A Framework for Analyzing Fairness, Accountability, Transparency and Ethics: A Use-case in Banking Services

Ettore Mariotti, Jose M. Alonso
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
Santiago de Compostela, Spain
{ettore.mariotti, josemaria.alonso.moral}@usc.es

Roberto Confalonieri
Faculty of Computer Science
Free University of Bozen-Bolzano
Bozen-Bolzano, Italy
roberto.confalonieri@unibz.it

Abstract—We introduce a novel framework to deal with fairness, accountability and explainability of intelligent systems. This framework puts together several tools to deal with bias at the level of data, algorithms and human cognition. The framework makes use of intelligent classifiers endowed with fuzzy-grounded linguistic explainability. As a result, it facilitates the exhaustive comparison of (white/grey/black)-box modelling techniques in combination with different strategies for handling missing values and unbalanced datasets. The proposal is evaluated on a real-world dataset in the context of banking services and reported results are encouraging.

I. INTRODUCTION

With the growing number of Artificial Intelligence (AI) systems and applications which are deployed in everyday life, more and more demands for better comprehensible and/or explainable AI and Machine Learning (ML) / Deep Learning (DL) systems are put forward regarding questions of responsible AI [1]. Responsible AI is concerned with fairness, accountability -liability-, responsibility, and privacy in partially or fully automated decision scenarios. These push for the development of responsible technologies [2] for ethical AI approaches in human-machine interaction, where eXplainable Artificial Intelligence (XAI[1] in short) plays a key role.

While interest in XAI had subsided together with that in expert systems after the mid-1980s [4], more recent successes in ML and DL have brought explainability back into the focus. This has led to a plethora of new approaches for explanations of black-box models (e.g., [5], [6]), aiming to achieve explainability without sacrificing system performances. In the vast literature available, several notions of an explainable system have been developed, each having different requirements and goals, and different impacts on the level of Fairness, Accountability, Transparency and Ethics (FAT-E) that can be achieved. Nonetheless, a holistic framework for analyzing FAT-E in AI systems by unifying different explanation techniques is yet to be developed.

In this work, we present a framework with the focus on how explainable classifiers can play a key role in analyzing FAT-E of AI systems and applications [1], [2]. This framework is supported by interactive natural language technology for XAI [7]. Namely, we have defined a fuzzy-grounded linguistic layer on top of AI classifiers to make them self-explaining in natural language.

It is worth noting that XAI is highly appreciated in banking services, especially since the European General Data Protection Regulation (GDPR[2]) became effective in 2018. GDPR states that European citizens have the "right to an explanation" of decisions affecting them even if such decisions are made by AI systems. As test-bed, we consider the LOAN dataset because it includes all the ingredients needed to validate our framework. This dataset was first used in a Kaggle[3] competition to solve the classification task of predicting if a bank client was to be granted (or not) a loan.

The rest of the manuscript is organized as follows. Section II introduces preliminary concepts. It also briefly reviews related publications and software tools. Section III presents the new framework for FAT-E. Section IV describes the experiments and discusses the reported results. Finally, Section V concludes the paper and outlines future work.

II. RELATED WORK

A. Fairness, Accountability, Transparency and Ethics (FAT-E)

Fairness is defined as "the quality or state of being impartial, i.e., lack of favouritism toward one side or another"[4]. In this work we refer to fairness in the context of bias detection and mitigation. Bias can affect data, algorithms and human cognition. Bias must be identified, mitigated and avoided if possible. Unfortunately, the use of opaque black-box models makes it more difficult to deal with bias [8].

---

[1]The acronym XAI was made popular with the XAI challenge launched by the USA Defense Advanced Research Projects Agency (DARPA) [3].

[2]See note (71) in the preamble of GDPR https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32016R0679

[3]https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset

[4]Definitions given in this section are taken from https://www.merriam-webster.com/dictionary/

Accountability is "the quality or state of being accountable, i.e., subject to giving an account and capable of being explained". Accordingly, accountability is mainly related to auditing algorithms and verifying they perform as expected.

