Grenze International Journal of Engineering and Technology, June Issue

Expanding the Knowledge on Machine Learning: Explainable Models for Smarter Decisions

Shreya Jha[1], Rajarshi Mondal[2], Niloy Avro Mondal[3], Debjit Mondal[4], Shreyashi Saha[5], Md Tauhid Alam[6], and Sagarika Chowdhury[7]
[1-7]Narula Institute of Technology/Department of CSE, Kolkata, India
Email: shrish.2613@gmail.com {www.rajm.07, niloyavromondal8, mondaldebjit263, shreyashisaha2004, md1398353, sagsaha2004}@gmail.com

***Abstract*— In this paper, our objective is to shed light on previously unexplored and overlooked aspects of explainable machine learning. We aim to illustrate the diverse domains where machine learning has quietly become a transformative force. Through an extensive survey, we delve into existing research and illuminate the potential avenues yet to be explored. In an era marked by technological advancement, machine learning has seamlessly integrated into nearly every facet of global activities. To remain abreast of these developments, it is imperative to deepen our understanding of this field. Our goal is to immerse ourselves in the realm of machine learning, leveraging this knowledge to undertake a significant project. Within these pages, readers will find a comprehensive overview of our survey findings alongside insights into our own future endeavors in this dynamic field.**

***Index Terms*— complexity, explain-ability, explainable artificial intelligence, interpretability, , interpretable artificial intelligence, machine learning, methods.**

I. INTRODUCTION

This paper talks about, Explainable Machine Learning, which is an essential field in AI that focuses on deciphering complex machine learning models that aids in various driving the conclusion-making methods. Its objective is to provide understandable insights into how these models arrive at predictions or decisions, crucial for trust and accountability [1]. XAI employs various techniques like feature importance analysis, Evaluating the impact of input characteristics on predictions is crucial in building robust, interpretable, and efficient machine learning models. Approaches like permutation advantages, SHAP or LIME help identify influential features, enhancing comprehension of model behaviour. Moreover, it promotes transparent model types (e.g., decision trees) whose structure is more interpretable than complex models like neural networks [2]. The importance of XAI lies in fostering trust, ensuring regulatory compliance, identifying biases, and enabling experts to verify and enhance model performance. Its role is pivotal across sectors like healthcare, finance, and justice, ensuring responsible and ethical AI deployment in diverse applications [3].

A. Neural Networks

Neural networks in machine learning are a class of algorithms inspired by the human brain. Composed of interconnected nodes (neurons), they learn patterns from data, enabling tasks like classification and prediction.

Grenze ID: 01.GIJET.10.2.554_4

They contain input, hidden, and output layers, employing weights to process information iteratively for learning and decision-making

B. Explainable AI

Explainable AI in machine learning focuses on making AI models transparent and understandable, emphasizing interpretability to permit users to retrospect and have faith in the conclusion constructed by the models. It aims to provide insights into how algorithms reach conclusions, fostering trust and facilitating human comprehension of AI-driven outcomes. This paper explains the idea that we focus on, which revolves around the base of machine learning impacting various fields and making life utmost easier. copy.

II. Related Work

Deep Neural Networks (DNNs) are known as black boxes as it is impossible to describe network itself or an external component via their output [4]. Interpretability and transparency are crucial in healthcare and finance for regulatory compliance and ethical considerations [5]. XAI is being applied in various domains such as medical, fitness, finance, cyber security, education system and civil engineering [6]. Explain-ability refers to delivering the main notes to solve the aroused need whereas, interpretability refers to the degree in accordance to the bullet marks that reach out to the audience to expand their education range [7]. XAI aims to make AI better than translucent, increase trust, and avoid limiting its acceptance in important genres, such as security [8]. In high- demand sectors like finance, insurance, and healthcare, this is particularly pivotal [9] In AI, interpretability, and explain-ability (XAI) refers to the capacity to comprehend the logic behind the results drawn by AI models. This capacity is pivotal for several reasons, including as promoting acceptance and confidence among stakeholders [10-11]. Indeed, with the progress made in XAI exploration, there are still several issues to be resolved, like balancing interpretability and practicality, guarding sequestration, and security, and considering the mortal perspective while developing AI explanations [12-13]. Addressing these issues and icing that AI is applied responsibly and immorally will bear collaboration between experimenters and politicians. XAI can grease informed decision-timber, foster trust in AI systems, and open the door to a future in which AI maximizes mortal implicit and advances society by placing a high precedence on mortal-cantered and ethical AI [14]. The environment information handed describes colorful aspects of resolvable AI(XAI). XAI refers to AI systems that can give explanation for their opinions or prognostications to mortal druggies [15].

