A Literature Review and Research Agenda on Explainable Artificial Intelligence (XAI)

Krishna Prakash Kalyanathaya [1] & Krishna Prasad K. [2]

[1] Research Scholar, College of Computer Science and Information Science, Srinivas University, Mangalore, India,
OrcidID: 0000-0002-0334-7056, E-mail: krishna.prakash.kk@gmail.com

[2] College of Computer Science and Information Science, Srinivas University, Mangalore, India, ORCID-ID: 0000-0001-5282-9038; E-mail: krishnaprasadkcci@srinivasuniversity.edu.in

Subject Area: Information Technology.
Type of the Paper: Literature Review.
Type of Review: Peer Reviewed as per |C|O|P|E| guidance.
Indexed In: OpenAIRE.
DOI: https://doi.org/10.5281/zenodo.5998488
Google Scholar Citation: IJAEML

How to Cite this Paper:
Kalyanathaya, Krishna Prakash, & Krishna Prasad, K., (2022). A Literature Review and Research Agenda on Explainable Artificial Intelligence (XAI). *International Journal of Applied Engineering and Management Letters (IJAEML), 6*(1), 43-59. DOI: https://doi.org/10.5281/zenodo.5998488

International Journal of Applied Engineering and Management Letters (IJAEML)
A Refereed International Journal of Srinivas University, India.

Crossref DOI : https://doi.org/10.47992/IJAEML.2581.7000.0119

A Literature Review and Research Agenda on Explainable Artificial Intelligence (XAI)

Krishna Prakash Kalyanathaya [1] & Krishna Prasad K. [2]

[1] Research Scholar, College of Computer Science and Information Science, Srinivas University, Mangalore, India,
OrcidID: 0000-0002-0334-7056, E-mail: krishna.prakash.kk@gmail.com

[2] College of Computer Science and Information Science, Srinivas University, Mangalore, India, ORCID-ID: 0000-0001-5282-9038; E-mail: krishnaprasadkcci@srinivasuniversity.edu.in

ABSTRACT

Purpose: *When Artificial Intelligence is penetrating every walk of our affairs and business, we face enormous challenges and opportunities to adopt this revolution. Machine learning models are used to make the important decisions in critical areas such as medical diagnosis, financial transactions. We need to know how they make decisions to trust the systems powered by these models. However, there are challenges in this area of explaining predictions or decisions made by machine learning model. Ensembles like Random Forest, Deep learning algorithms make the matter worst in terms of explaining the outcomes of decision even though these models produce more accurate results. We cannot accept the black box nature of AI models as we encounter the consequences of those decisions. In this paper, we would like to open this Pandora box and review the current challenges and opportunities to explain the decisions or outcome of AI model. There has been lot of debate on this topic with headlines as Explainable Artificial Intelligence (XAI), Interpreting ML models, Explainable ML models etc. This paper does the literature review of latest findings and surveys published in various reputed journals and publications. Towards the end, we try to bring some open research agenda in these findings and future directions.*

Methodology: *The literature survey on the chosen topic has been exhaustively covered to include fundamental concepts of the research topic. Journals from multiple secondary data sources such as books and research papers published in various reputable publications which are relevant for the work were chosen in the methodology.*

Findings/Result: *While there are no single approaches currently solve the explainable ML model challenges, some model algorithms such as Decision Trees, KNN algorithm provides built in interpretations. However there is no common approach and they cannot be used in all the problems. Developing model specific interpretations will be complex and difficult for the user to make them adopt. Model specific explanations may lead to multiple explanations on same predictions which will lead to ambiguity of the outcome. In this paper, we have conceptualized a common approach to build explainable models that may fulfill current challenges of XAI.*

Originality: *After the literature review, the knowledge gathered in the form of findings were used to model a theoretical framework for the research topic. Then concerted effort was made to develop a conceptual model to support the future research work.*

Paper Type: *Literature Review.*

Keywords: Artificial Intelligence, Machine learning, XAI, ML, Deep learning

1. INTRODUCTION :

Humans are naturally capable of thinking and performing certain tasks on their own due to brain. Human brain has the natural ability to perform some analytical tasks such as object recognition much faster than any computer. This has inspired researchers and scientist to build machines that can do similar tasks designed by human brain. Humans and animals have the ability to naturally learn, remember, make decisions and perform many complex tasks and this cognitive capability is called

Natural Intelligence. While humans and animals used the fuzzy (approximate reasoning) logic to learn and make decisions, computer machines are developed to do the same tasks using crisp (binary) logic. This process of developing intelligence using computers or similar machines is called Artificial Intelligence (AI). The ultimate goal of AI is to build a machine that can think and act like a human and automate complex tasks which can be performed efficiently. There are several branches of AI that makes a system comparable to Human intelligence and builds AI powered systems to digitize common mundane tasks and eliminate repetitive tasks. The following figure depicts the branches of AI comparable to Human body.

