Al-Nahrain Journal of Science
ANJS, Vol.28 (4), December, 2025, pp. 188-206

Explainable AI in the Medical Field: A Survey on Machine Learning Interpretability and Use Cases

Mahmood Thamer*, Zainab N. Sultani

Computer Science Department, College of Science, Al-Nahrain University, Baghdad, Iraq

Article's Information

Received: 18.12.2024
Accepted: 12.11.2025
Published: 15.12.2025

Keywords:
XAI
Machine Learning
Black-box
XAI methods
Metrics
Survey

Abstract

Explainable Artificial Intelligence (XAI) is a branch of Artificial Intelligence (AI) that focuses on developing tools, methodologies, and algorithms capable of delivering interpretable, intuitively understandable insights and rationales for human users. The growing need for XAI in medicine and other fields stems from the increasing demand for equitable and ethical decision-making. It has been established that AI systems largely depends on historical data, therefore, any existing bias or behavior will be perpetuated. As such, deep examination and interpretation are required in this process. Since these black-box models lack transparency and interpretability, several XAI models are developed for the respective domains, ranging from healthcare, military, energy, finance, and industry.The highly sensitive areas, such as healthcare, require knowing the underlying principles of model predictions. Emphasizing feature importance, XAI has improved machine learning models to identify the most critical variables that help improve accuracy and efficiency. With the use of appropriate XAI techniques, actionable insights can be derived that would support informed decisions. In the healthcare sector, the primary objective of XAI is to provide clinicians with tools to effectively evaluate AI-generated data for better patient care. This survey covers Explainable AI, its methods, and their applications in the medical domain.

http://doi.org/10.22401/ANJS.28.4.15
*Corresponding author: mahmood.thamer@nahrainuniv.edu.iq

1. Introduction

The field of artificial intelligence (AI) has experienced massive growth in recent years. It has advanced to the point that it can now outperform people in open competitions. A wide range of decision-making issues in the fields of criminal justice, finance, healthcare, engineering and agriculture are being addressed using machine learning (ML) approaches. However, these black-box AI algorithms make predictions without providing explainable insights, and hence, they raise reliability as well as credibility concerns because of their lack of transparency. This is a major concern that undermines trust in AI systems. Over the past several years, explainability and the associated ideas of interpretability, precision and transparency have emerged as obvious concerns for machine learning in the medical field Human health can be affected by judgments made in machine learning systems without understanding the rationale, it is therefore important to comprehend how these decisions are made. For instance, a disease diagnosis can lead to life-altering actions and outcomes. However, In precision medicine, specialists need much more detailed data rather than an easy dual prediction to validate their diagnosis. Users often have little to no understanding of how predictions are generated, rendering these systems metaphorical "black boxes". Figure 1, highlights the limitations of AI systems when explainability is absent. It shows a process where a learning algorithm trains an AI model, which then generates a prediction. However, these predictions are presented to end-users without any accompanying explanations. As a result, users are left with critical unanswered questions, such as why a particular prediction was made, why alternative predictions were not considered, whether there are other possible alternatives are there when the model succeeds or fails. When it can be trusted? and how errors can be corrected? This lack of transparency creates uncertainty and hinders reliance in the AI systems and modules. Relying solely on Machine learning

(ML) models has caused several tragic failures over the past few years, including fatal incidents from self-driving automobiles to crime-fighting robotic, and a child colliding. Alexa displays unedited audio rather than a child's music Additionally Amazon's facial identification algorithm mistakenly matched 28 persons in Congress with their criminal records. Furthermore, sophisticated AI-powered stock market trading softwares that triggered a trillion-dollar flash break. Ribeiro et-al, recent work introduced a model capable of distinguishing between different predictive behaviors and providing insights into the decision-making process, however when analyzing pictures of wolves and huskies, the model was only misclassified once. However, the algorithm took into consideration the presence of ice to deduce whether an image was a husky or a wolf, clearly pointing weakness in the way the model made a decision. Data-driven technology decisions in crucial areas, like parole hearings, are viewed with suspicion (for example, in 2016, convict Glen Rodrigues was refused freedom because he had an extremely risky COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) rating that was found to be incorrect. The use of "dirty data" in policy-making systems has resulted in distorted forecasts, as several case studies have shown in figure 14. It is therefore imperative to clarify black-box models and improve interpretation and transparency to offer reliable and accurate machine-learning outputs. An XAI framework facilitates the creation of an acceptable human explanation for black-box models. Knowledge of the "black box" is crucial because it helps physicians comprehend the internal workings of machine learning algorithms. Figure 2, shows how XAI improves interaction between end-users and AI by providing explanations for predictions. It uses surrogate models for black-box systems to clarify decisions, enabling users to understand why predictions were made, while alternatives were not, when the model succeeds or fails, and how to correct errors. This triggers trust, transparency, and usability in AI systems. Black box models should be explained either locally (focusing on a single prediction) or globally (addressing the entire model). Each approach may contribute to training future surrogate models. Surrogate models are understandable models that have been taught to resemble the predicted outcomes gained from black-box models. Several XAI approaches have been presented with each method taking a different approach and hence produces different results LIME (Local Interpretable Model-agnostic Explanation) and SHAP (SHapley Additive Explanations), are two known model-agnostic. Local explanation approaches for each black box classifiers. XAI plays an important role in modernizing healthcare by increasing transparency, reliable, and therapeutic results. This technique ensures the ethical and effective integration of AI systems into healthcare, leading to improved decision-making and more consistent patient care. This review has been structured into the following sections: The second part discusses the overview of Explainable Artificial Intelligence (XAI), the third part discusses the taxonomy of interpretability, the fourth part discusses the XAI methods and the fifth part discusses the benefits of feature with XAI. Furthermore, the sixth part discusses the evaluation of the algorithms. Followed by further discussion in the seventh part of the literature review, and finally, the eighth part concludes the paper.

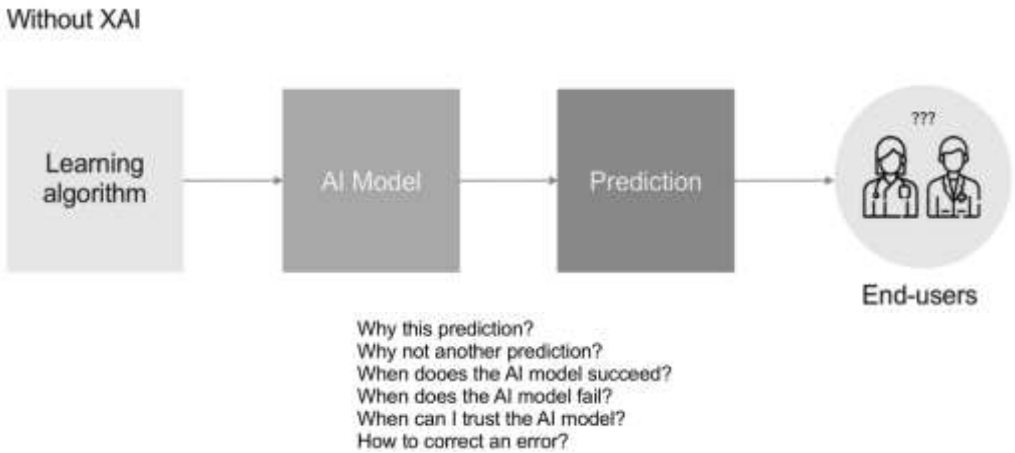


Figure 1. Challenges and limitations in communication between end-users and AI models without XAI [17].

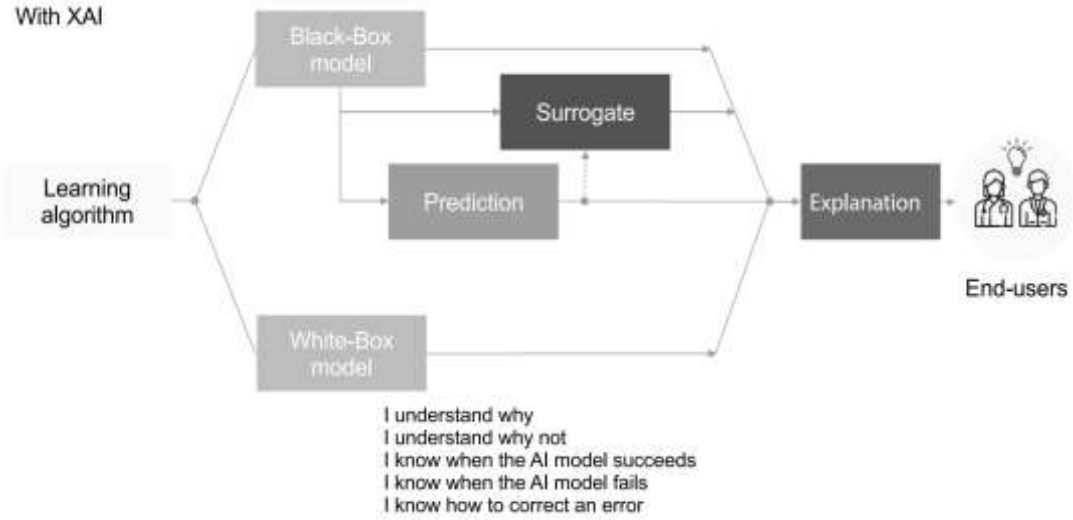


Figure 2. XAI facilitates effective communication among artificial intelligence models and end-users by enhancing transparency, trust, and understanding of AI predictions [17].

