Towards Model-Agnostic Ensemble Explanations

Szymon Bobek[1,2][0000−0002−6350−8405], Paweł Bałaga[1], and Grzegorz J. Nalepa[1,2][0000−0002−8182−4225]

[1] Jagiellonian Human-Centered Artificial Intelligence Laboratory (JAHCAI) and Institute of Applied Computer Science, Jagiellonian University, 31-007 Kraków, Poland

[2] AGH University of Science and Technology

{szymon.bobek,grzegorz.j.nalepa}@uj.edu.pl

Abstract. Explainable Artificial Intelligence (XAI) methods form a large portfolio of different frameworks and algorithms. Although the main goal of all of explanation methods is to provide an insight into the decision process of AI system, their underlying mechanisms may differ. This may result in very different explanations for the same tasks. In this work, we present an approach that aims at combining several XAI algorithms into one ensemble explanation mechanism via quantitative, automated evaluation framework. We focus on model-agnostic explainers to provide most robustness and we demonstrate our approach on image classification task.

Keywords: explainable artificial intelligence · machine learning · image processing

1 Introduction

Explainable Artificial Intelligence (XAI) has become an inherent component of data mining (DM) and machine learning (ML) pipelines in the areas where the insight into decision process of an automated system is important. Although the explainability (or intelligibility) is not a new concept in AI [16], it has been most extensively developed over the last decade. This is possibly due to the huge successes in black-box ML models such as deep neural networks in sensitive application contexts like medicine, industry 4.0 etc., but also a legal need of providing accountability and transparency to the reasoning process of AI systems [4]. A variety of algorithms for generating justifications for AI decisions and lack of explanations format standards, make it hard to integrate XAI methods into the standard ML/DM pipeline. Moreover, assessing quality of generated explanations is also non trivial task, as there is lack of unified metrics for evaluating XAI methods in an automated, quantitative manner.

The integration and evaluation of different ML methods into one pipeline is done via unified interfaces and metrics such as accuracy, F1 score, area under the ROC curve and many others. Different metrics may be relevant for different ML/DM tasks (recall over precision in medical diagnosis, F1 over accuracy in

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_4

imbalanced datasets, etc.). The same issue arises with explanability. Metrics such as stability, or consistency or comprehensibility may be relevant depending on who is the addressee of the explanation and what is a domain of explanations, or even what is the stage of the ML system development. The variety of explanation mechanisms makes their validation and inclusion into DM/ML process a non-trivial process.

Considering all of the above, the main goal of the work presented in this paper is to deliver a framework for calculating evaluation metrics for various XAI algorithms, and exploit these metrics in order to build an ensemble explanation mechanism that will combine explanations generated by different algorithms into one, comprehensive solution that can be easily included into the standard DM/ML pipeline. This approach can also be used to select the best explanation framework with respect to arbitrary selected criteria (metrics). We demonstrate our solution on the artificially generated, reproducible dataset and real-life scenario involving image classification task.

The rest of the paper is organised as follows. The overview of the achievements in the field of XAI with respect to assessing their quality was given in Sec. 2. The overview of our solution is given in Sec. 3. In Sec. 4 we present the results of our approach when applied to image classification taks. Finally, in Sec. 5 we discuss the limitation of the approach and describe future works.

2 Explainable Artificial Intelligence

The XAI methods are one of the most rapidly developed mechanisms in the last decade that main goal is to add transparency and accountability to machine learning (ML) and data mining (DM) models [1].

However, the analysis of explanations generated by the algorithms such as LIME [13] , SHAP [8] or Anchor [14] is most often reduced to feature selection. This is caused by the fact that such an analysis is a tedious task that involves generating multiple explanations with possibly multiple algorithms and then confronting them and assessing their quality by an expert judgement. Tools and methods for comparable analysis of results of explanations, and selecting or combining explanations are not fully investigated. Aforementioned frameworks provide basic methods for investigating explanations that are based mostly on visual presentation of results in a form of box plots, violin plots and other classical approaches. There are attempts at visualizing explanations which enhance the intelligibility of the explanations itself. However, these are mostly model-specific methods such as saliency maps for DNN [11] or task specific visualization [10].

What is more, although there are metrics used for assessing quality of explanations [9], the assessment is not automated by any framework, nor combined into a unified framework that will allow for reliable comparison of different explanation mechanisms.