III. Survey

XAI can be distributed into two main orders knowledge- driven and data- driven styles. ML allows computers to identify patterns, make predictions and refine themselves through experience without being explicitly programmed [16-17]. DL on the opposite side stands affected by the composition and working of the human brain and has shown remarkable predicting performance. The article highlights the efficiency of explain-ability in AI models to increase users' trust models allow for better decision-making [18]. The environment information describes the need for explain-ability in the domain of artificial intelligence (AI) systems and the colourful subareas of AI that were included in the hunt [19]. The paper outlines the delineations of confirmation, verification, and testing, and emphasizes the significance of assessing AI systems at the system position to ensure that the results match what the real world created [20],[21]. The authors bandy the literal development of AI assurance styles and punctuate the recent focus on areas similar as resolvable AI (XAI), computer vision, deep literacy, and underpinning literacy [22]. It discusses the significance of interpretability in machine literacy models and the need for different approaches to give explanations at both the original and global situations [23]. Specifically, the environment mentions two main approaches for interpretability Model-independent and Model-tailored styles. Model-agnostic styles differentiate the description from the machine learning model used, allowing for further inflexibility in furnishing explanations. On the other hand, Model-Specific styles make utilization of the inner framework of a specific system to give explanations [24],[25]. It mentions the benefits of interpretability, including the capability to justify prognostications, diagnose vulnerabilities, and ameliorate models [26].

This document discusses the role of communication and the impact of technology in remote areas. [27]. The lack of technology in remote areas poses challenges to communication. Lower Earth Orbit (LEO) satellites are presented as a solution, offering advantages such as lower latency and higher data transfer rates compared to geostationary satellites [28]. Mesh networks are discussed as a decentralized and self-configuring network topology that allows direct device-to-device communication. They offer flexibility, scalability, and redundancy. The role of computer science in developing effective communication strategies for remote teams is highlighted,

including the use of artificial intelligence language translators and wireless mesh networks [29]. This document discusses the balance between precision and comprehensibility of complex models within the framework of big data [30]. The document introduces the concept of “explanation models”, which are depictable approximations of the authentic model. It presents a unified framework known as SHAP (Shapley Additive Explanations) for elucidating predictions, which includes accumulated feature attribution ways [31]. It proposes new SHAP value estimation methods that resonate with human understanding and are more efficient in distinguishing among model output categories [32].

This document discusses the concepts of explain-ability and clarity in the realm of artificial intelligence (AI). Explain-ability refers to providing information for a specific audience to accomplish a specific need, while interpretability focuses on the extent to which these insights can be comprehensible within the audience's field expertise [33]. The document emphasizes the required explanations and interpretations in opaque AI systems. The document also mentions the obstacles and avenues for investigation in explainable AI (XAI) and highlights the importance of XAI in disclosing scientific insights derived from opaque AI models [34]. Interpretability and explain ability are important concepts in employment of machine learning. They serve an essential function in ML model design and development [35]. An Overview of the Explain Ability of Guided Machine Learning by Nadia Burkart and Marco F. Huber is an extensive review of the domain of explainable supervised machine learning (SML). The authors also discuss the different approaches to explaining SML models, including both local and global explanations [36]. One of the key contributions of the paper is its classification of SML approaches into three categories: clear models, substitute of model fitting, and explanation generation [37].

