TEST Engineering & Management

January-February 2020
ISSN: 0193-4120 Page No. 15970 - 15976

XAI Techniques: Review of Last 5 Years Explainable Artificial Intelligence Techniques

Vipin Bansal, Research Scholar, Chandigarh University, Gharuan

Article Info
Volume 82
Page Number: 15970 - 15976
Publication Issue:
January-February 2020

Abstract:
Due to availability of computational power and data, Machine Learning (ML) has gained lots of popularity in recent time. ML based solutions growing veryvastly, many industries already shown interest onit and using suchsolutions.ML solutions are significantly deployed in various key domains like healthcare, security, finance and has been used successfully with utmost human level of accuracy. With the availability of medical data, ML models are capable enough to identify the root cause of any ailment. Though ML models are producing high level of accuracy, still as a consumer of ML models no one is sure why the model take such decision, especially in case where ML models are based on Deep Neural Network (DNN), which is having very complex architecture.That's the reason, ML models are also known as black-box model. For various industries like healthcare, security etc.merely relying on ML model perditioncan be hazardous. One should definitely want to know why such decision has been made by the model before taking further actions. That's the reason, along with AI-ML research a new requirement is emerging in the market and hence lots of research is happening on explaining the decision made by model. Explainable AI (XAI) term is used for explaining the modeland identifying the reasons for its prediction. Lots of discussions and research is happening in converting a black-box model to white-box model (transparent model). Purpose of this paper is to explain what all concrete research work has been done in last 5 (2015-2019) years inXAI domain in terms of developing new approaches, methods, tools, frameworksusedfor explaining or interpreting ML models

Article History
Article Received: *18 May 2019*
Revised: *14 July 2019*
Accepted: *22 December 2019*
Publication: *28 February 2020*

Keywords: *Explainable AI, XAI, Interpreting AI, Black-box, Deep Learning, Fuzzy logic*

INTRODUCTION

With the increasing demand of automation, requirement of human less solutions, cost optimization etc., use of AI-MLbased solutions has been increased across different industry and domain. Due to availability of computational capabilities and requirement of developing highly trustable solutions, these AI-ML solutions become more complex and accurate. Complex ML model architecture add another level of complexity in terms of understanding the decision made by such solution. With the rising demand of ML solutions, research in XAI has also increased very significantly.

Even though many new things and research is happening in XAI, but still most of the research work are just providing merely an overview of XAI with respect to how it can impact the society, law associated with it without providing concrete approach or evidence.

Purpose of this paper to explore the research of last five years in XAI and review the different available literatures about the work done and present an overview of the techniques or framework developed and commonly used for XAI solution.

January-February 2020
ISSN: 0193-4120 Page No. 15970 - 15976

General ML queries:

This section divided intotwo parts. Part 1 highlights the key terms that has been frequently used in this paper and explained them in a question/answer format whereas part 2 explains the core approaches used in XAI.

Key terms:

S.No	Question	Answer
1.	What is machine learning?	It's a part or subset of artificial intelligence, where a system (computer program or mathematical equations) learn itself. It needs a huge data for understanding the pattern or correlation between different data elements.Based upon such data, system try tomapthisrelationandgenerate a mathematical model of it known as ML model. Based upon the dataset different algorithms used for generating such ML models. Complexity of this ML model depends upon the complexity of data-pattern along with the algorithm used.
2.	Why ML model is also known as black-box model?	We have various set of algorithms for ML model development. Selection of algorithm depends upon the data as well as the use case. Some of the ML algorithms are simple enough that are easy to explain its internal architecture and the prediction made by such ML model for example Decision Tree. Due to the availability of computational capability and requirement of higher accuracy, complex architecture preferred for ML models like *DNN*, which is having very complex architecture. Such complex architecture usually produces good accuracy but hide all the internal complexity. That's the reason such ML models called as black-box models.
3.	What is Explainable AI(XAI)?	It's a concept which try to convert such black-box model into white-box or transparent model with the help of some tools. It refers to the capability of explaining the functionalityand prediction made by such black box for better understanding and confidence onML system.

