Untangling Explainable AI in Applicative Domains: Taxonomy, Tools, and Open Challenges

Chapter · July 2023

DOI: 10.1007/978-981-99-1479-1_63

6 authors, including:

Sachi Chaudhary
Nirma University
9 PUBLICATIONS **51** CITATIONS
SEE PROFILE

Pronaya Bhattacharya
Amity University
170 PUBLICATIONS **3,054** CITATIONS
SEE PROFILE

Vivek Kumar Prasad
Nirma University
53 PUBLICATIONS **477** CITATIONS
SEE PROFILE

Rushabh Shah
Nirma University
7 PUBLICATIONS **89** CITATIONS
SEE PROFILE

Untangling Explainable AI in Applicative Domains: Taxonomy, Tools, and Open Challenges

Sachi Chaudhary, Pooja Joshi, Pronaya Bhattacharya, Vivek Kumar Prasad, Rushabh Shah, and Sudeep Tanwar

Abstract Recently, a paradigm shift is observed toward Industry 5.0, where tasks (processes) are automated at massive scales. This shift has initiated modern developments in artificial intelligence (AI) to support a plethora of applications like manufacturing, health care, vehicular net- works, and others. However, owing to the black-box nature of AI models, the research has shifted toward the proposal of novel techniques that aim toward the explainability and validity of these AI models. Thus, explainable AI (XAI) has become a norm in modern applicative domains, and the study of its frameworks and tools has become the buzzword among researchers. Thus, the paper intends to present the key concepts of XAI and aims at improving the model transparency. The survey systematically untangles the key concepts of XAI and presents a solution taxonomy in different applications. Modern XAI techniques are classified as self-explanatory, visual-based-model-agnostic, global surrogate, and local surrogate-model-agnostic. We also cover the tools and frameworks of XAI and discuss the open issues and challenges in practical realization. Thus, the survey intends to arm AI practitioners to design optimal solutions to realize XAI in practical use-case setups.

Keywords Explainable artificial intelligence · Machine learning · Applications · Domains · XAI tools · XAI frameworks

S. Chaudhary · P. Joshi · V. K. Prasad · R. Shah (✉) · S. Tanwar
Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat 382481, India
e-mail: rushabh.shah@nirmauni.ac.in

V. K. Prasad
e-mail: vivek.prasad@nirmauni.ac.in

S. Tanwar
e-mail: sudeep.tanwar@nirmauni.ac.in

P. Bhattacharya
Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University, Kolkata, West Bengal 700135, India
e-mail: pbhattacharya@kol.amity.edu

S. Tanwar et al. (eds.), *Proceedings of Fourth International Conference on Computing, Communications, and Cyber-Security*, Lecture Notes in Networks and Systems 664,
https://doi.org/10.1007/978-981-99-1479-1_63

1 Introduction

Industry 5.0 has shifted its operational core toward rapid personalization and customization, where the user is at the center of the ecosystem. Thus, automation in industrial processes has become a usual norm, and different applicative domains have shifted toward the use of artificial intelligence (AI) as a potential tool. However, the increased demand for AI deployment in applications is hindered due to potential limitations in data requirements, training process, and addressing system biases [1].

Biases in AI models occur as the model is trained only using a particular process or data type. Moreover, the cost involved in storing, classifying, mining, and training data are enormous, with little and no explainability of the obtained outputs. This opens the door for potential attackers to sabotage the system and launch severe adversarial attacks. Trust and safety in AI models are other crucial aspects of AI systems. Various AI architectures lack ethics in data processing and decision-making, and the privacy of the users is sometimes put at stake. The AI algorithms are necessary to be comprehended to identify the results of these algorithms. This process is called the "black box" which, makes the output infeasible to explain.

Thus, understanding these AI-based models can solve their drawbacks and help in complying to meet the regulatory standards, which builds confidence in the outputs. This is where the real need for explainable AI (XAI) arises. XAI is a process that facilitates the users to understand and trust the decision made by the framework implemented using machine learning and AI algorithms. XAI positively impacts the fairness, accuracy, outcome, and transparency of the AI-based model decision-making. XAI helps in two ways provide explanations to classify data in a better way and provide explanations to comprehend the AI framework better.

XAI provides a better user retention rate and superior inventory control. There are many techniques to execute the XAI framework. Some of the commonly implemented techniques are as follows.

