Comparative Study between Global and Local Explainable Models

Sidi Mohamed DAFALI
Laboratory LIM
Department of Computer Science
Faculty of Sciences and Techniques
University Hassan II Casablanca
Mohammedia, Morocco
smdafali13@gmail.com

Mohamed KISSI
Laboratory LIM
Department of Computer Science
Faculty of Sciences and Techniques
University Hassan II Casablanca
Mohammedia, Morocco
mohamed.kissi@fstm.ac.ma

Omar EL BEGGAR
Laboratory LIM
Department of Computer Science
Faculty of Sciences and Techniques
University Hassan II Casablanca
Mohammedia, Morocco
omar.elbeggar@fstm.ac.ma

2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA) | 979-8-3503-0821-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/SITA60746.2023.10373599

Abstract—Despite opaque machine learning models outperform transparent models, users are still unable to comprehend and trust their outcomes.

Therefore, Explainable Artificial Intelligence (XAI) becomes a prominent research area. The main objective of this discipline is providing new techniques, tools and models that explain how opaque machine learning models operate to give predictions or simply offer some information about system decisions.

In this article, we present an experimental evaluation of explainable models used in XAI, namely: LIME, SHAP, ANCHOR and EBM. Those models are applied on the results of two opaque machine learning models: Random Forest and XGBoost. Our experimental evaluation covers various aspects for comparison including:

- Examining the explainability coverage of methods.
- Comparing explanations provided by each method.

Keywords—LIME, SHAP, ANCHOR, EBM, Local explainability, Global explainability.

I. INTRODUCTION

In recent decades, Artificial Intelligence (AI) has made impressive strides, accomplishing tasks that were once deemed nearly impossible for human beings. Nonetheless, this growth has also given rise to substantial concerns, especially concerning the transparency and explainability of AI systems. Thus, the birth of (XAI) [1] wich aimes at making AI models understandable and interpretable for users and stakeholders. Instead of solely providing outcomes, explainable AI models furnish explanations on how decisions are provided especially for opaque models.

An opaque or black-box arrive at conclusions or decisions without providing any explanations as to how they were reached, while a transparent model is a perfectly known model because it is possible to build it entirely from previous knowledge [2].

XAI models can be categorized as local or global. Each of these models aims to explaining how the AI model makes predictions at different levels. The local model focuses on understanding the behavior of AI algorithms locally, its hierarchical level is low, often based on a single observation (or on a small subset of observations). These models provide metrics related to how each feature contributes to a given final prediction generated by the AI model. On the other hand, global model focuses on understanding the behavior of the AI algorithm at a high hierarchical level, i.e how features contribute to all predictions performed by the model [3].

Interestingly, there is an absolute difference between interpretability and explainability in XAI. Interpretability refers to the ability to understand the inner workings of the model, either the model as a whole or at least the parts of the model relevant to a given prediction. This might involve understanding decision rules, thresholds and the capability to manually derive model outcomes.

While in explainability, we believe that a model's prediction is explainable if a mechanism can offer (partial) information about the prediction. For instance, this could involve identifying which aspects of an input were most significant for the resulting prediction or recognizing what changes made to an input would lead to a different prediction. This understanding can be leveraged to enhance the model's abilityto predict more accurately in the future [4].

Among the objectives of XAI are:

- Enhancing users' confidence by enabling them to comprehend the rationale behind AI-driven decisions.
- Facilitating the detection and rectification of potential errors or biases in models, thereby enhancing the quality and reliability of AI models.
- Adhering to ethical standards and data protection regulations. Users have the right to know how their data is used and how decisions are made.

In this article, we present experimental studies on tabular data involving four methods: LIME, SHAP, ANCHOR and EBM.

Our work concentrates on quantifying measures to assess explainability techniques. The main contribution of this article is as follows: First, we introduce two powerful machine learning models which are Random Forest and XGBoost. We provide a detailed experimental evaluation of three recent and popular agnostic local explainability techniques: LIME, SHAP, and ANCHOR, and examine the behavior of predictions from each method. Second, we compare the results of local and global explanations. Finally, we introduce the EBM model and assess its performance (accuracy) along with its various local and global explanations.

Random Forest is one of the most popular and commonly used algorithms by Data Scientists. It is a Supervised Machine Learning Algorithm that is used widely in classification and regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. Among its advantages:

- It can be used in classification and regression problems.
- Having a larger number of trees in the forest leads to greater accuracy and helps avoid the issue of overfitting.
- It performs well even if the data contains null/missing values.

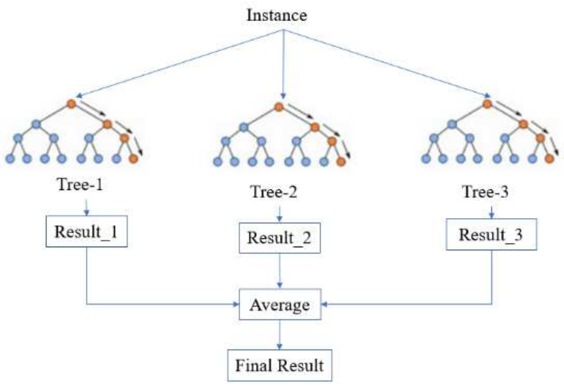

Figure 1: Simplified structure of Random Forest [5]

XGBoost stands for extreme gradient boosting. This approach is based on decision trees and builds upon other methods such as random forest and gradient boosting optimization. It performs well with large and intricate datasets by employing various optimization techniques.

