Explainable AI

Abdurrehman Syed
University of Oulu
abdurrehman.syed@student.oulu.fi

***Abstract*—Artificial Intelligence (AI), more precisely Deep Learning (DL) has made significant contributions to modern day problem solving in various domains. In comparison with traditional supervised and unsupervised Machine Learning (ML) algorithms, deep learning has achieved more precise and accurate models for predictions by utilizing big data. However, as the complexity of the models increased so did their inexplicability making it very challenging to interpret the decision making process. Many high-stake applications such as banking, autonomous vehicles, military, medical diagnosis require interpretability and explainability of their AI systems. This report discusses the emerging field of Explainable AI (XAI) which uses various techniques in dealing with such black-box architectures making these AI models transparent and trustworthy. Furthermore, emphasizing the limitations, challenges and future directions for this evolving discipline.**

***Index Terms*—Explainable AI, XAI, Big Data, Blackbox, Whitebox, DNNs, LIME, SHAP**

I. Introduction

With the advancements in AI and the usage of computer systems to handle complex day to day tasks has increased exponentially we have seen a rise in intelligent systems, systems that can recognize and learn patterns in data and easily perform complex computations or make decisive predictions. For instance, linear regression and decision trees which follow a systematic logic and based on that give responses that can easily be back-tracked. However, more complicated tasks required more convoluted structures and algorithms which gave rise to the issue of interpretability of these models [1]. The prime example is the rise of deep learning which consists of Deep Neural Networks (DNNs). These networks are a set of layers, each layer comprising hundreds or even thousands of neurons which can be interpreted as functions that take inputs and provide outputs but here these layers of neurons are interconnected with each other which hints that the output of one neuron will affect the entire structure of the network due to this complexity these models are referred to as "black-box", since it is implausible to understand how the network reaches to an output as you have to track the behavior of each and every neuron along with biases attached with them. This lack of interpretability problem hinders trust and limits the application of these powerful tools. Explainable AI emerges as a response to this challenge [2], aiming to make these sophisticated models more transparent and explainable. This transparency is crucial in applications where building trust within AI is important such as medical diagnosis, where the model generates a diagnostic report but without knowing how it reached that conclusion it can put the patient's life in danger. Similarly, loan systems use AI to find whether a person is eligible for a loan or not, understanding the decision will not only satisfy the customer but also reflect the performance of the model and its decision making capabilities. Explainable AI follows certain methodologies where it implements explainability within the model before training or uses algorithms after training to make the model interpretable [5]. It consists of a range of tools that can work on general purpose systems as well as specific models. Additionally, with the help of these techniques one can find how the overall structure and parameters of the model are contributing towards the output and how features are correlated with each other.

II. XAI and Big Data

Data is the key feature for any AI system. In recent years we have seen an astronomical increase in data generation with almost 328 terabytes of data being generated everyday. This surprising volume encompasses diverse sources, from social media interactions to sensor readings to healthcare records and financial transactions. In-order to handle and analyze these large volumes of mostly unstructured data we use Big Data methodologies. These methodologies tackle the complexities of data through a range of several tools and techniques designed to collect, store, process and analyze large volumes of data at high velocities. From distributed storage systems like Hadoop to scalable processing frameworks like Spark, allowing us to extract meaningful and relevant data.

In the scope of AI, Big Data serves as the foundation upon which the advanced machine learning and deep learning models are built. By training these models with massive amounts of labeled and unlabeled data, organizations can train AI systems to make predictions, recognize patterns, and derive meaningful outcomes. However, the abundance of data produces a particular set of challenges, in terms of interpretability and transparency. As the models process increasingly complex datasets they get more and more dense and understanding their underlying functionalities becomes very difficult. Explainable AI here bridges the gap between Big Data and the model's outcomes by providing insight on the AI decision making process. Big Data and Explainable AI together form a powerful connection, where XAI techniques not only handle the explainability but they also help in finding biases within the data as well as feature importance and relevance which greatly impacts the size of our datasets and overall performance of the models.

III. XAI Types

XAI can be categorized into two types [4]: *model-based*, which aims to integrate interpretability in the models during the training phase, and *post-hoc* models which involve interpretability after the training phase.