There were several attempts at providing methodological approaches for evaluation and verification of given explanation results [9, 18]. Among many qualitative approaches there are also ones that allow for quantitative evaluation.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_4

In [15] measures such as fidelity, consistency and stability were coined, that can be used for a numerical comparison of methods. In [22] the aforementioned measures were used to improve overall explanations. In [2] a measure that allows to capture stability or robustness of explanations was introduced. Another explanation framework that implements evaluation metrics is given in [21]. Authors present a local explainer with evaluation metrics: stability and correctness. However in neither of the above cases the evaluation is used in further context, limiting their usefulness only to quantify the explanations given by the particular framework. In [23] authors exploit the context of features within a training instance to improve explanations generated with LIME. In [6] a context of an instance that is being explained is generated for the purpose of up-sampling and generating explanations. A more advanced approach was discussed in [17], where an interactive explanation architecture was presented that allows for interactive verification and ad-hoc personalization of the explanations.

Further works on exploiting explanation mechanism as a part of ML/DM workflow include several papers. In [3] authors introduce ExplainExplore system which is an interactive explanation system to explore explanations that fit the subjective preference of data scientists. It leverages the domain knowledge of the data scientist to find optimal parameter settings and instance perturbations, and enable the discussion of the model and its explanation with domain experts. However it does not operationalize it into fully automated system, still relying in core aspect on human-in-the-loop component. Another example of auditing framework was presented in [12]. The framework is intended to contribute to closing the accountability gap in the development and deployment of large-scale artificial intelligence systems by including explainability into the process of auditing AI systems. Yet, it is more of a methodological approach rather than an automated system. Approach described in [7] shows a method for combining many local high quality explanations into one, which makes it similar to the work presented in our work, however the authors method usage is limited to tree-based models only. More generic approach was presented in [19], where authors present a Python toolbox that provides functionality for inspecting fairness, accountability and transparency of all aspects of the machine learning process: data (and their features), models and predictions. However, this framework is more focused on inspecting datasets rather than explanations generated for models trained with the dataset. Furthermore, similarly to other frameworks, it does not support combining explanations of arbitrary XAI algorithm into one explanation, nor compare them as long as they do not provide the same explanation format.

Taking into consideration the full landscape of the aforementioned methods and frameworks and their limitations, the motivation for our work arised. We aimed at filling in the gap between XAI methods and ML/DM pipeline by providing a framework that will allow not only to quantify explanations, but also use this measures to combine explanations into better ones, or to allow for automatic selection of the best model-explainer pair. More specifically, our goal was to introduce cross-platform solution that is independent of the XAI algorithm,

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_4

under the assumption that it provides measure of importance of feastures in the decision process. The following sections provides more detailed description of our framework.

3 Ensemble explanation framework

In this work we focus on three metrics delivered by the *InXAI* framework[3] developed by us. It can be used along with explanation frameworks either to choose best explanation mechanism that fits project requirements (high stability, high consistency, etc.), or to generate unified explanations according to specified objective metric. Although the description of the framework is out of the scope of this paper, it is worth mentioning the it follow the scikit-learn[4] interface, which allows the XAI methods to be included in the ML/DM pipeline not only in theoretical, but also practical way.

3.1 Metrics of explainability

In this paper we focus for simplicity only on three metrics of explainability implemented in the *InXAI* framework: consistency, stability and area under the loss curve.

For the sake of further discussion we assume following notation. The importance of feature i and instance j delivered by explanation model e for machine learning model m will be denoted as $\Phi_{i,j}^{e\rightarrow m}$. If we skip subscripts, we assume marginal value over missed subscripts. Therefore, a complete explanation matrix for every feature and every instance generated by explanation e for model m will be denoted as $\Phi^{e\rightarrow m}$.

Consistency. Consistency measures how explanations generated for predictions of different ML models are similar to each other. Therefore, it is more related to stability of ML models with respect to decision making rather than to explanation mechanisms directly. Assuming that $M(X)$ is a set of ML models with high accuracy, Eq. (1) depicts the consistency measure:

$$C(\Phi^{e\rightarrow m_1}, \Phi^{e\rightarrow m_2}, \ldots, \Phi^{e\rightarrow m_n}) = \frac{1}{\max\limits_{a,b\in m_1,m_2,\ldots,m_n} ||\Phi_j^{e\rightarrow m_a} - \Phi_j^{e\rightarrow m_b}||_2 + 1} \quad (1)$$

