient, enabling to use a smaller set of decision rules to obtain the same performance of state-of-the-art methods. The goal is to promote a novel methodology towards a more interpretable and personalized set of decision rules, while keeping a good predictive ability, improving its usability in the clinical field.

II. METHODS

A. Generation and extraction of decision rules

A decision tree can be converted to a set of decision rules. An individual decision rule corresponds to a path from the

[1]F. Valente, J. Henriques and P. de Carvalho are with the Centre for Informatics and Systems of the University of Coimbra pfcv@dei.uc.pt

[2]S. Paredes and T. Rocha are with the Polytechnic of Coimbra, Coimbra Institute of Engineering

[3]J. Morais is with the Cardiology Department of Leiria Hospital Centre

roof of the tree to its leaf. The condition of the rule, i.e. the IF segment, is defined by the set of conditions in that path. The prediction, i.e. the THEN segment, is defined by the class attributed to the leaf. Thus, there are as many rules as leaves in the tree. Furthermore, assuming a binary classification, we may consider that if the condition part of a given rule is not verified, then at that decision level the rule leads to a prediction of the opposite class (represented by an ELSE segment). Fig. 1 exemplifies a very simple DT. From that DT, the following decision rules can be extracted:

1) IF x1 $>$ 0.5, THEN y1 (ELSE y2).
2) IF x1 $\leq$ 0.5 AND x2 = false, THEN y2 (ELSE y1).
3) IF x1 $\leq$ 0.5 AND x2 = true, THEN y1 (ELSE y2).

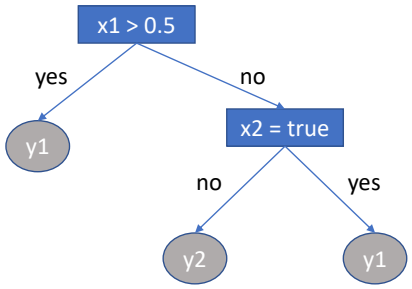

Fig. 1. Example of a decision tree. The split nodes are represented by the blue square boxes, and the leaf (or terminal) nodes are represented by the grey round boxes. $x1$ and $x2$ are two attributes (features), and $y1$ and $y2$ are two arbitrary classes (labels).

We take advantage of the random forest ability to generate a large group of DTs, and then get a set of decision rules as aforementioned. For interpretability issues, it is not enough to have decision rules, they must also be easy to understand, which implies those rules should be short. Therefore, in this study, we set to 3 the maximum depth of each DT, i.e. the condition part of each rule is composed of at most 3 elements (two AND connections).

B. Selection of a set of the best decision rules

In this study, the decision rules extracted from the individual trees of the RF are combined into a single set of rules. Then, we start by removing duplicated rules. Further, a subset of M decision rules is obtained using a logistic regression model with L1-regularization (LASSO), similar as performed in [9], as LASSO is a simple and intelligible heuristic selection method.

In this LASSO step, the input is a NxP matrix, where N is the number of samples and P is the number of decision rules extracted from the RF, and the output is the Nx1 outcome vector (observed outputs). The NxP matrix is binary, with a value of 1 if the rule condition (IF-part) of the rule p is verified to the patient n, and a value of 0 otherwise. Let us consider the rules of the DT of Fig. 1. If a given sample has the values $x1 = 0.3$ and $x2 = true$, only the IF-part of rule 3 is verified. Thus, the input vector for such patient and rules will be $condition\_verified([r1\ r2\ r3])$ = [0 0 1]. This procedure is applied to the N patients, taking into account the P rules, creating then the NxP matrix.

Therefore, the LASSO is applied as a rules selection procedure, shrinking towards 0 the coefficients of the rules that are least relevant to the output prediction as well as the ones of correlated rules. In our approach, we then obtain a subset of M (relevant and uncorrelated) rules by selecting the ones that present the M highest LASSO regression coefficient values. The M value can be defined by the user. For a better generalization ability, a 3-fold cross-validation LASSO is applied for that task.

