Decision Analytics Journal 7 (2023) 100230

Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends

Saranya A.[1], Subhashini R.[*]

School of Information Technology and Engineering, VIT University, Vellore, India

ARTICLE INFO

Keywords:
Artificial Intelligence
Machine learning
Deep learning
Explanation
Explainable Artificial Intelligence
HealthCare

ABSTRACT

Artificial Intelligence (AI) uses systems and machines to simulate human intelligence and solve common real-world problems. Machine learning and deep learning are Artificial intelligence technologies that use algorithms to predict outcomes more accurately without relying on human intervention. However, the opaque black box model and cumulative model complexity can be used to achieve. Explainable Artificial Intelligence (XAI) is a term that refers to Artificial Intelligence (AI) that can provide explanations for their decision or predictions to human users. XAI aims to increase the transparency, trustworthiness and accountability of AI system, especially when they are used for high-stakes application such as healthcare, finance or security. This paper offers systematic literature review of XAI approaches with different application and observes 91 recently published articles describing XAI development and applications in healthcare, manufacturing, transportation, and finance. We investigated the Scopus, Web of Science, IEEE Xplore and PubMed databases, to find the pertinent publications published between January 2018 to October 2022. It contains the published research on XAI modelling that were retrieved from scholarly databases using pertinent keyword searches. We think that our systematic review extends to the literature on XAI by working as a roadmap for further research in the field.

Contents

* Corresponding author.
E-mail address: rsubhashini@vit.ac.in (Subhashini R.).
[1] Research Scholar, SITE, VIT University, Vellore, TamilNadu, India.

https://doi.org/10.1016/j.dajour.2023.100230
Received 16 February 2023; Received in revised form 12 April 2023; Accepted 12 April 2023
Available online 17 April 2023

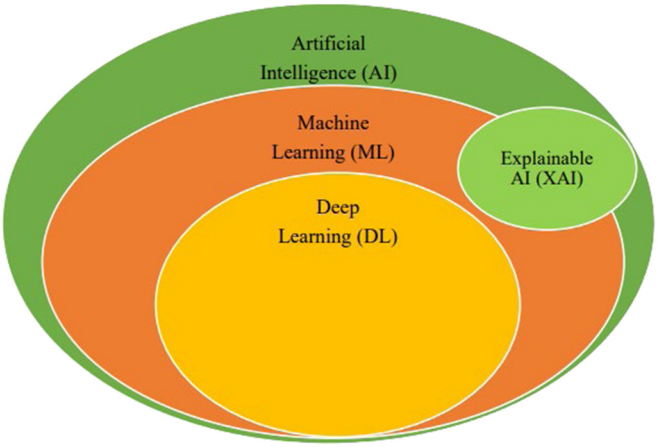

Fig. 1. AI vs. ML vs. DL vs. XAI.

1. Introduction

Artificial intelligence (AI) is a method that refers to a system or a machine that imitates human intelligence to perform functions in the real world. AI allows the system to be trained from the data and to think and learn from the experience to solve particular problems. It can heuristically refine itself based on the data used. AI applications include advanced web Search Engines, Automated Driving Cars, Games, Human Speech Recognition, Recommendation System, Healthcare etc.

AI was developed around 1950 in the computer science sector, and it copied the human mind to develop machines that can process, methodise, and perform based on the data given to the system, which will be useful when large amounts of datasets are used [1]. AI machineries widely being used in the industrial domain and prompted to do a more research works in engineering fields such as NLP (natural language processing), diagnosis of diseases and medicine, science [2]. AI machines used to learn from their previous experiences, which was helpful in solving problems, and it has been used in different application domains to increase the performance of the AI machines [3].

Fig. 1 explains the connection between the Artificial-Intelligence, Machine-Learning, Deep-Learning, and Explainable Artificial-Intelligence.

Machine Learning (ML) is a technique that allows a computer to identify patterns, make predictions that are more accurate, and refine itself via experience without being precisely programmed to do so. Machine Learning is used to build an AI-driven application. This process is done by using the ML Methodologies.

The process of ML:

The process of ML is described in Fig. 2. AI are used to make the decisions a lot. Integrated with AI, the system can perform tasks faster and predict the decisions needed to solve complex problems, evaluate risks, and evaluate business performance.

To measure the ML model's performance, different metrics have been used based on the ML algorithm used in the application domain. *Area under Curve (AUC)* and *Accuracy under the receiver operating characteristics (AUROC)* are the most commonly used performance metrics in ML tasks. AUC represents the performance of the model at separating classes, whereas accuracy denotes the model's overall correctness [1]. ML models have been trained with a larger amount of data and make the most accurate predictions [2]. ML is classified into two main groups: supervised and unsupervised learning. Further, supervised learning is classified as semi-supervised learning [4] and reinforcement learning [5].

Types of Learning methods in ML

Advanced machine learning styles have been explained by Supriya V. Mahadevkar et al. [6]. Some of the learning styles are listed in Fig. 3 [6].

Deep learning (DL) deals with algorithms influenced by the structure and function of the human brain. DL utilises artificial neural networks to create an intelligent model and solve critical problems. DL makes use of both structured and unstructured data to train a model (e.g., visual assistants like Siri, Alexa, and face recognition, etc.). DL is used for medical research and the prediction of life-threatening diseases. Recently, Deep Neural Networks (DNNs) have established remarkable predicting performance [7].

Nowadays, deep learning has made important progress due to its increased range of calculating power and because it provides a better solution for a larger number of datasets [2]. AI has a subclass called Deep Learning (DL) that is created on an *artificial neural network*. In this DL process, the input data will be trained by themselves over mathematical illustration. Some of the DL models are C*onvolutional Neural Networks (CNN), Visual Geometric Group Net (VGGNet), Residual Network (ResNet), Fully Convolutional Networks (FCNs), U-net* [8], *Deep feed forward networks, Siamese Neural Networks, Graph Neural Networks* [9]. Deep Learning models are divided into three modules called data pre-processing, feature extraction and recognition, and model optimisation [10].

The Artificial Intelligence algorithm was used to make the user take the decision in their business, but humans do not have any knowledge about the output of the AI or how it was reached. So, it is difficult for the users, to understand the output and process of the outcome. Hence, *Explainable Artificial Intelligence (XAI) is* being used.

Explainable AI (XAI) explains the inner process of a model i.e., used to provide the explanation of the methods, procedures and output of the processes and that should be understandable by the users. *The Defense Advanced Research Project Agency (DARPA)* invented the term "Explainable AI" (XAI). It will be called as "White box" because of explaining the process of the model.

Training data will be given as an input, and based on the requirements or application domain, you will have to select the methodology for the prediction and the XAI techniques being used to explain the inner workings of the models and the output with an explanation interface as mentioned in Fig. 4. Hence, the users had knowledge about the output of Explainable AI, which will increase their trust in AI models. Based on the knowledge of the output, users can improve the accuracy of the outcome and also expose the flaws in the model, which again will be useful for the users to make the right decision to improve the model. Explainable AI is used in critical sectors like healthcare, signal processing, etc.

Methods of Explainable AI is being classified into two categories: *knowledge-driven and data-driven* [11]. *Knowledge-driven* XAI needs knowledge about the methods to provide the explanation. *Data-driven* XAI needs local, global, and instance-specific methods for explanation. Most common XAI approaches have been classified based on an explanation of the scope of the model. Linear regression is self-interpretable with simple data called an "intrinsic model" (ex., Logic Analysis of Data (LAD)) and non-linear is interpretable with more complex data called "post-hoc approaches" (ex., Local Interpretable Model-agnostic Explanation and SHapley Additive exPlanations), "model agnostic", "model

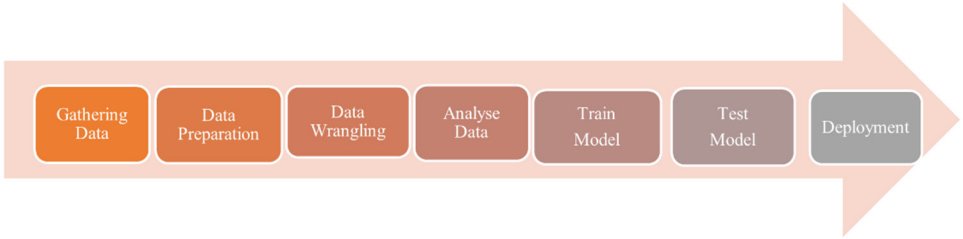

Fig. 2. Process of ML.

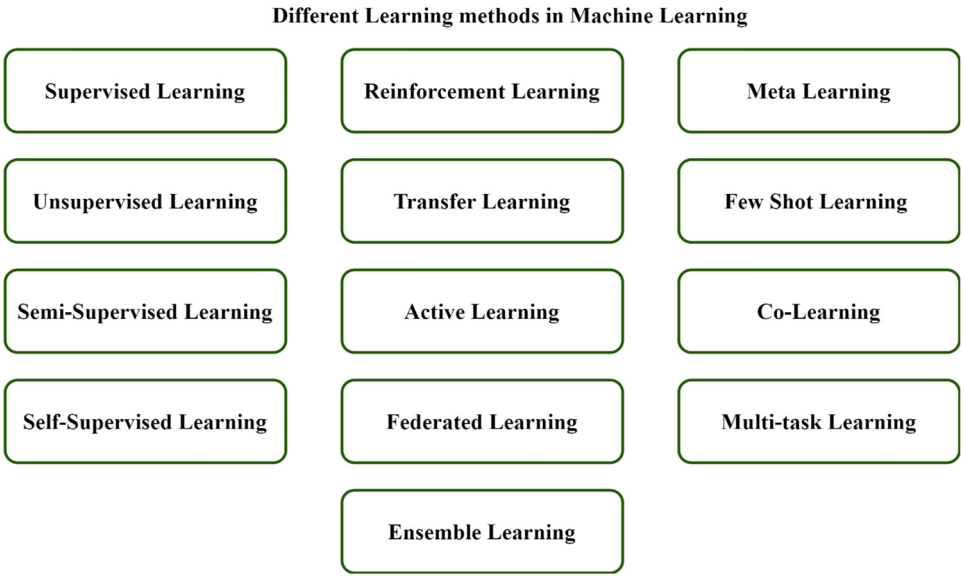

Fig. 3. Different learning methods in ML.

