Comparative Analysis of Explainability Techniques for Non-linear Models

Alessio Conti

Ethics in Artificial Intelligence 2023-2024

Abstract. The increasing complexity of machine learning models, particularly non-linear models, has increased the need for robust explicability techniques. While these models are powerful at making accurate predictions, they often operate as 'black boxes'. This opacity can interfere with the adoption of machine learning in critical domains such as healthcare, finance, and industrial applications, where understanding the rationale behind model predictions is paramount. As a result, various explicability techniques have been developed to shed light on these complex models. This report aims to provide a comparative analysis of the two most famous explainability techniques: LIME [4] and SHAP [1]. At the very end, PSyKE [5] will be taken into consideration to explore the surface of symbolic knowledge extraction approaches. These methods are evaluated across different types of machine learning algorithms and data modalities, including both images and tabular data, this study seeks to determine their effectiveness in providing comprehensive, consistent and reliable explanations.

1 Introduction

During the last few years, there has been a growing tendency in the use of Artificial Intelligence in different sectors, from healthcare to criminal justice and finance. This has a substantial impact on human lives, especially when used to make high-stakes predictions in the aforementioned sectors.
A considerable proportion of machine learning models lack transparency, especially the ones with billions of parameters used nowadays, failing to provide a clear and readily comprehensible explanation of their predictions.
Model explainability enhances transparency, builds trust with end-users, facilitates regulatory compliance, and aids in diagnosing and mitigating model biases. Without explainability, stakeholders may be reluctant to rely on model outputs, regardless of their predictive accuracy. The research inspects the major key players in black-box model explainability, trying to understand the cases when those methodologies are useful for explanations or just tools for superficial inspections.

2 Explainability techniques for tabular datasets

2.1 Dataset - Obesity Level

The dataset of choice [2] includes data for the estimation of obesity levels in individuals from the countries of South America (Mexico, Peru and Colombia) based on their eating

habits and physical condition. The data contains 17 attributes and 2111 records.
All the records are labelled with the class variable Obesity Level, allowing the classification of data as: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. All the other 16 features are the following:

- Gender: "Gender"
- Age: "Age"
- Height: Feature, Continuous
- Weight: Feature Continuous
- family_history_with_overweight: "Has a family member suffered or suffers from overweight?"
- FAVC: "Do you eat high-caloric food frequently?"
- FCVC: "Do you usually eat vegetables in your meals?"
- NCP: "How many main meals do you have daily?"
- CAEC: "Do you eat any food between meals?"
- SMOKE: "Do you smoke?"
- CH2O: "How much water do you drink daily?"
- SCC: "Do you monitor the calories you eat daily?"
- FAF: "How often do you have physical activity?"
- TUE: "How much time do you use technological devices such as cell phones, video games, television, computers and others?"
- CALC: "How often do you drink alcohol?"
- MTRANS:"Which transportation do you usually use?"

The preliminary inspection of the dataset assesses that all the classes are equally represented with not less than 250 instances each. Some features have great disproportion within the values, for example: SCC (calories intake daily monitor) is almost totally toward "no", and SMOKE is very disproportioned toward "yes". This feature would have been dropped in a normal machine learning scenario because being all toward one value lacks in useful information, in this situation instead is interesting to keep to see if the model (or the explainer) somehow relies on them to make predictions.

As pre-processing steps, all the continuous features have been standardized, while all categorical ones have been encoded.

2.2 Model Training

For the task two models have been trained: a self decision tree classifier, which is already a self-explanatory model, and a neural network. Both of them got the same accuracy of around 90% in the test set.

Decision tree classifier For what concerns the decision tree classifier an initial grid search has been done to find the best depth value, which has been set to 12. This resembled finding an accuracy of 92%.

Neural Network classifier The neural network is a standard network that performs classification: three dense layers with ReLu as activation function are alternated by three dropout layers. At the end, a dense layer with a Softmax activation function is used to map the different classes.
The accuracy of this model sat at 90% in the test set, being comparable to the decision tree.

2.3 Model Explanations

Straightforward Decision Tree Explanation Decision trees are inherently interpretable due to their straightforward structure and transparent partitions of data into subsets based on the value of input features.
The tree structure is easy to visualize and understand. Each path from the root to a leaf represents a sequence of decisions leading to a final prediction.
The tree structure can easily increase in size because of the depth of the tree. Inspecting the tree shows that the rules are based on the most important features that the model has identified (Weight, Height, Age and Gender). This is also clear by inspecting the feature importance plot in Figure 1.