C. Prediction of the correctness of each rule

In this study, we aim to classify each rule for each sample, i.e. to predict when a rule will give a correct output or not. In order to accomplish that goal, we create a classification model for each rule. More specifically, we train a model giving as features the variables used by all the selected rules, and as label a binary vector with a value of 1 if the rule gave a correct output and a value of 0 if the rule gave a wrong prediction. For example, considering the DT of Fig. 1, if a given sample has the values $x1 = 0.3$ and $x2 = true$, the predictions of the rules will be:

- rule 1: $y2$; rule 2: $y1$; rule 3: $y1$

So, if we assume that the true output is $y1$, rule 1 is incorrect (label=0) and rules 2 and 3 are correct (label=1). Therefore, if e.g. 5 rules are selected (M=5), 5 different binary label vectors will be created (with the information about this correctness for all the samples for each rule). Consequently, 5 classifiers are trained, each one to predict the individual rule correctness. The features used to train such classifiers are the original risk factors used by that selected rules. Thus, if those 5 rules use a total of K features, each of the 5 classifiers will have as input the NxK predictors matrix. Any classification algorithm can be used to predict the rule's correctness. In this study, a LASSO model was considered.

So, we not only extract and selected a subset of rules from an ET, as in state-of-the-art methods, but we also attempt to forecast if a given rule should be (or not) applied to a particular patient. This procedure introduces personalization to the methodology as the rules will be applied differently and more properly for each patient, which can contribute to the improvement of the individual predictions. Thus, this step is a novelty in relation to literature approaches, which apply the decision rules uniformly to all patients.

D. Computation of the probability of each class

For each new (validation) sample, two vectors of dimension 1xM are then generated, composed of binary values. The predicted rule's output informs if a given rule classifies the patient as having a disease (1) or not (0). The predicted rule's correctness informs if a given rule is expected to be correct (1) or not (0) for that patient. Table I presents an example of the vectors obtained for a given new sample, using 3 rules.

In order to generate the final output (probability that the sample is a positive class - e.g., the probability the patient has a given disease), the standard methodologies only do an averaging of the set of the M selected decision rules, usually:

TABLE I

PREDICTIONS OBTAINED FOR A NEW SAMPLE, USING A SET OF 3 RULES.

	Rule 1	**Rule 2**	**Rule 3**
Rule's output prediction	1	0	1
Rule's correctness prediction	1	1	0

$$Probability(class = 1) = \frac{1}{M}\sum_{i=1}^{M} rule\_output_i. \quad (1)$$

For the sample exemplified in Table I, those approaches would give a positive class probability of ∼0.66, because 2 rules out of 3 predict a positive class.

In contrast, in our approach, we also take into account the information about if each rule is expected to be (or not) correct for a given patient, i.e. the information of the last row of Table I. More specifically, the probability that a sample belongs to the positive class is given by a weighted average, which is personalized for each patient:

$$Probability(class = 1) = \frac{\sum_{i=1}^{M} rule\_output_i \cdot weight_i}{\sum_{i=1}^{M} weight_i}, \quad (2)$$

We could do only a simple averaging of the rules predicted to be correct, giving a weight of 1 if the rule is predicted to be correct, and a weight of 0 otherwise. However, it may happen that for some samples, none of the M selected rules is predicted to be correct, and thus no rule would be available to predict the final outcome. Therefore, we assume a weight of 2 if the rule is predicted to be correct, and a weight of 1 otherwise. This implies that the methodology can be generalized to all scenarios. Furthermore, it means that all the rules are considered but the ones predicted to be correct for that patient contribute twice more for the final output than the others. Therefore, for the example of Table I, rules 1 and 2 have a weight of 2, and rule 3 has a weight of 1. Thus, the probability of the positive class will be 0.6. Surely, such weights can be adjusted or optimized by the user.

E. Validation of the proposed approach

The proposed approach was validated in three public clinical datasets. Two of them are from UCI Machine Learning Repository (*https://archive.ics.uci.edu*): Heart Disease (prediction of presence/absence of heart disease) and Breast Cancer Wisconsin Diagnostic (prediction of benign/malign diagnosis of breast cancer). The third one is from Kaggle (*https://www.kaggle.com/*): Pima Indians Diabetes Database (prediction of presence/absence of diabetes disease). The datasets will be designated as Heart, Breast and Diabetes, respectively. Heart has some categorical variables with missing data, which was replaced by the most frequent value of the corresponding feature. A 10-times repeated 5-fold stratified cross-validation was used as the validation procedure.