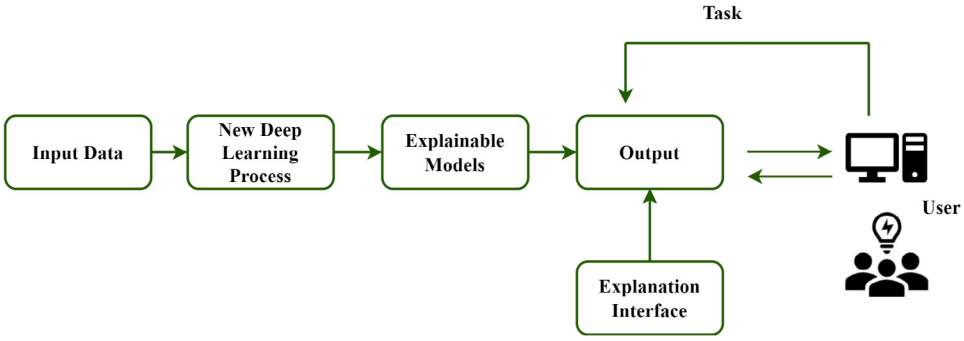

Fig. 4. Process of Explainable AI.

specific", "class activation map" (CAM), "layer-wise relevance propagation" (LRP), "gradient weighted CAM", Cluster-based, Filter-based, Attention Mechanism, Rule-Based, Knowledge-Based, Interpretable Model, Autoencoder-Based, Tree-Based [12].

Fig. 5 explains the explainable AI in different application domains. In this article, we explore XAI-related articles that were recently published on different applications and explain them using pie charts.

Most of the published articles were taken from the year 2022. 38 articles were taken in the year 2022, 28 articles from the year 2021, and 10 articles from the year 2020 and 10 articles from the year 2019, 5 articles from the year 2018 as mentioned in Fig. 6.

A PRISMA flowchart is used to select the articles from (2018, 2019, 2020, 2021, and 2022 of October), as mentioned in Fig. 7. XAI methodologies and their applications were discussed in this article, as were the taxonomy, principles, properties, concepts, and methods of XAI. This will allow the authors to explore the ongoing trends in XAI that can be identified, and it will be helpful to develop the new methodology using the XAI taxonomy.

2. Related works of XAI on different application domains

Explains about the growth of development and deployment of XAI in the recent years. Based on the domains, the related articles have been segregated.

2.1. Explainable AI in agriculture

Kaihua Wei, Bojian Chen, et al. [13] explored XAI in the agricultural classification field using DL models to detect Leaf Disease Classification. Five leaves dataset is used. Several categories of dataset have been arranged into 3 experiments such as VGG, GoogLeNet and ResNet models, respectively. In that, ResNet-attention model utilised with 3

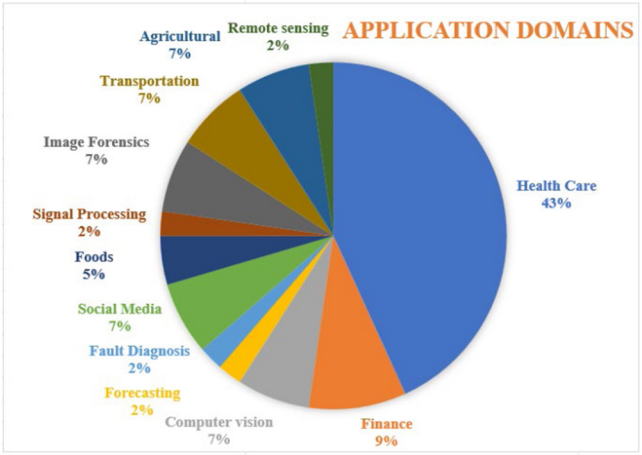

Fig. 5. Pie chart representation of XAI on different application domains with percentage.

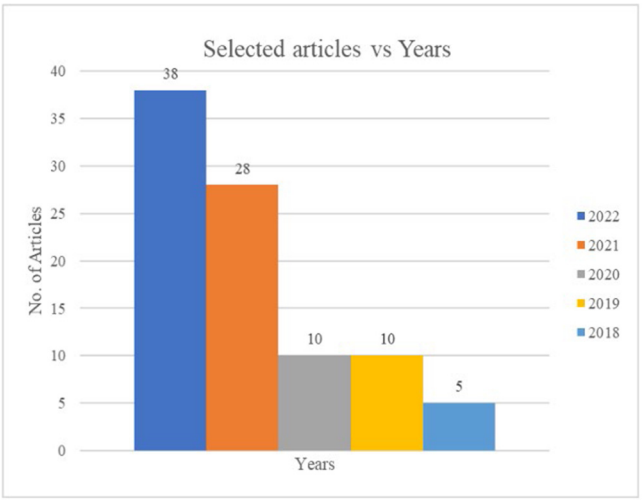

Fig. 6. Number of selected articles based on the year.

interpretable methods and thus it shows result of greatest accuracy rate of 99.11%, 99.4%, and 99.89% in the 3 experiments. Attention module also used to improve the feature extraction and clarify the focus of the model

2.2. Explainable AI in computer vision

Joshi et al. [14] Deep neural networks play an important role in Computers vision, NLP tasks, and many other domains, and researchers have investigated about Multimodal AI with XAI for better interpretability and understanding of the model. Hamad Naeem et al. [15] Inception-v3's CNN-based transfer-learned model is proposed to identify the malware using colour image malware display in Android's (DEX) Dalvik Executable File. Markus Langer et al. [16] explained the XAI concepts from the stakeholders' perspective. Gulsum Alicioglu, Bo Sun [17] Visual analytics been used for the well understanding of neutral networks for the end-users via XAI methods. Dang Minh et al. [18] explain the XAI methods in terms of three groups: pre-modelling explainability, interpretable models, and post-modelling explainability. Savita Walia et al. [19] The ResNet-50 architecture obtained an accuracy of more than 98% with different datasets to detect the manipulation of images. Ahmed Y. Al Hammadi et al. [20] proposed explainable with DL and ML models with EEG signals, used to identify the industrial insider threats.

2.3. Explainable AI in finance

Tanusree De et al. [21] proposed a method with combinations of clustering of the network's hidden layer representation and TREPAN decision tree and UCI Machine-Learning-Repository data sets used to predict the credit card default application. Implemented the methods using python programming language in PyCharm IDE. This method able to create better quality reason codes to make the human to understand the prediction of outcome from the neural network model. Future approach is to implement this method to the other machine learning algorithms.

2.4. Explainable AI in forecasting

Joze M. Rozanec et al. [22] proposed an architecture for the XAI using semantic and AI technologies, and it is used to detect demand forecasting and deployed in the real world. The Knowledge graph is used to provide the explanation about the process of demand forecasting at a higher level of explanation than the specific features. Hence, it is used to hide the sensitive information about the forecasting models used to provide the confidentiality. Their future work is to improve the quality of explanation in the domain of demand forecasting and media events Han-Yun Chen et al. [23] proposed XAI methods for the vibration signal analysis of the CNN model using fault classification. Initially, signals are converted into images using STFT called Short-time Fourier transform then the input is given to the CNN as classification model for the signal analysis with Grad-CAM, then the explanations are verified by using neural networks, adaptive network-based fuzzy inference system (ANFIS) and decision trees

2.5. Explainable AI in the healthcare domain

Health Care data [24] are collected from many different sources. Some data is collected directly from clinical trials and research, and some other data is collected via sensor. DenseNet and Convolutional Neural Network (CNN) models have been developed by V. Jahmunah et al. [25] to predict myocardial infarction (MI). An enhanced technique of Class Activation Maps (CAM) called Gradient-weighted CAM is used to visualise the productivity of the classification task for both models. This approach had the potential to diagnose MI in hospitals. AI and ML play a vital role in the health care domain. To gain the trust of AI models, XAI is used. It will bring the trust in the predictive modelling on the practical situation [24]. Jaishree Meena et al. [26] done the investigation about skin cancers of non-melanoma skin cancers for both the men and women. The Python SHAP (SHapley Additive exPlanations) programming language has been used on trained XGboost ML models for identifying the biomarkers for predicting skin cancers. Miquel Miro-Nicolau et al. [27] investigated the X-ray-image-related task using Explainable AI. Class Activation Maps (CAM) have been used to represent X-ray images. Lombardi, A. et al. [28] have investigated to predict the AD called as Alzheimer's-Disease and cognitive impairment with XAI. The ADNIMERGE R package and Random Forest Classifiers have been used to predict the disease.

Chang Hu et al. [29] proposed a ML model for the readmission of a septic patient in the Intensive Care Unit and SHAP values used to extract the related features which are used for accurate prediction and LIME used to explain about the models. Djordje Slijepcevic et al. [30] developed a Layer-wise Relevance Propagation method with XAI to analyse the clinical gait on a time series. They introduced XAI for the time series classification and will promote in future to detect the clinical gait for the accurate prediction. Jeremy Petch et al. [31] have been given information about the concepts and techniques of XAI in the field of cardiology. Nor et al. [32] explain the XAI in the fields of