Transparency is "the quality or state of being transparent, i.e., interpretable and readily understood". In this work, we refer to transparency or interpretability of models. Moreover, we assess transparency in terms of model complexity as it is usually done in the context of interpretable fuzzy systems [9].

Ethics are at the core of human-centric, trustworthy and responsible AI [2]. In this work, we adhere to the guidelines provided by the European High-Level Expert Group on AI [10], [11].

B. Explainable Models for FAT-E

Models endowed with explainability can assist users in dealing with issues related to FAT-E [1]. We consider a triple-axis taxonomy regarding model-agnostic/specific, local/global and intrinsic/extrinsic explanations. Namely, we can distinguish between white and grey models which are interpretable by design (e.g., decision trees or fuzzy rule-based systems) versus black-box models [12]. Regarding the explanation of black-box models, we can distinguish between looking at the internals (i.e., intrinsic explanations) of the models or paying attention only to their output (i.e., extrinsic "post-hoc" explanations).

Post-hoc explanation approaches amount to build or extract a surrogate model which is an interpretable proxy of the black-box. They usually aim at maximizing fidelity (i.e., the accuracy of the proxy in classifying instances as close as possible to the black-box). Post-hoc explanation approaches can be typically categorized in local and global methods.

In local explanation methods, individual predictions of a black-box model are approximated by generating local surrogate models that are intrinsically interpretable. This strategy has been implemented for example in the popular LIME approach [13]. A more recent approach to linear surrogate models is CLEAR which builds counterfactual local explanations via regression [14]. This approach seems to achieve a better fidelity than LIME w.r.t. the black-box models' decision boundaries. Several other local explanation methods have been proposed based on Shapley values (e.g., [15], [16]), which quantify the influence of a feature on a decision with a game-theoretic approach.

Global approaches aim at providing a more comprehensive understanding of how a black-box model works. A global model-agnostic method that explains black-box classifiers is TREPAN [17]. TREPAN is a tree induction algorithm that recursively extracts decision trees from a trained ML classifier (the oracle). The algorithm is model-agnostic, and it can be applied to explain any black-box classifier (e.g., Multi-layer Perceptron or Random Forest). At each learning step, the oracle's predicted labels are used instead of known real labels. The algorithm also generates new synthetic examples by sampling from distributions over the given data and constraints (conditions that examples must satisfy to reach a node), so that the amount of training data used to select splitting tests and to label leaves does not decrease with the depth of the tree. The use of the oracle serves two purposes: (i) it helps to prevent the tree from overfitting to outliers in the training data; and (ii) it helps to build more accurate trees. TREPAN Reloaded was proposed as an extension of TREPAN using ontologies to guide the decision tree generation [18]. A user study showed that explanations extracted by TREPAN Reloaded are more understandable by humans compared to those extracted by the original algorithm [19].

In addition, it is worth noting that explainable fuzzy systems [9] can pave the way from interpretable fuzzy systems to XAI systems. They provide users with both local and global but also intrinsic and extrinsic factual and counterfactual explanations in natural language [20]. They have already been successfully used for generating post-hoc explanations of black-box models [21].

C. Software Tools for FAT-E

There are many tools for building explainable models and coping with FAT-E issues. Here, we only cite the most representative tools in the context of our work.

On the one hand, we have generic tools which can be used for classification, regression, modelling, etc. For example, the AI Fairness 360 (AIF360[5]) toolkit is an open-source library that includes techniques for detecting and mitigating bias in ML models. It is available in R and Python. In addition, there are two very popular Python libraries which implement LIME[6] and SHAP[7]. They are aimed at generating model-agnostic local and global explanations. Notice that the previous tools neither manage fuzzy systems nor provide users with explanations in natural language.

On the other hand, GUAJE[8] [22] is an open-source software that supports the design of explainable fuzzy systems. GUAJE provides users with factual and counterfactual explanations in natural language which are semantically grounded thanks to the use of meaningful strong fuzzy partitions (SFPs) in agreement with expert knowledge. GUAJE complies with the IEEE Std 1855-2016 for fuzzy systems modeling [23].

