Application of explainable AI to healthcare: a review ⋆

Samuel Gbenga Faluyi, Yousra Chabchoub, Maurras Ulbricht Togbe, Jérémie Sublime

Application of explainable AI to healthcare: a review*

Samuel Gbenga Faluyi [a,*,†], Yousra Chabchoub [a,†], Maurras Togbe [a,†], Jérémie Sublime [a,†]

Isep, 10 rue de Vanves 92130 Issy Les Moulineaux, France

Abstract

The world of technology is advancing by the day, presenting innovative and efficient solutions across various sectors, with healthcare being no exception. This review study majorly focuses on eliciting the impact of machine learning and deep learning techniques to improve the delivery of healthcare. It investigates the different frameworks of previous research studies to establish facts regarding the application of machine learning and deep learning, as well as where enhancement of the model is required. The strengths and weaknesses of the techniques used are identified. Our review study shows that the impact of machine learning and deep learning techniques cannot be berated, notably in prediction modelling, pattern recognition, classification, regression, and image processing, among other applications of the models. Furthermore, the study identifies numerous benefits of model explainability and different model explanation techniques, such as Alibi, InterpretML, Explainerdashboard, etc. We also show that prospective studies could employ ensemble learning using boosting and deep learning algorithms as core learning units.

Keywords

Healthcare, Machine learning, Deep learning, Boosting algorithms, Model Explainability

1. Introduction

The involvement of technology in healthcare has achieved major improvements in resolving human health challenges. Explainability focuses on making an AI model's decisions understandable and accessible, providing user-friendly explanations that support causal reasoning [1]. In clinical medicine, it is crucial to have a clearer and deeper understanding of the stans made by the algorithms used to prevent occurrence of faulty conclusion and adverse patient outcomes [2]. To ensure these explanations are accessible to other professionals that are not computer experts and to obtain a greater level of fundamental comprehension among experts, straightforward explanation interfaces are essential [2].

The integration of Internet of Things (IoT) facilities for dataset collection along with patient monitoring has contributed to the improvement of healthcare mainly for medical staff decision support systems [3], [4]. Furthermore, with its prominent advantages like networking, sensing, expression, safety, and intelligence, the IoTs have become vital component of the healthcare industry [5]. The IoTs represents the interconnectedness of physical objects in cyberspace to exchange data. In addition to communicating, they are remotely controlled and observed. To maintain health records, data are gathered from a variety of devices, including blood glucose monitors, electrocardiograms (ECGs), and fetus monitors [5]. The IoT facilities help collect individual health relevant information in real-time. By leveraging data mining and ML/DL techniques, the data are often used to recommend health-related services or suggest lifestyle changes for the individual. [6]. Many modern medical sensors and gadgets are often linked over different networks, giving access to vital data regarding

IDDM'24: 7th International Conference on Informatics & Data-Driven Medicine, November 14 - 16, 2024, Birmingham, UK

* Corresponding author.

† These authors contributed equally.

samuel-gbenga.faluyi@ext.isep.fr (Samuel Gbenga Faluyi), yousra.chabchoub@isep.fr (Yousra Chabchoub), maurras.togbe@isep.fr (Maurras Togbe), jeremie.sublime@isep.fr (Jérémie Sublime),

0000-0002-2662-1599 (Samuel Gbenga Faluyi), 0000-0003-0772-1698 (Yousra Chabchoub), 0000-0001-9396-7947 (Maurras Togbe), 0000-0003-0508-8550 (Jérémie Sublime)

patients' status. The data can be employed for many functions, including remote patient monitoring, prognosticating disease, and recuperation by gaining a deeper understanding of symptoms and enhancing the diagnosis and treatment procedure through enhanced automation and mobility [3].

Machine learning and deep learning (ML/DL) are two subgroups of artificial intelligence (AI), that often learn from collected data, for intelligent decisions making. Their use in healthcare has improved the precision of diagnoses, customize treatment regimens, forecast patient outcomes, and accelerates operational efficiency [3]. There has been significant growth in the application of EHR resources recently which has accelerated the application of ML/DL to create patient phenotypes from EHR data [6]. EHR is regarded as rich spring of longitudinal experimental data that have the capability to house all important clinical and administrative data pertinent to a patient's care under a specific provider, including vital signs, prescriptions, medical history notes, demographics, and laboratory results [6].

ML models are often considered as black box algorithms, where the internal processes behind their predictions are not easily interpretable. In medicine, however, trust and explainability are crucial, as healthcare professionals and patients need to understand how decisions are being made. Explainability in machine learning refers to the ability to understand and articulate the inner workings of a model, its predictions, and the factors influencing those predictions. This transparency is essential for ensuring trust, accountability, and effective communication with stakeholders in automated decision-making systems. [7].

It is obvious that artificial intelligence with the use of ML/DL has contributed immensely to the rapidly growing achievements in healthcare. However, there is room for improvement. This review examines various ML/DL approaches previously used to extract critical information from state-of-the-art techniques. It also highlights their limitations and identifies areas where further solutions are necessary to enhance performance in healthcare.

This review paper is organized as follows. Section 2 presents the context and objectives of our study. Section 3 highlights the main research studies applying IA to healthcare. Detailed information regarding the ML/DL algorithms applied to healthcare can be found in Section 4. In section 5, we address healthcare data collection and data types. Explainability in the healthcare context is discussed in Section 6. Finally, Section 7 presents the conclusion and future works.

2. Context and Objectives

To develop individualized treatment routines, data is often collected and analyzed using ML/DL. Algorithms can predict a patient's response to different therapies based on their genetic composition. Moreover, predictive analytics is performed using models to forecast each patient's unique disease risk, making possible more individualized medical interventions.

