# Interpretable Model-Agnostic Methods for Image Classification: A Survey

Vano Mazashvili (Student ID - 1993251)

*DIAG - Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, Italy*

ARTICLE INFO

*Keywords*:
EXplainable AI (XAI)
Model-Agnostic Interpretability
Image Classification
Artificial Intelligence
SHapley Additive exPlanations (SHAP)
Local Interpretable Model-Agnostic Explanations (LIME)
Image Analysis

ABSTRACT

The rise of new machine learning technologies and sophisticated hardware coupled with an abundance of available data has propelled the field of machine learning to widespread adoption. Although the harder tasks usually require more complex architectures with many parameters, newer models have become opaque and thus hard or impossible to interpret. Its reasoning/decision making is a secret to us. This is a big challenge if we aim for the safety and transparency. Opaque architectures pose a risk to privacy, security, and safety, especially when important and even morally correct decisions must be made. Interpretable or eXplainable AI is a logical conclusion of the issues presented above. This type of model is easily understandable and interpretable by humans and can be used to implement a social right to explanation. Because the XAI field is as diverse as the solutions tend to be catored to the task, in this survey, we focus on more problem specific interpretable model-agnostic methods in complex image classification models. The main methods are presented: we delve into their theoretical foundations, applications and comparative strengths. Additionally, comprehensive overview of the newer approaches are presented. Through this analysis, strengths and limitations of each model are brought out to the light, enabling us to visualize the current challenges in the field and propose possible solutions/future directions

## 1. Introduction

The rapid advancements in sophisticated chip development, accompanied with the better understanding of the algorithm architectures and ever-growing available data, enabled us to tackle many challenges deemed impossible just several years ago. Fields like computer vision, medical diagnosis, natural language processing, and expert systems found great use of ever-developing machine learning and deep learning models. Although reaping the benefits, novel architectures pose unprecedented challenges - unlike the earlier methods with their interpretable mechanism, the intricate architectures of the cutting-edge neural networks are often deemed to be impenetrable black boxes; in short, there is clearly a trade-off between the capability and opacity. This issue is more prominent in fields where the decision-making process is crucial. We should be able to interpret the reasoning of the models used in healthcare, security, and autonomous systems, where critical and often morally challenging decisions must be made.

Thus, the emerging field of explainable AI (XAI) is the result of increased model complexity, growing dependence on AI systems, diminishing user trust and acceptance, and ethical and regulatory demands. With the aim of demistifying the hidden mechanisms of the contemporary machine learning and deep learning models, XAI is a way to interpret existing models and to build novel, inherently interpretable ones.

As generally accepted, XAI methods are categorized into two main families: posthoc and antehoc (8). Anthehoc methods incorporate explainability during the training phase. This means that the model is designed from the ground up with interpretability in mind. On the other hand, posthoc method completely excludes itself from the architecture and tries to approximate the already trained black box, in this case, the explainebility is either an afterthought, not the priority, or the design choice in favor of the performance - as the more intervention usually hinders the model's capability (8).

There are a plethora of approaches to "decode" the black box architecture. Usually, the methods change from model to model, as they can be adapted to specific types of neural network, their task, and objectives. However, the end goal, at least for the post-hoc methods, is a universal and versatile solution without the need for many adjustments. This is what model-agnostic methods offer. We get: (14; 16)

**1. Cross-Domain Applicability -**as the models and data become more complex and diverse, it is very important that the explainer can work with the models with the training and testing data sampled from the different distributions.

**2. Versatility across wide range of model architectures -** model-agnostic methods try to "mimic", or approximate the black-box structure by observing the input-output relationships. Thus, they can be applied to any machine learning model (16).

**3. Explainability AND accuracy -** posthoc methods can interpret the model without the need to change its architecture, thus, retraining is not necessary - removing the possibility of introducting the unintended alterations in model's predictive performance.