XAIAPPROACH:

Pre-hoc XAI, it's an approach where black-box model trained together with another model, which serves the purpose of explaining the black-box model. Generally, a complex architecture model like *DNN*, *Support Vector Machine(SVN)* trained together with explainable models like *Decision Tree*, *Logistic regression.*Some researcher also called its Hybrid approach. Here aim is to provide explanation with desirable accuracy. Since multiple models get optimized together, it has been observed that different model influenced each other in optimizing the objective and hence observed less accuracy, if we compare with the approach where predictive model is only developed.This approach is explained with the combination of black box model and tree-based model[1].

Post-hoc XAI, its an approach where a highly complex ML model is trained independently, considering that model accuracy is the primary factor for evaluating model performance. On top of it, techniques are used which extracts the information of already trained model for its explanation without considering the model inner structure. Since model development is independent with explanation, generally it has high level of accuracy.Visualizing the model prediction, natural language explanation generally used for explaining the black box model [2]. Framework like LIME(Local Interpretable Model-Agnostic Explanation)[3] used for such type of explanation.

REVIEW PROCESS:

Lot of research is happening in the field of AI, many hundreds of new research paper published every year. Evenin the field of XAI,observe the expansion in research. Despite that, very few papers which are actually taking care of handling the XAI problem. In Figure 1, I try to give comparative analysis between total papers written in general for AI+ML and specific to XAI, source "*Web of Science*".

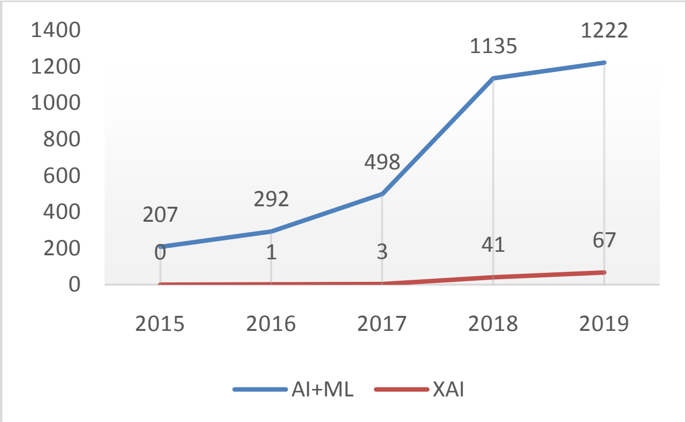

Figure 1: Comparison of papers between general AI+ML and XAI within 2015-2019

From the graph mentioned in Figure 1, it's clear that gradually researcher started showing interest towards XAI. Further review work presentenced on this paper based upon XAI only. For review process on XAI, I have used papers of two different sources, "*IEEE Xplore*" and "*Web of Science*". For more

January-February 2020
ISSN: 0193-4120 Page No. 15970 - 15976

accurate results, more specific key words and combinations are used. Common keywords used for paper search are *"Explainable Artificial Intelligence"*, *"Explainable AI"* and different combinations of it. On "*IEEE Xplore*" performed search in "*Document Title*" and "*Publication Title*". "*Web of Science*" performed search in "TI" and "TS" tags.

Search is done on the last five years of datai.e. 2015-2019. Total papers collected are 122(i.e. IEEE Xplore:15 and Web of Science 107). 10 of them are duplicate which makes effective number of papers for study are 112. After further analysis of all these papers, 30 papers found to be relevant for further studies.Table 1 present the list of these 30 papers along with publish year.

