universität
ulm

Seminar

Draft Explainable AI

Understanding Machine Learning Models

February 7, 2024

Andreas Lebedev
andreas.lebedev@uni-ulm.de

Supervision by
Prof. Dr. Stephan Schlüter

Contents

List of Figures

List of Tables

1 Introduction

Machine Learning (ML) models excel in various fields, often outperforming human capabilities. However, ML application requires careful consideration, especially in domains where critical decisions are made. Transparency and trustworthiness of models are crucial, for instance in healthcare, where their outcomes directly affect individuals' well-being or even their lives [9].
The field of Explainable Artificial Intelligence (XAI) has emerged as a response to this need, dedicated to exploring ML interpretability to provide transparency and trustworthiness in models. Highlighting the significance of XAI, consider the following examples.

Assume a scenario where you are faced with interpreting mammograms to identify potential lesions and determine their malignancy, thereby determining if a biopsy is necessary for the patient. Let's assume you have developed a Deep Learning model specifically trained for this task, achieving a high performance score on the test dataset. The question arises: Would the radiologist include it in his diagnostic practices? It seems unlikely if the model only provides a yes or no prediction or gives a numerical score without an accompanying explanation describing the reasoning behind the model's decision-making process. Alternatively, using an interpretable model, such as the case-based interpretable deep learning model *IAIA-BL* from [1], could be more advantageous. This model offers a structured explanation framework, that identifies relevant regions, links these regions to specific medical attributes and relies only on the evidence to make predictions [1].

Another scenario where XAI could prove valuable is in credit evaluation. In recent times, the loan acceptance rates of banks have descended, decreasing to 61%–70%, and diving further to 50% post-pandemic due to widespread financial setbacks and a higher rate of defaulters. However, what worsens the situation is the lack of transparent explanations provided to the customers, leaving them frustrated and confused. Providing clear explanations for loan rejections not only offers transparency to customers but also empowers them to understand and potentially modify their financial behavior. This transparency can improve customer satisfaction and trust in the banking system. [11]

Alone from these two examples, the significance of XAI becomes clear. This work aims to offer a brief understanding of XAI and explore various approaches within the field. In Section 2, an overview of XAI is provided, delving into its objectives, the different forms of explanations, and the taxonomy of XAI approaches. Following that, in Sections 3-5, different XAI approaches are presented, highlighting their interpretative capabilities and inherent limitations. Lastly, in Section 6, conclusions are drawn and four key insights derived from the study are presented. Much of the work relies on Molnar's book *Interpretable Machine Learning* [10], serving as a great entry point for those who are interested in interpretable ML. Additionally, in appendix A, brief lookup tables for the most popular XAI approaches are provided, containing key interpretability insights.

2 Definitions and Overview

XAI, short for Explainable Artificial Intelligence, is a term often used in the field of machine learning. It's all about going beyond a mysterious "black box" approach and aiming to understand how a trained model makes decisions.
In this chapter, we'll explore the uses of ML Interpretability. Next, we'll examine the different forms in which explanations can be presented. Finally, we'll categorize the types of machine learning models and algorithms.

2.1 Goals of ML Interpretability

Justify the model

The main aim of XAI is to make models easy to understand so that people can trust them more and feel confident in their predictions. This means making the model transparent so that we can understand how it makes its predictions.

According to [18], it's also important to consider who needs the explanation. Different people play different roles in how machine learning systems work. In a machine learning ecosystem they define six roles, illustrated in Figure 2.1, where individuals may fulfill more than one role simultaneously.

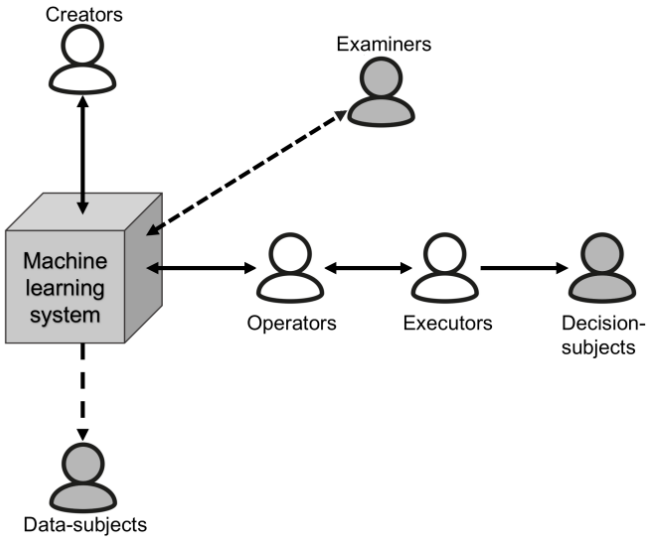

Figure 2.1: Illustration of a machine learning ecosystem.[18]

First, there are the creators who build and develop the machine learning model. Then, there are the operators who manage and run the model once it's implemented. Next, there are the executors who use the model to make decisions. After them, there are those who are impacted by the decisions made by the executors. This includes people whose personal data was used to train the model, as well as those who audit or investigate the machine learning system.

Understanding these different roles helps ensure that explanations are useful and relevant to everyone involved.

Get insights into data

Another aim in XAI is to gain insights from the data. Instead of just focusing on the model, the goal is to understand how the real-world process operates and how the features relate to the target. However, achieving this goal heavily relies on the model's performance. For XAI to be effective in uncovering insights, the model must perform well.[5]

Debug and improve model

One of the aims of XAI is to find and fix mistakes or unfairness in models and data. This involves spotting things like missing values or repeated information in the dataset, shifts in the data over time or finding redundant features that don't add any new useful information.[7, 18]

2.2 Different forms of explanations

XAI approaches can provide various types of explanations.

Features statistics + visualization

One form of explanation involves offering feature statistics, which can often be represented visually. For example, techniques such as Permutation Feature Importance (T. 4.1) or Shapley values (T.4.2) provide importance metrics indicating the influence of each feature on the model's prediction. Certain feature explanations are best understood through visualization, such as the saliency map produced by the Vanilla Gradient method (T.5.1).

