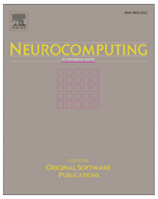
Contents lists available at ScienceDirect

Neurocomputing

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



Survey paper

Explaining deep neural networks: A survey on the global interpretation methods

Rabia Saleem [a,*], Bo Yuan [b,*], Fatih Kurugollu [a,c], Ashiq Anjum [b], Lu Liu [b]

[a] *School of Computing and Engineering, University of Derby, Kedleston Rd, Derby DE22 1GB, UK*
[b] *School of Computing and Mathematical Sciences, University of Leicester, University Rd, Leicester LE1 7RH, UK*
[c] *Department of Computer Science, University of Sharjah, Sharjah, United Arab Emirates*

ARTICLE INFO

Article history:
Received 18 May 2022
Revised 30 July 2022
Accepted 15 September 2022
Available online 23 September 2022
Communicated by Zidong Wang

Keywords:
Artificial intelligence
Deep neural networks
Black box Models
Explainable artificial intelligence
Global interpretation

ABSTRACT

A substantial amount of research has been carried out in Explainable Artificial Intelligence (XAI) models, especially in those which explain the deep architectures of neural networks. A number of XAI approaches have been proposed to achieve trust in Artificial Intelligence (AI) models as well as provide explainability of specific decisions made within these models. Among these approaches, global interpretation methods have emerged as the prominent methods of explainability because they have the strength to explain every feature and the structure of the model. This survey attempts to provide a comprehensive review of global interpretation methods that completely explain the behaviour of the AI models. We present a taxonomy of the available global interpretations models and systematically highlight the critical features and algorithms that differentiate them from local as well as hybrid models of explainability. Through examples and case studies from the literature, we evaluate the strengths and weaknesses of the global interpretation models and assess challenges when these methods are put into practice. We conclude the paper by providing the future directions of research in how the existing challenges in global interpretation methods could be addressed and what values and opportunities could be realized by the resolution of these challenges.

1. Introduction

Machine Learning (ML) has been central to AI research, as it has the ability to find patterns and categorise things. Deep Learning (DL) is the subset of ML that is mainly involved in the construction of the deep architectures known as deep neural networks (DNNs). From the past few years, the DNNs architectures have been frequently used in many computer-vision tasks such as action recognition [1], motion tracking [2], and object detection [3]. These tasks are performed by using various deep architectures such as convolutional neural networks (CNN) [4], deep Boltzmann machines [5], and deep belief networks [6]. DNNs have been extensively used in numerous critical applications such as audio processing [7], autonomous vehicles and robots [8], autism spectrum disorder [9–11], signal analysis [12,13] ophthalmology [14–16], cyber-security [17] and healthcare [18–20]. The DNNs use high performance computational resources to train multiple hidden layers and millions or billions of parameters that vigorously perform many crucial tasks with the best accuracy. However, the computation process of these DNNs models is opaque to human beings, so generally, these DNNs models are referred to as black box models [21]. We cannot explain the decision making process of these deep neural architectures leading to serious questions on the trust and transparency of these models.

The lack of transparency within deep neural architectures restricts the deployment of such models especially in healthcare and safety critical applications where a small possibility of the wrong decision could damage human life [22]. Therefore, an understandable explanation of the set of instructions behind every decision made by DL models is highly in demand. Many research papers have been published in the past few years that discussed the explainability issue of AI models. The explainability of the black box models has received so much importance in recent years that eXplainable Artificial Intelligence (XAI) has emerged as a specific domain within AI [23].

Abbreviations: AI, Artificial Intellegence; DNNs, Deep Neural Networks; ML, Machine Learning; XAI, eXplainable Artificial Intellegence.

[*] Corresponding author.

E-mail addresses: r.saleem3@unimail.derby.ac.uk (R. Saleem), b.yuan@leicester.ac.uk (B. Yuan).

https://doi.org/10.1016/j.neucom.2022.09.129
0925-2312/© 2022 The Author(s). Published by Elsevier B.V.

The purpose of XAI is to develop a simple, clear but logically explainable model that describes the inner functionalities of the black box models. The explanations produced by XAI should be understandable by human beings while maintaining high performance in terms of prediction accuracy. The generalised additive models (GAM) were initially introduced to explain the black box nature of the ML models by using a smooth function [24]. Later, a visual (tree-like graph) algorithm, decision-tree, was developed to provide a conditional and individual explanation of decisions [25,26]. Individual Conditional Expectation (ICE) explained the change in predictions with respect to the features [27]. A limited version of ICE, the Partial Dependence Plot (PDP) [28], has been initially put into practice to globally explain the nature and effect of only one or two features of the model prediction. The classification models such as k-NN and SVM were explained by a set of explanation vectors but these techniques could only explain the outcome of one instance [29].

The above discussion indicates that initially modest-size ML models were explained by using different explainability methods. However, a significant number of models use DNNs that have attained much importance in high-risk applications. The DNN models with high accuracy demand a better explanation that would lead us to produce more responsible and trustworthy AI systems [30]. Many local and visualisation techniques have been recently developed that help the AI experts to understand the decision procedure of the DNN model [31]. For example, De Graaf and Van Mulken [32] proposed a solution that imitates the intermediate process, connects the decision with the given piece of information, and is understandable by end-users. This approach however only provides a local explanation of the learning model.

This survey paper aims to provide a detailed state of the explainability methods available to AI researchers and practitioners as well as highlights the strengths and weaknesses of the global XAI methods that have been developed in the past ten years. Our contributions can be summarised as follows:

1. A comprehensive overview of the existing approaches used to globally explain the black box models of DNNs. have been described, particularly a critique of the visual and local XAI methods used in the global explanation of the black box models of DNNs has been provided. The latest and highly cited research papers are picked that have been published in renowned journals and conferences over the past ten years and highlight recent developments in explaining the global XAI approaches.
2. In order to systematically analyse the global XAI methods, a taxonomy has been introduced on the basis of ante-hoc and post hoc approaches, which provides clear recommendations on when to use a particular approach.
3. This paper highlights gaps in the global XAI methods and offers a way forward for the future direction of work by proposing a deterministic XAI model that will help to explain and address the existing gaps.

The structure of this paper is summarised as follows. Section 2 categorises terminologies that have been frequently and interchangeably used in the field of XAI. The knowledge of the XAI terminologies enables us to understand the three dimensions of XAI methods that have been critically evaluated in Section 3. Section 4 presents a comprehensive review and taxonomy of the available methods for the global interpretation of DNNs. Section 5 provides a summary of available Global XAI methods. Lastly, Section 6 concludes the papers with current research gaps and future directions.