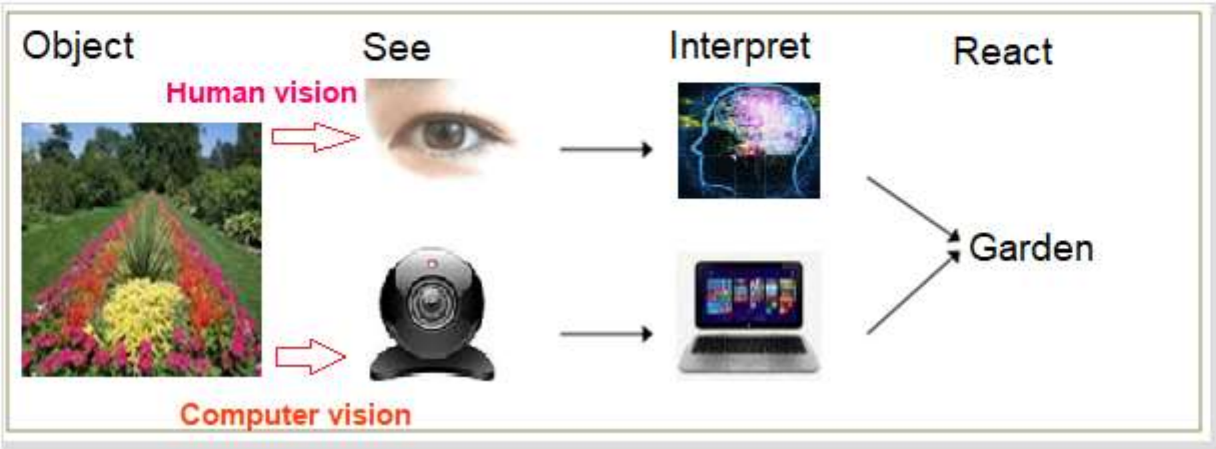

Fig. 1: Conceptual model of Artificial Intelligence as compared to Human Intelligence

The Figure 1 shown above compares the Human model to computer model of Intelligence taking an example of a real world scenario. A snapshot of real world object (picture of a garden) captured from an eye (Human model) compared to Web camera (Computer model). The picture or image is then transmitted to Brain (Human model) and Computer (computer model) for processing. After processing, Human and Computer model are reacting to the commands to provide the identification as Garden.

Table 1: Branches of AI

Human Body	Branch of AI that meets the actions of Human body	How?
Brain	Machine Learning and Deep learning	Ability of computers to “think”
Eyes	Computer Vision	Ability of computers to “see”
Ears and Mouth	Natural Language Processing and Speech recognition	Ability of the computer to “talk and listen”
Heart	Expert Systems	Ability of computers to maintain “control systems”
Arms and Legs	Robotics	Ability of computers to “walk”

The Table 1 above depicts the Human versus Artificial intelligence to explain the components of AI powered systems in simple terms. The machine learning and deep learning is a branch of AI that deals with various mathematical models of learning, Statistical techniques of learning, Biological models of learning such as Neural network, genetic programming techniques. The computer vision is a branch of AI that deals with various sensor devices for capturing images, pattern matching and algorithms that are used in recognition of objects. Natural Language Processing is a branch of AI that deals with various forms speech and text processing and translations. Expert system deals with automating control systems. Robotics is a branch of AI dealing with movement of computer controlled systems [1][2][3].

Types of learning in Artificial Intelligence

There are many approaches used by AI powered systems to learn and make decisions. AI powered systems started with rule based learning evolved into new approaches such as supervised learning, unsupervised learning, reinforcement learning, active learning, Inductive learning, Transfer Learning, Ensemble learning and Deep learning. All these learning approaches generally termed as machine learning. The process of developing a machine learning model involves data acquisition, data cleaning, model training, and model evaluations. Machine learning algorithms use historical data to identify patterns and statistical reasoning to make decisions. The process of identifying patterns and statistical reasoning is generally called as machine learning model training.

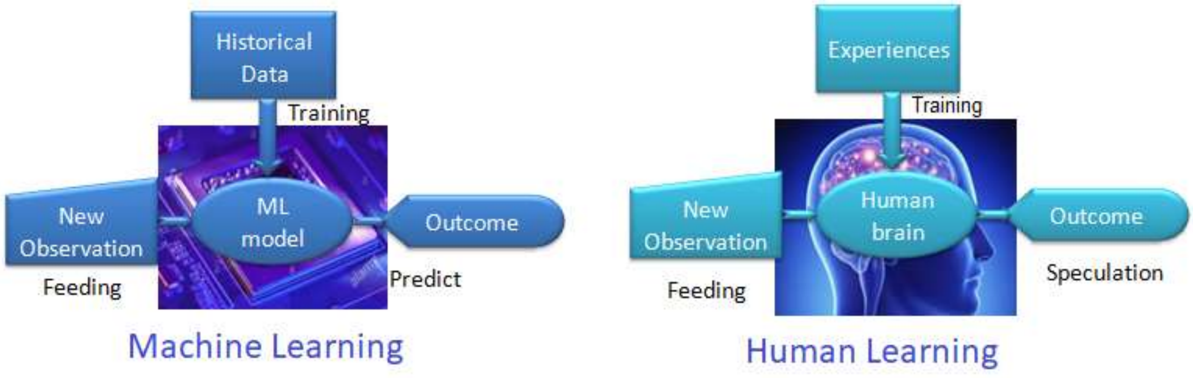

Fig. 2: Machine learning versus Human Learning

The Figure 2 shows the comparison of machine learning tasks with human learning. Human learning uses past experience (also called as Knowledge) gained either through training, observation or practically performing the tasks and stores in the brain. When new observation arrives, Humans applies this knowledge to speculate or respond to the tasks. In machine learning, historical data is used as training and stores the knowledge built in the ML model. When new observation arrives, the machine applies this knowledge (stored in ML model) to estimate or predict the outcome or respond to the task.

There are various algorithms are developed to identify patterns and statistical reasoning. Table 2 shown below provides an overview of most frequently used machine learning algorithms and their approaches.