- XAI through data visualization technique—It is an elementary and effective technique in providing explanations. However, it does not shows the importance of attributes that lead to a particular prediction.
- XAI through Shapley Additive exPlanations (SHAP)—It helps in decision-making as well as understanding the AI model and uses a standard approach for explaining different domain models, but it can have intensive computing.
- XAI through deep neural network models—It is used to explain the process as well as guide decision-making through the threshold values of the attributes.
- XAI through decision tree technique of machine learning (ML)-works similar to the neural network technique, but more number of levels may prove hard to understand.
- XAI through logistic regression model of ML—It is a widely used technique and also provides the weight of the factors in classification.

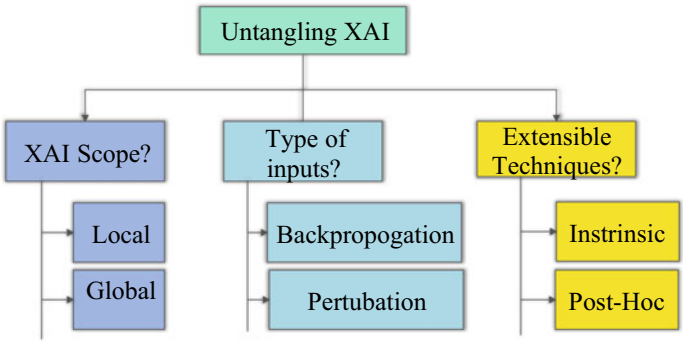

Fig. 1 Untangling XAI: The key pillars in applicative domains

Thus, it becomes imperative to untangle the XAI components and study its key pillars in applicative domains. Figure 1 presents an overview of the general categorization of XAI in terms of its scope, input, or methodology followed, and the techniques used to develop XAI.

The XAI-scope represents what the XAI framework is mainly being designed for. Is the method for local instance that is proposed for individual data instances or for one or more instances that is global scope used to study the whole model using XAI?

The type of inputs represents what methodology or algorithm is to be followed. The back-propagation is a core algorithm that provides gradients which are used in a back-propagated manner as the input to the next layer. Perturbation depends on the logic that changes done to the features present in input data are carefully selected and then implemented.

Lastly, the extensible techniques provide us with an overview of how an XAI model is developed. Is it customized for a particular use or can be used for any such application in general? The first technique is intrinsic where the XAI technique is implemented along with the neural network and cannot be further used in any other architectures. In the post-hoc technique, the XAI architecture is not dependent on any model and can be used with any of the trained neural networks.

1.1 Contributions and Layout

Following are the research contributions of the article.

1. The background and techniques of XAI (intrinsic and post-doc) are highlighted and key comparisons are derived.
2. An applicative-centric taxonomy of XAI is presented, and the potential benefits in diverse applications are discussed.

3. Tools and frameworks of XAI are presented as per the taxonomy to indicate its viability in real ecosystem setups. Next, we discuss the open issues and potential research directions for using XAI in diverse applicative domains.

The layout of the article is designed as follows. Section 2 presents the related work in XAI in different industrial domains. Section 3 presents the background of XAI which addresses the basic XAI algorithm in terms of scope, inputs, and extensible techniques. Section 4 presents a solution taxonomy in applicative domains. Section 5 presents the open issues and challenges in deploying XAI in applications, and Sect. 6 presents the concluding remarks and future scope of the survey.

2 Related Work

The section presents the use of XAI as a tool to support the mistrust against AI models. Authors in [12] proposed XAI for trust in AI-enabled decision models through a proposed data and knowledge engineering (DKE) framework which is cost-effective for added intelligence in model training. Li et al. [13] classified methods for the data-driven AI models where the explanation is due to task-oriented data, and knowledge-enabled techniques where external knowledge is included. Kuhn et al. [14] presented the combinatorial methods that have been used for developed form faulty location detection. Here, a simple technique is used to make the justification of classification easy by recognizing the combinations of attributes that are present as the members of justified classes and missing or the attributes that are rare in the non-member's classes. Ye et al. [15] developed the XAI model for computerized tomography (CT) scan classification that provided a qualitative and quantitative process to improve clinical decision-making. Authors in [16] proposed X-ray images for identifying tuberculosis and making required decisions using the Internet-of-Things (IoT) and AI methods in open environments.

In health care, authors in [17] proposed an XAI model to diagnose health statistics and provide transparency, explainability, reliability, and improvement in the medical domain. AI and ML models are also used for software engineering practices to carry out various decision-making practices for better software analytics. However, these models are still not completely actionable and explainable. Hence, Tantithamthavorn et al. [18] introduced XAI for software engineering that provides more actionable, interpretable, and practical prediction models.