XGBoost has achieved remarkable results in machine learning competitions, not only due to its principle of sequential self-improvement but also because it encompasses a significant number of hyperparameters that can be adjusted and fine-tuned for enhancement purposes.

This high degree of flexibility makes XGBoost a very robust choice [6].

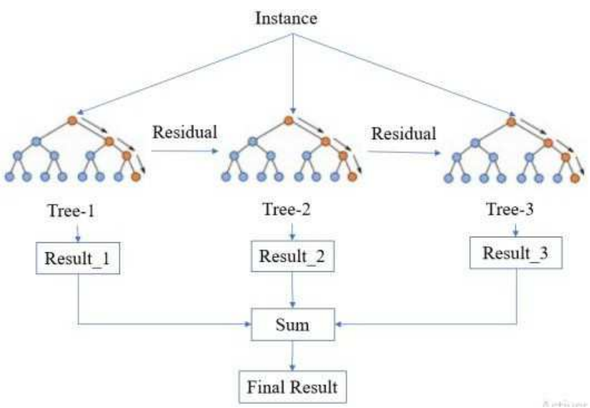

Figure 2 : Simplified structure of XGBoost [5].

In this article, we attempt to apply explainable models on opaque systems. The rest of this article is structured as follows: section 2 sheds light on the overview and the studies explainable model. Section 3 presents the experience made on the explainable models proposed as well as the results obtained. Section 4 provides some discussion. While section 5 is a conclusion.

II. OVERVIEW OF THE STUDIES EXPLAINABLE MODEL

A. LIME

The LIME technique, introduced as a method for local interpretability, operates on the premise that within the boundary of a complex machine learning model can be approximated as linear. This approach elucidates the instance in question by constructing an interpretable model based on perturbed samples derived from the input instance of interest. Specifically, LIME generates perturbed samples centered around the instance requiring explanation. Within this perturbed sample set, for each instance, LIME obtains predictions from the model to be expounded upon. This collection of perturbed samples and their associated predictions then serves as the training dataset for the interpretable model. Subsequently, the technique assigns weights to the instances within the new training dataset based on their proximity to the instance necessitating explanation. Ultimately, LIME fits an interpretable model using this newlyestablished training dataset [7].

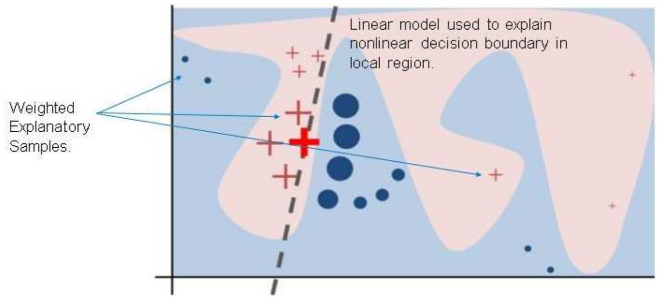

Figure 3: Presenting intuition for LIME [8]

B. SHAP

Vicinity of the instance under examination, the decision SHAP (short for SHAPley Additive exPlanation) for the post-hoc explanation of machine learning methods. The model generates a prediction value for each test sample and provides an explainable prediction. The main idea is to calculate the marginal contribution of the features added to the model, i.e., the SHAP value, which is equivalent to the impact of the features on the sample. In cooperative game theory, the SHAP value is calculated in equation (1) as follows [9]:

$$\Phi_m = \sum L \subseteq N\{m\}|L|!\,(M - |L| - 1)!\,M! \cdot [v(L \cup \{m\}) - v(L)] \quad (1)$$

Where Φ_m is the contribution of the m feature, L is thefeature subset, N{m} is the feature set, M is the total numberof input features, v (L ∪ {m}) is the predicted value of themodel when the sample has only the feature values in L ∪ {m} and v (L) is the predicted value of the model when the sample has only the feature values in L. In line with the additive eigen property approach, the linear function g is defined in equation (2) as follow [10]:

$$g(x) = \Phi_0 + \sum m = 1M\Phi_{mxm} \quad (2)$$

where g(x) is the explained model prediction for sample x, Φ_m is the mean of the model prediction and xm is the mth feature sample [11].

More concretely, the SHAPley value operates by justly dividing the variation between the prediction and the average prediction among the values of features for the instance under consideration.

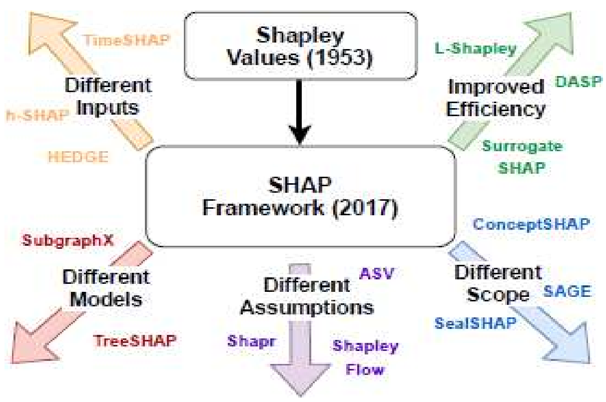

Figure 4 : the five research directions pursued by SHAPley-andSHAP-based approaches in XAI [12].

C. ANCHOR

An ANCHOR explanation is a rule that sufficiently "ANCHORs" the prediction locally – such that changes to the rest of the feature values of the instance do not matter. In other words, for instances on which the ANCHOR holds, the prediction is (almost) always the same.