Stability Stability (or robustness) assures generation of similar explanations for similar input. To obtain a numerical value to this property, modified notion of Lipschitz continuity has been proposed in [2]:

$$\hat{L}(\Phi^{e\rightarrow m}, X) = \max_{x_j\in N_\epsilon(x_i)} \frac{||x_i - x_j||_2}{||\Phi_i^{e\rightarrow m} - \Phi_j^{e\rightarrow m}||_2 + 1} \quad (2)$$

[3] See: https://github.com/sbobek/inxai.
[4] See: https://scikit-learn.org.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_4

where $N_\epsilon(x_i)$ is a set such as:

$$N_\epsilon(x_i) = \{x_j \in X | ||x_i - x_j|| < \epsilon\} \tag{3}$$

This optimization problem finds parameter describing most differing explanations *f(x)* for points in a vicinity of x_i, dictated by the set $N_\epsilon(x_i)$ proportional to the distance between the neighbours.

Area under the loss curve Area under the loss curve (AUCx) depicts the loss in accuracy (or other selected metric) when features are perturbed gradually according to their inverse importance returned by explanation algorithm.

Therefore, if the AUCx is high, it may imply that the importance of the features was set incorrectly, as perturbation caused large loss in accuracy. The loss in accuracy is defined as a difference in baseline accuracy obtained by a non-perturbed dataset and the accuracy obtained from perturbed dataset. In our work we used the trapezoidal rule to calculate it over set of accuracy losses for different perturbation rates.

It is worth noting that current version of the framework does only cover XAI algorithms that provide explanations in a form of feature importance assigned to particular features. This makes it more difficult to apply the framework to rule-based systems that does not provide such information out of the box. Currently we assume binary feature importance for rule-based explainer, meaning that features that were used for explanation have importance 1, while others features importance are 0. However, this feature is not yest provided out of the box, and needs to be programmed by the user.

3.2 Ensemble score

Calculating ensemble score is not limited to the metrics defined above and can be easily extended and modified as we will show in Section 4. The main goal of ES score is to capture the weighted importance of different metrics into one value. The definition of ES for a set of metrics M and weights w, was given in Eq. (4).

$$ES(M, w) = \sum w_i \cdot M_i \tag{4}$$

Having the ensemble score, we calculate a new, combined vector of explanation Φ^{ens} as a weighted sum of ensemble scores and associated with them original explanations $\Phi^{e_1}, \Phi^{e_2}, \ldots \Phi^{e_n}$. The weights are assigned arbitrary depending on the desired influence of a particular metric to the ensemble explanation.

Therefore, the final ensemble explanation is given by the Eq. (5).

$$\Phi^{ens} = \frac{ES(M, w) \cdot [\gamma_1 \Phi^{e_1}, \gamma_2 \Phi^{e_2}, \ldots, \gamma_n \Phi^{e_n}]}{\sum_{i=1}^{n} ES_i(M, w)} \tag{5}$$

Where $\gamma_1, \gamma_2, \ldots, \gamma_n$ are scaling factors that make it possible to compare and combine explanations obtained from different XAI frameworks. Note that

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_4

classic per-column or per-sample scaling will corrupt the internal dependencies between importance and features. Therefore the scaling is performed over the whole matrix of explanation generated by the same model. In our approach we used min-max normalization that is given in eq. (6).

$$\Phi' = \frac{\Phi - \min(\Phi)}{\max(\Phi) - \min(\Phi)} \tag{6}$$

The following plots and results were generated for the dataset presented in Fig. 1 along with two ML models and their decision boundaries used for calculating consistency measures.

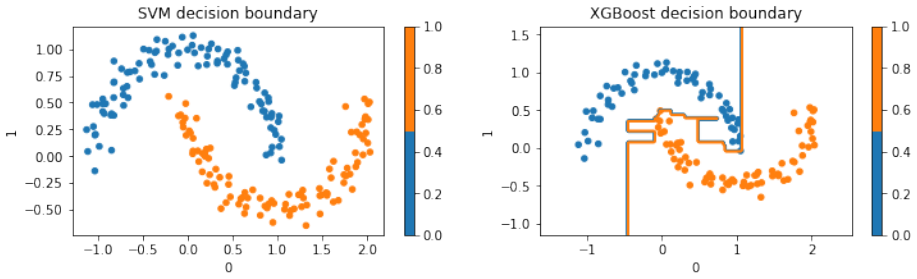

Fig. 1. Dataset and two classifiers with their decision boundaries used with InXAI framework for ensemble explanation generation.