Table 1 represents the relative study of the state-of-the-art XAI frameworks. All of these systems proposed by various researchers contain different techniques and have been used for different domains to promote practical, explainable, and trustworthy decision-making.

Table 1 Comparative analysis of different state-of-the-art XAI frameworks

Author	Year	Domain implemented	Purpose	Pros	Cons
D'Alterio et al. [2]	2020	Type-2 Fuzzy Classification	Presented a constrained meantime type-2 fuzzy recognition model for XAI	Highly interpretable, explanations are available in any natural language after classifications are made	Design of such systems is challenging
Deramgozi et al. [3]	n2021	Facial Expression Recognition	A hybrid AI explainable model, using CNN to recognize facial expression, and facial action unit for explainability	Good accuracy on the worked dataset, overall high precision	Lack complete explainability and reliability
Kuppa et al. [4]	2021	Cyber-Security	Proposes an attack methodology that supports XAI models from attacks and threats	Provides privacy and confidentiality, increases trust and builds confidence for models	New attacks that originate cannot be explained efficiently
Wang et al. [5]	2021	Feature Detector	Presented an XAI framework for feature detection in ultrasound copies	Achieves high robustness and accountability	Less performance as not linked with the advanced deep neural networks
Chen et al. [6]	2021	Breast Cancer Diagnosis	Proposed an XAI framework for breast cancer diagnosis using mammography results	Highly interpretable, efficiently extracts content from reports, high accuracy	Less correlation among the features used
Malhotra et al. [7]	2021	Blockchain Enabled Audit Trailing	Blockchain enabled architecture for authenticating XAI decisions	Highly secure auditing, good storage due to IPFS, supervision through smart contracts	High cost, traceability issues of stored explanations

(continued)

Table 1 (continued)

Author	Year	Domain implemented	Purpose	Pros	Cons
Lavrenovs et al. [8]	2021	Classifying Devices	Classification of various devices using XAI on Internet	Better classification accuracy, used for various devices, gives the most notable decisions	Increasing variety of devices proves to be challenging in classification
Zolanvari et al. [9]	2021	Trust XAI	Transparent, reliant based on the probability theory XAI model for explanations of ML and AI	Acceptable for many applications, issues explanations for any random sample, has high speed and performance	Challenging to design as fast as required XAI models
Pawar et al. [10]	2022	XAI and health care	XAI is described as a method for diagnosing and analyzing healthcare data	The usage of XAI in the healthcare industry has been highlighted for its transparency and traceability	The black-box functioning of AI model selections has an issue due to the lack of accountability and confidence
Adams et al. [11]	2022	XAI and Financial Sector	Type-2 fuzzy logic and its explainability were validated using the XAI	This approach intends to close the knowledge gap that exists between the fields of financial services and artificial intelligence	Limited knowledge of Type-2 Fuzzy Logic's capability as a highly flexible system in the financial sector

3 Explainable AI—Scope, Inputs, and Techniques

3.1 XAI—The Basic Preliminaries

XAI was coined by Lent et al. [19] in 2004. Later, in 2019, the defense advanced research projects agency (DARPA) developed the XAI Program, which has been instrumental in providing a comprehensive guideline to XAI, with descriptions of new techniques to make XAI more efficient and understandable.

There are mainly 2 terms for the data generated by the XAI system, which are explanations and interpretations, as explained by Palacio et al. [20]. These two terms are not clearly defined in the XAI community and literature wherein they are often used interchangeably despite having a difference in meaning.

- *Explanations*: Explanations are mainly a collection of relevant statements for human interpretation that are generated by the system and can give extra information. These statements provide insights into why the AI/ML model made a certain decision.
- *Interpretations*: Interpretations are mainly concerned with the characteristics that influenced the judgments. Interpretations are a useful tool which helps in identifying what data points and types of data are required by the system for it to provide a specific conclusion, using various metrics.

The US National Institute of Standards and Technology (NIST) provided four XAI principles that can serve as a basis for a good explanation.

- *Explanation*: Evidence-based explanations should be provided by AI systems.
- *Meaningful:* Explanations should be understandable in a way that users of the AI system can understand.
- *Accuracy*: Explanations of the AI system's process should be accurate.
- *Limits*: AI systems should work within the parameters for which they were created.

There are various techniques for XAI, which use the above-mentioned principles, and which have been broadly categorized into two—local and global. Some of these techniques are discussed in the next subsection.