In addition, ExpliClas [24] is a web service aimed at providing users with multimodal (textual + graphical) local and global explanations associated to fuzzy rule-based classifiers and decision trees. ExpliClas builds classifiers endowed with linguistic explainability employing a linguistic approximation module that translates the given fuzzy rules and trees into semantically grounded fuzzy classifiers like those generated by GUAJE.

III. FRAMEWORK FOR FAT-E ANALYSIS

We have designed a computational framework that aims at assisting users in climbing the DIKW[9] pyramid [25] while

---

[5]https://github.com/Trusted-AI/AIF360
[6]https://github.com/marcotcr/lime
[7]https://github.com/slundberg/shap
[8]The Java Environment for Generating Understandable and Accurate fuzzy systems (GUAJE) is available at https://gitlab.citius.usc.es/jose.alonso/guaje/
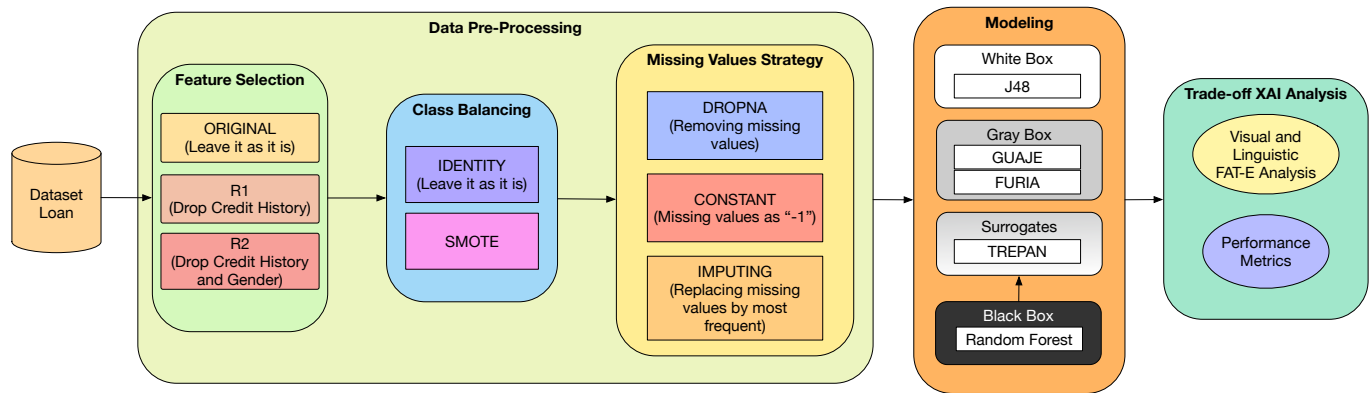[9]DIKW stands for Data, Information, Knowledge and Wisdom.

Fig. 1. Sketch of our framework for analyzing fairness, accountability, transparency and ethics in AI systems.

dealing with issues related to FAT-E: (i) extracting information from data (data pre-processing); (ii) creating knowledge from information (modelling); and (iii) gaining wisdom from knowledge (trade-off XAI analysis). The global pipeline is sketched in Fig. 1, and the three main tasks are described below.

A. Data Pre-Processing

This task aims for facilitating the analysis of fairness at the data level. Thus, we pay attention to detect and mitigate data bias from an empirical point of view. The two classes in the LOAN dataset are unbalanced (68.73% loan granted vs. 31.27% loan not granted). The dataset is made up of 614 instances (with missing values) involving 11 features which include nominal features (e.g., gender, married or credit history) and numerical features (e.g., applicant income, loan amount or loan term). Accordingly, starting from the given dataset we generate a pool of datasets to be used later in the experimental study. We consider the effects in data bias due to combining different strategies for feature selection (ORIGINAL vs. R1 vs. R2), class balancing (IDENTITY vs. SMOTE) and handling missing values (DROPNA vs. CONSTANT vs. IMPUTING). All in all, we generate 18 datasets with all possible combinations (3 x 2 x 3 = 18). For example, DS01 is ORIGINAL + IDENTITY + DROPNA, DS02 is ORIGINAL + IDENTITY + CONSTANT, and so on (see further details in Fig. 1).