2. Terminologies

One of the major issues while discussing XAI is the use of terminologies that are used interchangeably and complicate the understanding of concepts. Before diving into the deep ocean of the XAI field, we attempt to clarify the differences between the commonly used XAI concepts and terminologies and try to present a consistent version of (Fig. 1) their similarities and differences.

- *White box:* If there is complete information about the architecture and parameters of a model, it is known as the "white-box" model. This type of model is considered immensely helpful for endorsing trust, however, most of the time the amount of information is not adequate enough to explain the logical rationale behind the decisions.
- *Transparent box:* If a model can explain its design, parameters, or algorithm on its own and this justification is good enough for the end-user, then the model is named as the "transparent-box" model. One can check, evaluate, and improve the predictions by this transparency.
- *Black box:* A model with hidden and veiled architecture and parameters leads to an unknown process of decision-making, resulting into a "black-box" model. Generally, the DL models are black box models because their deep architecture makes them opaque [21]. Next, we define the three key terms that have been used as substitutes for one another in the XAI field. Hence, it becomes more difficult to differentiate these ideas.
- *Understandability:* It refers to the understanding of the model's characteristics, features, and function without knowing the internal process and procedure involved in the decision making [33]. This term provides an answer to the question "How the AI model works?."
- *Explainability:* It involves explaining the internal process and answering the question "How does the black box model of AI make certain decisions?."
- *Interpretability:* It entails the understanding of the internal functionality and characteristics of the model. This terminology provides a meaningful, clear, and logical reason(s) for the specific decision in a manner that is understandable to the targeted client. This term answers the question "What is the decision of the AI model and Why?." Research communities often can classify the explanation of AI models into the following three categories based on usage and scope.
- The *Complete Explanation* includes all potential features and facts while explaining the decision of AI models, while in *Compact Explanation* the decisions of the AI models can be explained by a limited number of factors. Mostly available XAI methods for the black box of DNNs models provide the compact explanation of DNN models for some particular instances [27].
- Some explanation approaches only explain the limited number of neural networks, for example, the explanation of the linear model by the regression weights. The limited ability to explain a certain type of black box model is known as *Model Specific*. The other approach is *Model Agnostic* which can explain each category of ML models. The model agnostic explanation tools are preferable because of their flexibility, however, they have no access to the inner information of the model [34].
- The *Ante-hoc* explanation scheme can capture all information from the input layer to the hidden and output layers of the given model [35] however, the *Post-hoc* scheme only highlights the route of a particular outcome. For example, the decision-tree method can explain the whole model but the LIME method can only explain the process behind the particular outcome of a ML model [36]. Hence, similar to the model-agnostic approach, the post hoc methods are more flexible and easier to apply on

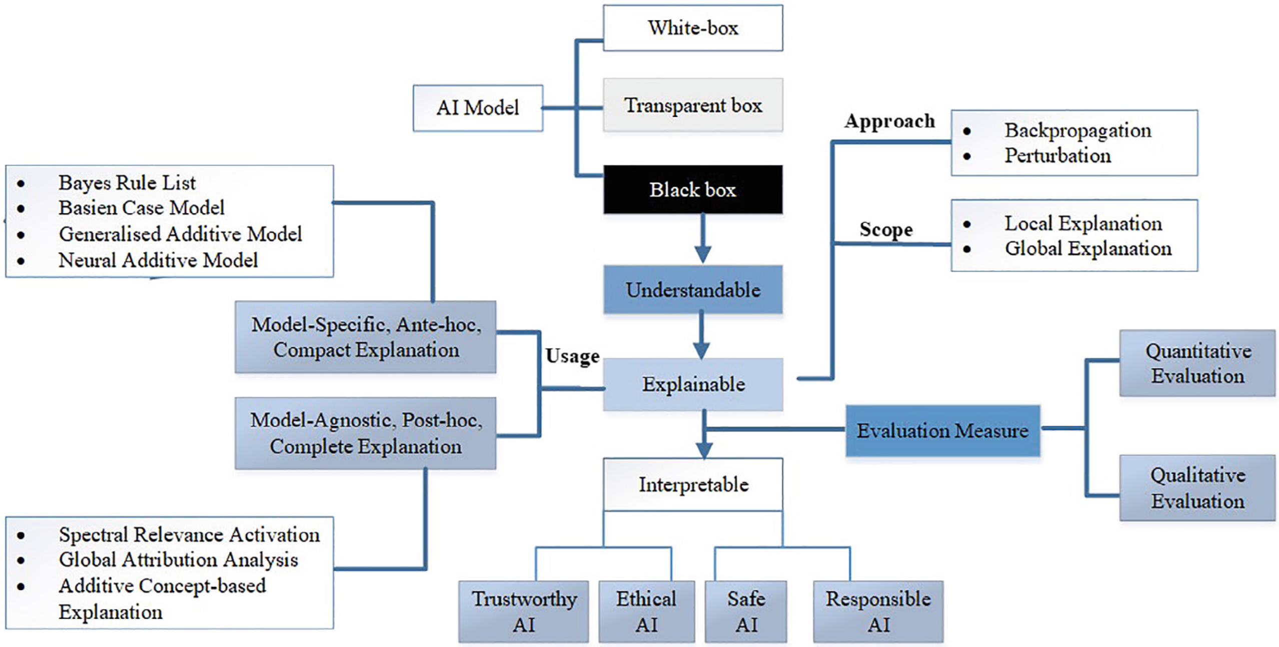


Fig. 1. Classification of XAI methods and their terminology map.

different models than the ante-hoc schemes. Many other terms have been frequently used interchangeably instead of XAI. Some prominently famous terminologies are mentioned and explained below.

- *Ethical AI and Responsible AI:* Both terms are very much associated with the XAI field. An AI model is said to be "ethical" if it does not break any defined rule or regulation by its user during the whole decision-making process. The implementation of "responsible AI" in real-world applications demands privacy, fairness, and ethics together with the explanation of the AI model [37].
- *Trustworthy AI and Safe AI:* An AI model is said to be "Trustworthy" if a user can anticipate the performance of the model positively. Trustworthiness is one of the primary goals of the XAI that would lead to another goal called Safe AI. The "Safe AI" has the potential to control the chances of unexpected decisions to minimise the risk of unintended harm during the interaction of systems and humans [38].

The answer to the question "How can you accurately explains the black box of AI model?" should be given by introducing an evaluation criterion for explainable methods. The following two are the main evaluation criteria for the validation of explainable methods.

