IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 1, April 2024

Explainable AI (XAI): History, Basic Ideas and Methods

Prachi Zodage, Hussain Harianawala, Hafsa Shaikh, Asad Kharodia
M.H. Saboo Siddik College of Engineering, Mumbai, Maharashtra, India

Abstract*: Explainable Artificial Intelligence (XAI) is a field that aims to make artificial intelligence (AI) processes more transparent, explainable, and understandable. As AI processes become more complex, the need to reveal the "black box" nature of these models and provide explanations for their results and decisions is increasing. XAI aims to bridge the gap between the opaque inner workings of AI and human understanding by creating AI that is accurate, useful, and can explain reasoning and decision-making processes in ways that humans can understand. XAI's importance stems from several factors. First, it addresses trust and accountability issues in AI systems, particularly in high-risk sectors like healthcare, finance, and technology. By providing explanations for AI decisions, stakeholders can better understand the logic behind them, detect inconsistencies, and ensure moral and administrative compliance. Second, XAI encourages collaboration and decision-making between people and intelligence, allowing experts and decision-makers to use their knowledge and experience to make better decisions. Thirdly, XAI plays a crucial role in modeling, debugging, and continuous improvement by identifying flaws, biases, or inconsistencies and working to improve performance standards and reliability. Various methods and techniques are used in XAI, each with their own advantages and limitations. Model-free explanations such as LIME, Anchor and SHAP are particularly important because they can be applied to any AI model, regardless of its design or complexity.*

Keywords: Explainable Artificial Intelligence, intelligible machine learning, Interpretability

I. INTRODUCTION

1.1 History of Explainable AI (XAI)

The evolution of Explainable AI (XAI) is deeply intertwined with the development of artificial intelligence itself. In the early days of AI research, from the 1950s to the 1970s, the focus was primarily on symbolic reasoning systems and rule-based approaches, which inherently provided transparency and interpretability. However, as machine learning and statistical methods gained prominence in the 1980s and 1990s, AI systems began to rely more heavily on complex models like neural networks and support vector machines. These models, while effective in solving complex tasks, operated as black boxes, obscuring their decision-making processes. As AI applications proliferated in the 2000s and beyond, concerns regarding the lack of transparency and interpretability became increasingly prominent. This led to the emergence of XAI as a distinct field of study in the 2010s, with researchers across various disciplines endeavoring to develop methods and techniques to make AI systems more transparent, interpretable, and accountable. The history of XAI is marked by significant milestones, including the development of post-hoc interpretation methods and inherently interpretable models, as well as advancements in neural network interpretability. Today, XAI techniques find applications across diverse domains, from healthcare to finance, offering insights into AI-driven decision-making processes and enhancing trust and accountability. However, challenges remain, including balancing interpretability with accuracy and ensuring the ethical deployment of XAI techniques. Looking ahead, interdisciplinary collaboration and continued research will be critical in advancing the field of XAI and addressing these challenges to realize its full potential in creating transparent and trustworthy AI systems.

1.2 Overview

Explainable AI (XAI) emerges as a transformative approach to address the opacity of AI systems, aiming to shed light on the decision-making processes and provide human-understandable explanations for AI predictions. At its core, XAI

DOI: 10.48175/IJARSCT-16988

seeks to bridge the gap between the complexity of advanced machine learning algorithms and the need for transparency, interpretability, and trustworthiness in AI-driven systems. The significance of XAI lies in its potential to democratize AI, empowering stakeholders, including domain experts, regulators, and end-users, to understand and validate the decisions made by AI models. By elucidating the factors influencing AI predictions and revealing the underlying decision logic, XAI enables individuals to assess the reliability of AI systems, identify potential biases or errors, and take informed actions based on AI-generated insights. XAI encompasses a diverse array of methodologies and techniques designed to render AI models more transparent and interpretable. These methodologies range from post-hoc interpretation methods, such as feature importance analysis and model-agnostic explanations, to inherently interpretable models, including decision trees (as illustrated in Fig. 1, starting at the top and going down, the solution path in the decision tree presents the reasoning of a final decision.), rule-based systems, and sparse linear models. Additionally, recent advancements in neural network interpretability, such as attention mechanisms and adversarial robustness techniques, have expanded the scope of XAI to complex deep learning models.