It is worth noting that our goal is to show the utility, advantages and disadvantages of the proposed framework regarding FAT-E analysis. Therefore, building the most accurate classifier for the given dataset is out of the scope of this work.

B. Modeling

This task aims for analyzing accountability and transparency at the model level. Accordingly, starting from each dataset produced in the previous stage, we have generated a pool of models with different level of transparency:

- **J48** is the implementation in Weka of the Quinlan's C4.5 algorithm [26]. J48 generates pruned C4.5 classifiers which are recognized worldwide as white boxes because their behavior can be understood by traversing the trees from root to leaves.

- **GUAJE** generates fuzzy classifiers which are interpretable by design [22]. They are linguistically grounded by means of SFPs, rules extracted from fuzzy decision trees and linguistic simplification.
- **FURIA** is the Fuzzy Unordered Rule Induction Algorithm [27]. It generates fuzzy classifiers with fuzzy sets and rules which lack linguistic interpretability.
- **TREPAN** generates decision trees as surrogates of black-box models [19]. In this work, we have extended TREPAN to take Random Forest [28] as a black-box oracle because, as explained in [29], Random Forest can produce the most accurate classifiers in many problems and is commonly taken as a baseline from the accuracy viewpoint. Random Forest is an ensemble learning method that combines randomly decision trees.

C. XAI and Performance Analysis

This task aims for facilitating the comparison among the models previously generated in terms of their balance between accuracy and explainability.

First, we build a linguistic approximation of each model. Second, we compare models in terms of the interpretability-accuracy trade-off exhibited by their linguistic approximations. We apply the same linguistic approximation procedure that is implemented in ExpliClas. The approximation is done via an interpretable fuzzy classifier which is endowed with explainability by means of semantically grounded SFPs.

Then, the classifiers can be loaded into GUAJE and compared in terms of accuracy (e.g., classification ratio or F-Measure) and interpretability (e.g., number of rules, total rule length, or total number of distinct concepts in the rule base). In addition, GUAJE can generate factual and counterfactual explanations for every single model and each single data instance. Moreover, GUAJE can check consistency among rules and explanations associated to different models.

D. Implementation Details

Our framework combines several tools and programming languages. First, data preprocessing is done in Java. Then, modelling is done with code in Java and C. Namely, J48, FURIA and Random Forest are implemented in Java and available

with the Weka data mining open-source software. GUAJE is also implemented in Java. In addition, TREPAN relies on a C implementation in which the oracle is a trained neural network. The neural network can be trained using any DL libraries (e.g., PyTorch, Keras, etc.). The trained network parameters are then imported in TREPAN (as well as the network structure), and they are used to draw predictions. Nonetheless, the oracle can be changed to other implementations. In this work, we modified the TREPAN code to call the Java implementation of Random Forest. Finally, the trade-off XAI analysis is done with GUAJE but also with a script in Python which supports the generation of plots to be presented in the next section.

IV. EXPERIMENTAL STUDY

This section summarizes the reported results when applying the proposed framework to the LOAN dataset. For each one of the 18 datasets (see Sect. III-A), we generate five models (see Sect. III-B). For the sake of generality and since our focus is on FAT-E, we applied default configuration parameters for each model. Then, the goodness of the generated models was evaluated with 10-fold cross-validation. Figs. 2 and 3 aggregate reported results (for all datasets) by means of box plots. We plot average Accuracy (see Fig. 2) and F-Measure (see Fig. 3) on test sets.
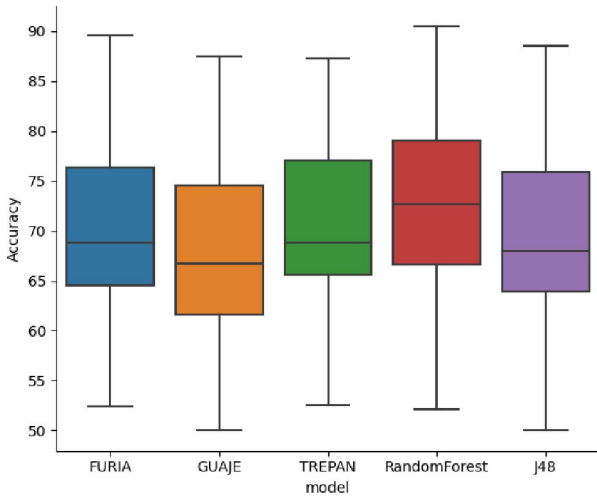