ANCHORs are intuitive, easy to comprehend, and have extremely clear coverage – they only apply when all the conditions in the rule are met, and if they apply the precision is high (by design) [13]. In practical terms, the greedy approach encounters certain limitations:

- it can only maintain a single rule at a time, thereby preventing modification of suboptimal choices,
- it yields the shortest ANCHOR, which might not coincide with the ANCHOR possessing the highest coverage.

To address these drawbacks, an alternative approach to constructing ANCHORs is the beam-search method. This technique maintains a collection of candidate rules and directs the search towards selecting the ANCHOR with the most extensive coverage among the myriad of possible ANCHORs.

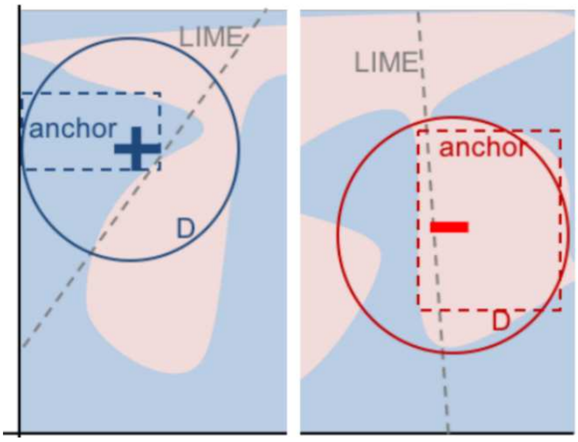

Figure 5: Different from LIME, ANCHORs uses the "local region" to learn how to explain the model [14]

D. Explainable Boosting machine (EBM)

The Explainable Boosting Machine (EBM) is a glassbox model engineered to achieve accuracy akin to leading machine learning methods such as Random Forest and Boosted Trees, while maintaining high explainability. EBM is structured as a generalized additive model indicate in equation (3) as follow [15]: $g(E[y]) = \beta_0 + \Sigma f_j(x_j)$ (3)

Where g is the link function that adapts the Generalized Additive Model (GAM) to different parameters such as regression or classification. EBM introduces several key enhancements over traditional GAMs (Hastie and Tibshirani, 1987). Firstly, EBM learns each function fj using modern machine learning techniques such as bagging and gradient boosting. The boosting procedure is carefully confined to training on one feature at a time in a round-robin manner, employing a very low learning rate to ensure feature order doesn't impact the results. It iterates through the features to mitigate collinearity effects and learn the optimal fj feature function for each attribute, showing how each feature contributes to the model's prediction. Secondly, EBM can automatically detect and include pairwise interaction terms indicate in equation (4) as follow [15]:

$$g(E[y]) = \beta_0 + \Sigma f_j(x_j) + \sum f_{ij}(x_i, x_j) \quad (4)$$

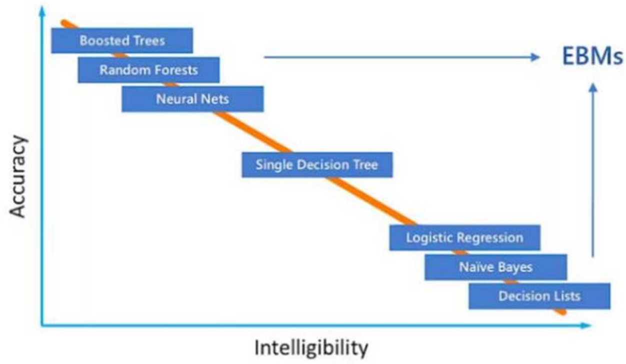

Figure 6: "The Science Behind InterpretML: Explainable Boosting Machine" on YouTube by Microsoft Research [16]

In terms of predictive capability, EBM consistently exhibits surprisingly robust performance, on par with state-of-the-art methodologies such as Random Forest and XGBoost. Another advantage of EBM surfaces in its lightweight inference procedure, making it particularly suitable for production environments where minimizing prediction latency is crucial.

III. EXPERIMENTS AND RESULTS

In this experiment, we used a "Mobile Price Classification" dataset from Kaggle, containing data tabular information on the mobile phone (2000 lines) namely: ram (random access memory in megabytes), battery life (longest time that a single battery charge will last), Wi-Fi (wireless networking technology) and 4G support, etc [17].

Our dataset contains categorical data such as: "dual_sim", "four_g" and "wifi" that we will use in the ANCHOR method. For the dataset split ratio between training and validation, we took 80% for training and 20% for validation.

We have a multi-class dataset classification. The aim of the prediction is to find the price range of the mobile device based on the mentioned features. We have four distincts categories which are: "Zero", "One", "Two", and "Three".

Regarding the hardware aspect of our experimentation, we used a 2019 MacBook Pro with the following configuration:

- CPU: Intel i5, 8th generation with 8 CPUs and a frequency of 1.4GHz.
- RAM: 16GB DDR3 clocked at 2133MHz.
- GPU: Intel Iris Plus Graphics Family with 8GB of VRAM.

We applied two machine learning models to our dataset: Random Forest and XGBoost, the notable performance results are shown in Table (1).

we have 400 occurrences of each class in y_true, the finding results noted in Table (1) indicate that Random Forest holds a slight edge over XGBoost in terms of performance including the classifier's ability to identify all positive samples(recall) and the weighted harmonic mean of precision(F1-score)[18].

	Precision	Recall	F1-score	Support
Random Forest	0.80	0.80	0.80	400
XGBoost	0.79	0.79	0.79	400

Table 1: The metrics of Random Forest and XGBoost for our model

A. *First Experiment: Local explainability*

The local explanations help us understand what is happening at each prediction level locally.