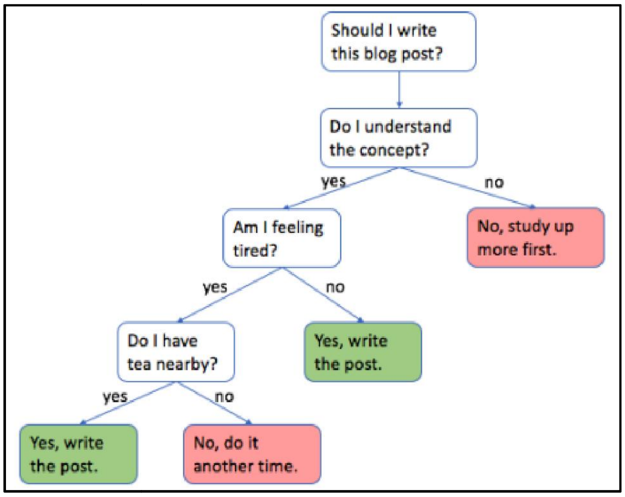


Fig. 1. An example of a decision tree, used by starting at the top and going down, level by level, according to the defined logic. (Image courtesy of J. Jordan [10])

II. WORKFLOW FOR DEVELOPING EXPLAINABLE AI (XAI) APPLICATIONS

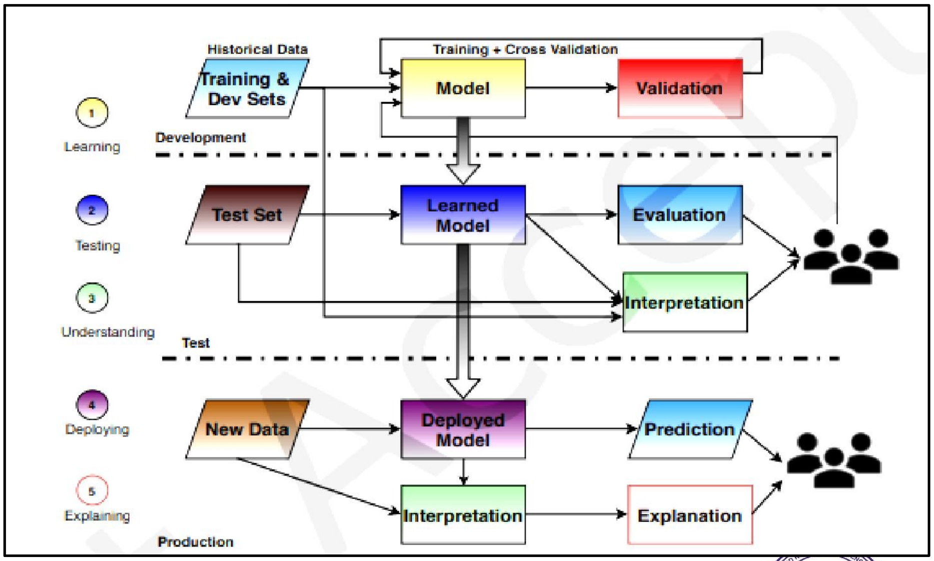


Fig. 2. Illustrates a revised ML life-cycle with the additional steps [6]

While an ML pipeline can provide accurate predictions, it lacks two important phases: understanding and explaining. The Understanding phase involves the training and quality assurance of an AI model whereas the Explaining phase is important when an ML model is deployed and used in real-world applications.

2.1 Understanding Phase.

This phase occurs during the testing stage, after the model has been trained and evaluated on a test set. The understanding phase involves applying XAI techniques to gain insights into the learned model's decision-making process and understand the reasoning behind its predictions on the test set.

The understanding phase is crucial for:

- Model Validation: By interpreting the model's reasoning on the test set, developers can identify potential flaws, biases, or inconsistencies in the model's behavior. This allows for further model refinement or selection of better-performing models.
- Explainability: Understanding how the model arrives at its predictions lays the foundation for developing effective explanations that can be provided to stakeholders during the deployment phase.
- Human-AI Collaboration: Insights gained from the understanding phase can facilitate human-AI collaboration, allowing domain experts to leverage their knowledge and experience in conjunction with the AI model's predictions and reasoning.