Table 2: Most frequently used Machine learning algorithms

Sl. No.	Machine Learning Algorithm	Approach used by the algorithm	Common Applications
1	Linear regression [4]	Uses least squares approach to model the correlation between two variables by fitting a linear equation on the observed data.	Regression and Time series analysis
2	Logistic regression [5]	Uses a probability model to output a value between 0 and 1 by fitting regression equation on observed data and labeled data.	Classification
3	Decision Tree [6]	Uses a data structure to build a decision tree and map the class (or category) that the observed data belongs to.	Classification
4	Support Vector Machines [7]	Construct a hyperplane that separates the set of observed data and provides a means to categorize the data.	Classification and regression

5	Naïve Bayes [8]	Developed using Bayes' theorem with the "naive" assumption. It calculates the posterior probability of a new class based on the prior probability of the observed traits. It is assumed all features are independent of each other in the observation.	Classification
6	K-Nearest Neighbor [9]	Designed using the distance calculation between the points(Ex: Euclidean, Manhattan) and finds closest neighbor data points to estimate the value or the label	Classification and regression
7	K-Means [10]	This unsupervised algorithm clusters data by separating samples in to groups. The clustering is based on calculation of a centroid value and distance from centroid to each sample.	Clustering and segmentation
8	DBSCAN[11]	Unsupervised algorithm works on the principle of grouping together the points that are closer (forming high density regions), points lie apart are marked as outliers (forming low density regions).	Clustering and segmentation
9	Hierarchical Clustering [12]	An unsupervised algorithm that builds hierarchy of clusters by merging similar observations and splitting dissimilar observations.	Clustering and segmentation
10	Random Forest [13]	An ensemble learning method built with multitude of decision trees to fit the observed data and estimate the value or label	Classification and regression
11	Adaboost [14]	An adaptive boosting algorithm that converts weak learners in machine learning to strong one by readjusting the weights.	Classification and regression
12	Stochastic Gradient Descent [15],[16]	Uses a gradient descent optimization algorithm by iteratively replacing the actual gradient with estimated gradient value.	Classification and regression
13	XG Boost [17]	XGB is an ensemble method with gradient boosting for optimizing the loss functions.	Classification and regression
14	Hidden Markov model [18]	This algorithm is a statistical model based on Markov chain, discovers the sequence of states from the observed data.	Forecasting, speech recognition
15	Multilayer Perceptrons (MLPs) [19]	A simplest form of feed forward deep neural network built with input layer, one or more hidden layer and output layer.	Image processing
16	Convolutional Neural Networks (CNNs) [20],[21]	A deep learning algorithm, extracts features by using moving windows (convolutions) and selecting important features of in each window.	Image processing

17	Long Short Term Memory Networks (LSTMs) [22]	An improved variant of recurrent neural networks (RNN) which works on a sequence data with memory. The memory keeps the computations performed in the previous stage that influence next element in the sequence.	Time series, forecasting
18	Generative Adversarial Networks (GANs) [23]	Built on a pair of neural networks complementing each other. Pair of networks called as Generative network that generates the candidate outcomes while the other Discriminative network evaluates them.	Image processing
19	Radial Basis Function Networks (RBFNs) [24]	RBFNs are the type of neural networks in which Radial Basis Function used as activation function and is specialized for non-linear classification tasks.	Classification and regression
20	Self-Organizing Maps (SOMs) [25]	SOMs are Kohonen maps that are used to perform clusters of observations.	Classification and regression
21	Autoencoders [26]	Autoencoder uses encoding of input to compress the inputs to process and decoding the output to reconstruct the outputs.	
22	Linear discriminant analysis (LDA) [27]	This algorithm uses a generalization of Fisher's linear discriminant analysis in statistics similar to analysis of variance (ANOVA)	Dimensionality reduction
23	Latent Dirichlet Allocation (LDA) [28]	This unsupervised machine learning algorithm builds a generative statistical model that explains hidden (latent) distributions (Dirichlet) of the data and groups (allocation).	Topic modeling, NLP
24	Monte Carlo Methods [29]	A reinforcement Algorithm that uses Monte Carlo analysis algorithm to mimic what if analysis using probability distributions.	Autonomous AI agents
25	Q learning [30],[31]	A reinforcement algorithm that works on principle of stochastic process that finds an optimal policy to maximize the estimated value of the total reward over any and all the successive steps, starting from the current state.	Autonomous AI agents
26	Transfer learning [32],[33]	A method of storing knowledge gained while solving a problem and applying them separately but to the relevant problem.	Image Processing

2. OBJECTIVE :

(1) Review the current challenges and opportunities to explain the decisions or outcome of AI model
(2) Examine present status of various implementations of Explainable AI techniques
(3) Identify research gap between the present status and ideal status

(4) Conceptualize a framework to implement XAI that may fulfill the current challenges

3. EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) :

What is explainable AI?

The explosion in data, recent advancement in computing power and communication technology has led to a new generation of AI applications offers a great promise of changing our business model and real life. But the outcomes produced by these applications unable to explain their decisions and actions to human users. In 2016, DARPA launched an initiative on the project Explainable Artificial Intelligence (XAI) to solve this problem so called "Black box models" making decisions. The "Black box models" are the machine learning models that produce an outcome based on patterns found in the historical data given as input to the model. Though these outcomes of "Black box models" are nearly reliable, they were difficult to trust and validate [34][35].