Internal parameters of a model

Another form of explanation involves gaining insight through internal model parameters. For instance, this can be observed in the weights of Linear Regression models (T. 3.1), which serve as both feature statistics and components of the model's internal workings. Similarly, the internal structure of Decision Trees (T. A.2) offers valuable insights into how the model makes decisions.

Points from dataset distribution

This form of explanation includes all approaches that uses data points from the existing data distribution or newly generated data points to improve model interpretability. For example, the model utilizes available data points from the dataset to make predictions, like with the ProtoPNet architecture discussed in subsection 3.2.

Approximation of model with an inherently interpretable model

Another form of explanation involves approximating a complex model with an interpretable one, either a globally using a global surrogate (T. A.7), or locally as seen in methods like LIME (T. A.10). This interpretable model can then be analyzed by examining its internal parameters or interpreted by obtaining feature statistics.

2.3 Taxonomy of Explainable AI Approaches

The taxonomy of XAI approaches is illustrated in the following figure:

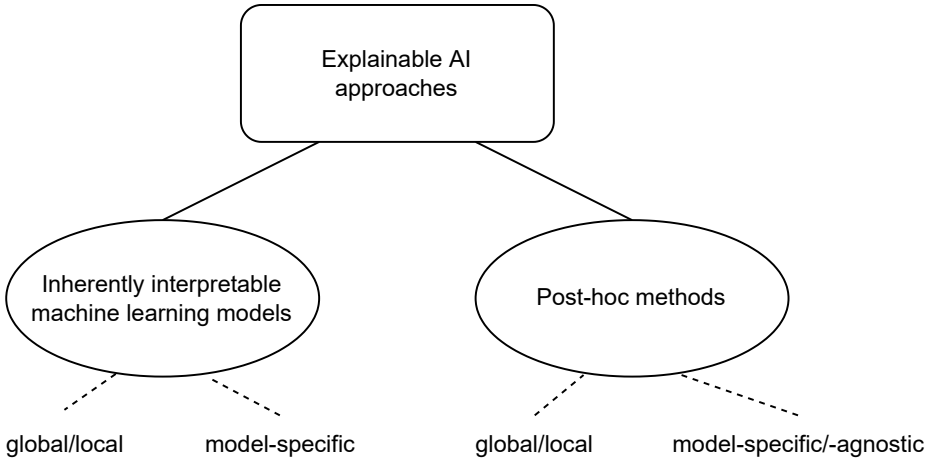

Figure 2.2: Overview over XAI.

XAI approaches can be divided into two main types. Firstly, there are **inherently interpretable ML models**, characterized by either a straightforward structure that is easy to interpret or models where interpretability is integrated into the architecture itself. Secondly, there are **Post hoc Methods**, which are applied to trained models.
An important distinction to note is whether the approach entails global or local interpretability. **Global interpretability** aims to comprehend the overall behavior of the model in making predictions. For instance, in a Linear Regression model, insights can be derived by analyzing the weights, while in a Decision Tree, one can analyze the nodes to understand the prediction mechanism. On the other hand, **local interpretability** aims to understand the reasoning behind individual or grouped predictions.
Furthermore, another important distinction is whether an approach is model-specific or model-agnostic. **Model-specific** approaches work only with particular model classes. Inherently interpretable ML models fall into this category by design. Additionally, post-hoc methods can also be model-specific, such as the Vanilla Gradient method (T. 5.1), which is only applicable to models allowing gradient computation. Conversely, **model-agnostic** approaches are post-hoc methods capable of being applied to any model.

3 Inherently interpretable machine learning models

Some models are really easy to understand because they're simple, like Linear Regression. With these models, it's straightforward to see how they make predictions. Then there are models where interpretability is built right into their design, like ProtoPNet. This means that even though they might be more complex, they're designed in a way that makes them easier to interpret.

3.1 Linear Regression

Form of explanation	Internal model parameters
Interpretation	An increase or change of feature x_k by one unit increases the prediction for y by β_k units when all other feature values remain fixed.
Advantages	• transparent • solid statistical history • extensions: e.g. GLM, GAM (less interpretable)
Disadvantages	• can only represent linear relationships • not suitable if features correlations to strong • interactions must be added manually
Properties	intrinsic, model-specific, global

Table 3.1: Overview XAI - Linear Regression.

Linear regression models have a simple structure and often offer a clear and understandable explanation of how the input variables influence the output. In certain scenarios, they can even outperform more complex nonlinear models for prediction tasks. The linear regression model is mathematically represented as follows:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{3.1}$$

where $X^T = (X_1, \ldots, X_p)$ denotes the input vector. One of the most commonly used techniques for estimating the parameters β_i is called *least squares*. In this method, the objective is to minimize the sum of the squared differences between observed and predicted values:

$$\min_{\beta} \sum_{t_i}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_j \beta_j \right)^2 \tag{3.2}$$

where $(x_1, y_1) \ldots (x_N, y_N)$ represent the training data pairs.[8]

Because the combination of features is linear, their effects are additive, allowing for easy separation. An increase of one unit in a feature results in a change in the estimated outcome by the weight assigned to that feature.
Figure 3.1 presents a visual representation of interpretability using data from a diabetes dataset[3]:

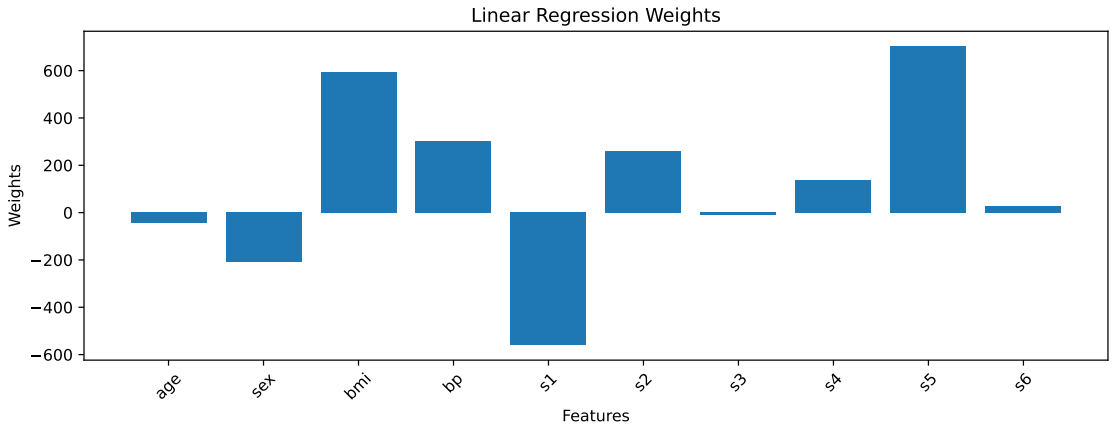

Figure 3.1: Weights obtained from a Linear Regression model trained on a diabetes dataset. *bp* means blood pressure, while $s_1, \ldots, s_6$ represent various laboratory values.

We observe that the variables s_5, s_1, and *bmi* significantly influence the outcome of the model.

However, when dealing with a larger number of features, such as thousands, interpreting the model becomes more challenging. To address this, techniques like variable subset selection or shrinkage methods are used to exclude features with minimal impact on the model's output.
A commonly used shrinkage method is the Lasso regression. This approach introduces a regularization term into the linear regression equation 3.2 to penalize the coefficients of less influential features:

$$\min_{\beta} \sum_{t_i}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_j \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{3.3}$$

Here, $\lambda > 0$ represents the regularization parameter.[8]

Figure 3.2 shows a trace plot of coefficients fit by lasso regression, where the regularization parameter λ varies within the range of 10^{-4} to 10^{1}, applied to the diabetes dataset and demonstrates the relationship between λ and the coefficients, as well as the resulting mean squared error:

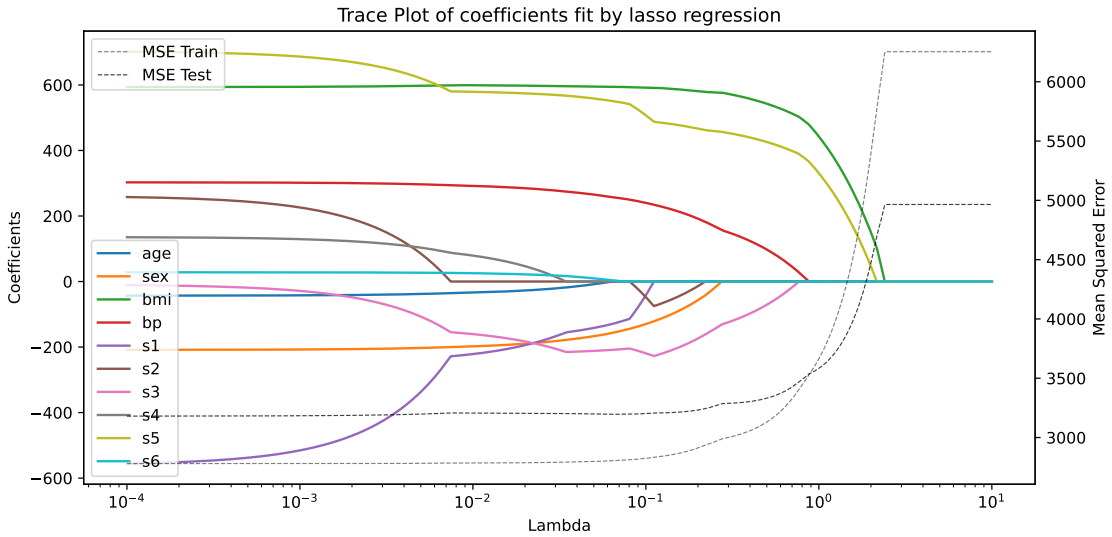

Figure 3.2: Trace Plot of coefficients fit by lasso regression.

When we observe a value of λ slightly greater than 10^{-1}, we notice that the mean squared error (MSE) on the test dataset stopped decreasing. At this point, the model employs only six features, making it simpler to interpret.
It's noteworthy that s_1 is not among these six features, despite its significant influence on the model output without the regularization term, as displayed in Figure 3.1 (or at $\lambda = 10^{-4}$ in Figure 3.2). This observation encourages the consideration of the limitations inherent in linear regression.

Linear regression is limited in its ability to represent complex relationships as it can only model linear relationships. Nonlinear interactions between variables need to be explicitly defined and provided to the model as input features.
Additionally, linear regression encounters challenges with correlated features. In cases where features are highly correlated, the increase of one feature by one unit might not accurately correspond to a proportional increase in the prediction for the output by its associated weight, as the model assumes all other feature values remain constant. This can lead to unintuitive weights in the model.[10]

3.2 ProtoPNet - *This* Looks Like *That*

Form of explanation	Points from dataset distribution.
Interpretation	Learned prototypical parts are compared to points from training dataset.
Advantages	• can achieve comparable accuracy with its analogous non-interpretable counterpart • provides a level of interpretability that is absent in other interpretable deep models
Disadvantages	• (minimal) trade-off in accuracy to best-performing deep models
Properties	intrinsic, model-specific, local

Table 3.2: Overview XAI - ProtoPNet.[2]

In this study [2], the authors have established a form of interpretability in image processing, which aligns with how humans naturally explain their reasoning in classification tasks. They introduce a model called the *prototypical part network* (ProtoPNet). The architecture of the network is shown in the following figure:

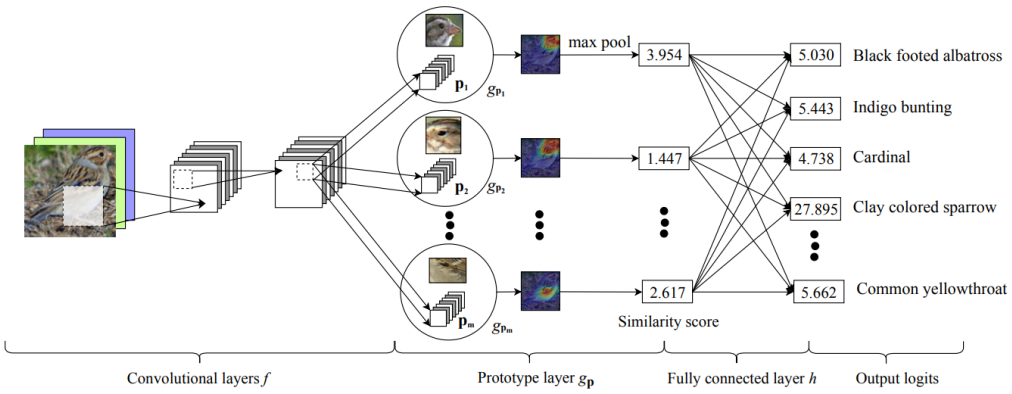

Figure 3.3: ProtoPNet Architecture.[2]

This network is designed to break down training images into prototypical parts that represent common identifiable prototypes, such as the head, beak, or wings, which are characteristics of a specific class, like the clay-colored sparrow. Figure 3.4 shows a demonstration:

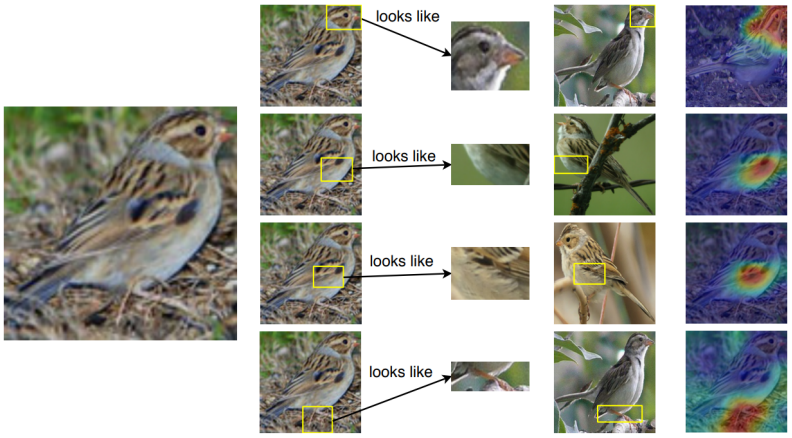

Figure 3.4: Image of a clay colored sparrow and how parts of it look like some learned prototypical parts of a clay colored sparrow used to classify the bird's species.[2]

The model then calculates a weighted sum of the similarity between different parts of an input image and the learned prototypes for each class.

The experiments demonstrate that ProtoPNet can achieve similar accuracy to its comparable non-interpretable counterpart without the Prototype layer. Additionally, ProtoPNet offers a level of interpretability that is lacking in other interpretable deep models.

4 Model-Agnostic Post-hoc Methods

4.1 PFI - Permutation Feature Importance

Form of explanation	Feature statistics
Interpretation	A feature is "important" if shuffling its values increases the model error.
Advantages	• provides a highly compressed, global insight into the model's behavior • takes into account all interactions with other features
Disadvantages	• when the permutation is repeated, the results might vary greatly • can be biased by unrealistic data instances • adding a correlated feature can decrease the importance of the associated feature by splitting the importance between both features
Properties	post-hoc, model-agnostic, global

Table 4.1: Overview XAI - Permutation Feature Importance.

Permutation Feature Importance (PFI) is a method that measures how important a feature is to the model by evaluating how the model's prediction error changes when that feature's values are randomly shuffled. The idea is that if a feature holds importance for the model, randomly rearranging its values will lead to a high increase in prediction error, which would indicate the model's dependency on the feature for accurate predictions.
The choice between using the training or test dataset to calculate the PFI depends on what goal you have. If you aim to measure the extent to which the model depends on a particular variable for making predictions, you should use the training data. On the other hand, if your goal is to evaluate the contribution of a feature to the model's performance on unseen data, the test data should be used.

To compute the Permutation Feature Importance (PFI) as outlined in [4], the process involves several steps. First, you compute the model error $e(X)$, on the dataset X. Then, for each feature, you shuffle that feature n times to create n permuted datasets $X^p_{perm_i}$ and calculate the model error $e(X^p_{perm_i})$ for each permuted dataset. Finally, to derive the feature importances, you subtract these errors from the original model error: $e(X) - e(X^p_{perm_i})$.

Figure 4.1 presents a visualization of PFI using Linear Regression on training and test set of a diabetes dataset:

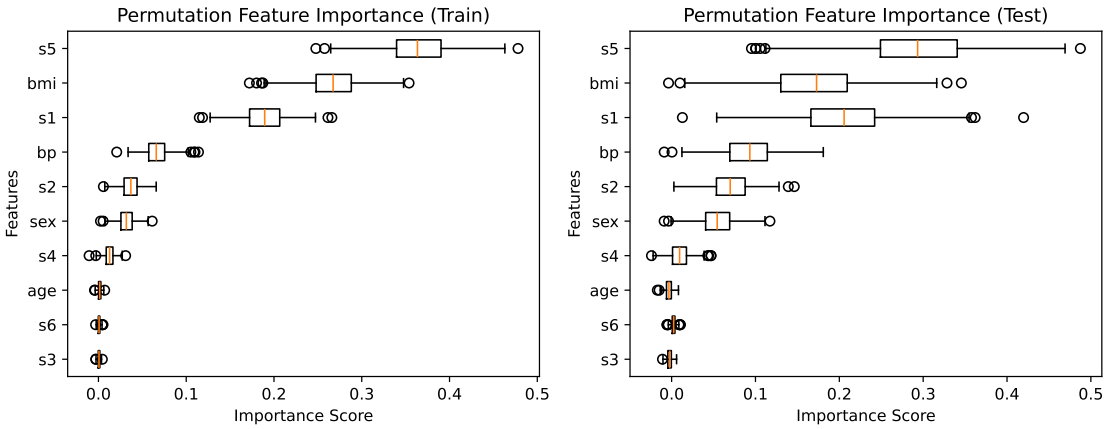

Figure 4.1: Visualization of Permutation Feature Importance using Linear Regression on training and test set of a diabetes dataset.