2.2 Explaining Phase.

This phase occurs during the production stage, when the deployed model is used to make predictions on new, unseen data. The explaining phase involves applying XAI techniques to generate explanations for the model's predictions, aiming to provide transparency and interpretability to stakeholders (represented by the human figures in Fig 2).

The explaining phase is crucial for:

- Trust and acceptance: By providing explanations for the model's predictions, stakeholders (e.g., domain experts, decision-makers, end-users) can better understand the rationale behind the decisions, fostering trust and acceptance of the AI system.
- Accountability and fairness: Explanations help identify potential biases or unfair decisions made by the AI model, enabling corrective actions and ensuring adherence to ethical principles and regulatory requirements.
- Human oversight: Interpretable explanations facilitate human oversight and decision-making, allowing stakeholders to incorporate their domain knowledge and contextual understanding in the final decision-making process.
- Continuous improvement: Feedback from stakeholders based on the provided explanations can be used to further refine and improve the AI model over time.

III. BLACK-BOX VS WHITE-BOX TECHNIQUES

Black-box models are non-transparent in nature whereas white-box models are straightforward and comparatively simple to understand.The black-box demonstrate is too named as intrinsic as it is accomplished by restricting the complexity of an AI model, whereas a white-box model is moreover named as post-hoc as it is applied on the model after training.

3.1 Black-Box model Techniques.

Black box models such as neural networks or complex ensembles of much lower complexity (like a gradient boosting model based on decision-trees). The architecture of these models is difficult to decode, as it is not clear how critical a part any given feature plays in the prediction model or how it interacts with other features. For example, in a completely associated neural network, tracing the output features rendered by a model against a particular causative input feature remains a challenge.

DOI: 10.48175/IJARSCT-16988

3.2.1. Tree ensembles

While some models are inherently interpretable (like decision trees), others become more opaque as their complexity increases. Tree ensembles, which combine multiple decision trees for better performance, fall into this category. Their intricate structure makes understanding their decision-making process difficult. Here are some methods developed to address this challenge:

- Simplified Tree Ensemble Learner (STEL): This technique converts complex tree ensembles into a more interpretable rule-based model. However, the ensemble approach averages over individual models, making it harder to pinpoint the contribution of specific trees to the final output.
- Tree Interpreter: This method provides explanations for decision trees and random forests by breaking down each prediction into individual feature contributions and a bias term. This offers a clearer picture of how specific features influence the final prediction.

3.2.2. Support Vector Machines (SVMs):

While effective, SVMs, a type of machine learning algorithm used for classification and regression, can also be challenging to interpret. They operate in high-dimensional spaces and create separation boundaries (hyperplanes) between different classes. Understanding the rationale behind this hyperplane placement can be difficult, especially for complex datasets.

3.2.3. Explainable Neural Networks:

Neural networks, particularly large and deep ones with multiple layers, are notorious for being "black boxes." Their complex internal structure makes it difficult to understand how they arrive at specific outputs for a given input. This lack of transparency is referred to as "black box explainability." Explainable neural networks aim to address this challenge by providing insights into the model's inner workings. These post-hoc explanation techniques can be applied to various neural network architectures, including single/multilayer networks, convolutional neural networks (CNNs), and recurrent neural networks (RNNs).

An enormous problem is that deep learning methods turn out to be uninterpretable ”black boxes,” which create serious challenges, including that of interpreting a predictive result when it may be confirmed as incorrect. For example, consider Figure 3, which presents an example from the Nature review by LeCun, Bengio, and Hinton [2][1]. The figure incorrectly labels an image of a dog lying on a floor and half hidden under a bed as “A dog sitting on a hardwood floor.” To be sure, the coverage of their image classification/prediction model is impressive, as is the learned coupling of language labels. But the reality is that the dog is not sitting.

