How to produce complementary explanations using an Ensemble Model

Wilson Silva*
INESC TEC and
Faculdade de Engenharia, Universidade do Porto
Porto, Portugal
wilson.j.silva@inesctec.pt

Kelwin Fernandes*
NILG.AI and
INESC TEC
Porto, Portugal
kelwin@nilg.ai

Jaime S. Cardoso
INESC TEC and
Faculdade de Engenharia, Universidade do Porto
Porto, Portugal
jaime.cardoso@inesctec.pt

***Abstract*—In order to increase the adoption of machine learning models in areas like medicine and finance, it is necessary to have correct and diverse explanations for the decisions that the models provide, to satisfy the curiosity of decision-makers and the needs of the regulators. In this paper, we introduced a method, based in a previously presented framework, to explain the decisions of an Ensemble Model. Moreover, we instantiate the proposed approach to an ensemble composed of a Scorecard, a Random Forest, and a Deep Neural Network, to produce accurate decisions along with correct and diverse explanations. Our methods are tested on two biomedical datasets and one financial dataset. The proposed ensemble leads to an improvement in the quality of the decisions, and in the correctness of the explanations, when compared to its constituents alone. Qualitatively, it produces diverse explanations that make sense and convince the experts.**

***Index Terms*—Interpretable Machine Learning, Explainable Machine Learning, Ensemble Model, Scorecards, Random Forests, Deep Neural Networks, Dermoscopics, Aesthetic Evaluation, Credit Scoring**

I. Introduction

Machine learning models are increasingly present in our day-to-day lives. This strong appearance is mainly due to the incredible performances obtained in the last years in the most diverse tasks, from vision [1]–[3] to language [4]–[6]. The mastery of tasks like those by machine learning models inspired their use in areas of great importance to society, such as medicine [7]–[9] and finance [10]–[12]. However, these areas are highly regulated, which means that performance is not enough. It is necessary to have great performances and at the same time, to understand the decisions provided by the models.

However, the task of generating explanations for the decisions that models make is not easy, due to the possible complexity of the models, and especially to the variability in application domains and target public/users. Moreover, even for a fixed domain, the most adequate type of explanation may vary with the example under evaluation. Thus, having a model capable of providing diverse and complementary types of explanations is imperious. For that, one has to combine different machine learning methods and different types of explanations, satisfying the needs of the decision-maker. In this sense, an ensemble of models appears as the obvious approach.

While it is true that an ensemble of models is less interpretable than its constituents alone, it is also possible to provide an explanation for its decisions, focusing our attention on the parts of the ensemble that are in agreement with its global decision.

This work was partially supported by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e Tecnologia within project POCI-01-0145-FEDER-028857, and PhD grant number SFRH/BD/139468/2018.

* These two authors contributed equally to this work.

II. Literature Review

The literature is rich in the use of different machine learning methods and in different approaches to obtain interpretability, or in a broader sense, to produce explanations for the decisions that models make. Nevertheless, one can think of interpretability, as a three-stage process, closely related to the development cycle of a data science solution. In accordance with this idea, Kim and Doshi-Velez [13] grouped the different strategies in pre-, in-, and post-model.

A. Pre-Model

The first stage, pre-model, focus on trying to understand the data itself and happens before the construction of the machine learning model. In here, visualization and exploratory data analysis play a significant role.

Visualization is a quite common technique in the business intelligence community, and it basically consists on visualizing the behaviour/distribution of the data according to the different features available [14]. Exploratory data analysis, concept firstly introduced by Tukey [15], also focus on the understanding of the data but it is more general than visualization, also including quantitative techniques apart from the graphical ones. These two strategies can be fundamental for building trust in the subsequent machine learning model. Furthermore, an understanding of the behaviour of the data in accordance with the features available might help in the construction of new features (hand-crafted), which is especially important when one is working with simpler models. When dealing

with highly complex data distributions, the use of prototypes, i.e., examples that characterize well the data or a particular class of elements, is beneficial. However, in the context of having a distribution, in which some data points are not well characterized by a given prototype, prototypes are not enough and the MMD-critic framework proposed by Kim *et al.* [16], which also selects the criticisms, is fundamental. In summary, pre-model interpretability is not enough when isolated, but it plays an important role when integrated into the general context of interpretability. Only understanding first the data distribution that we are dealing with, one can be confident with the posterior decisions and explanations that a given machine learning model is capable of providing.

