XAI Healthcare: A Comprehensive Survey of Explainable AI Techniques in Healthcare

[1]Amrita Koul , [2] NP Singh
[1,2] Department of Computer Science and Engineering, MVN University Palwal, India

Corresponding Author: Amrita Koul

Abstract: Explainable Artificial Intelligence (XAI) has risen as a pivotal advancement in dealing with the challenges of interpretability and transparency in AI-driven healthcare systems. AI's rapid integration into healthcare has shown immense potential in diagnostics, prognosis, personalized medicine, and decision-making. However, the absence of proper explainability in traditional AI models has raised critical concerns regarding trust, ethical accountability, and clinical adoption. The research paper examines both XAI methodologies and healthcare applications in present scenario, emphasizing how these techniques help increase the interpretability in case of complex models without compromising predictive accuracy. It delves into the pivotal part of XAI in improving clinical decision support systems, risk stratification, and patient engagement, while also addressing regulatory compliance and ethical considerations. By analyzing recent advancements, challenges, and future prospects, this paper provides insights into how XAI can bridge the gap between real-world healthcare applications and AI innovations, cultivating trust while enabling safer, more effective healthcare delivery.

Keywords: XAI, CNN, LORE,CDSS, DCIP

Introduction

Explainable Artificial Intelligence (XAI) is being progressively more recognized as essential for the adoption of AI-driven medical systems, ensuring veracity, trust, and decision-making. Research has evaluated the role of explain ability from diverse perspectives—technical, legal, medical, and patient-centered, highlighting its importance in fostering informed and better decision-making. The ethical implications of XAI have been evaluated using biomedical principles, reinforcing its necessity for just and proportionate healthcare practices. XAI helps build confidence in AI tools, encourages ethical use, and supports better decision-making in clinical settings. Studies have also explored the cognitive gap between the developers and the clinicians in designing explainable AI solutions, identifying key variations in their mental models and goals. To account for these challenges, several

methodologies have been proposed such as causal inference models, personalized explanations, and balancing exploratory and quantitative approaches. Various XAI techniques, including local rule-based explanations, interpretable ML models, and data visualization methods, have been utilized to further enhance model transparency in such clinical settings. In this paper multiple techniques of XAI have been explored so as to make the model decisions more interpretable and how its applied in the medical field. This paper provides a method-oriented review of XAI techniques in healthcare. We categorize, compare, and critically evaluate existing approaches based on their technical foundations, interpretability capabilities, and relevance in clinical applications. By doing so, we aim to support both AI researchers and healthcare professionals in selecting and applying the right explain ability tools for safe, transparent, and effective AI deployment in medicine. This paper tells about the research in the field of Explainable Artificial Intelligence and how it's helping to make the healthcare system more transparent and reliable for the patients. Through Explainable Artificial Intelligence we can bridge the gap between high level technology and the patient.

Related Work

The research by Aman et al. [1] examined medical AI explain ability needs thoroughly while conducting an ethical assessment about explain ability's role in AI tool implementation in clinical care. The medical AI explainability assessment by these AI-based clinical decision support systems used multidisciplinary surveys to evaluate explainability importance across medical, patient and legal and technological domains. The authors conducted ethical investigations to determine medical AI require explainability after their conceptual analysis and employing the Beauchamp and Childress principles as an evaluation tool (autonomy, beneficence, nonmaleficence, and justice). The researchers emphasized explainability serves as a means for patients together with healthcare providers to make wise independent healthcare choices which maintains patients at the center of their care. The implementation of explainability systems helps achieve equal resource distribution among patients.

The authors of Bienefeld et al. [2] published their research results from a multi-method longitudinal study which brought together 112 developers and clinicians who worked together to build an XAI solution for clinical decision support systems. Researchers revealed three fundamental distinctions which exist between mental models held by developers and medical clinicians about XAI through their study. These differences involved conflicting objectives between model interpretability and clinical plausibility as well as separate truth sources between data and patients alongside different approaches to new versus existing knowledge exploitation. To tackle this, the authors proposed a more collaborative, human-centered approach to building XAI systems. They argue that developers should work side-by-side with

clinicians, involving them early in the design process. This ensures the tools that get built are not just explainable in theory, but actually helpful in practice.

For AI to truly support healthcare, the explanations it provides must be designed with the end-user in mind. Their research findings enabled them to propose solutions which incorporated causal inference models together with personalized explanations alongside dual mindsets that support exploration and exploitation. Both developer and clinical perspectives need attention in XAI system design according to this research which also offers useful guidelines for healthcare XAI effectiveness improvements. They interviewed 112 healthcare professionals (nurses and physicians) along with three software developers (one data scientist and two senior software engineers and one senior visualization designer) who worked in Switzerland at a large university hospital during the co-creation process of the DCIP (an ML-based CDSS for aSAH patient Delayed Cerebral Ischemia prediction). The analysts used IBM SPSS version 23 for their statistical analysis by performing descriptive measurements and regression models on survey data. High-fidelity interface development incorporated an interactive UI prototype that drew its design from the DCIP system.

Figure 1 visually summarizes the differences in mindset and interpretability needs between developers and clinicians, and provides actionable recommendations. It supports discussions on interdisciplinary challenges, stakeholder perspectives, and strategies for bridging the gap when designing XAI for clinical use.