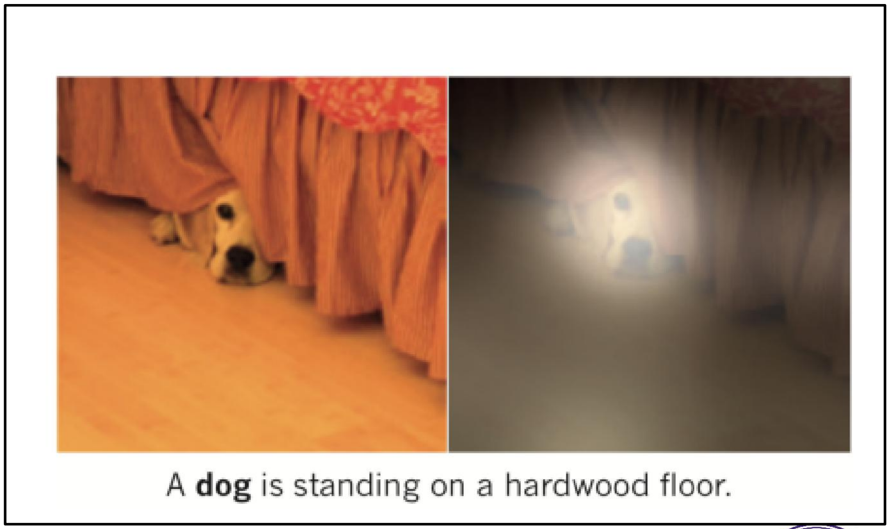


Fig. 3. Segment of an example from LeCun, Bengio,Hinton, Science [2][1]

DOI: 10.48175/IJARSCT-16988

3.2 White-Box Model Techniques.

A few AI models are straightforward and self-explanatory. For example, the predicted result y can be scientifically communicated as a weighted entirety of all of its highlights x. It is visualized as a straight line graph, with a as the slope of the line and b as the intercept on the y-axis. A linear model is a white-box model since its mechanism is straightforward and basic (as opposed to a black-box whose mechanism is not promptly caught on). In spite of the fact that basic, these are less capable of representing a bigger dataset highlighting complex interactions. Subsequently, for higher accuracy, we require more complex and expressive models.

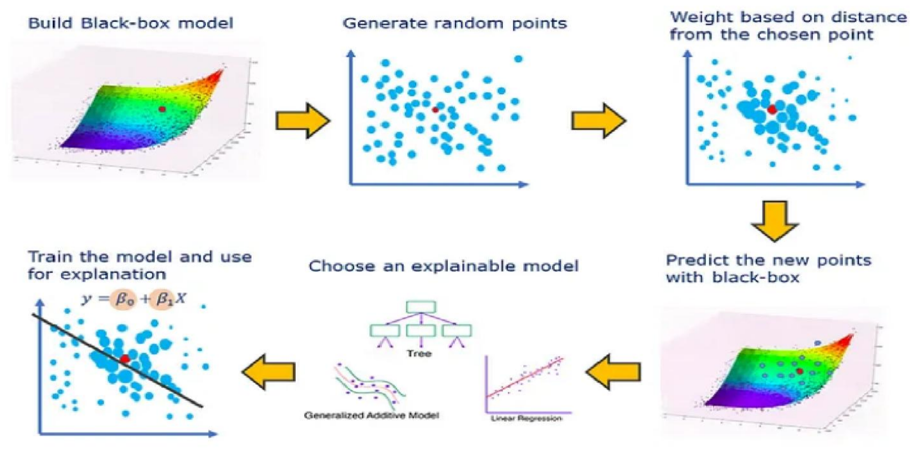
Few white-box models and programming methods used for interpretation.

- **Linear models:** Explainability of linear models involves a linear combination of feature values, adjusted by the coefficients of the model. For example, in y = mx + c, m is coefficient of feature x 1 and c is coefficient of x 0 so a polynomial of degree 1 is a linear polynomial. Similarly, Logistic Regression is one of the most interpretable linear ML models for certain classes of events. seaborn, matplotlib, sklearn and ALE libraries can be used to unfold and visualize a Logistic Regression Model.
- **Decision Tree:** A decision tree predicts the value of a target variable against multiple input variables. The terminal node, also called the leaf node, depicts the value of the target variable based on the input variable. A key benefit of decision trees lies in establishing the input and target variable relationship with a logic similar to Boolean. Scikit-learn library includes methods that can be used for interpretation of trees, e.g. sklearn.tree.export_text, sklearn.tree.plot_tree, sklearn.tree.export_graphviz, and dtreeviz and graphviz package. Sklearn also provides a way to evaluate feature importance – the total decrease entropy due to splits over a given feature.
- **Generalized Additive Models (GAMs):** Generalized Additive Models (GAMs) are an extension of Generalized Linear Models (GLMs) with a smoothing function. GAMs offer a trade-off between simple, interpretable models such as logistic regression and more complex, sophisticated models such as neural networks, which (usually) offer better accuracy and predictive power as compared to simple models. Over-fitting is unlikely in GAMs due to the regularization of prediction functions.