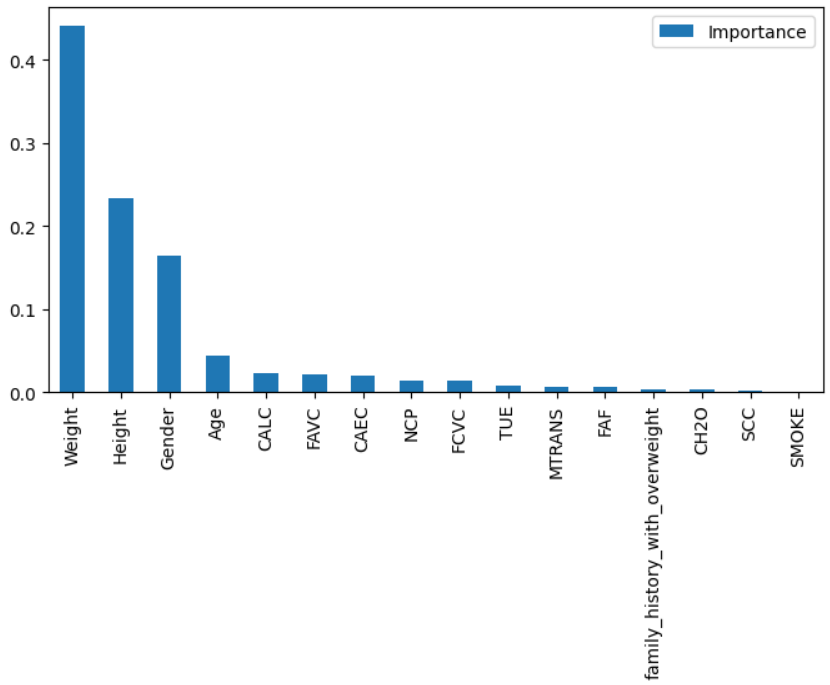


Fig. 1: Feature importance plot extracted from a decision tree classifier.

Looking at some instances it's straightforward to understand the reasons that led the model to come in one conclusion instead of another. Both are correctly predicted

Predicted class: 3 (Obesity_Type_II) - True class: 3 (Obesity_Type_II)

	Feature	Threshold	Decision
0	Weight	0.493202	Instance > 0.49320176243782043
1	Gender	0.5	Instance > 0.5
2	Age	-0.302418	Instance > -0.30241774022579193
3	Weight	0.863534	Instance > 0.8635341227054596
4	Leaf	Leaf	Leaf Node

(a) Obesity Type II

Predicted class: 1 (Normal_Weight) - True class: 1 (Normal_Weight)

	Feature	Threshold	Decision
0	Weight	0.493202	Instance <= 0.49320176243782043
1	Weight	-1.029966	Instance <= -1.0299664735794067
2	Height	-0.454428	Instance <= -0.45442794263362885
3	Weight	-1.540055	Instance > -1.540054976940155
4	Height	-1.97318	Instance > -1.9731800556182861
5	NCP	0.784498	Instance <= 0.7844981700181961
6	Age	-1.143028	Instance > -1.1430277228355408
7	SMOKE	0.5	Instance <= 0.5
8	Leaf	Leaf	Leaf Node

(b) Normal Weight

Fig. 2: Explanations of two predictions from a decision tree classifiers.

instances and we can see from Figure 2 that different features are used for identifying the different instances. During training the model has learnt how to divide the hyperspace on which the datapoints are placed. Some boundaries may be smooth and clearly defined, while others could be fractal decision boundaries. In Figure 3 an example of decision

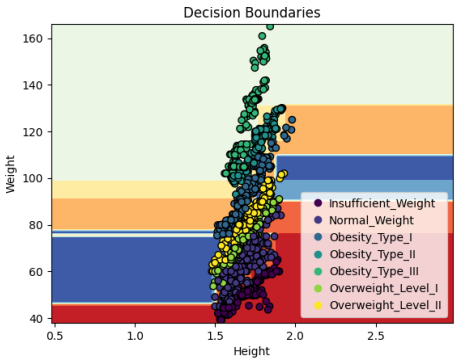


Fig. 3: Decision boundaries for *Weight* and *Height*.

boundary of the main feature *Weight* with respect to *Height*. From this we can understand that the model relies a lot on the *Weight* feature to guide the prediction for two main reasons (I) the decision boundaries are marked discerning one class from the others, (II) for all the classes there is a linear correlation between *Weight* and *Height* of the observed data points, which leads to a reasonable stratification of classes

Local feature explanation with LIME and SHAP In this phase the two explainability methods LIME and SHAP are used to get insights on single instance explanations.