The *ES* can also serve as a confidence measure of the explanation for single instance, or explanation framework with respect to selected metrics. Based on this confidence the ensemble explanations are created. Fig. 2 presents *ES* for LIME and SHAP. The transparency of a data point depicts the uncertainty of the explanation.

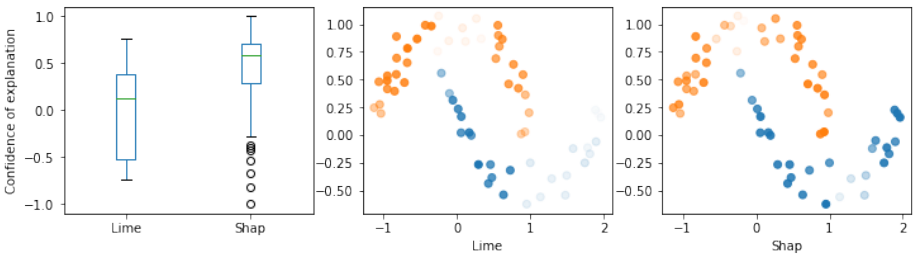

Fig. 2. Confidence of explanations where the maximum weight was put on the consistency metric.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_4

The metrics for an ensemble explainer built with Eq. (5) and its two components (LIME and SHAP) were given in Fig. 3. It is worth noting that the high value of weight for consistency, improves the consistency measure (upper row) in the ensemble. The same can be observed with stability (middle row). However, comparing to SHAP, only some of explanations were improved in terms of stability, while the overall (global) measure remained intact, or even worsen (see Sec. 5 for details of this phenomenon). With the maximum weight set on AUCx metric, we observe the overall reduction in the curve area. However, due to the fact that both explainers are correct with respect to this measure, the difference is not that apparent.

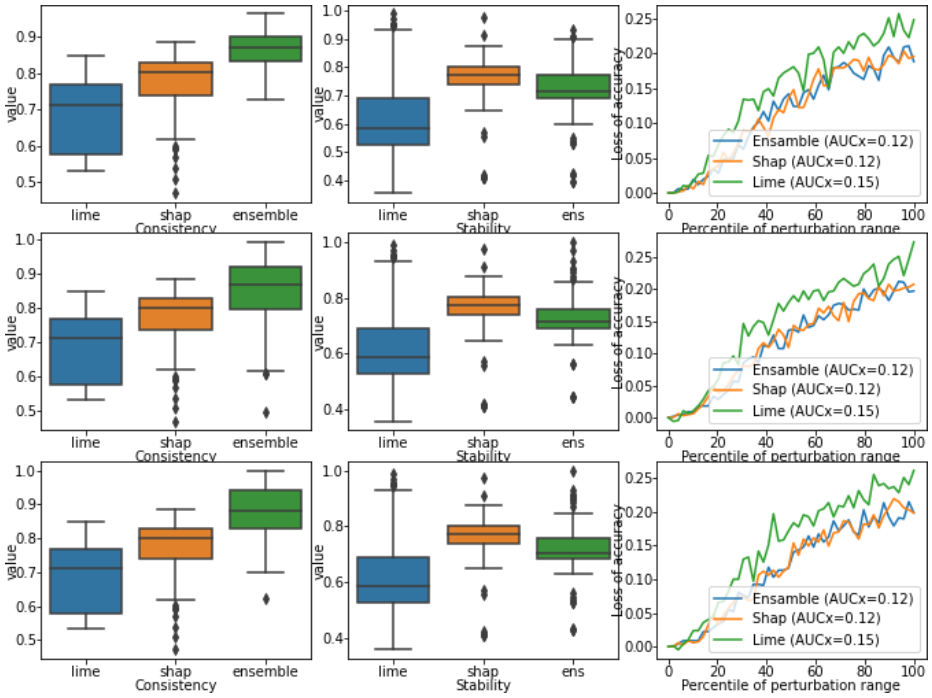

Fig. 3. Metrics for the Ensemble explanations generated for sample dataset. The upper row presents results for weights $[0.8, 0.1, 0.1]$ assigned to consistency, stability and AUCx metrics. Middle row gives results for weights $[0.1, 0.8, 0.1]$. Bottom row presents results for weights $[0.1, 0.1, 0.8]$. Single point represents metric value of single datapoint (local), while the global measure can be considered a spread of local values.