1) LIME

LIME presents the impact of each variable on the final decision. For instance: Considering that the models' predictions yield the "Three" class as the outcome. We can observe that the variables "ram" and "battery" have a positive influence whereas the variables "wifi" and "four_g" have a negative influence on the prediction for Random Forest (Figure 7), but only variable "four_g" have a negative impact on the prediction for XGBoost (Figure 8)

Then, we observe a minor influence of the "power_battery" and "mobile_wt" variables on the same prediction between the two models (Random Forest and XGBoost). Overall, the results are quite similar.

2) SHAP

Regarding the explanations of visualizations. There are three alternatives: force plot, decision plot, and waterfall plot. Firstly, the force plot is useful for visualizing where the output value stands relative to the base value (we noticethat there is a difference of f(x) between the two models). We can also observe which variables have a positive impact (inred: "ram" contributes to making the prediction higher than the baseline value) or a negative impact (in blue) on the prediction and the magnitude of this impact (Figure 9, Figure 10).

Secondly, the waterfall plot also enables us to observe the magnitude and nature of a variable's impact along with its quantification. It also helps reveal the order of the variable importance and the values taken by each variable for the studied instance (Figure 11 and Figure 12).

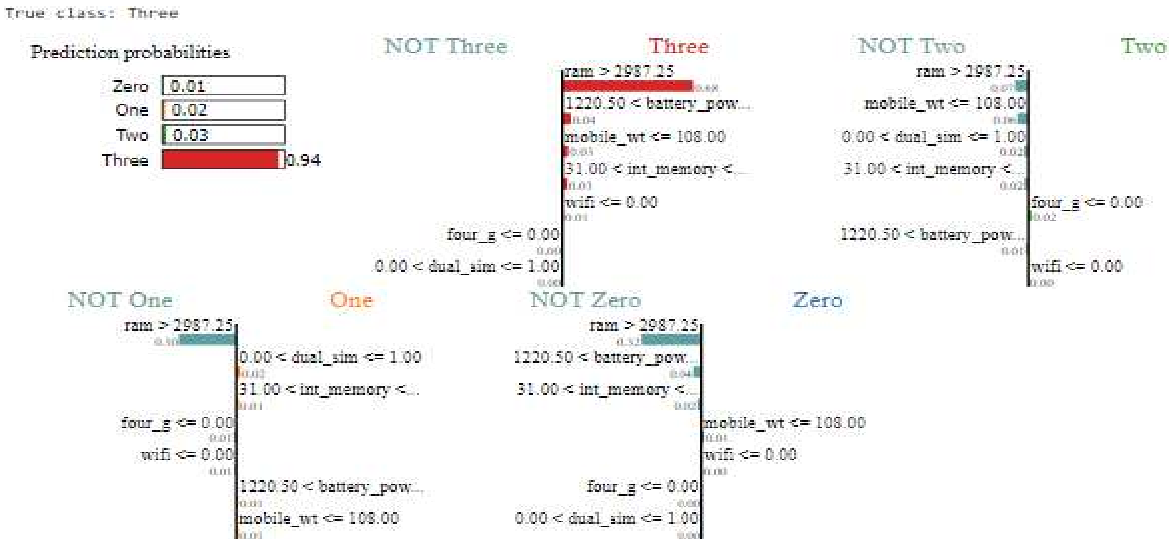

Figure 7 Example of an explanation describing the influential features for predicting (LIME with Random Forest)

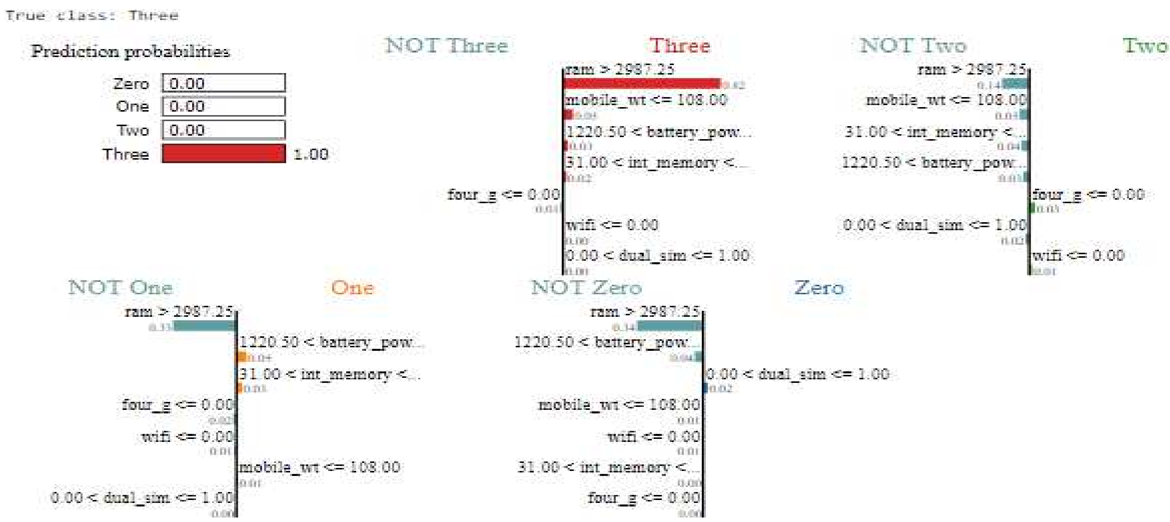

Figure 8 Example of an explanation describing the influential features for predicting (LIME with XGBoost)