III. RESULTS

The ability of the proposed approach to correctly classify the data into positive (presence of disease) or negative (absence of disease) samples was assessed through the area under the ROC curve (AUC). Those results are presented in Table II. The results of the proposed methodology are presented for different sets of selected rules, i.e. for different M values. More specifically, sets with the 3, 5, 10, 15 and 20 best decision rules were considered, as selected by the LASSO approach. The initial set of P rules was obtained by building a RF with 100 DTs. The results are also compared with two standard rules-based machine learning models: random forest and decision tree. For the RF methodologies, two versions are considered: one where the RF and its trees can grow without any constrictions; and a simpler and more interpretable version, where the RF can have at most 5 trees, each one with a maximum depth of 3. The parameters of those RF and DT models were optimized for each prediction task, using a cross-validated technique where the best parameters were chosen and applied in the final model.

TABLE II

AREA UNDER THE ROC CURVE (AUC) VALUES FOR THE PROPOSED APPROACH AND COMPARISON MODELS. THE RESULTS PRESENTED ARE RELATED TO THE MEAN AND ITS 95% CONFIDENCE INTERVAL.

	Heart	**Breast**	**Diabetes**
Random forest (no constraints)	0.90±0.01	0.99±0.00	0.83±0.01
Random forest (simpler)	0.87±0.01	0.98±0.00	0.79±0.01
Decision tree	0.79±0.02	0.94±0.01	0.73±0.01
Proposed approach (3 rules)	0.82±0.02	0.97±0.00	0.70±0.02
Proposed approach (5 rules)	0.85±0.01	0.98±0.00	0.74±0.02
Proposed approach (10 rules)	0.89±0.01	0.99±0.00	0.79±0.01
Proposed approach (15 rules)	0.90±0.01	0.99±0.00	0.80±0.01
Proposed approach (20 rules)	0.90±0.01	0.99±0.00	0.80±0.01

Table III provides additional information, presenting the cross-validation mean number of rules used by the RF and DT methods, in order to compare it with the amount used by our approach (3 to 20 rules).

TABLE III

MEAN NUMBER OF RULES USED BY THE RANDOM FOREST AND DECISION TREE MODELS.

	Heart	**Breast**	**Diabetes**
Random forest (no constraints)	1230	1105	969
Random forest (simpler)	27	29	26
Decision tree	24	18	21

Furthermore, in Fig. 2, the performance of the models following our approach is presented, i.e. final prediction based on the mean of the rules output weighted by its predicted correctness (2), versus the performance considering only a simple mean of the rules, as in the standard methodologies (1). The weighted mean values are represented by solid lines and the baseline mean values by dashed lines.

Finally, we present an example of a selected subset of decision rules. More specifically, a set for M=3 used in the outcome prediction for the Diabetes dataset:

1) IF $nr\_pregnancies \leq 6.5$ AND $[glucose] \leq 124.5$ AND $age \leq 34.5$, THEN no-diabetes (ELSE diabetes).

2) IF $age \leq 26.5$ AND $BMI \leq 37.2$ AND $tricep\_skin\_tickness \leq 28.5$, THEN no-diabetes (ELSE diabetes).

3) IF $blood\_pressure \leq 69.0$ AND $[glucose] > 119.5$ AND $diabetes\_pedigree \leq 0.23$, THEN diabetes (ELSE no-diabetes).

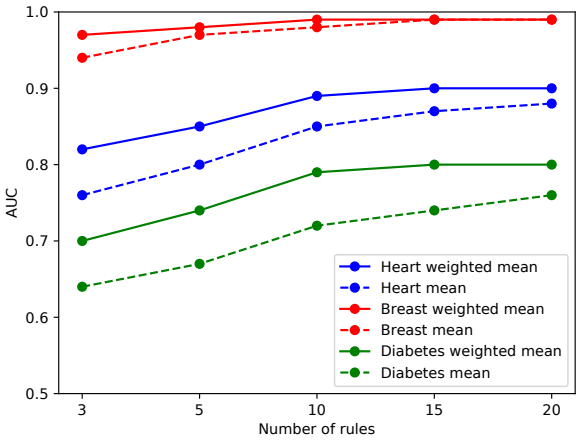

Fig. 2. Area under the ROC curve (AUC) values for the obtained set of rules, considering a simple mean (standard approaches) and a weighted mean (proposed approach) of their predictions, for the validation datasets.