2. Overview of Explainable Artificial Intelligence (XAI)

The branch of artificial intelligence called "explainable AI" deals with converting advanced "black-box" artificial intelligence systems and models results into understandable depictions, simple and explainable Black-box models may now be transformed into glass-box models in the medical field because of the development of explainable AI (XAI). Understanding the "black box" is crucial because it helps physicians comprehend how ML models operate inside According to Gunning D."A collection of machine learning algorithms will be developed by Explainable AI to enable users as individuals must comprehend, trust, and manage the future era of artificially intelligent partners." Explainable AI-based models are more dependable and reliable and perform equivalently with professionals. Explainable AI systems aim not just to enhance task efficiency and accuracy but also to provide explanations for how the decision was reached and give insight into the model's complex logic. XAI may be defined essentially as the goal of making AI systems appear human-like, which aims to build an AI system that is comprehensible, observable, reliable, responsible, interpreted, and explained. In the following paragraph, we will clarify the definitions of these terminologies:

2.1. Explainability

Explainability in artificial intelligence is defined as the ability to communicate an AI model's internal workings or outputs in understandable language. It makes complex AI choices clear and dependable. Explainability is important in domains such as healthcare and finance, where understanding why a model made a specific decision has ramifications. Explainability promotes accountability and aids in the diagnosis and correction of model flaws. It is a principle that a machine learning model and its output can be explained in an approach that makes sense to a person on an acceptable level.

2.2. Interpretability

Interpretability provides people with explanations or intelligible interpretations of concepts simultaneously Miller [defined interpretability as "the degree to which an observer can understand the cause of a decision". Doshi-Velez and Kim, describe interpretation as "the ability to explain or provide the meaning in understandable terms to a human". A particularly similar definition is provided in [3]. In machine learning, knowledge may be described more explicitly as [29], "Machine learning techniques are used to extract meaningful understanding of domain connections from data". As defined by Biran and Cotton [30], "Systems are interpretable if their operations can be understood by a human, either through introspection or a produced explanation". Gilpin et-al. [31], stated that "Interpretability is an important aspect of explanation". A good method for evaluating interpretation is that it must "respond to its ability to communicate the trained model outputs behaviors in an actual human-understood way" [32, 33].

2.3. Transparency

One of the most significant elements of building trust is a complete grasp and familiarity with the system [34, 35]. This is where the concept of transparency for a trustworthy AI first emerges. Transparent AI strives to give detailed explanations and communication of an AI model's output [36]. Transparency can refer to a comprehensive distinguished status in providing collaborators with useful knowledge about the way the models performs, such as evidence of training practices, evaluation of data from training distribution, coding releases, and algorithm-specific simplicity about how the model works, as compared to opaqueness [5, 37]. Table 1, displays a comparison model by evaluating the degree model of transparency.

3. Taxonomy of Interpretability

Explaining approaches in addition to methodologies for machine learning interpretation may be categorized based on many characteristics [39]. Figure 3, outlines Explainable AI (XAI) approaches, divided into pre-modeling (focusing on dataset understanding and white-box models) and post-modeling (interpreting black-box models using post-hoc techniques). XAI methods can be model-specific (focused on a single model type) or model-agnostic (applicable to any model). It explains individual predictions (local logic), entire models (global logic), and properties like sensitivity, enhancing transparency, trust, and usability in AI systems. Black-box models inherently lack explainability, hence approaches such as model properties, local logic, global logic, and so on are used to make the black-box model explainable through internal logic or model output [40].

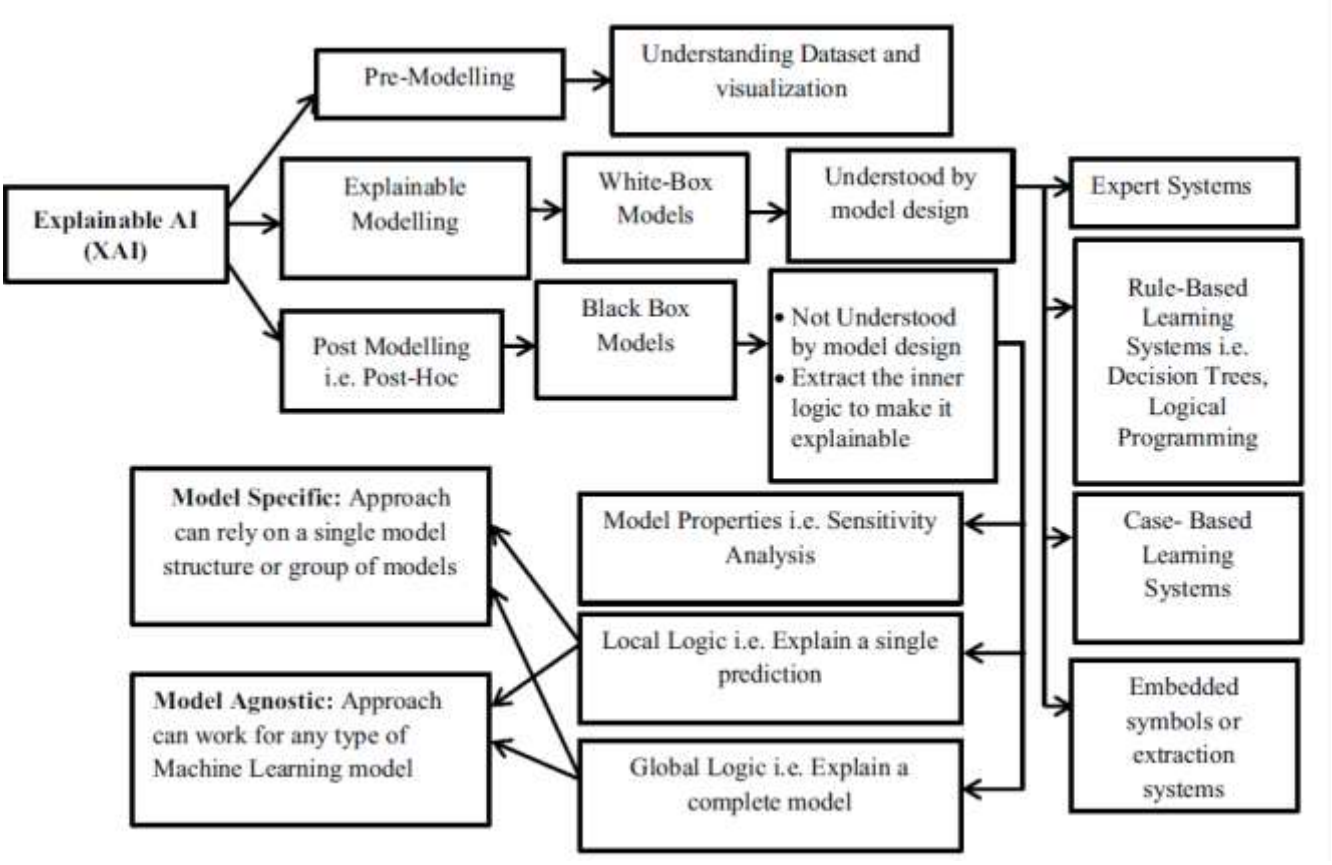


Figure 3. Categorizing XAI Approaches: Pre-Modelling and Post-Modelling Techniques for Interpreting Black-Box and White-Box Models [40].

3.1. Local explanations vs. Global explanations

Several studies in literature attempt to clarify opaque models with one of the following goals, giving a sample explanation for some particular decision utilizing a local approach [12, 41, 42, 43] or offering broad explanations that can elucidate the overall reasoning of this black box [44, 45]. Local explanations for just singular predictor are provided

to explain why the model selected a certain conclusion in that scenario [46]. Global explanations for the complete system help us know all of its model's behavior and logic, which leads to predicted consequences [46]. In Table 2, we compare local and global explanations.

3.2. Pre-Model vs. In-Model vs. Post-Model

Interpret approaches could be categorized based on when they are used: Prior (pre-model), throughout (in-model), or afterward (post-model), producing the model for machine learning [47]. Pre-model interpreting strategies were not dependent on the model. These are limited for data. Key intuitive features, such as sparsity (a minimal number of characteristics), are traits that enhance data interpretability. Pre-model interpreting is therefore strongly tied to data interpretation, which includes exploratory data analysis approaches [48]. In-model interpreting means machine learning (ML) models that are intrinsically readable (whether due to constraints or otherwise). Post-model interpretability is the process of increasing interpretability after the model was built (post hoc) [49].

3.3. Intrinsic vs. Post Hoc

The main criteria used to check if interpreting can be accomplished are limitations upon the complexity of the machine learning models. (Intrinsic) or by using post-training methods to evaluate the model's interpretability and decision-making process [49]. Intrinsic interpretation involves models that can be interpreted by themselves [49]. This in-model interpretation may be accomplished by applying constraints models, including sparsity, monotony, causation, or physical limitations related to knowledge of the domain [2]. Post-hoc interpretation is an explanatory approach utilized after the model is trained [49].

3.4. Model-Specific vs. Model-Agnostic

A further significant distinction is model-specific vs. model-agnostic. Model-specific interpretation techniques can only apply to certain model classes that exist because every method depends on the which of the specific model [49]. Model-agnostic techniques may be utilized for every machine learning model, black box or otherwise after it has been trained. By definition, these approaches are unable to see the model's internal workings, which include weight, otherwise structure details [49]; otherwise, they wouldn't be divorced from the black box models. The criteria covered in the two following sections, sections 3.1–3.3, have been linked in some way. Table 3 displays this relationship.