Explainable AI is a technique of explaining how a Machine learning model performs the actions and makes predictions. It is aimed at explaining the rationale of decision making process (compare this with a judicial system pronounces a judgment after thoroughly evaluating all the evidence submitted to the system). However, In the case of machine learning models, the evidences submitted to the model is evaluated by complex algorithm pronouncing the outcome as a prediction. In this case the ML model does not provide all the corroboration of evaluated evidences and hence complex algorithm is a black-box (or opaque model) [36][37].

Why explainable AI is important?

We identify five reasons why explainable AI is important [38].

1. Accountability: When a model built by AI algorithm makes a decision, knowing the factors that caused the decision and validity of the decision making process that is reasonably acceptable. The consequence of such decision should not make the system unreliable. A person should be able to assume the responsibility with full visibility of the process used to make such decisions.
2. Reliance: Trust or Reliability is critical in domains like healthcare or finance. So, all stakeholders must fully understand how AI tool is making decisions. The level of transparency of decision making process may impact the trust factor. The decisions made by the AI tool should be supported by adequate evidences gathered from the decision making process.
3. Compliance: The outcome and decision-making process of the AI model must ensure compliance with company policies, industry standards and government regulations. According to Article 14 of the European Data Protection Law (GDPR), when a company uses automated decision making tools, it must explain the rationale, importance and consequences of such processing for the data subject.
4. Performance: A lot of effort in building AI model is spent on tuning the performance and process of model development without much transparency in the AI tool. The transparency of the AI tool can also help to improve the performance. For example, a neural network can be optimized if an activation function can give a better result.
5. Control: Monitoring and Control are important aspect of making the AI tool operational. Understanding the decision-making process of the AI model is needed to identify vulnerabilities and flaws in AI model and fix them in the operational system.

4. LITERATURE REVIEW :

In this section we will provide the gist of most recent research papers published on XAI topics. AI powered systems are gaining popularity in various domains such as Healthcare, Automotive, Finance, Manufacturing, Retail, Education was discussed by K. P. Kalyanathaya et al. (2019) [5]. As the AI powered systems are gaining popularity, the importance of explainable machine learning model also increasing as the AI model solve critical problem areas. For example, during COVID-19, several AI based approaches were developed to detect the COVID signatures. An explainable DL approach for COVID-19 classification (detection) via computed tomography (CT) scans was proposed by Angelov, P., & Soares, E. (2020) [39]. In this paper, author proposed model specific explainable Deep Neural Network Architecture (xDNN) with 5 components to provide interpretability. The five components are named as Feature descriptor layer, Density layer, Criticality layer, Prototypes layer, Mega Cloud layer. In Publio et al. (2018) [40], authors present ontology based approach and developed ML schema model for representing the model characteristics, important parameters, characteristics of the data and other information required for explanations. In Burkart, N., & Huber, M. F. [41], authors presents different approaches to generate explanations from the ML model viz, explanation generators, interpretable models, surrogate model learning, data independent and data dependent approaches. The author also proposes the use of ontology to generate better explanations. For example, ontology can be used to generate uniform structure to represent knowledge created by the model and their relations with domain. This will help to develop a standard p5ocess for exchanging explanations from different models to other systems. Chari. S. et al. (2020) [42] discussed ontology based model of explanation for user centered AI. The authors presented the design of the ontology around the central explanation class (ep: Explanation) and included entities and attributes required for the explanations of model. The authors designed the ontology model with 13 classes capturing attributes of explanations of the model.

5. SUMMARY OF RELATED WORK :

Table 3: Summary of most recent research papers and findings on Explainability of Machine Learning

Sl. No.	Author	Year	Findings/Inventions/Results
1	Gunning, D. et al. [34]	2019	This paper discusses about basic understanding of interpretability and explainability. It also discusses the users expectation of explainability from a machine learning model.
2	Burkart, N., & Huber, M. F. [41]	2020	This survey paper provides the necessary definitions, an overview of the various principles and methodologies of explainable Supervised Machine Learning (SML). The paper presents a linear approximation method to a complex model as a sort of representative-analysis to illustrate the features for the predicted class based on the most important features.
3	Chari. S. et al. [42]	2020	This paper focus on explainability of knowledge-enabled systems, spanning the expert systems, cognitive assistants, semantic applications, and other machine learning domains. The authors propose semantic web technology for semantic representation of explainability of knowledge enabled hybrid AI systems.
4	Umang Bhatt et al. [43]	2020	The study examines how various stakeholders visualize and understands explainability for consumption. Discusses its important as a concept in legal and ethical guidelines for data and ML. It also references to Articles 13-15 of the European General Data Protection Regulation (GDPR) for expected consequences of processing the data subject. The study finally recommends a framework for establishing a base in explainability and then include concerns associated with AI models.
5	Arrieta, A. B. et.al. [44]	2020	The paper discusses the various taxonomy used in Explainable AI (XAI) and presents a rising trend of various research articles in Explainable AI recently. This shows an upsurge of research activity also triggered by research agendas of national government and regulatory bodies on Explainable AI. This paper thoroughly analyzes the literature on XAI covering approximately 400 contributions. The paper summarizes the various uses of explainability in ML models with two focused objectives: need for model understanding, and regulatory compliance.
6	Rudin. C.,et al [45]	2021	The paper discusses 5 fundamental principles of interpretability of machine learning and provides insights some of important challenges in this area. The authors argues that the works over the last few years