3.2 Techniques of Explainable AI

XAI methods are classified based on two parameters—scope and implementation level. XAI techniques can be classified into two categories based on the scope of the explanation-local and global explanations. Local explanations provide explanations of individual outputs of a model, by focusing on a single input data instance. On the other hand, globally explainable techniques involve multiple inputs to generate explanations for the general behavior of the ML model.

Figure 2 presents a representation of the local and global explainability [21]. The subfigure on the left side indicates that the ML model is a black box B, on which certain inputs I are supplied. An input instance I_s is selected for which the XAI module presents an explanation, termed as I_q. Alternatively, on the right side, we see that inputs I_1, I_2,..., I_k are supplied to the ML model, where we find the casual (conditional) effect of all the inputs I on the output $O = O_1$, O_2, …, O_k by the XAI model.

We further deep-dive into XAI to understand the XAI at the implementation level, where the explainability algorithms are divided into two categories, namely

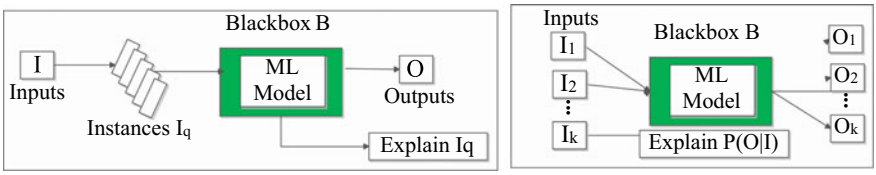

Fig. 2 Unwinding XAI-Scope: Local instance explanations (Left) vs. global explainability (Right)

the intrinsic and post-hoc explainability algorithms. The intrinsic explainability algorithms are self-explanatory models which are inherently interpretable. Such techniques are specific to the model and the explainability algorithm differs for different ML models. The post-hoc explainability algorithms, in contrast, are model-agnostic and the algorithm remains the same across various ML Models. Figure 3 presents the two types of explainability algorithms [21].

Intrinsic Models Using models that are simple and self-explanatory is the first and most crucial step toward XAI. There are a few models that are inherently interpretable, such as logistic regression (LR) models, decision tree-based models, neural network models, generalized linear rules, and generalized additive models (GAMs). The details are presented as follows.

- *Logistic Regression*: Models based on LR provide an efficient technique for the explanation of predictions generated by the models. The plotted results of LR attribute the features which are significantly more important for a given application.
- *Decision Tree-Based Models*: While LR models do provide information that helps in the identification of important features, they do not specify any threshold values. Decision tree-based models solve this issue, by generating results which show not only the important features but also the threshold values for each of the features.
- *Neural Network Models*: ML models use neural networks, which consist of neurons, which can be visualized for interpretation. These neurons send signals to each other, which can either be positive or negative, depending upon the relationships between the features. Furthermore, the neurons which are active during

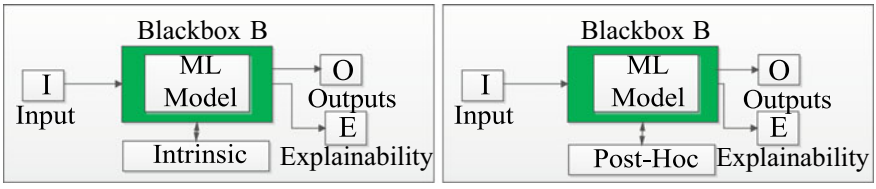

Fig. 3 XAI interpretation in terms of extensibility: The left figure shows the intrinsic XAI computation, and the right figure shows the post-hoc computation

prediction, called activated neurons, are also represented, making it easier to interpret the prediction by evaluating the relationships between the activated neurons and the nature of the signals sent to and from them.

Visualization-based, Post-hoc techniques These methods leverage visualization techniques and are independent of the underlying model. The training data that the ML model uses to train is visualized against a given data point. The difference between them is used to interpret the meaning and process behind the predictions generated by the model, which can help explain, to some extent, the features that the predicted outputs depend upon. Such methods are simple and often relatively faster, but do not provide information as to which features are more significant toward the predicted results.

Global Surrogate, Post-hoc This model, like a black-box model, is an interpretable approximation model that is trained with the same data set to forecast. The surrogate model can help us understand the behavior of the black-box model. By designing a new black-box model, we can fix the problem. These models are also blamed for offering a holistic understanding of the system, which cannot be classified as interpretation.