ML/DL techniques are often used to examine EHR data to predict patient outcomes, readmission rates, and notify patients of high danger for certain situations. The EHR, formerly known as Clinical Information System, was described as a warehouse for healthcare big data [8]. The data can be numerical, text (for NLP), or medical imaging (e.g. Positron Emission Tomography, X-ray, Computed Tomography, and Ultrasound identification of tumors, fractures, and lesions) [8]. Most of the previous research studies have adopted EHR for their analytics, as an example, we can cite the studies named “Prediction of mortality in paralytic ileus patients” [9], “predicting post-pneumonia using deep neural network approaches”, [10], or also “Predicting the onset of type 2 diabetes” [11], among other research studies.

Real-time health data, like blood pressure, glucose levels, and heart rate, are gathered by wearable technology and IoTs sensors. These data are analyzed using ML/DL algorithms to monitor patient health and send out notifications to the concerned stakeholders in case of any abnormalities. AI-powered chatbots and virtual assistants facilitate telehealth by setting up appointments, making initial diagnoses, and responding to patient questions. Moreover, IoT was used for the collection of live data in real-time (time series), which is the case of the study [5] about heart disease prediction where IoT devices were used to collect live data.

To guarantee that medical professionals understand and can rely on AI-driven decisions, there is an increasing emphasis on creating models that yield clear and comprehensible outcomes. ML/DL models are sometimes considered black boxes, due to difficulties in explaining the internal operations of the models. However, through interpretable models, model-agnostic methods, visualization techniques, and a balance between complexity and interpretability; giving practitioners a clear understanding of the results, fostering better and more responsible use of machine learning technologies [12].

There are numerous techniques for achieving explainability, but it is crucial to grasp the key themes underlying different types of explainers. These include factors such as scope (local vs. global), model type (black box vs. white box), task (e.g., classification or regression), data type (tabular, images, text, etc.), and insights (feature attributions, counterfactuals, influential training instances, and more) [13]. Furthermore, explainers serve as interfaces that can work together with the model. For black-box techniques, this interaction typically involves analyzing the inputs and outputs. In contrast, for white-box techniques, explainers can access and interpret the internal workings of the model [14].

3. State of the art

Ahmed et al. [9] purported a predictive model that combines statistical techniques with machine learning algorithms to enhance the predictive performance of the models. Statistical data analysis was performed on the data obtained using IBM SPSS to examine the statistical significance of the features, for features reduction. The machine learning algorithms used are decision tree, linear discriminant analysis, K-Nearest Neighbors, gaussian naive bayes, and support vector machines with linear kernel and radial basis functions. Based on the results of the model, SVM with (radical basic function, RBF kernel) has the highest accuracy and ROC-AUC score. However, there was no method of data validation such as K-fold validation. Hyper-parameter tuning could be used for better optimization of the model.

Ge et al. [10] developed a post-stroke pneumonia predictive model applying ML/DL techniques, which combine both time series and time-insensitive attributes. As part of the preprocessing, the numerical observations of the data were normalized to tackle data sparseness and convert the laboratory test to categorical. The data were divided into two parts (in the proportion of 85% and 15%) for training and testing respectively, and three machine learning (LR, SVM, XGBoost) and deep learning (MLP and attention augmented GRU) algorithms were tested. 10 (k-fold) cross-validation was applied to the training set. To assess the performance of the model ROC, AUC, sensitivity, and specificity were applied to the test set. Based on the results, deep learning outperformed all the machine learning methods employed, where attention augmented gated recurrent unit (GRU) model achieved the highest AUC score. Subsequently, hyper-parameter tuning for optimization such as grid search or Bayesian optimization could be applied to identify the best hyper-parameter values. Moreover, exploratory data analysis may be administered to the dataset to aid the feature selection and to identify the statistical significance of the data.

Gupta et al. [5] proposed ML-based models for heart disease detection. The statistical correlation matrix was used on the collected data to examine the significance of the features. The model phase was carried out by dividing the entire collected data into two proportions for training and testing of the model. The algorithms used include K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF) and Decision Tree (DT). Furthermore, to determine the K-value of the K-NN, a score graph method was used, where the highest score was identified at the K-value of 3. The performance metrics were accuracy, sensitivity, miss rate, and confusion matrix. However, the best model was selected to validate the live dataset. Thus, real-time data are gathered by attaching several sensors to the body of the patients to measure different parameters. This data is then fed into the trained model to predict the outcome. The result of training and testing identified K-NN (3-NN) as the best algorithm. Therefore, K-NN was used for the prediction of heart diseases on the collected live data.

Nguyen et al. [11] developed a prediction model that can identify patients at high risk for developing type 2 diabetes using electronic health record data. The collected data is divided into training and tests (respectively 70% and 30%). 10-fold cross-validation was applied to the training set. Due to an imbalance in the data, SMOTE [11] was adopted. Also, the stochastic gradient descent optimizer and the binary cross-loss function were adopted for model training, with an ensemble learning model. The sensitivity, specificity, and ROC AUC were the performance metrics used for this study. The predictive model for T2DM was developed and comparisons of the algorithms used were made. The outcome of data with the application with and without SMOTE were compared. However, models using SMOTE showed higher sensitivity but no significant improvement for the other metrics. Moreover, ensemble models without SMOTE showed higher AUC and specificity compared to SMOTE-enhanced models. No explainability techniques were adopted for a better understanding of the model as well as for transparency.