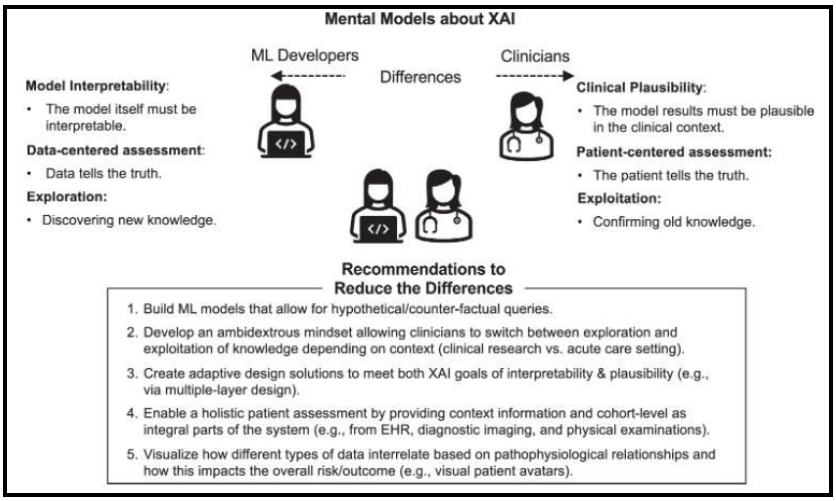

Figure 1: Mental Models about XAI – ML Developers vs Clinicians, Recommendations to Reduce the Differences

Metta et al. [3] research involves Local XAI methods and specifically the LORE i.e. Local Rule-Based Explanations technique used in healthcare and medical fields. This research revealed the sensitivity to clarification of the AI system and its transparency in detection methods in terms of diagnosis accuracy as well as prediction and generating treatment. The efficient LORE framework produces

efficient local explanations and interpretable explanations for machine learning models. In LORE, we use a genetic algorithm to develop synthetic data to set up a training base for the development of local interpretable prediction models. Built in logic provides the meaning through which the model decides on its decisions and understanding their interpretation is the strength of the predictor. Its approach was to build local interpretable models to cover analytical predictions and as a result, LORE functioned to create customized explanations. The first instance to be explained is picked by the method itself. The synthetic dataset generated by the genetic algorithm was found to replicate the local characteristics of original instance. The local dataset trained a decision tree for approximate modeling of the complex model behavior within this area. This brought up critical understanding on the complex health care XAI systems and potential benefits, as well as the significant problems health care professionals face.

Farahani et al. [4] developed a data extraction sheet that they applied to random studies to adapt it to its final form. The data were obtained in a procedure through which one review author (KF) checked the selected studies while a second review author (FF) confirmed the collected data. The analysis consisted of taxonomic topic, first author and year of publication, essential contributions, used XAI model along with sample size (if appropriate). All the conflicting views of the two review authors were resolved through discussion between them or BL or WK as required for final review. The co-occurrence relationships among the common points of the analyzed studies (such as XAI methods, diseases and ML/DL terminology as well as imaging modalities) were reviewed and analyzed by the authors. The development of the co-occurrence network requires first finding text keyword and then calculating co-occurrence frequency for network examinations and word clustering with central terms detection.

XAI visualization was used by Papanastasopoulos et al. [5] to understand features trained by their DCNN classifying ER+ versus ER- breast status of DCEMRI images. Minimum scanning with two contrast scans and total collection of 1395 ER+ and 729 ER- regions-of- interest (ROI) on 148 patients were obtained. Based on Alex Net model trained on Image Net, a dual domain transfer trained DCNN architecture was developed by the researchers as they used spatial and dynamic data from each DCE-MRI ROI as the input to that architecture. The performance of AUC is measured for network evaluation is obtained from a leave-one case out cross validation. In this context, XAI methodologies and its attribution methods (Integrated Gradients and Smooth Grad noise algorithm) were applied on training set ROIs to visualize DCNN learning. However, our DCNN extracted suitable spatial and dynamic domain features and the contributing features differed between domains according to their analysis. Their observation: DCNN learnt irrelevant pre-processing artifacts and concluded that to extract features from these.

Boutorh et al. [6] developed Support Vector Machine (SVM) and Random Forest (RF) through analyzing Bio-data symptoms in combination with deep Convolutional Neural Network (CNN) analysis of chest computed tomography (CT) images to detect COVID-19 positive cases. We needed to prove the possibility of explaining XAI systems through LIME as an interpretable framework for model explanations about positive virus patients. The experimental outcomes exceeded the present standard of practice. Evaluation of the CT-scan image revealed accuracy of 96% along with F1-score, and SVM outperformed RF as 90% accuracy and 91% specificity were recorded over the Bio-data. Finally, for the XAI-Img and XAI-bio models, we obtained interpretable results of trained SVM and CNN black box models after arriving COVID19 dataset of different kinds through LIME explanations. This improved examination procedure proves to enhance trust degrees, whereas experts find new patterns of the pandemic.

In the context of pulmonary nodule diagnosis Wang et al. [7] set up a deep learning model that is explainable as a multi-task system. In addition to identifying diagnostic signs, the neural system makes both lesion malignancy predictions. Researchers can then display the location of each manifestation for the purposes of visual interpretability. Experimental results obtained from the LIDC public database demonstrated a test AUC value of 0.992 and a test AUC of 0.923 using the in-house dataset. Experimental outcomes demonstrate that integrating manifestation identification tasks into the multi task model enhances the accuracy level in cancer classification. Effective systems can improve radiologist–clinical activities through the proposed multi task explainable model.

In healthcare, Explainable AI (XAI) is recognized as 'necessary' by S. S. Band et al. [8], in the context of challenges related to interpretability in AI driven decision making. Our team evaluated different XAI evaluation methods including LRP, LIME, SHAP, Grad-CAM and t-SNE for medical diagnosis application. The results were found to show widely used explainability methods of various nature along with visual descriptions in the medical field then rule based and numerical approach. Attention maps, saliency maps, heat maps and other visualizations demonstrate better effectiveness in helping medical professionals with the decision making process during diagnosis. It was discovered that explainability can help align deep learning models better with clinical decision making, by increasing trust, transparency and adoption.

The use of XAI in smart healthcare applications allowed U. Pawar et al. [9] to enhance AI system transparency and responsibility according to their research. By using the clinical knowledge (and dedicated work) in combination with existing XAI models, they were able to get greater advantages in the system based on AI to validate prediction, improve the model and make decisions. It was stressed that intuitive interfaces perform well in supporting intuitive interpretation, fulfilling

regulatory requirements for traceable decisions of health care based AI. Finally, it was found that continuous advancements in XAI are needed to allow its seamless integration in the AI enabled healthcare systems and increase adoption and usefulness.