Figure 9: Force plot gives us the explainability of a single model prediction (class "Three": SHAP with Random Forest)

Figure 10: Force plot gives us the explainability of a single model prediction (class "Three": SHAP with XGBoost)

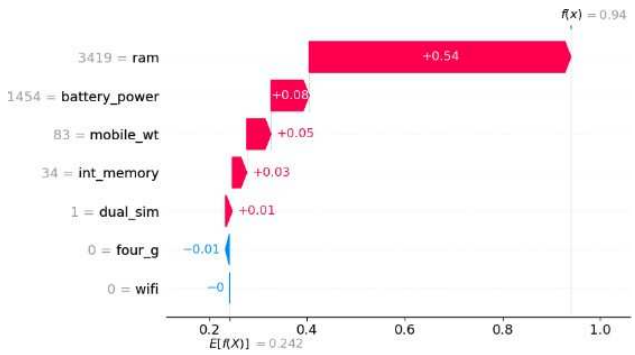

Figure 11: Waterfall plot gives us the explainability of a single model prediction (class "Three": SHAP with Random Forest)

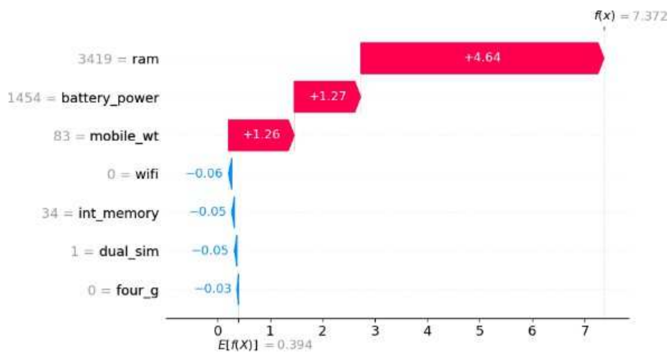

Figure 12: Waterfall plot gives us the explainability of a singlemodel prediction (class "Three": SHAP with XGBoost)

3) ANCHOR

First and foremost, it's necessary to specify the categorical variables (the index of their column and the values they take). In our case, we have three categorical variables: "wifi", "dual_sim,"and "four_g," with values ranging between 0 and 1. The explanation provided for the chosen class (the predicted class "Three") reveals the variables influences this prediction.We observe that a RAM value relatively higher than the average (RAM > 2985.25) and a battery power value greater than 1220.50. So, this explanation achieves remarkable accuracy(97.3% for Random Forest (Figure 13) and 97.9% for XGBoost(Figure 14)) and includes an additional piece of information called coverage. Coverage describes the proportion of observations that satisfy the ANCHOR; in this case,14% of the tested data possesses the characteristics defined by the ANCHOR.

4) EBM

Dashboard is a great feature from Interpret ML which allows you to see all the results in one view (Figure 15). The explanations available are split into tabs each one is covering an aspect of the pipeline.

- Data covers exploratory data analysis which is designed mostly for feature-level.
- Performance covers both model performance and user-defined groups.
- Global explains model decisions.
- Local explains a model decision for every instance/observation in one view [19].

We notice that the predictions provided by the Random Forest, XGBoost, and even EBM models are similar (Table 1), with EBM having an advantage in accuracy reaching 81%.

In classification, the intercept is the log of the base rate (for example -2.8 if the base rate is 10%). So, this graph lets us explain model prediction on an individual sample byshowing us a bar chart of how much individual features contributed to this prediction (Figure 15). It accepts a list of samples and their predictions as an input for generating an explanation object.

Also, we observe the positive influence of the "ram" variable on the "Three" class.

B. Second Experiment: Global explainability.

Global explainability lets the model owner determine to what extent each feature contributes to how the model makes its predictions over all the data. Given that LIME and ANCHOR are methods for local interpretation, we are left with two methods to consider: SHAP and EBM.

1) SHAP

The summary plot for multiclass classification (Figure 16) shows the overall importance of the variables calculated by the absolute SHAP values for each class.

In our case, we can see that the class drop hardly uses "wifi", "dual_sim" and "four_g" functionality. We can also see that the classes allow and deny the same functionality almost equally. This is the reason why the confusion between them is relatively great. For a better separation of the classes of authorization and refusal, it is necessary to generate new functionalities only dedicated to these classes.

Each point in the summary plot (Figure 17) is a SHAPley value for a feature and an instance. The position on the y-axis is determined by the characteristic and on the x-axis by the SHAPley value. We can see that the "wifi" feature is the least important feature, and it has low SHAPley values. The color represents the value of the feature from low to high (red dots for high values and blue dots for low values).