The PFI calculated on the training set aligns with the weights of the linear regression model (Figure 3.1) as expected. On the test set, the computed PFI indicates that when applying the model to new data, the feature *BMI* is less important than feature s_1.

As observed, Permutation Feature Importance (PFI) provides a clear and comprehensive understanding of how the model behaves globally. Additionally, PFI considers not only the primary impact of features but also their interactions, as it disrupts these interactions when permutating individual features.
However, there are some drawbacks to consider. There are $n!$ ways to permute a vector. For instance, with just 100 data points, there are approximately 9.33×10^{157} different permutation possibilities. This exponential increase in computational time forces choosing a fixed number of permutations to approximate the overall PFI. Consequently, the results may vary greatly each time PFI is repeated. Another issue is that permuting a feature can lead to unrealistic data points, especially when two features are strongly correlated. Furthermore, adding a correlated feature can decrease the importance of the original correlated features by splitting the importance between them. This scenario can lead to a misleading importance ranking of features, potentially causing incorrect interpretations of a feature's relevance in the model.

4.2 Shapley Values

Form of explanation	Feature statistics.
Interpretation	Given the current set of feature values, the contribution of a feature value to the difference between the actual prediction and the mean prediction is the estimated shapley value.
Advantages	• prediction is fairly distributed among the feature values • has a solid theoretical foundation in game theory • can be used for global model interpretations
Disadvantages	• requires a lot of computing time • problem with correlated features, can include unrealistic data instances • shapley values can be misinterpreted • no sparse explanation, always use all the features
Properties	post-hoc, model-agnostic, local

Table 4.2: Overview XAI - Shapley Values.

Shapley values, first introduced by Shapley [13] in 1953, present a method from coalitional game theory. This approach aims to assign fair payouts to players based on their individual contributions to the overall payout within a game. When applied to machine learning, shapley values attribute the prediction of a model to its individual features. This is achieved by considering all potential combinations of features and evaluating how the prediction changes when each feature is included.
A shapley value can be expressed as follows:

$$\phi_j(val) = \sum_{S \subseteq \{1,\dots,p\} \setminus \{j\}} \overbrace{\frac{|S|!(p-|S|-1)!}{p!}}^{\text{Weight}} \underbrace{\left(val(S \cup \{j\}) - val(S)\right)}_{\text{Marginal Contribution of Feature } j \text{ to Coalition } S} . \quad (4.1)$$

Here, S represents a subset of features used in the model, p indicates the total number of features, and *val* represents the prediction for feature values within set S, marginalized over features not included in set S.
To illustrate this more clearly, let's apply it to a Linear Regression model: $f(x) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j$. The contribution ϕ_j of the j-th feature to the prediction $\hat{f}(x)$ can be expressed as

$$\phi_j(\hat{f}) = \beta_j x_j - E(\beta_j X_j) \quad (4.2)$$

In this equation, $E(\beta_j X_j)$ represents the mean effect estimate for feature j. The contribution, therefore, is calculated as the difference between the effect of the feature and its

average effect.
Summing up all shapley values results in the predicted value for the data point x minus the average predicted value:

$$\begin{aligned}\sum_{j=1}^{p} \phi_j(\hat{f}) &= \sum_{j=1}^{p} (\beta_j x_j - E(\beta_j X_j)) = (\beta_0 + \sum_{j=1}^{p} \beta_j x_j) - (\beta_0 + \sum_{j=1}^{p} E(\beta_j X_j)) \\ &= \hat{f}(x) - E(\hat{f}(x)).\end{aligned} \tag{4.3}$$

In Figure 4.2, the shapley values for a single prediction from a diabetes dataset are displayed:

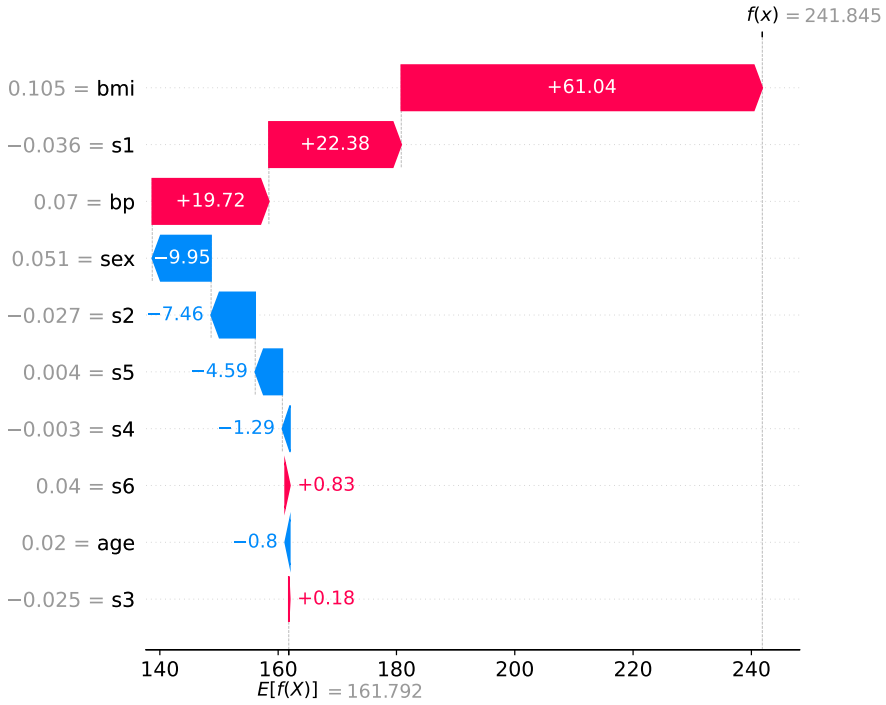

Figure 4.2: Visualizing shapley values through a Waterfall Plot with the SHAP library using Linear Regression on a test data point of a diabetes dataset.

We observe that the *bmi* feature significantly influenced this particular prediction. However, despite the large weight of the feature $s5$ in the model, its impact appears comparatively lower than that of other features.