The paper also discusses the different dimensions of SML, such as the category of model (black box or interpretable), the kind of explanation (local or global), and the purpose of the explanation (understanding, debugging, or trust) [38]. Explainable Artificial Intelligence (XAI) emerges as a crucial field dedicated to addressing this challenge by providing intuitive thoughts into the inner mechanism of AI models, allowing users to comprehend and trust their decisions [39]. This paper delves into the diverse landscape of XAI approaches, classified into three primary groups: Inherently Interpretable Approaches, Model-Independent Techniques, and Instance-Based Explanations [40]. Model-agnostic methods, however, cater to a wider range of models, employing descriptions that fit to the model irrespective of its architecture [41]. Fictitious details can be used to provide practical insights into model predictions by allowing us to change discrete instances as a path to reach a desired result [42]. Machine learning (ML) models are becoming increasingly powerful and widespread, but their complexity often makes it difficult to acknowledge how they make decisions. This absence of interpretability can lead to concerns about trust, fairness, and debugging. Interpretable Machine Learning (IML) is an area of exploration that achieves to build methods for making ML models more explainable [43]. It begins with an introduction to the field, discussing the importance of interpretability and the distinct types of interpretability techniques [44]. Tuxedos and suits may seem similar, but their differences stand out in several aspects. Tuxedos distinguish themselves with satin elements. In contrast, suits emphasize a broader range of colours and patterns using their fabric [45]. Ultimately, tuxedos rule the formal scene, especially at black-tie events, epitomizing sophistication. Suits, versatile in colour and style, adapt easily to diverse occasions [46-47].

Building trust and understanding around machine learning (ML) requires effective explanations for external stakeholders. This paper sheds light on the crucial role of community engagement, context-specific explanations, and robust evaluation methods [46], [48]. Battling ICU alarm overload, researchers propose a targeted approach using explainable machine learning. This roadmap for reshaping ICU alarms prioritizes age-specific factors, ultimately aiming to elevate patient care in critical settings [49]. The paper delves into the utilization of knowledge graphs within Explainable Machine Learning. It focuses on integrating knowledge graphs, which inherently offer domain-specific facts in a machine-readable layout, into XML methodologies [50]. Acknowledging the rising interest in combining Knowledge Representation (KR) techniques with machine learning, the study hypothesizes that knowledge graphs can play a pivotal role in enriching XML approaches. The concept of neuro-symbolic AI, merging symbolic methods for knowledge representation and sub symbolic approaches capable of handling vast data, serves as a foundation for this exploration [51]. The study employs a systematic literature review approach, exploring various real life machine learning applications that integrate definite educational stuff through graphs to incorporate descriptions [52-53]. By evaluating the integration of knowledge within XML through systematic analysis across various machine learning domains, the study aims to delineate the benefits and constraints of employing knowledge graphs as a supportive background for explainable AI systems [54].

The potential risks are the interpretability and audibility of AI and Machine Learning in different industries including financial affairs [55]. The authors highlight the topics which are needed for human agency and oversight in decision-making techniques. Particularly, the relevancy lies in the credit scoring models where the response

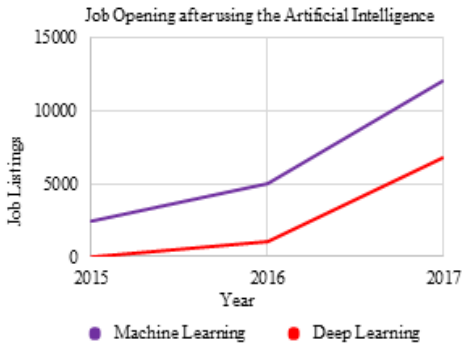
variable is influenced by the explanatory variables. Compound models of machine learning like neural networks and tree models can offer high level of accuracy which often lacks interpretability [56]. Many companies have embraced the idea of explainable AI models. These models convey the details and causes to the proper functioning of AI clearly and easily to understand. The methods focus on the analysing during post-processing phase, rather than processing [57]. Empirical analysis examines 15,000 small and medium-sized companies seeking credit. Risky and non-risky borrowers are assembled based on financial characteristics which explains the credit scores and predicts future behaviour [58]. Explain-ability methods aim to shed light on the knowledge about the procedure of the model to make it more interpretable, but popular models like neural networks and random forests are too complex to be interpretable, even in small variations [59]. Therefore, the explain-ability methods rescue the idea of interpretation to be more understandable. Due to the task's context-sensitive nature, it is quite difficult for the development of the automatically deployed explain-ability method [60].
Explainable machine learning (ML) models are in two different categories- Interpretable Models and Explainable Models. Interpretable models are defined as models that are understandable by humans while Explainable models provide explanations for the decision-making process [61]. These approaches consist of hybrid information sources that combine prior knowledge and data, representation, and addition of the knowledge to the ML pipeline. Current research in explain-ability focuses on three main classifications: Integrating knowledge into the ML pipeline, integrating knowledge into the explain-ability method, and deriving knowledge from explanations [62].