- *Qualitative Evaluation:* In this evaluation, the measure of XAI methods depends on the satisfaction of end-user inquisitiveness, safety, understanding, and usability [39]. This evaluation has importance as it provides the feedback of explainable methods from the point of view of human understanding and usability.
- *Quantitative Evaluation:* This is used if the evaluation measure emphasises the performance of the explainer that how closely the explainer mimics the black box model [40]. This approach can validate the explainable methods by quantitatively evaluating the outcomes of the AI model for all or some specific tasks.

3. Dimensions of XAI Problem

As discussed earlier that for critical tasks, just predictions from AI models are not enough. These models should be able to explain the whole functionality of the black box model that would eventually help to explain every reason behind each prediction (Global interpretation) or at least explain the reason behind a single prediction (Local interpretation). An explanation process can be divided into two parts (i) Extract information (ii) Exhibit information. The exhibition of information has extraordinary importance as it could directly connect the system with the novices as a client [33]. Besides the above two dimensions of explaining the black box model, another approach is also in demand which visually inspects the model and explains the reason behind the prediction.

Various high-performed approaches and methods for explainable AI have been developed in the past ten years. Mostly complex and deep architectures of neural networks would have a post hoc explanation that interprets certain predictions. However, there are few ante-hoc methods with the limited capacity of explaining the simple and small-sized AI models. The following section logically differentiates these approaches and discusses some prominent methods developed in the past decade.

3.1. Outcome Explanation or Local Interpretation

The outcome explanation, also referred to an instant-wise explanation, aims to explain the reasons behind a single prediction using a specific set of input–output (Fig. 2). Although this explanation is not considered suitable for the non-experts, most methods of outcome explanation would help the AI experts to scrutinize many edge cases of ML models. The understanding of decision routes for the edge cases such as prediction of an autonomous vehicle to react to an unusual behaviour of cyclist, pedestrian or any object on road, are becoming more important to validate the response of vehicle in every situation [36].

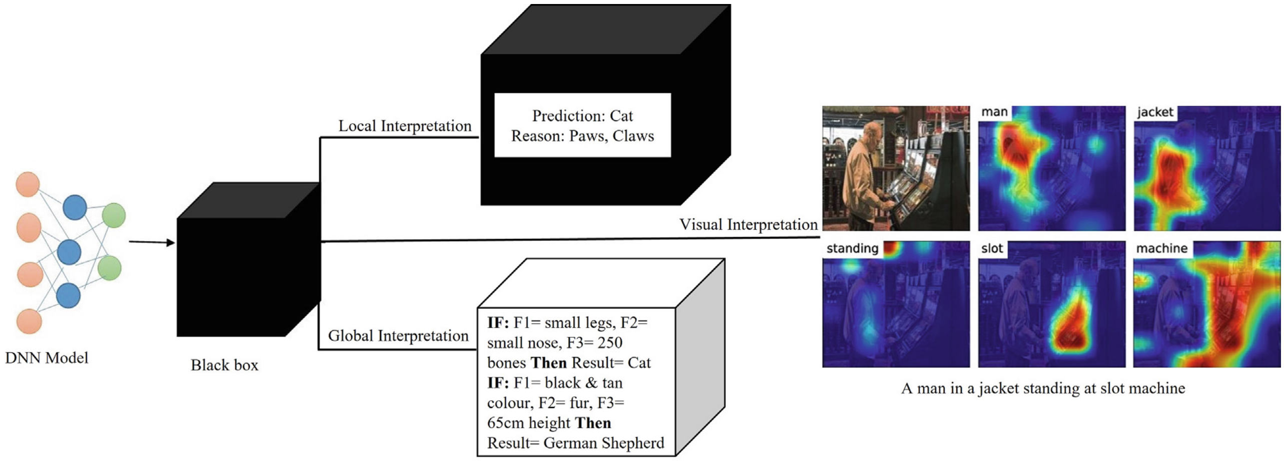

DNN Model
Black box
Local Interpretation
Prediction: Cat
Reason: Paws, Claws
Visual Interpretation
Global Interpretation
IF: F1= small legs, F2= small nose, F3= 250 bones Then Result= Cat
IF: F1= black & tan colour, F2= fur, F3= 65cm height Then Result= German Shepherd
man
jacket
standing
slot
machine
A man in a jacket standing at slot machine

Fig. 2. Three dimensions (Global, Local and Visual) of XAI methods. Visual interpretation source paper [58].

Precisely, an explainable function, say ”OE” is generated that explains the given mapping $NN : D(X_A) \rightarrow O(Y_A)$ for a particular instance A in the real-valued matrices (or vector). Note that in the above mathematical mapping, NN is the neural network applying to dataset D to produce O output. Initially, the outcome explanation methods established a framework to understand the significance of features and their relations by using feature importance matrices, heatmaps, Bayesian and rule-based techniques. This understanding has been further improved by designing game theory and graph-based models and attribution maps [41]. *Activation maximisation (AM)* [42] is an outcome explainer method that explains the convolutional neural network (CNN) by highlighting layer-wise feature importance. This method was introduced by Erhan et al. in 2010 suggesting an optimisation problem to maximise the unit activation by considering the input patterns. For fixed parameters, the activation mapping of i^{th} unit from j^{th} layer, $Z_{i,j}(X, \phi)$ can be defined as following:

$$X^* = argmax_{(X,X=\rho)} Z_{i,j}(X, \phi) \tag{1}$$

In 2013, Simonyan [43] presented *saliency maps*, which is another way of explaining the output by summarising the pixel importance of the input image. This approach uses a perturbation method and computes the gradient of input images to create resultant feature maps. By modifying the input data, perturbation-based methods discover the changes in output of the DNN model, where gradient-based methods can detect very tiny changes in the input data. Pixels in an image, words in text, or columns in a table can be perturbed by applying an occlusion mask, blurring or replacing parts of images, replacing a word with a synonym in text, or shuffling or inserting rows/columns in tables. It is important to choose these changes in input data very carefully to get better results. These maps can be sharpened by SmoothGrad that randomly perturbs the input and presents the result as the average of the resultant maps. *Layer-Wise Relevance BackPropagation (LRP)* [44] is an outcome explanation method for the DNNs introduced in 2015. This method decomposes the output of the DNN model and finds out the relevance scores of each feature in the given input data. The LRP method is widely used to explain outcome predictions of many DNN models such as Convolution Neural Network (CNN) and Recurrent Neural Network (RNN) from the past few years. However, many new researchers use this technique to prune the network system by understanding the attribution of each layer [45]. A method, called Class Activation Mapping (CAM), creates heat maps that highlight those parts of input which are responsible for an outcome [46]. Based on this theory, *Gradient- weight Class Activation Mapping (Grad-CAM)* was developed that highlights the key features of the last convolutional layer of CNN by using gradient information [47].