4.XAI METHODS

Explainable AI (XAI) approaches provide an understanding of the decisions made by machine learning models, making them more transparent and understandable to users. This section defines multiple XAI techniques and compares them using the taxonomy provided.

4.1 Local Interpretable Model-agnostic Explanation (LIME)

Local Interpretable Model-agnostic Explanations (LIME) is a popular post-hoc XAI approach that was created to explain predictions for any machine learning classifier by determining FI scores based on assumptions that might not necessarily be true across various types of classifiers [12]. This is suitable just for local interpretations. The level of explanation is difficult and is commonly employed in recognition of pattern application [12]. The local surrogate model (LIME) [12] is a model-independent technique utilized to clarify individual predictions from black-box machine learning algorithms. The main concept behind LIME is to employ an interpreted linear model to simulate the unexplained model locally. Yan et al. [50] introduced LEMNA, which has a similar underlying concept to LIME. LEMNA (Local Explanation Method using Nonlinear Approximation) refined choice boundaries by training the combined regression models and incorporated LASSO (Least Absolute Shrinkage and Selection Operator) to address the issue of feature dependency, which compensates for LIME's defect. The LIME approach facilitates the finding of blocks of pixels that affect the categorization process, making interpreting more comprehensive and informative. Figure 4, illustrates how LIME helps explain a histopathological image classification. Beginning with an original colon tissue image, and then segments it into super pixels (shown with yellow boundaries). LIME analyzes these regions to determine which parts influenced the model's prediction. In the final panel, only the key super pixels remain visible-highlighting areas most important to the model's decision, enhancing transparency, and trust in AI-driven medical diagnoses.

Table 1. Evaluation of model transparency [38].

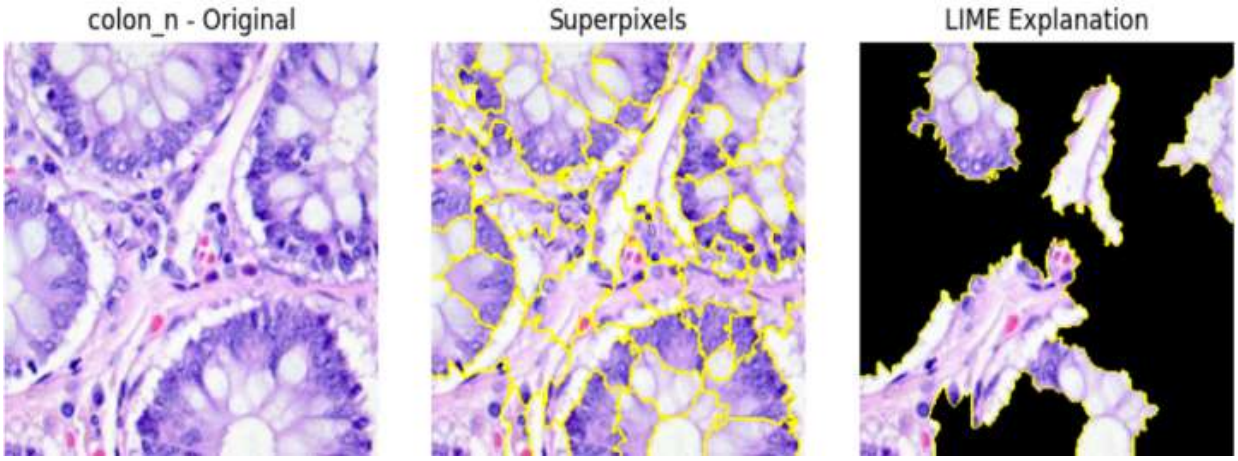
Models	Simulatability	Decomposability	Transparency Algorithmic	Post-hoc
Linear/Logistic Regression	Predictions are readable by humans, and connections among them are kept to a minimum.	A lot of connections and predictions.	The interactions and variables have become complicated for analysis without mathematics instruments.	Not required.
Decision Trees	Humans can comprehend without a proper understanding of math background.	Rules must not change data so that they are comprehensible.	Humans may grasp a Predictive model by walking the tree.	Not required.
K-Nearest Neighbors	Model difficulty matches human's naïve simulation skills.	Too several variables, yet the measure of similarities and the collection of variables may be examined.	Difficult likeness measurement; multiple variables to assess without mathematical instruments.	Not required.
Rule Based Learners	Readability variable, regulation size controllable by a person.	The scale of the regulations was too enormous to examine.	Rules have gotten so complicated that math tools are necessary.	Not required.
General Additive Models	Variable, connections also functions should be comprehensible.	Relationships are exceedingly difficult to duplication.	Variables and even interactions are too complicated for analysis without mathematics instruments.	Not required.
Bayesian Models	Statistics connections and variables must be understood by the intended audience.	Relations have a lot of variables.	Interactions and predictions are so complicated that mathematical approaches are required.	Not required.
Tree Ensembles	Not useful.	Not useful.	Not useful.	Features relevancy as well as modeling simplicity.
Support Vector Machines	Not useful.	Not useful.	Not useful.	Features relevancy as well as modeling simplicity.

Table 2. Local Explanations vs. Global Explanations.

Explanation Type	Purpose	XAI Techniques
Local Explanations	Explain particular forecasts, offering insights on feature influence in unique circumstances.	LIME, SHAP
Global Explanations	Offer insights into the overall model behavior, displaying broad trends and feature relevance all through the dataset.	Global Feature Importance, Partial Dependence Plots

Table 3. A relationship among interpretation criteria [49].

Pre-model	N.A.	N.A.
In-model	Intrinsic	Model-specific
Post-model	Post hoc	Model-agnostic



j

Figure 4. LIME-Based Visual Explanation of Histopathological Image Classification

4.2. Anchors

Is a model-independent tool that generates local interpretations of medium complexes for explanations [41]. They may be used in a range of machine-learning applications, including classification, structured prediction, and text production. Anchors provide explanations of local scope, these are if-then rules that simulate a black box within the area of a specific incident [52]. The main premise Individual predictions from any black-box model of classification can be explained by defining a rule of decision that is suitable "anchors" the predicted, therefore referred to as "anchors." [51].

4.3. Partial Dependence Plot (PDP)

It represents a global technique that considers every scenario and draws conclusions about the global relation between the features and predicted consequences [53]. It has the advantage that will understanding every kind of machine-learning issue in less time [54]. A PDP can show if the target's connection with a feature is linear, repetitive, or intricate [10]. For instance, when used with a linear regression model, partial dependence graphs at all times display a linear relationship. Table 4 shows the advantages of PDP.

4.4. Individual Conditional Expectation (ICE) .

This is a plot equivalent to PDPs for individual data examples [55]. ICE plots are more flexibility than PDP plots, which depicts how the model responds to a single example [54]. Local model-agnostic interpreters can be utilized for verifying monotony. Demands with a variety of explanation complexity levels [55]. Table 4 displays the advantages of ICE. The PDP for the average influence of a feature is a global method because it looks at the overall average rather than individual instances. Individual conditional expectation (ICE) plots are the equivalent of PDPs for individual data instances [55]. Individual conditional expectation curves are easier to understand than partial dependence charts. One line depicts the predictions for one occurrence if we change the feature of interest. Unlike partial dependence charts, ICE curves can reveal heterogeneous patterns [54].

4.5. Accumulated Local Effect (ALE)

A plot demonstrating significance for attributes in predictive machine learning [55]. Is a model-agnostic, locally interpretable models that perform analysis at the instance level, explaining predictions for specific data points, which could be utilized for explaining each machine learning issue type [57]. The plots are unbiased; therefore, they function even when characteristics are linked. Partial dependency plots are ineffective in such a scenario due to their marginalized implausible or physically impossible combinations of various feature values. ALE charts are quicker for calculation instead of PDPs, and they scale approximately O(n). Plots have their center at zero [51]. This improves their understanding since the value at every location within the ALE curve reflects the deviation from the mean prediction. The 2-dimensional ALE plot illustrates only the relationship; if both features have no relationship, the plot displays nothing [54]. Table 4 displays the advantages of ALE.

Table 4. PDP vs ICE vs ALE vs LIME [56].

Approach	Advantages	Disadvantages
PDP	Intuitive Easy applying Offers straightforward and causal explanations	Suitable for up to three features. Supposes no relationship among features. Might conceal heterogeneous effect.
ICE	Intuitive Specific show heterogeneous connections	The overcrowding might result in unreadability. Demands PDP for viewing the average. Unstable at high intervals.
ALE	Unbiased toward associated features quicker computation	Unstable at an extensive amount of intervals. Compared to more complicated. Does not have ICE charts.
LIME	Inherent interpretation Human-friendly interpretable	Not acceptable for global assumptions. Vulnerable to manipulations to conceal bias.