Therefore, in this survey, we dive into model-agnostic interpretability (posthoc) methods, as they allow us to explain diverse machine learning models in the field of image classification. We present various contributions that describe different architectures and methods and create a good understanding of their applications, strengths, and weaknesses. The derived understanding of these methods should be a

mazashvili.1993251@studenti.uniroma1.it (V. Mazashvili)

| Local Model-Agnostic Explanations | Global Model-Agnostic Methods |
|---|---|
| MAIRE | SHAP (SHapley Additive exPlanations) |
| Local Surrogate(LIME) | KernelSHAP |
| Counterfactual Explanations | Permutation Feature Importance |
| Scoped Rules (Anchors) | Global Surrogate |
| DLIME | |
| LEMON | |

**Table 1**
Existing Model-Agnostic global and local methods categorized

good indicator of the importance of the deep comprehension in complex image classification models.

## 2. Problem Description

The study of eXplainable Artificial Intelligence (XAI) and model-agnostic interpretability in not just a virtue, or theoretical pursuit; the necessity of it became apparent gradually, from the historical examples in various fields. While the simpler methods like decision trees and linear models (e.g linear SVM, linear regression) can provide inherently explainable predictions, modern architectures, especially deep neural networks, while extremely accurate, have become opaque and non-interpretable.

The tragic incidents in 1980s involving Therac-25 medical radiation machine are one of the early and prominent examples of disastrous outcome of machine's opaque decision-making and faulty human-centric interaction. This is a stark reminder, that every possible safety measures must be taken when human health and well-being are at stake.

Another example, fortifying the argument against AI non-transparency, would be the 37th move from AlphaGO in historic Go match between Lee Sedol, one of the best Go players in the world, and a groundbraking AI developed by the Google DeepMind. The move was so unexpected and unexplainable at first that it left everyone baffled - even Sedol himself left his seat to take a break. Interestingly, the inhumane play deemed as a mistake at first, turned out to be a deciding move for AlphaGO's victory (17).

Algorithmic trading models' role in market volatility was also scrutinized in the financial field after the "Flash Crash" in 2010, where the markets experienced sudden and violent downturn. In medicine, the AI must not misidentify the medical condition, comprehensive and interpretable explanation should be provided for the decisions made in criminal justice systems, autonomous vehicles and general human-centric designs.

Thus, with this survey, our goal is to embark on an exploration of the interpretable models in the realm of image classification with the aim of underline the necessity of explainability and contribute to the development of the truthworthy and ethically sound AI systems. We focus on model-agnostic methods, as they do not depend on the specific system architectures. Methods like Interpretable Model-Agnostic Explanations (LIME), surrogate models, Model-Agnostic Iterative Refinement (MAIRE), SHapley Additive exPlanations (SHAP) are surveyed and documented, providing explanations, challenges, shortcomings and future directions.

### 2.1. Existing Methods Categorized

Table 1 should serve as a good compass for the discussed methods in this survey. While the local methods try to interpret the predictions one by one, Global methods tend to derive how individual features influence the prediction on average. The survey will focus on local methods first, and investigate the global ones after.

#### *2.1.1. Local Model-Agnostic Methods*

**Local Interpretable Model-Agnostic Explanations (LIME) -** is a local surrogate interpretable model used to explain individual predictions from the black box. This is achieved by mimicking the behavior of the target neural model locally with an interpretable surrogate model. This can be any explainable model, ranging from linear regression to decision trees. The surrogate model is trained with the data created from the perturbed sampled instances around the input (14).

$$\xi(x) = \underbrace{\underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x)}_{\text{Minimized fidelity function}} + \underbrace{\Omega(\mathrm{g})}_{\text{Complexity measure}} \tag{1}$$

Equation 1 captures the essence of the fidelity-interpretibility trade-off for local model-agnostic models. High local fidelity means a locally faithful explanation, while interpretable explanation takes into account user's limitations, as it is clear and understandable for a human (14; 16). This means that complex explanations are more faithful but less interpretable. Thus, to ensure the best solution with high interpretibility and faithfulness, we minimize the sum of the value of the fidelity function and the complexity measure.

To train the surrogate model, we select the input instance we want the explanation of, then we perturb the dataset and get the new predictions, these new samples will be weighted and summed up - the samples closer to the input will have higher weights. With that, we train the interpretable surrogate model. It is noteworthy that we need to first choose the number of features for the explainer. Higher number of features ensures good local fidellity, but makes it hard to train the model.