IV. DISCUSSION

As we may observe in Table III, the non-constrained RF models (first row) consider a large number of decision rules to generate their outputs. In fact, it would be unfeasible for a physician to analyze thousands of rules in order to interpret the models generated by the RF. Consequently, that is a critical limitation for its application in clinical practice. Further, as presented in Table II, a significant drop in the predictive performance when we use a simpler RF (at most 5 trees with a maximum depth of 3) or a DT instead.

In Tables II and III, it is possible to observe that the proposed approach achieves better AUC than decision trees using at most 5 rules (while DTs use on average 18 to 24 rules), and equivalent or better AUC than the simpler random forest using at most 10 rules (while simpler RF uses on average 26 to 29 rules). Furthermore, the results show that using only 15 rules, the proposed procedure approaches the AUC performance of the more complex RF models. Therefore, the methodology presented in this paper seems to offer a better compromise between interpretability and prediction performance than the comparison models, which may facilitate its translation to clinical practice.

Fig. 2 shows how the proposed approach compares to the standard methodologies that select rules from ensembles of trees. It is possible to analyze that using the predicted rule's correctness to weigh the final outcome prediction for each patient (giving more importance to the rules predicted to be correct at the individual level) can significantly improve the predictive ability. As expected, this improvement is most noticeable for the smaller sets of rules, as a large set attenuates the effect of a single rule.

Lastly, the prediction of each rule's correctness offers a personalized element in the proposed approach, informing if a given rule is expected to be correct or not for a particular patient. This information can be used by the physicians to further assess the condition of each patient, i.e., they can better evaluate how each rule may be applied individually.

V. CONCLUSIONS AND FUTURE WORK

In this study, we introduced an innovative step to the methodologies that aim at extract and select decision rules from ensembles of trees in order to improve its interpretability, while assuring its good prediction performance. Such novelty is related to the prediction of the correctness of each rule for a given patient, which is then used to weigh the final output prediction. This personalization capability seems to improve the ability of the models to correctly classify the patient's outcome. In short, the development of this approach in the clinical domain might assume great importance.

The proposed methodology can be further improved. For example, different methods for rules' subset selection and rules' correctness prediction may be applied, which may increase the performance ability of the models. Finally, its extension to multiclass and regression problems, which are common in the clinical context, may also be considered.

ACKNOWLEDGMENT

This work was supported by the lookAfterRisk research project (POCI-01-0145-FEDER-030290).

REFERENCES

[1] I. G. Stiell and C. Bennett, "Implementation of Clinical Decision Rules in the Emergency Department," *Academic Emergency Medicine*, vol. 14, no. 11, pp. 955–959, nov 2007.

[2] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, jan 2019.

[3] B. A. Mateen, J. Liley, A. K. Denniston, C. C. Holmes, and S. J. Vollmer, "Improving the quality of machine learning in health applications and clinical research," *Nature Machine Intelligence*, vol. 2, no. 10, pp. 554–556, oct 2020.

[4] O. Sagi and L. Rokach, "Explainable decision forest: Transforming a decision forest into an interpretable tree," *Information Fusion*, vol. 61, pp. 124–138, 2020.

[5] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?" *Journal of Machine Learning Research*, vol. 15, no. 90, pp. 3133–3181, 2014.

[7] A. Moore, V. Murdock, Y. Cai, and K. Jones, "Transparent Tree Ensembles," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, ser. SIGIR '18. New York, NY, USA: ACM, jun 2018, pp. 1241–1244.

[8] S. Hara and K. Hayashi, "Making Tree Ensembles Interpretable: A Bayesian Model Selection Approach," in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Storkey and F. Perez-Cruz, Eds., vol. 84. PMLR, 2018, pp. 77–85.

[9] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 916–954, sep 2008.

[10] M. Mashayekhi and R. Gras, "Rule extraction from decision trees ensembles: new algorithms based on heuristic search and sparse group lasso methods," *International Journal of Information Technology and Decision Making*, vol. 16, no. 6, pp. 1707–1727, 2017.

[11] N. Meinshausen, "Node harvest," *The Annals of Applied Statistics*, vol. 4, no. 4, dec 2010.