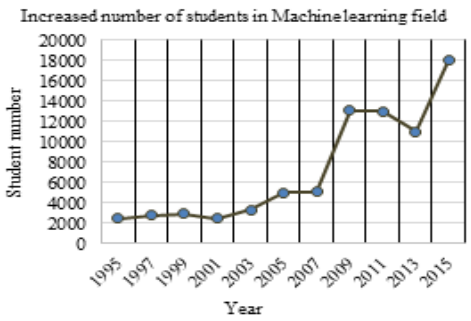
IV. Results And Discussion

Here, we talk about the various advantages put forward by various authors and the future scope that each paper had. A proper comparison allows us to properly introspect the details and gain knowledge. The representation through bar graphs shows how the world is affected by the advent of artificial intelligence. The pie chart shows the progress in the generative machine learning field over the decade. We also have put forward a comparative graph based on revenue.

A. Graphical Analysis

The graphs show the analysis of how the employment increased after the advent of artificial intelligence into the job sector. Where Blue line indicates Machine Learning and the Red line indicates the Deep Learning.
Both Fig 1 and Fig 2, vividly illustrate the global upswing in job opportunities.





B. Revenue Comparison

Every country has started spending a good amount on implementation of machine learning in various sectors for their development. The graph shows the statistical comparison between different countries and the amount they spend over artificial intelligence and machine learning.
The graph shows the money spent on machine learning by different countries. Fig 5 talks about the difference in money spent by various regions.

C. Vivid Comparison

The following table enlists the advantages that each paper brought and the advancements caused by the research work of each author in their respective paper domain. It also talks about the areas that can still be explored and with which better facilities can be provided in the respective genres. Table I. shows the comparison between the advancements done and the areas still left to explore.

Figure 5. Comparison between different countries on the basis of revenue

TABLE I. DETAILED ANALYSIS BETWEEN THE PROS AND THE FUTURE ADVANCEMENT AREAS OF THE PAPERS WRITTEN BY THE FOLLOWING AUTHORS

Paper Name	Advantages	Future Scope
Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges [5]	Enhanced public confidence in AI: Lifting the veil on AI decision-making fosters trust and acceptance of its outcomes. Improves debugging and model improvement.	Development of new explain-ability metrics. Increased accessibility for non-experts. Integration into AI development workflows.
An Empirical Survey on Explainable AI Technologies: Recent Trends, Use-Cases, and Categories from Technical and Application Perspectives [9]	Promotes acceptance and trust among stakeholders. Supports effective human-AI collaboration. Enables debugging and system improvement.	Balance interpretability and practicality. Protect privacy and security. Consider the human perspective while developing AI explanations.
A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends [15]	Increases trust and acceptance of AI systems. Enables better debugging and optimization of AI systems. Enhances the comprehensibility and clarity of AI models.	Development of more accurate and efficient XAI methods. Application of XAI to a wider range of AI domains. Development of XAI methods for explaining complex models.
A survey on artificial intelligence assurance [19]	Increased trust and acceptance of AI systems. Improved debugging and optimization of AI systems. Enhanced interpretability and explain-ability of AI models.	Development of more accurate and efficient assurance methods. Application of assurance to a wider range of AI domains. Development of assurance methods for complex AI systems.
Individual Explanations in Machine Learning Models: A Survey for Practitioners [23]	**Fortified user trust and facilitated acceptance of machine learning user friendly models.** Improved debugging and optimization of machine learning models. Enhanced fairness and non-discrimination of machine learning decision-making.	Increased user trust and extensive acceptance of different types of machine learning projects**.** Enhanced troubleshooting and enhancement of machine learning models.
Explainable Artificial Intelligence: A Survey [27]	Easy to understand and interpret. Applicable to any machine learning model. Can offer causal explanations for individual predictions.	Develop new XAI methods that are more accurate and efficient. Apply XAI to a wider range of applications, such as healthcare and finance. Develop XAI methods that can explain complex models, such as deep artificial neural networks.

Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities [30]	**Increased trust:** XAI helps to improve trust in AI systems by increasing their transparency and comprehensibility. This will be important for the extensive utilization of AI, specifically in critical areas like wellness program and finance.	**Integration into the AI development process:** XAI techniques are likely to become more integrated into the AI development procedures, starting from the beginning phases of design and modelling to the deployment and maintenance of AI systems.
A Unified Approach to Interpreting Model Predictions [33]	Increases trust and understanding of AI systems Enables debugging and improvement of AI systems.	Develop more accurate and efficient XAI methods. **Utilize XAI in a broader spectrum of applications, including healthcare and security.**
Interpretable and explainable machine learning: A methods-centric overview with concrete examples [35]	Builds trustworthy and reliable ML models. Various interpretable models and techniques exist.	Develop more accurate and efficient interpretable models and techniques Apply explainable models and methodologies to a broader range of applications.

D. Graphical Analysis

This section talks about the increase in machine learning users in the recent decades. Fig 3 and Fig 4 illustrate the increase in usage of machine learning through the years.

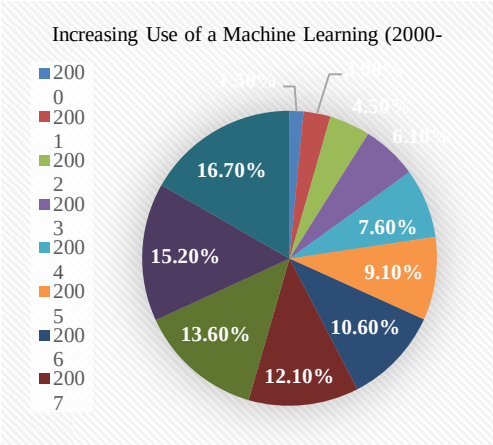


Figure 3. Increase in the use of ML from the year 2000-2010

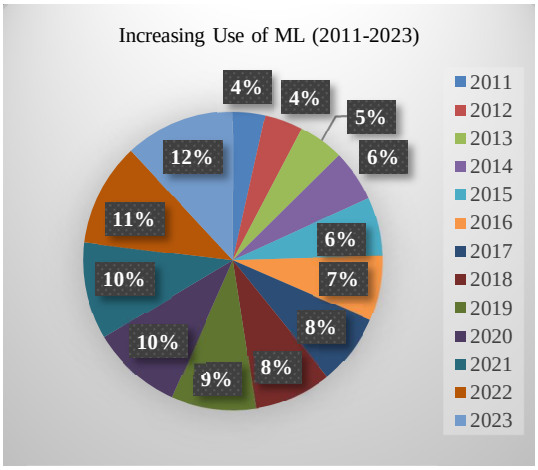


Figure 4. Increase in the use of ML from the year 2011-2023

V. Conclusion And Future Scope

In conclusion, Explainable Machine Learning (XAI) stands as an important pursuit within the monarchy of artificial intelligence, notifying the imperative need for translucent, accountability, and trust in increasingly difficult machine learning models. By elucidating the controlling processes of these models, XAI techniques contribute significantly to fostering trust among users, stakeholders, and regulatory bodies [5], [14].