Sood et al. [4] proposed healthcare IoT-fog technology to diagnose patient's hypertension stages and make prediction of hypertension occurrences based on users' health data collected. The study aims to leverage fog computing [15] to provide continuous monitoring for hypertensive patients and establish an efficient mechanism for sharing medical records and implementing precautionary measures. The system consists of three subsystems: an IoT-based subsystem for users, a health smart gateway (fog subsystem), and a cloud subsystem. The user subsystem utilizes various IoT devices to capture hypertension-related data, which is then transmitted to the health fog subsystem for real-time processing and diagnosis. Upon identifying a potential health issue, the health fog subsystem generates an alert message, sent directly to the user's mobile phone, allowing for timely precautionary action. Simultaneously, the analysis results and compiled medical records are stored in a cloud system, where they can be shared with authorized medical professionals, including doctors, pharmacies, hospitals, and healthcare providers. The cloud subsystem facilitates data storage and sharing, enabling domain experts to take swift action and offer precautionary advice in emergencies. The algorithms used for classification and prediction are Artificial Neural Network (ANN), K – Nearest Neighbours (K-NN), Multi-Layer Perceptron (MLP), and Logistic Regression (LR) [4]. The evaluation metrics include accuracy, time, sensitivity and precision. According to the system result, ANN outperforms all other classification algorithms in terms of accuracy, time, and standard metrics, for predicting the classification of hypertension attacks. The alert-generating result also reveals high values for sensitivity, specificity, precision, and coverage and low values for the Mean Absolute Error (MAE), Root Absolute Squared Error (RASE), Relative Absolute Error (RAE) and Root Relative Squared error (RRSE). These two latter are given by the following formulas:

$$Relative\ Absolute\ Error\ (RAE) = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i - \bar{y}_i|}$$

$$Root\ Relative\ Squared\ Error\ (RRSE) = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}}$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, n is the number of instances, and $\bar{y}$ is the mean of the actual values.

Furthermore, when it comes to alert production efficiency, fog monitoring-based alerts have the lowest delay times when compared to alerts based on cloud monitoring and alerts based on manual monitoring. Nonetheless, the security and privacy of data created by several layers of fog and cloud computing could be added in future studies.

Nguyen et al proposed in [16] three deep ensemble learning (DEL) approaches, for different data types (statistical, image-based and sequential). These are deep-stacked generalization ensemble learning, gradient deep learning boosting, and deep aggregation learning. Following the data reading phase, preprocessing of datasets was carried out using various techniques to convert the data into appropriate forms (e.g., converting the pictures to numerical data with the application of Convolutional Auto-Encoder, CAE). Subsequently, in the next phase, the suggested models and

additional conventional machine learning models were constructed following the dataset comprising various data. During this phase, the models' hyperparameters were also adjusted. Ultimately, they assessed the models and the suggested model in the last phase and compared the models' performances. A confusion matrix and metrics derived from it were used to evaluate the predictions' performance. The accuracy, Matthew's correlation coefficient (MCC), precision, F1-score, recall, and AUC metrics were obtained from the confusion matrix. MCC is defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The findings demonstrate that, across all dataset types, the deep ensemble learning (DEL) family of techniques outperforms all other based deployed models in terms of performance. The experiment results show that when DL is used as the CLU, the GDLB strategy is appropriate for numerical dataset, the DAL method is suggested for image dataset, and the DeSGEL [16] technique is suggested for sequential data. For more transparency of the developed model, the application of explainability techniques could be adopted. Also, developing an XGBoost algorithm for Neural networks which drops layers instead of pruning branches in decision trees.

4. ML/DL algorithms applied to healthcare

Research studies have shown that there are various ML/DL algorithms, which can be applied to healthcare. This section presents basic concepts of some common algorithms.

Ensemble Learning: Boosting, Bagging and Stacking

Ensemble methods hybridize several algorithms to build more powerful predictive models to obtain better performance than an individual based model [17]. The most typical kinds are bagging, boosting, and stacking.

Tree-based gradient boosting technique, XGBoost is flexible [18]. It integrates weak classifiers to produce prediction with high level of precision. While the classic gradient boosting approach optimizes using only the first derivative, XGBoost conducts the second-order Taylor expansion of the cost function and, for improved efficiency, adds a regularization item to the cost function [10].

Bagging is an ensemble machine learning technique, where many models are trained simultaneously on a random subset using the bootstrap aggregating approach. Using samples generated through sampling with replacement, the multiple algorithms are trained in the bootstrapping approach. The predictions from each model are then averaged. This algorithm considers the most evolved algorithms' categorization capability based on the voting mechanism. Reducing variance during training minimizes the likelihood of overfitting, which is a well-known specification of this approach [17], [19]. Random forest algorithm is one the major examples of bagging.

Boosting trains weak learners in a stepwise manner, where the mistakes made by earlier learners in the series are improved by each succeeding learner. First, a subset of the original dataset is selected. Following its training on this data, the first model predicts the outcome. Predictions about the samples may be accurate or inaccurate. For training the subsequent model, the incorrectly predicted samples are presented again. This allows the improvement of errors in earlier models by later versions. By aggregating the outcomes at each stage, boosting gathers the results earlier than bagging, which does it at the conclusion. A weighted average is used to combine them. Based on how successfully a model predicts the future, it is given a varying weight via weighted averaging [17], [19], [20]. Row and column resampling are also used by the boosting algorithms; these methods are used to prevent overfitting [21].

Support Vector Machine (SVM)

SVM algorithms, which are considered supervised learning are often used to create a model that classifies objects or data into different categories, by finding the maximum-margin hyperplane for the binary classification of new data points from known classified data points [3], [8], [22], [23]. This aids

the new input sets to be predicted more quickly than in other predictive models, regardless of the size of the training set in the domain. The goal of SVM is to locate the largest margin hyperplane, to divide the data and provide the best fit to arrange it [8]. By applying classical statistical learning theory, the SVM produces a model that is easily interpreted and provides good generalization of fresh data. Since they support the placement of the dividing hyperplanes, the nearest points are known as support vectors. This implies that the hyperplanes cannot be changed by shifting the nonsupport vectors, and vice versa [22]. Due to its ability to classify objects, SVM is often utilized in clinical imaging analysis to classify or categorize diagnoses [23].