			have contributed to the new terminology, which replaced the older field of interpretable machine learning into the new field of "XAI,".
7	Rai, A [46]	2019	The paper approaches the problem defining the ML model as "black box" model and explains various approaches to convert to a "glass-box" model. Two approaches has been discussed in this paper are Model specific approaches and Model agnostic approach. The model specific approach enforces interpretability constraints into the structure and learning mechanism. Model agnostic approach develops an alternative interpretable surrogate model to understand specific features of a "black-box" prediction model.
8	Xu, F., et al. [47]	2019	This paper summarizes the process of explainability in basic machine learning and deep learning. Further goes to explain the reasoning of deep learning models are black-box.
9	Liao, Q. & Kush, Ramazon. [48]	2021	This paper presents Human Centric view of Explainability (XAI).It discusses the roles that HCI play in developing the explainable AI models by helping navigate, assess and expand the XAI toolbox to provide conceptual frameworks for human-compatible XAI.
10	Barberan, C. J. et al. [49]	2021	This paper present new family of Deep Neural Network Architecture called "Neuro View" that are interpretable and explainable by design. In this architecture, each member of the family derives from standard deep neural network architecture that calculates unit output values by vectors and feeds them into a global linear classifier. The resulting structure establishes a direct, causal relationship between the position of each entity and the classification decision.
11	Adadi, A., & Berrada, M. [50]	2018	This paper presents a survey of existing approaches on XAI and explains the need for AI as four quadrants: Explain to Justify, Explain to Control, Explain to Improve, Explain to Discover. The paper also discusses technical challenges in XAI and summaries current approaches: Intrinsic, Post-hoc, Model-specific and Model-agnostic.
12	Pawar, U. et al.[51]	2020	In this paper, XAI is discussed as a proposed approach aimed at achieving the technology and accountability used by AI-based systems in the analysis and diagnosis of health data. Transparency in health care, outcome tracing and model reform.
13	Pfahler, L., & Morik, K. [52]	2021	This paper proposes a method that explains the decisions of a deep neural network architecture by analyzing the intermediate representations of each layer in the deep network that were refined during the training process. The author presented two types of visual explanations: One based on most-influential individual training instance, the other based on aggregated statistics over all training steps.
14	Seeliger, A., et al. [53]	2019	This paper provides an provide a literature survey of various usages of Semantic Web Technologies with ML methods to design explainability. In this paper, authors conclude that explainability is highly dependent on the

			usage of domain knowledge. So Semantic Web Technologies might be a critical knowledge to achieve truly explainable AI-systems.
15	Linardatos, P., et al. [54]	2021	In this paper, the study explores various methods of machine learning interpretability with a literature review and methods and their programming implementations.
16	Carvalho, D.et al. [55]	2019	This paper provides a review of the current state of the research field on the interpretation of machine learning, focusing on social impact and developed methods and metrics.
17	Angelov, P., & Soares, E. [58]	2020	The study proposes xDNN architecture that builds reasoning and learning approach to address explianability challenges in deep neural networks.
18	Hussain, F., et al. [59]	2020	This paper presents mathematical approach to explain ability of deep learning model with use case of a autonomous car.
19	Gerlings, J., et al. [60]	2020	This study proposed a explainable model that can interact with human stakeholders.
20	van der Velden, B. H., et al. [61]	2021	In this paper, a framework of XAI approach is introduced to train and classify deep learning-based medical image analysis methods.

6. RESEARCH GAP :

There are two important goals that needs to be addressed to solve the current challenges in XAI. First the model should be able to produce the trace of the steps it performed to map the inputs to predictions. This is the technical part frequently referred to interpretability. This will include some of metrics such as weights, bias and any other parameters that are used by the model to achieve the result. This information can be used to understand the dataset and algorithm (model) better for further optimization. The second goal is to produce the domain specification information such as features, factors that influence the predicted outcomes. The goal is to explain the decisions with adequate evidences supporting the decisions.