B. In-Model

In-model approaches, on the other hand, focus on integrating interpretability inside the model. In order to build an interpretable model or to make a machine learning model more interpretable, there are several different strategies available.

One of the first strategies that comes to mind when thinking about making an interpretable model, because it is closely related to, and inspired by, human nature, is to build a model based on rules, which are able to characterize the different classes in question. A widely known example of such a model is a decision tree [17]. However, other models like decision lists [18], and rule sets [19] are also valid options. Also related are the per-feature based models, like the generalized additive models [20] in general and their discretized version, scorecards [21], which are extensively used in industry, and in particular in finance. Nonetheless, the interpretability of these models is limited by the semantic meaning of the original features, the complexity of the rules, and by the size of the model, or depth in the case of decision trees.

Another strategy, closely related to the way human beings think, is to build models based in cases instead of in rules. Exploring again the idea of prototypes, decisions and explanations can be obtained through cluster divisions, with each cluster being characterized by a prototype [22]. However, the quality of an explanation generated using this approach is limited from the beginning by the representativeness of the prototypes. Moreover, the formation of the clusters typically depends on the distance metric considered, which may not be the more suited for the context being explored.

Now, instead of thinking in natural/obvious ways to base models in, we can search for procedures to increase the interpretability of a certain model, which is not interpretable in its usual implementation. For example, complex and non-interpretable models, like deep neural networks, can be made more interpretable using some regularization techniques that simplify the model. One of such techniques, which aims to achieve sparsity, is the well known $L1$ regularization [23]. This regularization technique consists on the sum of the absolute values of the model individual parameters, w_i, and is mathematically described in Equation (1).

$$\Omega(\theta) = ||\omega||_1 = \sum_i |\omega_i| \tag{1}$$

In the context of linear regression, the addition of this term results in the famous LASSO model [24]. With this addition, a subset of weights becomes zero, which means that some features are discarded, a property that obviously increases the interpretability of the method.

Another property with great interest in the interpretability domain is a monotonic relationship to some or, ideally, all of the input features [25]. In the context of neural networks, it can be obtained constraining the weights to be positive, or negative, depending on the increasing or decreasing nature of the function to be learned [26]. However, it is important to note that the introduction of these regularization techniques ($L1$ and monotonicity) that help improving interpretability due so in expense of model complexity and therefore can have a significant negative impact in model performance.

C. Post-Model

The last stage, post-model, has the aim of understanding the model decisions but already after a model has been built. In here, a possible strategy can be the perturbation of the input provided to the model and the analysis of the consequent impact on the model output. This strategy is known as sensitivity analysis. When working with images, a possible perturbation is the occlusion of some parts of the image [27]. Related with this approach are the gradient-based methods. Instead of occluding regions of the image, these methods use gradient information to identify the areas of the image that mostly contribute to the final decision (e.g., class that the images belong to). In the last years, several works inspired by this idea were presented. Examples of this are the works of Baehrens *et al.* [28], Simonyan *et al.* [29], Smilkov *et al.* [30], and Springenberg *et al.* [31]. Nonetheless, there is no guarantee that changes made to the input represent a realistic scenario, and spatial explanations typically do not have a rich semantic meaning.

A different strategy is to mimic a more complex model with a simpler one. Being simpler, a model is consequently more interpretable. In the context of neural networks, an example of this would be to try to imitate the behaviour of a very deep model with a more shallow one [32]. Two issues with this approach are the fact that a simpler model may not exist, and that it is difficult to verify if the mimic model is really representative of what the more complex model is doing.

Lastly, one can try to understand what the model is doing by looking to feature representations learned by the model when trying to solve a certain task. In case of being working with neural networks, this is done by observing the latent or hidden units of the network. However, an understanding of what is being represented in the learned semantic space is usually not easy. Therefore, some techniques were developed, which mainly consist of inverting the representations back to the input pixel space [33], [34] or to connect the representations to semantic concepts [35].

D. Conclusions

As pointed out, there are several different strategies to try to interpret the behaviour of a machine learning model. Nonetheless, all of them have some weaknesses. As such, it appears that the best approach to follow is the integration of various techniques, which will increase confidence in the decisions and explanations provided by the model and interpretability techniques, respectively. Therefore, a holistic approach to interpretability is fundamental, conjugating the three stages of interpretability (pre-, in-, and post-model) and different strategies inside each of them.

III. The Proposed Model

The combination of different models inside a global one allows the use of different interpretability strategies at the same time, which improves the quality of the explanations. Moreover, having different models results in a diversity of types of explanations, from the ones based on rules to the ones based on examples.

