International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Survey of Explainable AI (XAI) in Healthcare

Sanath NS[1], Dr Sivakumar[2]

[1]Undergraduate Student, Department of BCA, School of CS & IT, Jain [Deemed-To-Be-University], Bengaluru, Karnataka, India, sanath.nijagal@gmail.com
[2]Associate Professor School of CS & IT Jain [Deemed-To-Be-University], Bengaluru, Karnataka, India, sivakumar.n@jainuniversity.ac.in

Abstract:

Explainable AI is one of the trending concepts of research in the field of Artificial Intelligence. It mainly deals with providing an explanation by a system / machine about taking decisions or arriving at a solution to a problem. This paper gives an explanation about Explainable AI, Its need, Healthcare, discussing its history, Working of Explainable AI, the four principles used in XAI, Currently used approaches in Healthcare, and the challenges faced currently. The research areas as well as potential future directions are provided in the paper's conclusion.

Keywords: Explainable AI, AI, Healthcare, Machine Learning.

Introduction :

The word “Explain” is to make someone understand about a particular topic by telling in a way that he/she can easily understand. AI is a technology in which machines are fed with the intelligence, simulation to human intelligence. “Explainable” means how well a terminology, facts, technical terms, information and data can be conveyed to people.

"Explainability" refers to the ability of a system or process to provide understandable and transparent explanations for its decisions or actions. It is particularly relevant in fields such as artificial intelligence (AI), machine learning (ML), and data science, where complex algorithms are used to make decisions that impact individuals or organizations.

Explainability is crucial for several reasons such as Trust and Transparency, User Understanding, Bias and Fairness and so on. Understanding of the concept of Explainable AI will be further discussed.

Explainable AI :

The progress of artificial intelligence (AI) technology, particularly machine learning models, has led to noteworthy innovations in a variety of fields, including banking, healthcare, autonomous cars, and cybersecurity.

But the opaqueness of many complex models—like deep neural networks—has sparked questions about AI systems' accountability and interpretability. Explainable AI (XAI), an interdisciplinary subject that tries to provide insights into the decision-making processes of AI models and make them more comprehensible, justifiable, and accessible to both professionals and non-experts, has gained importance in order to overcome this challenge.

Explainable AI is a concept of Artificial Intelligence, which is used to clearly define the decisions taken by the system to arrive at a solution to a problem. For instance, while solving a ML problem, we need to first study the data and understand the relationship between input and output. Explainable AI (XAI) is used in this case to answer questions like “how the input and output variables are related to each other?”, “which feature has great importance on output and why?”, “what is the meaning of output?” and so on. It is applied in various sectors such as Healthcare, Automobile, HR, Insurance, Manufacturing, Defence, and so on.

Health Care :

Healthcare refers to the organized system of services, professionals, and facilities that aim to maintain, improve, or restore people's health. It includes a range of activities such as medical check-ups, treatments for illnesses or injuries, preventive measures, and the distribution of medicines. Healthcare involves various professionals like doctors, nurses, and specialists who work together to ensure the well-being of individuals and communities. The ultimate goal is to promote good health, prevent diseases, and provide care and support for those who are unwell.

Advanced Healthcare :

Advanced healthcare involves the application of cutting-edge technologies, sophisticated medical techniques, and innovative strategies to enhance the quality of healthcare. It consists of a wide range of advancements that contribute to more accurate diagnostics, effective treatments, and improved patient outcomes. Some aspects of advanced healthcare include:

- **Precision Medicine**: Tailoring medical treatment and interventions based on an individual's unique genetic, environmental, and lifestyle factors, allowing for personalized and more effective healthcare.
- **Telemedicine and Remote Monitoring**: Using technology to provide medical consultations, monitor patients remotely, and offer healthcare services from a distance, improving accessibility and convenience.
- **Use of AI in Healthcare**: Implementing AI algorithms for tasks such as medical imaging analysis, diagnostic assistance, drug discovery, and personalized treatment recommendations.
- **Robotics in Surgery**: Utilizing robotic systems to assist surgeons in performing minimally invasive surgeries, enabling greater precision and shorter recovery times, and so on.

Advanced healthcare strives to leverage technological innovations to improve patient outcomes, increase efficiency in healthcare delivery, and address healthcare challenges in a more personalized and effective manner.

History of Explainable AI :

Explainable AI is not a new topic in the present, but was found out 40 years ago.
When researchers were doing surveys on expert systems in AI, many scientists and researchers argued that the intelligent expert systems should not just accept input, process and produce output but also should provide clear explanation about the results obtained.
For instance, If an intelligent system predicts that a customer is not eligible to obtain a loan, then it has to explain the reason for why he/she is denied for the same.
However, Explainable AI (XAI) has become the present topic of research in the field of Deep Learning and Neural Networks as they are complex and there is a need for explanation of such complex models like ANN, CNN, RNN and so on.

