Hindawi
Mobile Information Systems
Volume 2022, Article ID 8167821, 20 pages
https://doi.org/10.1155/2022/8167821

Review Article

From Blackbox to Explainable AI in Healthcare: Existing Tools and Case Studies

Parvathaneni Naga Srinivasu,[1] **N. Sandhya**,[1] **Rutvij H. Jhaveri**,[2] **and Roshani Raut**[3]

[1]*Department of Computer Science and Engineering-AIML, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana 500090, India*
[2]*Department of Computer Science & Engineering, Pandit Deendayal Energy University, Gandhinagar, India*
[3]*Pimpri Chinchwad College of Engineering, Savitribai Phule Pune University, Pune, India*

Correspondence should be addressed to Rutvij H. Jhaveri; rutvij.jhaveri@sot.pdpu.ac.in

Received 29 March 2022; Revised 29 April 2022; Accepted 10 May 2022; Published 13 June 2022

Academic Editor: Saqib Hakak

Introduction. Artificial intelligence (AI) models have been employed to automate decision-making, from commerce to more critical fields directly affecting human lives, including healthcare. Although the vast majority of these proposed AI systems are considered black box models that lack explainability, there is an increasing trend of attempting to create medical explainable Artificial Intelligence (XAI) systems using approaches such as attention mechanisms and surrogate models. An AI system is said to be explainable if humans can tell how the system reached its decision. Various XAI-driven healthcare approaches and their performances in the current study are discussed. The toolkits used in local and global post hoc explainability and the multiple techniques for explainability pertaining the Rational, Data, and Performance explainability are discussed in the current study. *Methods*. The explainability of the artificial intelligence model in the healthcare domain is implemented through the Local Interpretable Model-Agnostic Explanations and Shapley Additive Explanations for better comprehensibility of the internal working mechanism of the original AI models and the correlation among the feature set that influences decision of the model. *Results*. The current state-of-the-art XAI-based and future technologies through XAI are reported on research findings in various implementation aspects, including research challenges and limitations of existing models. The role of XAI in the healthcare domain ranging from the earlier prediction of future illness to the disease's smart diagnosis is discussed. The metrics considered in evaluating the model's explainability are presented, along with various explainability tools. Three case studies about the role of XAI in the healthcare domain with their performances are incorporated for better comprehensibility. *Conclusion*. The future perspective of XAI in healthcare will assist in obtaining research insight in the healthcare domain.

1. Introduction

Explaining Artificial Intelligence (XAI) [1, 2] is a field of machine intelligence engineering that ensures that complicated technics are made accessible and customizable for effective decision-making in the field of science and technology. Currently, artificial intelligence is applied in the divergent areas of the medical and the healthcare domain, which includes the prediction of the future illness from genomic data [3] diagnosis of critical diseases by analyzing the electronic health records at earlier stages [4] recognizing the crucial illness in advance through the AI-based Early Warning Score (EWS) [5], and smart clinical decision supporting models [4].

The medical data is vast and divergent, which includes both structured data and unstructured data that is obtained as the diagnosis reports in the form of text, imaging, and signal data, patient clinical records, electronic healthcare records, ward records, individual staff records, and data from the healthcare gadgets and the wearable devices alongside the data acquired from the sensor in the smart healthcare environment. Artificial intelligence-based algorithms are mainly influential in the biomedical and healthcare domains. The tremendous involvement of the AI

models has led to the rising issues concerning the deployed models' lack of clarity and explainability. Also, some of the predictive models are biased and tend to produce misleading outcomes [6].

The challenges of the adaptability, transparency, and biased behavior of the existing models have provoked transparency in the design models in conventional Artificial Intelligence. The goal is to assist the professionals with the comprehensibility and customizability of the existing mechanism through explainable AI models by adhering to transparency through various techniques. The current models in healthcare and medicine have limitations in grappling with real-world complexities and uncertainties [7]. The medical decision support paradigm must be substantially robust and precise as it deals with human survival, and the models must be sure enough to make an appropriate decision. The smart healthcare and intelligent machine models in the smart diagnosis have become interdisciplinary with other domains like cloud technology, data science, data analytics, the Internet of things, pattern recognition for precise prediction, and diagnosis of the abnormality, through the enlightenment of the Explainable Artificial Intelligence, which will assist in designing and developing the models that are transparent, traceable, and customizable in concerning to the implementation environment. Figure 1 represents the various applications of XAI technology.

Since the inception of smart healthcare and cognitive diagnosis in the medical field, researchers have been striving to achieve human-level AI Development over the past few years. They have shown that it is very complex. The model's adaptability has been sluggish despite enormous growth in the healthcare and medical data from clinical reports, electronic healthcare records, data from wearable devices, and data from ambient healthcare sensors. There has been a spike of interest in XAI approaches that can describe any AI model to solve explainability in various AI applications. Bringing simplicity to ML models by including precise information on why the model adopted a particular action has been one of XAI's goals, developing further understandable and transparent ML models while retaining high output standards is another [8]. The AI and ML models have been extensively used in various fields for promising outcomes, and the models need to be more transparent and accountable for working in the areas like healthcare and medicine.

The current study is primarily motivated by the demand for system design and models that are more transparent and interpretable. Several robust approaches have been employed as "black boxes" and have not supplied any knowledge on how and why specific assessment, classification, and prediction methods are used. This lack of transparency may not directly impact consumers' ability to operate the applications or tools that these models equip, but specialists may nevertheless comprehend their structure. But when it comes to medical applications since there is a significant link between the effectiveness of a diagnosis or therapy and the acceptance of the approach utilized, this has paved the way for the current study to discuss the various explainable approaches, tools, and case studies that would assist the young researcher in explaining the models used in the healthcare field. The overall contributions of the studies are listed as follows:

(i) Presenting the various domains where the XAI technology could be incorporated makes the models evident.

(ii) Discuss the various technologies like SHAP and LIME-based explainable models.

(iii) Discuss the various toolkits available for making the model explainable under various explainable factors.

(iv) Precisely present various types of explainability associated with the decision models.

(v) Presenting the case studies in the healthcare domain and the statistical analysis would assist in better comprehensibility of the study.

The expandability of the model has several further consequences, including the technological capabilities of various existing models, even though evidence shows that artificial intelligence algorithms will outperform humans in some computational tasks. To attain healthcare professionals' confidence and various stakeholders, the predictive models and the smart diagnosis approaches must be transparent, understandable, and explicable. The design model must be clear and understandable about the decision-making and the rationale for its actions. Such frameworks are called the white-box models that can specify the sequence of actions [9]. In this chapter, the contents are restricted to the methodology to encourage explainability in the Artificial Intelligence models and make it easy to understand its research. The implementation of the explainable AI in coherence with the softy computing, optimization models, cognitive models, and intelligent machine approaches in the present and the future is being presented. Various metrics like Local Interpretable Model-agnostic Explanation (lime), Shapley Additive Explanations (SHAP), Yellowbrick, ELI5, and PDP are used to evaluate the explainability of the model. But in the current study, the lime and SHAP models are used in the evaluation. Smart diagnostic and biomedical engineering over the explainable artificial intelligence in the previous study are presented in Table 1. The performances of various post hoc models using the approaches like lime, SHAP, and PDP concerning the Accuracy (Acc), Sensitivity (Sen), Specificity (Spec), Recall, and F1-Score are analyzed in the current study.

This review will cover the XAI technology background that elaborates on the transformation and characteristics of the XAI models in section two. Then, the literature that presents the various existing studies on the XAI technology and tool kits and the approaches that deal with various categories of XAI technology for healthcare is introduced in section three. The methodology for conducting this review is presented in section three. Section four presents the gathered results for the current implementation models of the XAI technology in the lime lite of explainability and includes three case studies.

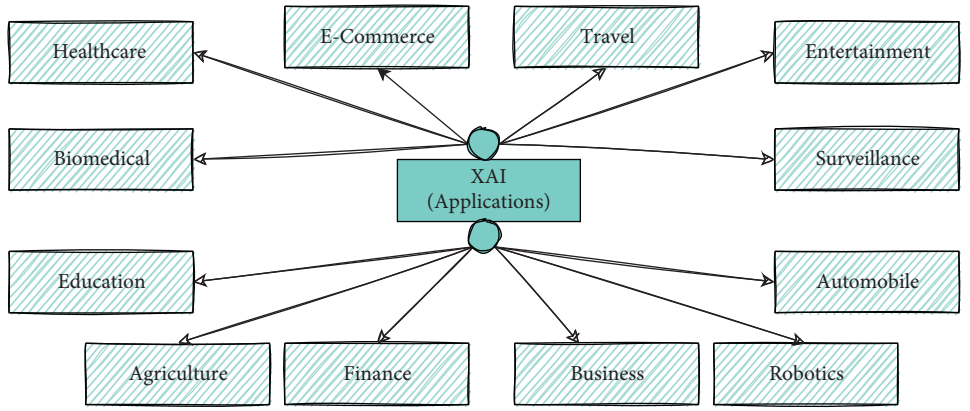

Figure 1: Representing the various applications of XAI.

Table 1: The performances of various XAI-driven models in healthcare.

Approach	Performance	Application	XAI framework
Entity-aware convolutional neural networks [10]	Sen: 88.8%	Used in clinical diagnosis and decision making	Bayesian network framework
Glioblastoma multiforme prediction [11]	Acc: 97%	Used in the diagnosis of glioblastoma multiforme	Lime
Convolutional neural network [12]	Acc: 93.3% Sen: 91%	Used in diagnosis of COVID-19 from chest X-ray	GSI
Random forest [13]	Acc: 93.95%	Used in predicting the alzheimer disease	SHAP
CNN: Over VGG-16 [14]	Pre: 95% F1-score: 94% Recall: 94%	Used in diagnosis of chronic wounds	Lime
Explainable cumulative fuzzy class membership criterion [15]	Acc: 91.08% Pre: 91.44% Recall: 91.04%	Used in diagnosis of colorectal cancer	Visual explanation
XG boost [16]	Acc: 78.9%	Used in decision-making strategies for evidence-based recommendation system for surgeries	SHAP
Random forest, support vector machine, decision tree [17]	Sen: 84% Spec: 67%	Used in diagnosis of Alzheimer	Sparse high-order interaction model with rejection
Convolutional neural network [18]	Acc: 95.2% Sen: 97.5% Spec: 90.9%	Used in the prediction of Parkinson's illness	Lime
Logistic regression, random forest, RF-AdaBoost, multilayer perceptron [19]	Acc: 71% Pre: 64% Recall: 26% F1-score: 59%	They were used in the analysis of post-stroke hospital discharge	Lime

Section six presents the conclusion and future perspective for XAI technology in the healthcare and medical domain.

2. Background

The Artificial Intelligent procedures involve several steps: data acquisition, data preprocessing, configuring, simulation, assessment, tuning, and recalibration. The models need to be transparent and understandable at each evaluation to make the outcome trustworthy. Nevertheless, various fields of science and technology are preferable to presenting the model in a completely opaque. However, in healthcare and the medical domain, the designed models must be transparent, and the rationale of a given decision to all relevant stakeholders should be explained [20]. This field benefits AI for various activities, including decision making, predictive evaluation, analysis, risk evaluation, and regulation [21]. Figure 2 represents the transformation of the XAI from the convectional black-box model to the feature white-box model.

The domain-specific features of the machine intelligent models in healthcare and biomedical engineering justify the decisions made in the assessment process. In recent years, complexity and the associated explanation have been presented in several studies to identify the situations where a concept is unclear to render the models more accessible [22].