XAI is used by T. Shahzad et al. [10] to explain a decision made by AI, when diagnosed using diabetic retinopathy using retinal images dataset. For the convergence history plots of the model, it is validated that the model can achieve 94% accuracy using LIME technique of XAI. XAI integration in the deep learning algorithm for the diagnosis of diabetic retinopathy facilitates for clinicians and stakeholders to access to the reasoning account behind algorithmic decisions built their trust by the use of AI to support their patient's care.

The research team led by S. Shridevi [11] developed an XAI-integrated machine learning framework to forecast and study Neck direction in head impact events through muscle force evaluation with suitable explanations for proper decision-making. The implementation of six Machine learning algorithms includes Logistic Regression among others while XGBoost presents the highest accuracy at 98.6% in the model framework. The XAI technique called LIME applied to XGBoost models provides explainable features to healthcare workers who require transparent AI predictions for optimal clinical choices.

The research of S. Ahmed et al. presented a Logistic Regression model for Diabetes predictions that utilized XAI methods [12]. Both LIME and SHAP techniques boost explainability through an accuracy rate of 86% in the model. The explainability of models increases through LIME because it delivers local explanations for individual instances which provides deeper understanding of model operations. SHAP leverages Shapley values to deliver explanations that remain consistent while being applicable on all datasets and produce stable interpretations for enhanced model interpretability. When both approaches are merged it generates complete understanding of model operation which leads to better healthcare outcomes while encouraging user trust in AI-based diabetes prediction.

P. A. Moreno-Sánchez et al. [13] designed an explainable machine learning model which predicted Chronic Kidney Disease from data within the CKD UCI-ML repository. The research aimed to demonstrate XAI capabilities for improving predictive models in medicine by increasing their accuracy and interpretability standards. XGBoost technique proved suitable because it delivers high accuracy along with explainability while maintaining important features for healthcare models requiring interpretability. This model secures 99.2% training accuracy together with 97.5% accuracy on new data that outperforms alternative classifiers. XAI methods confirm that hemoglobin stands as the main risk factor for Chronic Kidney Disease based on their evaluation of key features. This approach uses XAI

while helping healthcare providers trust AI diagnostic systems which deliver comprehensive information about early detection processes.

G. V. Aiosa et al. [14] created an XAI-based advanced clinical decision support system (CDSS) for forecasting obesity-related co morbidities risk factors and detect non-direct links between these diseases and non-communicable conditions. The predictive analysis employed different kinds of ML algorithms consisting of Multi-Layer Perceptron (MLP), Extreme Gradient Boosting (XGB), Logistic regression (LR), Nearest Neighbors (NN), Random Forest (RF), Decision Tree (DT), and Linear Support Vector Machine (LSV). SHAP plots within XAI systems helped explain predictions from the most effective models by demonstrating how features contribute to individual results in specific cases along with overall SHAP bee swarm plot distributions. The XAI-CDSS included a graphical user interface which allowed medical staff to see how obesity affects patient health throughout time while linking it to other potential medical conditions. The CDSS obtained greater transparency when XAI worked together with SHAP values to demonstrate model predictions and service explanations for healthcare experts who relied on feature contributions for building trust.

R. Kumar et al. [15] proposed a machine learning framework in integration with XAI techniques for predicting chronic pediatric respiratory diseases. Numerous machine learning algorithms processed extensive datasets of clinical variables to discover patterns and co-relationships within the information which human professionals may not foresee easily. Four XAI methods were used on Random Forest that performed the best to identify key features like ICU transfer, kaliemia, creatinine levels, cyanosis, and natremia. This approach shows potential for improving clinical diagnosis and decision-making across various diseases through ML and AI.

M. Radhakrishnan et al. [16] developed an XAI-integrated Deep Learning model for Ovarian Cancer Classification using 500 histopathological images 100 for 5 subtypes of Ovarian Cancer. Dataset augmentation was integrated and the examination revealed InceptionV3 as the most accurate DL model among MobileNetV2, VGG19, ResNet18, ResNeXt, Xception and Efficient Net by reaching 97.96% accuracy. XAI techniques, including grad-cam, saliency map, integrated gradient, and Deep Lift, were integrated to enhance interpretability and trustworthiness. These methods provided visual explanations of the model's decisions, highlighting key image regions influencing predictions, thereby increasing transparency, identifying biases or errors, and supporting clinician trust and decision-making.

Based on dataset in UCI ML repository, M. Azad et al. [17] built an XAI-driven ML framework for the accurate obesity estimation and insight into the factor on most influencing obesity. Utilities of the stacking ensemble technique with final estimator

itself being SGD classifier with a 98.82% accuracy were used in the model. The XAI method LIME was implemented to advance model interpretability and reliability through its widespread usage as it imparts model interpretability by finding the contribution of each feature in individual predictions thus providing explanation of obesity classification and revealing important factors that shape the decisions and improve trust in the reliability of the model.

Shiva Prasad Koyyada et al. [18] developed a methodology through an explainable artificial intelligence model that detects local indicators and lung diseases in X-ray images. They used the COVID-19 radiography dataset from Kaggle that contains 2396 images belonging to COVID-19 class and 1341 images from Normal class. The Custom CNN model received training through 3437 images which included 2396 COVID images along with 1041 Normal images from the available dataset. The remaining images were used for testing the Custom CNN model. The result of the CNN model utilised LIME to become more explainable.

It would work, model would be used in helping identify what the preferred features are and then be masked and shown. By using these local discriminant features along with normal images another CNN model was trained which increases the accuracy and made the model more understandable. The testing Accuracy of the final CNN model was 99.97 and training Accuracy was 99.63.

Sagheer Abbas et al. [19] developed an efficient method for predicting eye diseases through Explainable Artificial Intelligence. They selected Ocular Disease Recognition Dataset from Kaggle for their case study because it features left and right eye fundus images for 5000 patients. The information contained in this dataset was divided into sections amounting to three. The training stage included 60% of images while testing was done with 20% of images and validation utilized another 20% of the dataset images. The model creation employed Efficient Net as a pre-trained model together with LIME as explainable artificial intelligence to achieve accurate results which humans can understand and reproduce. LIME selected the crucial image areas to display for better human understanding. During model training the Machine Learning Model achieved 0.9996 accuracy while the validation scores reached 0.9574.