For images, LIME does not perturb the individual pixels, but the groups of them, refered as superpixels (14). These groups are the areas of an image with same color, with a higher likelihood of them to be of a same class.

Although its ease of use and flexibility come with several downsides - instability of explanations and poor performance with high-dimentional data (19; 2). (1) experimentally investigated these instabilities by "focusing on small perturbations that have minimal (or no) effect on the underlying model's predictions, yet have significant effects on the explanations given..." (1). Meaning that, in experimental settings, the explanations for two very close points had significant variations.

**DLIME -** or Deterministic LIME, is designed to address the instability issues presented above by utilizing agglomerative hierarchical clustering (HC) to partition the data set (19). While random perturbations used in LIME seemingly introduce instability in explanations, DLIME's HC is a deterministic way to partition the dataset into different clusters - ensuring consistent partitioning (clustering) from the same dataset.

After clustering is done, the KNN classifier determines the appropriate cluster for the new instance with similar data points. These data points from the given cluster are then used to train a linear regression model - generating the explanations (19).

The model yielded better results than LIME in data sets from the UCI repository, underlining its importance in Computer-Aided Diagnosis systems (19).

**LEMON -** or Local Explainable MOdel explanations using N-ball sampling, aims for more robust and faithful local model-agnostic explanations by introducing novel sampling methods, guaranteeing more relevant samples in the neighborhood of the instance. This is attained through approximation of the sample density by leveraging Kernel Density Estimation (KDE) within the certain radius around the instance (2). In this way, the chances of obtaining samples close to the instance to be explained remain high (2). The theory was verified in experimental setting using real-world datasets and various ML models, where LEMON outperformed LIME in faithfulness measures (50.8% lower RMSE and less affected by the data variation) (2).

**Anchors -** are another model-agnostic method from the creators of LIME (14). The main goal is to provide a high-precision explanation designed to be intuitive and easy to understand. Generated anchors represent the group of instances in the hyperspace with the same "sufficient" base conditions for predictions (16). That is, changes to the rest of the feature value for the sampled instance do not matter (16) - for that anchored instances prediction is almost always the same. Thus, an anchor captures the minimal set of conditions that are necessary and sufficient for the model to make a particular prediction for a given instance (16). This is why scoped rules achieve very high, human-friendly interpretability - as it can capture non-linearity with clearly defined precision and coverage.

The anchors can be found by generating candidates first, then evaluating their precision empirically on the perturbated dataset - querying black-box with those instances. After finding the anchor with the desired precision, we refine it by "perturbing" the features to improve precision, coverage, or both (16). In the end, we have scoped if-else rules, non-linearly explaining the chosen sample locally.

Anchors have a plethora of advantages over classical perturbation-based linear model-agnostic methods like LIME. Firstly, the resulting interpretibility is very understandable for an average person. Then, there is its high precision and clearly defined coverage. While for LIME, or its derivatives, local explanation means converting models' nonlinearity to surrogate models' linear explanation, Anchors are able to preserve the nonlinear nature even locally.

However, all of these positives come with some disadvantages. Although anchors can be used for every domain (therefore, it's model-agnostic nature), its extremely high configurability demands in-depth customization in order to yield any meaningful explanation. On top of that, the algorithm is computationally expensive, especially for the large datasets or instances with many features, as it requires probing the machine learning model many times.

Another big problem with Anchors is that discretization becomes necessity as with too many rules, coverage may suffer. Its trivial to imagine that too detailed continuous feature explanation will ask for discretization. After doing so, although we achieve explanations, the boundaries get blurry, therefore, it is hard to explain odd and very specfic instances.

The reasons provided above, gives us the arguements against Anchor's use case in image classification task. Although still possible, it is a challenge with many variables.

**MAIRE -** is another rule-based model method designed to give local explanations. It tries to address the problem with Anchors - the need to discretize the sample hyperspace. Also, the model is characterized with fewer parameters. For example, it can automatically generate the boundaries for the optimal orthotope containing the instance. In other words, it can generate the hypercuboid aligned with the largest axis for the same class label as the instance being explained (18). Accordingly, the explainer can work with high-dimensional datasets without any need to cater the orthotope parameters. Without the requirement for the dimensionality reduction, the explainer can be more suitable for image datasets.