Those methods approach the problem differently:
LIME modifies a single data sample by tweaking the feature values and observes the resulting impact on the output, this perturbations aims to understand the hyperspace configurations in the near surroundings of the point. The perturbed points are then used to train a local surrogate model inherently interpretable.
SHAP instead is based on the approximate computation of Shapely values. A metric used in game theory to weigh the different contributions of each player.
In the ML scenario, each feature is a player and the values are computed by averaging the marginal contributions of each of them across all possible permutations. This involves assessing every possible combination of features and determining the impact that each has on the model's prediction.
To investigate the difference between LIME and SHAP explanations, consider two distinct instances from the classification problem. Given a first instance where both models predict correctly, can be seen that the two methods outputs similar explanations. In the following Table 1 are summarized the most influential features for the prediction.

LIME		**SHAP**	
Tree	***NN***	***Tree***	***NN***
Weight	Weight	Weight	Weight
Gender	Age	Gender	Age
NCP	Gender	Age	Gender
Age	FCVC	NCP	Height
FCVC	NCP	FCVC	MTRANS

Table 1: Most influential features extracted by local predictions of LIME and SHAP for test datum n.13.

Can be seen that both LIME and SHAP are in accordance with each other on the top features for explaining the same instance with the two different models. This high correspondence, compared also to the ranking with the best decision tree's features [Image1], reassures on the correctness of model interpretation. When looking at another trickier instance instead Table 2, where the two models disagree, can be seen that the concordance rate decreases.

LIME		**SHAP**	
Tree	***NN***	***Tree***	***NN***
Weight	Weight	Weight	Weight
Height	Height	Height	FamilyHistory
CALC	NCP	NCP	Gender
TUE	Gender	Age	NCP
FCVC	Age	FAF	Height

Table 2: Most influential features extracted by local predictions of LIME and SHAP for test datum n.24.

One feature *Weight* is always present, while the others change even repeating different explanations of the same instance. There are few, besides from *Weight* however that appear more frequently than others: *Height, Age, Gender and NCP*.

To better capture the agreements of the two models some broad tests have been made. Two metrics are taken into account to measure the global model accordance on different instances: Intersection over Union (IoU) and concordance rate. The experiment was run on 100 random samples of the test set. For each instance explanation the list of the top five (5) relevant features for the model explanations are saved and then compared. For both methods (LIME and SHAP) the obtained results are similar, see Table 3. The

	LIME		**SHAP**	
	IoU	***Corr. Rate***	***IoU***	***Corr. Rate***
mean	0.40	0.56	0.43	0.58
25%	0.25	0.40	0.25	0.40
50%	0.43	0.60	0.43	0.60
75%	0.43	0.60	0.49	0.65

Table 3: Metrics for LIME and SHAP IoU and concordance rate.

values are below 0.5 for IoU: setting around 0.4 for the mean, with the 75% quartile having a similar score to the 50% quartile. For what concerning concordance rate this has a mean slightly above 0.5, setting 50% and 75% quartile at a score of 0.6 underliyng that for half of the instances tested at least 3 out of the 5 most frequent features are in accordance between the two models. This data expresses information about the high variability present in the "most relevant" features used to drive the explanations. This observation needs a deeper understanding. Even if the two proposed models have similar accuracy does not mean they act the same: it is more likely that they build two very different hyperspaces. On the other hand LIME and SHAP try to infer this boundaries by looking at models behaviours. Another interesting interaction that those two methods

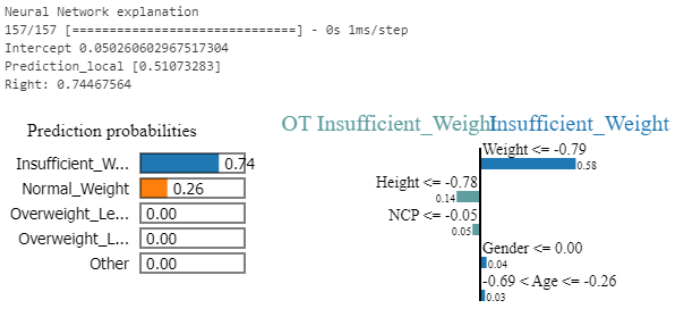


Fig. 4: LIME instance n.24 explanation for being misclassified.