Diversity is important because people and application domains vary a lot, and what is suited for a certain person or for a particular application domain may not be suited for others.

Decisions regarding the interpretability strategies to be used, and the way we propose to generate explanations, were based in a framework that we have presented in a previous work [36]. This framework aims to provide a quantitative evaluation of the explanations, turning a usually subjective subject into a more objective one. The framework was named as **"the three C's of interpretability"**, and refers to three interesting properties that an explanation should have:

- **Completeness:** an explanation should be complete, i.e., it should be general enough for it to be applied to more than one observation. Defining the covered set as the set of training cases covered by the explanation, completeness is the ratio between the sizes of the covered set and the training set (presented in percentage terms).
- **Correctness:** an explanation should be correct, in the sense that if we consider the explanation itself as a model, it should be able to correctly identify the class to which the current observation belongs to. Quantitatively, it means the percentage of the covered set that is correctly classified, or in other words, it is the accuracy of the explanation as a model.
- **Compactness:** an explanation should be compact, or in other words, it should be succinct. If an explanation is very long, it is explaining nothing. Quantitatively, compactness is measured as the size in bytes of the explanation.

Having this framework into consideration, we are able to propose a method to generate explanations for the predictions of ensemble models. For an ensemble model with N models (Fig. 1), the prediction made by the ensemble is given by the majority vote within its elements. Thus, we can look to the subset M of models that agree with the global decision and select the explanations that these models provide to a pool

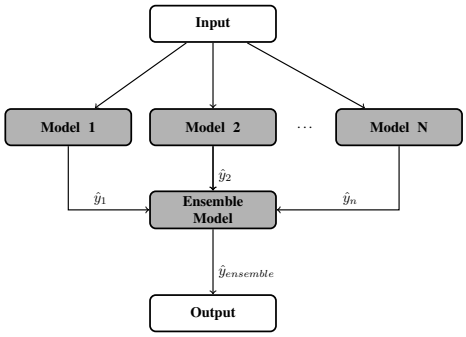


Fig. 1: Ensemble Model.

of candidate explanations for the ensemble. The final global explanation of the ensemble for a given test point x is the one from the pool of M candidate explanations ($E_n(x)$) that has the highest correctness ($corr(E_n(x))$ - Eq. (2).

$$E_{global}(x) = \underset{E_n(x)}{\operatorname{argmax}}(corr(E_n(x)), n \in \{1, ..., M\} \wedge M \leq N \tag{2}$$

In this work, we also propose a concrete example of an ensemble model to provide complementary explanations, which is constituted by the models that we present in the following subsections.

A. Deep Neural Networks

In a previous work [36], we proposed a deep model capable of generating complementary explanations regarding style and depth. The model, which is presented in Figure 2, consists of two different streams, one for the monotonic features and another for the non-monotonic features. The first stream, monotonic one, had a non-negative constraint in its weights. The second stream, on the contrary, does not have any constraint. However, the final layers have that non-negative constraint for the weights as well, which means that we are forcing the model to map the originally non-monotonic features to a latent space where they are also monotonic. For that to happen, we are considering a higher number of layers for the second stream.

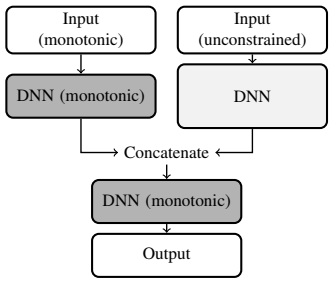


Fig. 2: DNN architecture [36].

Using the model described, we thought about extracting two different kinds of explanations: a rule-based explanation and a case-based explanation. To generate the rule-based explanation

we find the contribution of a particular feature, C_{ft}, through an adversarial example. After computing the contribution of all features, we can construct a rule explaining the decision made by the model. To generate the case-based explanation, we look for the nearest neighbors in a learned semantic space, which is adapted to the task being performed. As such, the neighbors have a much more semantic meaning than if they were found in the original space of features, which means that they are a better explanation for the decision being made. Thus, we found it useful to include two types of neighbors:

- **Same class:** the nearest neighbor from the same class, and the explanation for why they are from the same class.
- **Opposite class:** the nearest neighbor from the opposite class, and the explanation for why they are from different classes.

The explanations for why the current observation is of the same or opposite class than its neighbors are extracted using sensitivity analysis, and observing the impact of each feature to the distance in the latent space being analyzed.