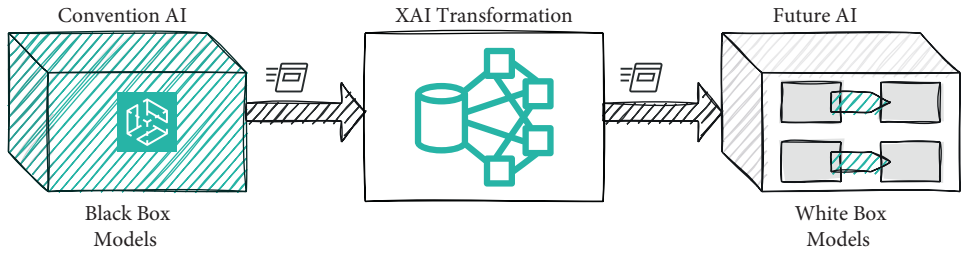

Figure 2: Image representing the transformation of XAI.

The explainability of the models is offered across the various aspects of the model like comprehensibility of the framework, understandability of the design of the system, comprehensibility of the objective functions, comprehensibility of the loss functions, comprehensibility of the training set, and the parameters associated with a training set of the model, comprehensibility of the validation set, comprehensibility of the parameters for the analysis, and the comprehensibility of the probabilistic assessment and the visual outcomes.

2.1. Local vs. Global Explainability. Artificial Intelligence has now been designed to automate choices, which may influence either good or bad business outcomes. Explainability may assist since it helps reveal the thought processes models use. In addition to being interpreted from the end user's perspective, the local interpretability of models includes giving comprehensive explanations about why a particular decision was obtained. The local explainability addresses the problem of explainability by splitting up the feature subset into smaller subsets and providing subsequent explanations for such simpler subsets that are important to the overall development of the precise model. The local explanation uses methods that differentiate the operation of a particular model component despite explaining just a subset of the system's overall function. Local Interpretable Model-Agnostic Explanations (lime) are one of the most predominantly used technologies in local explainability.

Determining the overall distribution of the desired result is enhanced by using a global model. For a multiclass classification model, comprehending the categorization of "good" or "bad" starts with the partial dependence model. In the real-time explainable global model, trained, understanding the method and data is essential. Seeing the model's choices holistically, from the feature set and each learned element like weights, biases, parameters, and structures, this is the degree of interpretability that assists in understanding the model.

2.2. Model-Specific vs. Model Agnostic. Model-specific models are the ones that are customized to a specific model or set of models specifically. Such models are very dependent on the functionality and features of a particular model. Model-specific algorithms perform their tasks by examining or providing detailed information about the model's internal working procedure [23].

An interpretation technique that is model-agnostic is much more versatile than one exclusive to a particular model. Model-agnostic techniques operate through understanding the correlation among input-output pairs of trained models rather than on the modeling techniques. The models are not dependent on the underlying structural framework used in determining the outcome. The model-agnostic models are further classified as the flexible model approaches, flexible explanation approaches, and flexible representation approaches.

2.3. Characteristics of XAI Models. In developing the machine, intelligence-operated models need to verify their behavior, as well as the working procedure before developing the model, and ensure its performance in a real-life environment. The explainability of the model is better understood by the terminology associated with the XAI model and the various assessment metrics used in evaluating the comprehensibility of the XAI model. The characteristics of the Explainable Artificial Intelligence model are presented as follows:

Explainability is synonymous with interpretation as a means of communication between human experts and the decision-maker, representing the decision-maker that is understandable to the expert.

Understandability is the model's ability to successfully communicate its latest research findings and its knowledge about formulating the solution to a human learner by elaborating the characteristics. The characteristics and the behavioral aspects of the individual components are analyzed to understand the structure and design of the model.

Fidelity determines the appropriateness and clarifies the implementation procedures by telling what the model performs internally. It describes the operational procedure of each component of the system.

Transparency: a paradigm is called transparent if it is understandable on its own due to the logic that every model has a varying degree of understandability and legibility on the decisions that are taken as part of yielding a better outcome.

Adaptability: the model's adaptability is all about the learning capability of the environment where the model is deployed to function. The model must be capable of rapidly adapting to new knowledge and gaining insight into the significance of that information.

The concept of explainable AI distinguishes two broad types of explainable techniques: straightforward and post hoc interpretability [24]. It shows to an extent how a machine arrived at production. This means that interpretations will represent a models' internal working. Frequently, developers assert that their model representations are meaningful. On the other hand, others would use metrics like human simulatability to determine if a human could appropriately take input and accurately interpret the model's output whenever combined with the prototype [25]. The performance of the models is generally assessed through the performance evaluation metrics like Accuracy, Sensitivity, Specificity, F1-Score, Jaccard Similarity Index, Mathews Correlation Coefficient from the assessed values of True Positive, True Negative, False Positive, and False Negative values of the proposed model in the assumed context [16–18].

3. Methodology

In the current study, the explainability of the artificial intelligence model is implemented through the techniques like Local Interpretable Model-Agnostic Explanations (lime) [26] and Shapley Additive Explanations (SHAP) [27] for understanding the explainability of the black-box model. SHAP is a standardized method for determining the relative significance of features in different models. The explanations are associated with the Shapley Values, which provide the average fractional contributions of the particular feature value conceivable feature combinations. Therefore, the SHAP calculated values correspond to the Shapley values of the model's dependent probability function.

3.1. Local Interpretable Model-Agnostic Explanations. The lime technique for the explainability of the models will enrich the potential of the machine learning model and the comprehension of its predictions. Therefore, the approach elucidates the classifier for a single instance and is appropriate for evaluating a local model. With a single prediction value, every classification model may yield various perceptions. For each occurrence and their associated prediction, the synthetic randomly selected facts are produced in the proximity of the input instance under which the forecast is generated. For each such instance, the probability is calculated and weighted by the closeness of the created instance to the input instance.

The working principle of the lime is to check for the local transparency and the comprehensibility of the model regardless of the technology and the algorithms that are involved in the design process. The features selected in the prediction process are significantly important to evaluate the local transparency of the model. However, the local explainability makes the prediction process more vigilant but may not perfectly suit the model globally. Let the explainable model be represented as $\mathbf{h}: \mathbb{R}^{x'} \longrightarrow \mathbb{R}, \mathbf{h} \epsilon \mathbf{H}$, where the variable $\mathbf{H}$ denotes the explainable artificial intelligent model like a decision tree or a rule-based classifier. For any given model $\mathbf{h} \epsilon \mathbf{H}$ that represents the model through various text-based and graphical illustrations, it is known that every $\mathbf{h} \epsilon \mathbf{H}$ model is completely explainable. Hence, the variable $\Omega\mathbf{h}$ represents the complexity of the interpretability concerning the soft and hard constraints [28].

The explainability of the classification model can be expressed as $\mathbf{f}: , \longrightarrow \mathbb{R}$, where the variable $\mathbf{f}(\mathbf{m})$ denotes the probability of the variable $\mathbf{m}$ belonging to a particular class. And the variable $\Pi_{\mathbf{m}}(\mathbf{p})$ presents the proximity among the instances $\mathbf{m}$ and $\mathbf{p}$ to determine the neighborhood of the $\mathbf{m}$. The variable $\vartheta(\mathbf{f}, \mathbf{h}, \Pi_{\mathbf{m}})$ determines the degree of unfairness $\mathbf{h}$ in predicting the $\mathbf{f}$ in the locality determined through $\Pi_{\mathbf{m}}$. It is desired that the value of $\vartheta(\mathbf{f}, \mathbf{h}, \Pi_{\mathbf{m}})$ be kept minimum for better explainability and local fidelity. Moreover, the value of the variable $\Omega\mathbf{h}$ is to be minimum for better explainability and the understandability of the model. The explainability through the lime model is presented through the variable $\varepsilon(\mathbf{m})$ in the following equation:

$$\boldsymbol{\varepsilon}(\mathbf{m}) = \sum_{\mathbf{h} \epsilon \mathbf{H}} \vartheta(\mathbf{f}, \mathbf{h}, \boldsymbol{\Pi}_{\mathbf{m}}) + \boldsymbol{\Omega}\mathbf{h}. \tag{1}$$

From equation (1), the variable ϑ denotes the fidelity function, and the notation Ω denotes the complexities. Estimating the value of the variable ϑ decomposed from the sample $\mathbf{m}$, assist in predictions over the black box model $\mathbf{f}$, and tuning the model in association with $\Pi_{\mathbf{m}}$. The same model can be applied to various classification and prediction models [29]. The process of the lime could be understood with a suitable healthcare-related that is discussed below.

In the process of assessing the future possibility of type-2 diabetes, numerous features like stabilized Glucose (stab. glu), age, total cholesterol (chol), gender, height, weight, hip, High-Density Lipoprotein (HDL), etc. are taken into consideration. The prediction model will provide the features as follows in the input dataset:

$$\textbf{Features} = [\textbf{stab.glu}, \textbf{age}, \textbf{chol}, \textbf{gender}, \textbf{height}, \textbf{weight}, \textbf{hip}, \textbf{HDL}]. \tag{2}$$

The features are associated with the values that will assist in determining the probability of being affected by type 2 diabetes.

$$\textbf{Features} = [\textbf{stab.glu} = \textbf{yes}, \textbf{chol} = \textbf{yes}, \textbf{weight} = \textbf{yes}, \textbf{HDL} = \textbf{yes}]. \tag{3}$$

The predictions are made in concern to the values that are associated with the features mentioned above set. The local explainability determines the transparency of the proposed model in evaluating the probability of type 2 diabetes based on the key features and the weights assigned to those parameters in the forecasting process. Lime model generally focuses on the internal functionality of each associated component in the prediction model.

3.2. Shapley Additive Explanations. SHAP [30] is a technique associated with lime for attributing the additive features expressed mostly by Shapley value explanations. SHAP is a technique to provide an overview of each forecast as it happens within conceptual game-theoretic optimum Shapley values. The Shapley values have gained widespread

popularity in collaborative game theory because they are valuable approaches accompanied by advantageous characteristics. The results offer a different additive feature set associated with approximating the preciseness of the precision and the consistency of the local model. SHAP effectively works with both model-agnostic and model-specific explanations. SHAP interprets the associated prediction model's output as the summation of its imputed values with each input feature vector. Figure 3 represents the framework of the SHAP-based explainable model.