offer while explaining single instances observation, is to explore the reasons why a model has chosen one prediction over the other possibilities, especially for those instances where has been mistakenly predicted Figures 4 5. In guiding this choice seems that

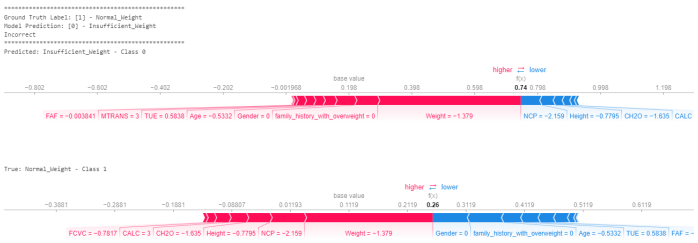


Fig. 5: SHAP force plot of instance n.24 being misclassified.

"Weight" has mainly guided the prediction toward *Insufficient Weight*, while *Height* tried to push it back. Unfortunately, the other features seem to be of less interest to the model and do not provide any real support to the decision. Surely having a feature (*Weight*) far more predominant than the others in guiding the prediction is a limitation of the dataset itself.

Global feature exlanation SHAP capabilities go beyond local instances interpretability and can be used also to explore the overall generalities of the model performance. Some

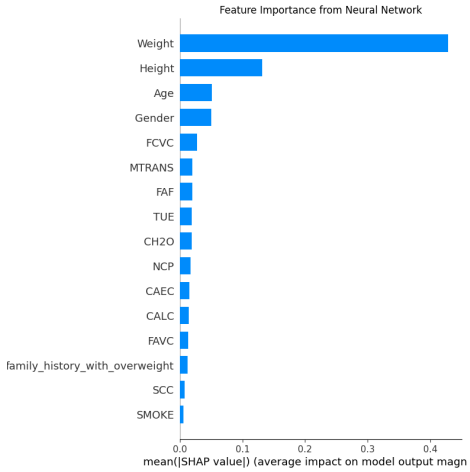
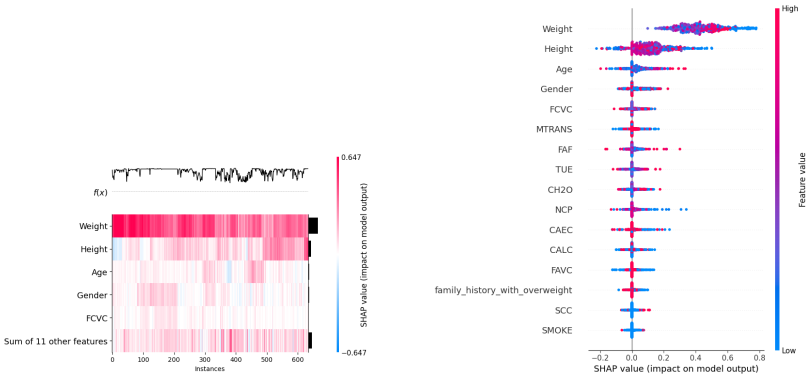


Fig. 6: SHAP feature importance ranking for NN model.

detailed information can be gathered by heatmap plot and summary plot 7. Those

images represent how much impact each feature has on the model output. Can be seen that for each instance "Weight" is the predominant feature, followed by "Height". This is in accordance with the decision tree straightforward explanation and with previous observations on local features. With some minor differences is others features' ranking Image 6.



(a) Heatmap (b) Summary plot

Fig. 7: Heatmap and summary plot of best features extracted by SHAP from the NN model.

Even if the top features concordance for single instance explanations is not always coherent, the overall feature importance ranking of the model is very similar to the one extracted from the decision tree standard model. This gives us some reassurance in the overall model understanding.

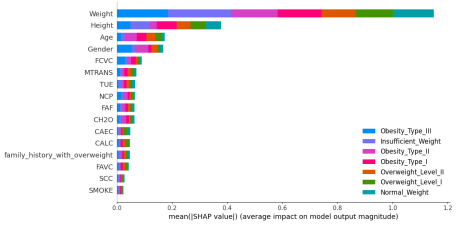


Fig. 8: SHAP feature importance ranking for NN model with classes attribution.