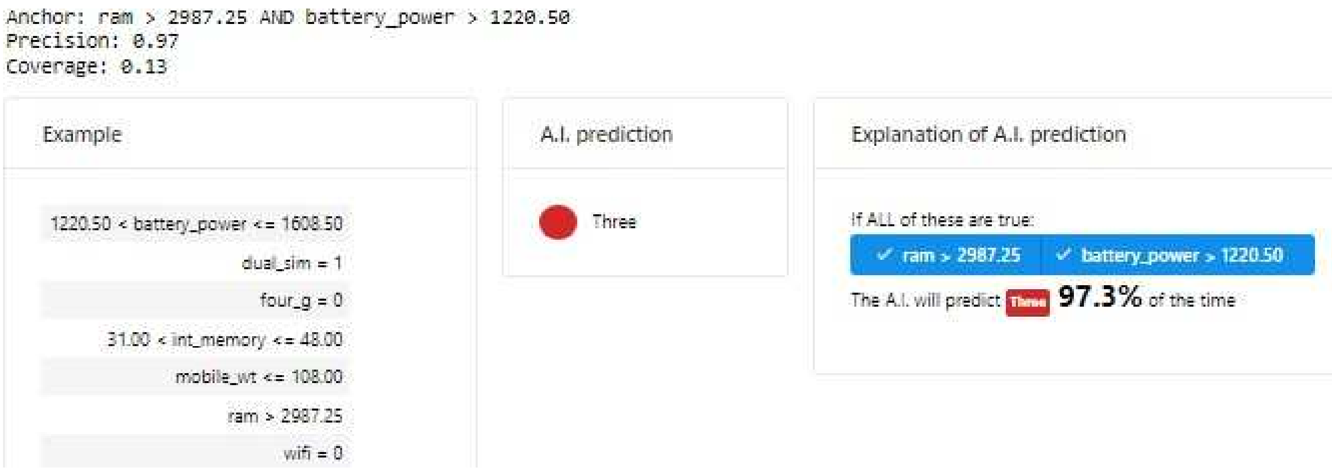

Figure 13: Local prediction (ANCHOR with Random Forest)

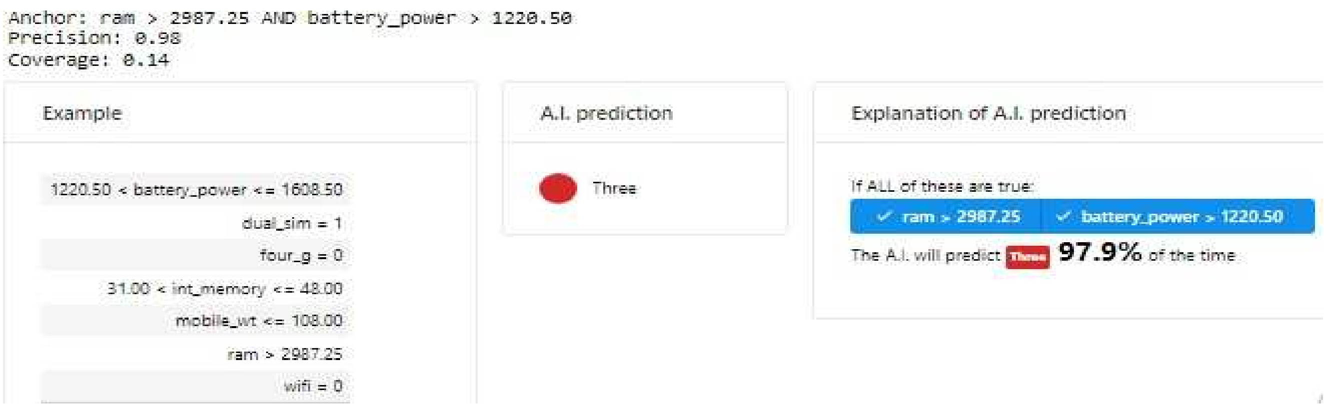

Figure 14: Local prediction (ANCHOR with XGBoost)

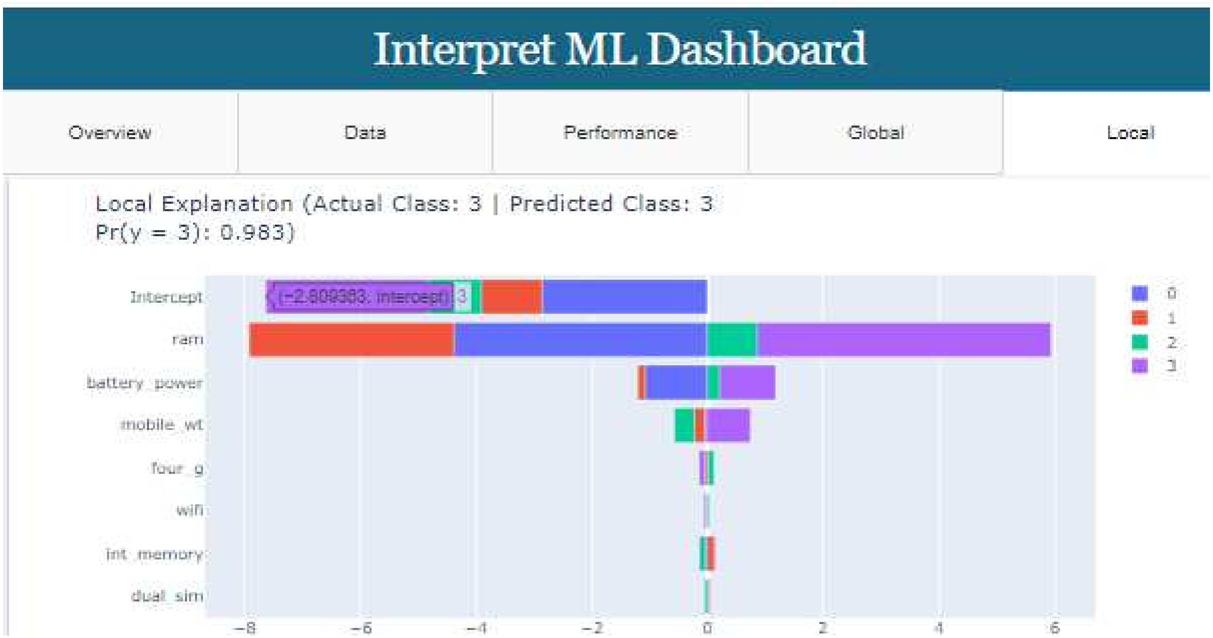

Figure 15: This can be used to view local explanation of all the models being used for predication in one view

The dependence plot is a scatter plot that shows the effectof a single feature on the predictions made by the model.