All potential feature coalitions, either with or without the feature set, should be used to determine the precise Shapley value. Moreover, the number of potential coalitions increases drastically [31]. The resultant outcome of the explainable model is influenced by the associated feature set, which is explained through equation (4). The variables "$\mathbf{p}$" and "$\mathbf{q}$" are the independent variables whose values are interdependent, used in approximating the outcome.

$$\boldsymbol{\Theta}(\mathbf{p}) = \left\{ \frac{\partial \mathbf{p}}{\partial \mathbf{q}_1} + \frac{\partial \mathbf{p}}{\partial \mathbf{q}_2} + \cdots + \frac{\partial \mathbf{p}}{\partial \mathbf{q}_n} \right\}. \tag{4}$$

From equation (4), the variable $\Theta(\mathbf{p})$ denotes the output of the prediction model, and the variable $\mathbf{n}$ indicates the number of feature set values. The variable ω indicates the change in the weights associated with each feature vector. The changes in the feature vector in association with weight are expressed in the following equation:

$$\boldsymbol{\delta p} = [\boldsymbol{\omega}_1 \times \mathbf{q}_1, \boldsymbol{\omega}_2 \times \mathbf{q}_2, \ldots, \boldsymbol{\omega}_n \times \mathbf{q}_n]. \tag{5}$$

The weight ω is determined based on the K-Nearest Neighborhood, updated over the iterations. The value of weight ω is determined as shown in equation (6). To assess the weight, alter the parameters to decrease with each step, considering the present instance and its neighboring hits and misses. You may change a certain instance's closest neighbors by calculating the nearest hits and misses of every chosen instance at any given moment. The variable $\mathbf{e}^{\mathbf{b}_x}_{\mathbf{a}_{\text{hit}}}$ is associated with $\mathbf{a}^{\mathbf{th}}$ feature of the nearest hit over the sample $\mathbf{b}$. The $\rho^{\mathbf{R}}$ is associated with the likeliness of the instances over the constant learning rate $(1/\mathbf{p} \times \mathbf{m})$.

$$\boldsymbol{\omega} = \boldsymbol{\omega} - \left[\frac{1}{\mathbf{p} \times \mathbf{m}} \left(\sum_{x=1}^{p} \boldsymbol{\rho}^{\mathbf{R}} \left(\mathbf{e}^{\mathbf{b}}_{\mathbf{a}}, \mathbf{e}^{\mathbf{b}_x}_{\mathbf{a}_{\text{hit}}} \right) + \left(\sum_{x=1}^{p} \boldsymbol{\rho}^{\mathbf{R}} \left(\mathbf{e}^{\mathbf{b}}_{\mathbf{d}}, \mathbf{e}^{\mathbf{b}_x}_{\mathbf{d}_{\text{miss}}} \right) \right) \right] \right]. \tag{6}$$

The change in outcome is associated with the change in associated weights with each of the elements in the feature vector, which is represented through equation (7) with $\mathbf{v}$ number of feature vectors in the considered dataset.

$$\boldsymbol{\theta}(\mathbf{p}) = \frac{1}{\mathbf{v}} \sum_{i=1}^{v} \left[\frac{\partial \mathbf{p}}{\partial \mathbf{q}_1} \times \boldsymbol{\omega}_1 \times \mathbf{q}_1, \frac{\partial \mathbf{p}}{\partial \mathbf{q}_2} \times \boldsymbol{\omega}_2 \times \mathbf{q}_2, \ldots, \frac{\partial \mathbf{p}}{\partial \mathbf{q}_n} \times \boldsymbol{\omega}_n \times \mathbf{q}_n \right]. \tag{7}$$

SHAP then calculates the Shapley values for each feature and utilizes them to determine the feature's significance. SHAP is an excellent tool for assessing high-fidelity, reliable, and exhaustive explanations for classification and prediction models. The SHAP model relies on the feature vector of the model.

4. Results

In the result section, various existing toolkits for analyzing the explainability and the interpretability are discussed. Three case studies concerning to explainability of the prediction model in the healthcare domain are presented in this section, with a detailed framework of each of those models. The case studies focus on predicting the future illness using the classification models, Explainable deep learning models, and the Explainable model for Earlier Warning Score (XAI-EWS).

4.1. Toolkits and Frameworks. Explainable models are frequently supplementary modeling frameworks for the predictive and diagnosis approaches that try to expound upon on properties and components of the existing version. Most XAI frameworks in the healthcare domain are dedicated to retrofitting approximation approaches onto more sophisticated original AI models. Humans mostly focus on the evidence of the choices made over the comprehensive explanation of the model. Probabilities are less important than causation. Therefore, a statistical representation without a causal explanation would be less gratifying than a causal explanation.

Explainable Deep learning models [32] for the image analysis in the biomedical imaging domain have yielded reasonable performance, and the model is interpretable. The workflow is evident at each phase of the implementation of the diagnosis model. The features that influence the process of decision-making in the deep learning model are evident to the user. Recognizing the features that contribute to a particular decision enables concept makers to address reliability issues, allowing end consumers to build confidence and make more informed decisions. The explainability and the transparency of the decisions are exceptionally important in the healthcare models. Regarding human-interpretable explanations [33] through a hybrid model of network hidden layer with TREPAN decision tree [34], the model can produce higher-quality reason codes that provide concise, human-interpretable reasons for model outcomes in the instance stage.

Regarding improvised Trust Management system in the Intrusion Detection System (IDS) [35], the XAI would enhance the trust and confidence through the feature interpretability of the decision tree model used in IDS. And the role of trust management is equally important in healthcare models like earlier warning scores and disease forecasting models. The guidelines are explainable and accompany vulnerability management professionals in enriching their confidence by recommending a safety plan in the event of detection of malicious traffic. The classification of the estrogen receptor is using the CNN ensemble model through the XAI technology for better interpretability of the model [36]. XAI-based visualization was used to obtain insight into the characteristics of a DCNN equipped to identify the group

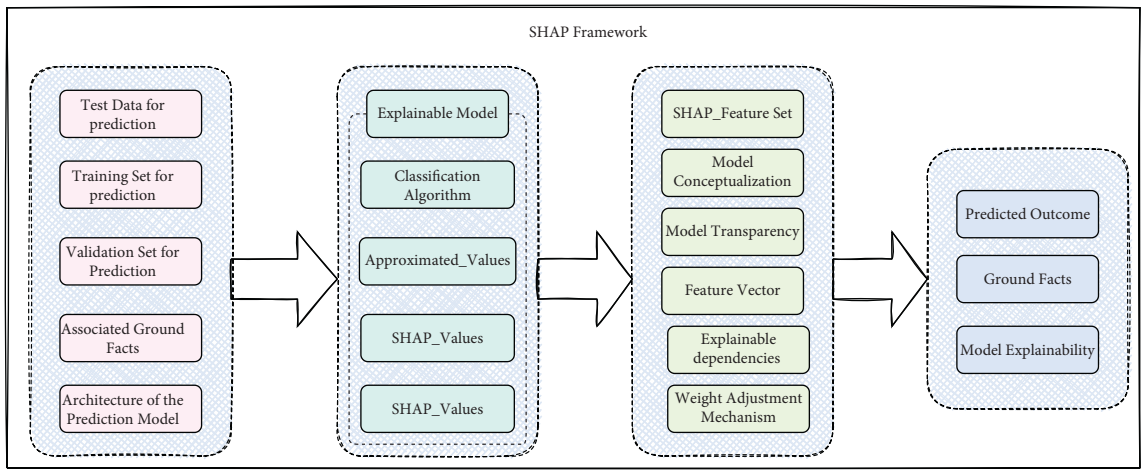

Figure 3: SHAP framework for explainable model.

of proteins generated inside the breast's cell using dynamic contrast-enhanced magnetic resonance imaging.

The process of gene expression patterns is being analyzed using the XAI technology. XAI emerged as a response to this issue, as rule-based methods are extremely well suited for empirical validation of the predictions made through gene analysis. XAI strategy focused on rules, including preprocessing, information retrieval, and operational validation for discovering biologically meaningful sequential rules from clinical gene expression data. Data Explainability presents the explainable aspects that are transparent and interpretable with the data that is associated with the model; for example, the training data and their associated data labels must be justified. Post hoc interpretability is a novel method for retrieving information from trained models. While post hoc interpretations may not always explain how a model works, they might provide helpful information to professionals and end consumers of machine learning. Natural language explanations, visualizations of acquired representations, or models are some typical ways of post hoc interpretations. Global explainable models may explain the overall behavior for any black-box models, whereas local explainable models can be used to interpret individual predictions. Table 2 presents the explainability of various other models concerning Data Explainability, Model Interpretability, Local Post Hoc, and the Global Post Hoc features of the toolkit.

A toolkit is a productivity software module that collects utility routines or an extended suite of software utilities for creating and maintaining applications and data. Arya et al. [37] have proposed a toolkit AI Explainability 360, which is available as open-access software with eight state-of-the-art features in handling the design and execution of the explainable model that includes Boolean Decision Rules via Column Generation (BRCG), Generalized Linear Rule Models (GLRM), ProtoDash, ProfWeight, Teaching Explanations for Decisions (TED), Contrastive Explanations Method (CEM), Contrastive Explanations Method with Monotonic Attribute Functions (CEM-MAF), and Disentangled Inferred Prior Variational Autoencoder (DIP-VAE) for the model. Table 3 represents the various characteristics like the acceptable input type, a local and global explainable toolkit feature, and the model-specific and agnostic feature associated with the toolkits [38].

The toolkits are efficient in evaluating the explainability and understandability of the model through various evaluation metrics [39], which are categorized as Rationale explanation, Data explanation, and Safety and performance explanation [40]. The Rationale explanation mainly focuses on the "why" of an artificially intelligent agent's decision and discusses the factors contributing to the decision. If the artificial intelligent agent judgment differed from what users predicted, this form of clarification enables stakeholders to determine if they think the decision's logic is faulty. Otherwise, the justification assists in developing logical reasons for a particular situation. Data Explanation demonstrates to the stockholders that an artificial intelligence algorithm is secure and effective by testing and monitoring the model's performance. Table 4 represents the various explanation categories and types considered in the design and development of the models. A neural network interpretability approach for NLP neural networks has been presented. This method attempts to provide prediction outcomes as the actual input by extracting smaller, customized parts of the same input text and then using them as input. The rationales, or justifications, are the little parts that explain and justify the outcome concerning the input. The safety and the performance explainability are the other significant explainability components desired to be explainable for any interpretable machine learning model. These metrics elucidate the relations among the individual constructs in the process of decision making, and the performance explainability is desired to make it evident about the summarizations that are made by the decision model.

Besides differential outputs caused by prejudice, AI performance often contradicts the output of professional clinicians. This unambiguous entitlement to factual knowledge proving the presence, reasoning, and projected implications of automated decision-making mechanisms

Table 2: Table representing various toolkits that implement explainable AI.

Toolkit	Data explainability	Model interpretability	Local post-hoc	Global post-hoc
AIX-360	✓	✓	✓	✓
Alibi			✓	✓
Skater		✓	✓	✓
H2O		✓	✓	✓
InterpretML		✓	✓	✓
EthicalML-XAI				✓
DALEX			✓	✓
iNNvestigate			✓	

Table 3: Features associated with various toolkits.

Toolkit	Input type	Local/global explainability	Specific/agnostic
AIX-360	Tabular and image data	Local/global	Agnostic
Alibi	Tabular and image data	Local/global	Agnostic
Skater	Tabular and image data	Global	Agnostic
H2O	Tabular and statistical data	Global	Agnostic
InterpretML	Tabular data	Local/global	Agnostic
EthicalML-XAI	Tabular data	Local/global	Agnostic
DALEX	Tabular and statistical data	Global	Agnostic
iNNvestigate	Image and tabular data	Local/global	Agnostic

Table 4: Table representing the various approaches associated with the explainability.

Approaches of explainability	Rational explanation	Data explanation	Safety and performance explanation
Saliency approaches	✓		
Neural network visualization	✓		
Knowledge distillation approaches	✓		
Restricted neural network	✓		✓
Feature relevance approaches		✓	
Exemplary approaches		✓	✓
High-level feature learning		✓	

remains unestablished, sparking controversy about the significance of interpretability and its limitations.

4.2. Case Studies. The prevalent assumption is that the more advanced artificial intelligence models are more accurate, implying that a sophisticated black box is required for superior prediction performance. This is often not true, even when the data are organized and well-represented in the context of inherently relevant characteristics. When dealing with structured data containing substantial characteristics, sometimes, there is no productivity distinction among the sophisticated classifiers (deep neural networks, boosted decision trees, and random forests) and relatively simpler classifiers after data preprocessing. When it comes to data science issues, where structured data with relevant characteristics is created as part of the information science process, there are few differences across methods. The data analyst follows a consistent method for data analysis. Three case studies of explainability are presented in this section to understand explainability from the healthcare perspective.