One interesting alternative to the plain feature importance histogram can be the one in Figure 8, showing the impact of each feature in predicting the different class labels. For example *Age* has much more impact in predicting *Obesity Type I/II* rather than *Normal Weight* or *Obesity Type III*.

3 Symbolic Knowledge Extraction with PSyKE

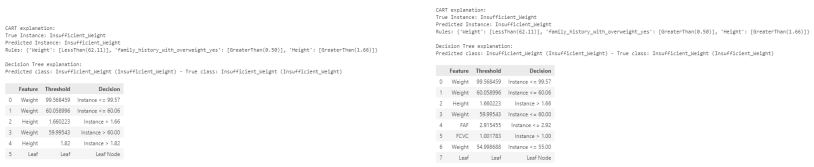
In this section, the aim is to scratch the surface of Symbolic Knowledge Extraction (SKE) and try to get some interesting explanations on tabular datasets with this approach.
With SKE we refer to the field that wants to explore different algorithms capable of extracting symbolic knowledge into logic rules from a model. Two options are available: (I) treating the model as a black box (as in this particular case) or (II) performing decomposition of the model itself. For the following experiments, the algorithm CART is used. CART algorithm applies to both classification and regression tasks and leads to explanations which are machine and human-interpretable. This method builds a surrogate model, mimicking the behaviour of the given classifier. The extracted rules can then be exploited to derive predictions or better understand the behaviour of the original classifier.

The model chosen as a black-box for this experiment is a decision tree classifier, this has been trained in two variations at different maximum depths to get different accuracy. CART is exploited in order to reach similar accuracy to the ones of the original classifiers. For this reason, different parameter configurations are explored. The following table summarises the accuracy reached in training and test set by the surrogate models (explainers) and the number of rules that have been extracted Table 4.

	Model accuracy	**Explainer acc. train**	**Explainer acc. test**	**n. rules**
I	0.94	0.91	0.88	50
II	0.63	0.61	0.61	20

Table 4: Explainer accuracy and the corresponding number of extracted rules.

To reach the same performance a decent amount of rules has been extracted, and those upper limits have been tested to prevent the model from overfitting. The explainer_1 is able to map the hyperspace similarly to the real model. By manually inspecting some of the instances can be seen that the extracted rules are very similar to the decision nodes of the actual model Figure 9.



(a) (b)

Fig. 9: CART explanations compare with Decision Tree explanation.

For what concerns explainer_2 instead, it reached the same accuracy as model_2 with just 20 rules. This indicates the good generalization abilities of CART. Here, for the instances that have been correctly predicted the explanations are similar to the one proposed by the model, while for the other cases, there is a greater degree of uncertainty. This is due to the fact that the decision boundaries has not been adequately defined, resulting in a sub-optimal division of the hyperspace. This led to the presence of numerous misclassified elements, as well as noise and confusion in the explanations. With respect to LIME and SHAP explaining through SKE brings more trustable and less uncertain results.

CART, and SKE in general, offer insights into how the model makes decisions across all data points, thereby providing a global understanding rather than investigating the surroundings of a single instance. However, it should be noted that the extracted rules may not perfectly capture the behaviour of the original model, leading to potential inaccuracies, particularly if the knowledge extraction comes from a very complex model. Consequently, finding solutions can be challenging and may lead to oversimplification. From my preliminary observations, SKE approaches appear to have considerable potential, particularly when a comprehensive understanding of the entire model is required. Moreover, the explanations are not merely simple representations; they are tangible, comprehensive rules that can be queried to retrieve predictions. Additionally, it is important to highlight that the output of this approach is highly comprehensible and intuitive, even to those without technical expertise.