Aggregating all shapley values into a single figure can provide us with comprehensive insights into the model's global behavior, as illustrated in the following figure:

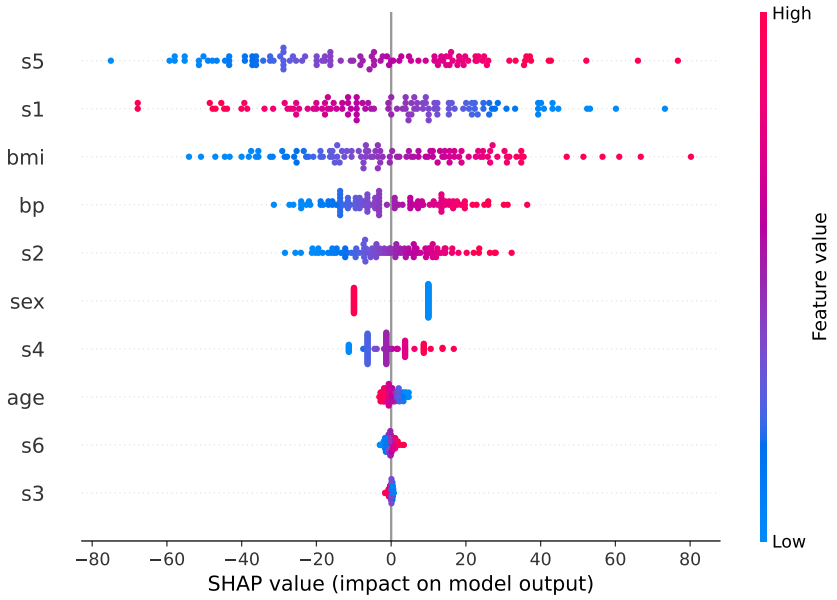

Figure 4.3: Visualizing shapley values through a Summary Plot with the SHAP library using Linear Regression on a test dataset of a diabetes dataset.

In addition to this, various other shapley value global interpretation methods exist, delivering feature importance values, feature dependence analysis, exploration of interactions, and clustering techniques.

The problem with shapley values lies in their computational complexity, which grows exponentially as more features are included. This expansion occurs due to the exponential increase in possible coalitions with each additional feature. Moreover, using more complex, non-linear models further complicates the computation. In addressing this challenge [16] proposed an approximation technique using Monte Carlo sampling. Another issue arises when dealing with correlated features, which can lead to the inclusion of unrealistic data instances. Additionally, there is a potential for misinterpretation, wherein shapley values might be mistakenly perceived as the difference in predicted values after removing a feature from the model training. A less desirable aspect of shapley values is their inherent requirement to use all features.

5 Model-Specific Post-hoc Methods

Model-specific post-hoc methods are limited to a specific type of model. In the following section, we examine the vanilla gradient method, which is only suitable for models that allow gradient computation. For instance, this method is not feasible for decision trees due to their inherent structure.

Image-Specific Class Saliency (Vanilla gradient)

Form of explanation	Feature visualization
Interpretation	Pixels that strongly influence the prediction of the neural network are identified.
Advantages	• easy to immediately recognize the important regions of the image to the model's prediction
Disadvantages	• tells only where the network is looking not why the prediction was made • can be highly unreliable
Properties	post-hoc, model-specific, local

Table 5.1: Overview XAI - Image-Specific Class Saliency (Vanilla gradient).

The Vanilla gradient, also known as "Image-Specific Class Saliency" by its authors, falls within the category of gradient-based pixel attribution techniques. Saliency maps generated by these methods highlight the pixels considered significant by a neural network during the process of image classification.
The computation of a saliency map using the Vanilla Gradient is relatively straightforward. Instead of calculating the gradient of the output regarding the network's weights, as done in typical backpropagation, the gradient of the output regarding the input image is computed: $\frac{\delta f(I;W)}{\delta I}$. Therefore, each pixel in the image has a corresponding gradient value, which can afterward be used for visualization. Pixels with higher absolute gradient values are considered more important in influencing the network's decision. An example of the Vanilla gradient method is shown in the following figure:

Figure 5.1: Image-specific class saliency maps for the top-1 predicted class in ILSVRC-2013 test images. The maps were extracted using a single back-propagation pass through a classification ConvNet.[14]

As you can see the interpretation is easy and by performing only one backpropagation step the computation is fast and efficient.

However, what meaningful conclusions can one derive from the highlighted pixels? [6] demonstrates that minor perturbations to the input image result in nearly identical saliency maps, yet the neural network may classify the output differently. For instance, as illustrated in Figure 5.2, an image of a husky was misclassified as a flute, despite having very similar saliency maps.

Figure 5.2: Image is labeled as a dog and a musical instrument when the saliency maps look essentially the same. [6]

So while the saliency map provides insight into the regions of focus within a neural network, it fails to explain the reason behind the network's specific prediction[12].

[6] also demonstrated that saliency maps display considerable unreliability. Even minor perturbations to an image can lead to substantial changes in the saliency map, despite the network's prediction remaining the same. The example shown in Figure 5.3 illustrates this phenomenon using three images:

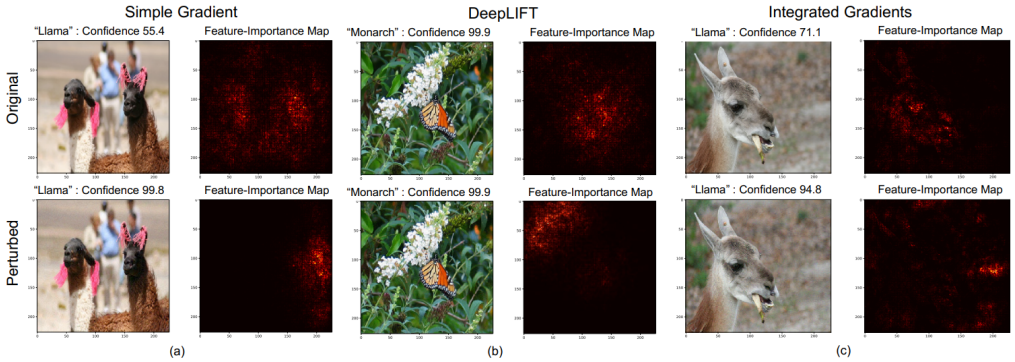

Figure 5.3: Adversarial attack against feature-importance maps. The top row shows the the original images and their saliency maps and the bottom row shows the perturbed images and the corresponding saliency maps.[6]

In each of the three images, the predicted classification remains the same despite the introduced perturbation. However, the saliency maps of the perturbed images have a notable shift towards pixels that typically wouldn't be considered as influential according to human perception.