Local Interpretable Model Explanation (LIME) was presented in the year 2016 to explain the outcomes of DNN models that would be interpretable in common man language [40]. The ultimate idea is to fit a surrogate model such as Linear regression or Ridge regression [48] on the perturb input dataset and generate the local explanation. A binary vector $x^*0, 1$ used to assure the absence and presence of superpixels that would be captured with the help of LIME from the input image. Different versions of the LIME method have been developed to explain the various kind of models. For example, Sound-LIME (SLIME) is the extended version of the LIME method used to explain the deep voice detector’s predictions by time–frequency and temporal segmentation [49], Modified Perturbed Sampling operation (MPS-LIME) uses clique operation for picking superpixels that reduce the run time because of the low number of perturbed samples [50] and KL-LIME uses Kullback Leibler divergence to explain the prediction of Bayesian model [51]. A similar method known as *Shapley Additive explanation (SHAP)* was proposed by Lundberg and Lee to enlighten the importance of the individual part of input data while explaining the prediction [52]. As described in Eq. 2, SHAP explains the prediction by suggesting an additive model with M simplified and alliance features, where $z' \epsilon \{0,1\}^M$ is alliance vector, and $\phi_k \epsilon\ R$ is the k^{th} feature attribution.

$$g(z)' = \phi_0 + \sum_{k=1}^{M} \phi_k(z'_k) \tag{2}$$

The contribution of each feature with the sum of bias is represented as $g(z)')$ in the above equation. Similar to the LIME method, SHAP also has different versions such as Low-Order SHAP, Kernel-SHAP, and Deep-SHAP [53]. Due to the better performance of SHAP as compared to the other methods, there is an extensive use of this approach in the medical field [54].

3.2. Visual Explanation or Model Inspection

The inspection approach aims to provide the textual or visual explanation of many ML models that would help to understand the reason for the model’s prediction (Fig. 2). A survey paper on the DNN visualisation techniques published in 2017 mentioned that most researchers use pixels to display their research outcomes [55].

This paper classifies the inspection methods in four categories based on feature importance (i) Visual explanation with Activation Maximisation (AM) (ii) Visual explanation with Partial Dependence Plot (PDP) (iii) Visual explanation with Decision Tree (iv) Visual explanation with Sensitive Analysis.

As discussed earlier that the area or pixels responsible for the decision can be highlighted by using the *Activation Maximisation (AM)* technique. Yosinki et al. proposed tools for visual explanation and interpretation of the DNNs such as CNN. One of them explains the computation process at the intermediate layers of CNN and the other tool highlights the active part or feature of each CNN layer during the training process [56]. Another similar method discussed in [57], traces back the computation process to detect those parts of the image responsible for neuron activation. The Deconvolutional approach is used to visualise the intermediate layers of the simple CNN model [58]. Many methods and their variants were proposed in different papers for a visual explanation of DNN [59–61]. *Partial Dependence Plot (PDP)* is a manner to visualise the relationship between different neurons in the feature space. In [27,62] extension of PDP was applied on the tabular dataset which was able to evaluate and visualise the interaction between the neurons and the relationship between the feature and prediction.

A relevant method known as Tree View uses the *Decision Tree* to visually explain the connection between the decomposed K- cluster of model features [63]. The *Quantitative Input Influence (QII)* is another method that captures the information about the input features and used decision tree approach to explain the prediction [64]. The uncertainty of the input–output pair can be studied under the measure of sensitivity. Initially, the *Sensitivity Analysis (SA)* was used to understand the mechanism of a neural network on the basis of sensitivity and the Neural Interpretation Diagram (NID) was used to confiscate the non-essential parts and connection of the neural network [65]. Based on sensitivity analysis, the *Gaussian Process Classification (GDP)* explain and visualise the local outcome by the explanation vector ω [29] and the *Variable Effect Characteristic curve (VEC)* draws the bar plots between the features and their response based on the importance of features [66].

The above discussion indicates that some explanation methods prefer to use different visual tools such as heatmaps and salience masks to display the information. During this type of explanation, two dimensions have been used for the interpretation of the black box model, namely local interpretation, and global interpretation. As the main focus of this survey, is the global interpretation methods so next section has a detailed discussion on this state of the art.

3.3. Model Explanation or Global Interpretation

One way to give details about the black box of the AI models is the global interpretation. This type of explanation describes the complete logic of all the outcomes by mimicking the behaviour of the black box model (Fig. 2). The inside view of a model would help to understand the nature of model features and their correlations that leads to the outcomes. Many model explanation methods divide the whole model into parts to make it easier for the explanation. Tree-based and Rule-based models are inherently fully explainable. The black box models consisting of the neural network were initially explained by approximating a single tree [67,68] and rule extraction [69]. In [70], an approach REFNE has been introduced to interpret generated instances from the trained neural network by extracting symbolic rules.

In recent years, the approaches that explain any type of AI model (agnostic) are in high demand. The *Generalised Additive Model (GAM)* is the first attempt toward the model-agnostic approach that explains regression splines, tree-ensembles, and single-trees by highlighting the contribution of each feature [71]. Same researchers refined this method in [72] and a case study on health care was published in 2015 as the application of refining GAM [73]. GoldenEye is an iterative method that explains the model by grouping those features whose interaction produces results [74].

The *Feature Importance Ranking Measure (FIRM)* is another method that finds the most relevant features by correlating the structure of features [75]. FIRM is the extension of the method called Positional Oligomiter Importance Matrices (POIMS) that uses a scoring system to rank and visualise each K-mer [76]. The extended version of FIRM, named, Measure of Feature Importance (MFI) is the non-linear classifier that detects those unobtrusive features whose interaction with other features impacts the outcomes [77].

Before discussing the global interpretation method, the Table 1 provides a quick view on the interpretation tools that are used to explain the AI models at the global level. Based on the model usage and type, we also interpret these tools to create white-box or to enhance the fairness in AI models. The next section synthesises the global model explanation methods, proposed in the past ten years.

4. Methods for Model Explanation

The AI explanation methods can be categorised (Fig. 1) on Scope: Does the XAI method explain the whole model or only a certain outcome (local or global)? Approach: What is the focus of the algorithm, input data, or model parameters (backpropagation or perturbation)? Usage: The developed method can explain any type of model architecture or just be applicable to the specific architecture (post hoc or ante-hoc)? This paper focuses only on global interpretation methods based on model usage (Fig. 3).

