Pratibodh
A Journal for Engineering
A free and Open Access Journal
Homepage: https://pratibodh.org

Decoding Machine Learning: Transparency Matters

Anshika Jain[1], Dr. Manju Vyas[2]
Department of Artificial Intelligence & Data Science
Jaipur Engineering College & Research Centre,
[1]anshikajain.ai24@jecrc.ac.in, [2]manjuvyas.cse@jecrc.ac.in

Abstract

Machine learning models are often complex and difficult to understand, which can make it difficult for users to trust them and ensure that they are fair. This is especially important as machine learning is increasingly being used in real-world applications that have a significant impact on people's lives. One way to address the problem of transparency in machine learning is to use explainable and interpretable models. These models can provide users with insights into how the model operates and justifies its predictions to users. This can help users to build trust in the model and to identify any potential biases or unfairness.

Keywords: Black-Box, Machine Learning, Deep Neural Network, Artificial Intelligence, Transparency, Logistic Regression, Reverse Engineering.

Article Status

Available online :

1. Introduction

In our social context, decision-making spans various spheres like business and health, now facilitated by the remarkable advancement of Machine Learning (ML). ML is a branch of Artificial Intelligence (AI) that equips computers to autonomously learn from data, enabling complex model creation beyond human comprehension. Despite its potential, challenges arise, notably in transparency. Critical decisions occur within opaque ML models, hindering trust, fairness, privacy, and security. Transparency demands understanding of internal operations and outcome justifications. However, most ML outputs remain non-human interpretable, fostering concerns around security, safety, and bias.

Sophisticated machine learning techniques, including Deep Neural Networks, often pose the challenge of being opaque (the "black box problem"). Despite their superior performance, specific models like LSTM neural networks surpass standard logistic regression in complexity and interpretability. Moreover, the presence of biased training data perpetuates prejudicial outcomes, reflecting the undesired influence of inherent biases within the algorithms' decision- making process.[1]
The research paper examines existing discourse on decoding machine learningtransparency

matters, highlighting emergent threads and potential solutions, particularly the role of explainable AI/ML. It focuses on discerning alternative configurations within technical and policy dialogues. Given the absence of precise accountability mechanisms, the paper advocates for a nuanced exploration of transparency, addressing calls for algorithmic transparency across varied implementations.

The paper's second section evaluates prior research on transparency in machine learning, followed by an examination of challenges posed by opaque systems. Subsequently, the problem statement concerning black- box models and transparency is delineated. Sections five and six delve into explainable and interpretable ML models, proposing solutions to non-transparency issues. The conclusion integrates key findings, supported by a comprehensive reference list.

2. Background work analysis:

Machine learning algorithms play a pivotal role in decision-making across diverse life domains. The growing demand for fair, accountable, and transparent (FAT) machine learning algorithms is evident. Despite advocating transparency in machine learning, it has faced criticism due to several reasons. This section critically reviews previous scholarly works on transparency in machine learning and highlights the associated debates. It aims to shed light on the multifaceted discourse surrounding the concept's implementation and effectiveness.

2.1 Transparency and interpret-ability

In her book, Cathy O'Neil underscores the significance of transparency, advocating for the disclosure of source code to address concerns about opacity. She posits that opacity can foster a sense of unfairness and highlights the demand for transparency in machine learning processes [2]. O'Neil argues that understanding both input data and source code is essential for enhancing the interpretability and confidence of machine learning algorithms. This transparency, she suggests, can play a crucial role in identifying and rectifying biases and unfair practices within the system.

On the contrary, Frank Pasquale offers a different perspective by framing opacity as a strategic move for self-preservation by corporations. He supports the idea of opening the "black box" of machine learning algorithms, contending that disclosing the source code will unveil the program's behavior [3]. Pasquale's argument is rooted in the assumption that transparency, achieved through revealing the source code, can provide insights into the decision-making processes of machine learning programs.

Computer scientists, exemplified by Diakopolous,

propose methods such as making source code available for scrutiny and suggesting that journalists employ reverse engineering to bolster accountability [4]. However, this proposal faces criticism from other quarters, including computer scientists and lawyers, who refute the idea by asserting that the inherent complexity of machine learning systems makes it challenging to comprehensively understand the algorithm in action, especially during the learning process. They label the disclosure of source code as an "obvious, but naive" solution, suggesting that the very nature of machine learning complicates direct comprehension [4].