Fig. 2. Comparison of models regarding performance (accuracy is measured as the correct classification ratio).

As expected, Random Forest (the oracle black-box) is the most accurate model, followed by TREPAN (which was built as a surrogate model of Random Forest), then coming to both grey models FURIA (non-linguistic fuzzy model) and GUAJE (linguistic fuzzy model). J48 (white box) has accuracy only slightly better than GUAJE. Accordingly, we observe the characteristic accuracy-interpretability trade-off with the highest accuracy associated to the darkest models. However, regarding F-Measure (what is important here because we are dealing with an unbalanced dataset), we observe how TREPAN and GUAJE turn up as the best models, i.e., the models which
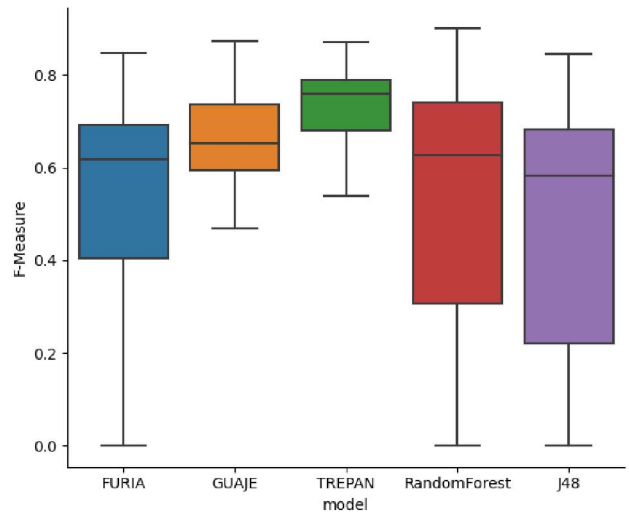

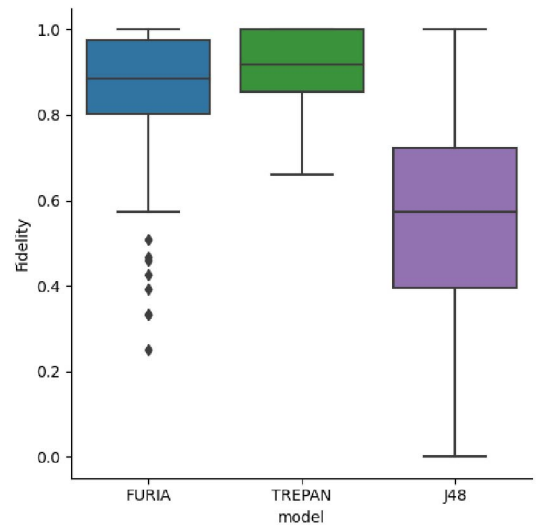Fig. 3. Comparison of models regarding performance (F-Measure).

Fig. 4. Comparison of linguistic approximations.

better classify both classes. On the contrary, the rest of the models seem to be affected by overfitting to the majority class, thus yielding wider boxplots and smaller average F-Measure.

In addition, Fig. 4 shows the goodness of the linguistic approximations in terms of fidelity (i.e., the matching ratio between the original model and its approximation; we count 1 when both outputs are the same and 0 otherwise). The more similar the output of the approximated model to the output of the original model the better. Both FURIA and TREPAN achieve good fidelity values. This means their linguistic approximation is trustworthy and reliable. From now on, we will pay main attention to FURIA, GUAJE and TREPAN (while keeping Random Forest in the plots just for comparison). Notice that, GUAJE is intrinsically linguistic so we can assume fidelity equals one.

