A survey on Unveiling the Black Box: The Role of Explainable AI in Transforming Health Care

[1]D.Swetha,
[1] Assistant Professor, Department of AI &DS ,Guru Nanak Institute of Technology, Ibrahimpatnam,Hyderabad.

[2]Sudha Singaraju
[2] Senior Assistant Professor, Department of CSE, Geethanjali College of Engineering & Technology, Cheeryal ,Hyderabad.

[3]Telugu Prathiba
[3] Assistant Professor, Department of IT ,Sridevi Women's Engineering college, Near Wipro , Gopanpally Campus, Hyderabad.

ABSTARCT:

Explainable Artificial Intelligence (XAI) in Healthcare Systems aims to improve transparency, trust and interpretability in AI's healthcare decisions. Although traditional black box key models are lacking very accurately, they often do not have the ability to explain predictions and limit the introduction of critical health treatments. XAI technologies such as shape, lime, attention mechanisms, and explanations of inconsistencies can help physicians understand how AI models can reach diagnosis, treatment recommendations, and risk assessments. XAI ensures ethical and fair use of AIS in healthcare by improving trust, regulatory compliance and error awareness. This article examines the importance of methods, applications, challenges and future directions from XAI, highlighting its role in improving patient outcomes and promoting human collaboration.

Keywords: XAI, SHAP, LIME,ANCHOR

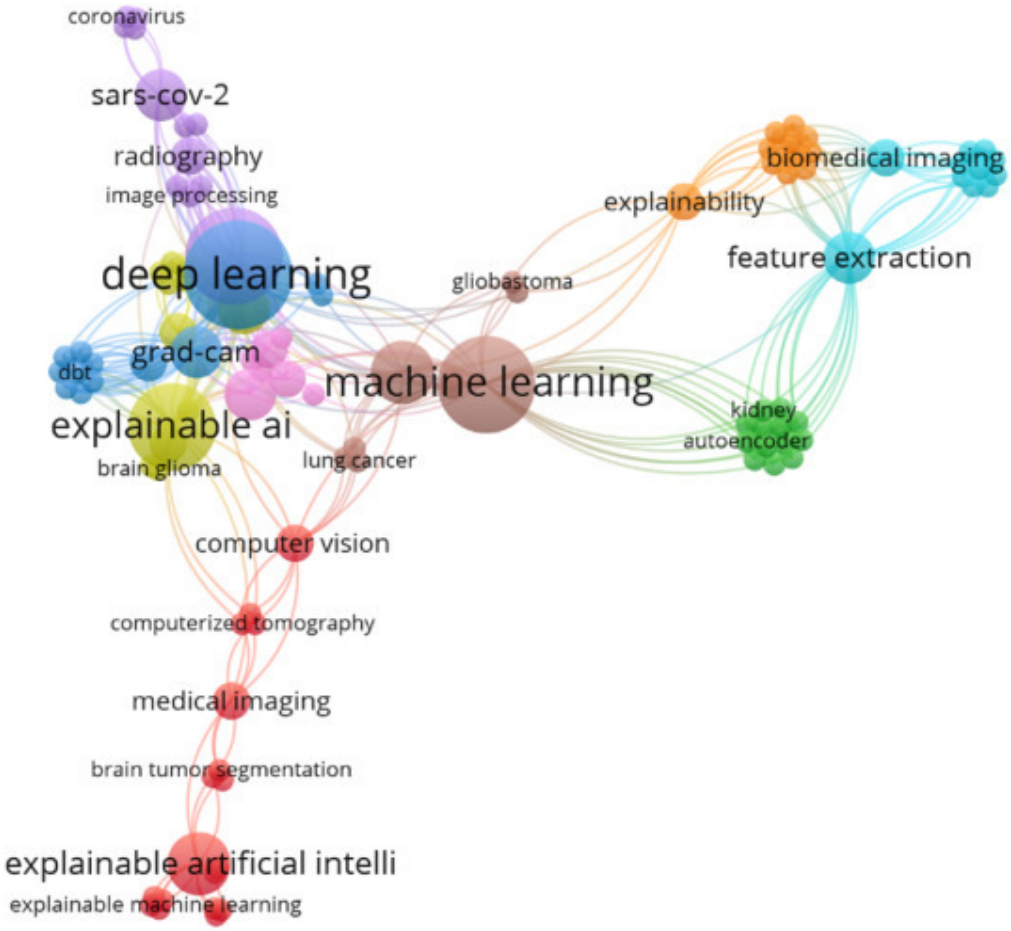
1. INTRODUCTION:

Artificial intelligence (AI) demonstrates the key potential of healthcare, from diagnosis of disease to predict patient outcomes and recommendations for treatment. However, many AI models, especially deep learning-based models, act as "black boxes," making it difficult for physicians to interpret their decisions (Doshi-Velez & Kim, 2017). Explanatory AI (XAI) deals with this issue by providing a transparent and interpretable decision process that is essential for trust, ethical considerations, and formal health care compliance (Ghassemi et al., 2021).