4.1. Ante-hoc methods

Ante-hoc methods are mostly model-specific and this might be seen as a drawback because they consider only a limited number of models. In many papers, this explanation is also known as intrinsic explanation. Hence by definition, model intrinsic explanation methods depend on a certain design and cannot be used again for any other architectures. This section provides a detailed discussion on those ante-hoc methods which have been established in the past ten years. A quick view of ante-hoc methods are given in Table 2 and Table 4 that summarise information about type of data, frameworks and methodologies, as well as their merits and demerits.

Definition 1. Let the DNN represent as a function $F : R^{D_1} \rightarrow R^{D_k}$ with an input $x\ \epsilon R^{D_1}$. Another representation of x could be ante-hoc explanation, $eXp(F,x)\ \epsilon R^{T}$ such that $T \leqslant D_1$.

4.1.1. Bayesian Case Model (BCM)

Studies show that just revealing of rules behind the decisions of AI models are not enough for achieving user's confidence, the example-based reasoning and interpretation improve the level of confidence significantly. The case-based reasoning (CBR) is considered exemplar-based modelling that involves the most effective tactics such as matching and prototyping as humans like to look at examples rather than recommendations. For example, the naturalistic studies mentioned that the decision-makers in a fire service use recognition-primed decision making that matches new situations with similar cases and decides appropriate manoeuvres to handle the situation [87]. Therefore, with the knowledge of CBR new situations can be represented successfully by analysing the previous situations.

Table 1
Global Interpretation Tools to create White-box (W) or to enhance Fairness (F) in AI models.

Model type (Scope)	Year	Data type	Category	Interpretation tools
Ante-hoc (Model Specific)	2010	Tabular	F	Fairness-Comparison[78]
	2015	Tabular	W	Interpret ML [73]
	2016	Tabular	W	Slim [79]
	2019	Tabular	W	AI-360 [80]
	2019	Tabular	F	ML-Fairness Gym [81]
Post-hoc (Model Agnostic)	2017	Tabular	F	AIF-360 [82] and Fair Classification [83]
	2018	Tabular	F	Fair Learning [84]
	2018	Tabular	F	AI-360 Gerry Fair [85]

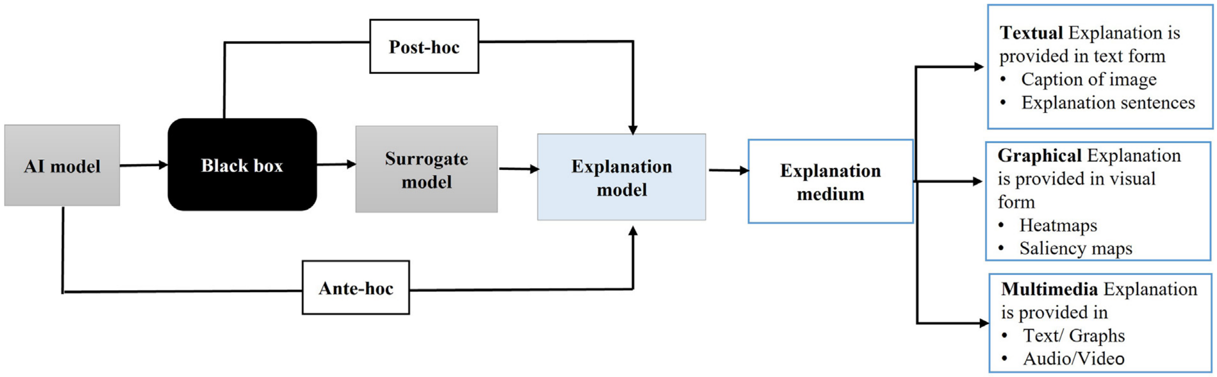
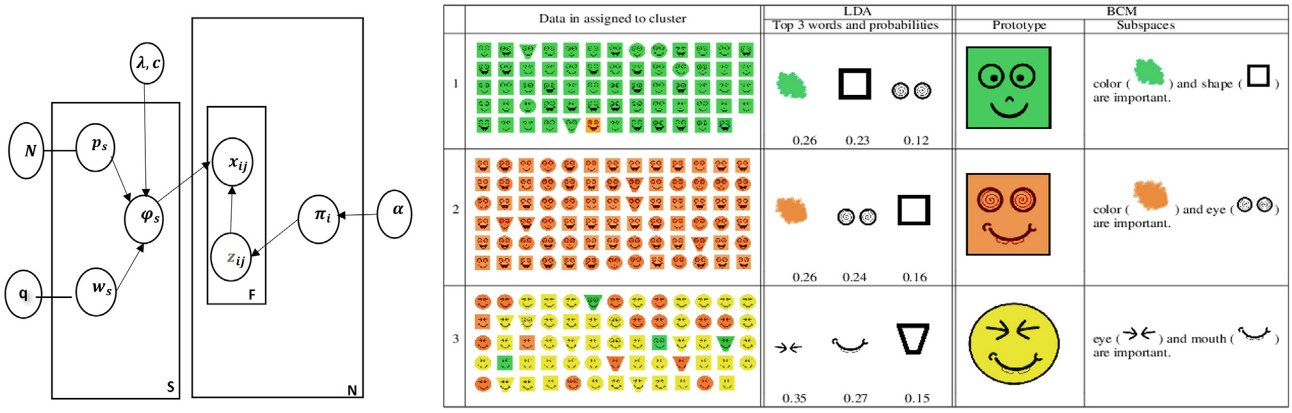


Fig. 3. Workflow of Post-hoc and Ante-hoc Global XAI methods.

Table 2
Summary of research papers published in the past ten years for the Global explanation (Ante-hoc) of AI models

Model usage (type)	Year	Methods	Data type	Methodologies	Explanation medium	Frameworks	XAI evaluation
Ante-hoc (Model-Specific)	2014	BCM [75]	Any	Perturbation-based	Multimedia	Python (PYMC)	Qualitative
	2015	GAM [76]	Tabular	Perturbation-based	Graphics (heatmaps)	R (PyGAM)	Qualitative
	2015	BRL [86]	Tabular	Rule-based	Textual	Python	Quantitative
	2020	NAM [70]	Image	Cluster-based	Graphics (heatmaps)	Pytorch	Quantitative


λ, c
N
p_s
φ_s
q
w_s
x_{ij}
z_{ij}
π_i
α
S
F
N
Data in assigned to cluster
LDA
Top 3 words and probabilities
Prototype
BCM
Subspaces
1
0.26 0.23 0.12
color ( ) and shape ( ) are important.
2
0.26 0.24 0.16
color ( ) and eye ( ) are important.
3
0.35 0.27 0.15
eye ( ) and mouth ( ) are important.

(a) Flow of Bayesian Case Model (BCM)

(b) BCM and LDA interpretation for dataset of smiley faces [77]

Fig. 4. Ante-hoc explanation by Bayesian Case Model.