Using machine learning and explainable artificial intelligence, to diagnose aplastic anemia from iron deficienct anemia, B. S. Dhruva Darshan, et. al. [20] suggested a method to analyze Blood. The dataset of AA and IDA dataset was used for the study and the researchers obtained blood test attributes from the Kasturba Medical College, Manipal Academy of Higher Education based in India. The research dataset consisted of 500 samples among which 266 were associated with IDA and 234 were associated with AA. Before starting machine learning development the research team divided their data into 80% train data and 20% test data distribution. The investigators applied the first stack deployment that involved

combining Logistic Regression with K Nearest Neighbors alongside Decision Trees and Random Forest to create a better classifier. The second stacking model integrated Adaptive Boost, Extreme Gradient Boosting, Light Gradient Boosting, Categorical Boosting among others, as integrated algorithms. With the first stack and the second stack, there was created the final ensemble stack. The alternative baseline models performed no better than tree-based approaches when handling the dataset because the data needed nonlinear methods for processing. SHAP and LIME, along with ELi5, Anchor and Q Lattice make up the Explainable Artificial Intelligence framework which tries to make the decision-making process transparent.

A methodology for discovering disease biomarkers for Ovarian Cancer through explainable methods of the PLCO Ovarian Biomarkers dataset was proposed by Weitong Huang et al. [21]. The clinical trials recognize this dataset that contains 113 case samples alongside 894 non-case specimens after eliminating null value entries. The dataset provides multiple domain-specific metadata together with different protocol types and variable descriptions which enables clear and sustained examination of the modelling process. The research paper applied random search optimization techniques along with 10-fold cross validation for optimizing decision trees as well as random forest models while implementing logistic regression as another modeling approach. The classification capabilities of the model were assessed through UC-ROC score evaluation alongside Shapley Additive explanations which generated explanations at local and global levels. Random forest model achieved the highest results based on the AUC-ROC Score measurements.

Zubaira Naz et al. [22] discusses about a method that can explain various lung pulmonary disease type classification results. The datasets to which this information applies belong to two: COVIDCT and COVID Net. The COVID-CT dataset contains a complete collection of 349 positive COVID 19 CT scans from 216 patients and 397 negative COVID 19 CT scans from 397 patients. Transfer learning models using CNN and the ResNet50 are last pretrained structure working on Image Net data, but their classification process of input images take place. Among the various available pretrained models, the ResNet 50 pretrained model achieved a better performance rate. The interpretable model LIME provided explanations showing which image characteristics mostly affected results. The designed models achieved 93% accuracy in detecting COVID CT images while obtaining 97% accuracy with COVID Net images.

The work of Mohammed Saidul Islam et al. [23] researched and developed a methodology about the XAI Model for Stroke Prediction using EEG signal. In one case study, the research used Biopac MP 160 Module to record EEG data from two or four channels in stroke patients and healthy adults. The research included 75 healthy adults with a mean age of 77 years and a representation of 31% males

whereas 48 stroke patients had a median age of 72.2±5.6 years with 62% male participants. EEG data was obtained for the patients that had stroke and that apparatus themselves during 3 months after their confirming that they had ischemic stroke, in the active state. The dataset divided the components into two parts out of which 80 were used as training model and 20 were trained as the testing model. The ML models used with LIME (Local Interpretable Model-agnostic Explanation) are Adaptive Gradient Boosting, XGBoost and Light GBM to explain the classification result by feature weights distribution to show feature importance for the classification result. The accuracy achieved by Adaptive Gradient Boosting, XGBoost, Light GBM amounted to 0.80, 0.77, 0.78 respectively.

In fact, Gangani Dharmarathne et al. [24] developed an explainable artificial intelligence approach to integrate machine learning for chronic kidney disease diagnosis. For the purpose of this experiment, the researchers studied using the data available on 400 individuals and 25 different attributes present on the UCI repository. The limitation of the dataset is small and thus it may result to repeated features. 70% of the data cases were partitioned into training and testing sections for testing the model performance where 30% of the data cases were reserved for testing the model performance and the rest of 70% was used for training the model. The Machine Learning models included K-Nearest Neighbors, Decision Tree, Random Forest, XGB, Artificial Neural Network together with explainable artificial intelligence methods Shapley Additive Explanations (SHAP) and Partial Dependency Plots (PDP) which were used in this study. The K-Nearest Neighbor decision tree together with Support Vector Classifier yielded accuracy results of 0.975, 0.975, 0.975, 0.975, and 0.975 from Random Forest, XGB as well as Artificial Neural Network.

At Eram Mahamud et al. [25], different explainable artificial intelligence models using fine tuned transfer learning were developed for classifying multiple lung diseases in chest X ray images with the choices more interpretable. For their study, the Lungs Disease dataset became the basis with the three sections according model evaluation (training - 80%, validation - 10%, and testing - remaining amount). Testing was done on 2027 images, validating with 2015 images and a total of 20,012 augmented training examples were used in the examination. The proposed CNN model reached 0.99 accuracy while the highest performance of 0.94 was obtained by using the pre trained EfficientNetBo model. Beyond heatmaps (Grad-Cam, Grad-CAM++), SHAP and LIME generated feature saliency maps in order to help locate the important features responsible for model outcome analysis. During phase models all were able to test at 0.967.

Alberto Ramírez-Mena et al. [26] conducted research on explainable artificial intelligence which involved using gene expression to identify and forecast prostate cancer tissue. The research team selected data from the TCGA-PRAD section of the

TCGA consortium which could be accessed through the GDC Legacy Archive at the National Cancer Institute (NIH). The dataset included information from 550 prostate patients, 9 tumoral samples and 52 controls who were at International Society of Urological Pathology grade levels between 1 to 5. The researchers selected KNN as well as rpart (classification and regression Trees-CART) and Random Forest models for their case study analysis. The Random Forest model took the top position in all G-mean, F1, AUC, Specificity measurements while exhibiting superior Sensitivity performance according to Random Forest's APV values. Through the implementation of SHAP researchers could visualize the significance of the 20 most important predictors in the algorithm. Each dot within the representation displayed the contribution value of specific genes to the final prediction made by the classifier.