Year of publish	Papers list
2016	[4]
2017	[5]
2018	[6][7][8][9][10][11][12][13][14][15][16][17][18][19][20]
2019	[21][22][23][24][25][26][27][28][29][30][31][32][33]

Table 1: Papers listing selected for further review

Criteria used for rejecting those papers are, lots of paper providing overview of XAI and talking about its use case. Many of them are survey papers, talking about future scope of XAI. Few of them also using existing frameworks like LIME [34][35] for ML model explainability on some different data set, which is not relevant for this survey. Purpose of this review is to explain the core work done for XAI that includes new techniques, frameworks, tools developed for ML model explainability.

I have followed very basic paper compilationapproach, I have considered only those papers where some tool, approach or method is provided for ML model explanation rest of the paper considered irrelevant for this review.

REVIEW REPORT:

In this section I'll present the analysis done on the research papers. Review report includes the geographical presentation where maximum XAI research work is done in last 5 years. It also includes the research paper category and its document type. Review report also present the brief overview of the various tools, frameworks used or developed for the purpose of XAI.

GEOGRAPHICAL REPORT:

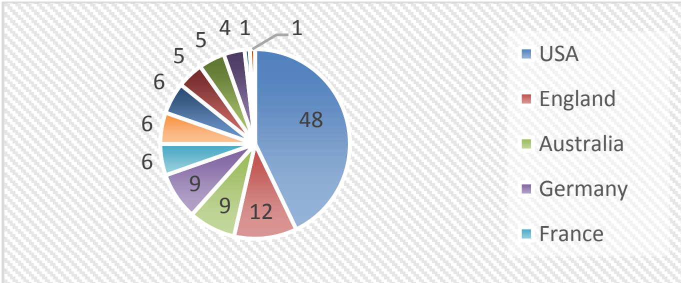

Figure 2: Papers published by each country

Figure 2, listing the papers according to the publishing country. More than 40% of the research only done in USA. Other countries like England, Australia and Germany also shown an interest in XAI research. Many agencies in USA funding a lot in this area. Agency like National Science Foundation, UNITED STATES DEPARTMENT OF DEFENSEare the two top contributors and published 10 and 6 papers respectively.

DOCUMENT TYPE REPORT:

In 2018 and 2019, lots of research has happened in XAI. Figure 1, explaining more about the quantitative value of the research happened during that tenure. Most of the research work talking about "WHY" XAI and talking about the importance of XAI. Very few papers talking about "HOW" XAI can be solved.

Figure 3 explain more about this. Out of 110 papers published in 2015-19 only 30 suggesting a concrete solution of the problem.

January-February 2020
ISSN: 0193-4120 Page No. 15970 - 15976

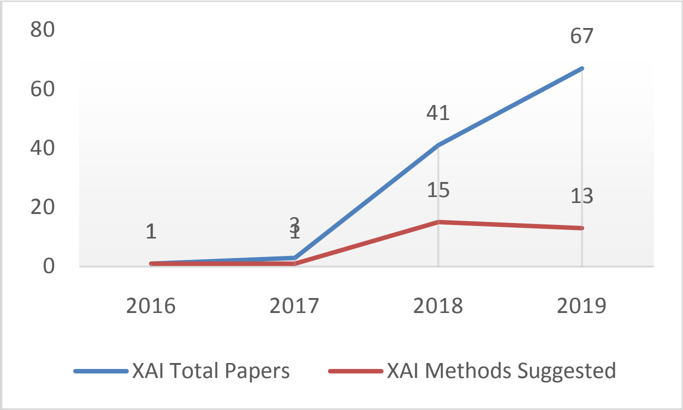

Figure 3: XAI paper comparison between Total papers and Solution suggested within 2015-2019

Further analysis of 30 papers as said above, it says research focused significantly on three different methods used for ML model explanation. *Fuzzy logic, Data Visualization and Natural Language Processing* are the three major contributing methods for XAI.