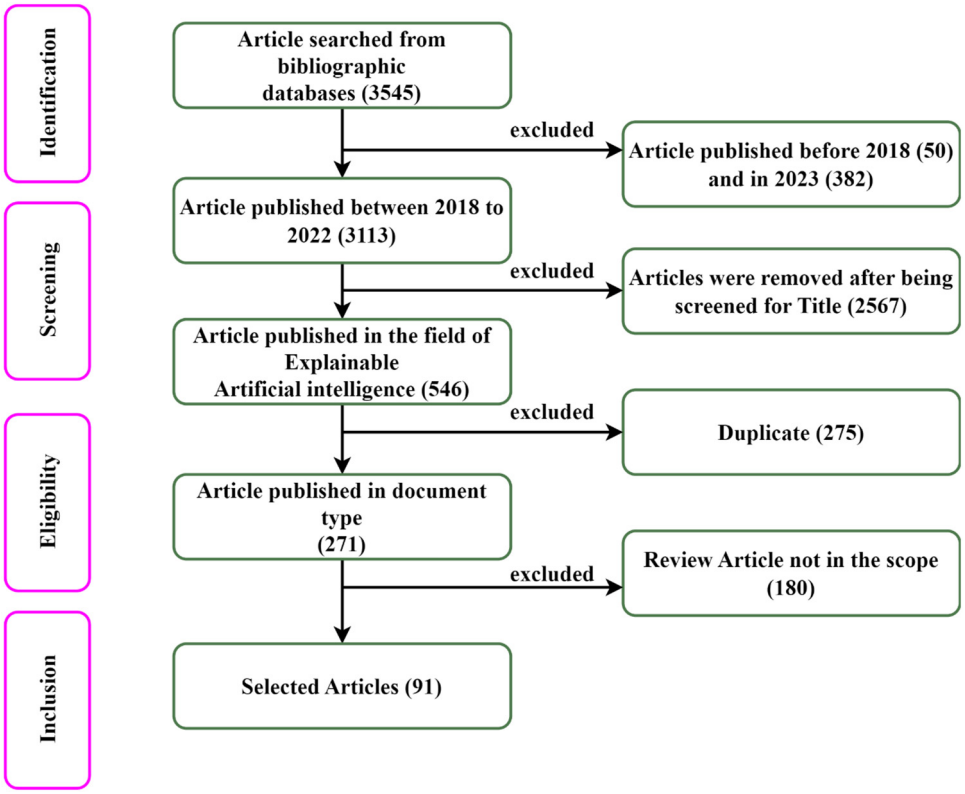

Fig. 7. Article selection using the PRISMA regulations.

medicine, psychology, clinical traits, and others. Prognostics and Health Management (PHM), however, failed to provide the analytical process of how XAI works in PHM. XAI used to provide knowledge about the diagnosis and detection of abnormal activities. Marwa Obayya et al. [33] proposed XAI methods in the field of teleophthalmology to reduce the waiting time of the patients, improve the services, increase accuracy, increase speed, and increase productivity in the field of telehealth. Using classification model, obtained 98.24% of accuracy. Nikolaos I. Papandrianos et al. [34] developed a deep XAI method used to predict CAD (coronary artery disease) by using the SPECT MPI images. David Pertzborn et al. [35] proposed a deep learning model for the detection of cancers. Using a deep learning model, researchers obtained 80% greater accuracy with a little bit of pre-processing data along with explainable artificial intelligence.

Ramy A. Zeineldin et al. [36] proposed the NeuroXAI framework for analysing the brain image and developed explanation methods for the visualisation map to diagnose and detect the brain tumours in the clinical sectors. Atul Anand et al. [37] proposed a number of deep neural networks using the PTB-XL dataset, which was available publicly, for the recognition of cardiac disorders using ECG signals. Giorgio Leonardi et al. [38] proposed CNN classifiers for the detection of strokes and developed and trace saliency maps used to trace the output of the model to make the model explainable. Zia U. Ahmed et al. [39] developed a stack-ensemble Machine Learning model outline with XAI that can visualise the spatial distribution of lung and bronchus cancer (LBC) and could visualise the relationship between the risk factors for LBC. Michael Merry et al. [40] proposed a new definition of explanation of XAI. Hao Sen Andrew Fang et al. [41] Clinical risk prediction models (CRPMs) uses the characteristic of patients to find the chance of evolving the particular disease but it is failed to had a scope in clinical practice due to lack of transparency. Andreu-Perez et al. [42] proposed multivariate pattern analysis for functional near-infrared spectroscopy (fNIRS) with XAI, which had been developed for the study of the development of the human brain in infants.

2.6. Explainable AI on remote sensing and signal processing

Dongha Kim and Jongsoo Lee [43] proposed a method called the optimal data augmentation method with XAI. A CNN model was developed to detect the quality of vehicle sounds. Optimal data augmentation method developed based on changes obtained for each selected characteristic. Optimal data augmentation method obtained 94.22% of accuracy than the existing method with improvement of 1.55–5.55%. On account of datasets used, the accuracy of standard deviation obtained was 2.13% which is more accurate result. Giorgio Leonardi et al. [44] proposed the XAI method with a classification task of remote sensing using an explanation of the model used a trained DL model with datasets. 10 Explainable AI methods been used in the field of remote sensing along with performance metrics to find the method's performance. Many experiments were done to analyse the overall performance of XAI in different cases (e.g., misclassification, multi-labels and prediction models). Grad-CAM given high-resolution outputs with 0.03 computational time apart from ten XAI methods.

Table 1. Describes the related articles that were taken for review and explores the data about the methods and applications. From the Refs. [45–57], methods and applications were taken from Ref. [63] to compare the methods and domains with recently published articles that we were considering for the systematic review.

2.7. Explainable AI in social media

Harshkumar Mehta and Kalpdrum Passi [58] proposed XAI methods to detect hate speech using DL models and used them to explain how the complex model of AI works with interpretation and explanation. Google Jigsaw and HateXplain datasets were used to detect the hate speech using XAI and LSTM reached 97.6% of accurate outcome. LIME XAI method applied to HateXplain dataset. Variants of BERT (bidirectional-encoder-representations from transformers) + ANN (Artificial-Neural-Network) achieved 93.55% of accuracy and BERT + MLP (Multilayer-perceptron) attained 93.67% of accuracy in terms of

Table 1
List of various applications using XAI models.

References and domain	Architecture	XAI models	Applications
Agriculture [13]	Very deep convolutional networks, GoogLeNet, ResNet models	Gradient-weighted class activation mapping, SmoothGrad, Local interpretable model-agnostic explanations	Leaf disease classification
Computer vision [15]	CNN, Inception-v3	Gradient-weighted class activation mapping	Malware detection
[19]	ResNet-50	Kernel-Shapley Additive exPlanations	To identify the manipulations in an image
[20]	One or Two-dimensional CNN, Adaptive boosting, Random-forest classifier, K-nearest neighbours	Shapley Additive exPlanations	To evaluate industrial internal security
Finance [21]	Feed forward Neural network	TREPAN model	Business applications
Forecasting [22]	Deep learning	Knowledge graph	To detect the demand forecasting
[23]	CNN	Gradient-weighted class activation mapping	Vibration signal analysis
Healthcare [45]	CNN	Gradient-weighted class activation mapping, Guided back-propagation.	Predicting brain tumour grade from imaging data
[46]	CNN	Visualising feature maps	Skin lesion classification
[47]	CNN	Gradient input, Guided backpropagation, layer-wise relevance, propagation, and occlusion	Alzheimer's disease classification
[48]	Inception version 4 model	Integrated gradients	Grading for diabetic retinopathy
[49]	CNN: VGG16 and GoogleLeNet	Expressive gradients	Age-related macular degeneration
[50]	Deep learning	Grad-CAM, Kernel SHAP	Skin cancer classification
[51]	CNN, fully convolutional network, U-shape network, other hybrid computational methods	Guided back-propagation, and Shapley Additive explanations	Diagnosis of ophthalmic diseases
[52]	Inception version 3	Attribution based XAI	Classification of retinal disease
[53]	Deep-learning convolutional neural networks, AlexNet	Integrated gradients attribution method, and smooth-grad noise reduction algorithm	Classification of estrogen receptor status from breast MRI
[54]	Fully convolutional network	Guided back-propagation	Segmentation of colorectal polyps
[55]	Fully convolutional neural networks	Shape attentive U-Net	Volume estimation of cardiac bi-ventricular
[56]	DL model ASD-DiagNet	Auto-ASD-Network	Diagnosing autism spectrum disorder
[57]	Six-layer convolutional neural network	Gradient-weighted class activation mapping	Detection of Covid-19
[25]	DenseNet and CNN	Gradient-weighted class activation mapping	To diagnose myocardial infarction
[26]	XGboost ML	Python Shapley Additive exPlanations	Non-melanoma skin cancers
[27] [28]	Deep learning Threefold classification	Class activation maps Shapley Additive exPlanations	X-ray-image related task To diagnosis the Alzheimer's disease
[29]	9- Machine learning models	Local interpretable model-agnostic explanations, Shapley Additive exPlanations	Readmission of a septic patient in the ICU
[30]	CNN, Support vector machine, and Multi-layer perceptron classification	Layer-wise relevance propagation	To analyse the clinical gait on time series

(*continued on next page*)

Table 1 (*continued*).

References and domain	Architecture	XAI models	Applications
[34]	CNN	Gradient-weighted class activation mapping	To predict coronary artery disease
[35]	Deep learning	densMAP	Salivary gland carcinomas
[36]	CNN	NeuroXAI framework	To diagnose and detect the brain tumours in the clinical sectors
[37]	Deep neural networks	Shapley Additive exPlanations	To detect and diagnose the cardiac disorders
[38]	CNN classifiers	Saliency map	Detection of strokes
[39]	Stack-ensemble-machine learning model	Model-agnostic approaches	Visualise the spatial distribution of Lung and bronchus cancer
[41]	Modified KNN	Mental model	Clinical risk prediction
Social media [58]	Deep learning	Local interpretable model-agnostic explanations	To detect the hate speech
[59]	Deep learning	Post-hoc	Deepfake voice detection
[60]	Deep learning, Natural language processing	Local interpretable model-agnostic explanations, Anchors	Fake news detection
[61]	Deep learning	Shapley Additive exPlanations, Local interpretable Model-agnostic explanations	Food delivery service
Signal processing [43]	CNN model	Local interpretable Model-agnostic explanations	Vehicle sound quality
[44]	Deep learning	Gradient-weighted class activation mapping	Classification task of remote sensing
Transportation [62]	Deep learning	Sensitivity analysis technique	Self-driving cars

explainability using the (Evaluating-rationales and simple English reasoning) ERASER. Suk-Young Lim et al. [59] presented the LIME and anchors XAI method for deepfake audio detection. Mateusz Szczepański1 et al. [60] proposed novel XAI approach in BERT-based fake news detectors. Anirban Adak et al. [61] proposed SHAP and LIME XAI methods for the Food Delivery Service (FDS) using DL models. XAI is used in the recommendation system [64].