4 Explainability Techniques in Image Classification

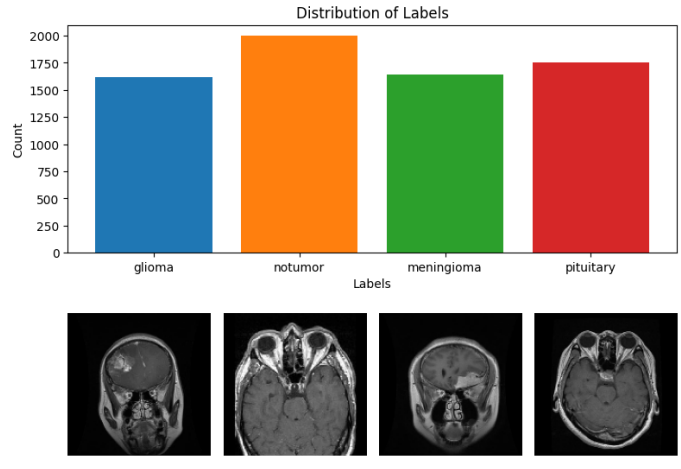


Fig. 10: MRI brain tumor class distibution.

Dataset - MRI Brain Tumor The dataset contains brain MRI images [3], which have undergone to some image processing techniques including: conversion to grayscale images, cropping, normalization and resizing to 256x256.
The dataset contains 7023 images divided into four (4) classes: *glioma, notumor, meningioma and pituitary* Figure 10.

The label's distribution is balanced throughout the dataset for all the label classes.

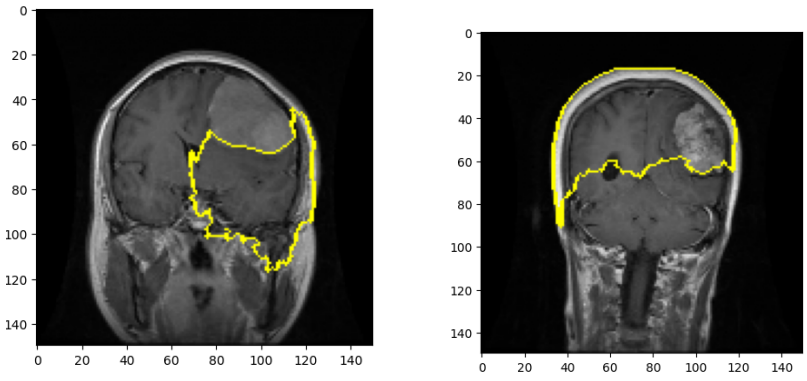
4.1 Model Training

The model of choice is EfficientNetB0, a model pre-trained on ImageNet without the top layers. This is chosen because of its modest size compared to others, being capable of capturing spatial hierarchies such as edges and textures during the first layers and refining the recognition of more complicated features deeper in the network. At the pre-trained model is attached a classification head consisting of two dense layers alternated by average pooling and a dropout layer. The final model has 5,365,415 total parameters, which admits training in a reasonable time leveraging a very high classification accuracy of almost 100%.

4.2 Model Explanation

To explain this dataset the experiments are made by manual observations of different images and trying to gain as much information as possible from the explainers.

LIME explainer LIME has a built-in explainer for images that works similarly to the one used for tabular data. After perturbing the input image, utilize a similarity kernel to measure distances between images. With this technique, LIME tries to identify "super-pixels" that are more likely to be used by the model to predict the outcome. This



(a) Misclassified meningioma (b) Broad meningioma detection

Fig. 11: Brain tumor position explanation with LIME.

approach may be successful for clear RGB and distinct images as shown in LIME paper [4], but this approach performed very poorly in this scenario. Figuring out some good parameters for getting some results was not trivial. By allowing the model the flexibility to take into account a small number of features (number of superpixels to include in the explanation), the results obtained were significantly below the expected ones, this can be attributed to the complexity of the image being explored. Conversely, when considering a greater number of features (500-1000), with a low min-weight (minimum weight of the superpixels to be included in the explanation), better results were obtained Figure 11. However the explainer still struggles in identifying clearly the tumor.

SHAP explainer Initial experiments with SHAP libraries went far better than the ones done with LIME. A more structured approach has been developed to go beyond just comparison. The steps followed are as follows:

1. Show the feature attributions for a subset of the test set
2. Review the explanations for both: the correct label and the other classes, to better understand the model choices
3. Repeat the steps with the same CNN, but significantly less accurate

To interpret the attributions: red pixels have to be considered positive features, increasing the probability of a class being predicted; blue pixels instead decrease the probability of a class being predicted.