The new research has shifted focus from technology to how technology is responsibly used to improve health services and patient outcomes. Many of today's diagnostic tools for artificial

intelligence (AI) and machine learning (ML) that provide an explanation of why a particular patient has a particular type of illness [3,4]. The black box concept of ML and AI could lead to a small use of such methods in healthcare [5]. A keynote speech entitled "Next Boundary: AI, We Are Really Trusted" and a paper from the bibliography. Subsequent research from the references highlights the importance of robust and explainable systems of artificial intelligence (AI) especially in key areas such as drugs with low data quality. The lack of robustness and explanation among the most powerful learning methods makes it difficult to rely on and explain why certain results were achieved. This article highlights the need for explanation and robustness of AI technology to achieve reliability and trust, ensuring that people control their decision-making processes. As highlighted in this study, the integration of conceptual knowledge and information can be used in a comprehensive and integrative way to help develop more robust, easier to explain, and unbiased ML and AI models.

Over the past decade, a wide range of deep learning algorithms have been proposed to address the issues mentioned. For example, the CNN method was applied to Vivo guest images with Shapley Additive Description (Shap), Locally Interpretable Model Tag Description (LIME), and Context-Related Importance and Usefulness (CIU) [10]. Other studies have presented a method called DBSCAN (EMR) for electronic medical records (EMR) using XAI methods with lime, shape and anchor .



2. RESEARCH METHODOLOGY:

Explanatory AI (XAI) techniques have the ability to explain their own behavior, identify their strengths and weaknesses, and communicate their understanding of future behavior. XAI's strategy is to pursue numerous techniques for creating many ways to provide future developers with a wide range of design options that cover performance and explanation compromises Expert systems descriptions not only explain what the system is doing, but also why it does it, its knowledge, and why it requires design and compilation of the system. The Explanatory Professional System (EES) framework was developed to capture the design aspects of expert systems that are important for generating excellent explanations. EES focuses on determining how general principles were used in a particular area or in a particular case, allowing general principles to be expressed from what a system was derived, and how a system was derived from these principles. To support proper explanations, the system must record additional knowledge and present the explanations flexibly and quickly. Additionally, the system needs to understand how to understand how the description is designed if it does not understand or

fails. The increased complexity of AI systems and behavioral models in military simulations has made it difficult for users to understand the activities of computer-controlled units. The simulator has been added to the prototype declaration system, but it is not modular and not portable. The designer does not consider the teachings drawn from the work when describing the behavior of the expert system. New modular and general domain-independent architectures for explaining the behavior of simulated entities have been proposed in several studies with the ability to explain the motivations behind entity actions and modularity to enable external components such as GUIs, natural language generators, and external components that can be moved using the system

2.1 SHapley additive exPlanations (SHAP)

Shap (Shapley Additive Explanations) is an interpretable method for machine learning (Lundberg and Lee 2017). SHAP was proposed to find spaces from the perspective of color diagrams and explain predictions to visualize the importance of input variables (e.g. surface data, pixels, etc.) for prediction. This method, for example, distinguishes between meaningful forecasting strategies. This classifies images of "normal or abnormal abnormalities" for detection of Covid-19 , or classifies the classification of antifungal peptides (positive samples) and non-antifunctional peptides (negative samples).

2.2 Local Interpretable Model-agnostic explanations (LIME)

The concept of an interpretability model - Lime (local interpretability model - impaired explanation) was proposed [9]. Lime can continuously handle patterns of properties and provide local estimates for interpretation of individual predictions. Therefore, we show the effect of each characteristic on the model results. For example, lime can recognize that this property has the greatest impact on the medical image.

2.3 Gradient-Weighted Class Activation Mapping (Grad-CAM)

Class Activation Assignment (CAM) is a technique that uses gradient information from the final layer of a CNN to create rough cards for important areas in a classification-based photograph. Grad-Cam is a wider, more common version of CAM and can be used in CNN-based architectures if necessary. Grad-Cam comes in two categories: Grad-Cam and Grad-Cam++. Both use class-specific gradient information, but Grad-Cam++ includes additional features to improve localization cards.