A. Model Based

These models are designed in such a way that they're explainable in nature, they follow a simple and systematic structure, providing transparency and straightforward understanding of their decisions. These models are quite simple and often referred to as white-box models.

1) Linear Models: These types of models have a very plain architecture, expressing relationship between features and target variables as an equation, where each feature is assigned a certain weight to influence the predicted output and individuals can easily understand the working by examining the coefficients. Fig-1 shows a simple linear correlation between two features.

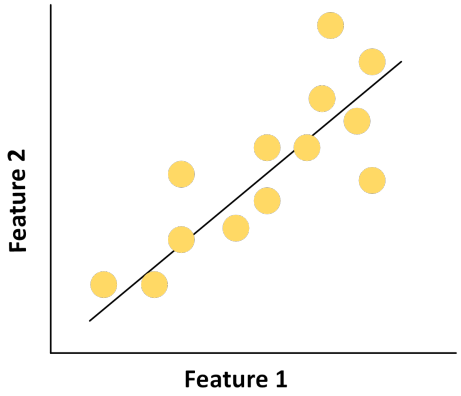

Fig. 1. Linear Model

2) Decision Trees: Decision Trees express their classification or regression logic in the form of a tree consisting of nodes and branches, each node represents the decisions and outcomes based on input features. The decisions are easy to understand and the flow of execution towards a particular decision is pretty self-explanatory as seen in Fig-2. Moreover, decision trees naturally provide us information about the important features that have the biggest impact on the final prediction. However, decision trees can become very inexplicable and difficult with large datasets in such scenarios they can be considered as black-box models.

B. Post Hoc

These models are first trained and then XAI techniques are applied to understand their decision-making structure; they don't modify the model itself but rather analyze its behavior after training. Post-hoc techniques can be applied to any model, the explanations are often approximations and generating them can be very computationally expensive for complex models.

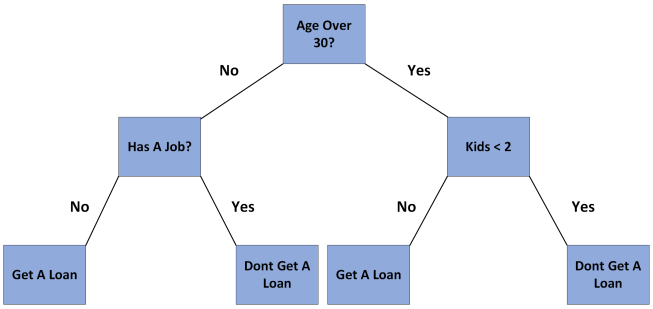

Fig. 2. Decision Tree

1) Deep Neural Networks: DNNs are powerful models that are used in learning complex patterns and relationships. Their dense multi-layered architecture with hundreds and thousands of features within the network make them very complicated and opaque. Additionally understanding the activation of the neurons for each layer and biases makes the problem exponentially difficult to comprehend. Fig-3 shows a simple form of DNN.

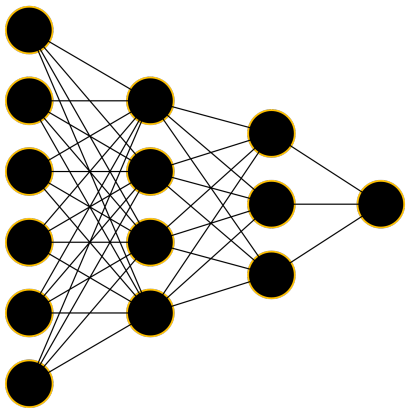

Fig. 3. Neural Network

IV. XAI Techniques

XAI methods can analyze models in two ways: model-agnostic or model-specific. Model-agnostic methods aren't limited by the structure or features of the AI algorithm and can be used for any black-box model, in other ways they're generic. Model-specific methods on the other hand are specific to certain algorithms for which they were created. They provide a more detailed understanding of the working and decision making of their respective algorithms they are created for.

Different XAI methods perform their analysis with varying perspectives and in different ways. Certain methods tell us about the features of the data and their correlations while others can highlight their relevance for the model and contributions.