To achieve superior performance, the algorithm first selects the most relevant features by calculating feature importance scores, then it identifies the range of values which contribute to the decision for the selected feature. This is done because only feature importance scoring is not sufficient, as one range might describe one class and other range - different one (18). Therefore, just feature relevancy is incomplete and will not capture all mechanisms. Only after that we get to the rule generation, which, like in Anchors, involves combination of these features and respective range values into an if-then-else statements. These rules are then refined and simplified by removing redundancies.

To sum up, although being a complex, computationally expensive, non-trivial continuation of Anchors method, MAIRE provides a high quality, flexible model-agnostic explanations, by leveraging strong optimization framework designed to maximize the multi-dimensional coverage and precision. It is effective in its high-dimensional rule extraction and explanation interpretability.

**Counterfactual Explanations -** are different from the methods presented above. The key deviation is that while (19; 18; 14; 16; 2) solutions try to interpret specific predictions, counterfactual explanations aim to justify the given prediction by comparing an instance from the predicted class to one of different category. For image classification task, this would imply generating a counterexample for the given instance, not to describe the prediction, but to explain the behavior of an AI system. The feature deviation to achieve the counterexample is minimal - thus, we can call it perturbations.

In theory, a naive trial and error method can be used to generate the counterfactual explanation, although it is not feasible. However, there are various heuristics to limit the search space (13). By introducting the loss function, where the inputs are the chosen instance and the desired counterfactual explanation, we can use approaches like gradient-based (Score Counterfactual Explanations - SCFE, REVISE), search-based (Casual Counterfactual Explanation with Variational AutoEncoder (VAE) - C-CHVAE, MINT) methods to minimize the given loss (13).

### *2.1.2. Global Model-Agnostic Methods*

There have been attempts to create global model-agnostic explanations (8). Although it is computationally complex task to create a global interpreter capable to deal with the multi-dimensional, large datasets and highly nonlinear models. Complex pixel interactions are also hard to interpret. Theoretically, though, we can apply a global surrogate to target models. The surrogate can be *any* inherently interpretable model (decision trees, linear regression, etc.). First, we choose the dataset from the same distribution on which the model was trained, get the predictions for that dataset, and train the desired model on the given dataset and its predictions (12).

As we approximate the model, we can calculate the similarity measure for the surrogate method. (12) mentions R-squared measure:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(\hat{y_*}^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n}(\hat{y}^{(i)} - \bar{\hat{y}})^2} \tag{2}$$

Where Sum of Squared Errors measure the prediction difference between surrogate and target models, squared sums. Sum of Squared Total is a total variability in the predictions of the target model, $\bar{\hat{y}}$ being the average prediction made by the complex model. We can tell from the high R-squared (close to 1) measure that the surrogate approximates the model well; in contrast, we have a low R-squared value (close to 0).

It is worth to note that these methods will suffer with high-dimensional data, have difficulty capturing nonlinear relationships, approximate one subset of dataset, while diverge for another (12; 10).

**SHAP -** or SHapley Additive exPlanations is a unified approach to interpret model predictions. The concept is inspired by Shapley values from cooperative game theory. The proposed SHAP values are a unified measure of the importance of the features (10). The method is of an additive feature attribution class, satisfying the desired property of *local accuracy*, *consistency*, and redundant *missingness* (10). That means that the method can approximate the original model for a specific input $x$ (10), the input attribution does not decrease if the model change increases the simplified input's contribution regardless of the other inputs (10), and the missing features in the original input have no impact if the simplified inputs represent feature presence (10).

Although theoretically sound, it is very computationally expensive to find the exact SHAP values, as it is needed to compute every feature "contribution" individually, thus the need for the approximation. The tradeoff is that we get better performance (or feasability in many cases) for accuracy.

In the same paper, *Lunderberg et al.* (10) propose **kernel SHAP** to address the problem described above. It estimates the contributions of each feature for the given instance. Using weighted linear regression (LIME-like design), the model can jointly approximate Shapley values with fewer evaluations (10).

**Permutation Feature Importance -** is proposed to score feature's importance in a machine learning model. The idea behind this method is to perturbate or permute the given feature's value and then observe the difference in model's performance by measuring the prediction error. (11). This way, we can rate the feature's importance based on how their permutation affects the model performance - bigger change means higher importance.