Fuzzy logic [5][7][9][19][22][31], set of predefined rules also known as Fuzzy rules are used for explaining the model predictions. Research paper [5], using *DNN* for extracting features of the data set. These features output clubbed with fuzzy rules and this clubbed system produce two outcomes, one used for prediction and another one used for explanation. Paper [7]is using approach where ML model is explained by driving the relation between attributes of data set. Some of the papers[19][31] also using approach where intermediate variable of the ML model is considered for model explanation.

Visualization based explanation commonly used for *Convolution Neural Network(CNN).* Papers [11][21][23][26][29][30] presenting ML model explanation by visualizing the model result.Visualization method commonly used for *CNN* based ML models in papers[11][21][26][29].

Another common method which is generally used is based upon *NLP*[4][18][28]. ML model trained to generate explanation in human understandable language. Such methods generally used Pre-hoc approach and trained multiple models together.

Research papers [12][15][27]using Pre-hoc approach for ML model explanation. Paper [12] using *Gradient boosting Tree model and Logistic regression* based combined model for prediction and explanation respectively. Paper [15] used *DNN* for prediction and this model fused with *Regression Models using Lasso* regularization for ML model explanation.Paper [27] human gesture recognition-based system, where gesture is used for ML model explanation. *Kinematic equations* are used for evaluation where parameters of these equations will be the data generated with different gesture position. These parameters further used for ML model explanation.

Paper [20] used the concept of minimal changes done on the sample dataset which is misclassified by the ML model. These modification in feature values presented visually that helps in identifying the most important features responsible for the misclassification and helps in generating explanation.

Paper [8] based upon the method of identifying the text similarity. Researcher used *TFIDF*(Term Frequency-Inverse Document Frequency) and *Word-Embedding* approach to identify the important and similar words for explanation.

Paper [14] used ML algorithm K-Nearest neighbor for explanation. Paper used method where training data set which already has some explanation and based upon the model outcome and neighbor data, explanation is generated.

Paper [30], developed a framework *RetainVis*, visual tool for *RNN* network. Tool present raw data which helps in explaining the prediction made by ML model.

Paper [6] used *Genetic Algorithm* based method for ML model explanation. Paper used an approach where multiple sub trees of features encapsulated for explanation. Genetic Programming used for crossover different sub treefor evaluating different combination for explanation.

CONCLUSION:

With the growing usage and dependency on AI-ML based solutions, people start thinking and raising the questions on the credibility of the predictions provided by these ML models. Explanation of these model become necessary for building a faith in AI system. Have seen lots of research is happening in model explanation domain. Most of the research in XAI are not providing clear description of the content and quantitative analysis of techniques used for XAI. Out of 122 papers, 30 papers presenting a solution for solving XAI problem. This paper presents an overview of different techniques used for ML model explainability.

REFRENCES:

[1] T. Van Gestel, B. Baesens, P. Van Dijcke, J. Suykens, and J. Garcia, "Linear and non-linear credit scoring by combining logistic regression and support vector machines," *J. Credit Risk*, vol. 1, no. 4, pp. 31–60, 2005.

[2] S. Krening, B. Harrison, K. M. Feigh, C. L. Isbell, M. Riedl, and A. Thomaz, "Learning From Explanations Using Sentiment and Advice in RL," *IEEE Trans. Cogn. Dev. Syst.*, vol. 9, no. 1, pp. 44–55, Mar. 2017.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, vol. 13-17-Augu, pp. 1135–1144.

[4] N. Wang, D. V Pynadath, and S. G. Hill, "The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams," in *AAMAS'16: PROCEEDINGS OF THE 2016 INTERNATIONAL CONFERENCE ON AUTONOMOUS AGENTS & MULTIAGENT SYSTEMS*, 2016, pp. 997–1005.

[5] D. Bonanno, K. Nock, L. Smith, P. Elmore, and F. Petry, "An approach to explainable deep learning using fuzzy inference," in *NEXT-GENERATION ANALYST V*, 2017, vol. 10207.