The Random Forest combined with XGBoost algorithms adopted by Rathore et al. [27] produced interpretability and causability of decisions using SHAP values on standard datasets. The Random Forest model achieved 98.21% accuracy in its predictive classification tasks. The introduction of explainability made results more transparent and helped reveal the fundamental disease cause in studied subjects. The study recommends a quantitative analysis to help healthcare adopt artificial intelligence while dealing with ethical issues in diagnosis through transparency and causability and interpretability.

Reddy Soora et al. [28] made their research by putting together image quality while identifying spatial relationship between features to create the modified capsules networks for the segmentation tasks. To solve this task, the proposed loss function used the active contours model as an integration of external forces with regional information and used the functions that are used in segmentation extraction, and let curves evolve by continuously minimizing the energy. An analysis of Dice-score and mean-IoU metrics was conducted on evaluation of a performance comparison on a brain tumor segmentation dataset by the authors.

In the research of Khera P et al. [29], EEG and surface EMG signals were used to create a model hierarchy (LRG-2L-LSTM) to recognize lower limb ankle movement using brain signal for muscular activity estimation of 4 EMG channels and 12 EEG channels. Our proposed model trained to an $R = 0.742 \pm 0.03$ and RMSE of 0.067 ± 0.002 for estimating EMG so as to obtain an average accuracy of recognition of $84.86 \pm 0.27\%$ of estimated EMG for ankle movements to develop a control system for lower limb prostheses and exoskeletons for those amputees with minimal muscular strength.

The figure 2 adds depth to your discussion about bridging the gap between developers and clinicians, and supports your review's recommendations and future work.

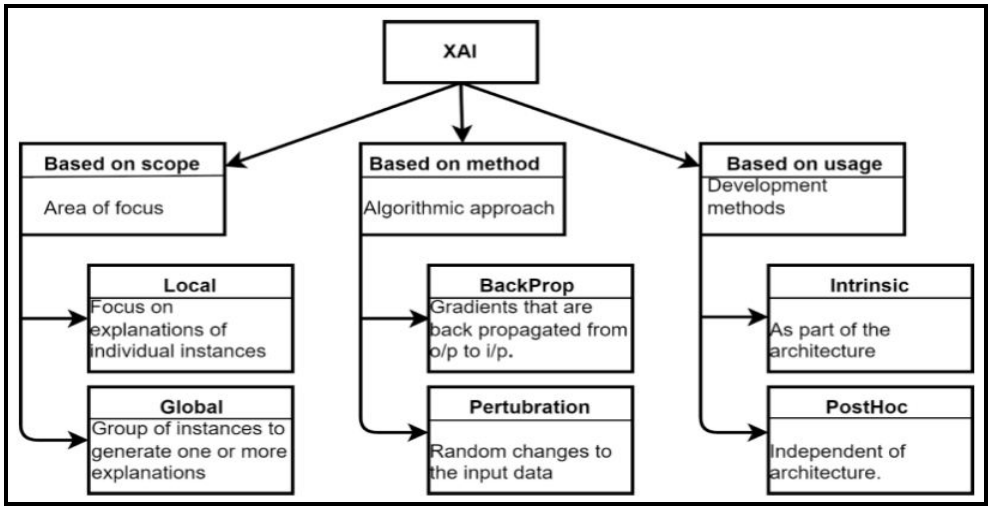

Figure 2: Mental Models, Recommendations, and Quality Criteria for XAI (ML Developers vs Clinicians, Recommendations, Key Concepts Like Transparency, Privacy, Fairness)

In this work, Pradhan et al. [30] developed their Block chain and AI enabled COVID-19 vaccine tracking system by using the Inter Planetary File System (IPFS) as decentralized storage and Truffle and Ganache Tool to combine with it to create its use within the Ethereum Virtual Machine (EVM). Finally, the proposed framework is tested using Keccack 256 transaction hash along with the number and the constraint (contract gas consumption metric). Performance measurements involving framework throughput in addition to the blockchain framework latency and memory utilization and CPU use and traffic counts both in and out were recorded by the team. The analysis task needed an artificial neural network (ANN) within the vaccination group for classification.

Methodology

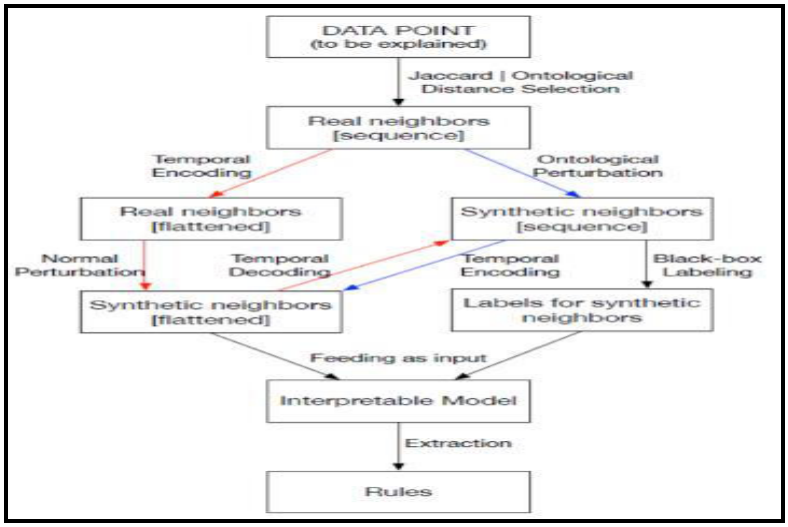

Figure 3: Data Point Explanation Framework – Real & Synthetic Neighbors, Interpretability Pipeline

This review paper follows a structured process for identifying and evaluating research on the use of XAI in healthcare. This methodology comprises of three core stages: (1) comprehensive literature search, (2) rigorous study selection based on inclusion criteria, and (3) structured data extraction and qualitative analysis.