In the next section we demonstrate the solution on the example of a system for emotion classification based on face expressions photographs.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_4

4 Image classification use case scenario

A strong advantage of the Ensemble Score defined in Eq. 4 is that, since it is problem-agnostic, it can accommodate any metric. Similarly, the formula given by Eq. 5 does not depend on model-specific assumptions. It provides a flexible foundation for building ensemble explanations involving a variety of methods. This section shows how this generic framework can be utilized to generate ensemble explanations for the problem of image classification.

Classification of affective images relies on recognizing facial features in photographs of human subjects. Model-agnostic methods such as SHAP, LIME or ANCHOR are all capable of explaining predictions related to such images, however, due to differences in how they operate, their outcomes are not easily comparable. We demonstrate that it is possible to find analogies between results produced by these methods and to integrate them into an ensemble.

A common characteristic of SHAP, LIME or ANCHOR is that all these methods allow identifying subareas of the image that contribute to a given prediction. To reduce dimensionality of the feature space to a computationally feasible level, images must be preprocessed by a segmentation algorithm of choice. Effectively, feature importance $\Phi_{i,j}^{e\to m}$ determines how individual superpixels in the segmented image contribute to model prediction. $\Phi_j^{e\to m}$ denotes feature importance vector for explanation e, which consists of as many elements as there are superpixels for instance j. In this work the SLIC segmentation algorithm was used as it performed well on the task of isolating facial features in images in our dataset.

4.1 Metric definition for images

Since Eq. 4 allows arbitrary definition of metric set M we introduce an alternative formulation of stability and consistency for the purpose of this example.

Firstly, we define an auxiliary metric called similarity in Eq. 7, which evaluates how close two explanations are to each other:

$$\mathrm{Sim}(\Phi_i^{e\to m}, \Phi_j^{e\to m}) = 1 - \frac{||\Phi_i^{e\to m} - \Phi_j^{e\to m}||_1}{S} \tag{7}$$

where S specifies the number of features (image segments). It is assumed that explanations are mapped to range $[0, 1]$ so as to guarantee similarity to be a normalized value between 0 and 1. Value of 1 identifies two explanations as identical, whereas 0 corresponds to no similarity.

Alternative definition of consistency is given by Eq. 8. In this formulation consistency is to be interpreted as average similarity between assessed explanation $\Phi^{e\to m_0}$ and a set of reference explanations $\Phi^{e\to m_1}$ through $\Phi^{e\to m_n}$ obtained for independently trained ML models.

$$C(\Phi^{e\to m_0}, \Phi^{e\to m_1}, \Phi^{e\to m_2}, \ldots, \Phi^{e\to m_n}) = \frac{\sum_{k=1}^{n} \mathrm{Sim}(\Phi_j^{e\to m_0}, \Phi_j^{e\to m_k})}{n} \tag{8}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_4

Stability measure can also be defined in terms of similarity between the assessed explanation $\Phi_i^{e \to m}$ and explanations obtained for perturbed instances in the vicinity of x_i, as given by Eq. 9.

$$\hat{L}(\Phi^{e \to m}, X) = \sum_{x_j \in N_\epsilon(x_i)} \frac{\mathrm{Sim}(\Phi_i^{e \to m}, \Phi_j^{e \to m})}{|N_\epsilon(x_i)|} \tag{9}$$

Consistency and stability metrics proposed in this section are fully compatible with Eq. 4. In this example the Ensemble Score was calculated with equal weights for both consistency and stability.

4.2 Model assumptions

The image classifier was built on top of a neural embedding network. The core of the network was based on Inception Resnet V1 model that was adjusted and fine-tuned for facial expression classification task. Training and computation of explanations was performed on RGB images of size 160x160 pixels. Images were sourced from a dataset [20] where no explicit label information was provided for individual instances, because the dataset was designed for triplet learning. Therefore, the classifier was built on top of labels generated artificially according to the following procedure instead:

1. Compute embeddings $e_1, \ldots, e_n$ for all n instances in the training set.
2. Perform k-means clustering with the number of clusters selected by optimizing silhouette score.
3. Build a K-NN classifier that maps any given embedding to index of the cluster that it fits best. For sake of this research $K = 200$ was used.

Summarizing, cluster indices were used directly as labels for the purpose of classifying unknown instances.