[6] D. Howard and M. A. Edwards, "Explainable AI: the promise of Genetic Programming Multi-run Subtree Encapsulation," in *2018 INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND DATA ENGINEERING (ICMLDE 2018)*, 2018, pp. 158–159.

[7] R. Pierrard, J.-P. Poli, and C. Hudelot, "Learning Fuzzy Relations and Properties for Explainable Artificial Intelligence," in *2018 IEEE INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS (FUZZ-IEEE)*, 2018.

[8] J. Landthaler, I. Glaser, and F. Matthes, "Towards Explainable Semantic Text Matching," in *LEGAL KNOWLEDGE AND INFORMATION SYSTEMS (JURIX 2018)*, 2018, vol. 313, pp. 200–204.

[9] R. Chimatapu, H. Hagras, A. Starkey, and G. Owusu, "Interval Type-2 Fuzzy Logic Based Stacked Autoencoder Deep Neural Network For Generating Explainable AI Models in Workforce Optimization," in *2018 IEEE INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS (FUZZ-IEEE)*, 2018.

[10] E. I. Sklar and M. Q. Azhar, "Explanation through Argumentation," in *HAI'18: PROCEEDINGS OF THE 6TH INTERNATIONAL CONFERENCE ON HUMAN-AGENT INTERACTION*, 2018, pp. 277–285.

[11] M. Bock and A. Schreiber, "Visualization of Neural Networks in Virtual Reality using UNREAL ENGINE," in *24TH ACM SYMPOSIUM ON VIRTUAL REALITY SOFTWARE AND TECHNOLOGY (VRST 2018)*, 2018.

[12] T. Ding, F. Hasan, W. K. Bickel, and S. Pan, "Interpreting Social Media-based Substance Use Prediction Models with Knowledge Distillation," in *2018 IEEE 30TH INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE (ICTAI)*, 2018, pp. 623–630.

[13] B. Murray, M. A. Islam, A. J. Pinar, T. C. Havens, D. T. Anderson, and G. Scott, "Explainable AI for Understanding Decisions and Data-Driven Optimization of the Choquet Integral," in *2018 IEEE INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS (FUZZ-IEEE)*, 2018.

[14] D. V Pynadath, N. Wang, E. Rovira, and M. J. Barnes, "Clustering Behavior to Recognize Subjective Beliefs in Human-Agent Teams," in *PROCEEDINGS OF THE 17TH INTERNATIONAL CONFERENCE ON AUTONOMOUS AGENTS AND MULTIAGENT SYSTEMS (AAMAS' 18)*, 2018, pp. 1495–1503.

[15] W. Guo, D. Mu, J. Xu, P. Su, G. Wang, and X. Xing, "LEMNA: Explaining Deep Learning based Security Applications," in *PROCEEDINGS OF THE 2018 ACM SIGSAC CONFERENCE ON*

COMPUTER AND COMMUNICATIONS SECURITY (CCS'18), 2018, pp. 364–379.

[16] A. Leroux, M. Boussard, and R. Des, "Inducing Readable Oblique Decision Trees," in *2018 IEEE 30TH INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE (ICTAI)*, 2018, pp. 401–408.

[17] H. Bharadhwaj and S. Joshi, "Explanations for Temporal Recommendations," *Kunstl. INTELLIGENZ*, vol. 32, no. 4, pp. 267–272, Nov. 2018.

[18] V. S. Silva, A. Freitas, and S. Handschuh, "Recognizing and Justifying Text Entailment through Distributional Navigation on Definition Graphs," in *THIRTY-SECOND AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE / THIRTIETH INNOVATIVE APPLICATIONS OF ARTIFICIAL INTELLIGENCE CONFERENCE / EIGHTH AAAI SYMPOSIUM ON EDUCATIONAL ADVANCES IN ARTIFICIAL INTELLIGENCE*, 2018, pp. 4913–4920.

[19] L. Magdalena, "Designing interpretable Hierarchical Fuzzy Systems," in *2018 IEEE INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS (FUZZ-IEEE)*, 2018.