With regards to this model, we consider in-model interpretability strategies and post-model interpretability strategies. Considering in-model interpretability, the network is regularized by the imposition of monotonic constraints, i.e., by constraining the weights of some layers to be non-negative. This assumes that at least some input features are monotonic, more precisely increasingly monotonic. Moreover, we search in an hidden/latent space for the nearest neighbors, which is considered a post-model interpretability strategy.

In this work, we will only consider the case-based explanations provided by the deep neural network, being the rule-based explanations generated by the other models of the ensemble.

B. Scorecards

A scorecard is an intrinsically interpretable model widely used in financial applications, particularly regarding credit scoring [21]. In Table I, we exemplify how this model works. Considering a new observation, depending on the values of its features, the observation will get a particular number of points per feature, resulting in a total score that will determine the class that it belongs to.

TABLE I: Example of a scorecard.

	Bins	**Points**
Feature X		
	Up to **x1**	10
	x1 to **x2**	25
	x2 to **x3**	38
	x3 and up	43
Feature Y		
	Up to **y1**	50
	y1 to **y2**	65
	y2 and up	70
Total Score		**88**

One of the critical aspects when building a scorecard is defining the way the discretization of the features is made. Typical strategies include performing the discretization using equal-width or equal-frequency algorithms. Equal-width consists on dividing the range of feature values in a pre-defined number of bins with the same width. On the other hand, equal-frequency consists on dividing the range of feature values also in a pre-defined number of bins but with each bin having the same number of training observations. Both algorithms belong to the unsupervised category, i.e., division of the bins does not take into account the class of the observations. In this work, we perform the binning of the scorecard using a decision tree, and its thresholds as the cut-off points of the bins. Figure 3 provides an illustration of a decision tree applied to a particular feature X. The discretization of feature X is done accordingly to the thresholds previously computed, resulting in the illustrative example presented in Table II.

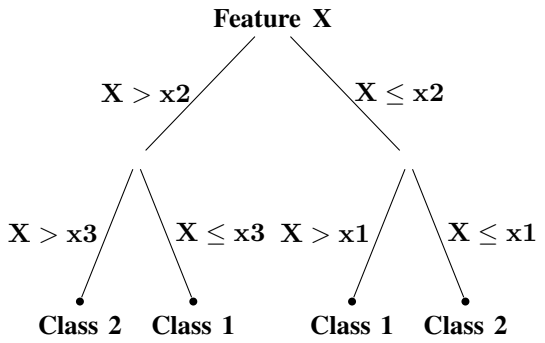


Fig. 3: Example of a Decision Tree applied to a particular feature X.

TABLE II: Example of a scorecard with discretization based on Decision Tree of Fig. 3. The scores are only illustrative.

	Bins	**Points**
Feature X		
	Up to **x1**	50
	x1 to **x2**	25
	x2 to **x3**	30
	x3 and up	60

Our implementation of the scorecard was based on neural networks, with the weights of the neurons being the weights of the scorecard (Fig. 4). Regarding in-model interpretability, we have considered three regularization techniques. The first one was to use differential-coding in the bins [37], which promotes a smooth variation in the points attributed to consecutive bins. The second one was to use $L1$ regularization over the differential-coding of the bins to ensure a sparse number of scores. Finally, the third one was to impose non-negative constraints in the weights of the neural network/scorecard, which in conjunction with the differential-coding ensures monotonicity. Moreover, we also increased the interpretability of the model after the same has been built (post-model interpretability). For this, we merge neighboring bins that are monotonic on the decision for a given local sample in an attempt to improve the completeness of the explanations.

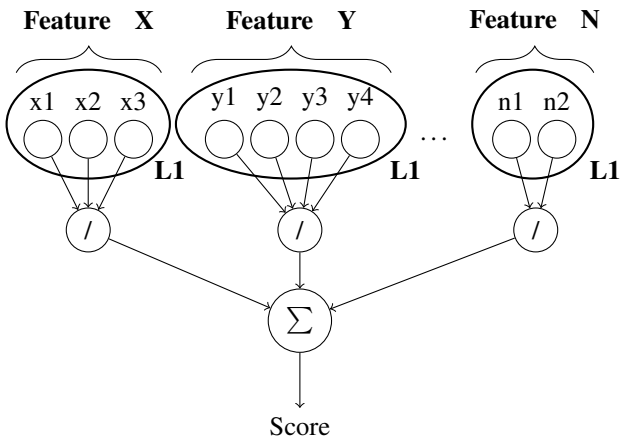


Fig. 4: Scorecard implemented with a neural network. Input neurons represent the bins, and / the linear activation function.