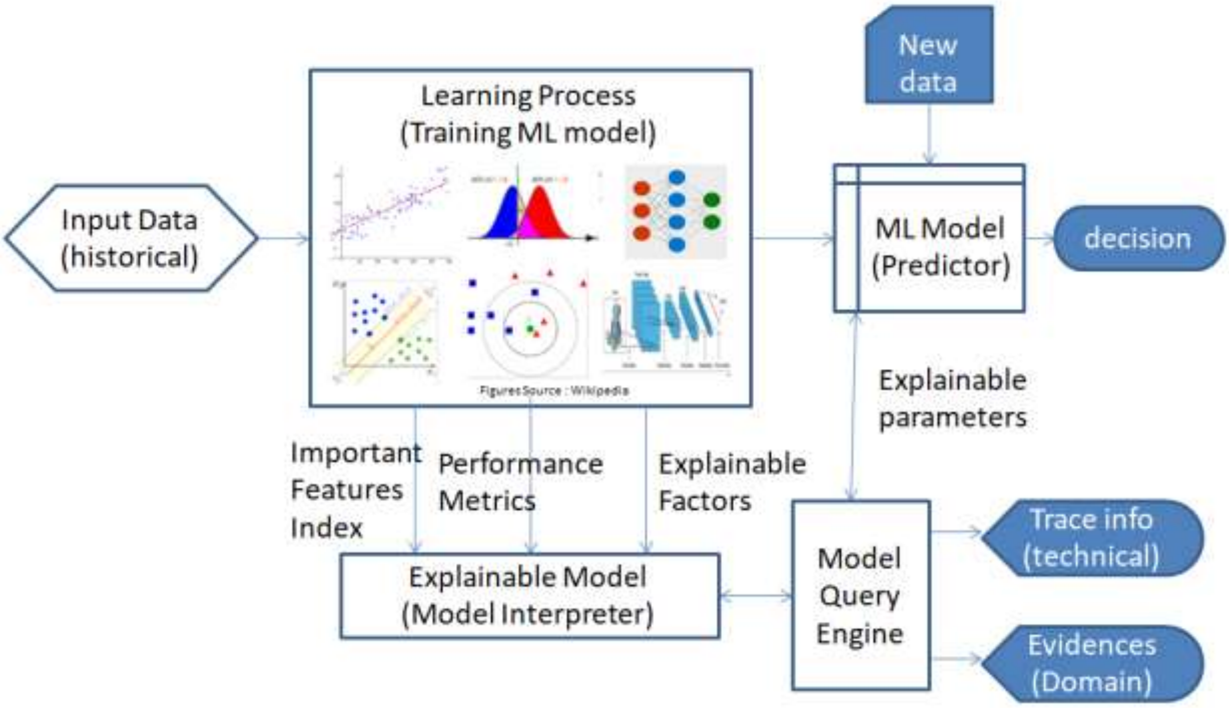

Fig. 3: Conceptualized Explainable AI model

The Figure 3 shown above explains a conceptualized model for an ideal solution to Explainable AI tool. In this tool, we create a model interpreter through the learning process after capturing crucial information and named it as Explainable model. Then we build a common model query engine that can translate a user query (technical or domain) to appropriate commands and retrieve the information from model interpreter and model predictor to explain the model outcomes. The queries can be broadly categorized into two types. First, query about algorithm and model metrics or parameters used to get the output results. The objective of these queries is to explain the model and devise a strategy to optimize the model. These queries are technical in nature and can be used to improve the model performance and a hence improved solution. These information's are expected to be useful to data analysts/data scientists. The second type of query is about the model output and corroborates the evidences gathered by the predictor from the new data to produce the result. This will explain logical relationship between the observations (new data) and outcomes. The objective of this query is to explain the outcome of the ML model to business user. The information gathered here purely domain related to domain of the solution such as important features, mapping between the features to output results. For example, a doctor using a ML model for cancer diagnosis may want to know the reasoning for the prediction of the model.

7. RESEARCH AGENDA :

There are several challenges that make decision involving machine learning models. Some samples of the challenges are as follows:

- In self-driving car, how can we explain the abnormal acts of the machine learning model in real traffic scenario?
- In medical diagnosis, how can we trust the machine learning model to treat the patients as instructed by a black-box model?
- In criminal investigation, How can we justify a decisions predicted by machine learning model and actions followed by the machine?
- The financial industry is obligated by regulation and financial market entities are mandated to make fair decisions and explain their models to provide justifications for their actions such as rejection of credit and fraud detection.
- Can Machine Learning provide legal evidences for the predicted decisions from the Machine learning model?
- How to explain the decisions predicted machine learning model in real world terms to various stakeholders?
- Can machine learning models are subjected to audit of predictions made by them?

The conceptual model shown above Figure 3 can address many of the challenged discussed in this paper. There are several approaches to build explainable models based on the above conceptual model. Some of the identified and reviewed approaches are given below: [62]

1. Interpretability is built-in the model algorithm so that results are explainable (Ex: Decision trees)
2. Interpretability is built outside the model algorithm so that results are explainable (No examples).
3. Generate data driven interpretations with generic model to explain the results (No examples)
4. Developing rule based systems to generate explanations from the model
5. Develop model specific individual interpretable units to generate the explanations.

7.1 Analysis of Research Agenda:

While there are no single approaches currently solve the explainable ML model challenges, some model algorithms such as Decision Trees, KNN algorithm provides built in interpretations. However there is no common approach and they cannot be used in all the problems. Developing model specific interpretations will be complex and difficult for the user to make them adopt. Model specific explanations may lead to multiple explanations on same predictions which will lead to ambiguity of the outcome.

ABCD Analysis of Chosen Research proposal

XAI techniques can be critical to some of domains such as Medical, Finance, Legal. Some domains such as supply chain, content filtering etc XAI technique is not critical. However there are unique advantages if AI model are explainable. The detailed analysis of XAI approaches can be seen in the ABCD analysis [63][64][65] as follows:

Advantages

- Provides the explanations to the outcome. Hence decision making becomes transparent.
- We can trust the AI models and use them with confidence of taking appropriate actions.
- Improves reliability of AI system and business system can expand the scope of the AI models

Benefits

- Transparent ML models can provide more details about the performance of the model.
- This can help to improve the performance of the model.
- This can help automating the decision making process.

Constraints

- Thousands of parameters used in the algorithm may pose a problem in developing XAI system
- Different algorithms use different approaches and hence there is no standard process to apply XAI
- The varied quality of data in each decision making process may impact the outcome

Disadvantages

- ML models becomes more complex due to additional code inserted to make the model transparent
- Performance may be affected due to complexity of the algorithm
- Poor quality of data may have adverse effect on the outcome

8. RESEARCH PROPOSAL :

The discussions and analysis of recent developments in on the Explainable AI models reveal that there is no common approach to solve these challenges. Also it is not feasible to provide model specific explanations to XAI challenges. So the scope of the research is to investigate the items 2, 3, and 4 above and identify a common approach to explainable ML models to generate the explanations. This research proposal is to investigate the following agenda:

1. Can Interpretability is built outside the model algorithm so that results are explainable?
2. Is it possible to generate data driven interpretations with generic model to explain the results?
3. Can we develop a rule based systems to generate explanations from the model?
4. Can we build a explainable AI model using the combination of the above 3 scopes?