A new CBR based unsupervised Bayesian model, 'Bayesian Case Model (BCM)' [86], was introduced in 2014 that learns about the notable features to create prototypes that produce accurate and interpretable outcomes on the standard datasets. To understand the generative process of BCM, some mathematical notations used in the BCM flow diagram (see Fig. 4(a)) such as x_i ,$i = 1, 2, \ldots, N$ random mixture over cluster with N observations and S known clusters. For i^{th} observation, π_i denotes the mixture weight over a cluster and x_{ij} indicates observation for j^{th} feature as each observation has F features. Each x_{ij} comes from a cluster denoted as z_{ij} and Z is the full set of the clusters formed by the observation-feature pair. All hyperparameters such as λ, α, q and c are fixed that specifies how much we can copy a prototype to make explanations. B.

Kim and her team members divide the explanatory process of BCM into three parts:

(i) *Prototype* (p_s): is a classic observation in x that is used to represent a cluster say s. For some i and element $j, p_{sj} = x_{ij}$ that maximizes $p(p_s|w_s, z, x)$, where w_s is a feature indicator that we discussed below as a next step of BCM explanatory process.

(ii) *Feature indicator* (w_s): is a binary vector $w_s \, \epsilon \, \{0, 1\}^Q$ of size Q that activates the important features to characterize clusters and prototypes. This vector also indicates the presence of feature j in subspace s.

(iii) *Feature outcome distribution* (ϕ_s): ϕ_{sj} is a vector that explains the discrete probability distribution of possible outcomes for feature j of length U_j, where U_j is the number of possible outcomes.

Fig. 4 demonstrates a graphical representation of the discrete mixture of BCM and Latent Dirichlet Allocation (LDA) method. The authors of the paper [86] used the dataset of a mixture of smiley faces to demonstrate each part of BCM and compare its interpretability accuracy with another method called LDA. The feature set (colours, shapes, and types of mouths and eyes) produces three clusters and each cluster has two features. The instance of BCM with the fixed value of hyperparameters produces 240 smiley faces. However, LDA and BCM represent their outcomes very differently but note that in both approaches selection of important features is the same for each cluster (Fig. 4(b)). It is also worth mentioning that the interpretability of BCM was also verified by performing the human subject experiment in which participants need to understand the formation of clusters for the recipe dataset without any training. In order to evaluate the effectiveness of the learning process, twenty-four participants were divided into two groups and asked to complete the BCM and LDA questionnaires consisting of eight questions each. As an explanation, LDA provided a cluster of top ingredients for each recipe, while BCM presented prototype ingredients without noting the recipe name or subspace. The number of top ingredients from LDA is set as the number of ingredients from BCM prototype and perform Gibbs sampling for LDA until the ground-truth clusters become identifiable. Results show that the explanation run-through BCM achieved higher accuracy (85.9%) than LDA (71.3%) which uses the same Gibbs sampling inference scheme as BCM.

4.1.2. Generalised Additive Model (GAM)

For many years, the Generative Additive Model (GAM) has been used to explain many ML models however, there is a trade-off between the accuracy and intelligibility of these models. In 2015, Caruana et al. introduced an intelligible model by integrating standard GAMs and another model called GA^2Ms. to improve accuracy [88]. Let's assume a training dataset of size N, denoted as $D = (x_k, y_k)_1^N$, where x_k is the feature set with p features and y_k is the response. If G is the link function then the pairwise interaction of GAM is written as

$$G(E[y]) = \gamma_0 + \sum_l f_l(x_k) + \sum_{k \neq l} f_{kl}(x_k, x_l) \tag{3}$$

To make the model intelligible the contribution of each feature can be determined by inspecting f_l , where $E[f_l] = 0$. Eq. (3) helps to understand the strategic flow of the above-paired model as follows: (i) build the finest GAMs model (ii) detect all possible pairs of interaction (iii) rank all top n-potential pairs. The evolution of any interpretable approach in ML can be assumed more valuable if its performance is validated on critical tasks such as healthcare. The performance of GA^2Ms. has been validated by discussing two healthcare problems: 30-day hospital readmission and pneumonia risk. In the 30-day hospital readmission, a large dataset was used with 3956 features such as patient 313 history, doctor's prescription, notes and recommended lab reports. The intelligible model provides a risk score and sorts the important features according to the doctor's requirement. On the other side, the dataset involved in a pneumonia case study is much smaller than the 30-day hospital readmission task and only 46 features participate in predicting the death rate from this lung disease. The GA2Ms model predicts the death rate by learning and editing all possible patterns that would be abandoned even in complex ML models. Hence the combination of standard GAMs and GA^2Ms. are more understandable because the unambiguousness of the model outcome can be visualised by a heat map. The above studies conclude that the GA^2Ms. balanced the trade-off between accuracy and interpretation very well and show practicable accuracy as compared to many ML models [72].

In 2018, a method called CONTENT [89] has been introduced that produces context vectors by transforming the patient's electronic health record into the clinical concept embedding. This method presented the refined but complex hidden knowledge in the context vector by distilling the complex relationship between risk factors present in the patient's record and readmission predictions. Primarily, the distillation technique transfers knowledge from an accurate but complex model to a fast and simple model that have same level accuracy. The distillation approach was also used to explain ICU outcomes by highlighting the important learned features that transfer from complex model to simple model. Many other medical applications such as diabetes classification and breast cancer used knowledge distillation approach to build the interpretation prediction models [90]

4.1.3. Bayes Rule List (BRL)

Mostly rule-based models like decision lists and decision trees are inherently interpretable and many other explainable approaches used them as a part of their algorithms. In 2015, predictive models were introduced that are based on the idea of a decision list and are known as 'Bayes Rule List (BRL)' [91]. The model of the decision list consists of a series of rule statements such as 'if-then' which automatically explains many obvious reasons behind every prediction. BRL is an associative classification method that starts the process by producing the posterior distribution over permutations of 'if-then' large but pre-minded set of rules say, $\mathfrak{R}$. If a data set has N observations $\{x_n, y_n\}$ and let r represent as $r(.) : x_n \in X \rightarrow y_n \in \{0, 1\}$. Let R be a rule set define as

$$R(x) = \begin{cases} 1 & \exists \, r \in R, \; r(x) = 1 \\ 0 & otherwise \end{cases} \tag{4}$$

With $R(.)$ classifier, x is classified as positive if it obeys at least one rule defined in Eq. 4. It is assumed that the interpretability of rules is associated with the number of conditions or length L of rules that are derived from a set of pre-mined rules $\mathfrak{R}$. $\mathfrak{R} = \cup_{l=1}^{L} \; \mathfrak{R}_l$ as $\mathfrak{R}$ is divided into maximum length L that a user allows. In a generative BRL model, the decision list and rules are defined with words like 'if', 'else if', and 'else', accumulation of such words gradually clarifies the rules which make the model understandable. An accurate decision list can be derived when a pre-minded set of rules is sufficiently expressive. MarketScan Medicaid Multi-State Database (MDCD) data of 11.1 million patients were used to explain stroke chances using the BRL method. Besides extracted features (atrial fibrillation condition, gender, age), additional information was collected such as medicines and other medical conditions. This information was used to generate binary predictor variables that confirm the presence or absence of drugs and conditions. The priori distribution helps to add, edit features and rules to create a sample

Table 3
The trustworthy set of rules for predicting future stroke [82].