2.2 Transparency and accountability.

Some people argue that transparency is necessary for accountability. They argue that we cannot hold machine learning systems accountable for their decisions if we cannot understand how they work. Nevertheless, comprehending these algorithms can be challenging due to their intricate nature and complexity. They argue that we also need to have mechanisms in place to ensure that machine learning systems are used in a fair and ethical way.For example, Professor Alan Winfield argues that we should always be able to find out why an AI system made a particular decision. He argues that this is necessary for accountability.[5]. However, other professors disagree. They argue that transparency is not always possible in machine learning systems, and that it is not enough to ensure accountability.[6]. Overall, the debate over transparency and accountability in machine learning is complex. There is no easy answer, but it is an important issue to consider as machine learning becomes increasingly pervasive in our lives.

2.3 Transparency and public/private sector

Transparency and accountability, vital in democracy, clash with machine learning's proprietary nature. Intellectual property conflicts with transparency; safeguarding sensitive data hinders public access. While some transparency suits government machine learning, complete openness is impractical. Ethical and legal calls for accountability lack defined mechanisms [7]. Public law guarantees stem from good governance, anti-corruption, and constitutional commitments tied to public duties. Private law relies on horizontally applied rights, seen in privacy policies and data protection regulations [8]. This complexity highlights challenges in integrating machine learning into democratic processes.

2.4 Transparency and power structure

In the current structure of machine learning, transparency requires impacted groups to proactively call for disclosure and consistency. This is because thedesigners of machine learning frameworks have an overwhelming advantage. They can choose to be responsible, but there are no guarantees that they will. Some academics have worked to improve the current structure, but others are pushing against it.[3,4]. For example, Watcher et al. argue that transparency is insufficient because it is a post-factor measure, meaning that it only occurs after harm has already been done.[9]. Datta et al. found that Google showed more paying job positions for men than for women and called for machine learning algorithms that would prevent such harms by avoiding segregation and providing transparency.[10].

3. Lack of transparency:

When machine learning models are not transparent, it means that we cannot understand why they make the decisions that they do. This can be a problem because it can lead to bias and unfairness in the results. For example, a machine learning model that is used to predict loan approvals might be biased against certain groups of people, such as minorities or women. If the model is not transparent, we cannot identify and fix this bias. Non-transparent machine learning models can also break laws and ethics. For example, a model that is used to predict crime rates might be biased against certain neighborhoods. If the model is not transparent, it is difficult to hold accountable the people who designed and deployed the model.Ensure your paper commences with a clear, detailed title, avoiding abbreviations. List authors' full names, specifying the corresponding author. Subsequently, provide affiliations with numerical linkage to authors.

3.1 Trust Issues

Machine learning (ML) models are used to make predictions about the future. They do this by taking in data from the past and learning from that. However, ML models can be complex and difficult to understand. This can make it difficult for users to trust the results of the model. It is important to have transparent ML models so that users can understand how they work and why they make the predictions that they do. This is especially important for business analysts who need to be able to explain the results of the model to their stakeholders. Transparency is also important for iterating on ML models. By understanding how the model works, it is easier to identify and solve problems.

3.2 Bias

Machine learning models play a critical role in crucial decision-making processes concerning various aspects of people's lives, such as who gets a job, who gets parole, and who gets a loan. However, these models can be biased, meaning that they may make unfair decisions against certain groups of people. This is a serious problem because it can lead to discrimination and unfair treatment. It is important to mitigate bias in machine learning models so that they can make fair and unbiased decisions. There are a number of ways to mitigate bias in machine learning models. One way is to make the models more transparent so that we can understand how they work and why they make the decisions they do. Another way is to use techniques to clean the data that the models are trained on. It is especially important tomitigate bias in machine learning models in fields such as finance and healthcare, where these models can have a significant impact on people's lives.[11].