9. CONCLUSION :

Explainable ML models are critical for the future of AI application in all walks of life. In this paper we discussed and brought some recent development and identified challenges in this area. We developed conceptualized model for an ideal solution to Explainable AI tool. In this tool, we create a model interpreter through the learning process after capturing crucial information and named it as Explainable model. Then we build a common model query engine that can translate a user query (technical or domain) to appropriate commands and retrieve the information from model interpreter and model predictor to explain the model outcomes. We have identified approaches to build explainable models based on the above conceptual model. A model specific explanation may bring multiple explanations to the same prediction and user may need to put additional effort to understand the model specific explanations. Hence we discussed the combination of several other approaches may be investigated and applied to the explainable ML model challenges.

REFERENCES :

[1] Korteling, J., van de Boer-Visschedijk, G. C., Blankendaal, R., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human- versus Artificial Intelligence. *Frontiers in artificial intelligence*, *4*(1), 622363-622364.
Google Scholar↗ Crossref↗

[2] Chatterjee, Rupen. (2020). Fundamental concepts of artificial intelligence and its applications. *Journal of Mathematical Problems, Equations and Statistics, 1(2),* 13-24.
Google Scholar↗

[3] Kalyanathaya, K. P., Akila, D., & Rajesh, P. (2019). Advances in natural language processing–a survey of current research trends, development tools and industry applications. *International Journal of Recent Technology and Engineering*, *7*(1), 199-202.
Google Scholar↗

[4] Freedman, D. A. (2009). Statistical models: theory and practice. *Cambridge university press. 2*(1), 1-133.
Google Scholar↗ CrossRef↗

[5] Tolles, J., & Meurer, W. J. (2016). Logistic regression: relating patient characteristics to outcomes. *Jama*, *316*(5), 533-534.
Google Scholar↗ Crossref↗

[6] Kamiński, B., Jakubczyk, M., & Szufel, P. (2018). A framework for sensitivity analysis of decision trees. *Central European journal of operations research, 26*(1), 135-159.
Google Scholar↗

[7] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.
Google Scholar↗

[8] Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, *3*(22), 41-46.
Google Scholar↗

[9] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175-185.
Google Scholar↗ Crossref↗

[10] Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. *Computer Science & Engineering (CS&E) Technical Reports*, *1*(1), 1-20.
Google Scholar↗

[11] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. *In kdd*, *96*(34), 226-231.
Google Scholar↗

[12] Nielsen, F. (2016). Hierarchical clustering. In Introduction to HPC with MPI for Data Science, *Springer*, *1*(1), 195-211.
Google Scholar↗ Crossref↗

[13] Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, *1*(1), 278-282. IEEE.
Google Scholar↗ Crossref↗

[14] Rojas, R. (2009). AdaBoost and the super bowl of classifiers a tutorial introduction to adaptive boosting. *Freie University, Berlin, Tech. Rep. 1*(1), 1-6.
Google Scholar↗

[15] Polyak, B. T., & Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, *30*(4), 838-855.
Google Scholar↗ Crossref↗

[16] Tsuruoka, Y., Tsujii, J. I., & Ananiadou, S. (2009). Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, *1*(1), 477-485.
Google Scholar↗ Crossref↗

[17] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 1*(1), 785-794.
Google Scholar↗

[18] Satish, L., & Gururaj, B. I. (1993). Use of hidden Markov models for partial discharge pattern classification. *IEEE transactions on electrical insulation, 28*(2), 172-182.
Google Scholar↗ Crossref↗

[19] Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms. *Cornell Aeronautical Lab Inc Buffalo NY*. 1-621.
Google Scholar↗

[20] Venkatesan, R., & Li, B. (2017). Convolutional neural networks in visual computing: a concise guide. *CRC Press*. 1-186.
Google Scholar↗ Crossref↗

[21] Le Callet, P., Viard-Gaudin, C., & Barba, D. (2006). A convolutional neural network approach for objective video quality assessment. *IEEE transactions on neural networks, 17*(5), 1316-1327.
Google Scholar↗ Crossref↗

[22] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.
Google Scholar↗

[23] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S. & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144.
Google Scholar↗

[24] Broomhead, D. S., & Lowe, D. (1988). Radial basis functions, multi-variable functional interpolation and adaptive networks. *Royal Signals and Radar Establishment Malvern (United Kingdom). 1*(1), 1-34.
Google Scholar↗

[25] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, *43*(1), 59-69.
Google Scholar↗ Crossref↗

[26] Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*. *1*(1), 1-18.
Google Scholar↗

[27] Tahmasebi, P., Hezarkhani, A., & Mortazavi, M. (2010). Application of discriminant analysis for alteration separation; sungun copper deposit, East Azerbaijan, Iran. *Australian Journal of Basic and Applied Sciences*, 6(4), 564-576.
Google Scholar↗

[28] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of machine Learning research, 3*(1), 993-1022.
Google Scholar↗

[29] Shonkwiler, R. W., & Mendivil, F. (2009). Explorations in Monte Carlo Methods. *Springer Science & Business Media.* 1-241.
Google Scholar↗

[30] Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence* 30(1), 2094–2100
Google Scholar↗

[31] Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4), 279-292.
Google Scholar↗

[32] George Karimpanal, T., & Bouffanais, R. (2019). Self-organizing maps for storage and transfer of knowledge in reinforcement learning. *Adaptive Behavior*, 27(2), 111-126.