Rules and conditions	Chances of strokes	Credible intervals
If and Hemiplegia Age > 60	then 58.9%	53.8%-63.8%
else if Cerebrovascular disorder	then 47.8%	44.8%-50.7%
else if Transient ischaemic attack	then 23.8%	19.5%-28.4%
else if Occlusion and stenosis of the carotid artery without infarction	then 15.8%	12.2%-19.6%
else if The altered state of consciousness	then 16.0%	12.2%-20.2%
else if Age > 70	then 4.6%	3.9%-5.4%
else	8.7%	7.9%-9.6%

Table 4
Potential merits and demerits of Global (Ante-hoc) XAI methods.

Ante-hoc methods	Merits	Demerits
BCM	• Ability to capture good information to improve predictions. • Results are easy to explain. • Can achieve more accuracy than LDA.	• Can not handle uncertainty with prior probability. • No correct way to choose a prior. • High computational cost.
GAM	• Able to deal with non-linear and non-monotonic relationships between the response and the predictor variables. • Can deal categorical predictions.	• Computational complexity with a high propensity of overfitting. • Python package is not available.
BRL	• Can handle both continuous and discrete data. • Easy to interpret by highlighting relevant features. • Fast, robust, and used to make real-time predictions.	• Rules focus on classification and almost completely neglects regression. • Bad in describing linear relationships. • Only deal categorical features.
NAM	• Can learn arbitrarily complex relationships between input feature and the output. • Flexible, scalable, and easy to extend. • Can explain result to larger community.	• Great chances of overfitting with the standard initializer. • Produce inconsistent results with Relu activation function.

rule list then BRL tries to optimise these rules. Table 3 highlights the small set of trustworthy rules to predict future chances of stroke. The first three rules are based on other medical disorders such as hemiplegia, cerebrovascular, and transient ischaemic, the chances of strokes seem remarkably high. In the last three columns, vascular disease, and age play an important role to predict the future risk of stroke and chances are comparatively low.

The BRL method is trustworthy because it applies to real medical data where risk is too high as patients with heart disease are so vulnerable. To assure BRL performance and level of accuracy, it was compared with CHADS$_2$ score system to predict chances of stroke in a patient with atrial fibrillation condition. In 2017, the scalability of the BRL method was enhanced by using improved theoretical bounds and tuned language libraries [92]. Hence, the optimised, concise, and reliable rules list generated by the BRL method allows to communicate with the domain experts and implement ML models in other fields such as industry, science, and engineering.

4.1.4. Neural Additive Model (NAM)

A combination of intelligibility of GAMs and expressivity of DNN yields a novel class of model, called 'Neural Additive Model (NAMs)' [93]. One can introduce the Neural Additive Model (NAMs) as an improved, accurate and scalable version of GAMs. NAMs can train multiple DNNs and learn a linear combination for each single input feature. The architecture of NAMs for binary classification is explained in Fig. 5.

The generalisation of the NAMs method can be achieved by parameterising the function f_l , see Eq. 5, in the presence of various hidden layers and neurons. Because of the failure of the ReLU activation function, researchers introduced exp-cantered (ExU) hidden units that help NAMs to learn jagged functions with standard initialisation. This new class of model should learn jagged functions to handle abrupt changes in the datasets relating to real-world problems.

The unit function for ExU hidden units can be calculated with input (x), bias (b) and, weight (w) parameters as:

$$h_u(x) = f(e^w * (x - b)) \tag{5}$$

NAM explains the contribution of each feature from every neural network by calculating the average value of the shape function as a positive and a negative value. Positive values increase the class probability while negative value reduces the chances. The visualisation of shape function by a shape plot can help to understand the model and allow to edit the model as well as the dataset before the final implementation. The interpretation by the NAM method is beneficial for DL as they used several hidden layers and units. From these hidden layers and hidden units, one can compute more complex but accurate shape functions and allow subnets to learn non-linear functions that are required to improve the accuracy of the model's intelligibility.

4.2. Post-hoc methods

The existing and pre-trained AI models can be more valuable if the interpretation of their decision path is understandable along with the accuracy. The post hoc explanation methods required an algorithm to look inside the black box of any DNNs architecture without losing its accuracy. Due to this key advantage of post hoc methods, this approach is also known as *"model-agnostic*". A quick view of post hoc methods are given in Table 5 and Table 6 that summarise information about type of data, frameworks and methodologies, as well as their merits and demerits.

Definition 2. Let the DNN represent as a function $F : R^{D_1} \rightarrow R^{D_k}$. The post hoc explanation, $eXp(F)$, consist of two functions G_1 and G_2 where $G_1 : R^{d_1} \rightarrow R^{d_2}$ represents F such that $d_1 \leqslant D_1$ and $d_2 \leqslant D_k$ and $G_2 : R^{D_1} \rightarrow R^{d_1}$ maps the original input to the valid inputs of the function G_1.

4.2.1. Global Interpretation from local interpretation methods (LIME, LRP, SHAP)

Recently, some researchers extended the existing ideas of local explanation of AI models and deployed them in a way so that they can be used for global or model explanation. This section discusses such methods that have been initially introduced as the local interpretation methods. Originally, a novel explanation method, *'Local Interpretable Model Explanation (LIME)'*, was introduced to explain prediction yield by a single instance. However, to solve real-world problems such as predictions about medical diagnosis, the explanation of a single prediction would not make these models trustworthy. To make the user more confident while using these models, researchers introduced the extended version of LIME called 'Submodular Pick (SP-LIME) [40]. SP-LIME is used to understand the single data instances to understand the global correlations of models.