2.8. Explainable AI in transportation

Hong-Sik Kim and Inwhee Joe [62] separate the category of images with more accurate in self-driving cars. Sensitivity Analysis technique used to explain the necessary part of authentic images are used to separate the category. Thus, the XAI needed to separate the category of images with more accurate in self-driving cars

3. Review on explainable AI

This section describes the data about the methods to build the explainability taxonomies, needs, properties, principles, concepts, and common methodology of explainable AI.

3.1. Ideas associated with the concept of explainability

A form of work focused on studying and trying to express the perception of explainability, which led to different kinds of description and the development of some features and structures. Some of the groups have been developed as given below.

- "Attributes by Explainability"
- "Types of explanation",
- "Structure of an explanation" [65].

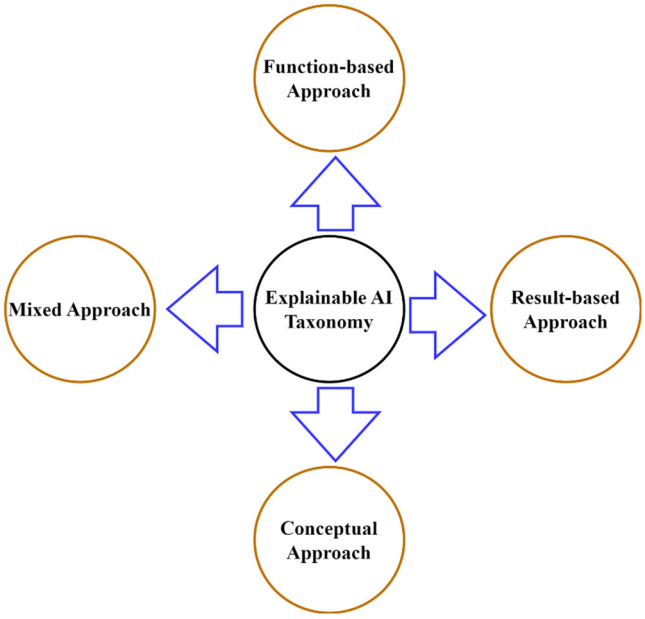

Fig. 8. Methods for building the explainable taxonomy.

3.2. Methods for building explainability taxonomies

To build the explainability taxonomy, four common approaches are being used and described in Fig. 8 [66].

Functioning-based approach:

A function-based approach works based on the basic functions of an explainability method as the important component of the classification model. Functions used to extract information about the model.

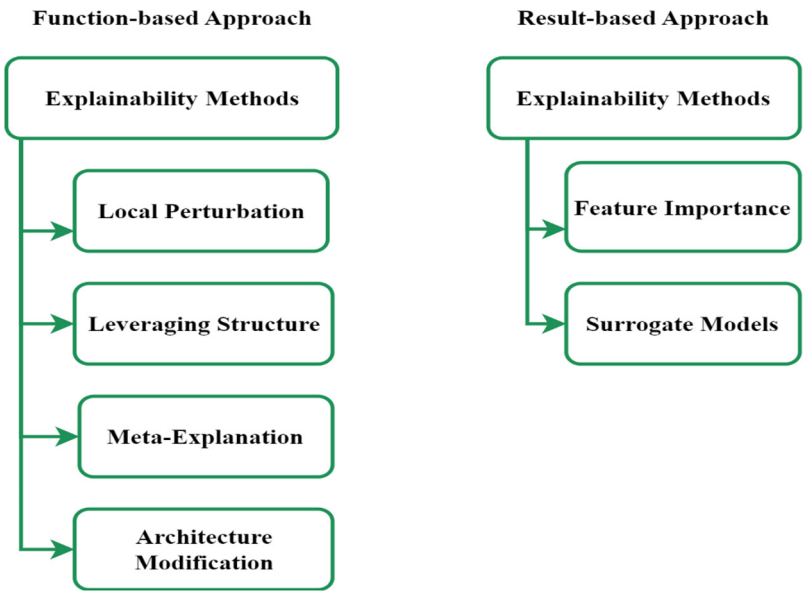

Fig. 9. XAI Taxonomy of function-based and result-based approach [66].

Explaining the model with *local perturbation* for the classification model: i.e., in order to predict the important feature in an input, it is used to perturb the input value, which is called local perturbation.

Leveraging structure is used to explain some important properties of models that are necessary to construct the explanation. Methods that are used to influence the structure are called "*feature importance* attributes".

Meta-explanation was used to aggregate the explanation for the better performance of the methods than the other method used. The local perturbation, feature importance, and meta-explanation are the explanations that represent the functions of the model alone.

Architecture Modification is used to simplify the complex structure by changing the architecture.

Fig. 9 describes the taxonomy of explainable AI. First one is function-based approach and another one is result-based.

Result-based Approach:

The *Result-based Approach* works based on the result of an explainability model, which is an important component of a classification model.

Feature Importance refers to the important feature for the output of an input.

Surrogate Model try to construct the model which are specific part of original model. It can be produced in different ways, and it can search for the innovative model through perturbations and the structure of leveraging. Hence, it depends on the result of functioning [66]. LIME (Local Interpretable Model-agnostic Explanation) is an example of a surrogate model [67], and it gives a solution to the surrogate model [68].

Conceptual Approach:

The *conceptual approach was* used to split up the classification of the method into many different *conceptual dimensions.* Based on the study, XAI methods are categorised into different types, as explained in Fig. 10.

Stages of Explanation Modelling:

They are 3 levels of explanation in XAI and two types of stages are present in XAI. Ante-hoc and post-hoc methods Pre-Modelling and During Modelling comes under ante-hoc methods.

Ante-hoc methods:

- Pre-modelling: A pre-modelling explanation is used to explain the development process of a model. It includes the kind of understanding and analysis of the data using explanatory data, dataset documentation, summarisation, and design to ensure clarity [14].
- During modelling, it explains the design of the model.

Ante-hoc method does not require any approach for an explanation. During training itself, model comes along with explanation.

Post-hoc modelling: it is proposed after the model is developed. The post-hoc method requires an additional approach to extract the explanation.

Scope of XAI:

Global and local are the two types of scope of explainability [69,70]. It depends on whether the description is extracted from a single instance or the whole model [14]. During the global scope of explanation, derive the explanation from the whole model and make it transparent to the user [71]. On the other hand, during the local scope of explanation, derive the explanation from a single instance of a model.

Problem Type of XAI:

Based on the problem, methods for explainability getting vary: *Classification* or *Regression*.

Input Data:

Based on the different application domains, it demands input data in different types: *numerical, categorical, pictorial, textual, time series, and vectors.* It is used to construct the methods of XAI [69].

Output Format:

Like input data, different types of output formats are required for the explanation based on the domains: *numerical, rule-based, textual, visual, or mixed* [69].

Mixed Approach

The mixed approach constructed with all the above 3 approaches like function-based, result-based and conceptual approach.

First level in mixed approach taxonomy represents the conceptual approaches and applicability like model-agnostic and model-specific as mentioned in Fig. 11. On the lower level, the applicability of the conceptual approach is connected with *"Explanation by Simplification,*

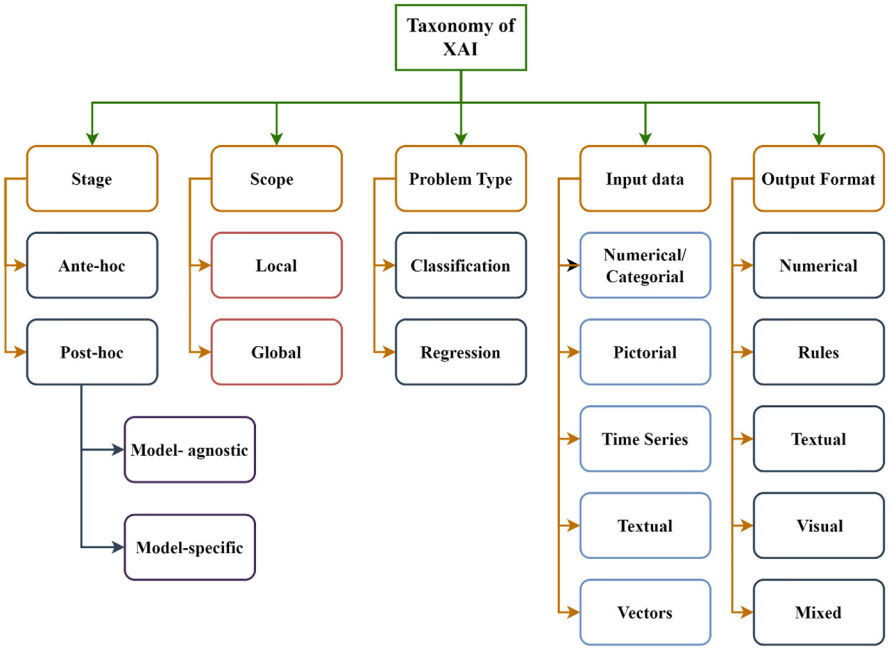

Fig. 10. XAI taxonomy of conceptual approach [69].

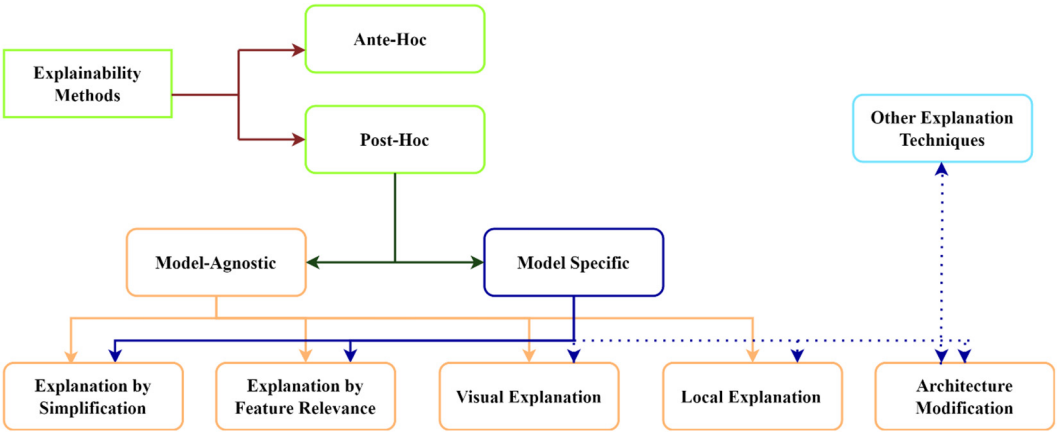

Fig. 11. XAI taxonomy of mixed approach [66]. In some cases, ante-hoc requires post-hoc methods (i.e., represented in dotted lines)

Explanation by Feature Relevance, Visual Explanation, Local Explanation and Architecture Modification" [66]. This mixed approach is the best one for new XAI developers when compared to other approaches.