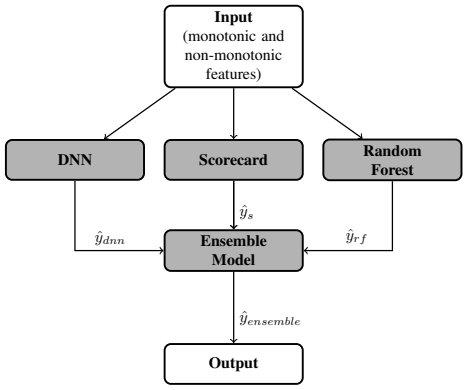


Fig. 5: Ensemble Model Proposed.

C. Random Forest

The last model we considered as part of our ensemble is itself an ensemble, a Random Forest [38]. A Random Forest is an ensemble of decision trees, and it was created because decision trees alone tend to overfit to the training data. When using a multitude of trees and averaging the predictions from all trees of the ensemble, we are able to reduce overfitting, and therefore, to have a more robust model. This robustness comes at the expense of more complexity, and consequently less interpretability. However, although random forests are not considered interpretable, the individual trees within the ensemble are (at least given a small limited depth). Thus, we can find an explanation for the ensemble decision using the approach we proposed, selecting the explanation from the tree with the path from its source to the tree leaf that leads to a more correct explanation.

Considering post-model interpretability, we prune the tree branches that do not lead to further class refinement, and so we are able to produce more complete explanations.

D. Ensemble Model

The model we propose in this work is then constituted by the previously presented models: Deep Neural Network, Scorecard, and Random Forest. The prediction made by the ensemble, $\hat{y}_{ensemble}$, is given by the majority vote computed based on the predictions $\hat{y}_{dnn}$, $\hat{y}_s$, and $\hat{y}_{rf}$, which are the predictions made by the Deep Neural Network, Scorecard, and Random Forest, respectively (Figure 5). Final explanation is chosen accordingly to the proposed method.

IV. Experimental Assessment

We validate the performance of the proposed ensemble model on three datasets, one from the financial domain and the others from the biomedical/medical domain.

Unlike the medical datasets, the original features of the financial dataset are "raw" features, defined without the help of an expert. Thus, it is important to consider pre-model interpretability strategies, with a subsequent step of feature engineering that will define new extended features[1]. Our results are computed using the extended feature version.

The financial dataset being used is the 2018 FICO Explainable Machine Learning Challenge's Credit dataset [39]. This dataset is an anonymized dataset of Home Equity Line of Credit (HELOC) applications. The problem related to the dataset is a binary classification with the target variable being called of RiskPerformance. RiskPerformance can be:

- **Good:** which means that the applicant made his/her payments without ever being more than 90 days overdue.
- **Bad:** which means that the applicant was 90 days past due or worse at least once over a period of 24 months from when the credit account was open.

Regarding the medical datasets, we considered one relative to dermoscopy image classification [40] and another to aesthetic evaluation of breast cancer treatments [41].

The first medical dataset, dermoscopic image classification [40], has 14 high-level features acquired from 200 patients. Features describe the presence of certain colors on the nevus and abnormal patterns. The goal of the problem related to this dataset is to classify each observation in three different classes: Common, Atypical, and Melanoma. For this work, as we only consider binary classification, we have binarized the problem in two different ones: Common vs. Atypical and Melanoma, and Common and Atypical vs. Melanoma. The second medical dataset, aesthetic evaluation of breast cancer treatments [41], has 23 high-level features acquired from 143 patients. Features describe breast asymmetry in terms of shape, and local and global differences in color. Local differences in color aim to detect scars in the breasts. The aesthetic evaluation of breast cancer treatments consists on an ordinal classification problem composed of four different classes: Poor, Fair, Good, and Excellent. In here, we considered the three binary classification tasks:

- **Excellent** vs. **Good, Fair, and Poor**
- **Excellent, and Good** vs. **Fair, and Poor**
- **Excellent, Good, and Fair** vs. **Poor**

[1]Code used in this work, and description of the feature engineering process considered for the FICO dataset is available at https://github.com/wjsilva19/Complementary_Explanations_Ensemble.

TABLE III: Quality of the predictions in terms of area under the ROC and Precision-Recall curves. Quality of the explanations in terms of correctness (Corr), completeness (Compl), and compactness (Compt).