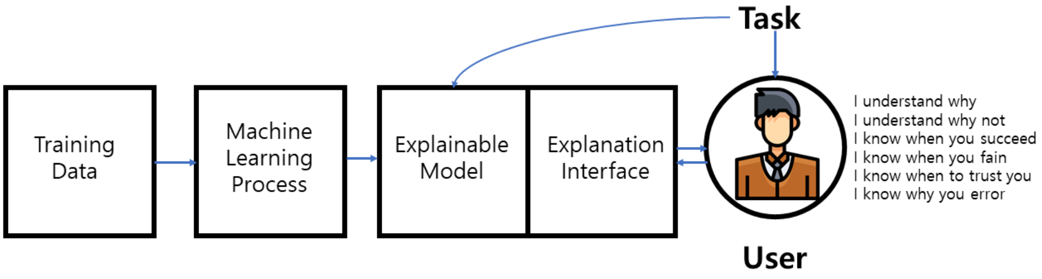
Need for Explainable AI :

With the advancement of Artificial Intelligence in the current industry i.e Industry 4.0, many algorithms are being used in Machine Learning to solve problems in various domains such as Education, Healthcare, Finance, Manufacturing, Supply Chain, Entertainment and so on.
However there is a need for a proper explanation about the logic / reason behind a system taking a decision to solve a problem.

For instance, if we need to predict the customer churn in OTT Platforms like Netflix, Amazon Prime, Hotstar, and others then there is a need to know which algorithm is best suited and how the system applies the algorithm step-by-step in detail.
Hence in this case, XAI is a required technology of AI, to fulfill the limitations in existing traditional machine learning models.

Explainable AI (XAI) will not just provide explanations to the users, but also provide specificity in decisions i.e., it answers questions like "Which decision is taken to get what solution?". It is also used for evaluation of decisions taken to solve the problem and in case of wrong decision, it can refigure the problem and provide accurate decision to obtain the desired solution, hence creating a great satisfaction to users and programmers.

Working of Explainable AI:



- ➢ **Training Data** - It is the first step where the data is collected from different sources, combined into a dataset and prepared and given to the system for solving Machine Learning problems.

- The data should be cleaned, numerical, transformed before feeding it to the system.

- **Machine Learning Process** - Machine learning is like teaching computers to learn from examples and experiences, and it involves a simple process:

- **Input Data**: We need to start with a bunch of information or data. It's like showing the computer lots of examples or giving it a bunch of experiences.
- **Training the Model**: Teach the computer by using algorithms (sets of instructions). The computer learns patterns and relationships in the data. It's similar to how we learn from examples and get better at something.
- **Testing and Evaluation**: After learning, the computer needs to be tested. We give it new data it hasn't seen before to see how well it can make predictions or decisions. It's like giving a test to check if it learned the right things.
- **Adjustments**: If the computer makes mistakes, we need to make some adjustments to the algorithms. It's like giving feedback to help it improve.
- **Prediction or Decision Making**: Once the computer has learned well, it can make predictions or decisions based on new data. It's like using what it learned to solve problems or give insights.

Explainable Model –

Once the result is obtained the model should be able to communicate the meaning of results and processes taken in a Human-Understandable Format to the end users.

Explainable Interface –

It acts as a medium between User and Explainable AI model to see the explanations and to attend the user queries about the results.

Principles of Explainable AI (XAI):

Explainable AI (XAI) works on mainly 4 principles - Explanation, Meaningful, Explanation Accuracy, Knowledge Limits.

1. **Explanation** is the 1st principle of Explainable AI. It means to provide proper evidence for processes carried out by the system. It involves providing an explanation for each and every process at every stage so that users can easily understand the logic used to solve the problem. Explanations given about the various processes carried out by the system should vary and should be different for different scenarios. This principle of XAI is independent of whether the explanation is useful, informative or sometimes even correct. It is also independent of any metric to measure the explanations.

2. **Meaningful** is the second principle of Explainable AI (XAI). The AI System successfully fulfills the 2nd principle only if the user / person has clearly understood the system's explanation. Explanations given to users should be user-friendly, in simple terms so that even a person from non-technical background can understand. Also the explanations given should be relevant and easy to understand. For example, The system first can provide an explanation as to Why one decision is taken? rather than telling Why the other decision is ignored?. So explanations can be made meaningful when it addresses the questions in a proper sequence. When it comes to giving meaningful explanations, different people find different sets of information as meaningful, hence the system should know well about preferences about different users and in different scenarios. This is an on-going area of research which is quite challenging.