The field of XAI not only facilitates regulatory compliance but also empowers experts to collaborate effectively with AI systems. By providing understandable insights into model behaviour, XAI enables domain experts to validate, refine, and improve model performance, leveraging their expertise for the benefit of the community [19].

In essence, the pursuit of Explainable Machine Learning aligns along with broader aim of creating responsible, translucent, and beneficial AI systems that enhance decision-making, enable ethical practices, and contribute positively to various facets of human life [35].

The Future of XAI lies on transparency, trust, and beyond. XAI's future is about going beyond “explain-ability” to true AI understanding.

After a vivid survey of the papers, we observe that there are still fields that can be improved and updated further using machine learning.

This future XAI fosters trust, transparency, and responsible AI development. It empowers users, fuels scientific progress, and paves the way for a greater way of living where AI and humans collaborate seamlesssly.

REFERENCES

[1] Zhang, Yu, Peter Tiňo, Aleš Leonardis, and Ke Tang. "A survey on neural network interpretability." IEEE Transactions on Emerging Topics in Computational Intelligence 5, no. 5 (2021): 726-742.

[2] Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information fusion 58 (2020): 82-115.

[3] Angelov, Plamen P., Eduardo A. Soares, Richard Jiang, Nicholas I. Arnold, and Peter M. Atkinson. "Explainable artificial intelligence: an analytical review." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11, no. 5 (2021): e1424.

[4] Yang, Wenli, Yuchen Wei, Hanyu Wei, Yanyu Chen, Guan Huang, Xiang Li, Renjie Li et al. "Survey on Explainable AI: From Approaches, Limitations and Applications Aspects." Human-Centric Intelligent Systems 3, no. 3 (2023): 161-188.

[5] Xu, Feiyu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. "Explainable AI: A brief survey on history, research areas, approaches and challenges." In Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8, pp. 563-574. Springer International Publishing, 2019.

[6] Gunning, David, and David Aha. "DARPA's explainable artificial intelligence (XAI) program." AI magazine 40, no. 2 (2019): 44-58.

[7] Saeed, Waddah, and Christian Omlin. "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities." Knowledge-Based Systems 263 (2023): 110273.

[8] Adadi, Amina, and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)." IEEE access 6 (2018): 52138-52160.

[9] Nagahisarchoghaei, Mohammad, Nasheen Nur, Logan Cummins, Nashtarin Nur, Mirhossein Mousavi Karimi, Shreya Nandanwar, Siddhartha Bhattacharyya, and Shahram Rahimi. "An empirical survey on explainable ai technologies: Recent trends, use-cases, and categories from technical and application perspectives." Electronics 12, no. 5 (2023): 1092.

[10] Preece, Alun. "Asking 'Why'in AI: Explain-ability of intelligent systems–perspectives and challenges." Intelligent Systems in Accounting, Finance and Management 25, no. 2 (2018): 63-72.

[11] Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." Artificial intelligence 267 (2019): 1-38.

[12] Lundberg, Scott M., Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston et al. "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery." Nature biomedical engineering 2, no. 10 (2018): 749-760.

[13] Vilone, Giulia, and Luca Longo. "Classification of explainable artificial intelligence methods through their output formats." Machine Learning and Knowledge Extraction 3, no. 3 (2021): 615-661.

[14] Agarwal, Garvita, Lauren Hay, Ia Iashvili, Benjamin Mannix, Christine McLean, Margaret Morris, Salvatore Rappoccio, and Ulrich Schubert. "Explainable AI for ML jet taggers using expert variables and layerwise relevance propagation." Journal of High Energy Physics 2021, no. 5 (2021): 1-36.

[15] Saranya, A., and R. Subhashini. "A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends." Decision analytics journal (2023): 100230.

[16] Minh, Dang, H. Xiang Wang, Y. Fen Li, and Tan N. Nguyen. "Explainable artificial intelligence: a comprehensive review." Artificial Intelligence Review (2022): 1-66.

[17] Walia, Savita, Krishan Kumar, Saurabh Agarwal, and Hyunsung Kim. "Using xai for deep learning-based image manipulation detection with shapley additive explanation." Symmetry 14, no. 8 (2022): 1611.