6 Conclusion

Overall, there is no singular optimal method for explaining models or constructing interpretable ones. While XAI has gained significant attention, it is crucial to acknowledge the inherent limitations of each approach and refrain from using them blindly. When used properly, XAI can yield substantial advantages. However, it remains an area of active research.

In conclusion, I would like to provide four key takeaways that I have learned from my study.

1. The selection of the XAI approach is goal and risk-specific.

In Section 2.1, we discussed the various goals associated with XAI. Different XAI approaches may be favored depending on the specific goal pursued.
Additionally, the choice of an approach depends on the context in which machine learning is applied and the particular task at hand. For instance, in environments where there are high risks involved and critical decisions must be made, it is advisable to use inherently interpretable models, as suggested by [12].

2. For the same model different explanations for different people.

As mentioned in Subsection 2.1, the audience you're addressing significantly influences the approach you should take in providing explanations. Certain techniques may not be appropriate for individuals with limited knowledge of machine learning. On the other hand, more complex approaches can be employed for experts in the field.

3. Many XAI methods have problems with correlated features.

As observed in techniques such as Linear Regression or Permutation Feature Importance, various methods are sensitive to the presence of correlated features, which can greatly affect the result of the explanations provided. To address this issue, one can use feature engineering techniques to reduce feature correlation before training. This can involve either introducing new features or excluding highly correlated ones.

4. No linear relationship between accuracy and interpretability.

Some people may believe that highly complex models offer the best accuracy, assuming that a complex "black box" approach is essential for achieving top predictive performance. However, this isn't always the case, especially when dealing with well-structured data containing naturally meaningful features. Figure 6.1 displays the misconception:

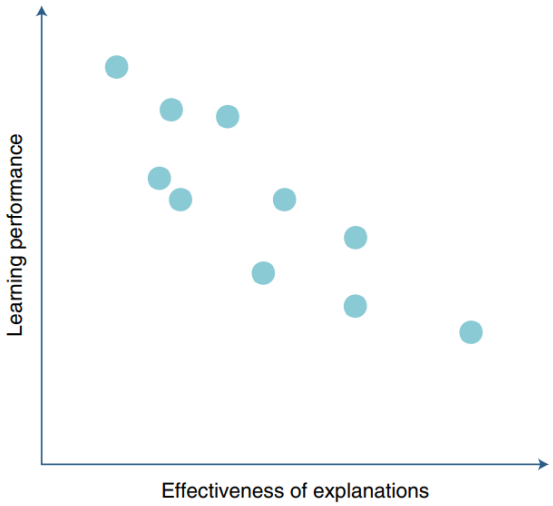

Figure 6.1: A fictional depiction of the accuracy–interpretability trade-off.[12]

According to [12], when addressing problems with structured data featuring significant features, there's often no notable difference in performance between complex models like deep neural networks and much simpler ones like logistic regression.

References

[1] A. J. Barnett et al. “A case-based interpretable deep learning model for classification of mass lesions in digital mammography”. In: *Nature Machine Intelligence* 3.12 (2021), pp. 1061–1070.

[2] C. Chen et al. “This Looks Like That: Deep Learning for Interpretable Image Recognition”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.

[3] *Diabetes Dataset*. Accessed on: February 7, 2024. URL: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html#sklearn-datasets-load-diabetes.

[4] A. Fisher, C. Rudin, and F. Dominici. “All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously.” In: *J. Mach. Learn. Res.* 20.177 (2019), pp. 1–81.

[5] T. Freiesleben et al. *Scientific Inference With Interpretable Machine Learning: Analyzing Models to Learn About Real-World Phenomena*. 2022. arXiv: 2206.05487 [stat.ML].

[6] A. Ghorbani, A. Abid, and J. Zou. “Interpretation of Neural Networks Is Fragile”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (2019), pp. 3681–3688.

[7] V. Hassija et al. “Interpreting black-box models: a review on explainable artificial intelligence”. In: *Cognitive Computation* 16.1 (2024), pp. 45–74.

[8] T. Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[9] A. F. Markus, J. A. Kors, and P. R. Rijnbeek. “The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies”. In: *Journal of Biomedical Informatics* 113 (2021), p. 103655.

[10] C. Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed. 2022. URL: https://christophm.github.io/interpretable-ml-book.

[11] M. Nallakaruppan et al. “An Explainable AI framework for credit evaluation and analysis”. In: *Applied Soft Computing* 153 (2024), p. 111307.

[12] C. Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature machine intelligence* 1.5 (2019), pp. 206–215.

[13] L. S. Shapley et al. “A value for n-person games”. In: (1953).

[14] K. Simonyan, A. Vedaldi, and A. Zisserman. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. 2014. arXiv: 1312.6034 [cs.CV].

[15] D. Slack et al. “Fooling lime and shap: Adversarial attacks on post hoc explanation methods”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 180–186.

[16] E. Štrumbelj and I. Kononenko. “Explaining prediction models and individual predictions with feature contributions”. In: *Knowledge and information systems* 41 (2014), pp. 647–665.

[17] *The Best of Both Worlds: Linear Model Trees*. Accessed on: February 7, 2024. URL: https://medium.com/convoy-tech/the-best-of-both-worlds-linear-model-trees-7c9ce139767d.

[18] R. Tomsett et al. “Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems”. In: *CoRR* (2018).