In the Figure18 and Figure19, we can observe a distinct vertical pattern of color coding that indicates the relationship between the attribute's "ram" and "battery_power".

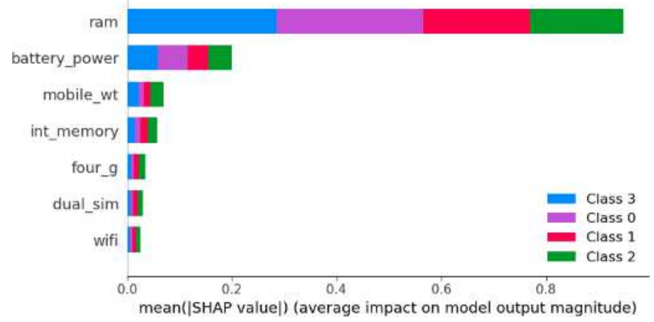

Figure 16: Summary plot shows the importance of the variables foreach class

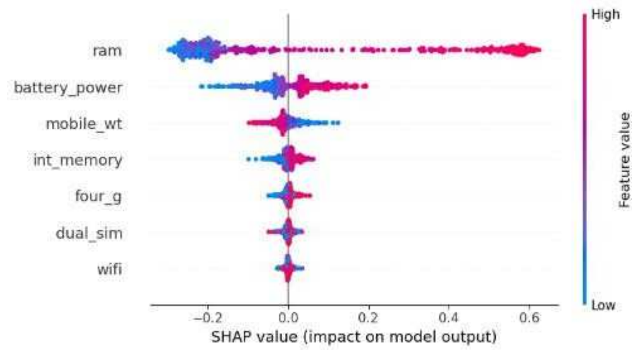

Figure 17: Summary plot shows the values of SHAP represented for each variable in their order of importance

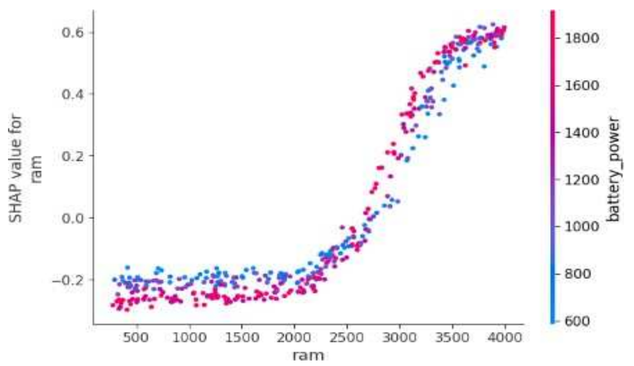

Figure 18: Dependance plot shows interaction between the features, "ram" and "battery_power" (Random Forest)

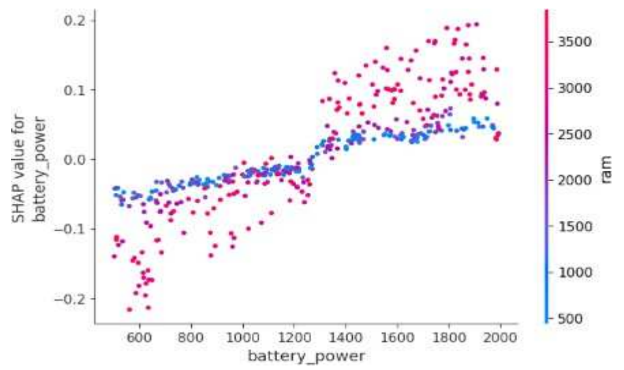

Figure 19: Dependance plot shows interaction between thefeatures, "ram" and "battery_power" (XGBoost)

2) EBM

The feature importance summary (Figure20) shows that two features "ram" and "battery_power" are very essential features. We can also look at individual characteristics to notice the impact.

Figure 21 for example shows the variations of the features "ram" according to the classes.