4.6. Shapley Additive Explanations (SHAP)

The explanation is usually time-consuming during calculation, so it may be used for both model-specific and model-agnostic procedures [58, 20]. Shapley values are another post-hoc XAI approach that was first developed in game theory to give a player's average expected marginal contribution to attaining a reward after all conceivable player combinations are taken into account [20]. The objective of SHAP is to determine the influence of each attribute on prediction [59]. Figure 5, shows how SHAP explains a block box model's prediction. Input variables (Var1–Var4) go into the model, which outputs a prediction. SHAP breaks down, this prediction by showing how each variable contributed to the final result. Starting from a base value (0.1), variables either increase (red bars) or decrease (blue bars) the prediction leading to a final probability of 0.4. This helps make model decisions more transparent and understandable [51]. Figure 6 presents a Shapley value plot, illustrating how different features impact a model's prediction (24.41). The horizontal axis shows the prediction range (14.34 to 30.34). The red section indicates features that decrease the output (PTRATIO at 15.3, LSTAT at 4.98), while the blue section highlights features that increase the output (RM at 6.575, NOX at 0.538, AGE at 65.2, RAD at 1). Arrows indicate the magnitude of each feature's impact, with longer arrows signifying greater influence. The plot reveals that PTRATIO, LSTAT, and AGE negatively impact the prediction, while RM, NOX, and RAD have a positive effect. Figure 7 shows a global SHAP explanation, illustrating the average impact of each feature on the model's output. The y-axis lists features from the Boston Housing Dataset, such as LSTAT, RM, and NOX, while the x-axis represents the SHAP values, indicating each feature's impact on the model's prediction. Positive values increase the output, and negative values decrease. The Dots represent individual data points, colored from blue (lower values) to red (higher values). The plot provides insight into the importance of each feature in predicting house prices [10]. Table 5 displays the main differences between SHAP and LIME.

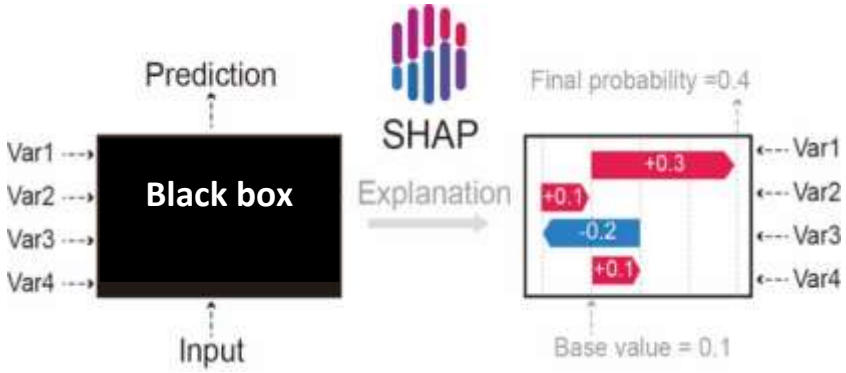


Figure 5. SHAP-Based Interpretation of Block-box Model Predictions [51].

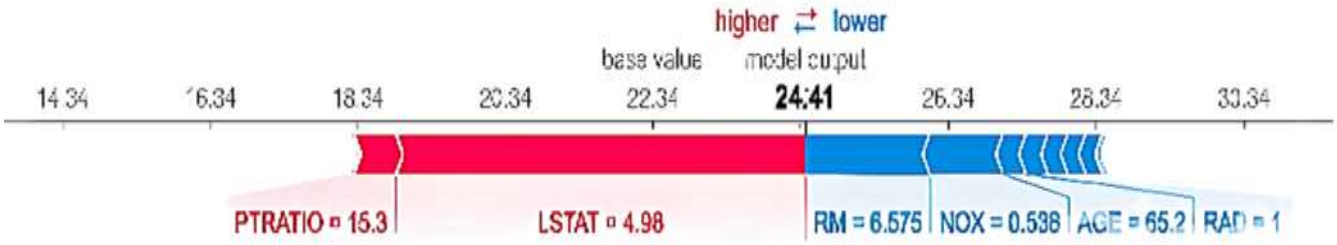


Figure 6. Shapley value plot analyzing feature Impact on model prediction [60].

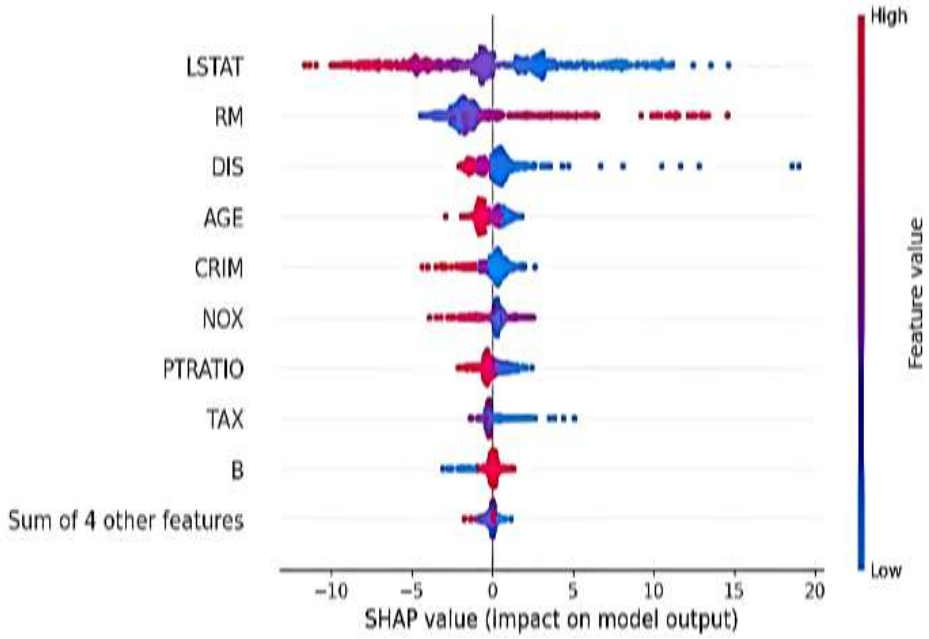


Figure 7. Global SHAP explanation visualizing feature impact on boston housing prices [60].

Table 5. Compare between LIME and SHAP [10].

Metrics	SHAP	LIME
Idea	Applied for the model as is.	Uses a local surrogate model to describe the complicated model.
Theories	Additive feature attribution using game theory.	Feature perturbation method
Type	Post-hoc model-agnostic	
Data type	Images, tabular data, and signals	
Explanation	Global, local	Local
Nonlinear decision	Depends on the used model	Incapable
Computing time	Higher	Lower
Illustration	A waterfall, beeswarm, as well as overview plots.	Just one plot.

4.6 Kernel SHAP

An integration of Shapley Value and Linear LIME [55]. The fundamental shortcoming of the Kernel SHAP explanations is that, it takes a long time for an explanation. However, it provides local visualization and may be applied to both diagnostic and model-specific techniques [58]. Table 6 displays the properties of each XAI model.

Table 6. Comparing for XAI Methods [10].

XAI Models	Consistency	Visualizations	Techniques for Interpretation	Scale for Interpretation.	Dependence Domain
LIME	Consistent	Local	Model Agnostic	Medium	Independent
Anchors	Consistent	Local	Model Agnostic	Medium	Independent
PDP	-	Global	Model Agnostic	Any	Independent
ICE	-	Local	Both	Medium	Independent
ALE	-	Local	Model Agnostic	Any	Dependent
SHAP	Consistent	Local	Both	Medium	Independent
Kernel SHAP	Consistent	Local	Both	Medium	Independent

5. Advantages of Feature Analysis in Images using XAI

Focusing on features in explainable AI increases openness, trust, and fairness by describing how models make judgments. It aids in the detection of biases, guarantees regulatory compliance, and simplifies complicated models for interpretation. This enables users to better understand model behavior, discover faults, and act on insights, resulting in more ethical and dependable AI systems. These advantages highlight the value of feature analysis using XAI for understanding, enhancing, and verifying models in image-based applications. Table 7 summarizes the benefits of applying feature analysis to images through Explainable AI (XAI).

5.1. XAI Techniques

Optimizing important features in machine learning models with Explainable AI (XAI) techniques is critical for increasing performance, improving interpretability, and fostering trust in predictions. Table 8 provides a comprehensive summary of popular XAI techniques, outlining their features and the substantial advantages they provide to machine learning model building and improvement. These tactics are very useful in guaranteeing models are both effective and transparent in their decision-making processes. Table 8, enhancing model performance by optimizing features through XAI techniques is presented.

6. Evaluation of the algorithms performance

There are various criteria for evaluation that can be utilized to evaluate outcomes for machine learning supervised models [67]. The most frequently applied classification binary metrics for evaluation are sensitivity, specificity, accuracy, and precision, which represent a percentage of instances correctly identified in the collection of all situations, truly positive instances, truly negative instances, and positively classified instances, respectively. Sensitivity is known as well as recall [68].

6.1. Confusion matrix

The four variables in the confusion matrix are the important quantities for determining the metrics for a binary classifier. True positives (TP) are accurately anticipated positive outcomes, true negatives (TN) are successfully predicted negative outcomes, false positives (FP) are negative instances projected to be positive, and false negatives (FN) are positive instances predicted to be negative [69].