3.3. Methods

Some of the Explainable AI methods are discussed below. The methods are SHapely Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), Gradient-weighted Class Activation Mapping (GradCAM) [72], Layer-wise Relevance Propagation (LRP), Fuzzy Classifier [32,73], and Partial Dependence Plots (PDP) [74].

Fuzzy Classifier: It results from fuzzy logic that copies human decision-making with different opportunities among yes and "no" [75], and it has four components: *"knowledge base, fuzzifier, fuzzy inference engine, and de-fuzzifier"*.

Knowledge base: it holds the *conditional if-then rules* given by human experts.

Fuzzifier: input data can be turns into the *input fuzzy sets* (small, medium negative, large negative, medium positive, largely positive).

Fuzzy inference engine: To generate *fuzzy output sets, the* inference engine will match the fuzzy input sets that were produced by the fuzzifier to the knowledge base rules.

De-fuzzifier: finally, *crisp values* will be produced by using output sets from the fuzzy inference engine [73].

A fuzzy classifier is one of the most transparent methods in XAI [76].

Gradient-weighted Class Activation Mapping (Grad-CAM):

Grad-CAM is an enhanced version of Class Activation Mapping and it is designed to differentiate the partial feature critical for the prediction of model, this is mostly interpreted with medical imaging technique and clinic. It is applicable to CNN (convolutional neural network) models alone without fully connected layers. It is good at the final convolutional layer. Without losing the complexity of Grad-CAM, used to preserve the CNN models' interpretability and generate heatmaps [73].

Grad-CAM is used to provide a visual explanation of the CNN model, and it uses the class-specific gradient information (because of the gradient-related method) going to the final convolution layer to produce the localisation map in the image of an important region for classification models and make it more transparent. In order to build a high-resolution class-discriminative visualisation, the Grad-CAM authors also showed how the technique can be integrated with currently available pixel-space visualisations, *Guided Grad-CAM*. Limitations of *Grad-CAM and CAM*: Due to its partial derivative assumptions, it is unable to localise objects that appear in many places in an image. Due to the frequent up- and down-sampling processes, it is unable to exactly regulate class-region coverage in an image and the probable loss in signal [77].

Grad-CAM++ is an upgradation of Grad-CAM methods and it provides the good visual explanation of CNN models. It is more reliable in multi-label classification problems, whereas the varying weight given to each pixel enables the gradient feature map to record the significance of each pixel distinctly [77].

Grad-CAM is a post-hoc explanation, and it reproduces the values from the output layer towards the final convolutional layer [78].

Layer-wise Relevance Propagation (LRP)

LRP, like a Grad-CAM, can also generate the heatmap based on the critical regions. LRP goes backward from the model's classifier layer by first calculating the relevance score for a particular output, which is also a final layer. Until it reaches the input image, it continues backward to calculate the relevance score of individual neurons in each layer. At that moment, it generates a heatmap based on the relevance score to highlight the significant regions that are necessary for the prediction. It has three rules, namely *"basic rules", "epsilon rules",* and "gamma rules", and these can be applied to the model of the upper layer, middle layer, and lower layer, respectively. But it has less exposure when compared to the Grad-CAM [73].

LRP does not depend on gradient information, but it defines the relevance score for the neuron's output [78]. To allocate a relevance score to each individual feature, it uses redistribution rules for the prediction of a model [79].

Local Interpretable Model-agnostic Explanations (LIME):

It is used to generate only local explanations of the model and provides the explanation with significant features that are necessary for prediction. Like SHAP, it does not depend on Game theory. Hence, it refers to the direct approach by varying the input data of models, and it will observe the changes made by the prediction of the model, and the explanations rely on single instances instead of a whole data set [73].

An example of a surrogate model is LIME, which is used to predict the opaque model. It will the train the model locally and gives the explanation about individual prediction of model. The main drawbacks of LIME are that it is "non-deterministic" due to perturbing the points randomly and lacking "stability". Muhammad Rehman Zafar presents the DLIME, or *Deterministic Local Interpretable Model-agnostic Explanations,* used to produce the reliable explanation for a test instance [80].

To recover the stability of LIME, instead of using random perturbation, it uses hierarchical clustering to set the training data, then selects the more relevant cluster for the instances [81].

One of the most commonly used interpretable approaches for black-box models is the LIME method [77]. Two CNN models were presented explanations using the LIME methodology using Patch Camelyon dataset [82]. LIME gives the solution to the surrogate model [67]. LIME gives a local explanation alone. Hence, it is been enhanced into two variations [79]:

- Sub-modular pick LIME (sp-LIME)
- k-LIME

Sub-modular picks are used to deliver the global explanation of the model using sub-modular picks to find the occurrences of the prediction model. sp-LIME is a global surrogate, model-agnostic, and post-hoc, which develops the LIME methods.

k-LIME: It is applied to the black-box model, where the clustering algorithm is used to find the k-clusters.

Partial Dependence Plots (PDP):

It is one of the most common post-hoc methods. PDP will plot the effects of the feature subsets on the prediction of models in order to explain the black-box model [74]. A tool for showing the relationship between the response and predictor variables in a limited feature space is the partial dependence plot r returned by function f [83].

SHapely Additive explanation (SHAP):

SHAP is like a "*feature attribution mechanism*" [84]. SHAP is like a game theory method used to improve the interpretability of each specific prediction by calculating the significance values for each and every feature. Three important properties are upheld by the SHAP values as a combined degree of feature importance: *"accuracy, consistency, and missingness".* SHAP is easier to calculate and more natural with respect to interpretation [77]. It is used to generate both local and global explanations, is not only model-agnostic, and is more reliable with any type of data.

Shapley values are used to allocate the rewards for players by calculating the involvement of each player, and they satisfy the four axioms for calculating the involvement of each player: *"Efficiency, Symmetry, Dummy, Additive"* [73]. SHAP was coined by *Shapley* in *1951,* and SHAP is used to describe the particular output based on the involvement of each input in a prediction. It will apply to the surrogate models [79]. Proposed SHAP explanations into two different versions:

- Kernel SHAP,
- Deep SHAP.

Kernel SHAP: Shapley values and LIME are integrated in a *model-agnostic framework* for outcome explanation. It is similar to *Local Interpretable Model-agnostic Explanation (LIME)*; it is also used to train the local surrogate model, but the model is limited to weighted linear regressions, and the training is also done on different data sets. Further, Kernel SHAP studies the instances of interest. Shapley values used to give the chosen explanation subsequently reflect each feature's contribution to output. Kernel SHAP has a limitation; it is not suitable for global explanation models [81].

Deep SHAP is a *model-specific framework* used to explain the output of DL models by integrating the Shapley values with DeepLIFT. Deep SHAP used the idea of DeepLIFT, but the difference is instead of assigning a value to input feature, it steals idea from game theory by each input feature of player and their involvement in a game using the Shapley value [81].

SHAP able to provide the consistent results and the drawbacks is the high calculation speed [85]. Proposed Tree-Explainer to solve the problem for tree-based Artificial Intelligence algorithms like *Random Forest, Decision Tree and Gradient boosted trees. SHAP and Tree Explainer* are similar based on their functionality. In Tree Explainer, predictions can be made tree-based with respect to the reference point of artificial intelligence and also used to analyse global data [85]. Shap explanations can be computed in polynomial time by the algorithm [84]. Compared to LIME scores, SHAP values demonstrated much greater intra-consistency values [86]. Lev Utkin et al. [87] categorised SHAP into Ensemble of Random SHAPs, Ensemble of Random weighted SHAPs, Ensemble of Random SHAPs generated by the Random Forest.

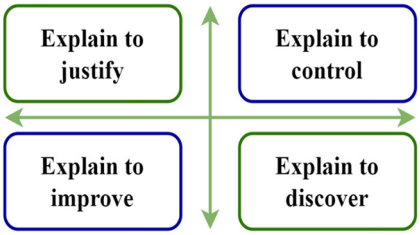

Fig. 12. Need for XAI [88].

3.4. Need for explainable AI

XAI is necessary and emerging, especially in the field of medical research, which will increase trust in AI. Because, explainability leads to generate the insights in the process of models and also explain about the results which leads to better knowledge towards the problem statement. Hence, it is helpful to improve the model and provide better application domain and also XAI explain the answer for the question "why" which is able to provide in the traditional model that gives trust towards AI using XAI [28] used to lessen the impact of model bias.

Explainable AI will be helpful for making the collaboration between the professionals of AI and outside gatherings and will support for the development of model, when it is not dependable yet to be measured [12].

There are four common reasons why we need to go for XAI (see Fig. 12).

Explain to Justify: To justify the result, explainable AI is used to deliver the information, mainly when it comes to real-world critical situations with the name of instances called "Right to Explanation" guidelines involved in GDPR (General Data Protection Regulation).

Explain to Control: Apart from justification, it is also necessary to control the explanation from going wrong and used to identify the vulnerabilities and flaws.

Explain to Improve: Another needs to build an explainable model is to improve them continuously by knowing the flaws. This can be achieved by using the explainable model.

Explain to Discover: It is used to discover the new facts for the problem, thus gaining knowledge by gathering the information, which will create a new insight.

Also, explanations can be categorised into three dimensions called source, scope, and depth [89] (see Fig. 13).

3.5. Principles of explainable AI

National Institute of Standards and Technology (NIST) conducted a workshop with AI community and proposed *4* Explainable AI principles. They are

"Explanation, Meaningful, Accuracy, Knowledge limits" [90]

Explanation: The system provides an explanation for the process and output of the model. The process states that the actions, design and workflow of the system and the output states that the outcome or the action performed by the system. The output will differ according to the task which we have taken [65].