[18] Meena, Jaishree, and Yasha Hasija. "Application of explainable artificial intelligence in the identification of Squamous Cell Carcinoma biomarkers." Computers in Biology and Medicine 146 (2022): 105505.

[19] Batarseh, Feras A., Laura Freeman, and Chih-Hao Huang. "A survey on artificial intelligence assurance." Journal of Big Data 8, no. 1 (2021): 60.

[20] Chen, Han-Yun, and Ching-Hung Lee. "Vibration signals analysis by explainable artificial intelligence (XAI) approach: Application on bearing faults diagnosis." IEEE Access 8 (2020): 13324246-134256.

[21] Dağlarli, Evren. "Explainable artificial intelligence (xAI) approaches and deep meta-learning models." Advances and applications in deep learning 79 (2020).

[22] Dodge, Jonathan, and Margaret Burnett. "Position: We Can Measure XAI Explanations Better with Templates." In ExSS-ATEC@ IUI, pp. 1-13. 2020.

[23] Carrillo, Alfredo, Luis F. Cantú, and Alejandro Noriega. "Individual explanations in machine learning models: A survey for practitioners." arXiv preprint arXiv:2104.04144 (2021).

[24] Adadi, Amina, and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)." IEEE access 6 (2018): 52138-52160.

[25] Apley, Daniel W., and Jingyu Zhu. "Visualizing the effects of predictor variables in black box supervised learning models." Journal of the Royal Statistical Society Series B: Statistical Methodology 82, no. 4 (2020): 1059-1086.

[26] Dabkowski, Piotr, and Yarin Gal. "Real time image saliency for black box classifiers." Advances in neural information processing systems 30 (2017).