Actually Positive (1)	Actually Negative (0)	
True Positives (TPs)	False Positives (FPs)	Predicted Positive (1)
False Negatives (FNs)	True Negatives (TNs)	Predicted Negative (0)

6.2. Accuracy

Determined as the combined value of the two correct predictions (TP + TN) divided by the entire amount of data sets (P + N). The maximum accuracy is 1.0, whereas the lowest is 0.00 [70].

$$Acc. = TP + TN - TP + TN + FP + FN \ [0,1] \dots (1)$$

6.3. True Positive Rate (Sensitivity or Recall)

Determined as the amount of correct positive predictions (TP) divided by the total amount of positive predictions (TP + FN). Also referred to as sensitive or recall (REC). The highest possible TP rate is 1.0, whereas the lowest rate is 0.0 [70].

$$Sen. = Rec. = TP - TP + FN \quad [0,1] \dots (2)$$

Table 7. Benefits and Techniques of Explainable AI (XAI) for Model Transparency [12, 61, 62, 63].

Benefit	Description	XAI Technique
Model Interpretability	Explains model predictions to boost reliability and usability in essential applications.	LIME, SHAP
Local and Global Explanations	Provides insights at the instance and feature levels, allowing for greater comprehension of specific predictions and generally model behavior.	SHAP
Feature Importance Analysis	Identify the important qualities that influence predictions, and allow input data to be prioritized or refined.	SHAP
Debugging Models	Identifies biases, mistakes, and overfitting within machine learning models, allowing for enhancement of the model.	Anchors, LIME
Domain Adaptation	Enables domain specialists to connect predictive models with domain-specific knowledge via transparent reason.	Anchors, Counterfactuals
Trust in Automation	Increases trust in AI adoption in sensitive fields such as healthcare and finance by providing clear explanations for actions.	LIME, SHAP
Fairness Assessment	Identifies and mitigates discriminatory trends in predictions, resulting in ethical AI systems.	LIME, SHAP, Fairness-Specific XAI
Improved Decision-Making	Helps make smarter decisions by offering actionable insights to AI model predictions as well as data.	LIME, SHAP

Table 8. XAI Techniques for Optimizing Important Features [64 - 66].

XAI Technique	Description	Benefits
SHAP (SHapley Additive Explanations)	Gives every feature a significance rating depending on how it contributes to the predictions using cooperative game theory principles.	Provides an appropriate as well as comprehensive awareness of feature contributions, guiding feature selection and optimization.
LIME (Local Interpretable Model-agnostic Explanations)	Creates local surrogate models that interpret individual predictions by perturbing the input data.	Enhances understanding of model behavior for specific predictions, allowing targeted improvements in feature selection.
Grad-CAM (Gradient-weighted Class Activation Mapping)	Makes use of gradients from convolutional layers to produce visual explanations highlighting important regions in images.	Validates model focus on relevant image features, ensuring alignment with domain expertise and improving interpretability.
Integrated Gradients	Calculates the integral of gradients a route through an initial input towards the actual input, assessing feature importance.	Offers a robust method for understanding feature contributions in deep learning models, particularly useful for image classification tasks.
Attention Mechanisms	Uses attention scores from models like Transformers to determine which input features are most relevant for predictions.	Allows for dynamic feature importance assessment based on context, improving model adaptability and performance.

6.4. Precision

Determined as the amount of true positive predictions (TP) divided by the total amount of positive predictions (TP + FP). The accuracy that is greatest is 1.0, whereas the least accurate is 0.0 [70].

$$Pre. = TP — TP + FP \quad [0,1] \ldots (3)$$

6.5. True Negative Rate – TNR (Specificity)

Determined as the amount of true negative predictions (TN) divided by the whole amount of negative (N). The highest specificity is 1.0, while the lowest value is 0 [70].

$$Spe. = TN — TN + FP \quad [0,1] \ldots (4)$$

6.6. FP Rate - False Positive Rate

Determined by dividing the amount of false positives (FP) by the whole amount of negatives (N). The optimal rate for false positives is zero, while the lowest rate is 1.0. It can also be identified as one-specificity [71].

$$FPR = FP - TN + FP \quad [0,1] \ldots (5)$$

6.7. F-Measure or F-score (F1)

The F-measure is now frequently used as a performance indicator for a range of prediction tasks, including multi-label classification (MLC), binary classification, and structured output prediction [72]. The F1-score is limited to the range [0, 1], with 1 representing maximum accuracy and recall values and 0 representing zero precision and/or recall [73].

$$F1\,Score = 2 * Precision * Recall - Precision + Recall \ldots (6)$$

6.8. ROC Area - Receiver Operating Characteristic Area

The ROC curve is a graph that depicts the compromise between true positive and false positive rates. The higher the rate of true positives and the smaller the rate Of the false positives at every threshold, the more favorable. The region underneath the ROC curve is known as the ROC AUC score, which is a statistic that indicates how excellent the ROC curve is [75]. The ROC AUC Score displays the success of the model is current rating results. The likelihood that a randomly generated positive example ranks more highly than a selected-at-random negative example [71, 73]. Figure 8 shows ROC curve, which evaluates a classifier's performance. The x-axis represents the False Positive Rate (FPR), while the y-axis represents the True Positive Rate (TPR). The red curve illustrates the trade-off between sensitivity and specificity. A model with a curve closer to the top-left corner performs better. The dotted lines indicate a specific threshold's FPR_a and TPR_a. ROC curves help compare different models and determine the optimal decision threshold.

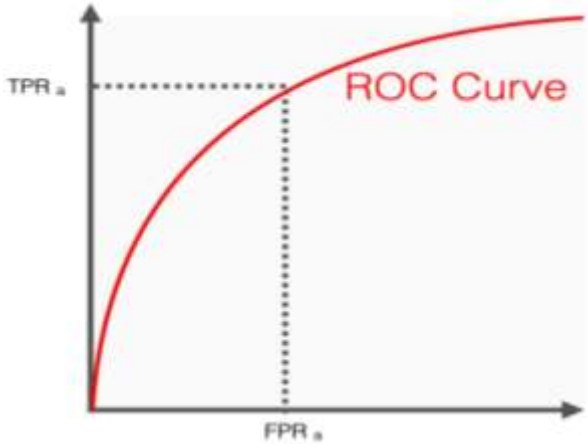


Figure 8. ROC curve performance evaluation of a classification model [74].

6.9. AUC - Area Under the Curve

The area under the receiver operating characteristic (ROC) plot represents the discriminatory capacity of classification models [75]. The area under the curve (AUC) is a widely used summary metric of the receiver operating characteristic (ROC) curve. It describes a diagnostic test's overall performance in terms of accuracy at various diagnostic thresholds used to distinguish between cases and non-cases of disease. The AUC metric is also employed in meta-analyses, in which each component study estimates the test's sensitivity and specificity [76]. AUC effectively distinguishes between "good" and "bad" models, but not between "good" models [77]. AUC values near 1.0 implying that the marker has a high diagnostic accuracy [78]. Figure 9 displays an ROC curve, showing the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) at different thresholds. The Area Under the Curve (AUC) measures classifier performance, with a higher AUC indicating better accuracy. The diagonal line represents random classifier performance.

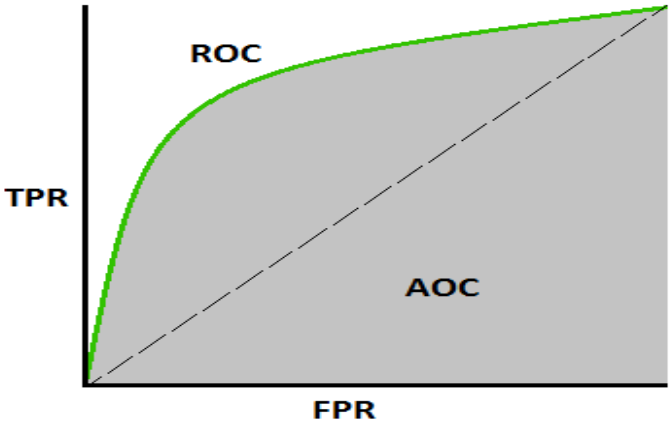


Figure 9. Area Under the Curve [79].

7. Literature Review

During the past few years, the utilization of machine learning (ML) techniques in healthcare has generated much attention, particularly in the field of predictive analytics for various medical diseases. In Table 9, many comparative analysis of XAI techniques in medical fields are presented. Guleria et-al. [46] proposed XAI frameworks for cardiovascular disease prediction that use a variety of models, which involve Support Vector Machines (SVM), AdaBoost, K-Nearest Neighbors (KNN), Bagging, Logistic Regression (LR), and Naive Bayes. Their solution used SHAP and LIME for feature interpretability, resulting in an astonishing 89% accuracy with ensemble models. Similarly, Dave et al. [80] aimed to increase trust in XAI for heart disease diagnosis; however, particular findings were not provided. Porto et al. [81] investigated the prediction of cardiovascular illness while improving the interpretability of models, obtaining 86% accuracy with decision trees, Naive Bayes, and SVM. Moreno-Sanchez et al. [82] focused on heart failure survival prediction using ensemble trees and ensemble interpretability approaches but did not report their findings. In a similar study, Peng et al. [83] created an XAI framework for hepatitis auxiliary diagnosis using XGBoost, SVM, and Random Forest, with Random Forest attaining an impressive accuracy of 91.9%. Ghosh and Khandoker (2024) [84] improved early chronic kidney disease (CKD) prediction with interpretable ML approaches, attaining a 93.29% accuracy with XGBoost and an AUC of 0.9689. Salekin et al. [85] also employed Random Forest for CKD prediction, attaining a significant F1 score of 99.3% via feature selection. Lundberg and Lee [61] underline the relevance of model interpretability in healthcare by using model-agnostic approaches, specifically SHAP, to diverse healthcare datasets and offering interpretable explanations. In oncology, Parmar et al. used Random Forest, Bayesian techniques, and SVM to identify predictive biomarkers in head and neck cancer, with an AUC of up to 0.72 [86]. Macyszyn et -al. employed SVM to predict patient survival and glioblastoma subtypes based on MRI imaging data, with an accuracy of as high as 80% in survival classification [87]. Tabari et al. [88] used Random Forest and SHAP to identify the pathological response of hepatocellular carcinoma (HCC) after ablation, with an AUC of 0.83. Severn et al. [89] created an explainable pipeline for medical imaging in radiomics using Elastic Net, Random Forest, and XGBoost, with an AUC of 0.969 for Elastic Net. Sadeghsalehi et al. [90] predicted primary tumors from brain metastasis MRI using XGBoost and Random Forest, achieving an impressive 99% accuracy using FOX-optimized XGBoost. In pediatrics, Amie J. Barda et al. [91] applied Random Forest and SHAP to predict in-hospital mortality risk for pediatric ICU patients, with an AUROC (Area Under the Receiver Operating Characteristic) of 0.94 and an AUPRC (Area Under the Precision-Recall Curve) of 0.78. Tae Keun Yoo et al. [92] used multiclass XGBoost and SHAP. To choose the best laser refractive surgery approaches, achieving an accuracy rate of 81.0% for internal validation and 78.9% for external validation. Finally, Furui Cheng et al. [93] used SHAP to improve clinicians' comprehension of ML predictions in healthcare across a variety of predictive tasks, although no precise accuracy data were offered. Collectively, these studies highlight the importance of explanation and interpretation in the applications of machine learning predictions to medicine, allowing physicians and patients to make better decisions, and build confidence.