Meaningful: The system offers explanations about the system that are interpretable to the proposed stakeholders.

Explanation Accuracy: The Principle of *Explanation Accuracy* imposes accuracy on an explanation of the system. Both the explanation and meaningful principles refer to the explanation, which does not verify the quality of the explanation. Because of these two principles, we do not need to explain the process of the system appropriately for its outcome. Explanation accuracy is different from decision accuracy. Decision accuracy defines whether the system's judgement is correct or incorrect.

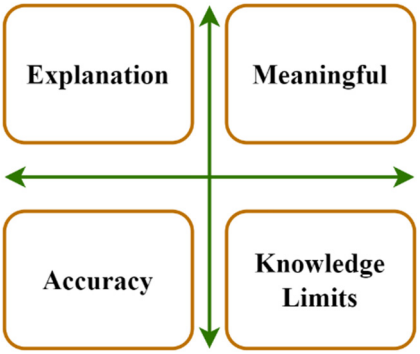

Fig. 13. Principles of Explainable AI.

Knowledge Limits: According to the knowledge limits principle, systems detect situations for which they were not intended or approved to function or in which the answers they provide are not trustworthy. By preventing inaccurate, unsafe, or biased outcomes, this principle can raise the level of trust in a system.

3.6. Properties of explanation

Properties of explanations can be categorised into two types based on the explanation. They are called as, Styles and purposes of explanation.

Styles: The properties of style labels and how an explanation is provided

Purpose: The purpose describes why the user needs an explanation.

The properties of styles are further divided into 3 types of elements as mentioned in Fig. 14. They are

- *Level of detail:* It describes the level, i.e., from "*sparse to extensive*". Sparse will provide limited information, whereas extensive will give detailed information.
- *Degree of interaction between the human and the machine:* It is divided into three types. "*Declarative explanation, one-way, and two-way interaction*".
- *Formats:* It includes information about the "*visual/graphical, verbal, and auditory/visual alerts*" [90].

4. Challenges of explainable AI

Explainable Artificial Intelligence idea highlights interconnected research issues:

- how to create models that are easier to explain,
- how to develop explanation interfaces, and
- how to comprehend the psychological conditions necessary for persuasive explanations [91].

5. Discussion

It is difficult to examine the opaque black box model and cumulative model complexity can be used to achieve the more accurate prediction in the field of Artificial intelligence. That is a significant issue that, if it is used in the real world, might have a considerable impact on the system's performance. We need the transparency to trust the model, especially when they are used for high-stakes application such as healthcare, finance or security.

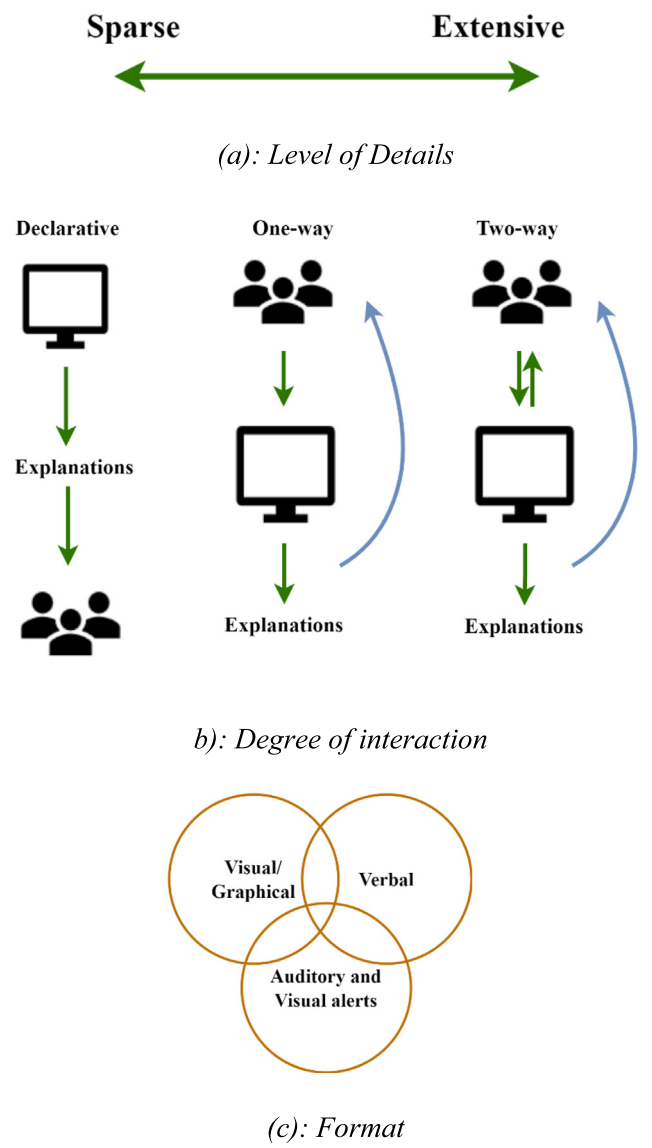

Fig. 14. Diagram of styles of explanation [90].

Deep Learning is gaining more attention because of its quality in terms of accuracy, having been trained with a larger amount of data, but it is considered an opaque model because we do not have any knowledge about the output of the model or how it was reached. To understand the output and process of the model, we are moving towards "Explainable artificial intelligence".

The study's findings make it clear that the impact of XAI on different applications is diverse. Further, we framed the table based on the applications with different methodologies used in the research and methods to build the XAI taxonomy with four common approaches, Ideas associated with the concept of explainability, methodology, the need for XAI, the principles of explainable AI, the properties of explanation, and challenges.

Different methods have been used to get the explanation of the model in different output formats (Numerical; Rule-based; Textual; Visual; Mixed). Methods have been selected based on the problem statement. The explanation may be varied according to the scope of the model. Post-hoc and ante-hoc methods have been used for stages of explanation.

SHapely Additive explanation is used to calculate the significance values for each and every feature and displays the impact of a feature on the prediction. *Local Interpretable Model-agnostic Explanations* offer excellent visuals. *Gradient-weighted Class Activation Mapping* is more transparent and produces a visual explanation as well. *Layer-wise Relevance Propagation* generates a heatmap and defines the relevance score. *A fuzzy classifier works based on if-then rules. Partial Dependence Plots show the connection between the target and a feature in a direct manner.*

Methods for building the explainability taxonomy. Four common approaches have been used (*Functioning-based; Result-based; Conceptual Approach; Mixed Approach)*

Is it true both function-based and result-based approaches are required to build the taxonomies of explainability methods? (Research Question 1). Local-perturbation techniques can produce surrogate models in addition to feature representation. Yet, in addition to being the outcome of local perturbations, surrogate models can also be produced by using a model's structure. Hence, to build the taxonomy, either functioning-based or result-based is not enough to sufficiently categorise a method, and it might withhold significant information. So, both approaches are required to classify a method.

Research question 2: Is exclusion of methods possible in the mixed approach? Some of the explanation methods do not require all the taxonomies. Model-specific methods can exclude local or visual description, but there is no chance to remove it. Model-specific depends on one output, but it produces both the outputs (local or visual) due to the mixed approach.

In this article, we reviewed the recently published manuscript using Explainable AI. However, greater study has been done in the domain of healthcare when compared to other high-stakes domains such as finance, industry, and academics due to recent trends in XAI.

5.1. Issues and future work

This paper offers a systematic literature review of XAI approaches with different applications published from 2018 to 2022 and finds a gap in the research are the datasets used for research are imbalanced and public datasets have been used instead of real-world data. The noisy data have used for the research. Transcriptome data may have certain inherent biases. Certain prospective characteristics that are strongly predictive were not captured by the study. Noisy saliency maps are produced by Vanilla Gradient. Guided backpropagation is limited to CNN.

Future work: In the future, we plan to do research using Explainable artificial intelligence in sensitive domains. Hence, this review will be more helpful to find the concepts and challenges of XAI in a critical situation where the users must take the correct decision based on knowledge about the output and process of the model given by the Explanation Interface.

6. Conclusion

This paper offered a systematic review of XAI in different applications published from 2018 to 2022 and observed 91 published articles on the development of XAI. In this study, our contributions are that we have listed the methods and applications that have been used in recent research and explained the methods used to build the XAI taxonomy with four common approaches: *Functioning-based; Result-based; Conceptual approach; Mixed approach; color-4the* concept of explainability; methodology; the need for XAI; the principles of explainable AI; the properties of explanation; and challenges have been described in this article, and it will be useful for the researchers to know about the recent developments on XAI. This will encourage the researchers to develop new techniques for XAI in high-stakes domains.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] K.P. Exarchos, et al., Review of artificial intelligence techniques in chronic obstructive lung disease, IEEE J. Biomed. Health Inform. 26 (5) (2022) 2331–2338, http://dx.doi.org/10.1109/JBHI.2021.3135838.

[2] F. Shi, et al., Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19, IEEE Rev. Biomed. Eng. 14 (2021) 4–15, http://dx.doi.org/10.1109/RBME.2020.2987975.

[3] E. Mohammadi, M. Alizadeh, M. Asgarimoghaddam, X. Wang, M.G. Simões, A review on application of artificial intelligence techniques in microgrids, IEEE J. Emerg. Sel. Top. Ind. Electron. 3 (4) (2022) 878–890, http://dx.doi.org/10.1109/JESTIE.2022.3198504.

[4] M.-P. Hosseini, A. Hosseini, K. Ahi, A review on machine learning for EEG signal processing in bioengineering, IEEE Rev. Biomed. Eng. 14 (2021) 204–218, http://dx.doi.org/10.1109/RBME.2020.2969915.

[5] Nabila Sabrin Sworna, A.K.M. Muzahidul Islam, Swakkhar Shatabda, Salekul Islam, Towards development of IoT-ML driven healthcare systems: A survey, J. Netw. Comput. Appl. 196 (2021).

[6] S.V. Mahadevkar, et al., A review on machine learning styles in computer vision—Techniques and future directions, IEEE Access 10 (2022) 107293–107329, http://dx.doi.org/10.1109/ACCESS.2022.3209825.