- **Explanation Accuracy** is the 3rd principle of Explainable AI. The AI system accomplishes this principle only if it is able to judge between correct and incorrect information and provide complete reasoning for it being correct / incorrect.

- For example, let's say a Machine Learning Model is to be prepared for predicting whether a person is sick or not. If the actual value is "Yes" and the machine predicts as "No", then the machine should understand and explain that the prediction is wrong and provide a reason why it has made a wrong prediction and what is the level of accuracy in correct prediction. Researchers have discovered some metrics to measure the decision accuracy of the model, however for Explanation Accuracy, it is still under research. Explanation Accuracy depends on the level of detailing of information given about the correct / incorrect result. These can vary in different scenarios. In some scenarios the level of detailing might be high but some information may / may not be relevant to the users. On the other hand, the level of detailing might be low but the information given might be useful for users.

- **Knowledge Limits** is the 4th principle of Explainable AI (XAI). It is successfully accomplished when a system knows the limit of information it has to provide to the user and it is reliable and relevant to the queries in different scenarios. The aim of this principle is to make sure that the system gives limited and accurate information about the results. For instance, if a machine learning model is developed to categorize fruits and it begins to identify newly collected data as vegetables, it has exceeded its capacity to provide accurate information. Another instance of Exceeding its limits is when the system is trained with low quality data. For example, the system is trained with images of apples but quality is very low, it may / may not provide correct explanations.

Literature Review :

Various methodologies of building an explainable AI model are discussed by Andreas Holzinger in the healthcare sector. Explainability of AI Models is more crucial than the mathematical logics and principles behind the models. Machine learning models are needed in the healthcare industry to explain diseases and problems, along with their causes, symptoms, and recovery strategies, in addition to making predictions about them. These models must provide transparency, proper interpretation and explanations about the information to be conveyed.

The Explainability of AI can be categorized into 2 types - Ante-Hoc and Post-Hoc Models. "Ante-hoc" means "before this." Ante-hoc XAI models are designed to be interpretable and explainable from the very beginning, during the model's development phase. In other words, they are the models which provide explanations about some decisions right from the scratch.
Benefits of using this method includes - Transparency from Scratch, User Confidence, Reduction in complexity, Easier Model Evaluation and so on.
Examples of these kinds of models include, Decision Trees, Linear Regression, Fuzzy-based models, etc.

"Post-hoc" means "after this." Post-hoc XAI models are applied after the AI model has already been trained and developed, to provide explanations for its decisions dynamically. In other words, they are the models which are already built and developed and provide explanations which are related to the present models.
Benefits of using this method includes - Relevant information is extracted, Model Understanding, Retrospective Explanations and so on.
Examples of these kinds of models include, BETA (Black-Box Explanations using Transport Approximations)
It has been discovered that the Local Interpretable Model-Agnostic Explanations (LIME) algorithm may explain predictions of any model. First the model selects the feature “x” for which the explanation is to be given. Introduce some small changes in the data to be similar to the original feature / instance in order to generate multiple instances of the chosen feature. Further a model is built to predict the outcomes. Finally, we interpret the model by understanding the relation between the input and output features. Once the insights are obtained the next and final thing is to provide explanations about the findings in user-understandable format.
Attention models refer to mechanisms that highlight and prioritize specific elements or features in the input data, shedding light on the factors that contribute to the model's decision-making process. It is a mathematical construct that assigns weights to different parts of the input data, emphasizing the most relevant information during the decision-making process of a machine learning model. In electronic health records (EHRs), the model might focus on specific symptoms, lab results, or temporal patterns that are crucial for diagnosis or prognosis. In medical imaging, spatial attention can highlight relevant regions in radiological images, such as tumors or abnormalities, assisting radiologists in understanding the AI model's decision. Likewise It can be used in many other ways in the Healthcare Sector.