Regarding data pre-processing, Fig. 5 shows almost no difference among the strategies for dealing with missing values. Only DROPNA seems slightly better than the others.
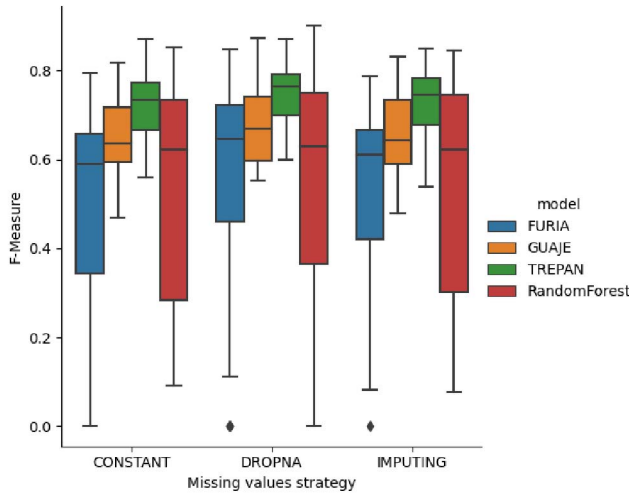

Fig. 5. Comparison of models in terms of missing values strategy.

Fig. 6. Comparison of models in terms of class balancing.

In addition, Fig. 6 shows how applying a balancing strategy as SMOTE contributes to a better F-Measure. Surprisingly, TREPAN seems to work better without balancing. This is likely to be due to the augmenting data process and rule pruning used by TREPAN which endows the model with very good generalization ability. Also, as expected, F-Measure is higher when considering the original dataset (see Fig. 7). In the case of removing credit history (R1), performance gets worse. However, there is no significant difference when removing also gender (R2). This is because credit history is by far the most relevant feature in the LOAN dataset. On the contrary, gender is ignored by the considered models. In this sense, we can say the generated models are fair because they do not rely

on features (e.g., gender) that should be controlled from the fairness viewpoint. Interestingly, this turns up automatically due to the data distribution but we may have identified and fixed it easily otherwise.
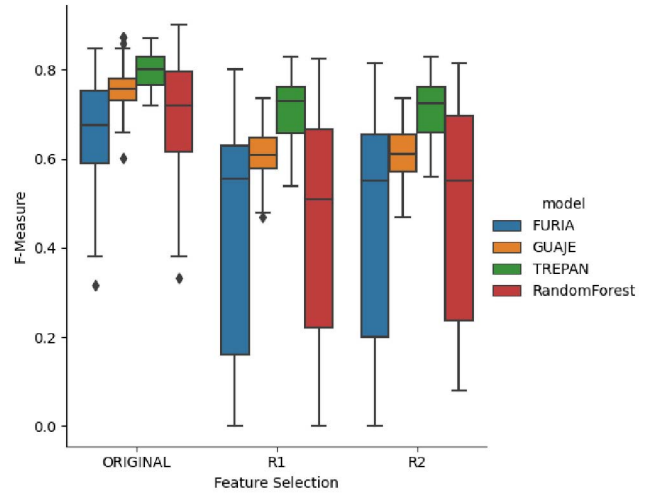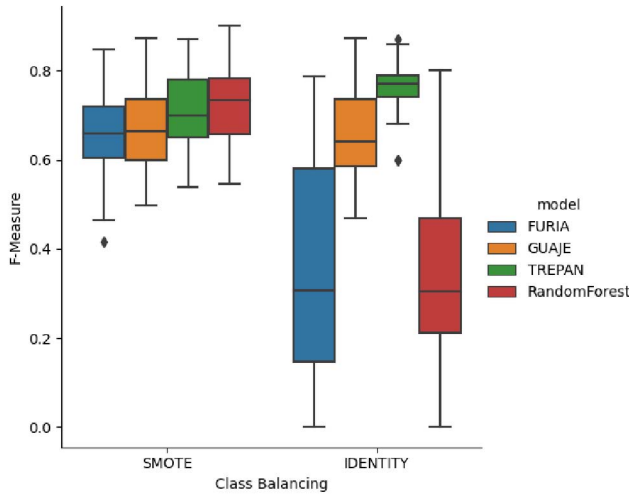