2.4 NeuroXAI

Neuroxai uses a deep neuron network (CNN) that analyzes brain images and generates special output calculations including foldable function cards and visualization of clusters for tumor classification/segmentation. Neuro XAI can perform classification and segmentation on two- and three-dimensional medical images. In the continuation of this phase, the results of the healthcare professional are evaluated to ask for explanations as needed. Previous studies have used interpretable methods to improve the interpretability of deep learning models for image classification. However, neuro XAI can be used to improve image segmentation by converting segmentation tasks into multi-label classification tasks. This is achieved by using the global average bus for each class obtained from the initial forecast layer.

2.5 The contextual importance and utility (CIU)

Context-related meanings and benefits (CIU) use the concept of attribute weights and its trust in other attributes. It considers correlations of the importance of attributes. If the combination of properties is appropriate or causes interactions in prediction, it is interpreted as a distinctive interaction, leading to explanation at a higher level.

2.6 ANCHOR (anchors: high-precision model-agnostic explanations)

The anchor method describes predictions in a black box classification model by finding decision limits that "anchors" the predictions well. The declaration of an anchor is related to rules that allow you to effectively "fix" predictions for a particular local context within the analyzed instance. This means that changes in other characteristic values of the instance have no significant impact on the ability to explain the prediction. Anchor technology is based on reinforcement learning methods and diagram search algorithms. The number of model calls required for this term must be minimized and efficiently restored from the local optimizer.

2.7 T-Distributed Stochastic Neighbor Embedding (t-SNE)

T-SNE, like other methods, is a dimension reduction method that visually represents relationships between samples. The result is a two-dimensional or three-dimensional map in which the proximity of the mapped points reflects similarity between corresponding samples. The goal of T-SNE is to maintain similar samples proximity and simultaneously increase the distance between different samples, as well as other methods of reducing dimensions. In contrast to other methods, T-SNE offers a nonlinear approach that allows for more flexible compromises between local and global relationships between data points. As a result, T-SNEs

often produce visually attractive clusters compared to other techniques.

3. Applications of explainable AI in healthcare

Today, artificial intelligence (AI) plays an important role in pursuing critical systems such as education, healthcare, renewable energy, transportation, and transportation that impact our daily lives. In healthcare, AI technology is constantly advancing However, transparency and explanation are required for AI applications and models in healthy practices as inaccurate predictions can have serious consequences . Clinics require that AI systems be understood as a prerequisite for predicting and establishing trust in AI applications acquisition . Reliability, accuracy and transparency are important requirements, especially for healthcare decision makers. Therefore, AI researchers and practitioners focus on explaining decisions for AI applications such as ML and Deep Learning (DL). AI algorithms should provide clinicians with an understanding explanation of the outcome . In the diagnosis of disease diagnosis, XAI can demonstrate properties that contribute to the AI model problem for the patient's condition. The Shapley Additive Description (SHAP) algorithm was used to understand the relationship between microbial communities and phenotypes. The motivation is that SHAP technology can explain the prediction of a particular prototype value depending on the output of effective parameters. Characteristics have a positive effect on predicting target values when SHAP values are positive. Dopaminergic pictorial techniques such as Spect Datscan were analyzed in early diagnosis of Parkinson's disease . Limestone technology was used to accurately classify Parkinson's disease from a particular Dutch can. Acute critical detection of disease is another important application of the XAI approach in medical research. For example, an early warning score system was proposed . The system was able to use SHAP technology to explain its predictions using data information from electronic health data. XAI for diagnosis of glioblastoma based on topological and textual features was also examined . An AI model for restoring liquid edges for diverse classification of glioblastoma has been validated. Local property associations of samples in the test set were calculated using the lime method.

3.1 Interoperability and Visualization

In addition to other application verticals, there is a lack of adequacy in interoperability and visualization, including the human attention's ability to comprehend explanation maps and the measures used to confirm the correctness and completeness of explanation maps produced by

the EXAI system. This necessitates the provision of improved explanations for enhanced visualization and interoperability of the explanation map in applications of vital importance to the mission. In such situations, EXAI models such as SHAP and LIME can elucidate the predictive variables (input) by calculating the contribution of features to the output.