3.3 Unexplained results

Machine learning models can sometimes produce unexpected results for some inputs. This is because machine learning models are trained on data from the past and may not be able to generalize to new data that they have never seen before. For example, a machine

learning model that is trained to predict loan approvals might be trained on data from a time when the economy was good. In times of economic downturn, machine learning models might yield unforeseen outcomes. Understanding their constraints is crucial, as is exercising prudence in their application. Regular monitoring and potential retraining of these models are essential practices for sustained efficacy.[12].

4. Problem Statement:

Consider a scenario where an individual's credit card application is denied via a website due to a newly implemented machine learning algorithm, leaving the applicant seeking an explanation. Regrettably, customer service is unable to offer any reasoning for the rejection, representing a situation increasingly observed in practical contexts. Recent incidents in the sector demonstrate the algorithmic opacity affecting credit approvals.

Machine learning and artificial intelligence substantially enhance business efficiency and innovation, yet the imperative of comprehensible decision-making in automated systems persists. This challenge primarily confronts extensive deep learning models and neural networks, which fragment problems into countless sub-components, obscuring the internal workings, denoted as the "black box dilemma." The ensuing lack of transparency impedes the necessary insights for algorithmic updates and instills a complex web of trust concerns. Establishing confidence in machine learning algorithms assumes heightened significance with their projected deeper integration into daily affairs. Clarifying the internal mechanisms of these systems is pivotal for fortifying trust, ensuring ethical practice, and fostering reliability in their operation within society.[13].

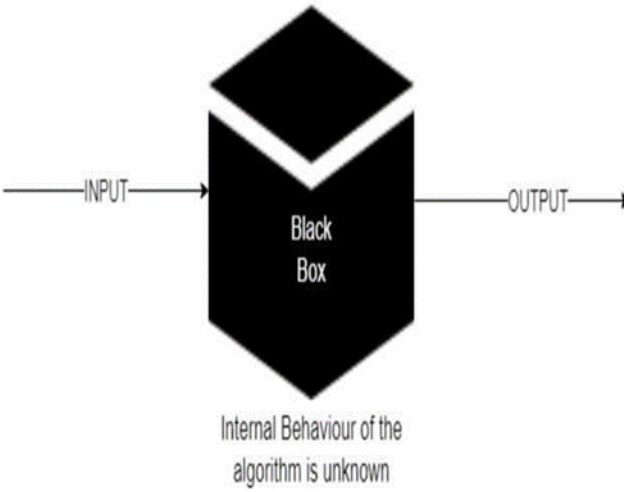


Figure 1. Opaque in decision-making processes[13].

5.Explainable ML/AI:

The concept of Explainable AI involves techniques and design strategies facilitating the addition of transparency to AI algorithms, validating their predictions. Through AI characterization, intended impacts, and bias assessment, human experts can comprehend and foster trust in the outcomes. The effectiveness of explanations rests upon their informative content, supporting a context-specific understanding. Two scenarios are conceivable:[14].

1.Establishing the AI model's lineage involves delineating its training process, data incorporation, potential biases, and corresponding mitigation strategies.

2.Elucidating the overall model through dual techniques:

(a) Proxy modeling entails the use of a simpler model, like a decision tree, to represent a more intricate AI model. Although it offers a simplified insight, it remains an approximation and might not precisely mirror the actual outcomes.

(b) Designing for interpretability involves configuring AI models to prioritize clear and comprehensible behavior. However, this approach could lead to less robust models as it restricts the developer's toolkit and potential complexity.[15].

NIST USA devised four AI principles for transparency and accountability:

1.The NIST's four AI explainability principles emphasize evidence-backed, comprehensible system outputs. Diverse explanation types include rationale- based, causal, and algorithmic justifications.

(a) Provide user-friendly, constructive explanations to enhance user understanding and engagement.

(b) Create justifications to instill confidence and trust in the system's operations.

(c) Develop justifications aligned with regulatory mandates to ensure legal compliance and adherence.

(d) Craft explanations facilitating algorithm development and upkeep for effective management and enhancement.

(e) Generate justifications advantageous to the model owner, as seen in movie recommendation systems.

2.The explanation's relevance is crucial for user guidance and task fulfillment. To cater to diverse user competencies, the system should offer multiple explanations suited to different proficiency levels within the user base.