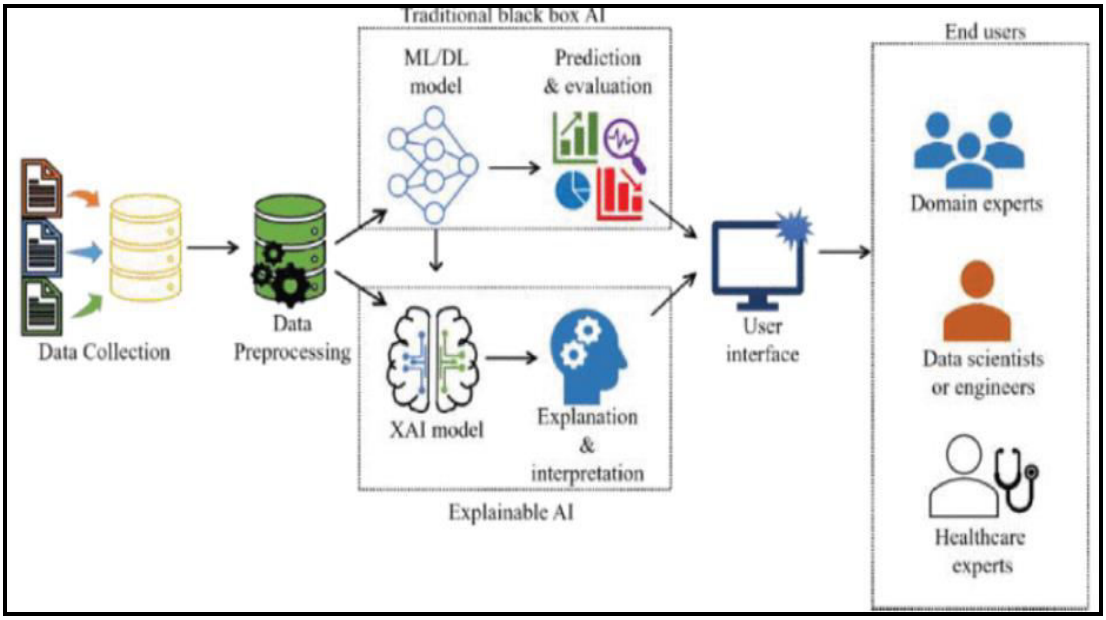

Figure 4: Taxonomy/Classification of XAI Methods (Scope, Method, Usage)

Figure 3 represents a flowchart that is ideal for illustrating how data points are selected, processed, and transformed for explanation purposes. Insert it when describing technical methodologies, XAI algorithm pipelines, or interpretability workflows used in surveyed studies.

Figure 4 visually categorizes how XAI methods are classified, supporting your survey's method-oriented review and evaluation of various approaches.

A. Review of Existing Studies:

A meticulous analysis was done for this review paper across multiple academic databases including IETE Journal, IEEE, Springer, Science Direct, Nature, and NCBI. For the search of these papers the following terms were mostly used:

- ("Explainable AI" OR "XAI") AND ("healthcare" OR "medical diagnosis" OR "clinical decision support" OR "Medical Tests" OR "Diseases")
- ("Interpretable Artificial Intelligence" AND "machine learning" AND "deep learning" in Healthcare)
- ("SHAP" OR "LIME" OR "Grad-Cam" OR "Grad-Cam++" OR "PCP graphs") Model-specific methods are tailored for a particular class of models (e.g., Grad-CAM works only for CNNs).Model-agnostic methods like SHAP or LIME can explain the predictions of any machine learning model by treating it as a black box.

B. Study Selection:

The selection process for the research papers involved two steps:

1. Title and Abstract Screening Phase:

The retrieved articles were reviewed based on their titles and content in abstract, and only those relevant to XAI in Healthcare were included, while the rest were discarded.

2. The Full Text Review Phase:

In this second phase the whole article was thoroughly read and analyzed. The articles which were relevant and provided substantial insights in Explainable Artificial Intelligence in Healthcare were included only. The factors that were used for the selection of articles were:

- Included of XAI techniques application in the field of Healthcare.
- Articles which discussed interpretability methods for machine learning models in medicine or healthcare.
- Articles focusing on the regulatory and ethical aspect of Explainable Artificial Intelligence in Healthcare.

Figure 5 clearly illustrates the overall workflow, from data collection and preprocessing through traditional "black box" AI and towards explainable AI, mapping to your discussion of methodologies, model pipelines, and stakeholders.