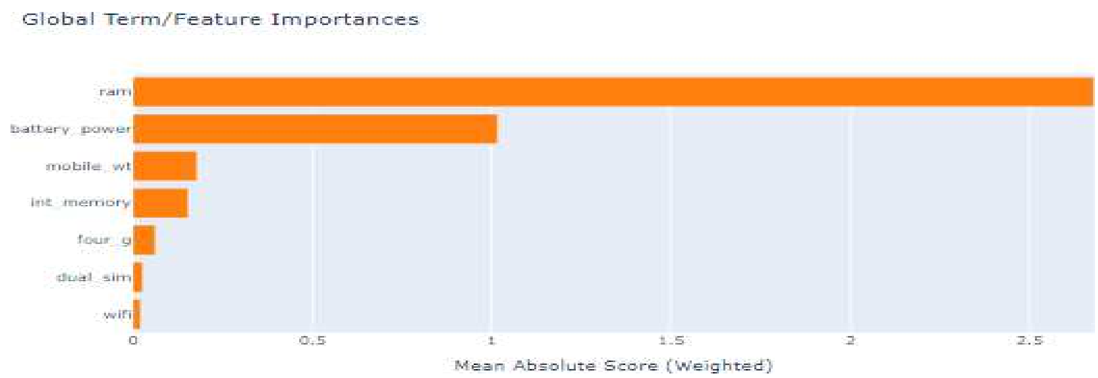

Figure 20: Summary of features in order of importance

Figure 21: Variations of "ram" functionality by class

Models / Criteria	LIME	SHAP	ANCHOR	EBM
Approach	Perturbed samples derived from the input instance of interest.	Calculate the average marginal contribution of a feature value across all possible coalitions (Shapley value)	The same approach as LIME, but the resulting explanations are expressed as easy-to-understand IF-THEN rules.	A glass box model, designed for high precision, uses modern machine learning techniques such as bagging and gradient boosting.
Hierarchical Level	☑ Local. ☐ Global.	☑ Local. ☑ Global.	☑ Local ☐ Global.	☑ Local. ☑ Global.
Results Presentation	• Predictions probabilities. • influence rate of features for each class.	• Force plot. • Waterfall plot. • Summary plot. • Dependence plot.	• Predictions in rules form. • Explanation of predictions.	• Interpret ML Dashboard. • Feature importance. • Variations of each feature by class.
Accuracy	Applied on results of: • Random Forest: 94%. • XGBoost: 100%	Not displayed	Applied on results of: • Random Forest: 97.3%. • XGBoost: 97.9%.	• EBM: 98.3%.
Execution time	Fast execution (0.07s)	Slower than other models especially if there are many features (0.85s)	Slightly slower than LIME (0.08s)	The fastest among these models (0.05s)

Table 2: Comparison between LIME, SHAP, ANCHOR and EBM

IV. DISCUSSION

The main similarities and differences between the XAI models studied are provided in Table 2.

The four models offer comparable results with variations in terms of presentation, accuracy and execution time.

- For local explanations: LIME presented results by displaying the predicted class and positive influencing degree of each feature. SHAP offered a variety of presentations that could enrich explanations. ANCHOR provided an intuitive presentation of results in a rule form with a greater precision compared to LIME. Meanwhile, EBM in addition to predicting the class, once demonstrated a higher accuracy than LIME and provided a more detailed display.
- For global explanation: SHAP and EBM present their results in the form of graphs that aid incomprehending the overall contribution of features to the model and how each feature is connected to the model. Additionally, SHAP offers a "dependence plot" that can illustrate interactions between two features.

V. CONCLUSION

In this study, we conducted a comparative study between global and local explainable model. We examined distinct models, namely LIME, SHAP, ANCHOR and EBM, the findings indicated that there wasn't a definitive winner, as the results provided by these explainability techniques are quite comparable, the difference lies in how these results are presented.

It is important to note that all models have identified the same features that influence positively the predictions, meanwhile SHAP present more graphical presentation.

The EBM model is the best fit for local and global explanations with high accuracy and fast execution time.

In the future works, we plan to propose our explainable model and we will compare it to studied models of this paper.

REFERENCES

[1] Peter E.D. Lovea, Weili Fangb, Jane Matthewsd , Stuart Portere, Hanbin Luof, Lieyun Ding. Explainable Artificial Intelligence (XAI): Precepts, Methods, and Opportunities for Research in Construction,

[2] Lennart LjungDiv. Black-box Models from Input-output Measurements : of Automatic ControlLink¨oping UniversitySE-58183 Link¨oping, Sweden

[3] Kleyton da Costa, Adriano Soares Koshiyama. Local and Global Explainability Metrics for Machine Learning Predictions: Pontifícia Universidade Católica do Rio de Janeiro

[4] Christian Kästner. "Interpretability and Explainability"

[5] Weilun Wang, Goutam Chakraborty and Basabi Chakraborty. Predicting the Risk of Chronic Kidney Disease (CKD) Using Machine Learning Algorithm

[6] Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System : : University of Washington.

[7] Radwa El Shawi, Youssef Sherif, Mouaz H Al-Mallah and Sherif Sakr Interpretability in healthcare: A comparative study of local machine learning interpretability techniques University of Tartu.

[8] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin "Why Should I Trust You?" Explaining the Predictions of Any Classifier : University of Washington

[9] Chauvin, C.; Lardjane, S.; Morel, G.; Clostermann, J.-P.; Langard, B. Human and organisational factors in maritime accidents: Analysis of collisions at sea using the HFACS. Accid. Anal. Prev. 2013, 59, 26–37.

[10] Zhang, J.; Teixeira, Â.P.; Soares, C.G.; Yan, X. Quantitative assessment of collision risk influence factors in the Tianjin port. Saf. Sci. 2018, 110, 363–371.

[11] Lin Chuan Fusing XGBoost and SHAP Models for Maritime Accident Prediction and Causality Interpretability Analysis Fuzhou University.

[12] Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, SHAP-Based Explanation Methods A Review for NLP Interpretability TU Munich, Department of Informatics, Germany

[13] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin ANCHORs: High- Precision Model-Agnostic Explanations University of Washington.

[14] Edward Ma , ANCHOR your Model Interpretation by ANCHORs Towards Data Science.

[15] Harsha Nori, Samuel Jenkins, Paul Koch, Rich Caruana, InterpretML A Unified Framework for Machine Learning Interpretability Microsoft Corporation1 Microsoft Way Redmond, WA 98052, USA

[16] Mayur Sand, Model Interpretation with Microsoft's Interpret ML Analytics Vidhya May 23, 2020

[17] Mobile Price Classification Classify Mobile Price Range, https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification : accessed on 22 April 2023.

[18] Metrics of models, precision, recall and fscore, https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html : accessed on 11 June 2023.

[19] Michał Oleszak, Explainable Boosting Machines Towards AI Jan 23, 2022.