PH^2: Dermoscopy Images [40]							
Binarization	**Model**	**Predictions**		**Explanations**			
		ROC	**PR**	**Type**	**Corr**	**Compl**	**Compt**
Common *vs.* Atypical, Melanoma	Random Forest	99.53	99.70	Rule	97.30	10.49	33.32
	Scorecard	99.17	99.60	Rule	82.97	24.23	23.96
	DNN	**99.74**	**99.83**	Similar Opponent	97.11 74.59	39.00 70.61	19.32 37.69
	Ensemble	99.64	99.76	Best	92.27	16.69	30.96
Common, Atypical *vs.* Melanoma	RF	96.33	87.41	Rule	95.01	13.59	32.88
	SC	95.86	87.85	Rule	82.94	38.10	24.00
	DNN	96.02	89.30	Similar Opponent	91.49 84.02	8.15 62.12	33.27 46.24
	Ensemble	**96.64**	**89.43**	Best	94.76	18.65	35.25
BCCT: Breast Aesthetics [41]							
Binarization	**Model**	**Predictions**		**Explanations**			
		ROC	**PR**	**Type**	**Corr**	**Compl**	**Compt**
Excellent *vs.* Good, Fair, Poor	Random Forest	90.06	75.28	Rule	97.68	7.97	33.14
	Scorecard	92.95	78.25	Rule	96.81	26.19	24.68
	DNN	91.03	73.00	Similar Opponent	87.25 92.82	1.46 67.86	79.79 157.81
	Ensemble	**93.72**	**79.58**	Best	94.91	14.98	74.36
Excellent, Good *vs.* Fair, Poor	Random Forest	86.00	82.27	Rule	94.80	4.87	46.32
	Scorecard	86.72	**83.31**	Rule	81.57	22.60	24.87
	DNN	86.78	82.82	Similar Opponent	72.52 81.16	17.34 31.28	80.36 138.00
	Ensemble	**87.28**	82.47	Best	86.61	21.43	87.69
Excellent, Good, Fair *vs.* Poor	Random Forest	83.49	97.32	Rule	97.50	13.95	27.12
	Scorecard	85.03	97.35	Rule	98.00	10.60	33.34
	DNN	80.61	96.55	Similar Opponent	85.69 92.04	95.20 46.87	124.94 149.68
	Ensemble	**85.61**	**97.72**	Best	96.27	12.67	63.73
FICO Explainable ML Challenge [39]							
Binarization	**Model**	**Predictions**		**Explanations**			
		ROC	**PR**	**Type**	**Corr**	**Compl**	**Compt**
Negative *vs.* Positive	Random Forest	**77.61**	75.46	Rule	84.18	5.77	57.87
	Scorecard	76.35	74.55	Rule	73.57	17.96	30.97
	DNN	76.71	74.88	Similar Opponent	87.22 49.70	0.38 99.01	114.64 199.24
	Ensemble	77.41	**75.58**	Best	82.11	30.76	101.74

We compared the performance of the proposed Ensemble Model against its constituents alone: a Deep Neural Network (previously proposed [36]), a Scorecard (which is a highly interpretable model), and a Random Forest. We used 10-fold cross-validation to choose the best hyper-parameter configuration and to generate explanations. Scorecard bins per feature were limited to 20 for FICO dataset, and 10 for the medical datasets. The depth of trees within the Random Forest was limited to 5, being a good trade-off between predictive performance and model interpretability. We show in Table III the performance of the four models regarding the quality of the predictions and their respective explanations. To measure the quality of the predictions, we considered the area under the ROC curve, and Precision-Recall (PR). Our proposed Ensemble model leads to a higher predictive performance in the majority of the scenarios considered, and when it does not, it is in line with the best performing model. Regarding the quality of the explanations, the evaluation was based in the "three C's of interpretability", considering the correctness, completeness, and compactness of the explanations generated by each model. The Ensemble proposed, despite increasing the diversity of the explanations, maintains a very high correctness in the explanations that it generates.

Figures 6 and 7 illustrate the explanations obtained by the proposed ensemble on the three datasets. As can be seen in the example, different models and explanation strategies tend to support the decision using different subsets of features. Namely, they offer complementary evidence of the predicted class.

V. Conclusion

The use of machine learning models in areas like medicine and finance is highly restricted due to interpretability concerns. Both clinicians and patients, and clients and regulators want to understand the decisions provided by the models. Given the diversity of target users, the variability in application domains and in examples within the same application, the existence of a method able to generate correct explanations along with diversity and complementarity is fundamental.