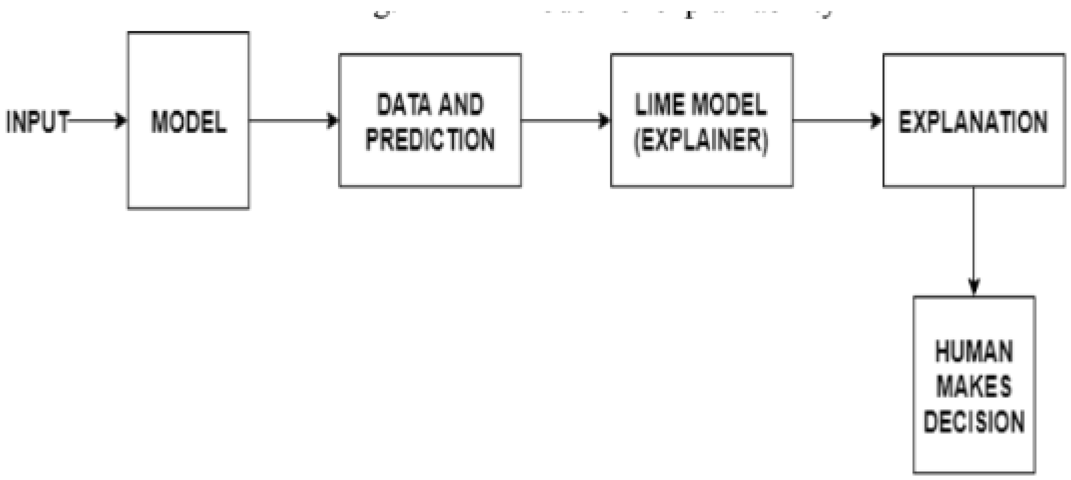
Current Approaches:

Several approaches are currently used to achieve explainability in healthcare AI systems:

Local Interpretable Model-Agnostic Explanations (LIME):

LIME Models are the post-hoc models which generate explanations for a simpler, interpretable model that approximates the complex model's behavior in a specific region. LIME can be applied to explain predictions made by complex models like neural networks in medical image analysis. Explanation to predictions involves providing textual and visual artifacts that help to understand the relation between data instances.

Steps for Implementation of LIME Method:-



1. Construct the RNN (Recurrent Neural Network) Model for Prediction.
2. Feed the model into the LIME Algorithm for creating an interpretable model. This will give textual / visual representation of explanations.

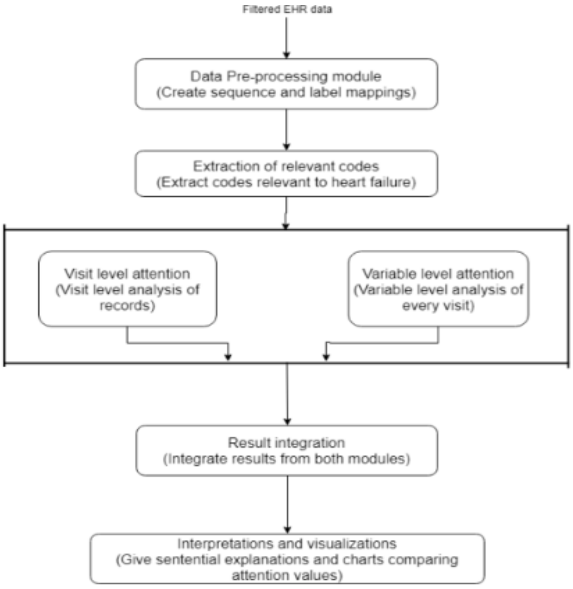
The advantage of using this method is that RNNs are particularly well-suited for sequential data, making them effective in tasks such as natural language processing or time-series analysis. When applied to sequential data, LIME can help interpret the decisions made by RNNs at specific time steps or for specific input sequences, providing insights into the model's reasoning.

According to a survey, the attention model gives an accuracy of 80%..

Attention Mechanisms:

This is a method used in deep learning models. The attention mechanisms highlight or undermine the input features that are most important to the model's decision, providing insights about required inputs. They can be applied in medical image analysis, where the model focuses on specific regions of an image for making a diagnosis.

Steps to Implement the Attention Mechanism Method:


Filtered EHR data
Data Pre-processing module (Create sequence and label mappings)
Extraction of relevant codes (Extract codes relevant to heart failure)
Visit level attention (Visit level analysis of records)
Variable level attention (Variable level analysis of every visit)
Result integration (Integrate results from both modules)
Interpretations and visualizations (Give sentential explanations and charts comparing attention values)

1. Create and Pre-Process the Data.
2. Extract the relevant features of the Input data.
3. Perform Data Analysis.

4. Predict the Outcomes.
5. Interpret and Visualize the Explanations for Results

According to a survey, the attention model gives an accuracy of more than 80%.

Challenges Faced:

LIME and Attention based models have proven to be effective in providing interpretable explanations for complex machine learning models, however both the models have also faced some challenges.