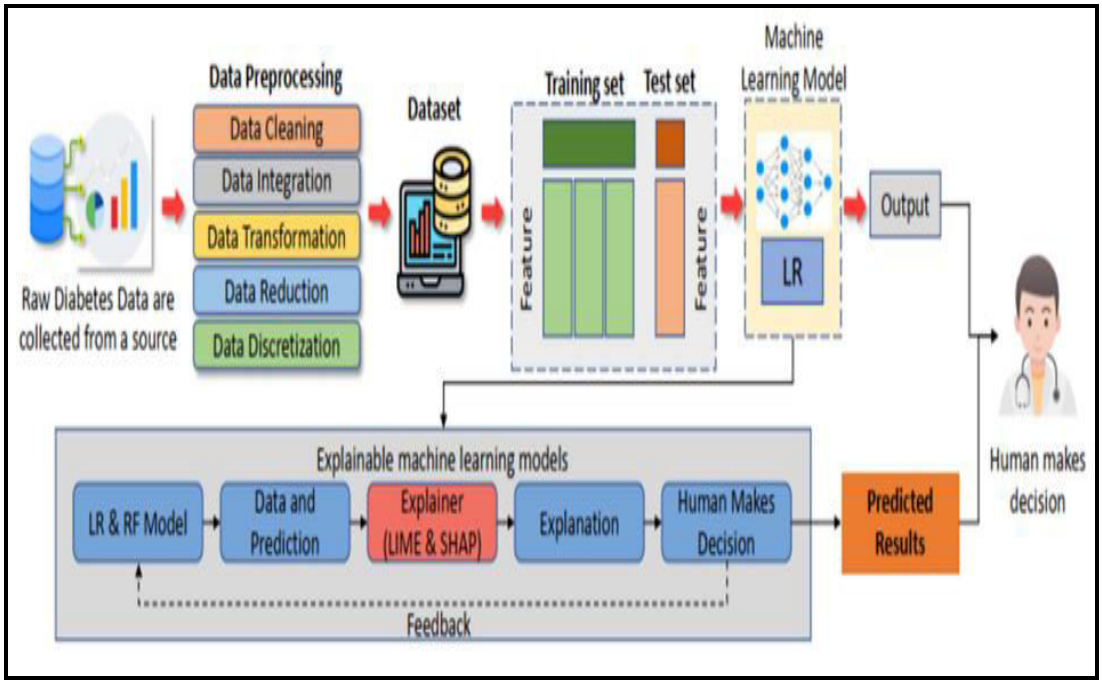

Figure 5: End-to-End XAI Workflows in Healthcare (Raw Data to Decision, Traditional vs Explainable AI, Healthcare Pipeline)

C. Data/Content Extraction:

From these research articles only the necessary content and data were taken or extracted, this extracted data included:

- The Explainable Artificial Techniques used or implemented and how it made the machine learning model more interpretable.
- The healthcare domain in which it was implemented for eg. radiology, cardiology etc.
- The dataset that was used for the machine learning models' training and testing.
- The performance as well as the evaluation of the machine learning models.

Conclusion

This Explainable Artificial Intelligence (XAI) is transforming healthcare by prioritizing interpretability, transparency, and trust in AI-driven clinical decision-making. The integration of XAI into machine learning pipelines enhances not only the accuracy of diagnostic models but also their transparency, making complex algorithms more comprehensible for clinicians, patients, and stakeholders. With methods such as LIME, SHAP, and visual tools like heatmaps and saliency maps, XAI enables the breakdown of opaque "black-box" predictions from classifiers like SVM, CNN, and XGBoost into actionable and understandable insights.

Figure 6 succinctly highlights the essential attributes that define high-quality, trustworthy, and user-centered XAI systems. It emphasizes what should be prioritized for safe and effective adoption in clinical environments and can reinforce your closing arguments.

In clinical practice, explainable models support safer, user-centered decisions by providing clear rationales for diagnoses, prognoses, and treatment recommendations. This interpretability is critical for regulatory compliance and ethical accountability, helping healthcare professionals assure that AI-guided interventions align with medical standards and patient values. User-oriented interfaces further facilitate smooth adoption and meaningful interactions with XAI systems, increasing the likelihood of successful implementation in real-world healthcare environments.

By revealing the key features driving predictions, XAI methods improve model reliability and foster deeper collaboration between humans and AI systems. Comprehensive XAI frameworks that combine multiple interpretability techniques allow healthcare programs to deliver more equitable outcomes, offering transparent justifications for resource allocation and clinical actions. Moreover, XAI-based solutions empower medical staff to validate and verify model outputs, reducing risks associated with erroneous or biased decisions.

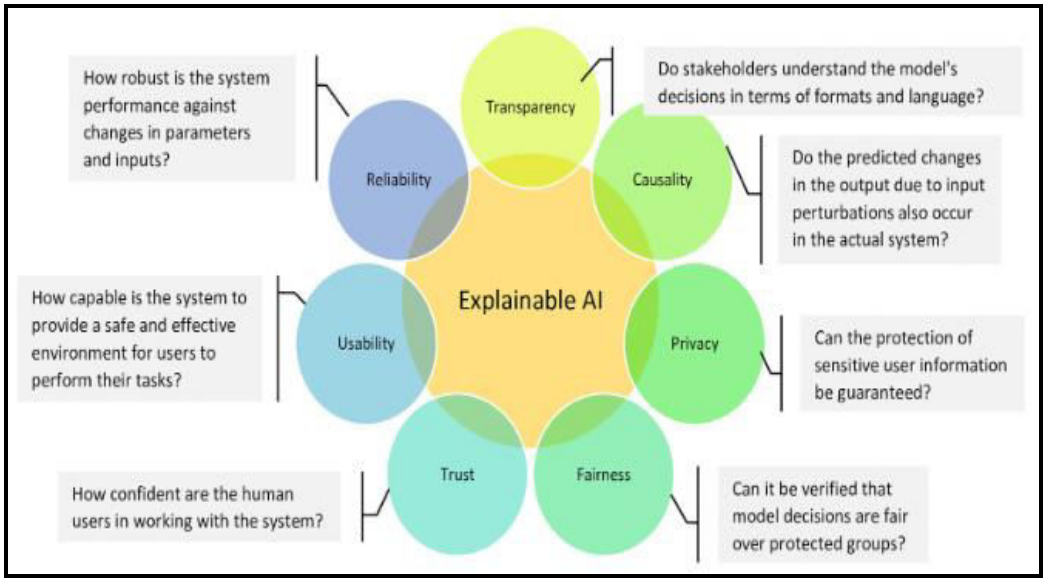

Figure 6: Key Criteria for Explainable AI – Transparency, Causality, Privacy, Reliability, Usability, Trust, Fairness

Looking ahead, the continued advancement of XAI will be essential for meeting emerging challenges in healthcare, such as addressing disparities, managing complex medical data, and supporting personalized medicine. Future research should focus on broadening the scope of XAI capabilities, integrating human-centric explanation mechanisms, and promoting cooperative systems where clinicians and AI tools work together seamlessly. These efforts will be paramount for increasing trust, ensuring safety, and driving effective adoption of AI across diverse medical domains.

References

1. J. Amann, A. Blasimme, E. Vayena et al., "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," BMC Medical Informatics and Decision Making, vol. 20, p. 310, 2020.
2. N. Bienefeld, J. M. Boss, R. Lüthy et al., "Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals," npj Digital Medicine, vol. 6, p. 94, 2023.
3. C. Metta, A. Beretta, R. Pellungrini, S. Rinzivillo, and F. Giannotti, "Towards transparent healthcare: Advancing local explanation methods in explainable artificial intelligence," Bioengineering, vol. 11, no. 4, p. 369, 2024.
4. F. V. Farahani, K. Fiok, B. Lahijanian, W. Karwowski, and P. K. Douglas, "Explainable AI: A review of applications to neuroimaging data," Frontiers in Neuroscience, vol. 16, 2022.
5. Z. Papanastasopoulos, R. K. Samala, H. P. Chan et al., "Explainable AI for medical imaging: Deep-learning CNN ensemble for classification of estrogen

receptor status from breast MRI," in Medical Imaging 2020: Computer-Aided Diagnosis, SPIE, 2020.