The added value here, more than fact-checking if the super-pixels found are reasonable or not, is to compare for each instance the predicted class with the others. In doing so, it is possible to understand what elements the network actually relies on for its predictions. Take for example the first image on Figure12 presented: a *meningioma* is clearly visible in the middle top portion of the brain. The network correctly predicts it and the explanation points out exactly that spot in red. But this might not be enough to understand what's happening underneath the hood of the black-box. Inspecting also the other classes can be seen that *notumor* and *glioma* have a blue spot corresponding to the *meningioma* position, meaning in that area some features that decrease the probability of being in those classes are detected. This happens for almost all the instances presented, with different levels of noise present. This can address the conclusion that even if this approach works better than LIME, sometimes minor importance features or misleading super-pixels (also wrong ones) are considered important in sustaining the prediction choice.

The experiment then involved the training of the same model with lower accuracy. To do this, EfficientNetB0 was trained for only one epoch, resulting in (Un)EfficientNetB0 Figure 13, with an accuracy of 75%. Almost 25% different with respect to the first network. Considering the same inspected instances now, even if the model correctly predicts the examples, the explanations are much more confusing. No clear regions of interest are spotted and the marking areas are broad and vague. Most of the time the explanation is not even in accordance with the actual tumor position.

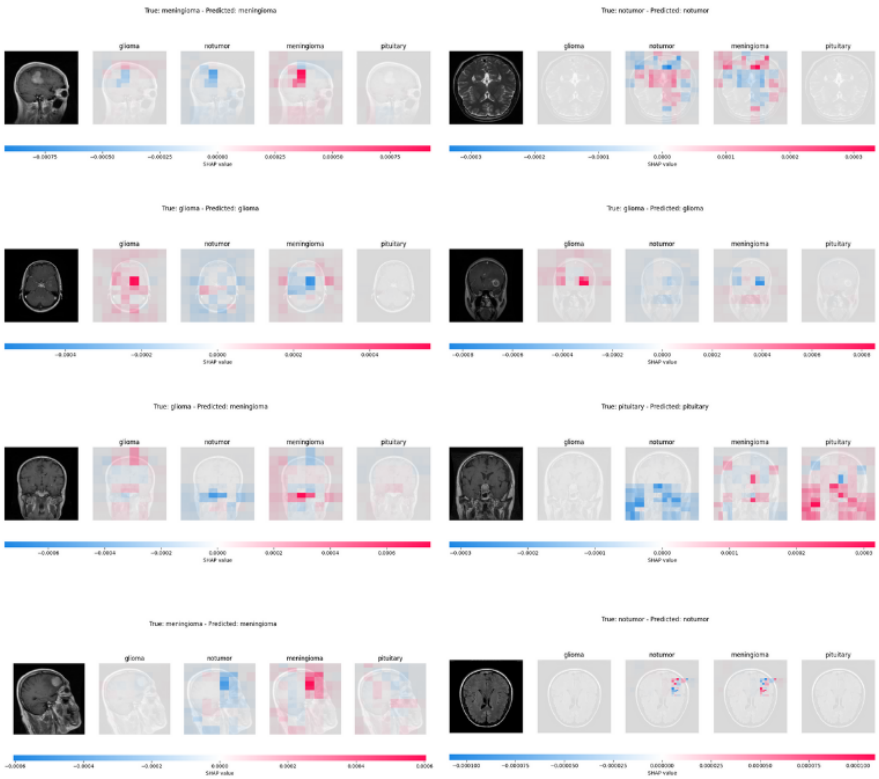


Fig. 12: EfficientNetB0 results with SHAP

5 Conclusion and future works

After inspecting those techniques and different models some concluding remarks. When considering a tabular dataset both methods perform accurately finding the best predominant feature, in this case *Weight*, while struggling to find univoque features supporting the predictions. Probably the feature importance disproportion (towards weight) is not helping the explainers to correctly evaluate the others, because predictions are mainly guided by that. I was expecting a major coherence in local feature extractions also for feature like *Height*, which plays a crucial role in the decision-making process. As seen both LIME and SHAP, even with slightly different outputs perform similarly and give insights on the local instance explanations. SHAP was tested to generalize the overall model expressivity, ranking the features by overall importance. This result is comparable and roughly in agreement with the output of the decision tree feature importance output. At least for the top elements.
PSyKE tested with CART algorithm outputs the most human-readable explanations, resembling the ones that can be derived directly from a decision tree path explanation.
Considering images: LIME was too general in finding super-pixels, while SHAP gave