IV. EXPLAINABLE AI METHODS

4.1. LIME (Local Interpretable Model Agnostic Explanations).

LIME, the acronym for local interpretable model-agnostic explanations, is a technique that approximates any black box machine learning model with a local, interpretable model to explain each individual prediction.[8]


Build Black-box model
Generate random points
Weight based on distance from the chosen point
Predict the new points with black-box
Choose an explainable model
Tree
Generalized Additive Model
Linear Regression
Train the model and use for explanation
$y = \beta_0 + \beta_1 X$

Steps of the LIME algorithm. Picture by Giorgio Visani

Fig 4. Steps of the LIME Algorithm (Image courtesy of Giorgio Visani [8])

LIME algorithm's workflow, divided into sections:

- Build Black-Box Model: This represents the initial machine learning model you want to explain. LIME can be applied to various models, making it agnostic to the specific model type (hence the term "model-agnostic").
- Generate Random Points: This step involves generating random data points around a specific instance (data point) you want to understand from the original dataset. Imagine this instance as a particular prediction you want to gain insights into. LIME creates new data points, similar to the instance of interest, but with slight variations.
- Weight Based on Distance: The algorithm assigns weights to these newly generated data points. Points closer in proximity to the original instance receive higher weights, ensuring the explanation focuses on the local area around the specific prediction we're trying to understand.
- Train the Model (for explanation): LIME utilizes these weighted data points to train a simpler, interpretable model (like a linear regression model) to approximate the complex black-box model's behavior in the local vicinity of the instance of interest.
- Choose an Explainable Model: Here, you select an appropriate interpretable model type to explain the complex model's behavior. The image shows options like decision trees or generalized additive models (GAMs) along with linear regression

Interpreting the Explanation

The result of this process is an explanation for the specific prediction made by the black-box model for the chosen instance. This explanation is derived from the simpler, interpretable model trained in step 4. The explanation should highlight the features or factors in the data that played a key role in the model's prediction for that particular instance.

Key Points about LIME

- LIME focuses on explaining individual predictions rather than providing a global explanation of the entire model.
- It is model-agnostic, meaning it can be applied to various machine learning models.
- LIME explanations are local, meaning they explain the model's behavior in a specific region around the instance of interest, rather than offering a universal explanation for all predictions.

4.2 Anchors

The anchors method focuses on explaining individual predictions made by any black-box classification model. It achieves this by identifying a specific decision rule that acts like an "anchor," holding the prediction in place. This means that even if other features change, the prediction remains the same as long as the anchor rule is satisfied.

The method cleverly combines reinforcement learning techniques with a graph search algorithm. This approach minimizes the number of times it needs to call upon the black-box model (reducing computation time) while still ensuring it finds a good explanation, even if there are multiple possible explanations (local optima). Interestingly, the anchors method was developed by the same researchers who created the LIME algorithm

Good anchors should have high precision and high coverage. Precision is the proportion of data points in the region defined by the anchor that have the same class as the data point being explained. Coverage describes how many data points an anchor's decision rule applies to. The more data points an anchor covers, the better, because the anchor then covers a larger area of the feature space and thus represents a more general rule. Anchors is a model-agnostic explanation method, i.e., it can be applied to any prediction model without requiring knowledge about the internals.[10]

4.3 SHAP (SHapley Additive exPlanations)

SHAP (SHapley Additive exPlanations) cuts through the complexity of machine learning models, revealing how individual features contribute to specific predictions. It achieves this feat by borrowing the concept of Shapley values from game theory.

DOI: 10.48175/IJARSCT-16988

Shapley Values:

Imagine a team effort, like a model making a prediction. Shapley values help us fairly distribute the credit (the model's prediction) among the team members (the features) based on their individual contributions. Here's the key idea: we consider all possible ways these team members could collaborate in different groups (coalitions) and see how much each member's contribution adds to the final outcome (the prediction) across these various scenarios.