4.2.1. Case Study 1. Forecasting the risk of having a heart stroke for individuals has always been a focus of study for numerous researchers globally, as it is a common occurrence. There is compelling evidence that early knowledge of that risk may aid in prevention and treatment. Local Interpretable Model-agnostic Explanations (lime) [41] are a method that would not attempt to explain the entire model; rather, it attempts to explain the method by disturbing the input of sample data and observing the resulting changes in predictions. Lime facilitates the interpretability of local models. The output effect of modifying a particular sample data by changing certain feature vectors is experienced. When a model's output is examined, this is frequently connected to individual interests. Lime is an algorithm that can explain the predictions of any classifier or regressor faithfully by approximating it locally with an interpretable model Figure 4. In the initial phase of the prediction model, the entire dataset is partitioned as the training set and the testing set. Preprocessing is performed to ensure the data used for the training model fits the model specifications. And feature selection is performed on the training data to recognize the pivotal features that assist in the data's precise perdition of the heart stroke. And the lime models are used in analyzing the insights and the operational mechanisms of the classification models.

The explainable model needs to be trained with the preexisting data in heart stroke prediction. In the current case study, the heart disease dataset from the University of

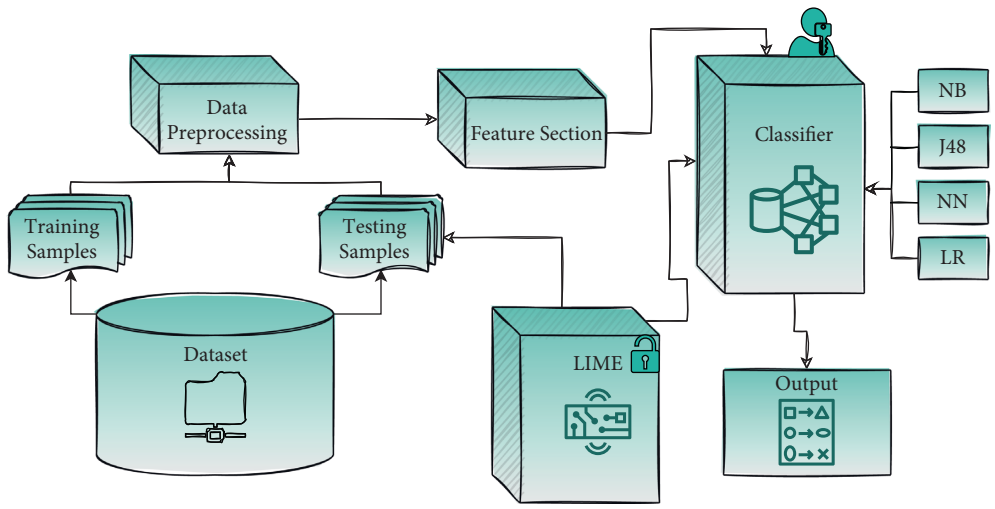

Figure 4: Image representing the block diagram of the lime based prediction model.

California is considered, and the dataset consists of 270 instances classified as present and absent. The present class determines that the particular instance had a heat stroke, and the other class indicates that no heart disease was associated with the instance. The dataset comprises 76 features, out of which 14 significant features are considered in the majority of the studies for evaluation [42, 43]. The considered features are being presented in Table 5.

Feature selection is exceedingly important in the precise illness assessment, and the feature selection process must be evident for any trustworthy model. The study on heart disease [44] on different feature sets has yielded divergent result. The feature set and accuracies obtained in the study are presented as follows:

$$\begin{aligned}
\text{feature}^{1}_{\text{set}} &= \{\text{sex, Fbs} - \text{As, Chest} - \text{Tdolor, Oldpeak} - \text{Dir, Restecg} - \text{Re, Thal} - \text{Tt, Ca} - \text{Nbp}\},\\
\text{feature}^{2}_{\text{set}} &= \{\text{age, Chest} - \text{Tdolor, sex, Fbs} - \text{As, Chol} - \text{Cs, Oldpeak} - \text{Dir, Ca} - \text{Nbp, Slope}\},\\
\text{feature}^{3}_{\text{set}} &= \{\text{Chest} - \text{Tdolor, sex, Thal} - \text{Tt, Fbs} - \text{As, Slope, Exang–Ei, Ca} - \text{Nbp, Thal} - \text{Tt}\},\\
\text{feature}^{4}_{\text{set}} &= \{\text{Chest} - \text{Tdolor, sex, Exang–Ei, Ca} - \text{Nbp, Thalach} - \text{Fcm, Slope, Thal} - \text{Tt}\},\\
\text{feature}^{5}_{\text{set}} &= \{\text{sex, age, Chol} - \text{Cs, Chest} - \text{Tdolor, Restecg} - \text{Re, Slope, Oldpeak} - \text{Dir, Ca} - \text{Nbp, Slope, Thal} - \text{Tt}\},\\
\text{feature}^{6}_{\text{set}} &= \{\text{sex, Trestbps} - \text{Pa, Chest} - \text{Tdolor, Restecg} - \text{Re, Fbs} - \text{As, Exang–Ei, Thal} - \text{Tt, Slope, Oldpeak} - \text{Dir, Ca} - \text{Nbp}\}.
\end{aligned} \tag{8}$$

The feature selection must be made in a more evident way, and the feature set which is assumed to be irrelevant is discarded from the consideration in the evaluation process. It focuses on reducing the number of superfluous features and dimensions of the data using Single Spectrum Analysis (SSA). The features are classified into two classes: the More Significant and Less Significant features. The fitness of the most significant feature is then updated using equations (9) and (10):

$$\text{Update}_{\text{best\_fit}} = \text{fit}_i + \alpha\left((ut_i - lt_i)\beta + lt_i\right)\gamma \geq 0, \tag{9}$$

$$\text{Update}_{\text{best\_fit}} = \text{fit}_i - \alpha\left((ut_i - lt_i)\beta + lt_i\right)\gamma < 0. \tag{10}$$

The position of the more significant feature set is determined by the variable $\text{Update}_{\text{fit}}$ and variables ut_i, lt_i denote the upper and lower threshold, and the variables α, β, γ represent the balancing factors. The value of the variable α that is used in exploration and exploitation is determined through the following equation:

$$\alpha = 2e^{-\left(4i/i_{\text{tot}}\right)^2}. \tag{11}$$

From equation (11), the variable i denotes the current iteration, and the variable i_{tot} represents the total number of iterations. The value of the variable α is scaled down lineally over the iterations. Similarly, the value of the second balancing factor recognized by β is influenced by the appropriate feature set, determined through the chaotic map.

$$\beta = \delta_c. \tag{12}$$

From equation (12), the variable δ_c is the resultant value of the chaotic gradient of the corresponding iteration. The fitness function for evaluating the feature set is determined through the following equation:

Table 5: Presents the features associated with the heart disease dataset.

Feature	Data type	Data category	Description
Age	Integer	Discreet	In years
Sex	Binary	Discreet	1 = male 0 = female
Chest-tdolor	Integer	Discreet	1 = typical angina 2 = atypical angina 3 = without angina 4 = asymptomatic
Trestbps-Pa	Float	Continuous	Resting blood pressure (mm hg)
Chol-Cs	Float	Continuous	Serum cholesterol
Fbs-as	Binary	Discreet	Fasting glucose
Restecg-Re	Integer	Discreet	Resting electrocardiographic
Thalach-fcm	Float	Continuous	Peak heat rate
Exang–Ei	Binary	Discreet	ExerExercise-inducedina
Oldpeak-dir	Float	Continuous	Exercise-induced depression in comparison to rest
Slope	Integer	Discreet	Slope associated with exercise1 = upward slope2 = flat3 = Downward slope
Ca-Nbp	Float	Discreet	Number of main vessels in real fluoroscopy
Thal-tt	Integer	Discreet	Thallium scan of heart muscles3 = normal 6 = nominal Irreversible7 = reversible defect
Class	Binary	Discreet	0 = healthy1 = possible heart disease

$$\text{fit}(x) = \omega \times \text{acc} + (1-\omega) \times \left(1 - \frac{I_g}{f_t}\right). \tag{13}$$

From equation (13), the variable ω denotes the weight factor, and the variable *acc* represents the accuracy. The variable I_g represents the information grain, and the variable f_t indicates the total number of features associated. The information gain is assessed using equation (14) with an assumption, where i denotes the all-possible feature set combinations shown through bins B_i

$$I_g(f) = -\sum_{f_t} \rho(B_i) \log \rho(B_i) + \rho(x) \sum_{f_t} \rho\left(\frac{B_i}{x}\right). \log \rho\left(\frac{B_i}{x}\right) + \text{fit}(x) \times \left(\rho(\hat{x}) \sum_{f_t} \rho\left(\frac{B_i}{\hat{x}}\right) \log \rho\left(\frac{B_i}{\hat{x}}\right)\right). \tag{14}$$

From equation (13), the variable ρ denotes the probability associated with the feature, the variable B_i denotes the probability associated with the corresponding bin of the feature set, and $\rho(x)$ denotes the probability of the feature. The variable $\rho(B_i/x)$ denotes the probability of the feature x in association with the bin of features B_i.

Lime is used to understand the working procedure and the internal setting for predictions using intelligent classifiers. The lime model goal is to train alternative models directly and to discover model-independent reasons for the predictions. The input I applies target n to instance x_m with likelihood y_{nm} in the lime mechanism. Lime provides two possibilities to interpret predictions for the x_m input instance: Y_1 and Y_2. Hence, Y_1 is favorably linked with a choice; on the other hand, Y_2 is adversely related to the decision. The mathematical model for the lime can be determined through the following equation:

$$\exp(x) = \min_{y \in Y} \text{Loss}\left(x, y, \pi_y\right) + \omega(y). \tag{15}$$

From equation (15), the variable "y" represents the model that is being considered for the classification over the sample that will be reduced through the loss function Loss is concerning the initial perdition value concerning model "u" with model intricacy of $\omega(y)$. The variable "y" represents the realizable explanations, and the variable π_y denotes the degree of locality concerning the sample "x" considered for the interpretation. The accuracy of the various classification models is of concern to lime. Table 6 presents the value from stroke prediction through interpretable classifiers, and Figure 5 represents the corresponding graph plotted from the accuracy values of various classification models [45].

The lime model is widely used with text, image, and tabular data, and the studies have shown a reasonable performance [4]. Table 7 presents the interpretability score associated with various classification models over the SHAP explainable model as presented in the empirical study presented by Moreno-Sanchez [46]. In the current case study, the LIME-based XAI models are used to evaluate the model's feature selection mechanism, which assists in finding the best feature set that maximizes the model's performance. It is desired to make the feature selection model evident as it is significant in making the decisions associated with the future illness prediction model.

4.2.2. Case Study 2. The importance of DL is that the black-box characteristics of these deep learning model and other pivotal considerations, like processing efforts, are making it difficult to trust the existing advanced intelligence models in the diagnosis process [47]. The fact is that, despite the existence of fundamental statistical principles, deep neural networks cannot clearly express the information associated with a specific task that is performed in the process of computer-aided diagnosis. The existing AI approaches,

Table 6: Values representing the accuracy of various classification models with lime.