3.2 Human Machine Interaction

Human-Machine Interaction It is crucial to design, develop, and implement responsible AI that focuses on human needs. For models to be comprehensively explained to the user, interaction between humans and machines is necessary. Context-aware explanations are generated by adaptive explainability modeling, which takes into account a variety of human profiles. EXAI research is influenced by the intersection of empirical studies in social science, human behaviour, and human-machine interaction. A human-enabled feedback mechanism based on derived machine explanations (whether visual or logical) can improve human-machine interaction by integrating transparency, ethics, judgment, and social norms. Enhanced human-machine interaction produces an explanation map tailored to the appropriate user, facilitating their comprehension of results—this is particularly relevant in clinical settings. This necessitates additional investigation to enhance the research techniques used to pinpoint the right problem.

3.3 Decision Support Systems

Healthcare tools are integrated into decision support systems to improve the decision-making process and supply knowledge as well as individualized data to all parties directly engaged in a health information system. It is essential to include collaborative contributions from all stakeholders across various domains when developing an ML-based clinical decision support system, as this influences the overall outcome. Nonetheless, different parties from other domains encounter the problem of unstructured medical data. By examining data mining and explainability within the model, privacy of the exchanged data is ensured while allowing for the extraction of significant information to capture essential medical details.

3.4 Integration With AI

Systems and algorithms based on AI demand vast amounts of data and energy. This creates

ongoing requirements for computing systems (like cores and graphical processing units) to function efficiently. Modelling based on ML and DL yields results that are not known in advance and cannot be predicted. For AI methods to reach human-level accuracy, they need hyper-parameter optimization, fine-tuning, strong computing resources, large datasets, and ongoing data training. These data originate from millions of user IoT devices and are susceptible to cyber-attacks. Data produced by AI-based systems also exhibits bias; therefore, EXAI offers clear explainability to facilitate human-level intelligence.

4. Conclusion:

Digital wellness is now the focus of healthcare 5.0 ecosystems, where analytics-driven decision models with real-time forecasts and informatics assistance are used. AI models have changed to incorporate EXAI decision modules due to concerns about interpretability and the validity of AI models.In addition to enabling interpretability and model debugging, which improve performance by reducing bias, EXAI increases confidence in clinical procedures.The survey provides an overview of EXAI's fundamentals, related KPIs, and various use-case implementations.For the purpose of categorizing and segmenting COVID-19 patients, an EXAIdriven architecture is suggested.Performance evaluations that confirm the advantages of EXAI in healthcare settings are included in the case study.Lastly, the survey's lessons learned, research obstacles, and outstanding topics are explored.

References:

1) Ahmed, A., Topuz, K., Moqbel, M., & Abdulrashid, I. (2024). What makes accidents severe! Explainable analytics framework with parameter optimization. European Journal of Operational Research, 317(2), 425–436. https://doi.org/10.1016/j.ejor.2023.11.013

2) Asilkalkan, A., Dag, A. Z., Simsek, S., & Aydas, O. T. (2023). Streamlining patients ' opioid prescription dosage: An explanatory bayesian model. Annals of Operations Research. https://doi.org/10.1007/s10479-023-05709-4

3) Bal, M. I., Iyigun, C., Polat, F., & Aydin, H. (2024). Population-based exploration in reinforcement learning through repulsive reward shaping using eligibility traces. Annals of Operations Research. https://doi.org/10.1007/s10479-023-05798-1

4) Bohlen, L., Rosenberger, J., Zschech, P., & Kraus, M. (2024). Leveraging interpretable machine learning in intensive care. Annals of Operations Research. https://doi.org/10.1007/s10479-024-06226-8

5) Borchert, P., Coussement, K., De Caigny, A., & De Weerdt, J. (2022). Extending business failure prediction models with textual website content using deep learning. European Journal of Operational Research, forthcoming. https://doi.org/10.1016/j.ejor.2022.06.060

6) Coussement, K. (2014). Improving customer retention management through cost-sensitive learning. European Journal of Marketing, 48(3/4), 477–495. https://doi.org/10.1108/EJM-03-2012-0180

7) Coussement, K., Harrigan, P., & Benoit, D. F. (2015). Improving direct mail targeting through customer response modeling. Expert Systems with Applications, 42(22), 8403–8412. https://doi.org/10.1016/j.eswa.2015.06.054

8) Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data Preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decision Support Systems, 95, 27–36. https://doi.org/10.1016/j.dss.2016.11.007

9) De Bock, K. W., Coussement, K., & Lessmann, S. (2020). Cost-sensitive business failure prediction when misclassification costs are uncertain: A heterogeneous ensemble selection approach. European Journal of Operational Research, 285(2), 612–630. https://doi.org/10.1016/j.ejor.2020.01.052

10) Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774