[20] D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An Adversarial Approach for Explainable AI in Intrusion Detection Systems," Nov. 2018.

[21] K. R. Mopuri, U. Garg, and R. V. Babu, "CNN Fixations: An Unraveling Approach to Visualize the Discriminative Image Regions," *IEEE Trans. IMAGE Process.*, vol. 28, no. 5, pp. 2116–2125, May 2019.

[22] B. M. Keneni *et al.*, "Evolving Rule-Based Explainable Artificial Intelligence for Unmanned Aerial Vehicles," *IEEE ACCESS*, vol. 7, pp. 17001–17016, 2019.

[23] H. Strobelt, S. Gehrmann, M. Behrisch, A. Perer, H. Pfister, and A. M. Rush, "SEQ2SEQ-VIS : A Visual Debugging Tool for Sequence-to-Sequence Models," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 353–363, Jan. 2019.

[24] J. Yeo, H. Park, S. Lee, E. W. Lee, and S. Hwang, "XINA: Explainable Instance Alignment using Dominance Relationship," in *2019 IEEE 35TH INTERNATIONAL CONFERENCE ON DATA ENGINEERING (ICDE 2019)*, 2019, pp. 2135–2136.

[25] H. Kuwajima, M. Tanaka, and M. Okutomi, "Improving transparency of deep neural inference process," *Prog. Artif. Intell.*, vol. 8, no. 2, pp. 273–285, Jun. 2019.

[26] M. Han and J. Kim, "Joint Banknote Recognition and Counterfeit Detection Using Explainable Artificial Intelligence," *SENSORS*, vol. 19, no. 16, Aug. 2019.

[27] A. Banerjee, I. Lamrani, P. Paudyal, and S. K. S. Gupta, "Generation of Movement Explanations for Testing Gesture Based Co-operative Learning Applications," in *2019 IEEE INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE TESTING (AITEST)*, 2019, pp. 9–16.

[28] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, "Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions," in *PROCEEDINGS OF IUI 2019*, 2019, pp. 263–274.

[29] D. R. Chittajallu *et al.*, "XAI-CBIR: EXPLAINABLE AI SYSTEM FOR CONTENT BASED RETRIEVAL OF VIDEO FRAMES FROM MINIMALLY INVASIVE SURGERY VIDEOS," in *2019 IEEE 16TH INTERNATIONAL SYMPOSIUM ON BIOMEDICAL IMAGING (ISBI 2019)*, 2019, pp. 66–69.

[30] B. C. Kwon *et al.*, "RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 299–309, Jan. 2019.

[31] L. Magdalena, "Semantic interpretability in hierarchical fuzzy systems: Creating semantically decouplable hierarchies," *Inf. Sci. (Ny).*, vol. 496, pp. 109–123, Sep. 2019.

[32] I. Garcia-Magarino, R. Muttukrishnan, and J. Lloret, "Human-Centric AI for Trustworthy IoT Systems With Explainable Multilayer Perceptrons," *IEEE ACCESS*, vol. 7, pp. 125562–125574, 2019.

[33] S. Jha, T. Sahai, V. Raman, A. Pinto, and M. Francis, "Explaining AI Decisions Using Efficient Methods for Learning Sparse Boolean Formulae," *J. Autom. Reason.*, vol. 63, no. 4, SI, pp. 1055–1075, Dec. 2019.

[34] I. P. de Sousa, M. M. Bernardes Rebuzzi Vellasco, and E. C. da Silva, “Local Interpretable Model-Agnostic Explanations for Classification of Lymph Node Metastases,” *SENSORS*, vol. 19, no. 13, Jul. 2019.

[35] K. Weitz, T. Hassan, U. Schmid, and J.-U. Garbas, “Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods,” *TM-TECHNISCHES Mess.*, vol. 86, no. 7–8, pp. 404–412, Jul. 2019.