Approaches	Accuracy (%)
Bayesian rule lists	75.6
Multilayer perceptron	76.4
Dempster-Shafer classifier	61.2
Recurrent neural network	66.9
Gradient descent	83.8

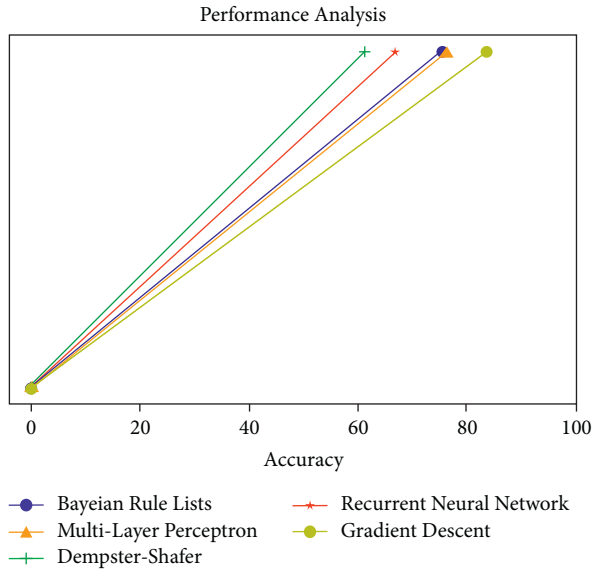

Figure 5: Graph representing the accuracy of various classification models.

such as linear regression and decision trees, are self-explanatory since the classification's decision boundary can be expressed after a few dimensions with the model's parameters [48].

Deep learning models are deemed opaque since the neurons' weights cannot be immediately interpreted as knowledge, neither the amplitude nor selectivity of activations nor their effect on network choices as adequate indicators of a neuron's significance for a particular task. There is a necessity for a thorough analysis of the underlying architecture, processes, and deterministic statics that can be achieved through explainable AI technology. There are essentially two ways to explain the findings of deep neural networks (DNN) in diagnostic imaging: those that rely on traditional attribution-based techniques and the models that depend on new architecture or domain-specific methodologies. The issue of attributing a value to every characteristic of such a network resulted in various attributing techniques. An attributing technique's objective is to ascertain the influence of an adaptive feature on the objective neuron, which is often the output neuron for the appropriate class in a classification issue. The organization of all of the other input characteristics' attributions inside the configuration of such input feature generates heatmaps referred to as attribution maps. These are the characteristics, or pixels in the case of images, that would provide a balance of contrary forces evidence of various intensities. Figure 6 presents the components of the XAI model in the computer-aided diagnosis. It can be observed that the XAI models are used in analyzing the diagnostic reports to make the results of the machine learning model to be more evident. Feature engineering performs the tasks like feature extraction and feature subset selection based on the feature weights. Later, the XAI model is deployed in the weight initialization and optimization for the features based on the value obtained by the loss function. The features assumed to be more significant in the evaluation process are given more weightage than the others. It is desired to be transparent to make sure the model is evident.

The features are associated with weights in each layer of the neural network, updated in each iteration. The accuracy of the prediction model relies on the initial weights and biases associated with each layer and their associated activation functions. The transparency of such weight approximation functions is important in making the operational procedure model more evident. In weight initialization, the features that are more significant in the prediction process are assigned more weights and optimized accordingly [49]. Initially, the weights are trained in the neural network as shown in equation (16), and the variable Int_w denotes the initialization weight associated with the feature.

$$\mathrm{Int}_w = \sum_{p=1}^{m}\sum_{q=1}^{n}\left|w_{p,q} \times w_{q,o}\right|, \tag{16}$$

where the variable $w_{p,q}$ denotes the weight linked with the network among the input node p and the corresponding hidden node q. The variable $w_{q,o}$ denotes the weights linked with the hidden node q and the output node o. One method for the most reasonable way of assessing the overall weight of features is by combining weights of all less important features equal to the elements of the entire features set, and all the corresponding weights are managed based on the significance of that feature. The inception weights of the less important features LIF_w is determined through the following equation:

$$LIF_w = \frac{1}{f_{\mathrm{tot}}}\sum_{i=1}^{f_{\mathrm{tot}}} LIF_w. \tag{17}$$

From the above equation, the variable f_{tot} denotes the total number of features that are associated with the training data. The values of these weights are refined over the iterations for better performances. The weight is optimized by considering the model-specific parameters and the associated loss functions [49]. Now, let us assume that the CAD model predicts the disease based on the input feature vector used in the training process in the current context. The featured pair (p, q) and $\{(p_v, q_v), 0 \le v \le f_{\mathrm{tot}}\}$ represent the training data. The validation set that is described as $\{(p'_v, q'_v), 0 \le v \le f_{\mathrm{tot}}\}$, which is used in fine-tuning the model. The smart diagnosis model is represented as $s_d(v, \varnothing)$, and the model-specific loss function is being recognized through $f_l(q', q)$, which is kept minimum. The loss incurred during

Table 7: Values representing the accuracy of various classification models with SHAP.

Approaches	Accuracy	Sensitivity	Specificity	Precision	F1 Score	Interpretability score
Random forest	87.6	79.2	91.6	82.0	80.4	0.50
Extra trees	87.1	79.3	90.8	80.7	79.9	0.58
Ada boost	85.2	75.1	90.1	78.7	76.4	0.50
Gradient boosting	84.7	78.0	88.0	76.0	76.6	0.50
XGBoost	87.1	79.2	90.8	80.6	80.6	0.25
Max voting	86.6	76.3	91.5	81.6	81.6	0.58

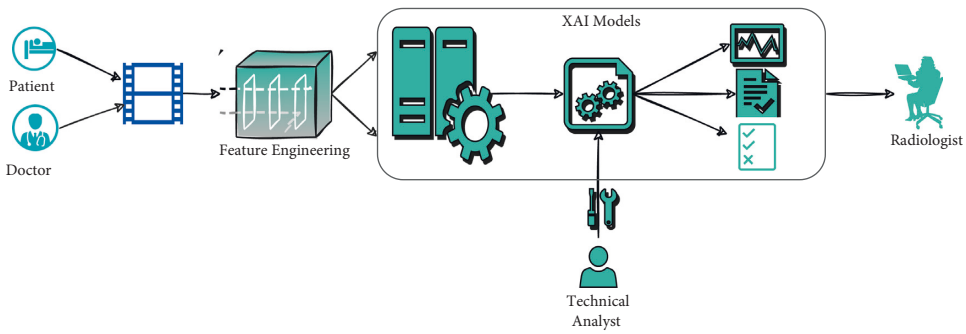

Figure 6: Image representing the role of XAI in CAD applications.

the training process is approximated using the variable tr_l that is elaborated in the following equation:

$$tr_l = \frac{1}{f_{tot}} \sum_{v=1}^{f_{tot}} f_l(q', q) = \frac{1}{f_{\text{tot}}} \sum_{v=1}^{f_{\text{tot}}} f_q(\varnothing), \tag{18}$$

where the variable $f_q(\varnothing)$ denotes the loss function, and the associated weights are fine-tuned to reduce the weight loss as shown in equation (19), and the variable $R(\varnothing)$ denotes the loss function.

$$R(\varnothing) = \varnothing' \sum_{q=1}^{f_{tot}} w_q f_q(\varnothing). \tag{19}$$

The variable w_q denotes the weight associated with the feature q, which is being updated through $\{w_q\}_{q=1}^{f_{\text{tot}}}$ using the training and validation components. The loss function associated with the weight of the features in training data is determined by equation (20), which is desired to be the lowest.

$$\widehat{w} = \frac{1}{f_t} \sum_{q=1}^{f_{\text{tot}}} f_q'(\varnothing \times w). \tag{20}$$

The weight optimization process is done throughout the training process, the value is fine-tuned, and the values are recommended to be greater than zero for a robust diagnosis model. The experimental results out of the XAI-driven smart diagnosis evaluated by Khodabandehloo et al. [50] are presented in Table 8.

It is a known fact that medical imaging is relatively of low cost and perceived usefulness, that is, the bulk of the medical imaging research that has examined the explainability of deep learning techniques utilized attribution-based approaches. Experts may train the appropriate neural network model without adding complexity by rendering it intrinsically explainable and using an existing attribution model. It enables the adoption of the current deep neural network, perhaps one built from scratch, to get reasonable accuracy. The conventional model framework enables the use of methods such as transfer learning. In contrast, the latter allows the emphasis on detailed information and avoids overfitting through fewer attributes. The model evaluation utilizes attributions to show whether the network is acquiring meaningful features or overfitting the input data by learning false features. It enables researchers to fine-tune the model framework and hyperparameters to obtain superior performance on validation samples, and subsequently, in a future real-world scenario. In the current case study, the process of initial weight assessment and the evaluations associated with the loss function is presented to make the decision more evident. The hyperparameter tuning is significant in analyzing the underfitting and overfitting scenarios related to the training data. The explainable weight assessment, updating, and optimization scenarios concerning the loss function are presented in the case study.

4.2.3. Case Study 3. The technology of Explainable AI for Earlier Warning Score (XAI-EWS) [5, 51, 52] consists of a strong and effective artificial intelligence model for forecasting critical diseases using electronic healthcare-related information. XAI-EWS model can be created to offer straightforward visual explanations for the predictions made. The XAI-EWS allowed for evaluating model interpretations from two perspectives: an individual-level and a population-level viewpoint. Individually, the explanation component enables the XAI-EWS to determine whichever clinical factors were significant at a particular moment in time for a specific forecast. In contemporary clinical practice,

Table 8: The performances of various diagnostic models concerning XAI-driven features.

Approaches	Accuracy	Precision	Recall	F1 Score
Nearest neighbourhood	49.9	49.7	49.9	49.0
Bayes classifier	58.7	59.0	58.7	58.5
C4.5	56.4	56.7	56.4	54.7
Neural network	58.4	58.4	58.4	57.6
Random forest	57.5	57.5	57.5	56.9
SVM	58.9	59.0	58.9	58.6
Decision Table(XAI)	60.1	60.0	60.1	59.4

physicians often notice either a high EWS or a rise in EWS. However, when the physician knows which, clinical factors contributed to the elevated EWS or change in EWS, the following focused therapeutic action about the possible critical disease occurs. This is one of the primary motivations for AI-powered EWS systems to rationalize such forecasts [53]. Figure 7 represents the XAI-based EWS model using neural network models to forecast the illness.

In the current case study, the Earlier Warning Scores are primarily dependent on the probabilistic measures associated with the SoftMax layer of the proposed model. The probabilities are exceedingly significant in the appropriate measure of the warning score. The transparency in the probabilistic assessment would help the model to be more evident in the evaluation process. The fundamental concept behind Batch processing is to regulate distribution for every hidden layer that assists in increasing the learning rate of the model. Furthermore, the Batch processing is performed with each convolutional layer of the assessment model. Each hidden layer's feature distribution constantly updates in response to parameter changes, and the distribution progressively updates the activation functions [54–56].

Batch processing would boost the variance of output feature gradients since output feature distributions generally possess higher gradients. Moreover, when the learning rate of the model is faster than the training phase of the model, it would be much faster with minimal effort. The formula for the same is determined through the following equation:

$$f'^{(x)} = \frac{f^{(x)} - E(f^x)}{\sqrt{v(f^{(x)})}}. \tag{21}$$

From equation (21), the variable $f^{(x)}$ denotes the outcome of the previous layer, where the outcome of each layer would yield a normal distribution with a variance of 1 and a mean of 0. To attain the nonlinearity in the batch processing, regarding two additional parameters, scale and shift, the resultant outcome $O^{(x)}$ is shown in the following equation:

$$O^{(x)} = p^{(x)} f'^{(x)} + q^{(x)}, \tag{22}$$

where the variables $p^{(x)}, q^{(x)}$ denote the scale and shift, respectively, which would yield the nonlinearity of the model. They would, with a better nonlinearity, yield a better outcome. The cross-entropy is determined by the softmax layer function, which is a.

The performance of the XAI-EWS is being assessed through 95% of the confidence interval, and the cross-validation of the model is being presented in Table 9 for both Area Under the Receiver Operating Characteristic Curve (AUROCC) and Area Under the Precision-Recall Curve (AUPRC) for the three diseases like sepsis, acute kidney injury, and lung injury.