In this work, we presented an approach to select the global explanation of an ensemble. Moreover, we also proposed an Ensemble Model capable of fulfilling the need for correctness and diversity.

Input image
(**Prediction:** {Poor, Fair})

Random Forest: Low inter-breast alignment (pBRA) and high color difference (cX2Lab)

Scorecard: High color difference (cEMDa) and scar visibility (sX2b)

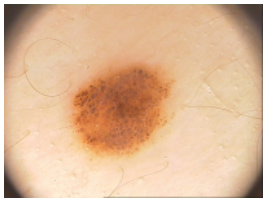
Similar case

Why?: Similar scar ($sEMDL$), inter-breast overlap ($pBOD$), color ($cEMDb$), contour difference ($pBCD$) and upward nipple retraction ($pUNR$).

Opponent case

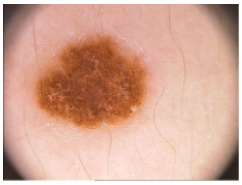
Why?: Strong difference on the scar visibility ($sX2a$), breast overlap ($pBOD$), upward nipple retraction ($pUNR$), compliance evaluation ($pBCE$) and lower contour ($pLBC$)

Input image
(**Prediction:** {Common, Atypical})



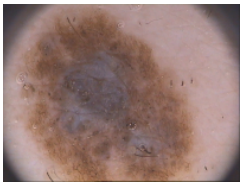
Random Forest: Lack of regression, presence brown color and lack of white color.

Scorecard: Presence of brown color, lack of atypical dots and asymmetry.



Similar case

Why?: Both images have light and dark brown color and atypical presence of dots/globules.



Opponent case

Why?: It doesn't have light brown color or atypical dots/globules. It has blue whitish veil and pigmented network.

Fig. 6: Visualization of the explanations. In the BCCT dataset we are considering the binary classification problem: {Poor, Fair} *vs.* {Good, Excellent}. Regarding the PH^2, the classification problem comes down to {Common, Atypical} *vs.* {Melanoma}.

Prediction: Rejected

- **Scorecard:** The consolidated version of risk markers is below 81.50, the maximum of the relative values of the average number of months in file and the percentage of trades never delinquent is above 0.67, the average number of months in file is below 97.50, the sum of the relative values of the average number of months in file and the percentage of trades never delinquent is above 0.57, and the average relative position of the client's features is above 0.50.
- **Random Forest:** Although the sum of the relative values of the percentage of trades never delinquent and the net fraction revolving burden is below 0.89, the following facts support the decision: condition of the number of months since most recent delinquency not met and the average relative position of the client's features is above 0.51.
- **Deep Neural Network:**
  - **Similar:** This client is rejected because No usable/valid trades or inquiries observed in the number of months Since Most Recent inquiries (excluding the last 7 days), the average number of months in file with value 41.0, the number of satisfactory trades with value 2.0, the number of months Since Most Recent inquiries (excluding the last 7 days) with value 0.0, and the number of Months Since Most Recent Delinquency with value 15.0 are similar to client in row 5662. Client 5662 could not payoff.
  - **Opponent:** The minimum of the relative values of the number of months Since Most Recent inquiries (excluding the last 7 days) and the Net Fraction Revolving Burden with value 0.0, the average number of months in file with value 41.0, the percentage of Trades Never Delinquent with value 100.0, the number of Months Since Most Recent Delinquency with value 15.0, and the Net Fraction Revolving Burden with value 0.0 are different to client in row 4340, with values 0.5, 219.0, 86.0, 1.0, and 31.0 respectively.. Client 4340 paid.

Fig. 7: Explanations obtained in the FICO Explainable ML Challenge.

The Ensemble Model is evaluated in three datasets, one financial (FICO Explainable Machine Learning Challenge) and two biomedical (Dermoscopic Image Classification, and Aesthetic Evaluation of Breast Cancer Treatments). Regarding the quantitative results, the proposed ensemble leads to a higher predictive performance when compared with the Deep Neural Network, Scorecard, and Random Forest alone. Moreover, the explanations that it generates have very high values of correctness, without loosing to much completeness, and maintaining a reasonable compactness. In its turn, qualitative results show that the goal of obtaining diversity in the explanations generated is fulfilled. Moreover, the explanations are in accordance with the analysis of experts (clinicians in the case of the medical applications, and bankers in the case of credit).