Note that from the perspective of the ensemble framework demonstrated further in this work, implementation details of the image classifier are not critical. However, defining a classifier was necessary to produce explanations with the underlying methods: SHAP, LIME and ANCHOR. The ensemble approach by itself can merge any set of explanations, if only they quantify importance of each feature of every explained instance.

In order to enable calculating consistency multiple models were trained, each with a different embedding space size and hyperparameter configuration. Due to these configuration differences, independent classifiers had to be built separately for each model. To ensure a fair comparison, cluster number optimization (according to silhouette score) was conducted only once for a specific reference model, and assumed equal for all other models. As a result, count of different class labels the same across all explanation models. To compute stability we sampled such instances from neighborhood $N_\epsilon(x_i)$, for which it was known that the explanation should remain unaltered.

In the next step a selection of instances was chosen from the validation dataset [20] and fed to SHAP, LIME or ANCHOR frameworks independently.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_4

Each framework-specific explanation was evaluated in terms of stability, consistency and, most importantly, Ensemble Score.

4.3 Explanation scaling and aggregation

Here we demonstrate how independent explanations obtained from SHAP, LIME and ANCHOR can be combined into one ensemble explanation.

SHAP and LIME feature importance attributed to each superpixel is a real number. On the other hand, ANCHOR determines which image segments have crucial contribution towards a specific model prediction; that is, were they not part of the image, the prediction would have been different. ANCHOR explanations can be seen as a vector of binary values, where 1 is assigned to the crucial features and 0 corresponds to features that have no impact on prediction.

To enable aggregation of results originating from different XAI frameworks, explanations need to be scaled. Recommended choice of scaling factor for explanations generated by k-th model is given by Eq. 10.

$$\gamma_k = \frac{1}{||\Phi^{e \to k}||_{max}} \tag{10}$$

As a consequence, it is guaranteed that – independently of the XAI framework – scaled feature importance values fit in a normalized range $[-1, +1]$ and that feature with the strongest contribution is assigned importance of ± 1. Sign depends on whether the maximum contribution is positive or negative.

Inserting the original explanations, their Ensemble Scores and scaling coefficients γ into Eq. 5 yields the final ensemble explanation.

4.4 Results

We present example results obtained according to the framework described in this work. Fig. 4 provides ensemble explanations for an array of facial images picked randomly from the validation dataset [20]. Visualization is based on a color mapping, where color intensity corresponds to importance of specific image features. Green-colored areas have a positive contribution towards particular explained label, and red overlay signifies negative contribution.

Fig. 4. Example ensemble explanations for facial expression images

Resulting ensemble explanations are visually appealing combinations of several underlying XAI approaches. Note that positive contributions are generally

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_4

dominating, which is partly due to the fact that ANCHOR never attributes negative contribution to features.

Ability to utilize strengths of multiple explanation models simultaneously is the primary advantage of the presented approach. We found that, in multiple cases, facial features captured by the ensemble explanations remained in stronger agreement with human intuition than in underlying explanation methods assessed individually. Probably this is because feature importance in ensemble explanations was derived as a weighted average of respective feature importance values in the underlying methods according to Eq. 5. Therefore potential errors in importance values might be canceled out by aggregating multiple partial solutions. The risk that the ensemble overestimates or underestimates contribution of a feature is lower than in case of any individual underlying explanation model.

On the other hand, values of stability, consistency and Ensemble Score obtained in our study were characterized with relatively low variance across example images considered in this research. A presumable root cause is that image areas where feature importance was close to zero (neutral impact on prediction) were usually much larger than areas where feature contribution was strongest. It is a desired characteristic because it makes explanations more specific and understandable, i.e. focused on critical features. However, the downside was that the similarity measure between two explanations was influenced predominantly by neutral areas, resulting in low variance of the similarity measure, on which stability, consistency and ensemble score are built. To alleviate this issue, alternative, more sensitive formulations of the similarity measure are also possible and can easily be used as a drop-in replacement in Eq. 8 and Eq. 9.

Another point is that quantifying stability, consistency and Ensemble Score allowed objective validation of the explanation model used in our research. Since each metric takes into account a different set of factors, using a combination of such metrics made the validation process more robust. For example, note that a faulty model might also yield high-stability explanations, although it is less likely to produce high-consistency results.

