International Journal of Future Generation Communication and Networking
Vol. 15, No. 1, (2022), pp. 43-49

To Develop Robust Algorithm for Security in Black box Using Explainable Artificial Intelligence (XAI)

Shruti Ajaykumar Yelne[1], Dr. A. N. Thakare[2]

[1]P. G. Student, Department of Computer Engineering, Bapurao Deshmukh College of Engineering, Sevagram Wardha

[2]Professor, Department of Computer Engineering, Bapurao Deshmukh College of Engineering, Sevagram, Wardha

ABSTRACT

Artificial Intelligence based systems is that they often lack transparency. Indeed, the black-box nature of these systems allows powerful predictions, but it cannot be directly explained. This issue has triggered a new debate on explainable AI (XAI). AI to continue making steady progress without disruption.

Explainable Artificial Intelligence (XAI) is a branch of AI that advocates a set of tools, strategies, and algorithms for generating high-quality interpretable, intuitive, and human-understandable explanations of AI judgments. This paper gives mathematical summaries of seminal work in addition to offering a holistic assessment of the present XAI landscape in deep learning. We begin by establishing a taxonomy and categorising XAI strategies based on the scope of explanations, algorithmic methodology, and level of explanation or application, all of which aid in the development of reliable, interpretable, and self-explanatory deep learning models. The key ideas employed in XAI research are then described, and a timeline of significant XAI studies from 2007 to 2020 is shown. We evaluate the explanation maps created by XAI algorithms using image data, highlight the limitations of this methodology, and suggest potential future routes to improve XAI assessment after thoroughly discussing each category of methods and methodologies.

Keywords : Explainable XAI, Interpretable Deep Learning, Machine Learning, Computer Vision, Neural Network,Black-box Model

I. INTRODUCTION

For many years, artificial intelligence (AI) was mostly a theoretical field with few applications with real-world significance. This has fundamentally altered during the last decade, as breakthroughs in Machine Learning

(ML) have been enabled by a combination of more powerful machines, improved learning algorithms, and better access to enormous amounts of data, resulting in widespread industrial adoption [1].

Deep Learning approaches [2] began to dominate accuracy benchmarks around 2012, hitting superhuman performances and continuing to improve over time. As a result, machine learning models are now being used to solve a wide range of real-world problems in a variety of areas, ranging from retail and finance [3,

4] to medicine and healthcare behavior that had languished in the scientific community in the years prior to its resurgence, with most research focusing on the predictive p Figure 1 shows the evolution of the popularity of the search phrase "Explainable AI" over time, as measured by Google Trends.

ISSN: 2233-7857 IJFGCN

Increased model complexity, on the other hand, is frequently used to obtain higher predicted accuracy. The deep learning paradigm, which is at the heart of most cutting-edge machine learning systems, is an excellent example. It enables machines to explore, learn, and extract the hierarchical data representations required for detection and classification tasks automatically. This hierarchy of increasing complexity, combined with the fact that large volumes of data are utilised to train and construct such sophisticated systems, although boosting the systems' predictive power in most circumstances, reduces their ability to explain their inner workings and methods naturally. As a result, the reasoning behind their actions becomes more difficult to comprehend, making their projections more difficult to interpret.

There is a clear trade-off between a machine learning model's ability to deliver explainable and interpretable predictions and its performance. On the one hand, there are "black-box" models such as deep learning [2] and ensembles. The so-called white-box or glass-box models, on the other hand, yield simply explainable results—common examples are linear [11] and decision tree-based [12] models. The later models, while more explainable and interpretable, are not as powerful as the former and fail to achieve state-of-theart performance when compared to the former. Their poor performance, as well as their capacity to be simply comprehended and explained, stem from the same source: their cost-cutting design.

It is difficult to trust systems whose conclusions are difficult to explain, especially in fields like healthcare or self-driving cars, where moral and justice questions have inevitably arisen. The need for trustworthy, fair, robust, high-performing models for real-world applications prompted the resurgence of the field of explainable Artificial Intelligence (XAI) [13]—a field devoted to the understanding and interpretation of AI system

The substantial growth in recent years, reflecting the field's regeneration, is mirrored in the increasing research output during the same time period.

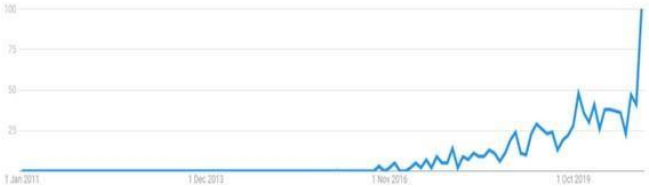

Figure 1. Google Trends Popularity Index (Max value is 100) of the term “Explainable AI” over the last ten years (2011–2020).