6. A. Boutorh, H. Rahim, and Y. Bendoumia, "Explainable AI models for COVID-19 diagnosis using CT-scan images and clinical data," in Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB 2021), LNCS, vol. 13483, Springer, 2022, pp. 210–223.
7. C. Wang, Y. Liu, F. Wang et al., "Towards reliable and explainable AI model for solid pulmonary nodule diagnosis," arXiv preprint, arXiv:2204.04219, 2022.
8. S. S. Band, A. Yarahmadi, C.-C. Hsu et al., "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," Informatics in Medicine Unlocked, vol. 40, p. 101286, 2023.
9. U. Pawar, D. O'Shea, S. Rea, and R. O'Reilly, "Explainable AI in healthcare," in Proc. International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), Dublin, Ireland, 2020, pp. 1–2.
10. T. Shahzad, M. Saleem, M. S. Farooq et al., "Developing a transparent diagnosis model for diabetic retinopathy using explainable AI," IEEE Access, vol. 12, pp. 149700–149709, 2024.
11. S. Shridevi and S. Elias, "Explainable AI based neck direction prediction and analysis during head impacts," IEEE Access, vol. 12, pp. 31399–31408, 2024.
12. S. Ahmed, M. S. Kaiser, M. S. Hossain, and K. Andersson, "A comparative analysis of LIME and SHAP interpreters with explainable ML-based diabetes predictions," IEEE Access, 2024.
13. P. A. Moreno-Sánchez, "Data-driven early diagnosis of chronic kidney disease: Development and evaluation of an explainable AI model," IEEE Access, vol. 11, pp. 38359–38369, 2023.
14. G. V. Aiosa, M. Palesi, and F. Sapuppo, "Explainable AI for decision support to obesity comorbidities diagnosis," IEEE Access, vol. 11, pp. 107767–107782, 2023.
15. R. Kumar et al., "Using explainable machine learning methods to predict the survivability rate of pediatric respiratory diseases," IEEE Access, vol. 12, pp. 189515–189534, 2024.
16. M. Radhakrishnan, N. Sampathila, H. Muralikrishna, and K. S. Swathi, "Advancing ovarian cancer diagnosis through deep learning and explainable AI: A multiclassification approach," IEEE Access, vol. 12, pp. 116968–116986, 2024.
17. M. Azad, M. F. K. Khan, and S. A. El-Ghany, "XAI-enhanced machine learning for obesity risk classification: A stacking approach with LIME explanations," IEEE Access, vol. 13, pp. 13847–13865, 2025.
18. S. P. Koyyada and T. P. Singh, "An explainable artificial intelligence model for identifying local indicators and detecting lung disease from chest X-ray images," Healthcare Analytics, vol. 4, p. 100206, 2023.

19. S. Abbas, A. Qaisar, M. S. Farooq et al., "Smart vision transparency: Efficient ocular disease prediction model using explainable artificial intelligence," Sensors, vol. 24, no. 20, 2023.
20. B. S. D. Darshan, N. Sampathila, and G. M. Bairy, "Differential diagnosis of iron deficiency anemia from aplastic anemia using machine learning and explainable artificial intelligence," Scientific Reports, vol. 15, p. 505, 2025.
21. W. Huang, H. Suominen, T. Liu et al., "Explainable discovery of disease biomarkers: The case of ovarian cancer," Journal of Biomedical Informatics, vol. 141, p. 104365, 2025.
22. Z. Naz, M. U. G. Khan, T. Saba et al., "An explainable AI-enabled framework for interpreting pulmonary diseases from chest radiographs," Cancers, vol. 15, no. 1, p. 314, 2023.
23. M. S. Islam, I. Hussain, M. M. Rahman et al., "Explainable artificial intelligence model for stroke prediction using EEG signals," Sensors, vol. 22, no. 24, p. 9859, 2022.
24. G. Dharmarathne, M. Bogahawaththa, M. McAfee et al., "Diagnosis of chronic kidney disease using a machine learning-based interface with explainable artificial intelligence," Intelligent Systems with Applications, vol. 22, p. 200397, 2024.
25. E. Mahamud, N. Fahad, M. Assaduzzaman et al., "An explainable artificial intelligence model for multiple lung disease classification from chest X-ray images," Decision Analytics Journal, vol. 12, p. 100499, 2024.
26. A. Ramírez-Mena, E. Andrés-León, M. J. Alvarez-Cubero et al., "Explainable artificial intelligence to predict prostate cancer tissue by gene expression," Computer Methods and Programs in Biomedicine, vol. 240, p. 107719, 2023.
27. A. S. Rathore, S. K. Arjaria, M. Gupta et al., "Erythemato-squamous diseases prediction and interpretation using explainable AI," IETE Journal of Research, vol. 70, no. 1, pp. 405–424, 2022.
28. N. R. Reddy Soora, E. U. Rahman Mohammed, S. W. Mohammed, and N. C. S. Kumar, "Deep active contour-based capsule network for medical image segmentation," IETE Journal of Research, vol. 69, no. 12, pp. 8770–8780, 2022.
29. P. Khera, T. Das, N. M. Kakoty, and N. Kumar, "AI-enabled hybrid model for lower-limb movement recognition using cortical brain signals," IETE Journal of Research, vol. 70, no. 5, pp. 5270–5279, 2023.
30. N. R. Pradhan, R. Mahule, P. K. Wamuyu et al., "A blockchain and AI-based vaccination tracking framework for COVID-19 epidemics," IETE Journal of Research, vol. 69, no. 11, pp. 7803–7815, 2022.