SHAP breaks down the explanation process:

1. Coalition Vector: A vector of 0s and 1s represents feature presence or absence. A 1 indicates a feature is "active" in a particular explanation, while 0 signifies it's "inactive."
2. Additive Explanations: SHAP presents the explanation as an additive model, similar to a linear model. This means individual feature contributions can be added together to explain the entire prediction.

The core formula for a SHAP explanation of a model f(x) for a specific instance x is:

$$g(z') = \varphi_0 + \Sigma(j=1 \text{ to } M)\ \varphi\_j * z'\_j$$

where:

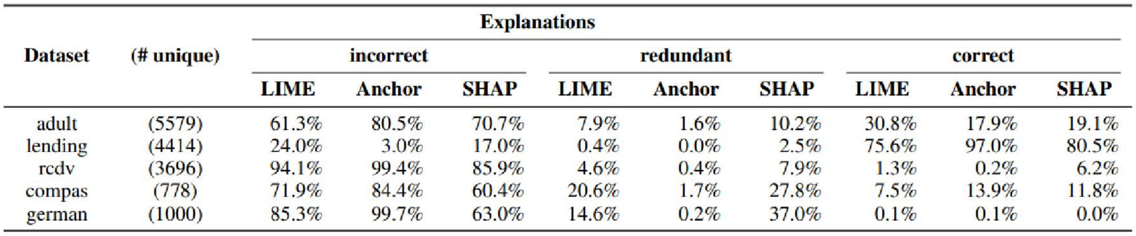
- g(z') is the explanation model (additive model representing feature attributions).
- φ_0 (phi_0) is the bias term, representing the average prediction of the model.
- M is the total number of features.
- φ_j (phi_j) is the Shapley value for feature j, representing its average contribution across all possible coalitions.
- z'_j is the jth element of the coalition vector, indicating the presence (1) or absence (0) of feature j in the explanation.

SHAP explanations adhere to specific properties that guarantee their trustworthiness:

1. Local Accuracy: The explanation (g(x')) should match the original model's prediction f(x) for the specific instance being explained.
2. Missingness: Features with missing values naturally have no contribution, so their Shapley value (φ_j) is zero.
3. Consistency: If a model changes such that a feature's influence increases or stays the same, its Shapley value should also increase or stay the same.

V. PERFORMANCE OF EXPLANATION METHODS ON VARIOUS DATASETS

The experiment by Ignatiev et al. (2019c)[7][11] focuses on a trained boosted tree that computes a heuristic explanation for each unique instance of an input dataset using LIME, Anchor, or SHAP. Five publicly available datasets are considered, including adult, lending, and recidivism. The experiment shows that most explanations computed by LIME, Anchor, and SHAP are inadequate from a global perspective. For 4 out of 5 datasets, all three explainers' explanations are mostly incorrect. For example, for recidivism and German, over 99% of Anchor's explanations are invalid. Similar results are found for LIME (SHAP), with 94.1% and 85.3% of explanations for recidivism and German being incorrect. The number of redundant explanations is usually lower, with the exception of SHAP. Overall, the number of correct explanations does not exceed 17.9% for Anchor, 30.8% for LIME, and 19.1% for SHAP, making it just a few percent.



Dataset	(# unique)	Explanations								
		incorrect			redundant			correct		
		LIME	Anchor	SHAP	LIME	Anchor	SHAP	LIME	Anchor	SHAP
adult	(5579)	61.3%	80.5%	70.7%	7.9%	1.6%	10.2%	30.8%	17.9%	19.1%
lending	(4414)	24.0%	3.0%	17.0%	0.4%	0.0%	2.5%	75.6%	97.0%	80.5%
rcdv	(3696)	94.1%	99.4%	85.9%	4.6%	0.4%	7.9%	1.3%	0.2%	6.2%
compas	(778)	71.9%	84.4%	60.4%	20.6%	1.7%	27.8%	7.5%	13.9%	11.8%
german	(1000)	85.3%	99.7%	63.0%	14.6%	0.2%	37.0%	0.1%	0.1%	0.0%

Table 1: Heuristic explanations assessed, for each data instance of the input datasets. The table shows the percentage of incorrect, redundant, and correct explanations provided by LIME, Anchor, and SHAP. The total number of unique instances per dataset is shown in column 2.(Table courtesy Alexey Ignatiev [7].)