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Advances in Neural Information Processing Systems*, 2015, pp. 1693–1701.

[5] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning*, 2014, pp. 1188–1196.

[6] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," *arXiv preprint arXiv:1602.02410*, 2016.

[7] K. Fernandes, D. Chicco, J. S. Cardoso, and J. Fernandes, "Supervised deep learning embeddings for the prediction of cervical cancer diagnosis," *PeerJ Computer Science*, vol. 4, p. e154, 2018.

[8] A. BenTaieb and G. Hamarneh, "Predicting cancer with a recurrent visual attention model for histopathology images," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, pp. 129–137.

[9] H. Tang, D. R. Kim, and X. Xie, "Automated pulmonary nodule detection using 3d deep convolutional neural networks," in *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*. IEEE, 2018, pp. 523–526.

[10] J. Sirignano, A. Sadhwani, and K. Giesecke, "Deep learning for mortgage risk," *arXiv preprint arXiv:1607.02470*, 2016.

[11] F. Butaru, Q. Chen, B. Clark, S. Das, A. W. Lo, and A. Siddique, "Risk and risk management in the credit card industry," *Journal of Banking & Finance*, vol. 72, pp. 218–239, 2016.

[12] M. Abe and H. Nakayama, "Deep learning for forecasting stock returns in the cross-section," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2018, pp. 273–284.

[13] B. Kim and F. Doshi-Velez, "Interpretable machine learning: The fuss, the concrete and the questions," 2017.

[14] K. R. Varshney, J. C. Rasmussen, A. Mojsilović, M. Singh, and J. M. DiMicco, "Interactive visual salesforce analytics," 2012.

[15] J. W. Tukey, *Exploratory data analysis*. Reading, Mass., 1977, vol. 2.

[16] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016, pp. 2280–2288.

[17] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks, 1984.

[18] R. L. Rivest, "Learning decision lists," *Machine learning*, vol. 2, no. 3, pp. 229–246, 1987.

[19] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille, "A bayesian framework for learning rule sets for interpretable classification," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 2357–2393, 2017.

[20] T. Hastie and R. Tibshirani, "Generalized additive models," *Statistical Science*, vol. 1, no. 3, pp. 297–318, 1986.

[21] FICO, "Introduction to Scorecard for FICO Model Builder," Tech. Rep., 2006.

[22] B. Kim, C. Rudin, and J. A. Shah, "The bayesian case model: A generative approach for case-based reasoning and prototype classification," in *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014, pp. 1952–1960.

[23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[24] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[25] M. Gupta, A. Cotter, J. Pfeifer, K. Voevodski, K. Canini, A. Mangylov, W. Moczydlowski, and A. van Esbroeck, "Monotonic calibrated interpolated look-up tables," *Journal of Machine Learning Research*, vol. 17, no. 109, pp. 1–47, 2016.

[26] J. Sill, "Monotonic networks," in *Advances in neural information processing systems*, 1998, pp. 661–667.

[27] K. Fernandes, J. S. Cardoso, and B. Astrup, "A deep learning approach for the forensic evaluation of sexual assault," *Pattern Analysis and Applications*, 2018.

[28] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. MÃžller, "How to explain individual classification decisions," *Journal of Machine Learning Research*, vol. 11, no. Jun, pp. 1803–1831, 2010.

[29] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[30] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[31] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[32] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014, pp. 2654–2662.

[33] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014*. Cham: Springer International Publishing, 2014, pp. 818–833.

[34] A. Dosovitskiy and T. Brox, "Inverting visual representations with convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4829–4837.

[35] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," *arXiv preprint arXiv:1704.05796*, 2017.

[36] W. Silva, K. Fernandes, M. J. Cardoso, and J. S. Cardoso, "Towards complementary explanations using deep neural networks," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Springer, 2018, pp. 133–140.

[37] P. F. Silva and J. S. Cardoso, "Differential scorecards for binary and ordinal data," *Intelligent data analysis*, vol. 19, no. 6, pp. 1391–1408, 2015.

[38] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[39] FICO, "Explainable machine learning challenge," https://community.fico.com/s/explainable-machine-learning-challenge, 2018.

[40] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, "PH2 - a dermoscopic image database for research and benchmarking," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, July 2013, pp. 5437–5440.

[41] J. S. Cardoso and M. J. Cardoso, "Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment," *Artificial Intelligence in Medicine*, vol. 40, pp. 115–126, 2007.