A Overview XAI Approaches

A.1 Inherently interpretable machine learning models

A.1.1 Linear Regression

Form of explanation	Internal model parameters
Interpretation	An increase or change of feature x_k by one unit increases the prediction for y by β_k units when all other feature values remain fixed.
Advantages	• transparent • solid statistical history • extensions: e.g. GLM, GAM (less interpretable)
Disadvantages	• can only represent linear relationships • not suitable if features correlations to strong • interactions must be added manually
Properties	intrinsic, model-specific, global

Table A.1: Overview XAI - Linear Regression.

A.1.2 Decision Tree

Form of explanation	Internal model parameters
Interpretation	If feature x is [smaller/bigger] than threshold c AND ... then the predicted outcome is the mean value of y of the instances in that node.
Advantages	• captures interactions between features in the data • interpretation easy, ends up in distinct groups, good explanations
Disadvantages	• fails to deal with linear relationships • lack of smoothness • unstable, few changes in the training dataset can create a completely different tree • number of terminal nodes increases quickly with depth, which makes trees more difficult to understand
Properties	intrinsic, model-specific, global

Table A.2: Overview XAI - Decision Tree.

A.1.3 Linear Tree

Form of explanation	Internal model parameters
Interpretation	Combination of Decision Tree and Linear Regression. If feature x is [smaller/bigger] than threshold c AND ... then the predicted outcome is the prediction of a Linear Regression model where the weights are interpretable.
Advantages	• interpretation easy • get insights into linear and non-linear relationships • identify subpopulations with different behavior
Disadvantages	• as the number of terminal nodes gets higher, makes Linear Tree more difficult to understand
Properties	intrinsic, model-specific, global

Table A.3: Overview XAI - Linear Tree.[17]

A.1.4 ProtoPNet - *This* Looks Like *That*

Form of explanation	Points from dataset distribution.
Interpretation	Learned prototypical parts are compared to points from training dataset.
Advantages	• can achieve comparable accuracy with its analogous non-interpretable counterpart • provides a level of interpretability that is absent in other interpretable deep models
Disadvantages	• (minimal) trade-off in accuracy to best-performing deep models
Properties	intrinsic, model-specific, local

Table A.4: Overview XAI - ProtoPNet.[2]

A.2 Post-hoc methods - Model-Agnostic - global

A.2.1 PFI - Permutation Feature Importance

Form of explanation	Feature statistics
Interpretation	A feature is important if shuffling its values increases the model error.
Advantages	• provides a highly compressed, global insight into the model's behavior • takes into account all interactions with other features
Disadvantages	• when the permutation is repeated, the results might vary greatly • can be biased by unrealistic data instances • adding a correlated feature can decrease the importance of the associated feature by splitting the importance between both features
Properties	post-hoc, model-agnostic, global

Table A.5: Overview XAI - Permutation Feature Importance.

A.2.2 PDP - Partial Dependence Plot

Form of explanation	Features statistics
Interpretation	Feature effect method.
Advantages	• clear interpretation
Disadvantages	• does not show the feature distribution • it is assumed that the feature(s) for which the partial dependence is computed are not correlated with other features • heterogeneous effects might be hidden
Properties	post-hoc, model-agnostic, global

Table A.6: Overview XAI - Partial Dependence Plot (PDP).

A.2.3 Global surrogate

Form of explanation	Approximation with an inherently interpretable model
Interpretation	The original model is replaced with a simpler model for interpretation.
Advantages	• any interpretable model can be used • with the R-squared measure, we can easily measure how good our surrogate models are in approximating the black box predictions
Disadvantages	• you conclude the model and not about the data • what is the best cut-off for R-squared? • could happen that the interpretable model is very close for one subset of the dataset but widely divergent for another subset
Properties	post-hoc, model-agnostic, global

Table A.7: Overview XAI - Global surrogate.

A.3 Post-hoc methods - Model-Agnostic - local

A.3.1 Shapley Values

Form of explanation	Feature statistics.
Interpretation	Given the current set of feature values, the contribution of a feature value to the difference between the actual prediction and the mean prediction is the estimated Shapley value.
Advantages	• prediction is fairly distributed among the feature values • has a solid theoretical foundation in game theory • can be used for global model interpretations
Disadvantages	• requires a lot of computing time • no sparse explanation, always use all the features • problem with correlated features, can include unrealistic data instances
Properties	post-hoc, model-agnostic, local

Table A.8: Overview XAI - Shapley Values.

A.3.2 ICE - Individual Conditional Expectation

Form of explanation	Feature visualization.
Interpretation	Shows how the prediction of one instance changes when a feature changes.
Advantages	• intuitive to understand • can uncover heterogeneous relationships
Disadvantages	• problem with correlated features, some points in the lines might be invalid data points • when many ICE curves are plotted, the visual representation may become chaotic
Properties	post-hoc, model-agnostic, local

Table A.9: Overview XAI - Individual Conditional Expectation (ICE)

A.3.3 LIME - Local Surrogate.

Form of explanation	Feature visualization.
Interpretation	Train a local surrogate model to explain individual prediction.
Advantages	• you can use interpretable models to provide explanations without necessarily using them for prediction purposes • local surrogate models can use features beyond those used in the original model during training.
Disadvantages	• for each application you have to try different kernel settings and see for yourself if the explanations make sense • data points are sampled from a Gaussian distribution, ignoring the correlation between features, which can lead to unlikely data points • by repeating the sampling process, the resulting explanations can vary • LIME explanations can be manipulated to hide biases[15]
Properties	post-hoc, model-agnostic, local

Table A.10: Overview XAI - Individual Local Surrogate (LIME).

A.4 Post-hoc methods - Model-Specific

A.4.1 Image-Specific Class Saliency (Vanilla gradient)

Form of explanation	Feature visualization
Interpretation	Pixels that strongly influence the prediction of the neural network are identified.
Advantages	• easy to immediately recognize the important regions of the image to the model's prediction
Disadvantages	• tells only where the network is looking not why the prediction was made • can be highly unreliable
Properties	post-hoc, model-specific, local

Table A.11: Overview XAI - Image-Specific Class Saliency (Vanilla gradient).