3.The explanation should be succinct and accurate, distinguishing it from the correctness of the output itself.

4.Operate within predefined knowledge boundaries to yield expected outputs that align with the system's capabilities and constraints.[16].

Various frameworks aid in addressing the black-box challenge and fostering transparency in machine learning algorithms.

1.SHAP

(SHapley Additive explanations) serves as a versatile tool across various machine learning (ML) models, ranging from basic linear regression to intricate deep learning architectures for tasks like image classification and natural language processing. Its adaptability extends to functions such as opinion mining, translation, and text summarization, offering in-depth insights into the intricate workings of complex models. Operating independently of the model type, SHAP leverages the principles of game theory's Shapley values to unravel the impact of diverse features on a model's output, thereby enabling a profound understanding of the contributing elements steering specific outcomes.

By delving into the intricate interplay between various characteristics and the final model outcomes, SHAP not only enhances interpretability but also contributes to bolstering the trust and reliability of AI-driven decision-making processes. Its ability to shed light on the nuanced contributions of different features not only aids in understanding model behavior but also assists in

refining and optimizing the decision-making process across multiple domains and industries.[17].

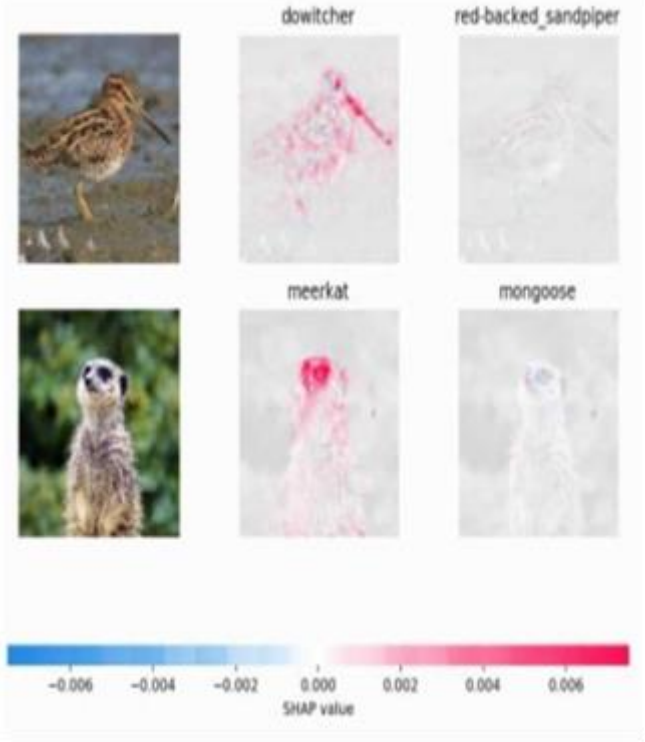


Figure 2: SHAP Framework[17].

2. LIME

LIME (Local Interpretable Model-agnostic Explanations), akin to SHAP, excels in rapid computations. Generating a range of explanations corresponding to each contributing aspect in a provided data sample's prediction, LIME efficiently elucidates black-box classifiers with multiple classes. Its functionality is contingent on the classifier's ability to construct a function that receives raw data or a NumPy array, offering probabilities for individual classes. Leveraging these attributes, LIME serves as a valuable tool for comprehending and interpreting the inner workings of complex models, facilitating swift and effective explanations across various classification scenarios.[18].