[7] Xiao Bai, et al., Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments, Pattern Recognit. 120 (2021) 108102.

[8] B. Goutam, M.F. Hashmi, Z.W. Geem, N.D. Bokde, A comprehensive review of deep learning strategies in retinal disease diagnosis using fundus images, IEEE Access 10 (2022) 57796–57823, http://dx.doi.org/10.1109/ACCESS.2022.3178372.

[9] R.I. Mukhamediev, Y. Popova, Y. Kuchin, E. Zaitseva, A. Kalimoldayev, A. Symagulov, V. Levashenko, F. Abdoldina, V. Gopejenko, K. Yakunin, et al., Review of artificial intelligence and machine learning technologies: Classification, restrictions, opportunities and challenges, Mathematics 10 (2022) 2552, http://dx.doi.org/10.3390/math10152552.

[10] Ning Wang, Yuanyuan Wang, Meng Joo Er, Review on deep learning techniques for marine object recognition: Architectures and algorithms, Control Eng. Pract., 118, http://dx.doi.org/10.1016/j.conengprac.2020.104458.

[11] D.V. Kute, B. Pradhan, N. Shukla, A. Alamri, Deep learning and explainable artificial intelligence techniques applied for detecting money laundering–a critical review, IEEE Access 9 (2021) 82300–82317, http://dx.doi.org/10.1109/ACCESS.2021.3086230.

[12] A.K.M. Nor, S.R. Pedapati, M. Muhammad, V. Leiva, Overview of explainable artificial intelligence for prognostic and health management of industrial assets based on preferred reporting items for systematic reviews and meta-analyses, Sensors 21 (2021) 8020, http://dx.doi.org/10.3390/s21238.

[13] K. Wei, B. Chen, J. Zhang, S. Fan, K. Wu, G. Liu, D. Chen, Explainable deep learning study for leaf disease classification, Agronomy 12 (2022) 1035, http://dx.doi.org/10.3390/agronomy12051035.

[14] G. Joshi, R. Walambe, K. Kotecha, A review on explainability in multimodal deep neural nets, IEEE Access 9 (2021) 59800–59821, http://dx.doi.org/10.1109/ACCESS.2021.3070212.

[15] Hamad Naeem, Bandar M. Alshammari, Farhan Ullah, Explainable artificial intelligence-based IoT device malware detection mechanism using image visualization and fine-tuned CNN-based transfer learning model, Comput. Intell. Neurosci. (2022) 7671967, http://dx.doi.org/10.1155/2022/7671967, 17 pages, 2022.

[16] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, Kevin Baum, What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research, Artificial Intelligence 296 (2021) http://dx.doi.org/10.1016/j.artint.2021.103473.

[17] Gulsum Alicioglu, Bo Sun, A survey of visual analytics for Explainable Artificial Intelligence methods, Comput. Graph. 102 (2022) 502–520, http://dx.doi.org/10.1016/j.cag.2021.09.002.

[18] D. Minh, H.X. Wang, Y.F. Li, et al., Explainable artificial intelligence: a comprehensive review, Artif. Intell. Rev. 55 (2022) 3503–3568, http://dx.doi.org/10.1007/s10462-021-10088-y.

[19] S. Walia, K. Kumar, S. Agarwal, H. Kim, Using XAI for deep learning-based image manipulation detection with Shapley additive explanation, Symmetry 14 (2022) 1611, http://dx.doi.org/10.3390/sym14081611.

[20] Ahmed Y. Al Hammadi, Chan Yeob Yeun, Ernesto Damiani, Paul D. Yoo, Jiankun Hu, Hyun Ku Yeun, Man-Sung Yim, Explainable artificial intelligence to evaluate industrial internal security using EEG signals in IoT framework, Ad Hoc Netw. 123 (2021) http://dx.doi.org/10.1016/j.adhoc.2021.102641.

[21] Tanusree De, Prasenjit Giri, Ahmeduvesh Mevawala, Ramyasri Nemani, Arati deo explainable AI: A hybrid approach to generate human-interpretable explanation for deep learning prediction, Procedia Comput. Sci. 168 (2020) 40–48.

[22] Joze M. Rozanec, Blaz Fortuna, Dunja Mladenic, Knowledge graph-based rich and confidentiality preserving Explainable Artificial Intelligence (XAI), Inf. Fusion 81 (2022) 91–102, http://dx.doi.org/10.1016/j.inffus.2021.11.015.

[23] H.-Y. Chen, C.-H. Lee, Vibration signals analysis by explainable artificial intelligence (XAI) approach: Application on bearing faults diagnosis, IEEE Access 8 (2020) 134246–134256, http://dx.doi.org/10.1109/ACCESS.2020.3006491.

[24] C.C. Yang, Explainable artificial intelligence for predictive modeling in healthcare, J. Healthc. Inform. Res. 6 (2022) 228–239, http://dx.doi.org/10.1007/s41666-022-00114-1.

[25] V. Jahmunah, E.Y.K. Ng, Ru-San Tan, Shu Lih Oh, U Rajendra Acharya, Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals, Comput. Biol. Med. 146 (2022) http://dx.doi.org/10.1016/j.compbiomed.2022.105550.

[26] Jaishree Meena, Yasha Hasija, Application of explainable artificial intelligence in the identification of Squamous Cell Carcinoma biomarkers, Comput. Biol. Med. 146 (2022) http://dx.doi.org/10.1016/j.compbiomed.2022.105505.

[27] Evaluating explainable artificial intelligence for X-ray image analysis, Appl. Sci. 12 (2022) 4459, http://dx.doi.org/10.3390/app12094459.

[28] A. Lombardi, D. Diacono, N. Amoroso, et al., A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of Mild Cognitive Impairment and Alzheimer's Disease, Brain Inf. 9 (17) (2022) http://dx.doi.org/10.1186/s40708-022-00165-5.

[29] Chang Hu, et al., Explainable machine-learning model for prediction of in-hospital mortality in septic patients requiring intensive care unit readmission, Infect. Dis. Ther. 11 (4) (2022) 1695–1713.

[30] Djordje Slijepcevic, Fabian Horst, Sebastian Lapuschkin, Brian Horsak, Anna-Maria Raberger, Andreas Kranzl, Wojciech Samek, Christian Breiteneder, Wolfgang Immanuel Schöllhorn, Matthias Zeppelzauer, Explaining machine learning models for clinical gait analysis, ACM Trans. Comput. Healthc. 3 (2) (2021) 14, http://dx.doi.org/10.1145/3474121, 2021, 27 pages.

[31] Jeremy Petch, Shuang Di, Walter Nelson, Opening the black box: the promise and limitations of explainable machine learning in cardiology, Can. J. Cardiol. (2021).

[32] Ahmad Kamal Mohd Nor, et al., Overview of explainable artificial intelligence for prognostic and health management of industrial assets based on preferred reporting items for systematic reviews and meta-analyses, Sensors 21 (23) (2021) 8020.

[33] M. Obayya, N. Nemri, M.K. Nour, M. Al Duhayyim, H. Mohsen, M. Rizwanullah, A. Sarwar Zamani, A. Motwakel, Explainable artificial intelligence enabled TeleOphthalmology for diabetic retinopathy grading and classification, Appl. Sci. 12 (2022) 8749, http://dx.doi.org/10.3390/app12178749.

[34] N.I. Papandrianos, A. Feleki, S. Moustakidis, E.I. Papageorgiou, I.D. Apostolopoulos, D.J. Apostolopoulos, An explainable classification method of SPECT myocardial perfusion images in nuclear cardiology using deep learning and grad-CAM, Appl. Sci. 12 (2022) 7592, http://dx.doi.org/10.3390/app12157592.

[35] D. Pertzborn, C. Arolt, G. Ernst, O.J. Lechtenfeld, J. Kaesler, D. Pelzel, O. Guntinas-Lichius, F. von Eggeling, F. Hoffmann, Multi-class cancer subtyping in salivary gland carcinomas with MALDI imaging and deep learning, Cancers 14 (2022) 4342, http://dx.doi.org/10.3390/cancers14174342.

[36] R.A. Zeineldin, M.E. Karar, Z. Elshaer, et al., Explainability of deep neural networks for MRI analysis of brain tumors, Int. J. CARS 17 (2022) 1673–1683, http://dx.doi.org/10.1007/s11548-022-02619-x.

[37] Atul Anand, Tushar Kadian, Manu Kumar Shetty, Anubha Gupta, Explainable AI decision model for ECG data of cardiac disorders, Biomed. Signal Process. Control 75 (2022) http://dx.doi.org/10.1016/j.bspc.2022.103584.

[38] Giorgio Leonardi, Stefania Montani, Manuel Striani, Explainable process trace classification: An application to stroke, J. Biomed. Inform. 126 (2022) http://dx.doi.org/10.1016/j.jbi.2021.103981.

[39] Z.U. Ahmed, K. Sun, M. Shelly, et al., Explainable artificial intelligence (XAI) for exploring spatial variability of lung and bronchus cancer (LBC) mortality rates in the contiguous USA, Sci. Rep. 11 (2021) 24090, http://dx.doi.org/10.1038/s41598-021-03198-8.

[40] M. Merry, P. Riddle, J. Warren, A mental models approach for defining explainable artificial intelligence, BMC Med. Inform. Decis. Mak. 21 (2021) 344, http://dx.doi.org/10.1186/s12911-021-01703-7.

[41] H.S.A. Fang, N.C. Tan, W.Y. Tan, et al., Patient similarity analytics for explainable clinical risk prediction, BMC Med. Inform. Decis. Mak. 21 (2021) 207, http://dx.doi.org/10.1186/s12911-021-01566-y.

[42] J. Andreu-Perez, L.L. Emberson, M. Kiani, et al., Explainable artificial intelligence-based analysis for interpreting infant fNIRS data in developmental cognitive neuroscience, Commun. Biol. 4 (2021) 1077, http://dx.doi.org/10.1038/s42003-021-02534-y.