PFI is a solid and straightforward approach to measure each feature's contribution in the given instance. It provides a good global interpretation and does not require retraining of the model (11). However, all this comes with high computational complexity, especially on the large multi-dimensional datasets. On top of that, by assuming each feature independent, might not be ideal for some tasks.

## 3. Conclusion

In this survey, we have explored the local and global model-agnostic methods to interpret the decisions made by the non-linear opaque networks. We have discussed the strengths and weaknesses of the models like SHAP, LIME, Anchors, and their derivatives. We showed that interpretability involving high-dimensional datasets like in image classification task is manageable locally but hard to achieve globally. While we have solid theoretical foundation, there are several hurdles to overcome before implementing it in real-life scenarios. High computational complexity, trade-off between performance and accuracy, sensitivity to distribution,

etc. plague this field. Nevertheless, the methods described above, are right step into interpretability, trustworthyness and safety.

## References


[1] ALVAREZ-MELIS, D., AND JAAKKOLA, T. S. On the robustness of interpretability methods, 2018.

[2] COLLARIS, D., GAJANE, P., JORRITSMA, J., VAN WIJK, J. J., AND PECHENIZKIY, M. LEMON: Alternative Sampling for More Faithful Explanation Through Local Surrogate Models. Springer Nature Switzerland, 2023, p. 77–90.

[3] FRIEDMAN, J. H., AND POPESCU, B. E. Predictive learning via rule ensembles. The Annals of Applied Statistics 2, 3 (Sept. 2008).

[4] GIANFAGNA, L., AND DI CECCO, A. Model-Agnostic Methods for XAI. Springer International Publishing, 2021, p. 81–113.

[5] GOLDSTEIN, A., KAPELNER, A., BLEICH, J., AND PITKIN, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics 24, 1 (Jan. 2015), 44–65.

[6] GUIDOTTI, R., MONREALE, A., RUGGIERI, S., TURINI, F., PEDRESCHI, D., AND GIANNOTTI, F. A survey of methods for explaining black box models, 2018.

[7] HARRIS, C., PYMAR, R., AND ROWAT, C. Joint shapley values: a measure of joint feature importance, 2022.

[8] KAMAKSHI, V., AND KRISHNAN, N. C. Explainable image classification: The journey so far and the road ahead. AI 4, 3 (Aug. 2023), 620–651.

[9] LAKKARAJU, H., KAMAR, E., CARUANA, R., AND LESKOVEC, J. Faithful and customizable explanations of black box models. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (Jan. 2019), AIES '19, ACM.

[10] LUNDBERG, S., AND LEE, S.-I. A unified approach to interpreting model predictions, 2017.

[11] LUNDBERG, S. M., ERION, G. G., AND LEE, S.-I. Consistent individualized feature attribution for tree ensembles, 2019.

[12] MOLNAR, C. Interpretable Machine Learning, 2 ed. 2022.

[13] PAWELCZYK, M., AGARWAL, C., JOSHI, S., UPADHYAY, S., AND LAKKARAJU, H. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis, 2021.

[14] RIBEIRO, M. T., SINGH, S., AND GUESTRIN, C. Model-agnostic interpretability of machine learning, 2016.

[15] RIBEIRO, M. T., SINGH, S., AND GUESTRIN, C. "why should I trust you?": Explaining the predictions of any classifier. CoRR abs/1602.04938 (2016).

[16] RIBEIRO, M. T., SINGH, S., AND GUESTRIN, C. Anchors: High-precision model-agnostic explanations. Proceedings of the AAAI Conference on Artificial Intelligence 32, 1 (Apr. 2018).

[17] SAMEK, W., WIEGAND, T., AND MÜLLER, K. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. CoRR abs/1708.08296 (2017).

[18] SHARMA, R., REDDY, N., KAMAKSHI, V., KRISHNAN, N. C., AND JAIN, S. Maire – a model-agnostic interpretable rule extraction procedure for explaining classifiers, 2020.

[19] ZAFAR, M. R., AND KHAN, N. M. Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems, 2019.