In conclusion, it was shown that the approach introduced in this work successfully unifies various XAI frameworks that were not initially designed with compatibility on mind. It is also inherently possible to extend this approach on methods other than SHAP, LIME or ANCHOR. The proposed framework has high potential to be used in automated assessment and comparison of different explanation models according to a set of well-defined objective measures. However, to fully utilize the automation potential, there is a need for comprehensive tests to confirm that ensemble explanation visualizations are consistent with human perception.

5 Summary and Future Works

The methods of Explainable Artificial Intelligence methods form a large portfolio of versatile frameworks and algorithms providing insights into the decision process of an AI system. However, their underlying mechanisms may be very

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_4

different. Thus it may result in very different explanations for the same tasks. In this paper, we presented an original approach that aims at combining several XAI algorithms into one ensemble explanation mechanism via quantitative, automated evaluation framework. We focused on model-agnostic explainers to provide most robustness. We provided an illustrative demonstration of our approach on image classification task.

Weights such as stability works in most of the cases only locally and can be used to weight single instance explanation. This means that combining several explanations with high stability does not assure the resulting ensemble will also have the high stability, as the neighbourhood of explanations was altered. We plan to use SMAC [5] or similar Bayesian optimizer to optimize explanations with respect to the selected metric in a way that they will be optimized globally.

We also plan to conduct observational studies with domain experts and real-life use-cases to validate the feasibility of our solution. Finally we will be evaluating this approach on different datasets, including industrial ones.

Acknowledgements

The paper is funded from the XPM project funded by the National Science Centre, Poland under CHIST-ERA programme (NCN UMO-2020/02/Y/ST6/00070).

References

1. A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.
2. D. Alvarez-Melis and T. S. Jaakkola. On the robustness of interpretability methods, 2018.
3. D. Collaris and J. J. van Wijk. Explainexplore: Visual exploration of machine learning explanations. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, pages 26–35, 2020.
4. B. Goodman and S. Flaxman. European union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*, 2016.
5. F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration (extended version). Technical Report TR-2010-10, University of British Columbia, Department of Computer Science, 2010. Available online: http://www.cs.ubc.ca/˜hutter/papers/10-TR-SMAC.pdf.
6. N. Liu, D. Shin, and X. Hu. Contextual outlier interpretation. *arXiv preprint arXiv:1711.10589*, 2017.
7. S. M. Lundberg, G. G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. Lee. Explainable AI for trees: From local explanations to global understanding. *CoRR*, abs/1905.04610, 2019.
8. S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777. Curran Associates Inc., 2017.
9. S. Mohseni, N. Zarei, and E. D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems, 2020.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_4

10. F. Mujkanovic, V. Doskoč, M. Schirneck, P. Schäfer, and T. Friedrich. timexplain – a framework for explaining the predictions of time series classifiers, 2020.
11. P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann. Explainability methods for graph convolutional neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10764–10773, 2019.
12. I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *CoRR*, abs/2001.00973, 2020.
13. M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.
14. M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Publications, Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
15. M. Robnik-Šikonja and M. Bohanec. Perturbation-based explanations of prediction models. In J. Zhou and F. Chen, editors, *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, pages 159–175. Springer, 2018.
16. R. C. Schank. Explanation: A first pass. In J. L. Kolodner and C. K. Riesbeck, editors, *Experience, Memory, and Reasoning*, pages 139–165, Hillsdale, NJ, 1986. Lawrence Erlbaum Associates.
17. K. Sokol and P. Flach. One explanation does not fit all. *KI - Künstliche Intelligenz*, 34(2):235–250, Feb 2020.
18. K. Sokol and P. A. Flach. Explainability fact sheets: A framework for systematic assessment of explainable approaches. *CoRR*, abs/1912.05100, 2019.
19. K. Sokol, R. Santos-Rodríguez, and P. A. Flach. FAT forensics: A python toolbox for algorithmic fairness, accountability and transparency. *CoRR*, abs/1909.05167, 2019.
20. R. Vemulapalli and A. Agarwala. A compact embedding for facial expression similarity. *CoRR*, abs/1811.11283, 2018.
21. M. Verma and D. Ganguly. Lirme: Locally interpretable ranking model explanation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1281–1284, New York, NY, USA, 2019. Association for Computing Machinery.
22. C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, and P. Ravikumar. On the (in)fidelity and sensitivity for explanations, 2019.
23. Z. Zhang, F. Yang, H. Wang, and X. Hu. Contextual local explanation for black box classifiers. *CoRR*, abs/1910.00768, 2019.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_4