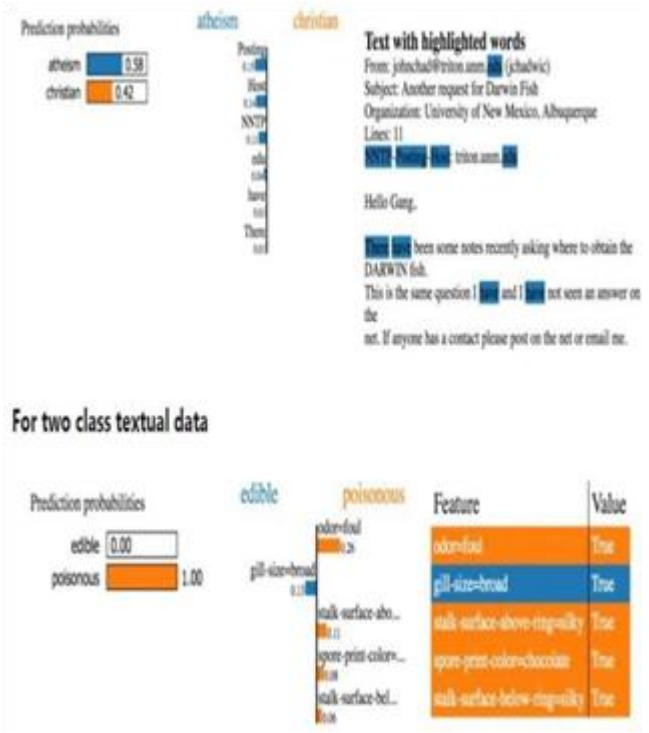


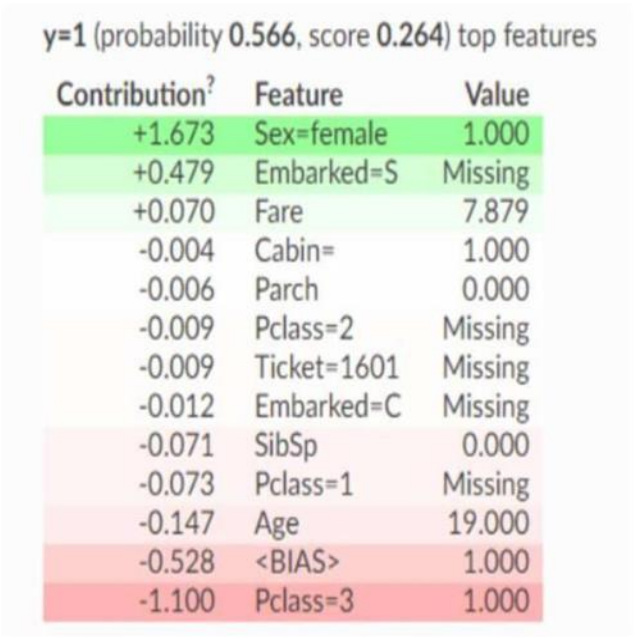
Figure 3: LIME Framework[18].

3. **ELI5**

ELI5, a Python library, aids in the debugging and depiction of ML classifiers, compatible with frameworks like Scikit-learn, Keras, XGBoost, Light GBM, and Cat Boost.[19].

It offers dual perspectives on classification or regression models, presenting them in accessible and informative formats.

(a) Analyze model parameters and attempt to understand how the model operates at a global level.
(b)Assess a specific ML model prediction to comprehend the factors influencing its final outcome and rationale.



y=1 (probability **0.566**, score **0.264**) top features

Contribution?	Feature	Value
+1.673	Sex=female	1.000
+0.479	Embarked=S	Missing
+0.070	Fare	7.879
-0.004	Cabin=	1.000
-0.006	Parch	0.000
-0.009	Pclass=2	Missing
-0.009	Ticket=1601	Missing
-0.012	Embarked=C	Missing
-0.071	SibSp	0.000
-0.073	Pclass=1	Missing
-0.147	Age	19.000
-0.528	<BIAS>	1.000
-1.100	Pclass=3	1.000

Figure 4: ELI5 Framework[19].

4. What-if Tool

Google's What-If Tool (WIT) enhances user comprehension of machine learning model operations by enabling performance assessment in diverse hypothetical scenarios. WIT facilitates the exploration of data impacts, visualization of model behavior, and evaluation of various fairness criteria. This Jupyter, Colaboratory, and Cloud AI Platform extension supports tasks like binary classification, multi-class classification, and regression, accommodating diverse data types including tabular, image, and text. Moreover, it integrates seamlessly with SHAP and LIME, amplifying its analytical capabilities and fostering a comprehensive understanding of complex model functionalities.