SVMs exhibit the advantage of being very preventive to overfitting issues. SVMs are not limited to employing linear classifiers; they can also be used to classify data using non-linear functions by utilizing non-linear kernels [3]. A model demonstrating an SVM classifier that performed better than previous classifiers for determining whether a person has influenza-like illnesses (ILI) often referred to as acute respiratory infections was proposed by [24]. In location verification, SVM is determined to be the most accurate, and it doesn't need channel characteristics data to function [3].

Random Forest (RF)

Multiple decision trees can be trained concurrently using random forests to generate a single output. Random Forest is one of the bagging ensemble machine learning that involves merging decision trees [3]. Al Hossain et al. [25] provided evidence of the use of a random forest algorithm that outperformed alternative models in estimating the number of influenza cases in public areas with a 95% accuracy rate. Because it can integrate the results from every decision tree, it demonstrates a high degree of accuracy.

Naive Bayes (NB)

The Bayes theorem operates as the conceptual foundation for Naive Bayes classification. The term "naive" describes the belief that each attribute is independent of the others. A response vector and a feature matrix are created from the data [3]. The whole set of data is presented in the feature matrix's rows as vectors, each of which denotes a different relative variable type. Conversely, every row in the response vector denotes a class of outcome. Naive Bayes classifier performed significantly well in classifying in controlling the social networks during pandemic catastrophe, where it outperformed other classifiers [26].

Extreme Gradient Boost

As an ensemble learning technique, where weak learners are combined to produce a stronger learner for better accuracy, boosting establishes decision boundaries for every weak learner and weights them according to how well the boundaries identified or approximated the data. Until a workable model is produced, this is repeated. Gradient boosting involves the sequential creation of numerous boundaries, or learners so that each learner can partially account for the errors of the preceding one [3]. Through parallel processing, pruning of decision trees, management of missing values, and reducing the likelihood of bias or overfitting in a model, extreme gradient boosting (XGBoost) is used to optimize gradient boosting techniques. Each iteration's tree is computed using the first- and second-order gradient of the loss function, and the shrinkage parameter is used to add the predictions to the current function and minimize the optimal node predictions made in each iteration [30]. XGBoost is a potent and adaptable algorithm that may be applied to a range of problems, including regression and classification, forecasting, and ranking. Ensuring that machines are operating as efficiently as possible in terms of mobility, scalability, and accuracy is the primary objective of the XGBoost model. On the other hand, extreme gradient boosting, or XGBoost, is renowned for its meticulous tuning to yield better outcomes with fewer resources while remaining effective [27]. Heart patients' irregular cycles were predicted with 92.1% accuracy using extreme gradient boosting [28]. Analogously, speech signals obtained from wearables can be utilized to identify Parkinson's disease symptoms in their early stages [29], while predictive analytics can be employed to identify diabetes [3], [30].

Artificial Neural Network (ANN)

ANN is an ML model that simulates the way the human brain learns. It consists of an input layer that accepts information, many tiers that analyze the input, and an output layer that gives results. If the outputs are inaccurate, they are propagated backwards through the preceding layers using a cost function to adjust the weights until the answers are received with a high enough degree of precision. "It calculates several weighted sums, which are then passed through layers with weights and sums until they reach the last layer, which uses an activation function to determine the output" [3]. ANNs are very flexible in application and are often used in pattern recognition-related fields. Sood and Mahajan [4] employed a fog-layer system to store patient data related to heart attacks and to detect, monitor and treatment of hypertension (BP) cases [3], [4].

Convolutional Neural Network (CNN)

CNN is regarded as a feed-forward neural network and is usually used in classification challenges. The input is decomposed into its constituent pieces, which are subsequently sent to a convolution layer, and then these parts are combined in various ways until patterns are produced (convolution) [3], [31]. The input images are then mapped against these patterns using a Rectified Linear Unit (ReLU) layer, creating a rectified feature layer, which is then passed on to a pooling layer. To create a pooled feature map, the map is reduced by the pooling layer. This map is then flattened to create a linear vector, which is then served into a completely linked network to classify the input. CNNs are widely utilized in fields where visual interpretation of grid-like-topped images is required [3]. Brain wave values acquired as a 2D-time series were converted to forecast epileptic incidents and immediately notify health authorities [32]. Ke et al. [33] suggested utilizing lightweight CNN to assess depression using raw electroencephalograms (EEGs). Ciocca et al. [34] utilized picture recognition to recognise food and, consequently, calories, a finding with implications for fitness and nourishment. Alhussein and Muhammad [35] applied deep learning on pitch tones in mobile healthcare frameworks to identify speech disorders. Using the LUNA16 dataset, Bansal [36] developed a resnet-based model for lung disease classification and 3D dissection, where an excellent accuracy of segmentation and classification was obtained [36].

Thus, numerous ML/DL techniques can be used to healthcare to perform various activities including prediction, classification, regression, among others. Some of the algorithms have been highlighted above and various research studies have identified the efficiency of these algorithms based on the adopted parameters.

5. Healthcare data collection and data types

In the application or development of ML/DL modelling, one of the major factors to be considered is the approach of data gathering and the type of collected data. This section presents some instances of data collected and used in previous research studies.