8. Conclusion

AI and machine learning have played important roles in the field of healthcare, and they have emerged as possible tools for building and implementing smart systems. Applications based on AI save individuals both money and time while also providing early medical assistance. Humans are unable to completely trust machine learning (ML)-based choices in high-stakes scenarios, including COVID-19 testing, detecting fraud, loan sanctions, scoring of credit, and so on, because standard machine learning (ML) algorithms lack transparency owing to our black-box structure. Explaining ML outcomes to assist humans in comprehending the core features of running ML models. This paper discusses the role of AI and ML in the medical field. In this review, we present the

transformative potential of explainable AI (XAI) in healthcare, providing critical insights into the application and benefits of interpretability techniques. Prominent XAI methodologies, such as SHAP and LIME, are frequently utilized for feature importance analysis and localized explanations, enabling clinicians to trust and understand predictive models. For instance, studies like Guleria et al. [10], Ghosh & Khandoker [83], and Salekin et al. [84] demonstrated that ensemble models like XGBoost and Random Forest excel in chronic disease prediction, achieving remarkable accuracy (e.g., 93.29% for CKD prediction by Ghosh & Khandoker) while offering interpretable outputs. Similarly, SHAP was pivotal in studies like those by Tabari et al. [87] and Severn et al. [88], where radiomic features from MRI data were explained to predict outcomes in hepatocellular carcinoma and glioma, achieving AUCs of 0.83 and 0.969, respectively. SHAP and LIME are prominent approaches for assessing model interpretability. SHAP generates consistent and theoretically founded explanations by computing each feature's contribution using Shapley values, although it can be computationally costly. LIME, on the other hand, locally approximates a model using a simpler surrogate and is compatible with any machine learning model, providing flexibility but possibly unstable explanations. SHAP excels in consistency and theoretical rigor, whereas LIME is more adaptable but may suffer from more fluctuation and instability. The decision between them is determined by the model and analytic requirements. These findings highlight the critical role of XAI in identifying relevant features, enhancing model transparency, and fostering clinician trust. Practical implications include creating clinician-friendly interfaces for XAI technologies and ensuring that explanations are clear and actionable. Collaboration between AI researchers and healthcare professionals will be critical for designing solutions that correspond with clinical goals, eventually improving patient care and results. The findings open the path for a more transparent and trustworthy use of AI in healthcare, bridging the gap between sophisticated technology and practical applications.

Table 9. Applies XAI techniques in heart disease and other healthcare applications.

Authors	Focus	Machine Learning Model	XAI Techniques	Dataset	Results
Guleria et al. [10]	XAI Frameworks for Cardiovascular Disease Predictions	SVM, AdaBoost, KNN, Bagging, LR, NB	SHAP, LIME for feature interpretability	Cardiovascular dataset (303 instances, 14 features)	Achieved 89% accuracy with ensemble models (Naive Bayes, LR, SVM)
Dave et al.[80]	Building trust in XAI for heart disease diagnosis	Black-box model	SHAP, LIME	Heart disease dataset	Not specified
Porto et al. [81]	Predict cardiovascular disease while enhancing the interpretability models of the machine learning used	Naive Bayes, decision trees, (SVM), and artificial neural networks (ANN).	Attribute reduction, interpretable model	Statlog Heart Disease Dataset	86% accuracy
Moreno-Sanchez et al. [82]	Heart failure survival prediction	Ensemble trees	Ensemble interpretability methods	Heart failure dataset	Not specified
Peng et al. [83]	XAI framework for hepatitis auxiliary diagnosis	XGBoost, SVM, Random Forest	XAI framework for transparent predictions	Hepatitis dataset	Random Forest achieved 91.9%
Ghosh & Khandoker (2024)[84]	Early CKD prediction using interpretable ML	XGBoost, Logistic	SHAP and LIME for model interpretability	491 CKD patients (UAE dataset)	XGBoost achieved 93.29%

Table 9. Applies XAI techniques in heart disease and other healthcare applications.

Authors	Focus	Machine Learning Model	XAI Techniques	Dataset	Results
		Regression, RF, DT, NB			accuracy, AUC 0.9689
Salekin et al. [85]	CKD prediction using RF with feature selection	Random Forest	SHAP for feature importance analysis	CKD dataset	99.3% F1 score, similar results with 10 attributes
Lundberg & Lee. [61]	Model interpretation in healthcare	Model-agnostic techniques	SHAP for localized and global explanations	Various healthcare datasets	Provided interpretable explanations using SHAP
Parmar et al. [86]	Prognostic biomarker identification in head and neck cancer.	Random Forest, Bayesian, SVM	Feature importance analysis	Head and neck cancer CT images; 196 patients	AUC: up to 0.72
Macyszyn et al. [87]	Predicting patient survival and glioblastoma subtype based on MRI imaging features.	Support Vector Machines (SVM)	Feature importance ranking	Glioblastoma MRI dataset	Accuracy up to 80% in survival classification
Tabari et al. [88]	Predicting HCC pathologic response post-ablation	Random Forest	SHAP for model interpretability	MRI dataset, 97 HCC patients	AUC 0.83
Severn et al. [89]	Explainable pipeline for medical imaging in radiomics	Elastic Net, Random Forest, XGBoost	SHAP for model interpretability and feature importance	MRI data from glioma patients	AUC 0.969 (Elastic Net)
Sadeghsalehi et al. [90]	Predicting primary tumor from brain metastasis MRI	Random Forest, XGBoost	SHAP	Brain Metastasis MRI dataset (75 patients)	99% (FOX-optimized XGBoost)
Amie Janeth Barda [91]	In-hospital mortality risk prediction for pediatric ICU	Random Forest	SHAP	Pediatric Intensive Care Unit (PICU) data	AUROC: 0.94, AUPRC: 0.78
Tae Keun Yoo et al. [92]	Selection of optimal laser refractive surgery technique	Multiclass XGBoost	SHAP	This dataset includes ophthalmic and interview data from Individuals who planned to get refractive surgery during B&VIIT Eye Center, South Korea	81.0% (internal), 78.9% (external)
Furui Cheng et al.[93]	Enhancing clinicians' understanding of ML predictions in healthcare	Various ML models used in predictive tasks (e.g., surgical complication prediction)	SHAP	Paediatric Intensive Care (PIC) Database	Accuracy not specified in the paper

Acknowledgments: The authors gratefully acknowledge all those who provided scientific, technical support that contributed to the completion of this study.

Conflicts of Interest: The authors declare no conflict of interest.

Funding: This research was self-funded.

References

[1] Li, X.H.; Cao, C.C.; Shi, Y.; Bai, W.; Gao, H.; Qiu, L.; and Chen, L.; "A survey of data-driven and knowledge-aware explainable AI". IEEE Trans. Knowl. Data Eng., 34 (1): 29–49, 2020.

[2] Rudin, C.; "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". Nat. Mach. Intell., 1(5): 206–215, 2019.

[3] Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D.; "A survey of methods for explaining black box models". ACM Comput. Surv (CSUR), 51(5): 1–42, 2018.

[4] Vellido, A.; "The importance of interpretability and visualization in machine learning for applications in medicine and health care" .Neural Comput. Appl., 32(24): 18069–18083, 2020.

[5] Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; and Herrera, F.; "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". Inf. Fusion, 58: 82–115, 2020.

[6] Adadi, A.; and Berrada, M.; "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". IEEE Access, 6: 52138–52160, 2018.

[7] Zednik, C.; "Solving the black box problem: A normative framework for explainable artificial intelligence". *Philos. Technol.*, 34(2): 265–288, 2021.

[8] Harris, M.; "The Radical Scope of Tesla's Data Hoard: Every Tesla is providing reams of sensitive data about its driver's life". *IEEE Spectrum*, 59 (10): 40-45, 2022.