Google Scholar↗

[33] Raina, R., Ng, A. Y., & Koller, D. (2006). Constructing informative priors using transfer learning. In *Proceedings of the 23rd international conference on Machine learning*, *1*(1), 713-720.
Google Scholar↗ Crossref↗

[34] Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, *40*(2), 44-58.
Google Scholar↗ Crossref↗

[35] Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), e1424.
Google Scholar↗ Crossref↗

[36] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, 4(37), 1-18.
Google Scholar↗ Crossref↗

[37] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*(1), 1-38.
Google Scholar↗ Crossref↗

[38] Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E. & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473, 1-58.
Google Scholar↗

[39] Angelov, P., & Soares, E. (2020). Explainable-by-design approach for covid-19 classification via CT-scan. *medRxiv*. *1*(1), 1-8.
Google Scholar↗ Crossref↗

[40] Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, *70*(1), 245-317.
Google Scholar↗ Crossref↗

[41] Publio, G. C., Esteves, D., Ławrynowicz, A., Panov, P., Soldatova, L., Soru, T. & Zafar, H. (2018). ML-schema: exposing the semantics of machine learning with schemas and ontologies. *arXiv preprint arXiv:1807.05351*. *1*(1), 1-5.
Google Scholar↗

[42] Chari, S., Gruen, D. M., Seneviratne, O., & McGuinness, D. L. (2020). Foundations of Explainable Knowledge-Enabled Systems. *arXiv preprint arXiv:2003.07520*. *1*(1), 1-26.
Google Scholar↗

[43] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., & Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, *1*(1), 648-657.
Google Scholar↗

[44] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion*, *58*(1), 82-115.
Google Scholar↗ Crossref↗

[45] Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, *16*(1), 1-85.
Google Scholar↗ Crossref↗

[46] Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137-141.

Google Scholar↗ Crossref↗

[47] Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019, October). Explainable AI: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing*, *1*(1), 563-574. Springer, Cham.
Google Scholar↗ Crossref↗

[48] Liao, Q. V., & Varshney, K. R. (2021). Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790*, *1*(1), 1-17.
Google Scholar↗

[49] Barberan, C. J., Balestriero, R., & Baraniuk, R. G. (2021). NeuroView: Explainable Deep Network Decision Making. *arXiv preprint arXiv:2110.07778*, *1*(1), 1-12
Google Scholar↗

[50] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, *6*(1), 52138-52160.
Google Scholar↗ Crossref↗

[51] Pawar, U., O'Shea, D., Rea, S., & O'Reilly, R. (2020, June). Explainable AI in healthcare. In *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, *1*(1), 1-2. IEEE.
Google Scholar↗ Crossref↗

[52] Pfahler, L., & Morik, K. (2021). Explaining Deep Learning Representations by Tracing the Training Process. *arXiv preprint arXiv:2109.05880, 1*(1), 1-8.
Google Scholar↗

[53] Seeliger, A., Pfaff, M., & Krcmar, H. (2019). Semantic Web Technologies for Explainable Machine Learning Models: A Literature Review. *PROFILES/SEMEX@ ISWC*, *2465*, 1-16.
Google Scholar↗

[54] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy*, *23*(1), 1-18.
Google Scholar↗ Crossref↗

[55] Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.
Google Scholar↗ Crossref↗

[56] Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018, March). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *In 2018 IEEE winter conference on applications of computer vision (WACV)*, *1*(1), 839-847. IEEE.
Google Scholar↗ Crossref↗

[57] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *In Proceedings of the IEEE international conference on computer vision*, *1*(1), 618-626.
Google Scholar↗

[58] Angelov, P., & Soares, E. (2020). Towards explainable deep neural networks (xDNN). *Neural Networks*, *130*(1), 185-194.
Google Scholar↗ Crossref↗

[59] Hussain, F., Hussain, R., & Hossain, E. (2021). Explainable Artificial Intelligence (XAI): An Engineering Perspective. *arXiv preprint arXiv:2101.03613*.
Google Scholar↗

[60] Gerlings, J., Shollo, A., & Constantiou, I. (2020). Reviewing the Need for Explainable Artificial Intelligence (xAI). *arXiv preprint arXiv:2012.01007*.
Google Scholar↗

[61] van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., & Viergever, M. A. (2021). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *arXiv preprint arXiv:2107.10912*.
Google Scholar↗

[62] Dieber, J., & Kirrane, S. (2020). Why model why? Assessing the strengths and limitations of LIME. *arXiv preprint arXiv:2012.00093*. *1*(1), 1-13.
Google Scholar↗

[63] Krishna Prasad, K. (2018). ABCD Analysis of Fingerprint Biometric Attendance Maintenance System. *International Journal of Applied Engineering and Management Letters (IJAEML)*, *2*(2), 53-70.
Google Scholar↗

[64] Aithal, P. S. (2017). ABCD Analysis of Recently Announced New Research Indices. *International Journal of Management, Technology, and Social Sciences (IJMTS)*, *1*(1), 65-76.
Google Scholar↗ Crossref↗

[65] Aithal, P. S. (2016). Study on ABCD analysis technique for business models, business strategies, operating concepts & business systems. *International Journal in Management and Social Science*, *4*(1), 95-115.
Google Scholar↗

****