Local and global explainers further categorize the models where local explainers explain the model's behavior for specific decisions, whereas global explainers take into account

the behavior of the model in general. For instance in a local explainer, a movie recommendation system can either recommend a movie based on the past movies a user has watched and based on that suggests a movie and explain why it did that or it will consider the global scope and analyze all the outcomes of the model and its general behavior, things like how much did the user's favorited movies affect the recommendation more than the genre or actors. A global explainer puts the model's inputs and their respective outputs together in such a way that it can understand how each underlying feature is affecting the outcome.

A. Model-Agnostic Techniques

These techniques do not take into account the structure of the model and can be applied to any machine learning algorithm. They work in a way where they obtain explanations based on tampering with the feature values and understanding the changes in the outcomes of the model. This approach gives interesting insights on the varying sensitivity of the data performance.

Below are some of the popular model-agnostic techniques which further explains the concept.

1) LIME: LIME or Local Interpretable Model-Agnostic Explanations is used in explaining individual predictions made by complex machine learning models. The primary objective of LIME is to understand and provide local approximations to the model's predictions and in enabling better understanding and interpretability.

The process begins with first choosing an output data prediction for which explanation is required and then creating a simpler linear model around it. LIME then varies the input features and analyzes the variations in outputs until it finds the approximation for features where it gets to the output of the selected data point. Through this approach LIME records all the weights of various features and their contributions for this particular output. By providing these local explanations LIME helps in gaining insight on the inner working of the system.

2) SHAP: SHAP or Shapely Additive Explanations [7] is used for understanding predictions from a global perspective, unlike local interpretations provided by techniques such as LIME. It achieves this by calculating "shapely values" for each data feature in the dataset, where each shapely value represents the weight for that feature on the predictions performed by the model.

The concept of Shapely values originates from game theory, where each data attribute is considered a player that contributes to a game – in this case, the game is the machine learning model and the predictions are the rewards as seen in Fig-4. Each feature receives a portion of the reward based on its contribution; the feature that contributes the most will have the highest weight or shapely value.

To calculate the shapely value for a specific feature, we first analyze the difference in predictions before and after removing that feature from the consideration. Importantly, SHAP not only considers direct impact but also the correlation between features is considered and counts towards the weight contribution ensuring a comprehensive and detailed explanation.

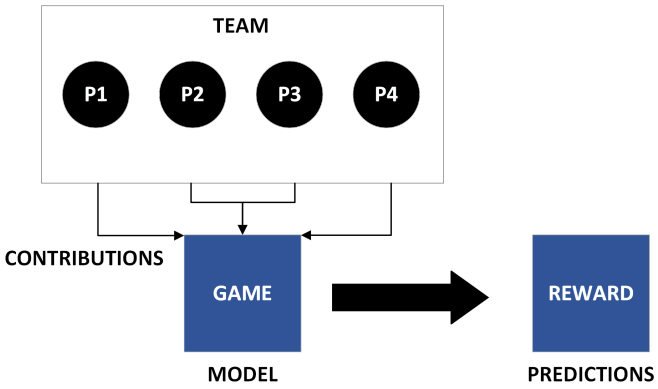

Fig. 4. SHAP - Game Theory Approach

B. Model Specific Techniques

These techniques are specific to certain model architectures and are used for understanding their internal working, like a specific Deep Neural Network (DNN). They apply reverse-engineering approach to provide explanations on the decision-making capabilities.

The benefit of this approach is that it allows us to get a deeper understanding and customizing the model, learning feature-importance. However, the negative of such models is that we will have to compromise the performance. For instance, de-convolution of a Convolutional Neural Network (CNN) which traverses the path of CNN in reverse direction starting from the output and going back to the input layer highlighting the contributions of neurons in each layer and what features of the input image contributed the most.

1) Attention Mechanism: Attention Mechanism is a model-specific tool for Natural Language Processing (NLP) and image recognition. It is used for assigning more importance to certain attributes based on their relevance among the features.

Attention mechanism uses a encoder-decoder approach where the encoder assigns weights to the input features based on their relevance and similarity scores with the query. It generates an array known as a context-vector which holds all the feature weights. These weights are then passed onto the decoder where it performs the attention mechanism by not only assuming the values in the context vector but also assigning its own attention values based on the data it got. Assuming an example of English to French translation each word in the english sentence is given some input weights these weights are then sent to the decoder where it assigns attention mechanism based on the relevance between those words for example, "pizza" and "eat" are related so it will get more attention values resulting in a more accurate translation. This way the attention mechanism can explain the understanding of the specific model by paying attention to the right features.