Table 1
Methods of data collection and data types

Study	Source of data	Component of the data	Data type
[9]	The data collected for the study contained different ICU records from 2001 to 2012.	The dataset made up of 46476 patients admitted to ICU, where 1021 patients were diagnosed with paralytic ileus and >= 18 years old was used for the prediction model.	Numerical and Categorical
[10]	The dataset was collected from the EHR of a hospital in the space of 10 years from 2007 to 2017.	The data contains 13930 records of patients, where 1012 had pneumonia while in the hospital. Some of the records are time sensitive (medication,	Numerical and categorical

		laboratory tests) and others are time insensitive (demographic information).	
[5]	Data used for the study was collected from the UCI repository.	The data consists of 303 instances of 14 features, which are grouped into numerical (such as age) and categorical (sex, chest pain type, etc) features.	Numerical and categorical.
[11]	The data was collected from the EHR of a hospital in the United States from 2009 to 2011.	The data comprises 9948 patients' records, where 1904 patients were diagnosed with type 2 diabetes mellitus.	Numerical and categorical
[4]	The data were collected via users' subsystems comprising several IoT facilities to obtain hypertension activities. They are then communicated to a fog system for concurrent processing and diagnosis. Alerts are generated and shared with the staff concerned. The data is stored in the cloud.	Data collected by this system are categorized into six groups, which are health data (such as obesity, SBP, DBP, etc), environmental data (room temperature, noise level, air quality), physical activity data (such as sleeping, sitting, walking, etc), behavioural data (anxiety level, restlessness, etc), dietary data (Diet type, quantity), and GPS data (location and time).	Numerical and categorical
[16]	Three different open datasets were used, which are Heart Disease UCI (HDU) data, X-ray data, and the Depresjon data.	HDU: 270 instances (containing 120 and 150 records of having and not having heart disease respectively) with 13 attributes. X-ray: 5856 samples (made up of 4273 and 1583 images with and without pneumonia respectively) Depresjon: 267 and 547 samples of depressed patients and non-depressed people respectively	HDU: Numerical X-ray: Image Depresjon: Numerical

We showed in table 1 that healthcare data for ML/DL modelling can be obtained from different sources. More so, part of the benefits of adopting ML/DL techniques is the capability to deal with different types of input data (numerical, categorical, image-based, text based, etc.). The data can be time sensitive or time insensitive. The selection of the most appropriate ML/DL model depends on all these characteristics of the dataset, in addition to the target result of the application (prediction, classification…). The trust of the obtained result is also closely dependent on its explainability which is discussed in the following part.

6. Explainability in healthcare context

Model explainability, often known as explainable AI, describes methods for making machine learning (black box and white box) models' predictions more comprehensible to human observers, particularly when there are difficulties in explaining the internal operations of the models [12]. A strong machine learning system must have the capacity to justify predictions to foster confidence in the decisions made in the model's process [37]. The explanation's target insights vary greatly depending on who uses them, from regulators auditing the models to data scientists troubleshooting them. Therefore, to meet the needs of the target audience, a variety of approaches are required. This is because stand-alone explanation techniques may produce explanations that are deceptive or lacking in context [38], [39]. This implies that explaining models holistically is necessary. Explainability in machine learning is essential for ensuring transparency, trust, and accountability in automated decision-making systems

[40]. It involves understanding how models make predictions and being able to communicate this knowledge to various stakeholders [41], [42]. The interpretation ability, which assesses the influence, relationship, and correlation of conditioning components within a model, highlights the benefits of XAI above traditional techniques. Explainability was proposed to enhance the prediction capability of infections related to healthcare in patients admitted into intensive care units while preserving the model. This goes beyond the artificial neural network black box paradigm by using a parsimonious and robust semi-parametric approach. More so, the saliency map was used to examine and justify the additional predictive capability of this model [7].

Explainability and causality in the medical field are also critical for regulatory compliance, further highlighting its relevance [43]. These interfaces not only help keep humans involved in the process but also permit for the incorporation of their experiential knowledge and conceptual understanding into AI operations. While the importance of a person-in-the-loop is sometimes undervalued, implicit knowledge and human expertise remain indispensable in medical diagnosis [44]. By following diagnostic steps, individual components that contribute to a diagnosis can be identified and applied to train and improve models prospectively [45].

Explainability frameworks

The rapid advancements in ML/DL technologies, along with the increasing adoption of AI, highlight the need for greater awareness of AI's operational mechanisms, making explainable modelling essential. There are various examples of model explainers, which include but are not limited to the ones highlighted in table 2.

Although there is a wide variety of approaches accessible for explainability, it is critical to comprehend the overarching themes of the many categories of Explainers. Among them are: scope (local (L) and global (G)), type of model (white box (Wbox) and black box (Bbox)), Task (regression (R), classification (C), time-series (TS), image (I), etc.), type of data (text (Tt), image (I), tabular (Tab), etc.) and Insight (attributions of features, counterfactuals, significant training examples, etc.). However, these systems exhibit data flow patterns like those of explainers functioning as interfaces. Particularly, a lot of them call for the users to enable them to communicate both with the model and the data it processes; in the case of black-box techniques, this refers to the inputs and outputs, while in the case of white-box techniques, it refers to the internal workings of the models [14].

Table 2
Explainability Frameworks

Explainer **Feature**	**Alibi**	**DALEX**	**Interpret ML**	**Explainer dashboard**	**SHAP**	**Alibi-detect**	**Captum**
Scope	G, L	G, L	G, L	G, L	G	G, L	G, L
Model type	Bbox, Wbox	Bbox, Wbox	Bbox, Wbox	Bbox, Wbox	Bbox,	Bbox,	Bbox,
Task	C, R	R, C, I	C, R	C, R	C, R	TS, I, C	I, TS, C, R
Data type	Tab, Tt, I	Tab, Tt, I	Tab	Tab	Tab	Tab, Tt, I	Tab, Tt, I
Insight	Feature attribution, influential training instances	Feature importance, PDP, residua diagnostic, surrogate model	Feature importance, EBM, SHAP, LIME, etc.	Feature importance, SHAP, PDP, Decision Path Visualization	Summary plots, PDP, interaction effect, feature importance	Local Outlier Factor, Isolation Forest, Visualization, anomaly score	Saliency map, layer-wise attribution, neuron attribution

In [7], authors introduced a two-step methodology for predicting ICU-acquired infections (ICU-AIs) using high-resolution longitudinal data combined with survival models. The study applied a saliency map model explainer to examine the images of signal present in the used data and the outcome of the model [7].