6. Conclusions:

Transparency serves as a cornerstone for establishing trust and reliability in artificial intelligence (AI), crucial for its widespread acceptance in markets and society. Despite the present limitations in achieving comprehensive transparency, addressing trust and accountability concerns assumes paramount importance. AI transparency necessitates a holistic examination, transcending individual algorithmic components and encompassing complexities arising from literacy variations, information asymmetries, and intricacies associated with explainability. Navigating these intricacies underscores the trade-offs and governance challenges inherent in AI transparency, requiring interdisciplinary research for effective management.

This research underscores the significance of transparency, highlighting the challenges posed by opaque systems and emphasizing the role of Explainable Machine Learning Frameworks in promoting transparency. Nonetheless, the critical reliance of these frameworks on training data presents a significant hurdle that demands careful consideration. Acknowledging these realities becomes imperative for understanding the far-reaching implications of data transparency and for devising effective governance strategies, signifying the evolving nature of AI transparency as a dynamic and multifaceted domain.

7. References:

[1].Cress, M. (2019). The Black Box Problem. http://artificialintelligencemania.com/2019/01/10 /the-blackbox-problem/
[2].O'Neil, C. (2016). Weapons of Math Destruction. https://dl.acm.org/doi/10.5555/3002861
[3].Pasquale, F. (2016). The Black Box Society. https://www.hup.harvard.edu/catalog.php?isbn=9 780674970847
[4].Diakopolous, N. (2016). Algorithmic Accountability Reporting: On the Investigation of Black Boxes. https://doi.org/10.7916/D8ZK5TW2
[5].Fox, J. (2007). The Uncertain Relationship between Transparency and Accountability. http://dx.doi.org/10.1080/09614520701469955
[6].Boyd, D. (2016). Transparency Does Not Equal Accountability. https://points.datasociety.net/transparencyaccoun tability-3c04e4804504
[7].Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. https://www.science.org/doi/10.1126/scirobotics. aan6080
[8].Brauneis, R., & Goodman, E. (2018). Algorithmic transparency for the Smart City. https://papers.ssrn.com/sol3/papers.cfm?abstract id=
[9].Wachter, L. F. S., & Mittelstadt, B. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. https://academic.oup.com/idpl/article/7/2/76/38 60948
[10].Datta, M. T. A., Datta, A. "Algorithmic transparency for the Smart City," 2015. https://www.sciendo.com/article/10.1515/popets- 2015-0007
[11].Schmelzer, R. "Towards A More Transparent AI," Forbes, 2020. https://www.forbes.com/sites/cognitiveworld/202 0/05/23/towards-a-more-transparent- ai/?sh=73edec43d937
[12].Fireman, K. "AI's lack of transparency triggers a debate over ethics," LSE Business Review, 2018. https://blogs.lse.ac.uk/businessreview/2018/06/1 4/ais-lack-of-transparency-triggers-a-debate-over- ethics/
[13].Card, D. "The "black box" metaphor in machine learning," Medium, 2017. https://dallascard.medium.com/the-black-box- metaphor-in-machine-learning-4e57a3a1d2b0
[14].Johnson, J. "Interpretability vs Explainability: The Black Box of Machine Learning," BMC Blog, 2020. https://www.bmc.com/blogs/machine-learning- interpretability-vs-explainability/#:%E2%88%BC:text=Interpretabilit y%20has%20to%20do%20with,Nets%2C%20to% 20justify%20the%20results
[15].Casey, K. "What is explainable AI?" The Enterprisers Project, 2019. https://enterprisersproject.com/article/2019/5/w hat-is-explainable-ai
[16].I. P. Ltd. "Explainable AI: A Transparent Future," Tech Break, 2021.https://medium.com/tech-break/explainable-ai-a- transparent-future-4a8b44bac564
[17].Dataman, D. "Explain Your Model with the SHAP Values," TowardsDataScience,2019. https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d
[18].Hulstaert, L. "Understanding model predictions with LIME," TowardsDataScience,2018. https://towardsdatascience.com/understanding-model-predictions-with-lime-a582fdff3a3b
[19].priyanka870, "Demystifying Model Interpretation using ELI5,"AnalyticsVidhya,2020. https://www.analyticsvidhya.com/blog/2020/11/ demystifying-model-interpretation-using-eli5/