V. XAI Frameworks and Tools

XAI frameworks and tools [3]: These are the essential components that are needed for developing, testing, and explaining machine learning models. They provide various XAI techniques and methods for solving the challenges faced by traditional AI.

A. ELI5

ELI5 [8] which stands for "Explain Like I am 5" is a python library which is designed for users who do not have a technical background and want to understand the working of complex machine learning models in an easier way. It does so by giving it the model and in return ELI5 analyzes that model and generates a plain text summary of its working.

B. InterpretML

Interpret ML [12] is an open-source python library developed by Microsoft Research designed for model interpretation and understanding. It offers various features including model-agnostic and model-specific techniques, ability to explain individual and global behavior as well as it provides an interactive user-interface where the models are explained with visualizations.

C. Alibi Explain

Alibi [9] is a powerful and reliable open-source python library for explaining machine learning models. It offers a range of techniques and explainability methods. Moreover, it provides integration with other Alibi frameworks like Alibi Detect to monitor the performance of the model and integration with other frameworks like TensorFlow and scikit-learn makes it an easy to use module.

D. AIX360

AIX360 or AI Explainability 360 [10] is IBM's open source toolkit designed for developers and researchers by providing them with a set of algorithms and tools to work on ML black-boxes. Furthermore, it is flexible in the way that users can mix and match explanation techniques and also provides evaluation metrics to compare the interpretability of different models.

E. What-If tool

What-If tool [11] is an open-source application designed to visually analyze the machine learning models. It provides an interactive interface where users can generate "what-if" scenarios by changing input parameters and visualizing how it affects the output, inspecting individual data points, and feature information. Additionally, it provides functionalities to perform fairness analysis and how ML predictions vary across different demographic groups.

VI. Applications

The fundamental difference between a traditional machine learning model and an XAI model is of explainability. Traditional models often operate as black-boxes providing accurate results but insufficient information for the predicted outcomes. XAI models on the other hand prioritize not only the accuracy but also the transparency and explanations for their results helping users in identifying potential errors or biases that they may have. Not only does it develop trust but also increase performance which is a concern for traditional AI. Fig-5 shows the different workflows of both models. For a traditional approach the data passes through the blackbox of the model and gives us an output, however, for XAI workflow the data passes through XAI block which creates explanations for what is happening inside of the model and then generating report alongside the predictions.

Fig. 5. Workflow

The following workflow of XAI has created a lot of applications for organizations and users that require transparency in their models.

PayPal uses machine learning to figure out the fraudulent activities from its transactions. Their models traversing through millions of transactions in real-time, they use XAI for understanding and labeling these fraudulent transactions for further review or modifying the decision.

BlackRock, a leading asset management company, uses AI for investment strategies. Their AI system analyzes vast amounts of data to search for investment possibilities, while Explainable AI (XAI) allows them to communicate these decisions to both internal investment managers and external clients.

IBM Watson, a healthcare analyzer which provides diagnostic reports along with its explanations and recommends optimal treatment to patients based on their conditions.

XAI is utilized in the legal sector as well, where it can provide its analysis over cases, legal advice and judgements with clear explanations behind it, officials can learn and understand while building a trust in their model. Moreover, the autonomous vehicle sectors are using XAI for their cars and for building trust within their customers. Companies like ZestFinance are using XAI for understanding and explaining their customers regarding eligibility of their loans and its requirements.

VII. Discussion

Explainable AI due to its complicated nature deals with a lot of challenges, this section highlights some of those key features.

a) Trade-off between Accuracy and Interpretability: There is often a trade-off between accuracy and interpretability when dealing with different types of models where increasing interpretability may result in loss of accuracy and vice versa. Finding the right balance for the given task with respect to this trade-off is a key challenge. Fig-6 shows the trade-off where the simplest models like the linear regression has the highest interpretability and as the complexity increases the accuracy as can be seen of the DNNs but at the cost of interpretability.