Local Surrogate, Post-hoc This category of models includes interpretable models that are used in black-box machine learning models to make specific predictions. Following are some examples of this technique:

- *Local interpretable model-agnostic explanations (LIME)*: ML models like LR, decision trees and neural networks are self-explanatory and provide nearly identical explanations. However, it is up to the user to interpret these explanations, which can mean that these XAI models may be interpreted differently. Ribeiro et al. [22] provided a solution to this problem by proposing LIME. LIME provides the same explanation and the same visualization for every supervised ML model. It is a model-agnostic method, which provides local explanations. It uses feature selection techniques for the identification of the main features contributing toward prediction.
- *Shapley Additive Explanations*: SHAP is proposed by Lundberg et al. [23], and the predicted output depends on all the different factors or features, and the amount each feature contributes toward the predicted results. It uses Shapley Values, a concept popular in game theory, to accomplish this. This provides an accurate interpretation for explaining the predictions of any ML model, as well as providing additional information on the important features that can help in deciding further action.

3.3 Tools and Frameworks

To explain sophisticated AI models, most XAI suppliers provide various explanation interfaces, presented as follows.

- *Google Cloud Platform*: Google Cloud's XAI platform scores each component to assess how it contributes to the ultimate result of predictions using your ML models. It can also be used to build scenario analyses by manipulating data. Some of its features include
- *AI Explanations*: Receive a score explaining how much each factor contributed to the model predictions in AutoML Tables, BigQuery ML, inside your notebook, or via Vertex AI Prediction API.
- *What-If Tool*: Using the What-If Tool coupled with Vertex AI, investigate model performance for a variety of features in your dataset, optimization methodologies, and even modifications to individual datapoint values.
- *Continuous evaluation*: Predictions from Vertex AI's trained machine learning models are sampled. Using the continuous evaluation capabilities, provide ground truth labels for prediction inputs. Model predictions are compared against ground truth labels in the data labeling service to help you enhance model performance.
- *IBM Cloud Pak*: IBM Cloud Pak's Watson OpenScale offers contrasting explanations for any classification models you build. That means it displays pertinent positives and pertinent negatives, which both help explain each model's behavior.

4 XAI in Applicative Verticals: A Solution Taxonomy

The section presents a solution taxonomy of XAI in different applicative verticals. The details are presented as follows. Figure 4 presents the solution taxonomy.

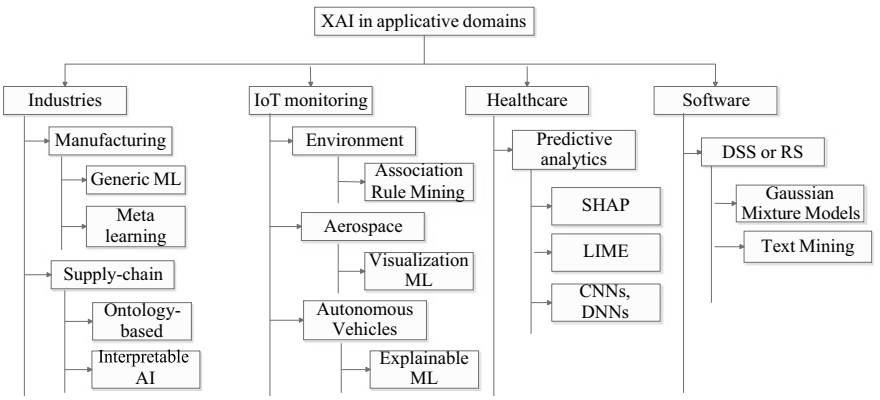

Fig. 4 Solution taxonomy of XAI in applicative verticals

4.1 Industry

With the emergence of Industry 5.0, the manufacturing sector has included automation and IoT as prime components to form cyber-physical systems, processes, and operations. Due to the "black-box" effect in manufacturing, it is less essential to understanding how AI systems generate any decisions, which is quite a challenge for business experts [24]. Industry systems can be further classified into manufacturing and supplied-chain systems.

Manufacturing In manufacturing systems, generic ML and meta-techniques are preferred. Generic ML evaluation allows the functionality to support a wide variety of ML algorithms, and decide the optimal splits for the ML model to improve accuracy. Meta-learning refers to ensemble ML where multiple algorithms and related predictions are bundled as a suite to learn the system behavior. Thus, they become equally important in manufacturing aspects, which helps in fault isolation, diagnosis, and detection [25].