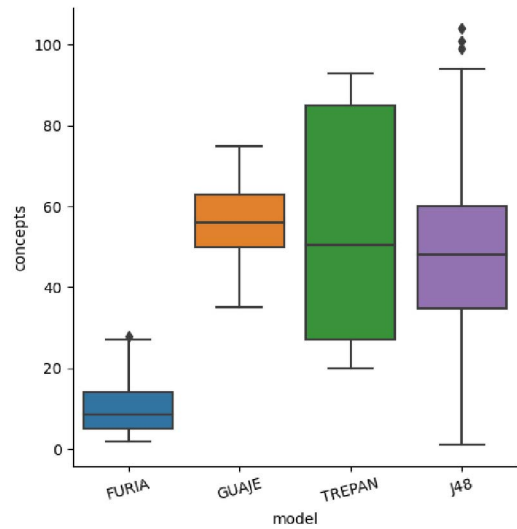Fig. 7. Comparison of models in terms of feature selection.

Fig. 8. Comparison of models in terms of distinct linguistic concepts.

In addition to the previous visual analysis, one of the main advantages of our framework is that we can analyze models at the linguistic level, with FURIA turning up as the most compact and concise model (see Fig. 8). For a brief illustrative example, let us consider DS01 (ORIGINAL+IDENTITY+DROPNA), fold CV0 and the first instance in the test dataset[10]. The explanation generated by GUAJE is as follows: "*The loan status is Granted. We have high confidence in the classification result [...] It is very likely that this loan status is granted because in accordance with rule 30, loan*

---

[10]Resources are available at https://gitlab.citius.usc.es/jose.alonso/xai

status is granted in case that Dependents is zero and Loan Amount is high and Loan Amount Term is 360 days and Credit History is yes and Property Area is semiurban. Loan status would be not granted if Credit History were not present.". In case of considering the linguistic approximation of TREPAN, then the explanation is very similar: *"The loan status is Granted [...] because in accordance with rule 12, loan status is granted in case that Loan Amount Term is 360 days and Credit History is yes and Property Area is semiurban. Loan status would be not granted if Credit History were not present.".* The explanation of FURIA is also consistent but shorter: *"The loan status is Granted [...] because in accordance with rule 2, loan status is granted in case that Credit History is yes. Loan Status would be not granted if Credit History were not present.".*

V. CONCLUSIONS AND FUTURE WORK

In this work, we have introduced a computational framework ready to assist users with visual and linguistic FAT-E analysis of AI systems. Notice that, linguistic analysis opens the door for analyzing inconsistencies among different models when paying attention to the classification of single data instances. As future work, we will validate this framework with more datasets. In addition, we will add more algorithms for data processing and modelling, as well as more quality metrics regarding both performance and explainability. Moreover, we will research for correlation between such metrics and human evaluation. With that aim, we will define a novel protocol for intrinsic and extrinsic human evaluation of AI systems.

REFERENCES

[1] A. B. Arrieta, N. D. Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): concepts, taxonomies,opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[2] N. Osman and C. Sierra, "Responsible technologies," in *GCAI-2018, 4th Global Conference on Artificial Intelligence, Luxembourg, September 18-21, 2018*, ser. EPiC Series in Computing, D. D. Lee, A. Steen, and T. Walsh, Eds., vol. 55. EasyChair, 2018, pp. 134–147.

[3] D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[4] M. R. Wick and W. B. Thompson, "Reconstructive expert system explanation," *Artificial Intelligence*, vol. 54, no. 1-2, pp. 33–70, 1992.

[5] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comp. Surv.*, vol. 51, no. 5, pp. 1–42, 2018.

[6] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, "A historical perspective of explainable artificial intelligence," *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 1, 2021.