[27] Došilović, Filip Karlo, Mario Brčić, and Nikica Hlupić. “Explainable artificial intelligence: A survey.” In 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), pp. 0210-0215. IEEE, 2018.
[28] Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Müller. “Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models.” arXiv preprint arXiv:1708.08296 (2017).
[29] Cortez, Paulo, and Mark J. Embrechts. “Using sensitivity analysis and visualization techniques to open black box data mining models.” Information Sciences 225 (2013): 1-17.
[30] Saeed, Waddah, and Christian Omlin. “Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities.” Knowledge-Based Systems 263 (2023): 110273.
[31] Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García et al. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.” Information fusion 58 (2020): 82-115.
[32] Gade, Krishna, Sahin Geyik, Krishnaram Kenthapadi, Varun Mithal, and Ankur Taly. “Explainable AI in industry: Practical challenges and lessons learned.” In Companion Proceedings of the Web Conference 2020, pp. 303-304. 2020.
[33] Lundberg, Scott M., and Su-In Lee. “A unified approach to interpreting model predictions.” Advances in neural information processing systems 30 (2017).
[34] Štrumbelj, Erik, and Igor Kononenko. “Explaining prediction models and individual predictions with feature contributions.” Knowledge and information systems 41 (2014): 647-665.
[35] Marcinkevičs, Ričards, and Julia E. Vogt. “Interpretable and explainable machine learning: A methods-centric overview with concrete examples.” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (2023): e1493.
[36] Burkart, Nadia, and Marco F. Huber. “A survey on the explain-ability of supervised machine learning.” Journal of Artificial Intelligence Research 70 (2021): 245-317.
[37] Su, Guolong, Dennis Wei, Kush R. Varshney, and Dmitry M. Malioutov. “Learning sparse two-level boolean rules.” In 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1-6. IEEE, 2016.
[38] Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Müller. “Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models.” arXiv preprint arXiv:1708.08296 (2017).
[39] Islam, Sheikh Rabiul, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. “Explainable artificial intelligence approaches: A survey.” arXiv preprint arXiv:2101.09429 (2021).
[40] Došilović, Filip Karlo, Mario Brčić, and Nikica Hlupić. “Explainable artificial intelligence: A survey.” In 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), pp. 0210-0215. IEEE, 2018.
[41] Tjoa, Erico, and Cuntai Guan. “A survey on explainable artificial intelligence (xai): Toward medical xai.” IEEE transactions on neural networks and learning systems 32, no. 11 (2020): 4793-4813.
[42] Verma, Sahil, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, and Chirag Shah. “Counterfactual explanations and algorithmic recourses for machine learning: A review.” arXiv preprint arXiv:2010.10596 (2020).
[43] Vellido, Alfredo, José David Martín-Guerrero, and Paulo JG Lisboa. “Making machine learning models interpretable.” In ESANN, vol. 12, pp. 163-172. 2012.
[44] Doshi-Velez, Finale, and Been Kim. “Towards a rigorous science of interpretable machine learning.” arXiv preprint arXiv:1702.08608 (2017).
[45] Bhatt, Umang, McKane Andrus, Adrian Weller, and Alice Xiang. “Machine learning explainability for external stakeholders.” arXiv preprint arXiv:2007.05408 (2020).
[46] Weller, Adrian. “Transparency: motivations and challenges.” In Explainable AI: interpreting, explaining and visualizing deep learning, pp. 23-40. Cham: Springer International Publishing, 2019.
[47] Doshi-Velez, Finale, and Been Kim. “Towards a rigorous science of interpretable machine learning.” arXiv preprint arXiv:1702.08608 (2017).
[48] Amarasinghe, Kasun, Kit T. Rodolfa, Hemank Lamba, and Rayid Ghani. “Explainable machine learning for public policy: Use cases, gaps, and research directions.” Data & Policy 5 (2023): e5.
[49] González-Nóvoa, José A., Laura Busto, Juan J. Rodríguez-Andina, José Fariña, Marta Segura, Vanesa Gómez, Dolores Vila, and César Veiga. “Using explainable machine learning to improve intensive care unit alarm systems.” Sensors 21, no. 21 (2021): 7125.
[50] Tiddi, Ilaria, and Stefan Schlobach. “Knowledge graphs as tools for explainable machine learning: A survey.” Artificial Intelligence 302 (2022): 103627.
[51] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. “Anchors: High-precision model-agnostic explanations.” In Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1. 2018.
[52] Biran, Or, and Courtenay Cotton. “Explanation and justification in machine learning: A survey.” In IJCAI-17 workshop on explainable AI (XAI), vol. 8, no. 1, pp. 8-13. 2017.
[53] Miller, Tim. “Explanation in artificial intelligence: Insights from the social sciences.” Artificial intelligence 267 (2019): 1-38.
[54] Bellucci, Matthieu, Nicolas Delestre, Nicolas Malandain, and Cecilia Zanni-Merk. “Une terminologie pour une IA explicable contextualisée.” In EXPLAIN'AI Workshop EGC 2022. 2022.

[55] Bussmann, Niklas, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. “Explainable machine learning in credit risk management.” Computational Economics 57 (2021): 203-216.
[56] Molnar, Christoph. Interpretable machine learning. Lulu. com, 2020.
[57] Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. “Definitions, methods, and applications in interpretable machine learning.” Proceedings of the National Academy of Sciences 116, no. 44 (2019): 22071-22080.
[58] Bracke, Philippe, Anupam Datta, Carsten Jung, and Shayak Sen. “Machine learning explainability in finance: an application to default risk analysis.” (2019).
[59] Beckh, Katharina, Sebastian Müller, Matthias Jakobs, Vanessa Toborek, Hanxiao Tan, Raphael Fischer, Pascal Welke, Sebastian Houben, and Laura von Rueden. “Explainable machine learning with prior knowledge: an overview.” arXiv preprint arXiv:2105.10172 (2021).
[60] Wexler, James, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. “The what-if tool: Interactive probing of machine learning models.” IEEE transactions on visualization and computer graphics 26, no. 1 (2019): 56-65.
[61] Von Rueden, Laura, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch et al. “Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems.” IEEE Transactions on Knowledge and Data Engineering 35, no. 1 (2021): 614-633.
[62] Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Müller. “Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models.” arXiv preprint arXiv:1708.08296 (2017).