It can be observed in Table 7 that the performance of the XAI-EWS model is reasonably good compared to that of the other traditional models. XAI-EWS demonstrates strong prediction accuracy, allowing physicians to justify the predictions by identifying critical input data. The feature engineering tasks like the feature selection and the batch processing are brought under the explainable concepts that make the model's predictions more evident and authentic. The performance of the XAI-driven model is on par with the conventional models, and it can be seen from the table above that the performance of the XAI-driven models is far better than the conventional black-box models.

5. Discussion

In the recent revolutionary advancement in artificial intelligence and soft computing models, the models face various challenges that prevent the widespread adoption of AI in some applications. A few challenges encompass many different components and the rapid energy rate required for today's powerful machine intelligent models. Lack of accountability and clarity about the machine intelligent models' decision-making impacts the AI system's confidence.

Several of the most efficient machine learning systems' properties seem suspicious in making the decisions [57]. Various approaches are being used in the diagnosis of the abnormality from the imaging technology [58–60], volumetric assessments, surgical procedure planning [61, 62], illness classification models [48, 63, 64], and predictive illness models [65, 66] that are efficient and reasonably good in producing the promising outcome. The judgments of the proposed models need to be transparent through the interpretable operational models. There are various machine learning techniques such as support vector machine, logistic regression, random forest, k-nearest neighborhood, Naive Bayes, deep learning models, cognitive techniques, Reinforcement learning models, and bioinspired optimization algorithms interpreting the internal functionality of each of those models. The major limitation to utilizing these sophisticated and backbox frameworks is that people lack confidence and trust in their forecasts. Figure 8 presents the generic architecture of the XAI model in concern to the explainable model, explainable function, and explainable interface.

The linear models are designed and developed over the linearly regulated systems that map the target values' feature values. The models are applied in the divergent fields of technology. Certain approaches produce the system's internal components and descriptive statistics in linear model weights that determine the specific outcome's probability. If characteristics are not usually distributed, these probabilities

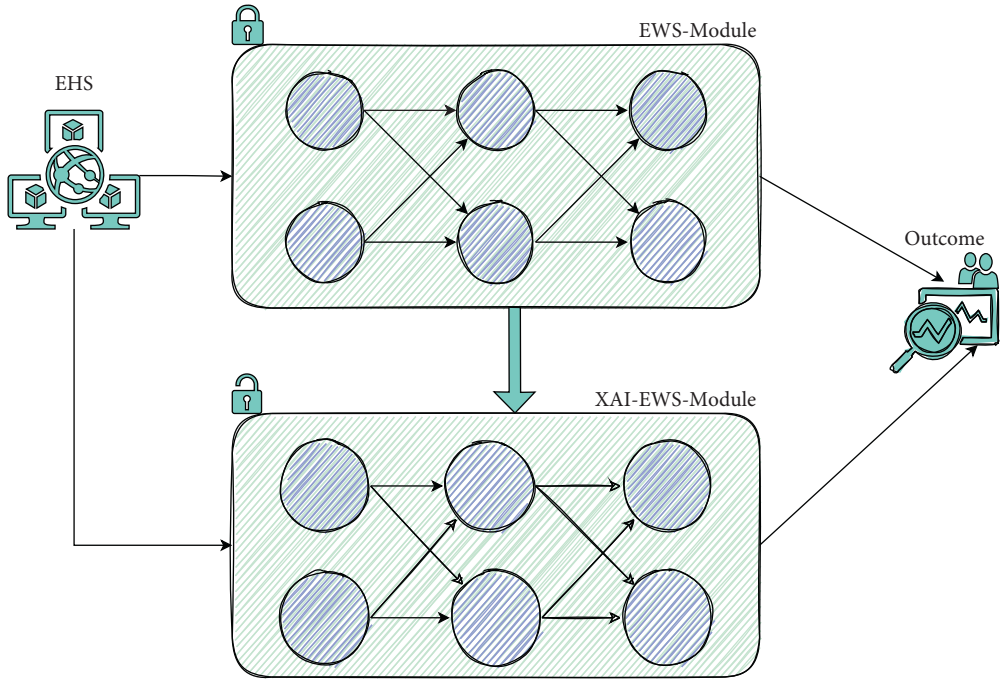

Figure 7: Image representing the neural network-based EWS-XAI in prediction models.

Table 9: Table representing the cross-validation of the XAI-EWS models.

Mechanisms	Cross-validation
AUROCC	0.92
	0.80
	0.88
	0.79
	0.90
AUPRC	0.43
	0.08
	0.22
	0.14
	0.23

would be calculated according to the sample size. For function value, the weights in a linear model are used to recognize the parameters and attributes that are globally or locally significant for accurate prediction. The linearity of the estimation method simplifies the estimation process and, most specifically, simplifies the insight into the association among variables and the weight administration [67]. On the other hand, the nonlinear models, including the Random Forest, Quadratic Discrimination analysis, Decision Tree K-Nearest Neighbors, and neural networks, perform well in divergent computer science fields, accurately classifying the data and making predictions. Though no coefficient reflects the difference in the performance of an objective function approximation, nonlinear and monotonic functions typically move in the same direction when an input variable is changed. Typically, they increase the possibilities of plots explaining their actions and explanation codes, and various statistical steps. Thus, nonlinear, monotonic objective functions are easily interpretable and used in supervised applications.

The decomposition of the multidimensional classification model methodology provides interpretability. The layer specifications, and the parameters, whether in terms of their count, scalability, or variance, are often chosen dynamically in a heuristic manner and are not guided through knowledge; thus, these decisions are not clear from the architecture point of view and interpretable. The selection of hyperparameters such as learning rate and batch size is more heuristic and opaquer. Deep medicine is based on three main processes: deep phenotyping, deep learning, deep empathy, and emotional interaction [68]. To fully exploit deep learning approaches implemented in medical studies, algorithms must be resilient to missing and incorrect values and capable of dealing with highly variable-sized datasets and long-term dependencies associated with human diagnoses, treatments, assessments, and drug prescriptions. Figure 9 presents the various categories of the models that are part of artificial intelligence.

The ensemble models reasonably predict predictions, classification, and future forecasting of illnesses [67]. Ensembles may be highly effective for mitigating prejudice because of the more classifiers (diverse and equal). Under this premise, an ensemble's discrimination sensitivity may be adjusted by varying the diversity of such classifiers. In contrast, the trade-off between consistency and discrimination in explainability is determined by the number of conflicts among the classification models and the amount of inappropriately classified instances [69–71]. Nonlinear models created by training-based artificial intelligence approaches allow more reliable predictions on previously unknown data. This usually results in increased financial

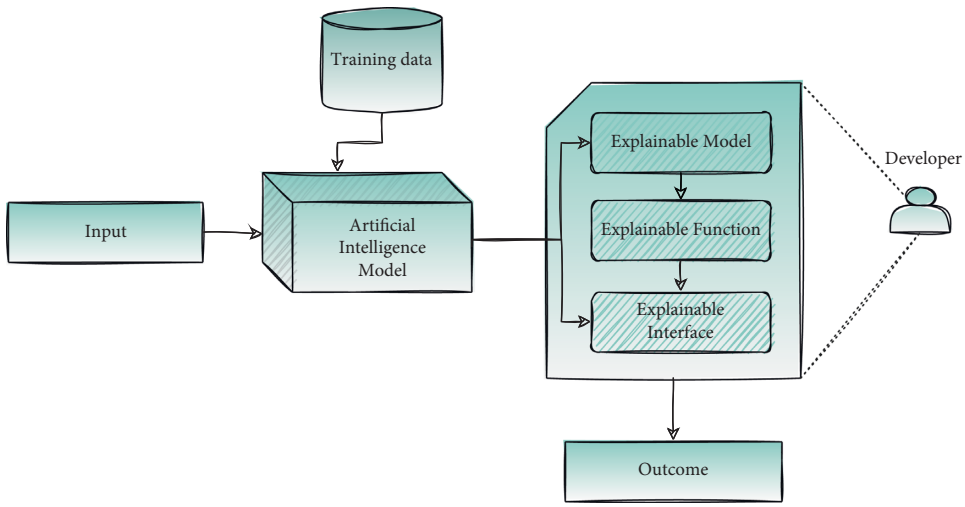

Figure 8: Image representing the generic architecture of the XAI model.

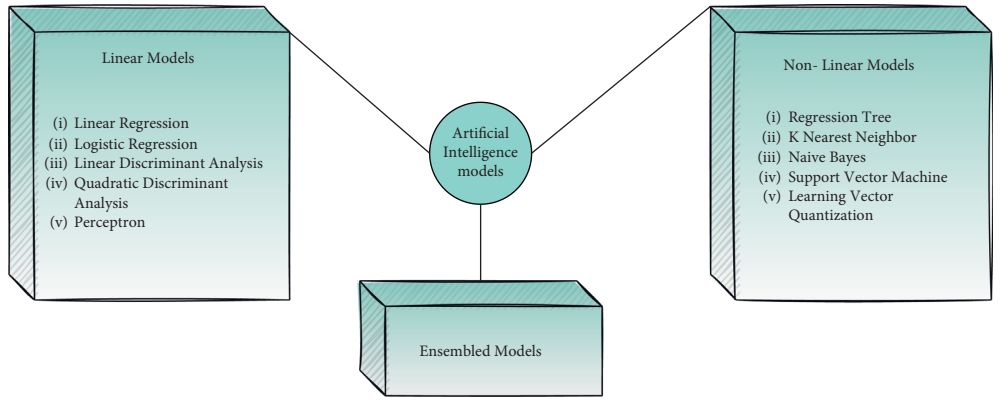

Figure 9: Image representing the various machine intelligent models.

margins, but only if internal validation teams adopt the model, business partners, and consumers. Interpretable artificial intelligent models and strategies for debugging, description, and fairness may help build awareness and confidence in newer or more rigorous machine learning methods, allowing for more complex and theoretically more reliable models instead of previously used linear models.

Hyperparameters are the other significant and fine-tuning components associated with supervised learning algorithms. Various elements like training loss, training accuracy, testing loss, testing accuracy, and learning rate [72] are the components that are adjusted for a better outcome. The models are modulated using weight and bias optimization for optimal performance. It is desired to make the model interpretable concerning the hyperparameters so that the internal operations of the model decision are transparent. It is highly desired that the decision mechanism for the healthcare models must be explainable and transparent to be more evident.

6. Future Perspective of XAI

In biomedical engineering and healthcare informatics, the pivotal agenda of the explainable Artificial Intelligence technology is to make humans aware of the working procedure of the prototype through transparency. The human-computer interaction is the other technology that has led the explainability of the AI models to an entirely new pace. The model-specific representation approaches provide insight into the model's internal parameters through analysis. Model agnostic methods may be generalized to any machine learning algorithms and thus are usually applied post hoc. The internal working framework and model parameters of black-box models are often overlooked. Typically, model agnostic interpretability is accomplished by studying a trustworthy evaluation of a complex black-box model from its outputs through a substitute or a simple representation model: the future technology of adhering to the explainability of such models in the clinical decision-making models and the healthcare analytical models. There is a significant demand for models that operate with negligible data. These models need to work over unstructured data like medical images, healthcare records, and the signal data like Electrocardiogram (ECG), electrocardiogram (EEG), and Electromyography (EMG).