[9] Gudla, R.; Telidevulapalli, V.S.; Kota, J.S.; and Mandha, G.; "Review on self-driving cars using neural network architectures". *World J. Adv. Res. Rev.*, 16 (2): 736-746, 2022.

[10] Das, S.; Agarwal, N.; Venugopal, D.; Sheldon, F.T.; and Shiva, S.; "Taxonomy and Survey of Interpretable Machine Learning Method." Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, Australia, 1-4 December 2020; IEEE: 2020.

[11] Liu, C.Y.J.; and Wilkinson, C.; "Image conditions for machine-based face recognition of juvenile faces". Sci. Justice, 60(1): 43-52, 2020.

[12] Ribeiro, M.T.; Singh, S.; and Guestrin, C.; "'Why Should I Trust You?' Explaining the Predictions of Any Classifier". In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, California, USA, August 13-17 2016; Association for Computing Machinery: New York, USA,1135–1144, 2016.

[13] Alangari, N.; El Bachir Menai, M.; Mathkour, H.; and Almosallam, I.; "Exploring evaluation methods for interpretable machine learning: A survey". Information, 14 (8): 469, 2023.

[14] Richardson, R.; Schultz, J.M.; and Crawford, K.; "Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice". NYUL Rev. 94: 15, 2019.

[15] Srinivasu, P.N.; Sandhya, N.; Jhaveri, R.H.; and Raut, R.; "From blackbox to explainable AI in healthcare: existing tools and case studies" .Mob. Inf. Syst., 2022(1): 8167821, 2022.

[16] Azodi, C.B.; Tang, J.; and Shiu, S.H.; "Opening the black box: interpretable machine learning for geneticists". Trends Genet., 36(6): 442–455, 2020.

[17] Borys, K.; Schmitt, Y.A.; Nauta, M.; Seifert, C.; and Krämer, N.; Friedrich, C.M.; Nensa, F.; "Explainable AI in medical imaging: An overview for clinical practitioners–saliency-based XAI approaches," Eur. J. Radiol., 162: 110787, 2023.

[18] Cao, Q.H.; Nguyen, T.T.H.; Nguyen, V.T.K.; and Nguyen, X.P.; "A Novel Explainable Artificial Intelligence Model in Image Classification Problem".arXiv Prepr., 2023.

[19] Lundberg, S.M.; and Lee, S.I; "A Unified Approach to Interpreting Model Predictions". Proceedings of the 31st International Conference on Neural Information Processing Systems, California, USA, 2017; Curran Associates Inc.: NY, USA, 2017.

[20] Nicodeme, C.; "Build Confidence and Acceptance of AI-Based Decision Support Systems – Explainable and Liable AI". *Proceedings of the 2020 13th International Conference on Human System Interaction (HSI)*, Tokyo, Japan, 6–8 July 2020; IEEE: Piscataway, USA, 2020.

[21] Gunning, D.; and Aha, D.; "DARPA's Explainable Artificial Intelligence (XAI) Program". AI Mag., 40(2): 44–58, 2019.

[22] Linardatos, P.; Papastefanopoulos, V.; and Kotsiantis, S.; "Explainable AI: A review of machine learning interpretability methods". Entropy, 23(1): 18, 2020.

[23] Aslam, N.; Khan, I.U.; Mirza, S.; AlOwayed, A.; Anis, F.M.; Aljuaid, R.M.; and Baageel, R.; "Interpretable machine learning models for malicious domains detection using explainable artificial intelligence (XAI)". Sustainability, 14(12): 7375, 2022.

[24] Holzinger, A.; Saranti, A.; Molnar, C.; Biecek, P.; and Samek, W.; "Explainable AI Methods - A Brief Overview". In xxAI - Beyond Explainable AI, Lecture Notes in Computer Science; Holzinger, A.; Goebel, R.; Fong, R.; Moon, T.; Müller, K.R.; Samek, W., Eds.; Springer: Cham, Switzerland, 13–38, 2022.

[25] Calegari, R.; Ciatto, G.; Dellaluce, J.; and Omicini, A.; "Interpretable Narrative Explanation for ML Predictors with LP: A Case Study for XAI". *Proceedings of the 20th Workshop "From Objects to Agents"*, Parma, Italy, 26th–28th June 2019; Bergenti, F., Monica, S.; CEUR-WS.org: Pisa, Italy, 2019.

[26] Linardatos, P.; Papastefanopoulos, V.; and Kotsiantis, S.; "Explainable AI: A review of machine learning interpretability methods" .Entropy, 23(1), 18, 2020.

[27] Miller, T.; "Explanation in artificial intelligence: Insights from the social sciences" .Artif. Intell., 267: 1–38, 2019.

[28] Kim, B.; and Doshi-Velez, F.; "Machine learning techniques for accountability". AI Magazine, 42(1): 47–52, 2021.

[29] Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; and Yu, B.; "Definitions, methods, and applications in interpretable machine learning". Proc. Natl. Acad. Sci., 116(44): 22071–22080, 2019.

[30] Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; and Yu, B.; "Definitions, methods, and applications in interpretable machine learning". *Proc. Natl. Acad. Sci.*, 116 (44): 22071-22080, 2019.

[31] Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L.; "Explaining Explanations: An Overview of Interpretability of Machine Learning". Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, 2018; IEEE: Turin, Italy, 2018.

[32] Gacto, M.J.; Alcalá, R.; and Herrera, F.; "Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures". Inf. Sci., 181(20): 4340–4360, 2011.

[33] He, C.; Ma, M.; and Wang, P.; "Extract interpretability-accuracy balanced rules from artificial neural networks: A review" .Neurocomputing, 387: 346–358, 2020.

[34] Loi, M.; Ferrario, A.; and Viganò, E.; "Transparency as design publicity: explaining and justifying inscrutable algorithms". Ethics Inf. Technol., 23(3): 253–263, 2021.

[35] Sachan, S.; Almaghrabi, F.; Yang, J.B.; and Xu, D.L.; "Evidential reasoning for preprocessing uncertain categorical data for trustworthy decisions: An application on healthcare and finance". Expert Syst. Appl., 185: 115597, 2021.

[36] Hamon, R.; Junklewitz, H.; and Sanchez, I.; "Robustness and Explainability of Artificial Intelligence". Publications Office of the European Union, Luxembourg, 2020.

[37] Bhatt, U.; Xiang, A.; Sharma, S.; Weller, A.; Taly, A.; Jia, Y.; and Eckersley, P.; "Explainable machine learning in deployment". In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 648–657, 2020.

[38] Belle, V.; and Papantonis, I.; "Principles and practice of explainable machine learning" .Front. Big Data, 4: 688969, 2021.

[39] Carvalho, D.V.; Pereira, E.M.; and Cardoso, J.S.; "Machine learning interpretability: A survey on methods and metrics". Electronics, 8(8): 832, 2019.

[40] Guleria, P.; and Sood, M.; "Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling". Educ. Inf. Technol., 28(1): 1081–1116, 2023.

[41] Wen, Y.; and Holweg, M.; "A Phenomenological Perspective on AI Ethical Failures: The Case of Facial Recognition Technology". AI & Society, 39(4), 1929–1946, 2024.

[42] Guidotti, R.; Monreale, A.; Giannotti, F.; Pedreschi, D.; Ruggieri, S.; and Turini, F.; "Factual and counterfactual explanations for black box decision making". IEEE Intell. Syst., 34(6): 14–23, 2019.

[43] Panigutti, C.; Guidotti, R.; Monreale, A.; and Pedreschi, D.; "Explaining Multi-label Black-Box Classifiers for Health Applications". In Precision Health and Medicine; Shaban-Nejad, A., Michalowski, M., Eds.; Springer: Cham, Switzerland, 155-167, 2020.

[44] Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; and Hussain, A.; "Interpreting black-box models: a review on explainable artificial intelligence". Cogn. Comput., 16 (1): 45–74, 2024.

[45] Deng, H.; "Interpreting tree ensembles with intrees". Int. J. Data Sci. Anal., 7(4): 277–287, 2019.

[46] Guleria, P.; Naga Srinivasu, P.; Ahmed, S.; Almusallam, N.; and Alarfaj, F.K.; "XAI framework for cardiovascular disease prediction using classification techniques". Electronics, 11(24): 4086, 2022.

[47] Singh, A.; Sengupta, S.; and Lakshminarayanan, V.; "Explainable deep learning models in medical image analysis". *J. Imaging*, 6 (6): 52, 2020.