Attacks on XAI Methods

Recently, some research [14] has begun to investigate adversarial robustness by examining the range of classification accuracy and network interpretability. Zhang et al. describe a family of white-box attacks that generate adversarial inputs that deceive target deep learning classifiers as well as their associated interpretation models [15]. They put the proposed strategy to the test using four distinct types of explainers. An unnoticeable adversarial perturbation to fool classifiers can result in a large change in a classspecific network interpretability map, as demonstrated in [14]. The sensitivity of explanation maps to tiny disturbances in the picture domain has been illustrated in [16]. In the area of picture categorization, there has been some recent study on modifying explanations. In [17], the authors demonstrate how to change inputs in a way that is unnoticeable to humans demonstrated that post-hoc explanations are unreliable, presenting merely correlations between the underlying computations. LIME and SHAP explanations are not intuitive in the case of structured data, as demonstrated by [13]. Dylan et al. [14] developed a unique approach for efficiently hiding discriminating biases in any black-box classifier and fooling post-hoc explanation approaches such as LIME and SHAP in a recent paper.

ISSN: 2233-7857 IJFGCN

Our research focuses on two types of adversaries and black box focused attacks against them. The first tries to threaten the integrity of the underlying classifier and explainer, whereas the second tries to attack only the explainer without affecting the classifier's prediction, i.e. modify the explanation map given a natural sample.

Adadi and Berrada [18] did an exhaustive literature analysis, collecting and assessing 381 distinct scientific papers between 2004 and 2018. They arranged all of the scientific work in the field of explainable AI along four primary axes and underlined the need for more formalism to be introduced in the field of XAI and for more interaction between people and machines. After underlining the trend of the community to study explainability exclusively in terms of models, they advocated embracing explainability in other elements of machine learning. Finally, they identified a prospective research route that would go towards the composition of existing explainability methodologies.

Another survey that sought to categorise the available explain ability methodologies is this of Guidotti et al. [19]. Firstly, the authors established four categories for each approach depending on the sort of problem that they were created to answer. One category for discussing black-box models, one for inspecting them, one for explaining their results, and, ultimately, one for developing transparent black box models. Subsequently, they provided a taxonomy that takes into account the sort of underlying explanation model (explanator), the type of data utilised as input, the difficulty the technique confronts, as well as the black box model that was "opened". As with works previously discussed, the lack of formality and need for a definition of metrics for evaluating the performance of interpretability methods was highlighted once again, while the incapacity of most black-box explain ability methods to interpret models that make decisions based on unknown or latent features was also raised. Lastly, the lack of interpretability techniques in the field of recommender systems is noted and a strategy according to which models might be trained directly from explanations is presented.

Upon noting the lack of formality and means to test the success of interpretability approaches, Murdoch et al. [20] published a study in 2019, in which they built an interpretability framework in the expectation that it would help to overcome the aforementioned gap in the area. The Prediction, Descriptive, Relevant (PDR) framework presented three types of metrics for grading the interpretability approaches, predictive accuracy, descriptive accuracy, and relevancy. To conclude, they dealt with transparent models and post-hoc interpretation, as they believed that post-hoc interpretability could be used to elevate the predictive accuracy of a model and that transparent models could increase their use cases by increasing predictive accuracy—making clear, that, in some cases, the combination of the two methods is ideal.

A more recent study carried out by Arrieta et al. [21] offered a different style of organisation that initially distinguished transparent and post-hoc approaches and subsequently formed sub-categories. An alternate taxonomy exclusively for the deep learning interpretability approaches, due to their huge volume, was devised. Four categories were proposed under this taxonomy: one for explaining deep network processing, one for explaining deep network representation, one for explaining producing systems, and one for explaining hybrids of deep network processing, representation, and production systems.

ISSN: 2233-7857 IJFGCN

Procedures that are transparent and methods that are black-box finally, the authors discussed the concept of Responsible Artificial Intelligence, which is a technique that introduces a set of criteria for integrating AI in businesses.

II. PROPOSED METHODOLOGY

The main focus of the proposed system is on three key areas:

- The defense mechanisms against the attack proposed;
- Extend the proposed method to compromise the privacy and confidentiality properties of explainable methods, and
- Examine the security robustness of other XAI with different neural network architectures.

In the context of the security domain, we divide the explainability space into - (a) explanations of predictions/data itself X-PLAIN ; (b) explanations covering security and privacy properties of predictions/data XSP PLAIN ; (c) explanations covering threat model of predictions/data under consideration XT PLAIN .