Supply-chain In supply-chain ecosystems, ontology-based ML improves classification accuracy due to the extraction of semantic features from text patterns. This includes building an ontology-based dictionary for domain knowledge. Another important XAI approach is interpretable AI, which fits normally to time- series data. Some useful techniques are partial-dependency plots, individual conditional expectation, and permuted feature importance. In partial-dependency plots, we find the average prediction for function change, and it deals with average marginal effects. In individual condition expectation, it explains the causal effect on model prediction when we vary a particular feature, and thus derives the interrelation among multi-features. In Permuted feature importance, we permute the feature values and find the variance in the prediction error, which defines how features contribute to the model output in different components.

4.2 IoT Monitoring

In IoT applications, we have considered environmental systems (precision agriculture), aerospace, and autonomous vehicles. The details are presented as follows.

Environment In environmental systems, a use-case of precision agriculture is considered. In such cases, the association between different sensor nodes in the network is analyzed, where evaluation metrics like confidence, support, and expectation values are considered. As a model case, Viana et al. [26] presented a model-agnostic method to explain the most dominant factors determining better growth of crops. The resultant factors show that the framework has great potential to be used as a decision-making tool.

Aerospace In aerospace environments like flight controllers, radars, marine engineering, and others, visualization ML improves the model explainability of different components. The collected data is enormous and requires deep neural nets to understand its inherent complexity, due to non-linear transformations. Visualization ML improves the model debugging and improves the performance by tweaking hyperparameters during and after the training process. Visualization ML provides the basic information provided by air-traffic controllers, pilots, and maintainers. It helped to consciously to predict the decisions in certain situations or times.

Autonomous Vehicles Autonomous vehicles communicate with each other through information exchanged between autonomous vehicles (AVs) through ad-hoc networks. Explainable ML algorithms have been used to detect malicious acts in information transmission. Mankodiya et al. [27] proposed an XAI-enabled semantic object detection for AVs. It justifies how integrating XAI with the deep learning (DL) model helps to achieve explainability and high accuracy in detecting and segmenting the roads on which AVs are moving.

4.3 Health Care

In health care, critical patient indicators are most-importance, and thus predictive analytics is performed. Due to the AI models' extensive use, there are growing concerns about how understandable and transparent the deployed healthcare models are is a challenge. In such cases, XAI methods like SHAP and LIME become helpful to understand the relative significance of health parameters. In image-based networks, explainable CNNs and DNNs are preferred. For example, Kapcia et al. [28] presented ExMed which is a tool that provides XAI-based data analysis to domain professionals without the requirement of any precise programming knowledge. This tool has been used in two real-life applications that are COVID-19 control measures analysis and predicting lung cancer patients' age, and it proved to be accurate and flexible for the medical domain.

4.4 Software

In software development projects, we consider decision support systems (DSS), or recommender systems (RS) for our study. DSS analyzes the software's internal data and presents key reports to allow the timely solution to customer demands. RS is mainly a subclass of information filtering systems, where the filter depends on the class group, item-based, and collaborative decision. In such cases, Gaussian mixture models are preferred where the data points depend on Gaussian distributions, and

it helps to tackle unknown variables. Another approach is text mining algorithms, where product reviews can be extracted, and important recommender summaries are generated.

Figure 4 represents various techniques of XAI which can be applied to solve problems in various application like health care, software, industries, and monitoring. However, there are various limitations too of using these techniques. Some of which are, people might lack expertise in understanding the explanations, also the experts can make some disputed choices. Explanations can also change over time with change in decisions and data over time. The algorithms used for different applications cannot be explained at a general level as results can be different for different problems. The applications generally have solutions that are ambiguous, but the algorithms provide only single solution that generally does not changes with time.

5 Open Issues and Future Directions

There has been significant progress in XAI over the past few years but there are still several obstacles to overcome. Some of the major challenges faced in XAI models and implementation are mentioned as follows.

- *Lack of standardized terminologies and understanding across teams*: To scale XAI solutions, the standards and policy guidelines should reflect the demands of users, stakeholders, and impacted communities. Terminology and vagueness of meanings are two important issues in the field of XAI. Thus, unique XAI standards are required to be designed by researchers.
- *XAI techniques are computationally costly*: A key issue is a trade-off between explainability and performance. DL model's intrinsic "non-transparency" creates a huge hurdle in making them explainable for XAI purposes as they get more complicated and successful at addressing learning challenges. Hence, improvement is required in the system's performance and accuracy along with explainability in equilibrium with it.
- *Fairness of XAI techniques*: The datasets used to train black-box models might provide biased results, which can lead to unjust outcomes. Other sources of biases, in addition to datasets, include restricted features, sample size discrepancies, and proxy features. To minimize biases in XAI strategies like optimal encoding, membership, and adversarial denoising systems are proposed.
- *Post-hoc explanatory approaches*: One of the major challenges with the post-hoc explainability methodologies is their transferability. These post-hoc approaches are usually inextricably linked to the ML model and network architecture. More generalized approaches and AI/ML architectures that are intrinsically explainable using various post-hoc methodologies are required to be formulated.