[48] Tukey, J.W.; "Exploratory Data Analysis". Reading/Addison-Wesley, 2: 131-160, 1977.
[49] Jiarpakdee, J.; Tantithamthavorn, C.K.; Dam, H.K.; and Grundy, J.; "An empirical study of model-agnostic techniques for defect prediction models". IEEE Trans. Softw. Eng., 48 (1): 166–185, 2020.
[50] Yan, F.; Wen, S.; Nepal, S.; Paris, C.; and Xiang, Y.; "Explainable machine learning in cybersecurity: A survey". *Int. J. Intell. Syst.*, 37 (12): 12305–12334, 2022.
[51] Wang, N.; Zhang, H.; Dahal, A.; Cheng, W.; Zhao, M.; and Lombardo, L.; "On the use of explainable AI for susceptibility modeling: Examining the spatial pattern of SHAP values". Geosci. Front., 15(4): 101800, 2024.
[52] Kök, I.; Okay, F.Y.; Muyanlı, Ö.; and Özdemir, S.; "Explainable artificial intelligence (XAI) for Internet of Things: a survey". IEEE Internet Things J., 10 (16): 14764–14779, 2023
[53] Roy, T.; Das, P.; Jagirdar, R.; Shhabat, M.; Abdullah, M.S.; Kashem, A.; and Rahman, R.; "Prediction of mechanical properties of eco-friendly concrete using machine learning algorithms and partial dependence plot analysis". *Smart Constr. Sustain. Cities*, **3** (1): 2, 2025.
[54] Friedman, J.H.; "Greedy function approximation: a gradient boosting machine" .Ann. Statist., 29: 1189–1232, 2001.
[55] Goldstein, A.; Kapelner, A.; Bleich, J.; and Pitkin, E.; "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation". J. Comput. Graph. Stat., 24(1): 44–65, 2015.
[56] Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; and Hussain, A.; "Interpreting black-box models: a review on explainable artificial intelligence". Cogn. Comput., 16(1): 45-74, 2024.
[57] Apley, D.W.; and Zhu, J.; "Visualizing the effects of predictor variables in black box supervised learning models". J. R. Stat. Soc. Ser. B: Stat. Methodol., 82(4): 1059–1086, 2020.
[58] Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; and Ranjan, R.; "Explainable AI (XAI): Core ideas, techniques, and solutions". *ACM Comput. Surv.*, 55 (9): 1–33, 2023.
[59] García, M.V.; and Aznarte, J.L.; "Shapley additive explanations for NO2 forecasting". Ecol. Informatics, 56: 101039, 2020.
[60] Keerthana, C.S.; Nalluri, S.C.; Muskaan, S.; and Sadagopan, P.; "Explainable AI in Credit Card Fraud Detection: SHAP and LIME for Machine Learning Models".Proceedings of the 2025 10th International Conference on Signal Processing and Communication (ICSC), Noida, India, 2025; Publisher: 2025.
[61] Lundberg, S.M.; and Lee, S.-I.; "A Unified Approach to Interpreting Model Predictions". Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017; Curran Associates Inc.: Red Hook, USA, 2017.
[62] Mothilal, R.K.; Sharma, A.; and Tan, C.; "Explaining Machine Learning Classifiers Through Diverse Counterfactual Explanations". 2020 Conference on Fairness, Accountability, and Transparency; Association for Computing Machinery: New York, USA, 2020.
[63] Salih, A.M.; Raisi-Estabragh, Z.; Boscolo Galazzo, I.; Radeva, P.; Petersen, S.E.; Lekadir, K.; and Menegaz, G.; "A perspective on explainable artificial intelligence methods: SHAP and LIME". Adv. Intell. Syst., 7(1): 2400304, 2025.
[64] Zhao, C.; Liu, J.; and Parilina, E.; "ShapG: New Feature Importance Method Based on the Shapley Value". Eng. Appl. Artif. Intell., 148(1): 110409, 2025.
[65] Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; and Müller, H.; "Causability and Explainability of Artificial Intelligence in Medicine". *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 9(4): 1312, 2019.
[66] Amin, A.; Hasan, K.; Zein-Sabatto, S.; Chimba, D.; Ahmed, I.; and Islam, T.; "An explainable AI framework for artificial intelligence of medical things". In 2023 IEEE Globecom Workshops (GC Wkshps), 2097–2102, 2023.
[67] Santafe, G.; Inza, I.; and Lozano, J.A.; "Dealing with the evaluation of supervised classification algorithms". Artif. Intell. Rev., 44: 467-508, 2015.
[68] Canbek, G.; Taskaya Temizel, T.; and Sagiroglu, S.; "PToPI: A comprehensive review, analysis, and knowledge representation of binary classification performance measures/metrics". SN Comput. Sci., 4 (1): 13, 2022.
[69] Zhu, W.; Zeng, N.F.; and Wang, N.; "Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS". Proceedings of the NESUG Conference: Health Care and Life Sciences, New York, USA, 2010; NESUG: 2010.
[70] Tharwat, A.; "Classification assessment methods". Appl. Comput. Informat., 17(1): 168-192, 2021.
[71] Dembczynski, K.; Waegeman, W.; Cheng, W.; and Hüllermeier, E.; "An Exact Algorithm for F-Measure Maximization". In: Adv. Neural Inf. Process. Syst., Proceedings of the 24th Annual Conference on Neural Information Processing

Systems (NeurIPS), Granada, Spain, 2011; Curran Associates: NY, USA, 2011.
[72] Hicks, S. A.; Strümke, I.; Thambawita, V.; Hammou, M.; Riegler, M. A.; Halvorsen, P.; and Parasa, S.; "On evaluation metrics for medical applications of artificial intelligence". Sci. Rep., 12(1): 5979, 2022.
[73] Kumar, R.; and Indrayan, A.; "Receiver operating characteristic (ROC) curve for medical researchers". Indian Pediatr., 48: 277–287, 2011.
[74] Haitsam, H.; "Human Detection With Plant Bioelectric Signal". J. Qua Teknika, 15(01): 35-42, 2025.
[75] Pepe, M. S.; "Receiver operating characteristic methodology". J. Am. Stat. Assoc., 95: 308–311, 2000.
[76] Walter, S. D.; "The partial area under the summary ROC curve". Stat. Med., 24(13): 2025-2040, 2005.
[77] Marzban, C.; "The ROC curve and the area under it as performance measures". Weather Forecast., 19(6): 1106–1114, 2004.
[78] Faraggi, D.; and Reiser, B.; "Estimation of the area under the ROC curve". Stat. Med., 21(20): 3093-3106, 2002.
[79] Carrington, A.M.; Manuel, D.G.; Fieguth, P.W.; Ramsay, T.; Osmani, V.; Wernly, B.; and Holzinger, A.; "Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation". IEEE Trans. Pattern Anal. Mach. Intell., 45(1): 329-341, 2022.
[80] Sheu, R.K.; and Pardeshi, M.S.; "A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System". Sensors, 22(20): 8068, 2022.
[81] Porto, R.; Molina, J. M.; Berlanga, A.; and Patricio, M. A.; "Minimum relevant features to obtain explainable systems for predicting cardiovascular disease using the statlog data set". Appl. Sci., 11(3), 1285, 2021.
[82] Moreno-Sanchez, P. A.; "Development of an Explainable Prediction Model of Heart Failure Survival by Using Ensemble Trees". Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, USA, 2020; IEEE: 2020.
[83] Peng, J.; Zou, K.; Zhou, M.; Teng, Y.; Zhu, X.; and Zhang, F.; and Xu, J.; "An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients". J. Med. Syst., 45(5): 61, 2021.
[84] Ghosh, S. K.; and Khandoker, A. H.; "Investigation on explainable machine learning models to predict chronic kidney diseases". Sci. Rep., 14(1): 3687, 2024.
[85] Salekin, A.; and Stankovic, J.; "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes". Proceedings of the 2016 IEEE International Conference on Healthcare Informatics (ICHI), Chicago, USA, 2016; IEEE: 2016.
[86] Parmar, C.; Grossmann, P.; Rietveld, D.; Rietbergen, M. M.; Lambin, P.; and Aerts, H. J.; "Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer". Front. Oncol., 5: 272, 2015.
[87] Macyszyn, L.; Akbari, H.; Pisapia, J. M.; Da, X.; Attiah, M.; Pigrish, V.; and Davatzikos, C.; "Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques". Neuro-Oncol., 18(3): 417–425, 2015.
[88] Tabari, A.; D'Amore, B.; Cox, M.; Brito, S.; Gee, M. S.; Wehrenberg-Klee, E.; and Daye, D.; "Machine Learning-Based Radiomic Features on Pre-Ablation MRI as Predictors of Pathologic Response in Patients with Hepatocellular Carcinoma Who Underwent Hepatic Transplant". Cancers, 15(7): 2058, 2023.
[89] Severn, C.; Suresh, K.; Görg, C.; Choi, Y. S.; Jain, R.; and Ghosh, D.; "A Pipeline for the Implementation and Visualization of Explainable Machine Learning for Medical Imaging Using Radiomics Features". Sensors, 22(14): 5205, 2022.
[90] Xie, Y.; Li, X.; Yang, S.; Jia, F.; Han, Y.; and Huang, M.; "Radiomics Models Using Machine Learning Algorithms to Differentiate the Primary Focus of Brain Metastasis". Transl. Cancer Res., 14(2): 731, 2025.
[91] Barda, A.J.; "Design and Evaluation of User-Centered Explanations for Machine Learning Model Predictions in Healthcare". University of Pittsburgh: Pittsburgh, USA, 2020.
[92] Yoo, T. K.; Ryu, I. H.; Choi, H.; Kim, J. K.; Lee, I. S.; Kim, J. S.; Lee, G.; and Rim, T. H.; "Explainable Machine Learning Approach as a Tool to Understand Factors Used to Select the Refractive Surgery Technique on the Expert Level". Transl. Vis. Sci. Technol., 9(2): 8, 2020.
[93] Cheng, F.; Liu, D.; Du, F.; Lin, Y.; Zytek, A.; Li, H.; Qu, H.; and Veeramachaneni, K.; "Vbridge: Connecting the Dots Between Features and Data to Explain Healthcare Models". IEEE Trans. Vis. Comput. Graph., 28(1): 378–388, 2021.