1) X-PLAIN: This space covers the following type of explanations:

- Static vs. interactive changes in explanations seen by user in response to feedback. • Local vs. global explanations.
- In-model vs. post-hoc model explanations that cover models, which are transparent by their nature vs. use of an auxiliary method to explain a model after it has been trained.
- Surrogate model is a second, usually directly interpretable model that approximates a more complex model, while a visualization of a model may focus on parts of it and is not itself a fullfledged model.

2) XSP-PLAIN: The XSP-PLAIN explanations include:

Confidentiality properties of data and model e.g. which features of the data are protected by system owner. Integrity properties of data and model e.g. when and how the data was collected and model was trained to accommodate domain shifts etc. Fairness property can be part of model integrity in which explanations can help expose fairness violations by providing insights intopossible biases in a model.

Privacy properties of data and model in the explanations e.g. which part of the data/predictions is exposed to whom. For the publicly released training data and models, have noise added to them so that data rights or model privacy are not compromised? Global explainability methods need to investigate ways to provide explanations about the model without providing details on model weights (directly or via feature importance scores).

3) XT-PLAIN: This space captures the properties of threat models considered at the time of training and deployment. e. g. data poisoning protection, thresholds used, etc. Below we list some of the properties of XAI methods that are relevant to threat modelling in the security domain.

- Correctness: Correctness evaluates the ability of an ex- plainer to correctly identify components of the input that contribute most to the prediction of the classifier.
- Consistency: It is the measure of the explainer's ability to capture the relevant components under various transformation to the input. Morespecifically, if the classifier predicts the same class for both the original and transformed inputs, consistency attempts to measure whether the generated explanation for the transformed input is similar to the one generated for the original input modulo the transformation.

ISSN: 2233-7857 IJFGCN

- Transferability: Explainability is an advocate for transferability, since it may ease the task of elucidating the boundaries that might affect a model, allowing for a bet- ter understanding and implementation. Similarly, the mere understanding of the inner relations taking place within a model facilitates the ability of a user/attacker to reuse this knowledge craft an attack.
- Confidence: as a generalization of robustness and stability, confidence should always be assessed on a model in which reliability is expected. As stated in [23]–[25], stability is a must-have when drawing interpretations from a certain model. Trustworthy interpretations should not be produced by models that are not stable. Hence, an explainable model should contain information about the confidence of its working regime.
- Fairness: From a social standpoint, explainability can be considered as the capacity to reach and guarantee fairness in ML models. One of the objectives of XAI is highlighting bias in the data a model was exposed to [5], [22]. The support of algorithms and models is growing fast in fields that involve human lives, hence explainability should be considered as a bridge to avoid the unfair or unethical use of the algorithm's outputs.
- Privacy: One of the by-products enabled by explainability in ML models is its ability to assess privacy. ML models may have complex representations of their learned patterns. Not being able to understand what has been captured by the model [6] and stored in its internal representation may entail a privacy breach. Contrarily, the ability to explain the inner relations of a trained model by

Ideally, XAI should be able to explain the knowledge within a model and it should be able to reason about what the model acts upon. However, the information revealed by XAI techniques can be used both to generate more effective attacks in adversarial contexts aimed at confusing the model, at the same time as to develop techniques to better protect against private content exposure by using such information.

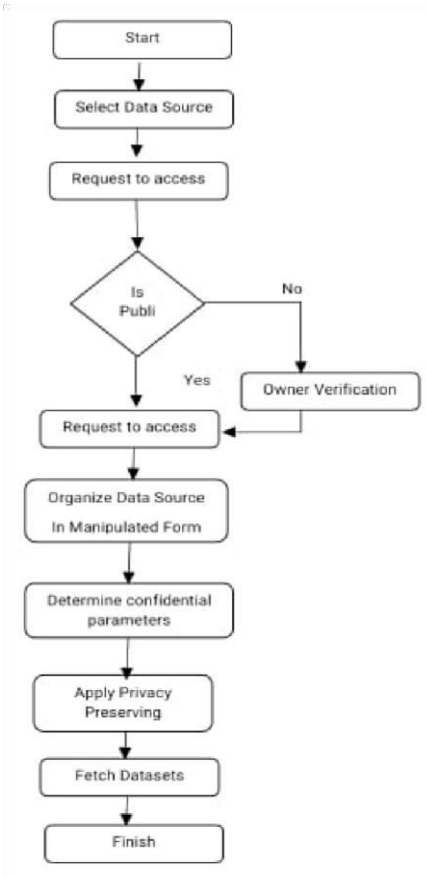

Figure 1. Flowchart for proposed system.