VI. CONCLUSION

The growing adoption of Artificial Intelligence (AI) across various industries necessitates addressing its "black box" nature. Explainable AI (XAI) emerges as a critical field dedicated to making AI models more transparent and interpretable. This report delved into the history, core concepts, and methodologies of XAI, highlighting its significance in fostering trust, accountability, and human-AI collaboration. The workflow for developing XAI applications involves two distinct phases: understanding and explaining. The understanding phase focuses on applying XAI techniques during the model development stage to gain insights into the model's decision-making process. This is crucial for model validation, identifying potential biases, and laying the foundation for effective explanations. The explaining phase, which occurs during deployment, utilizes XAI techniques to generate explanations for the model's predictions, enhancing trust and enabling human oversight. The report further explored the distinction between white-box and black-box models. White-box models, like linear models and decision trees, are inherently interpretable due to their simpler structure. Black-box models, such as complex neural networks, are more challenging to understand. Various techniques have been developed to address this challenge, including LIME, Anchors, and SHAP. LIME offers model-agnostic explanations by approximating complex models with simpler, interpretable ones for each individual prediction. Anchors identify specific decision rules responsible for a prediction, while SHAP leverages Shapley values from game theory to explain how individual features contribute to a prediction. The report concludes by summarizing the findings of a recent experiment by Ignatiev et al. (2019c) [7]. This experiment evaluated the performance of LIME, Anchor, and SHAP on various datasets. The results are concerning: for most explanations generated by all three methods, they were found to be incorrect from a global perspective. This highlights the ongoing challenge in developing robust XAI techniques, particularly for complex models.

In conclusion, while XAI offers a promising path towards achieving transparency and trust in AI systems, significant research efforts are still required. Future advancements in XAI methods, coupled with a deeper understanding of model behavior, are crucial for ensuring the responsible and ethical development and deployment of AI technologies.

REFERENCES

[1] Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P. & Holzinger, A. (2018). “Explainable AI: The new 42?”. Paper presented at the CD-MAKE 2018, 27-30 Aug 2018, Hamburg, Germany. doi: 10.1007/978-3-319-99740-7_21.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “ Deep learning. ” Nature, 521:436 EP –, 05 2015.

[3] Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek, “ Explainable AI Methods - A Brief Overview ” Springer Link, 17 April 2022.

[4] Robert R. Hoffman, Shane T. Mueller, Gary Klein, Jordan Litman, “Metrics for Explainable AI: Challenges and Prospects”, Presented with arXivLabs ,https://doi.org/10.48550/arXiv.1812.04608 ,(11 Dec 2018).

[5] Feiyu Xu, Hans Uszkoreit , Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu, “ Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges ”, DOI:10.1007/978-3-030-32236-6_51 In book: Natural Language Processing and Chinese Computing (pp.563-574) (September 2019).

[6] Dwivedi Rudresh, Dave Devam, Naik Het, Singhal Smiti, Rana Omer, Patel Pankesh, Qian, Bin, Wen Zhenyu, Shah Tejal, Morgan Graham and Ranjan Rajiv. “Explainable AI (XAI): core ideas, techniques and solutions. ACM Computing Surveys, Publishers page: http://dx.doi.org/10.1145/3561048 (2023)

[7] Alexey Ignatiev Monash University, Australia alexey.ignatiev@monash.edu , “ Towards Trustable Explainable AI ”, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) Early Career Track.

[8]GiorgioVisani, “LIME: explain Machine Learning predictions” , Published in Towards Data Science on Dec 18, 2020.

[9] “What is Local Interpretable Model-Agnostic Explanations (LIME)?” , blog on C3.ai What is Local Interpretable Model-Agnostic Explanations (LIME)?.

[10] Tobias Goerke & Magdalena Lang , “Scoped Rules (Anchors)” from book “Interpretable Machine Learning: A Guide for Making Black Box Models Explainable ”, Christoph Molnar (2023-08-21).

DOI: 10.48175/IJARSCT-16988

[11] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems (2017).

[12] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On validating, repairing and refining heuristic ML explanations. CoRR, abs/1907.02509, 2019.