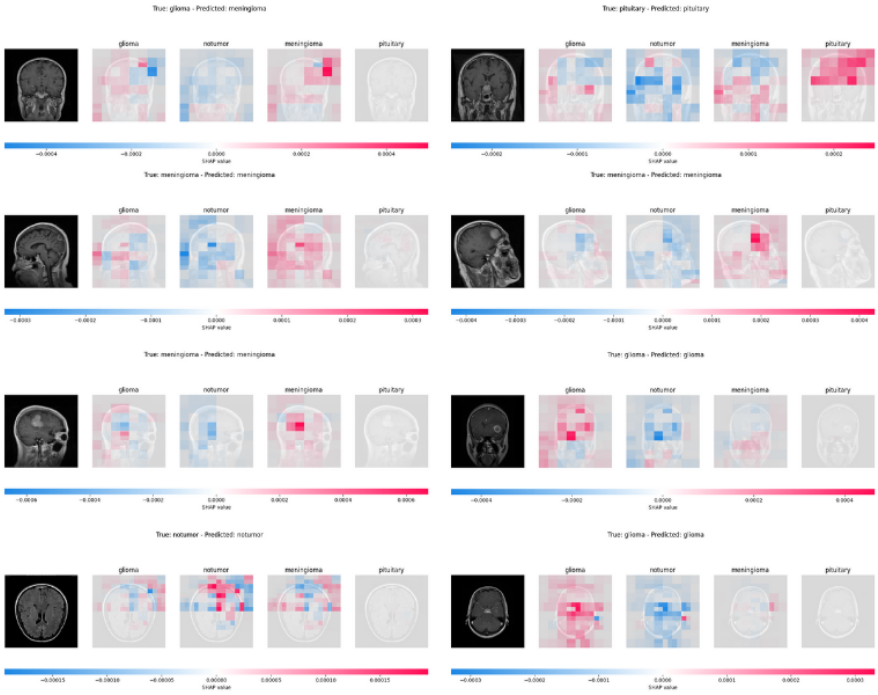


Fig. 13: (Un)EfficientNetB0 results with SHAP

interesting insights on areas that draw most attention to the prediction.

While using those methods it has to be kept in mind that those explainers are additional models, built upon the real ones. This leads to approximations of the actual model functionalities.
An explainer is as reliable as the underneath model is. Lowering the model accuracy reflects also the explainer's capabilities to achieve interesting insights from the instances. This is a consequence of increasing noise in between the features' boundaries, which are no longer sharp.

When trying to explain black-box models with methods like LIME and SHAP, one has to be sceptical about the obtained results; thus considering those answers as a more interpretable representation of the model behaviour instead of a real reliable explanation. Those are not to be used as an out-of-the-box tool. Explanations must be guided by critical thinking and a tendency to question the tool's actions and results. If real explanations are needed then the suggestion would always be to use models that are explainable by themselves. Nevertheless, the aforementioned tools can be beneficial in situations where the model is being debugged. They can be employed to inspect the approximated behaviour under certain limited circumstances, thriftily. As a matter of fact, I would not

recommend using this technique as oracle for explanations.

This is not the case with the extraction of symbolic knowledge. This method provides detailed rules, comparable to the output of a decision tree, which allows for greater accuracy and precision, increasing the confidence of the end-user. It would be of interest to explore the generalizable capabilities of SKE algorithms to exploit possibilities of teacher forcing techniques, where complex and obscure models are used as teachers, while the symbolic explanation can be used elsewhere as interpretable models.

References

1. Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.
2. Fatemeh Mehrparvar. Obesity levels dataset, Apr 2024.
3. Masoud Nickparvar. Brain tumor mri dataset, Sep 2021.
4. Marco Tulio Ribeiro, Sameer Singh, and Carlos 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, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.
5. Federico Sabbatini, Giovanni Ciatto, Roberta Calegari, and Andrea Omicini. On the design of PSyKE: A platform for symbolic knowledge extraction. In Roberta Calegari, Giovanni Ciatto, Enrico Denti, Andrea Omicini, and Giovanni Sartor, editors, *WOA 2021 – 22nd Workshop "From Objects to Agents"*, volume 2963 of *CEUR Workshop Proceedings*, pages 29–48. Sun SITE Central Europe, RWTH Aachen University, October 2021. 22nd Workshop "From Objects to Agents" (WOA 2021), Bologna, Italy, 1–3 September 2021. Proceedings.