III. SYSTEM REQUIREMENT

- Software Requirement

Eclipse IDE , JDK 7.0 , MYSQL

- Hardware Requirement

RAM:4 GB , Processor:i3(6th Gen)

IV. CONCLUSION

Findings showed that XAI is not just a research field, its impact is spanning in a large range of application domains. However, we have seen evidence throughout this work for the lack of formalism in terms of problem formulation and clear unambiguous definitions. Further more, it has been noted that the human's role is not sufficiently studied in existing explainability approaches. In essence, attention is devoted to interpreting ML models letting other promising AI system explainability under-explored. It has then been concluded that considerable effort will be required in the future to tackle the challenges and open issues with XAI.

XAI is indeed a key area of multidisciplinary AI research. This document offers a full background on this subject in the spirit of holism. Inspired by our way of understanding new issues, we concentrated on the five components of wisdom and how to cover all the factors associated to XAI. What, Who, Who, Why, Where and How. This poll also examined a portfolio of explanations from a variety of angles for the aim of mapping the wide terrain around XAI research.

Findings have shown that XAI is not only a laboratory field but has an impact on a wide variety of application fields. In the course of this effort, however, we have found evidence that problem formulation and clear, precise definitions have not been formalised. In addition, it was found that the function of the human person in present methods to explainability is not thoroughly studied. Basically, the focus is on the interpretation of the ML models which under-explore the explainability of another promising AI system. In the future, great effort has been concluded in order to confront XAI problems and open problems.

V. REFERENCES

[1]. International Data Corporation IDC. (2018). Worldwide Semiannual Cognitive Artificial Intelligence Systems Spending Guide. Accessed: Jun. 6, 2018. [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=prUS43662418

[2]. Statista. (2018). Revenues From the Artificial Intelligence (AI) Market Worldwide From 2016 to 2025. Accessed: Jun. 6, 2018. [Online]. Available: https://www.statista.com/statistics/607716/worldwide-artificialintelligence-market-revenues/

[3]. Gartner. (2017). Top 10 Strategic Technology Trends for 2018. Accessed: Jun. 6, 2018. [Online]. Available: https://www.gartner.com/doc/3811368?srcId=1-6595640781

[4]. S. Barocas, S. Friedler, M. Hardt, J. Kroll, S. VenkaTasubramanian, and H. Wallach. The FAT-ML Workshop Series on Fairness, Accountability, and Transparency in Machine Learning. Accessed: Jun. 6, 2018. [Online]. Available: http://www.fatml.org/

[5]. B. Kim, K. R. Varshney, and A. Weller. 2018 Workshop on Human Interpretability in Machine Learning (WHI). [Online]. Available: https://sites.google.com/view/whi2018/

[6]. A. G. Wilson, B. Kim, and W. Herlands. (2016). Proceedings of NIPS 2016 Workshop on Interpretable Machine Learning for Complex Systems. [Online]. Available: https://arxiv.org/abs/1611.09139

[7]. D. W. Aha, T. Darrell, M. Pazzani, D. Reid, C. Sammut, and P. Stone, in Proc. Workshop Explainable AI (XAI) IJCAI, 2017.

[8]. M. P. Farina and C. Reed, in Proc. XCI, Explainable Comput. Intell. Workshop, 2017.

[9]. I. Guyon et al., in Proc. IJCNN Explainability Learn. Mach., 2017.

[10]. A. Chander et al., in Proc. MAKE-Explainable AI, 2018.

[11]. S. Biundo, P. Langley, D. Magazzeni, and D. Smith, in Proc. ICAPS Workshop, EXplainable AI

Planning, 2018.

[12]. M. Graaf, B. Malle, A. Dragan, and T. Ziemke, in Proc. HRI Workshop, Explainable Robot. Syst., 2018.

[13]. T. Komatsu and A. Said, in Proc. ACM Intell. Interfaces (IUI) Workshop, Explainable Smart Syst. (EXSS), 2018.

[14]. Xu K. et al. (2018), Structured adversarial attack: Towards general implementation and better

interpretability, arXiv preprint arXiv:1808.01664

[15]. Zhang X. et al. (2018), Interpretable Deep Learning under Fire, arXiv preprint arXiv:1812.00891,2018

[16]. Ghorbani A.,Abubakar A., Zou J. (2019), Interpretation of neural networks is fragile, Proceedings of the AAAI Conference on Artificial Intelligence, 2019

[17]. Dombrowski A.-K. et al. (2019), Explanations can be manipulated and geometry is to blame, arXiv preprint arXiv:1906.07983 2019

ISSN: 2233-7857 IJFGCN