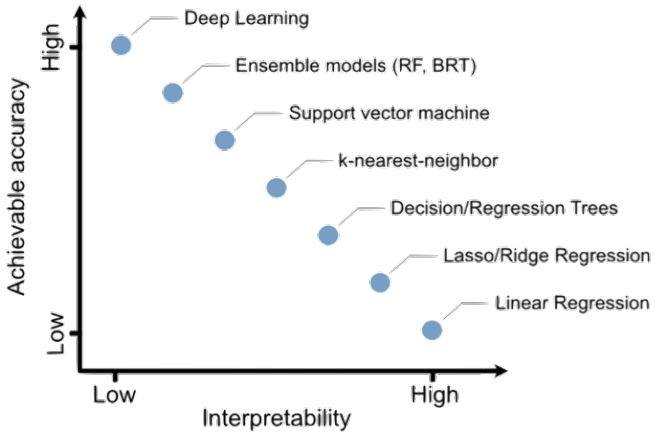

Fig. 6. Tradeoff: Accuracy - Interpretability (Image taken from [6])

b) Black-Box Complexity: Deep learning is an expanding field and with time there will be more and more complex DNNs. This highlights that the complexity of these black-boxes will keep on increasing which is a growing challenge as it will become much harder to interpret their decisions.

c) Approximation and Accuracy: Post-hoc explanations are an approximation and it is not necessary that they may always be right or capture the true inner workings of the model. Ultimately, leading to bad explanations and wrong decisions.

d) Scalability and Efficiency: XAI models should be scalable and efficient, capable of working with large scale applications and data, interpreting complex models while being efficient is an ongoing challenge.

e) Privacy Concerns: In some cases, the models may leak sensitive information about the data they have been trained on in their explanations. Encouraging explainability while being cautious of privacy is a key challenge.

f) Lack of Standardized Techniques: Another concern is the lack of standardized methods, XAI is still evolving and there is an absence of well-established techniques for explaining all types of machine learning models. This can lead to inconsistent explanations where each technique interprets the model differently causing untrustworthy explanations .

Despite these challenges, the future for XAI looks very promising. There are already efforts being made to subdue these obstacles. Additionally, we can expect more enhanced personalized learning in education, intuitive chatbots, trustworthy AI systems in businesses, banking, automobile industries and most importantly in the research area. Understanding how these intelligent computers think will not only impact Artificial Intelligence itself but all the domains connected with it.

VIII. Conclusion

Explainable AI has the potential to revolutionize how we understand, interpret and interact with Artificial Intelligence. By allowing trust and transparency within the decisions and for creating more efficient futuristic models with the help of its techniques. While XAI faces challenges due to the black-box nature of the models and big data complexities, ongoing research efforts are paving the way for a future where explainable models become the norm, enabling the creation of even more powerful and innovative AI systems.

References

[1] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, Rajiv Ranjan. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, 30 September 2023. https://doi.org/10.1145/3561048

[2] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv. 55, 13s, Article 295*. December 2023. https://doi.org/10.1145/3583558

[3] A. Lucic. XAI PAI Toolsheets. Available at: https://a-lucic.github.io/talks/xai_pai_toolsheets.pdf [Accessed May 3, 2024]

[4] F. K. Došilović, M. Brčić and N. Hlupić. Explainable Artificial Intelligence: A Survey. *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 2018. doi: 10.23919/MIPRO.2018.8400040. https://ieeexplore.ieee.org/abstract/document/8400040

[5] A. Adadi and M. Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, vol. 6, pp. 52138-52160. 2018. doi: 10.1109/ACCESS.2018.2870052. https://ieeexplore.ieee.org/abstract/document/8466590

[6] Accuracy-Interpretability Trade-off. Available at: [Accessed May 3, 2024]

[7] Alexander A. Huang and Samuel Y. Huang. Increasing Transparency in Machine Learning through Bootstrap Simulation and Shapely Additive Explanations. *PLOS ONE*, vol 18 pp 1-15. 2023. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0281922

[8] ELI5. https://eli5.readthedocs.io/en/latest/

[9] Alibi Explain. https://docs.seldon.io/projects/alibi/en/latest/ [Accessed May 3, 2024]

[10] AIX 360. https://aix360.res.ibm.com/ [Accessed May 3, 2024]

[11] What-if Tool. https://pair-code.github.io/what-if-tool/ [Accessed May 3, 2024]

[12] Interpret ML. https://interpret.ml/ [Accessed May 3, 2024]