[43] Dongha Kim, Jongsoo Lee, Predictive evaluation of spectrogram-based vehicle sound quality via data augmentation and explainable artificial Intelligence: Image color adjustment with brightness and contrast, Mech. Syst. Signal Process. 179 (2022) http://dx.doi.org/10.1016/j.ymssp.2022.109363.

[44] Ioannis Kakogeorgiou, Konstantinos Karantzalos, Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing, Int. J. Appl. Earth Obs. Geoinf. 103 (2021) http://dx.doi.org/10.1016/j.jag.2021.102520.

[45] S. Pereira, R. Meier, V. Alves, M. Reyes, C.A. Silva, Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, Springer, Cham, 2018, pp. 106–114, http://dx.doi.org/10.1007/978-3-030-02628-8.

[46] P. van Molle, M. de Strooper, T. Verbelen, B. Vankeirsbilck, P. Simoens, et al., Visualizing convolutional neural networks to improve decision support for skin lesion classification, in: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, Springer, Cham, 2018, pp. 115–123.

[47] F. Eitel, K. Ritter, Alzheimer's Disease Neuroimaging Initiative, Testing the robustness of attribution methods for convolutional neural networks in mri-based alzheimer's disease classification, in: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, Springer, Cham, 2019, pp. 3–11.

[48] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, et al., Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy, Ophthalmology 126 (4) (2019) 552–564, http://dx.doi.org/10.1007/978-3-030-02628-8_12.

[49] H.L. Yang, J.J. Kim, J.H. Kim, Y.K. Kang, D.H. Park, et al., Weakly supervised lesion localization for age-related macular degeneration detection using optical coherence tomography images, PLoS One 14 (4) (2019) e0215076, http://dx.doi.org/10.1371/journal.pone.0215076.

[50] K. Young, G. Booth, B. Simpson, R. Dutton, S. Shrapnel, Deep neural network or dermatologist? in interpretability of machine intelligence, in: Medical Image Computing and Multimodal Learning for Clinical Decision Support, Springer, Cham, Shenzhen, China, 2019, pp. 48–55, http://dx.doi.org/10.1007/978-3-030-33850-3_6.

[51] H. Leopold, A. Singh, S. Sengupta, J. Zelek, V. Lakshminarayanan, Recent advances in deep learning applications for retinal diagnosis using OCT, in: Tate of the Art in Neural Networks, Elsevier, NY, 2020.

[52] A. Singh, S. Sengupta, V. Lakshminarayanan, Interpretation of deep learning using attributions: Application to ophthalmic diagnosis, Appl. Mach. Learn. 2020 (2020) 115110A, http://dx.doi.org/10.1117/12.2568631.

[53] Z. Papanastasopoulos, R.K. Samala, H.P. Chan, L. Hadjiiski, C. Paramagul, et al., Explainable AI for medical imaging: Deep-learning CNN ensemble for classification of estrogen receptor statusfrom breast MRI, in: SPIE Medical Imaging 2020: Computer-Aided Diagnosis, Vol. 11314, Houston, Texas, USA, 2020, 113140Z, http://dx.doi.org/10.1117/12.2549298.

[54] K. Wickstrøm, M. Kampffmeyer, R. Jenssen, Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps, Med. Image Anal. 60 (2020) 101619, http://dx.doi.org/10.1016/j.media.2019.101619.

[55] J. Sun, F. Darbehani, M. Zaidi, B. Wang, SAUNet: Shape attentive u-net for interpretable medical image segmentation, in: International Conference on Medical Image Computing and ComputerAssisted Intervention, Springer, Cham, 2020, pp. 797–806.

[56] T. Eslami, J.S. Raiker, F. Saeed, Explainable and scalable machine learning algorithms for detection of autism spectrum disorder using fmri data, in: Neural Engineering Techniques for Autism Spectrum Disorder, Academic Press, Nevada, USA, 2021, pp. 39–54.

[57] S. Hou, J. Han, COVID-19 detection via a 6-layer deep convolutional neural network, CMES Comput. Model. Eng. Sci. 130 (2) (2022) 855–869, http://dx.doi.org/10.32604/cmes.2022.016621.

[58] H. Mehta, K. Passi, Social media hate speech detection using explainable artificial intelligence (XAI), Algorithms 15 (2022) 291, http://dx.doi.org/10.3390/a15080291.

[59] S.-Y. Lim, D.-K. Chae, S.-C. Lee, Detecting deepfake voice using explainable deep learning techniques, Appl. Sci. 12 (2022) 3926, http://dx.doi.org/10.3390/app12083926.

[60] M. Szczepański, M. Pawlicki, R. Kozik, et al., New explainability method for BERT-based model in fake news detection, Sci. Rep. 11 (2021) 23705, http://dx.doi.org/10.1038/s41598-021-03100-6.

[61] A. Adak, B. Pradhan, N. Shukla, A. Alamri, Unboxing deep learning model of food delivery service reviews using explainable artificial intelligence (XAI) technique, Foods 2022 (2019) 11, http://dx.doi.org/10.3390/foods%2011142019.

[62] H.-S. Kim, I. Joe, An XAI method for convolutional neural networks in self-driving cars, PLoS ONE 17 (8) (2022) e0267282, http://dx.doi.org/10.1371/journal.pone.0267282.

[63] Nilkanth Mukund Deshpande, et al., Explainable artificial intelligence–a new step towards the trust in medical diagnosis with AI frameworks: A review, Comput. Model. Eng. Sci. 133 (2022) 1–30.

[64] H. Bharadhwaj, S. Joshi, Explanations for temporal recommendations, Künstl Intell. 32 (2018) 267–272, http://dx.doi.org/10.1007/s13218-018-0560-x.

[65] Giulia Vilone, Luca Longo, Explainable artificial intelligence: a systematic review, 2020, arXiv preprint arXiv:2006.00093.

[66] Timo Speith, A review of taxonomies of explainable artificial intelligence (XAI) methods, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022.

[67] M. Nazar, M.M. Alam, E. Yafi, M.S. Mazliham, A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques, IEEE Access (2021).

[68] Romila Pradhan, et al., Explainable AI: Foundations, applications, opportunities for data management research, in: 2022 IEEE 38th International Conference on Data Engineering, ICDE, IEEE, 2022.

[69] G. Vilone, L. Longo, Classification of explainable artificial intelligence methods through their output formats, Mach. Learn. Knowl. Extr. 3 (2021) 615–661, http://dx.doi.org/10.3390/make3030032.

[70] P. Angelov Plamen, et al., Explainable artificial intelligence: an analytical review, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 11 (5) (2021) e1424.

[71] M.R. Islam, M.U. Ahmed, S. Barua, S. Begum, A systematic review of explainable artificial intelligence in terms of different application domains and tasks, Appl. Sci. 12 (2022) 1353, http://dx.doi.org/10.3390/app12031353.

[72] M. Han, J. Kim, Joint banknote recognition and counterfeit detection using explainable artificial intelligence, Sensors 19 (2019) 3607, http://dx.doi.org/10.3390/s19163607.

[73] Hui Wen Loh, Chui Ping Ooi, Silvia Seoni, Prabal Datta Barua, Filippo Molinari, U. Rajendra Acharya, Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022), Comput. Methods Programs Biomed. 226 (2022) http://dx.doi.org/10.1016/j.cmpb.2022.107161.

[74] Y. Zhang, Y. Weng, J. Lund, Applications of explainable artificial intelligence in diagnosis and surgery, Diagnostics 12 (2022) 237, http://dx.doi.org/10.3390/diagnostics12020237.

[75] B.M. Keneni, et al., Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles, IEEE Access 7 (2019) 17001–17016, http://dx.doi.org/10.1109/ACCESS.2019.2893141.

[76] Khalid Bahani, Mohammed Moujabbir, Mohammed Ramdani, An accurate fuzzy rule-based classification systems for heart disease diagnosis, Sci. Afr. 14 (2021) 01019.

[77] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable AI: A review of machine learning interpretability methods, Entropy 23 (2021) 18, http://dx.doi.org/10.3390/e23010018.

[78] M.P. Ayyar, J. Benois-Pineau, A. Zemmari, Review of white box methods for explanations of convolutional neural networks in image classification tasks, J. Electron. Imaging 30 (5) (2021) 050901.

[79] Nadia Burkart, Marco F. Huber, A survey on the explainability of supervised machine learning, J. Artificial Intelligence Res. 70 (2021) 245–317.

[80] Muhammad Rehman Zafar, Naimul Mefraz Khan, DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems, 2019, arXiv preprint arXiv:1906.10263.

[81] Maria Sahakyan, Zeyar Aung, Talal Rahwan, Explainable artificial intelligence for tabular data: A survey, IEEE Access 9 (2021) 135392-135422.

[82] I. Palatnik de Sousa, M. Maria Bernardes Rebuzzi Vellasco, E. Costa da Silva, Local interpretable model-agnostic explanations for classification of lymph node metastases, Sensors 19 (2019) 2969, http://dx.doi.org/10.3390/s19132969.

[83] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, Dino Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (5) (2018) 93, (2019) 42 pages. http://dx.doi.org/10.1145/3236009.

[84] Van den Broeck, Guy, et al., On the tractability of SHAP explanations, J. Artificial Intelligence Res. 74 (2022) 851–886.

[85] Thomas Ponn, Thomas Kröger, Frank Diermeyer, Identification and explanation of challenging conditions for camera-based object detection of automated vehicles, Sensors 20 (13) (2020) 3699.

[86] Angela Lombardi, et al., Explainable deep learning for personalized age prediction with brain morphology, Front. Neurosci. (2021) 578.

[87] Lev Utkin, Andrei Konstantinov, Ensembles of random SHAPs, Algorithms 15 (11) (2022) 431.

[88] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160, http://dx.doi.org/10.1109/ACCESS.2018.2870052.

[89] R. Sheh, I. Monteath, Defining explainable AI for requirements analysis, Künstl Intell. 32 (2018) 261–266, http://dx.doi.org/10.1007/s13218-018-0559-3.

[90] Phillips P. Jonathon, et al., Four Principles of Explainable Artificial Intelligence, Gaithersburg, Maryland, 2020.

[91] DW, Gunning D. Aha, DARPA's explainable artificial intelligence program, AI Mag. 40 (2) (2019) 44.