The future illness prediction models are primarily dependent on the nonlinear classification models in approximating the illness based on the probabilistic approach. The models may improvise the visibility and

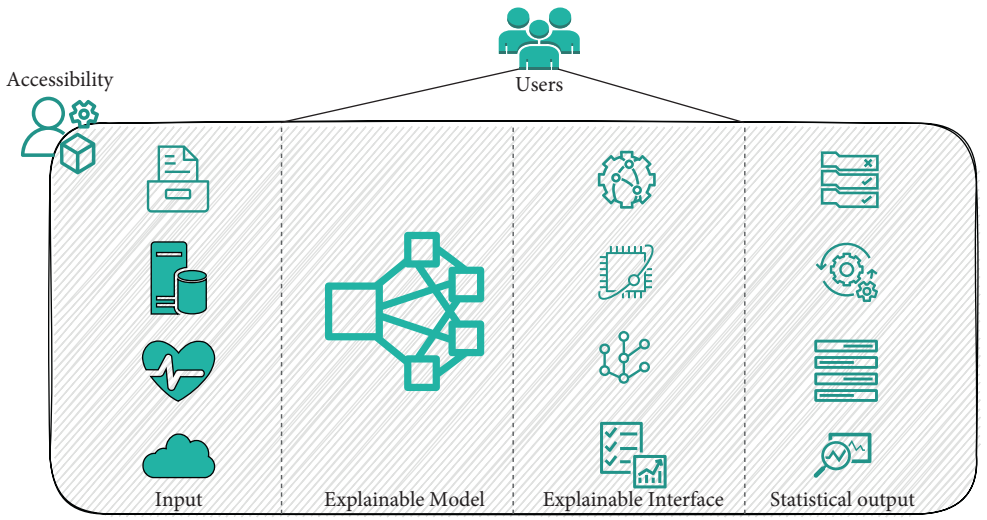

Figure 10: Image representing the explainable models and explainable interfaces.

the comprehensibility of the model concerning the feature selection models, weight assignment, bias assessment, and probabilistic evaluations. This selection of feature vectors that results from eliminating noisy, repetitive, and irrelevant features contributes to the problem's interpretability to some degree. Acquiring and reviewing obsolete but valuable functions may be essential for a complete view of various issues. Data acquisition, data preprocessing, data preparation, model building, and visualization are the vital steps in divergent fields of healthcare informatics used in predicting the illness category and future model risk analysis, as shown in Figure 10. At the dataset stage, function significance scores attempt to quantify the contribution of features to forecast within a framework. These ratings will focus on the model features defined as critical for outcomes and their relative significance.

Domain-specific requirements must be considered, such as a firm understanding of the system's purpose, the efficiency, and explainability of current systems, and the degree and complexity of the explanations desired. Artificial Intelligence and Machine Learning (AI and ML) are used in cutting-edge healthcare equipment. XAI is a collection of frameworks and tools for better understanding and interpreting predictions from machine learning (ML). In the healthcare and wellness market, XAI and Machine intelligent models are attracting a lot of attention because of their ability to uncover and anticipate previously undiscovered trends. XAI's current concepts and ideas in health care system building may permit more advanced algorithms, architectures, and sustainable strategies to be another future objective. Domain-specific requirements in the healthcare domain, such as a complete knowledge of a system's function, performance, and interpretability operations associated with the decision support system for medical assistance and the robotic surgery models, must be considered. That has paved a great research scope, and the field needs an empirical study that would assist in more interpretability of the intelligent models.

In some cases, the models need to work with inadequate data due to the lack of availability of the preexisting studies or the occurrence, resulting in the adoption of reinforcement learning, self-learning models, and weakly-trained models. All such models with inadequate data must be robust in prediction and unbiased to any internal or external dependencies [3]. There are divergent studies that precisely performed DNA-based analysis to predict the abnormality [73]. While this model is capable of strong predictive and descriptive precision, it is critical to keep in mind that the lack of explainability inherent in these models limits their utility. A series of self-learning and poorly trained interpretability techniques have been created to assist with such a process to provide insight into how the supervised model has been implemented without altering the underlying model. These techniques are especially useful in situations where the obtained data is multidimensional and dynamic, such as Signal data like ECG, EMG, EEG, medical imaging data, and the Internet of Things (IoT) based healthcare records of individuals. In these contexts, explainable approaches must contend that individual features lack contextual significance, rendering the task more difficult than on datasets with more meaningful features [74]. The models used in the IoT-based healthcare monitoring are demanding for the interpretability models in determining the internal operations of the model in concern to the scalability, privacy, and the authenticity of decisions by the model [75, 76].

7. Conclusion and Feature Scope

The trade-off between precision and explainability is evident in how confidence regulates AI adoption in sensitive healthcare and biomedical engineering domains. Although interpretable and explainable, AI algorithms have numerous advantages, including increased confidence and transparency for the application and the working frameworks. Moreover, explainability is not a

built-in feature of the AI-driven models, which must be mechanized following the situation. The requirement for the rationale for an output usually increases in proportion to the outcome's efficiency [77]. Moreover, considering the immediate costs levied on the model's explainability, the XAI has long-term legal advantages that make it investment-wise for healthcare innovators. Evidence of explainability's utility is still missing in practice and acknowledges that complementary interventions will be required to construct trustworthy AI.

Despite various advantages, training the proposed framework needs more labeling components to the data than other existing models, such as instance-level bounding box and mask labeling, restricting our method's scalability to large-scale generation problems. To address such challenges, a variety of interesting future research avenues exist. One possibility is to conduct an unsupervised analysis of intermediate semantic constructs via end-to-end instruction. Injecting the functional inductive bias into the construct by appropriately regularizing the intermediate output structures leads it to discover concrete structures from the data. It is concluded that omitting explainability from diagnostic and treatment structures jeopardizes fundamental medical ethical principles and can negatively impact human and global safety. Various limitations are associated with the XAI technology, and especially the lack of expertise in the technology, randomness, nonlinearity of the data, and context-dependency are a few of those challenges associated with the model to incorporate the XAI technology. The healthcare applications used in clinical assistance and decision-making with dependency inferences make it challenging to deal with the interpretability and explainability of the models.

The XAI models might be recommended to work under the control environment with clinical protocols, which need a human-machine intervention to a large extent. Clinical functions are often dynamic, and traditional statistical success appraisal metrics make integrating and measuring all beneficial properties in a model difficult. Explanations can assist by encouraging a human to be included in the process to detect and fix difficulties in some cases. Such models can be mechanized and fully independent in decision-making in a controlled environment without much human involvement. Explainable modeling is the concept where the user's functional and operational model is understandable. Technologies like reinforcement techniques, cognitive models, and self-learning models are intrinsically interpretable and considered a pivotal model in the model future of healthcare informatics.

Data Availability

Not Applicable.

Conflicts of Interest

The authors express no conflicts of interest.

References

[1] J. Amann, A. Blasimme, E. Blasimme, D. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 310, 2020.

[2] A. Holzinger, "Explainable AI and multi-modal causability in medicine," *I-Com*, vol. 19, no. 3, pp. 171–179, 2021.

[3] S. J. Schrodi, S. Mukherjee, Y. Shan et al., "Genetic-based prediction of disease traits: prediction is very difficult, especially about the futureâ€," *Frontiers in Genetics*, vol. 5, 2014.

[4] S. M. Lauritsen, M. Kristensen, M. V. Olsen et al., "Explainable artificial intelligence model to predict acute critical illness from electronic health records," *Nature Communications*, vol. 11, no. 1, p. 3852, 2020.

[5] S. Muralitharan, W. Nelson, S. Di, M. McGillion, P. Devereaux, and N. Barr, "Petch JMachine learning–based early warning systems for clinical deterioration: systematic scoping review," *Journal of Medical Internet Research*, vol. 23, no. 2, Article ID e25187, 2021.

[6] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk, "Potential biases in machine learning algorithms using electronic health record data," *JAMA Internal Medicine*, vol. 178, no. 11, pp. 1544–1547, 2018.

[7] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.

[8] M. A. Ahmad, C. Eckert, and A. Teredesai, "interpretable machine learning in healthcare," in *Proceedings of the 2018 ACM international conference on bioinformatics, Computational Biology, and Health Informatics*, pp. 559-560, Washington, DC, USA, August 2018.

[9] O. Loyola-Gonzalez, "Black-box vs. White-box: understanding their advantages and weaknesses from a practical point of view," *IEEE Access*, vol. 7, pp. 154096–154113, 2019.

[10] J. Chen, X. Dai, Q. Yuan, C. Lu, and H. Huang, "Towards interpretable clinical diagnosis with bayesian network ensembles stacked on entity-aware CNNs," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3143–3153, Stroudsburg, PA, USA, July 2020.

[11] M. Rucco, G. Viticchi, and L. Falsetti, "Towards personalized diagnosis of glioblastoma in fluid-attenuated inversion recovery (FLAIR) by topological interpretable machine learning," *Mathematics*, vol. 8, no. 5, p. 770, 2020.

[12] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," *Scientific Reports*, vol. 10, no. 1, Article ID 19549, 2020.

[13] S. El-Sappagh, J. M. Alonso, S. M. R. Islam, A. M. Sultan, and K. S. Kwak, "A multi-layer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease," *Scientific Reports*, vol. 11, pp. 1–26, 2021.

[14] S. Sarp, M. Kuzlu, E. Wilson, U. Cali, and O. Guler, "The enlightening role of explainable artificial intelligence in chronic wound classification," *Electronics*, vol. 10, no. 12, p. 1406, 2021.

[15] P. Sabol, P. Sinčák, P. Hartono et al., "Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images," *Journal of Biomedical Informatics*, vol. 109, Article ID 103523, 2020.

[16] T. K. Yoo, I. H. Ryu, H. Choi et al., "Explainable machine learning approach as a tool to understand factors used to select the refractive surgery technique on the expert level," *Translational Vision Science & Technology*, vol. 9, no. 2, pp. 8–14, 2020.

[17] D. Das, J. Ito, T. Kadowaki, and K. Tsuda, "An interpretable machine learning model for diagnosis of Alzheimer's disease," *PeerJ*, vol. 7, Article ID e6543, 2019.

[18] P. R. Magesh, R. D. Myloth, and R. J. Tom, "An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery," *Computers in Biology and Medicine*, vol. 126, Article ID 104041, 2020.

[19] J. Cho, A. Alharin, Z. Hu, N. Fell, and M. Sartipi, "Predicting post-stroke hospital discharge disposition using interpretable machine learning approaches," in *Proceedings of the 2019 IEEE International Conference on Big Data (Big Data)*, pp. 4817–4822, Los Angeles, CA, USA, December 2019.

[20] A. Lakhan, M. A. Mohammed, J. Nedoma et al., "Federated-learning based privacy preservation and fraud-enabled blockchain IoMT system for healthcare," *IEEE Journal of Biomedical and Health Informatics*, p. 1, 2022.

[21] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging*, vol. 6, no. 6, p. 52, 2020.

[22] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. Hansen, and KR. Müller, Eds., Springer, Cham, pp. 5–22, 2019.

[23] K. H. Abdulkareem, S. A. Mostafa, Z. N. Al-Qudsy et al., "Automated system for identifying COVID-19 infections in computed tomography images using deep learning models," *Journal of Healthcare Engineering*, vol. 2022, Article ID 5329014, 13 pages, 2022.

[24] Z. C. Lipton, "The mythos of model interpretability," *ACM Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[25] I. Lage, E. Chen, J. He et al., "Human evaluation of models built for interpretability," *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, no. 1, pp. 59–67, 2019.

[26] A. Malhi, S. Knapic, and K. Främling, "Explainable agents for less bias in human-agent decision making," in *Explainable, Transparent Autonomous Agents and Multi-Agent Systems. EXTRAAMAS 2020*, D. Calvaresi, A. Najjar, M. Winikoff, and K. Främling, Eds., Springer, Cham, pp. 129–146, 2020.