Model explainability can be categorized as either ante-hoc or post-hoc. Ante-hoc models are inherently self-explainable, while post-hoc models require the use of explainable AI (XAI) methods to provide explanations for their decisions [12], [43]. Once a machine learning model has been trained and has made its predictions, post-hoc explainer techniques examine and clarify its decision-making procedure to provide insights into how the model operates. In contrast, ante-hoc approaches are inherently interpretable; often referred to as intrinsically explainable, transparent, or glass-box models. Like interactive machine learning (iML), these approaches focus on embedding interpretability directly into architecture of the model, ensuring transparency and explainability from the outset [43], [46], [47], [48].

One common post-hoc method involves determining the significance of various attributes in producing a specific result [49]. Post-hoc methods based on game theory, such as Shapley values, can quantify the importance of individual features. Similarly, Anchors, another post-hoc approach, provides insights into coverage, and the region where the explanation is applicable, and helps define the boundaries of attributes. Anchors are particularly useful for classification models involving text-based, and tabular data [50]. According to Dandl et al. [51], counterfactuals are an XAI technique that describes the smallest modification to the attribute values that affects the prediction to explain specific forecasts [43], [51].

Decision trees (DT) are one of the well-known examples of interpretable machine learning models. They operate by repeatedly splitting the data based on specific threshold values of the features, creating distinct subsets of the dataset, with each instance assigned to one of these subsets [52]. These models are interpretable because their structure can be easily followed, commencing from the root node, through the subsequent nodes and edges, until the leaf node with the predicted outcome is reached. The DT algorithms are considered interpretable due to their structure of hierarchy such as if-then-else rules that can be easily visualized, understood, and interpreted by humans [43].

7. Conclusion and Future works

We focus in this paper on identifying the impact of ML/DL and their applications in providing solutions to health challenges. These include prediction modelling, enhancing models using ensemble machine learning and optimization techniques, application of fog facilities and cloud systems for collecting, sharing and storing data as well as easy retrieval of the data. We considered the main research studies applying ML/DL algorithms to healthcare. We highlighted the benefits of using EHR: some of the collected datasets adopted for the training, validation and testing in the simulation processes of the model development were stored in the EHR. Various data types are considered. Moreover, we examined the most well-known explainability frameworks and their different characteristics.

However, there are several challenges or shortfalls in the previous research studies, which require prospective investigation to further enhance the impact of ML/DL applications in the healthcare sector. For example, making predictions requires the involvement of optimization techniques such as hyperparameter tuning, grid-search techniques, etc. Further studies could engage the application of ensemble learning approaches along with model explainer techniques to enhance the adoption of ML/DL in healthcare. Furthermore, transparency in the model undeniably fosters trust, which in turn promotes its adoption and contributes to improving the developed model.

Acknowledgement

We would like to thank the Tertiary Education Trust Fund (TETFUND) for supporting this study.

References

[1] T. Evans *et al.*, "The explainability paradox: Challenges for xAI in digital pathology," *Future Generation Computer Systems*, vol. 133, pp. 281–296, Aug. 2022, doi: 10.1016/j.future.2022.03.009.
[2] M. Plass *et al.*, "Explainability and causability in digital pathology," *The Journal of Pathology CR*, vol. 9, no. 4, pp. 251–260, Jul. 2023, doi: 10.1002/cjp2.322.
[3] H. K. Bharadwaj *et al.*, "A Review on the Role of Machine Learning in Enabling IoT Based Healthcare Applications," *IEEE Access*, vol. 9, pp. 38859–38890, 2021, doi: 10.1109/ACCESS.2021.3059858.
[4] S. K. Sood and I. Mahajan, "IoT-Fog-Based Healthcare Framework to Identify and Control Hypertension Attack," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1920–1927, Apr. 2019, doi: 10.1109/JIOT.2018.2871630.
[5] R. Gupta, S. Tanwar, S. Tyagi, and N. Kumar, "Machine Learning Models for Secure Data Analytics: A taxonomy and threat model," *Computer Communications*, vol. 153, pp. 406–440, Mar. 2020, doi: 10.1016/j.comcom.2020.02.008.
[6] M. Bampa, I. Miliou, B. Jovanovic, and P. Papapetrou, "M-ClustEHR: A multimodal clustering approach for electronic health records," *Artificial Intelligence in Medicine*, vol. 154, p. 102905, Aug. 2024, doi: 10.1016/j.artmed.2024.102905.
[7] G. Lancia, M. R. J. Varkila, O. L. Cremer, and C. Spitoni, "Two-step interpretable modeling of ICU-AIs," *Artificial Intelligence in Medicine*, vol. 151, p. 102862, May 2024, doi: 10.1016/j.artmed.2024.102862.
[8] H. Habehh and S. Gohel, "Machine Learning in Healthcare," *CG*, vol. 22, no. 4, pp. 291–300, Dec. 2021, doi: 10.2174/1389202922666210705124359.
[9] F. S. Ahmed *et al.*, "A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (EHRs)," *J Ambient Intell Human Comput*, vol. 14, no. 10, pp. 14367–14367, Oct. 2023, doi: 10.1007/s12652-020-02509-7.
[10] Y. Ge *et al.*, "Predicting post-stroke pneumonia using deep neural network approaches," *International Journal of Medical Informatics*, vol. 132, p. 103986, Dec. 2019, doi: 10.1016/j.ijmedinf.2019.103986.
[11] B. P. Nguyen *et al.*, "Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records," *Computer Methods and Programs in Biomedicine*, vol. 182, p. 105055, Dec. 2019, doi: 10.1016/j.cmpb.2019.105055.
[12] C. Molnar, *8.6 Global Surrogate | Interpretable Machine Learning*. 2019. Accessed: Jul. 29, 2024. [Online]. Available: https://christophm.github.io/interpretable-ml-book/global.html
[13] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, Dec. 2021, doi: 10.1016/j.inffus.2021.05.009.
[14] A. Saucedo, "Production Machine Learning Monitoring: Outliers, Drift, Explainers & Statistical Performance," Medium. Accessed: Jul. 29, 2024. [Online]. Available: https://towardsdatascience.com/production-machine-learning-monitoring-outliers-drift-explainers-statistical-performance-d9b1d02ac158
[15] P. Verma and S. K. Sood, "Fog Assisted-IoT Enabled Patient Health Monitoring in Smart Homes," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1789–1796, Jun. 2018, doi: 10.1109/JIOT.2018.2803201.
[16] D.-K. Nguyen, C.-H. Lan, and C.-L. Chan, "Deep Ensemble Learning Approaches in Healthcare to Enhance the Prediction and Diagnosing Performance: The Workflows, Deployments, and Surveys on the Statistical, Image-Based, and Sequential Datasets," *International Journal of Environmental Research and Public Health*, vol. 18, no. 20, Art. no. 20, Jan. 2021, doi: 10.3390/ijerph182010811.
[17] Sumbatilinda, "Ensemble Learning in Machine Learning: Bagging, Boosting and Stacking," Medium. Accessed: Jul. 23, 2024. [Online]. Available: https://medium.com/@sumbatilinda/ensemble-learning-in-machine-learning-bagging-boosting-and-stacking-a00c6bae971f
[18] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in

KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[19] R. Mirzaeian, R. Nopour, Z. Asghari Varzaneh, M. Shafiee, M. Shanbehzadeh, and H. Kazemi-Arpanahi, "Which are best for successful aging prediction? Bagging, boosting, or simple machine learning algorithms?," *BioMed Eng OnLine*, vol. 22, no. 1, p. 85, Aug. 2023, doi: 10.1186/s12938-023-01140-9.
[20] S. Ganie, P. Dutta Pramanik, M. Malik, and A. Nayyar, "An Improved Ensemble Learning Approach for Heart Disease Prediction Using Boosting Algorithms," *Computer Systems Science and Engineering*, vol. 46, pp. 3993–4006, Apr. 2023, doi: 10.32604/csse.2023.035244.
[21] S. Faluyi, T. Balogun, G. Ojo, K. Fapohunda, and Akande Adeyemi, "Forecasting transaction card fraud using bossting algorithms," in *Communication and e-Systems for Economic Stability*, 2023.
[22] B. A. Akinnuwesi *et al.*, "Application of support vector machine algorithm for early differential diagnosis of prostate cancer," *Data Science and Management*, vol. 6, no. 1, pp. 1–12, Mar. 2023, doi: 10.1016/j.dsm.2022.10.001.
[23] E.-J. Lee, Y.-H. Kim, N. Kim, and D.-W. Kang, "Deep into the Brain: Artificial Intelligence in Stroke Imaging," *J Stroke*, vol. 19, no. 3, pp. 277–285, Sep. 2017, doi: 10.5853/jos.2017.02054.
[24] N. L. W. S. R. Ginantra, I. G. A. D. Indradewi, and E. Hartono, "Machine learning approach for Acute Respiratory Infections (ISPA) prediction: Case study Indonesia," *J. Phys.: Conf. Ser.*, vol. 1469, no. 1, p. 012044, Feb. 2020, doi: 10.1088/1742-6596/1469/1/012044.
[25] F. Al Hossain, A. A. Lover, G. A. Corey, N. G. Reich, and T. Rahman, "FluSense: A Contactless Syndromic Surveillance Platform for Influenza-Like Illness in Hospital Waiting Areas," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 1–28, Mar. 2020, doi: 10.1145/3381014.
[26] N. Assery, Y. Xiaohong, S. Almalki, R. Kaushik, and Q. Xiuli, "Comparing Learning-Based Methods for Identifying Disaster-Related Tweets," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Dec. 2019, pp. 1829–1836. doi: 10.1109/ICMLA.2019.00295.
[27] A. Khang, G. Rana, R. K. Tailor, and V. Abdullayev, Eds., *Data-centric AI solutions and emerging technologies in the healthcare ecosystem*, First edition. Boca Raton: CRC Press, 2024.
[28] H. Shi, H. Wang, Y. Huang, L. Zhao, C. Qin, and C. Liu, "A hierarchical method based on weighted extreme gradient boosting in ECG heartbeat classification," *Computer Methods and Programs in Biomedicine*, vol. 171, pp. 1–10, Apr. 2019, doi: 10.1016/j.cmpb.2019.02.005.
[29] S. Barbon Junior, V. G. T. Costa, S.-H. Chen, and R. C. Guido, "U-Healthcare System for Pre-Diagnosis of Parkinson's Disease from Voice Signal," in *2018 IEEE International Symposium on Multimedia (ISM)*, Dec. 2018, pp. 271–274. doi: 10.1109/ISM.2018.00039.
[30] F. Zafar, S. Raza, M. U. Khalid, and M. A. Tahir, "Predictive Analytics in Healthcare for Diabetes Prediction," in *Proceedings of the 2019 9th International Conference on Biomedical Engineering and Technology*, in ICBET '19. New York, NY, USA: Association for Computing Machinery, Mar. 2019, pp. 253–259. doi: 10.1145/3326172.3326213.
[31] A. Mehra, M. Mandal, P. Narang, and V. Chamola, "ReViewNet: A Fast and Resource Optimized Network for Enabling Safe Autonomous Driving in Hazy Weather Conditions," *IEEE Trans. Intell. Transport. Syst.*, vol. 22, no. 7, pp. 4256–4266, Jul. 2021, doi: 10.1109/TITS.2020.3013099.
[32] M. Alhussein, G. Muhammad, M. S. Hossain, and S. U. Amin, "Cognitive IoT-Cloud Integration for Smart Healthcare: Case Study for Epileptic Seizure Detection and Monitoring," *Mobile Netw Appl*, vol. 23, no. 6, pp. 1624–1635, Dec. 2018, doi: 10.1007/s11036-018-1113-0.
[33] H. Ke *et al.*, "Cloud-aided online EEG classification system for brain healthcare: A case study of depression evaluation with a lightweight CNN," *Softw Pract Exp*, vol. 50, no. 5, pp. 596–610, May 2020, doi: 10.1002/spe.2668.
[34] G. Ciocca, P. Napoletano, and R. Schettini, "CNN-based features for retrieval and classification of food images,'' Comput. Vis. Image Understand," vol. 176, pp. 70–77, 2018.
[35] M. Alhussein and G. Muhammad, "Voice pathology detection using deep learning on mobile healthcare framework," vol. 6, pp. 41034–41041, 2018.