- *XAI security*: Fewer studies are concentrated on building security of XAI systems. While this is an important step for the development and implementation of XAI systems, their security must not be overlooked. An XAI system must be made robust and resilient against adversary assaults to be incorporated in any applicative domain.

6 Conclusion and Future Works

The paper surveys the applicability of XAI in diverse industrial verticals of Industry 5.0. We systematically unfolded the key concepts of XAI in terms of data inputs, scope, and model explainability techniques. Further, we analyzed the importance of XAI in applicative verticals ranging from industry, IoT, health care, and software domains. We presented the open issues and challenges of XAI mechanisms in terms of cost, standardization, and fairness, which are crucial to designing valid XAI ecosystems.

As the future scope of the work, the authors intend to design a trusted XAI system for healthcare ecosystems, where the collected data is heterogeneous and have high statistical differences. Computational methods to reduce the variance in the models would be designed, which reduces the overall data and model complexity of the underlying system.

References

1. Saraswat D, Bhattacharya P, Verma A, Prasad VK, Tanwar S, Sharma G, Bokoro PN, Sharma R (2022) Explainable AI for healthcare 5.0: Opportunities and challenges. IEEE Access
2. D'Alterio P, Garibaldi JM, John RI (2020) Constrained interval type-2 fuzzy classification systems for explainable AI (XAI). In: 2020 IEEE international conference on fuzzy systems (FUZZ-IEEE), pp 1–8. https://doi.org/10.1109/FUZZ48607.2020.9177671
3. Deramgozin M, Jovanovic S, Rabah H, Ramzan N (2021) A hybrid explainable ai framework applied to global and local facial expression recognition. In: 2021 IEEE international conference on imaging systems and techniques (IST), pp 1– 5. https://doi.org/10.1109/IST50367.2021.9651357
4. Kuppa A, Le-Khac NA (2021) Adversarial XAI methods in cybersecurity. IEEE Trans Inf Forensics Secur 16:4924–4938. https://doi.org/10.1109/TIFS.2021.3117075
5. Wang Z, Zhu H, Ma Y, Basu A (2021) XAI feature detector for ultrasound feature matching. In: 2021 43rd annual international conference of the IEEE engineering in medicine biology society (EMBC), pp 2928–2931. https://doi.org/10.1109/EMBC46164.2021.9629944
6. Chen D, Zhao H, He J, Pan Q, Zhao W (2021) An causal XAI diagnostic model for breast cancer based on mammography reports. In: 2021 IEEE international conference on bioinformatics and biomedicine (BIBM), pp 3341–3349. https://doi.org/10.1109/BIBM52615.2021.9669648
7. Malhotra D, Srivastava S, Saini P, Singh AK (2021) Blockchain based audit trail ing of XAI decisions: Storing on ipfs and ethereum blockchain. In: 2021 international conference on communication systems networks (COMSNETS), pp 1–5. https://doi.org/10.1109/COMSNETS51098.2021.9352908