[27] A. Abdollahi and B. Pradhan, "Urban vegetation mapping from aerial imagery using explainable AI (XAI)," *Sensors*, vol. 21, no. 14, p. 4738, 2021.

[28] J. Jiarpakdee, C. K. Tantithamthavorn, H. K. Dam, and J. Grundy, "An empirical study of model-agnostic techniques for defect prediction models," *IEEE Transactions on Software Engineering*, vol. 48, no. 1, pp. 166–185, 2022.

[29] R. Marco, S. Singh, and G. Carlos, "Model-agnostic interpretability of machine learning," 2016, https://doi.org/10.48550/arXiv.1606.05386.

[30] A. Vij and P. Nanjundan, "Comparing strategies for post-hoc explanations in machine learning models," in *Mobile Computing and Sustainable Informatics*, S. Shakya, R. Bestak, R. Palanisamy, and K. A. Kamel, Eds., Springer, Singapore, pp. 585–592, Lecture Notes on Data Engineering and Communications Technologies, 2022.

[31] K. Zhang, P. Xu, and J. Zhang, "Explainable AI in deep reinforcement learning models: a SHAP method applied in power system emergency control," in *Proceedings of the 2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2)*, pp. 711–716, Wuhan, China, November 2020.

[32] J. N. Hasoon, A. H. Fadel, R. S. Hameed et al., "COVID-19 anomaly detection and classification method based on supervised machine learning of chest X-ray images," *Results in Physics*, vol. 31, Article ID 105045, 2021.

[33] T. De, P. Giri, A. Mevawala, R. Nemani, and A. Deo, "Explainable AI: a hybrid approach to generate human-interpretable explanation for deep learning prediction," *Procedia Computer Science*, vol. 168, pp. 40–48, 2020.

[34] R. Confalonieri, W. Tillman, R. Tarek, and Besold and Fermín Moscoso del Prado Martín, "Trepan reloaded: a knowledge-driven approach to explaining artificial neural networks," 2019, https://doi.org/10.48550/arXiv.1906.08362.

[35] B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model," *Complexity*, vol. 2021, Article ID 6634811, 11 pages, 2021.

[36] 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 *Proceedings of the SPIE 11314, Medical Imaging 2020: Computer-Aided Diagnosis*, p. 113140Z, Houston, Texas, USA, 16 March 2020.

[37] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, and M. Hind, "One explanation does not fit all: a toolkit and taxonomy of ai explainability techniques," 2019, https://arxiv.org/abs/1909.03012.

[38] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: a review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[39] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: a survey on methods and metrics," *Electronics*, vol. 10, no. 5, p. 593, 2021.

[40] H. Quinn, *Explaining Decisions Made with AI: Draft Guidance for Consultation—Part 1: The Basics of Explaining AI*, ICO & The Alan Turing Institute: Wilmslow/Cheshire, UK, 2019.

[41] P. Chen, W. Dong, J. Wang, X. Lu, U. Kaymak, and Z. Huang, "Interpretable clinical prediction via attention-based neural network," *BMC Medical Informatics and Decision Making*, vol. 20, no. S3, p. 131, 2020.

[42] S. Penafiel, N. Baloian, H. Sanson, and J. A. Pino, "Predicting stroke risk with an interpretable classifier," *IEEE Access*, vol. 9, pp. 1154–1166, 2021.

[43] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES 20)*, pp. 180–186, Association for Computing Machinery, New York, NY, USA, February 2020.

[44] J. Wankhede, M. Kumar, and P. Sambandam, "Efficient heart disease prediction-based on optimal feature selection using DFCSS and classification by improved Elman-SFO," *IET Systems Biology*, vol. 14, no. 6, pp. 380–390, 2020.

[45] A. M. Antoniadi, Y. Du, Y. Guendouz et al., "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review," *Applied Sciences*, vol. 11, no. 11, p. 5088, 2021.

[46] P. A. Moreno-Sanchez, "Improvement of a prediction model for heart failure survival through explainable artificial intelligence," 2021, https://arxiv.org/abs/2108.10717.

[47] A. J. London, "Artificial intelligence and black-box medical decisions: accuracy versus explainability," *Hastings Center Report*, vol. 49, no. 1, pp. 15–21, 2019.

[48] P. N. Srinivasu, J. G. SivaSai, M. F. Ijaz, A. K. Bhoi, W. Kim, and J. J. Kang, "Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM," *Sensors*, vol. 21, no. 8, p. 2852, 2021.

[49] M. I. Hertzog, U. Brisolara Correa, and R. M. Araujo, "SpreadOut: a kernel weight initializer for convolutional neural networks," in *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, Budapest, Hungary, July 2019.

[50] E. Khodabandehloo, D. Riboni, and A. Alimohammadi, "HealthXAI: collaborative and explainable AI for supporting early diagnosis of cognitive decline," *Future Generation Computer Systems*, vol. 116, pp. 168–189, 2021.

[51] P. N. Srinivasu, G. JayaLakshmi, R. H. Jhaveri, and S. P. Praveen, "Ambient assistive living for monitoring the physical activity of diabetic adults through body area networks," *Mobile Information Systems*, vol. 2022, Article ID 3169927, 18 pages, 2022.

[52] I. Ben Ida, M. Balti, S. Chabaane, and A. Jemai, "Self-adaptative early warning scoring system for smart hospital," in *The Impact of Digital Technologies on Public Health in Developed and Developing Countries. ICOST 2020*, M. Jmaiel, M. Mokhtari, B. Abdulrazak, H. Aloulou, and S. Kallel, Eds., pp. 16–27, Springer, Cham, Lecture Notes in Computer Science, 2020.

[53] M. Jemmali, M. Denden, W. Boulila, G. Srivastava, R. H. Jhaveri, and T. Reddy Gadekallu, "Novel model based on window-pass preferences for data-emergency-aware scheduling in computer network," in *Proceedings of the IEEE Transactions on Industrial Informatics*, IEEE, February 2022.

[54] Z. A. A. Alyasseri, M. A. Al-Betar, I. A. Doush et al., "Review on COVID -19 diagnosis models based on machine learning and deep learning approaches," *Expert Systems*, vol. 39, no. 3, Article ID e12759, 2021.

[55] L. Munkhdalai, K. H. Ryu, O.-E. Namsrai, and N. Theera-Umpon, "A partially interpretable Adaptive softmax regression for credit scoring," *Applied Sciences*, vol. 11, no. 7, p. 3227, 2021.

[56] Q. Zhu, Z. He, T. Zhang, and W. Cui, "Improving classification performance of softmax loss function based on scalable batch-normalization," *Applied Sciences*, vol. 10, no. 8, p. 2950, 2020.

[57] D. Lee and S. N. Yoon, "Application of artificial intelligence-based technologies in the healthcare industry: opportunities and challenges," *International Journal of Environmental Research and Public Health*, vol. 18, no. 1, p. 271, 2021.

[58] M. M. Badža and M. Č Barjaktarović, "Segmentation of brain tumors from MRI images using convolutional autoencoder," *Applied Sciences*, vol. 11, no. 9, p. 4317, 2021.

[59] P. Naga Srinivasu, T. Srinivasa Rao, and V. E. Balas, "A systematic approach for identification of tumor regions in the human brain through HARIS algorithm," *Deep Learning Techniques for Biomedical and Health Informatics*, Academic Press, pp. 97–118, Cambridge, MA, USA, 2020.

[60] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, "Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–28, 2022.

[61] M. P. Paing, S. Tungjitkusolmun, T. H. Bui, S. Visitsattapongse, and C. Pintavirooj, "Automated segmentation of infarct lesions in T1-weighted MRI scans using variational mode decomposition and deep learning," *Sensors*, vol. 21, no. 6, p. 1952, 2021.

[62] P. N. Srinivasu, A. K. Bhoi, R. H. Jhaveri, G. T. Reddy, and M. Bilal, "Probabilistic Deep Q Network for real-time path planning in censorious robotic procedures using force sensors," *Journal of Real-Time Image Processing*, vol. 18, no. 5, pp. 1773–1785, 2021.

[63] Y. Gupta, R. K. Lama, S. W. Lee, and G. R. Kwon, "An MRI brain disease classification system using PDFB-CT and GLCM with kernel-SVM for medical decision support," *Multimedia Tools and Applications*, vol. 79, no. 43-44, pp. 32195–32224, 2020.

[64] J. G. SivaSai, P. N. Srinivasu, M. N. Sindhuri, K. Rohitha, and S. Deepika, "An automated segmentation of brain MR image through fuzzy recurrent neural network," in *Bio-inspired Neurocomputing*, A. Bhoi, P. Mallick, CM. Liu, and V. Balas, Eds., pp. 163–179, Springer, Singapore, 2021.

[65] S. Vadupu, K. S. Kandala, A. Peddi, N. S. Yadav, G. V. Kumar, and P. A. Harsha Vardhini, "Skin pathology detection using artificial intelligence," in *Proceedings of the 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, pp. 373–376, Solan, India, October 2021.

[66] K. K. Srinivas, P. Vangara, R. Thiparapu, R. Sravanth Kumar, and K. A. Bhagavathi, "Artificial intelligence based forecasting techniques for the covid-19 pandemic," in *Proceedings of the 2022 International Mobile and Embedded Technology Conference (MECON)*, pp. 297–301, Noida, India, March 2022.

[67] T. Kim, S. Sharda, X. Zhou, R. M. Pendyala, and Ram, "A stepwise interpretable machine learning framework using linear regression (LR) and long short-term memory (LSTM): city-wide demand-side prediction of yellow taxi and for-hire vehicle (FHV) service," *Transportation Research Part C: Emerging Technologies*, vol. 120, Article ID 102786, 2020.

[68] A. Shaban-Nejad, M. Michalowski, and D. L. Buckeridge, "Explainability and interpretability: keys to deep medicine," in *Explainable AI in Healthcare and Medicine*, A. Shaban-Nejad, M. Michalowski, and D. L. Buckeridge, Eds., pp. 1–10, Springer, Cham, 2021.

[69] D. Ramesh and Y. S. Katheria, "Ensemble method based predictive model for analyzing disease datasets: a predictive analysis approach," *Health Technology*, vol. 9, no. 4, pp. 533–545, 2019.

[70] H. Jiang, K. Yang, M. Gao, D. Zhang, H. Ma, and W. Qian, "An interpretable ensemble deep learning model for diabetic retinopathy disease classification," in *Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2045–2048, Berlin, Germany, 2019 July.

[71] P. Guleria, S. Ahmed, A. Alhumam, and P. N. Srinivasu, "Empirical study on classifiers for earlier prediction of COVID-19 infection cure and death rate in the Indian states," *Healthcare*, vol. 10, no. 1, p. 85, 2022.

[72] A. Vulli, P. N. Srinivasu, M. S. K. Sashank, J. Shafi, J. Choi, and M. F. Ijaz, "Fine-tuned DenseNet-169 for breast cancer metastasis prediction using FastAI and 1-cycle policy," *Sensors*, vol. 22, no. 8, p. 2988, 2022.

[73] M. Smith, "DNA sequence analysis in clinical medicine, proceeding cautiously," *Frontiers in Molecular Biosciences*, vol. 4, p. 24, 2017.

[74] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22071–22080, 2019.

[75] L. K. Hansen and L. Rieger, "Interpretability in intelligent systems—a new concept?," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. Hansen, and KR. Müller, Eds., pp. 41–49, Springer, Cham, Lecture Notes in Computer Science, 2019.

[76] A. Barredo Arrieta, J. Del Ser, A. Bennetot et al., "Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[77] R. Hoffman, S. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: challenges and prospects," 2018, https://arxiv.org/abs/1812.04608.