[36] G. Bansal, V. Chamola, P. Narang, S. Kumar, and S. Raman, "Deep3DSCan: Deep residual network and morphological descriptor based framework forlung cancer classification and 3D segmentation," *IET Image Processing*, vol. 14, no. 7, pp. 1240–1247, 2020, doi: 10.1049/iet-ipr.2019.1164.
[37] J. Klaise, A. Van Looveren, C. Cox, G. Vacanti, and A. Coca, "Monitoring and explainability of models in production," Jul. 13, 2020, *arXiv*: arXiv:2007.06299. doi: 10.48550/arXiv.2007.06299.
[38] J. Klaise, A. Van Looveren, G. Vacanti, and A. Coca, "Alibi explain: algorithms for explaining machine learning models," *J. Mach. Learn. Res.*, vol. 22, no. 1, p. 181:8194-181:8200, Jan. 2021.
[39] J. Covell, "Project expl AI n - Interim report | Policy Commons," 2019, Accessed: Aug. 19, 2024. [Online]. Available: https://policycommons.net/artifacts/2440692/project-expl-ai-n/3462416/
[40] H. A. H. Al-Najjar, B. Pradhan, G. Beydoun, R. Sarkar, H.-J. Park, and A. Alamri, "A novel method using explainable artificial intelligence (XAI)-based Shapley Additive Explanations for spatial landslide prediction using Time-Series SAR dataset," *Gondwana Research*, vol. 123, pp. 107–124, Nov. 2023, doi: 10.1016/j.gr.2022.08.004.
[41] P. Biecek, "XAI in Python with dalex," Medium. Accessed: Aug. 19, 2024. [Online]. Available: https://medium.com/@ModelOriented/xai-in-python-with-dalex-4b173486aa92
[42] A. Dhinakaran, "A Look Into Global, Cohort and Local Model Explainability," Medium. Accessed: Aug. 18, 2024. [Online]. Available: https://towardsdatascience.com/a-look-into-global-cohort-and-local-model-explainability-973bd449969f
[43] C. O. Retzlaff *et al.*, "Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists," *Cognitive Systems Research*, vol. 86, p. 101243, Aug. 2024, doi: 10.1016/j.cogsys.2024.101243.
[44] J. M. Metsch *et al.*, "CLARUS: An interactive explainable AI platform for manual counterfactuals in graph neural networks," *Journal of Biomedical Informatics*, vol. 150, p. 104600, Feb. 2024, doi: 10.1016/j.jbi.2024.104600.
[45] M. Plass, M. Kargl, P. Nitsche, E. Jungwirth, A. Holzinger, and H. Muller, "Understanding and Explaining Diagnostic Paths: Toward Augmented Decision Making," *IEEE Comput. Grap. Appl.*, vol. 42, no. 6, pp. 47–57, Nov. 2022, doi: 10.1109/MCG.2022.3197957.
[46] A. Holzinger, "Explainable AI (ex-AI)," *Informatik Spektrum*, vol. 41, no. 2, pp. 138–143, Apr. 2018, doi: 10.1007/s00287-018-1102-5.
[47] A. Holzinger *et al.*, "Interactive machine learning: experimental evidence for the human in the algorithmic loop: A case study on Ant Colony Optimization," *Appl Intell*, vol. 49, no. 7, pp. 2401–2414, Jul. 2019, doi: 10.1007/s10489-018-1361-5.
[48] C. O. Retzlaff *et al.*, "Human-in-the-Loop Reinforcement Learning: A Survey and Position on Requirements, Challenges, and Opportunities," *jair*, vol. 79, pp. 359–415, Jan. 2024, doi: 10.1613/jair.1.15348.
[49] C. Glanois *et al.*, "A survey on interpretable reinforcement learning," *Mach Learn*, vol. 113, no. 8, pp. 5847–5890, Aug. 2024, doi: 10.1007/s10994-024-06543-w.
[50] R. Dwivedi *et al.*, "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–33, Sep. 2023, doi: 10.1145/3561048.
[51] S. Dandl, C. Molnar, M. Binder, and B. Bischl, "Multi-Objective Counterfactual Explanations," in *Parallel Problem Solving from Nature – PPSN XVI*, vol. 12269, T. Bäck, M. Preuss, A. Deutz, H. Wang, C. Doerr, M. Emmerich, and H. Trautmann, Eds., in Lecture Notes in Computer Science, vol. 12269., Cham: Springer International Publishing, 2020, pp. 448–469. doi: 10.1007/978-3-030-58112-1_31.
[52] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 3, pp. 660–674, Jun. 1991, doi: 10.1109/21.97458.