8. Lavrenovs A, Graf R (2021) Explainable AI for classifying devices on the internet. In: 2021 13th international conference on cyber conflict (CyCon), pp 291–308. https://doi.org/10.23919/CyCon51939.2021.9467804
9. Zolanvari M, Yang Z, Khan K, Jain R, Meskin N (2021) Trust XAI: model-agnostic explanations for AI with a case study on IIoT security. IEEE Internet Things J:1–1. https://doi.org/10.1109/JIOT.2021.3122019
10. Pawar U, O'Shea D, Rea S, O'Reilly R (2020) Explainable AI in healthcare. In: 2020 international conference on cyber situational awareness, data analytics and assessment (CyberSA), pp 1–2. IEEE
11. Adams J, Hagras H (2020) A type-2 fuzzy logic approach to explainable AI for regula tory compliance, fair customer outcomes and market stability in the global financial sector. In: 2020 IEEE international conference on fuzzy systems (FUZZ-IEEE). IEEE, pp 1–8
12. Nicodeme C (2020) Build confidence and acceptance of ai-based decision support systems—explainable and liable ai. In: 2020 13th international conference on human system interaction (HSI), pp 20–23. https://doi.org/10.1109/HSI49210.2020.9142668
13. Li XH, Cao CC, Shi Y, Bai W, Gao H, Qiu L, Wang C, Gao Y, Zhang S, Xue X, Chen L (2022) A survey of data-driven and knowledge-aware explainable AI. IEEE Trans Knowl Data Eng 34(1):29–49. https://doi.org/10.1109/TKDE.2020.2983930
14. Kuhn DR, Kacker RN, Lei Y, Simos DE (2020) Combinatorial methods for explainable ai. In: 2020 IEEE international conference on software testing, verification and validation workshops (ICSTW), pp 167–170. https://doi.org/10.1109/ICSTW50294.2020.00037
15. Ye Q, Xia J, Yang G (2021) Explainable AI for COVID-19 CT classifiers: an initial comparison study. In: 2021 IEEE 34th international symposium on computer-based medical systems (CBMS), pp 521–526. https://doi.org/10.1109/CBMS52027.2021.00103
16. Ameen ZS, Saleh Mubarak A, Altrjman C, Alturjman S, Abdulkadir RA (2021) Explainable residual network for tuberculosis classification in the IoT era. In: 2021 international conference on forthcoming networks and sustainability in AIoT era (FoNeS-AIoT), pp 9–12. https://doi.org/10.1109/FoNeS-AIoT54873.2021.00012
17. Pawar U, O'Shea D, Rea S, O'Reilly R (2020) Explainable AI in health care. In: 2020 international conference on cyber situational awareness, data analytics and assessment (CyberSA), pp 1–2. https://doi.org/10.1109/CyberSA49311.2020.9139655
18. Tantithamthavorn CK, Jiarpakdee J (2021) Explainable AI for software engineering. In: 2021 36th IEEE/ACM international conference on automated software engineering (ASE), pp 1–2. https://doi.org/10.1109/ASE51524.2021.9678580
19. Van Lent M, Fisher W, Mancuso M (2004) An explainable artificial intelligence system for small-unit tactical behavior. In: Proceedings of the national conference on artificial intelligence. AAAI Press, MIT Press, Menlo Park, Cambridge, London, pp 900–907
20. Palacio S, Lucieri A, Munir M, Hees J, Ahmed S, Dengel A (2021) XAI handbook: towards a unified framework for explainable AI. CoRR abs/2105.06677. https://arxiv.org/abs/2105.06677
21. Das, A., Rad, P.: Opportunities and challenges in explainable artificial intelligence (XAI): A survey (2020). https://doi.org/10.48550/ARXIV.2006.11371, https:// arxiv.org/abs/2006.11371
22. Ribeiro MT, Singh S, Guestrin C (2016) why should I trust you? Explaining the predictions of any classifier. CoRR abs/1602.04938. http://arxiv.org/abs/1602.04938
23. Lundberg SM, Lee S (2017) A unified approach to interpreting model predictions. CoRR abs/1705.07874. http://arxiv.org/abs/1705.07874
24. Lampathaki F, Agostinho C, Glikman Y, Sesana M (2021) Moving from 'black box'to 'glass box'artificial intelligence in manufacturing with xmanai. In: 2021 IEEE international conference on engineering, technology and innovation (ICE/ITMC). IEEE, pp 1–6
25. Terziyan V, Vitko O (2022) Explainable AI for industry 4.0: semantic representation of deep learning models. Procedia Comput Sci 200:216–226. https://doi.org/10.1016/j.procs.2022.01.220. https://www.sciencedirect.com/science/article/pii/S1877050922002290. (3rd international conference on industry 4.0 and smart manufacturing)

26. Viana CM, Santos M, Freire D, Abrantes P, Rocha J (2021) Evaluation of the factors explaining the use of agricultural land: A machine learning and model-agnostic approach. Ecological Indicators 131:108200. https://doi.org/10.1016/j.ecolind.2021.108200, https:// www.sciencedirect.com/science/article/pii/S1470160X21008657
27. Mankodiya H, Jadav D, Gupta R, Tanwar S, Hong WC, Sharma R (2022) Od-XAI: Explainable AI-based semantic object detection for autonomous vehicles. Applied Sciences 12(11). https://www.mdpi.com/2076-3417/12/11/5310
28. Kapcia M, Eshkiki H, Duell J, Fan X, Zhou S, Mora B (2021) Exmed: An AI tool for experimenting explainable AI techniques on medical data analytics. In: 2021 IEEE 33rd international conference on tools with artificial intelligence (ICTAI), pp 841–845. https://doi.org/10.1109/ICTAI52525.2021.00134