A. Challenges In LIME Model :

- It is necessary to carefully adjust LIME's hyperparameters, which include the amount of samples and the interpretable model's complexity. These hyperparameters may have an impact on how well the LIME Model explains things, thus experimenting a little to determine the ideal values may be necessary. Due to its interpretable model's simplicity, some aspects of the original model may be ignored in favor of a biased explanation.
- LIME was initially designed for tabular data. Hence its application to non-tabular data, such as images or text, may require additional modifications or adaptations.
- The assumption of Independence of Input Features might not hold good in all cases, especially when dealing with complex relationships between features. In scenarios where feature interactions are essential, LIME's local explanations may not accurately capture them, thus impacting the overall model behavior.
- LIME focuses on providing local explanations for individual predictions. This is useful for understanding specific instances, however it may not capture the overall behavior or trends of the model. Understanding the overall behavior of the model is important for the decision-making process.

B. Challenges In Attention Mechanism Model

- While attention mechanisms highlight the importance of certain parts of the input features or images, understanding why specific weights are assigned to particular elements might not always be straightforward. This lack of interpretability can limit the trust-worthiness of attention-based explanations.
- Attention mechanisms reveal only the correlations between input and output. Understanding why a model pays attention to a particular region or feature doesn't necessarily explain the causal relationship between that attention and the model's decision.
- This can lead to misinterpretations or incomplete explanations.
- Small changes in the input sequence or image may result in different attention patterns.
- This sensitivity to input variations can make it challenging to provide consistent and reliable explanations, especially in scenarios where subtle changes can have a significant impact on the model's decision.
- The visual representation of attention may not always align with human intuition of explanations, and different visualization techniques may result in inconsistent or misleading results.
- Attention mechanisms, especially in large models like transformers, can involve a vast number of parameters and computations. This scalability can pose challenges for both training and inference, making it computationally expensive and complex.

Future Directions:

LIME and Attention Mechanism Models for Explainable AI can be improved. LIME Models can be extended by training the model with non-tabular data like images and text to check for the explanations which the model provides.

Attention Mechanism Models can be extended by providing proper explanation as to why the model pays attention to a specific feature in the data.

A Hybrid model comprising of both LIME and Attention Mechanism Models can be tried out which have the following benefits :-

1. Proper Understanding of Correlations between input and output features using Attention Mechanism which is useful for solving complex relationship problems in the LIME Model.

2. The Hybrid Model / Ensemble Approach can be used to improve the overall accuracy in predicting the results which can be considered as an area of future research.

3. Proper Decision Making Process.

Conclusion:

In conclusion, Explainable Artificial Intelligence (XAI) plays a pivotal role in addressing the opacity and complexity associated with advanced machine learning models, particularly in critical domains such as healthcare. The need for Explainable AI systems is driven by the growing reliance on these technologies in decision-making processes, where the consequences can have significant real-world impacts such as predicting the possibility of heart failure with a proper explanation of causes and symptoms can help doctors by giving them alerts on time and solve the problem of re-admissions.

References :

[1] Andreas Holzinger, Chris Biemann, Constantinos S.Pattichis, Douglas B. Kell,"What do we need to build explainable AI systems for the medical domain",arXiv:1712.09923v1,2017.

[2] Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE transactions on neural networks and learning systems, *32*(11), 4793-4813.

[3] Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., & Przybocki, M. A. (2021). Four Principles of Explainable Artificial Intelligence. 2021. *Gaithersburg, Maryland*.

[4] D. Bahdanau, K. Cho, and Y. Bengio"Neural machine translation by jointly learning to align and translate",ICLR, 2015.

[5] Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, *263*, 110273.

[6] Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio,"Neural Machine Translation using Attention mechanism paper", ICLR, 2015

[7] [B¨ackstr¨om et al., 2013] B¨ackstr¨om, C., Jonsson, P., and St˚ahlberg, S. (2013). Fast detection of un- solvable planning instances using local consistency. In Sixth Annual Symposium on Combinatorial

[8] Search. [Baehrens et al., 2010] Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and MA˘zller, K.-R. (2010). How to explain individual classification decisions.˜ Journal of Machine Learning Research, 11(Jun):1803–1831.

[9] Gohel, P., Singh, P., & Mohanty, M. (2021). Explainable AI: current status and future directions. *arXiv preprint arXiv:2107.07045*.

[10] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. ProPublica, May 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[11] Islam, S. R., Eberle, W., Ghafoor, S. K., & Ahmed, M. (2021). Explainable artificial intelligence approaches: A survey. *arXiv preprint arXiv:2101.09429*.

[12] Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg More Employable Than Lakisha and Jamal?: A Field Experiment on Labor Market Discrimination. American Economic Review, 94(4):991–1013, 2004. doi: 10.4324/9780429499821-53.