[7] J. M. Alonso et al., "Interactive natural language technology for explainable artificial intelligence," in *1st Workshop on Foundations of Trustworthy AI integrating Learning, Optimisation and Reasoning, European Conference on Artificial Intelligence*. Springer, 2020.

[8] M. Hutson, "The opacity of artificial intelligence makes it hard to tell when decision-making is biased," *IEEE Spectrum*, vol. 58, pp. 40–45.

[9] J. M. Alonso, C. Castiello, L. Magdalena, and C. Mencar, *Explainable Fuzzy Systems. Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems*. Springer International Publishing, 2021.

[10] EU High Level Expert Group on AI, "AI Ethics Guidelines for Trustworthy AI," European Commission, Brussels, Belgium, Tech. Rep., 2019.

[11] ——, "The assessment list for trustworthy artificial intelligence (altai) for self assessment," European Commission, Brussels, Belgium, Tech. Rep., 2019.

[12] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proc. of the 22nd Int. Conf. on Knowledge Discovery and Data Mining*, ser. KDD '16. ACM, 2016, pp. 1135–1144.

[14] A. White and A. S. d'Avila Garcez, "Measurable counterfactual local explanations for any classifier," in *Proceedings of the 24th European Conference on Artificial Intelligence*, ser. Frontiers in Artificial Intelligence and Applications, vol. 325. IOS Press, 2020, pp. 2529–2535.

[15] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.

[16] L. S. Shapley, *Notes on the n-Person Game—I: Characteristic-Point Solutions of the Four-Person Game*. Santa Monica, CA: RAND Corporation, 1951.

[17] M. W. Craven and J. W. Shavlik, "Extracting tree-structured representations of trained networks," in *NIPS 1995*. MIT Press, 1995, pp. 24–30.

[18] R. Confalonieri, T. Weyde, T. R. Besold, and F. M. del Prado Martín, "Trepan Reloaded: A Knowledge-driven Approach to Explaining Black-box Models," in *Proceedings of the 24th European Conference on Artificial Intelligence*, 2020, pp. 2457–2464.

[19] R. Confalonieri, T. Weyde, T. R. Besold, and F. Moscoso del Prado Martín, "Using ontologies to enhance human understandability of global post-hoc explanations of black-box models," *Artificial Intelligence*, vol. 296, 2021.

[20] I. Stepin, A. Catala, M. Pereira-Fariña, and J. M. Alonso, "Factual and counterfactual explanation of fuzzy information granules," in *Interpretable Artificial Intelligence: A perspective of Granular Computing*, S.-M. C. Witold Pedrycz, Ed. Springer-Verlag, 2021.

[21] J. M. Alonso, J. Toja-Alamancos, and A. Bugarin, "Experimental study on generating multi-modal explanations of black-box classifiers in terms of gray-box classifiers," in *IEEE World Congress on Computational Intelligence*, 2020, pp. 1–8.

[22] D. Pancho, J. M. Alonso, and L. Magdalena, "Quest for interpretability-accuracy trade-off supported by fingrams into the fuzzy modeling tool GUAJE," *International Journal of Computational Intelligence Systems*, pp. 46–60, 2013.

[23] J. M. Soto-Hidalgo, J. M. Alonso, G. Acampora, and J. Alcala-Fdez, "JFML: a java library to design fuzzy logic systems according to the IEEE std 1855-2016," *IEEE Access*, vol. 6, pp. 56 952–56 964, 2018.

[24] J. M. Alonso and A. Bugarín, "ExpliClas: Automatic Generation of Explanations in Natural Language for Weka Classifiers," in *IEEE Int. Conference on Fuzzy Systems (FUZZ-IEEE)*, 2019, pp. 1–6.

[25] J. Rowley, "The wisdom hierarchy: representations of the DIKW hierarchy," *Journal of Information and Communication Science*, vol. 33, pp. 163–180, 2007.

[26] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[27] J. Hühn and E. Hüllermeier, "FURIA: an algorithm for unordered fuzzy rule induction," *Data Mining and Knowledge Discovery*, vol. 19, no. 3, pp. 293–319, 2009.

[28] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[29] 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, pp. 3133–3181, 2014.