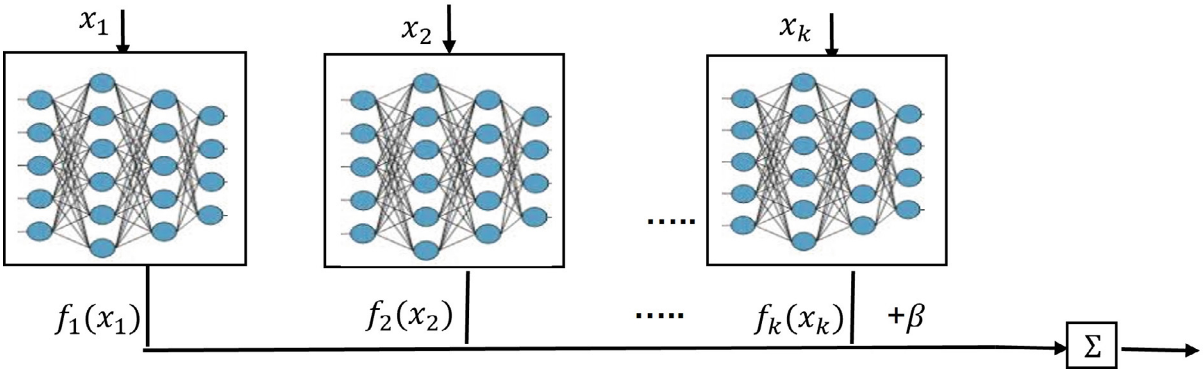


Fig. 5. The Neural Additive Model for binary classification.

Table 5
Summary of research papers published in the past ten years for the Global explanation (Post-hoc) of AI models

Model usage (type)	Year	Methods	Data type	Methodologies	Explanation medium	Frameworks	XAI evaluation
Post-hoc (Model-Agnostic)	2016	SP-LIME[40]	Any	Perturbation-based	Graphics	Python/R	Qualitative
	2015	LRP [86]	Image	Gradient-based	Graphics (heatmaps)	Caffe	Quantitative
	2017	SHAP [52]	Any	Perturbation-based	Multimedia	Python (XGBoost)	Quantitative
	2019	SpRAy [53]	Image	Gradient-based	Graphics	Caffe	Quantitative
	2019	GAA [72]	Image	Perturbation-based	Multimedia	Multi- dimensional	Quantitative
	2019	ACE [94]	Image	Concept-based	Graphics	TensorFlow	Qualitative

Table 6
Potential merits and demerits of Global (Post-hoc) XAI method.

Post-hoc Methods	Merits	Demerits
SP-LIME	• Fast implementation and less robust. • Easy to interpret by expert and non-expert.	• Inherently generate an explanation for local instances. • Do not guarantee the optimal solution.
SHAP	• Fast implementation. • Contrastive explanations. • Consistent interpretation.	• Slow computation. • Shapley values can be misinterpreted.
SpRay	• Detect any kind of anomaly. • Ability to explain complex DNNs by highlighting important feature (heatmaps). • Low computation and storage cost.	• Only qualitatively evaluation is available. • Heatmaps are sensitive to specific features.
GAA	• Can represent Nonlinear relationship. • Provides tunable subpopulation granularity. • Easy to implement.	• Computationally expensive due to large number of features.
ACE	• Generate meaningful,important, and coherent concept to explain DNNs.	• Need lot of image segmentation processes while generating explanation for image dataset. • Can not generate explanation for the complex concept.

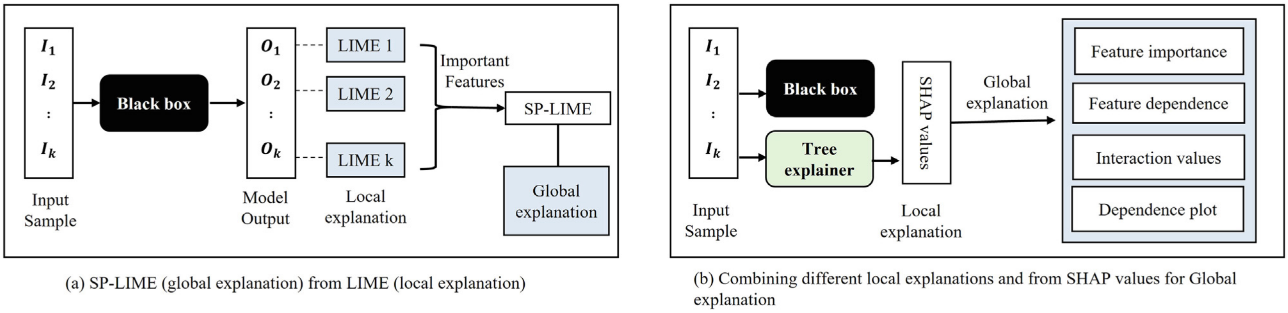
Fig. 6(a) describes the idea of the SP-LIME algorithm which provides the global explanation by fetching important features from each instance. If the given set of instances are *I*, we can choose *B* ‘Budget’ as the required number of explanations. Firstly, we can run the LIME algorithm for the available set of instances and save the explanation of each instant into the ‘explanation matrix’ say *L*. The explanation matrix helps to extract the important features of the given model. The greedy optimisation technique is applied to the new matrix of size *IB* generated by the SP-LIME algorithm which chooses the minimum number of inputs and extracts the maximum number of important features.

The behaviour of the SP-LIME is similar to surrogate models as it extracts useful information and independent explanation from the LIME method. SP-LIME would be preferable over LIME because it provides a non-redundant and global view of the model to trust those models. There is another method, called ‘Layer-wise Relevance BackPropagation (LRP)’, which was initially used to explain the single instance of prediction [94]. The decomposition of prediction helps to calculate the relevance score for every individual input feature. Many deep architectures of the neural network such as CNNs and RNNs use the backpropagation and update the relevance scores to explain the single prediction by generating heat maps. Recently, LRP utilises these heat maps as input for their global explanation algorithm. Network pruning is another way to use LRP that helps to reduce the memory cost of the AI model without sacrificing accuracy [45]. The relevance score generated by LRP highlights the least important features that are eventually removed from the model to prune it.

In [49], the *‘SHapley Additive ExPlanation (SHAP)’* method calculates Shapley values and explains the prediction on behalf of feature contribution towards a certain output. The calculation of Shapley value is based on the concept of coalition game theory, where a prediction is treated as ‘payoff, and the value of each feature is assumed ‘player’. Shapley values state the fair distribution of payoff (prediction) among the players (features). The mathematical formulation for computing contribution of each feature is given as:

$$g(z)' = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{6}$$

Where *g* is the explanation model, ϕ_j is feature attribution for $j^t h$ feature, $z' \in [0,1]^M$ and *M* is the maximum size of the coalition. Different versions of SHAP like KernelSHAP, LinearSHAP, and DeepSHAP were introduced to explain the individual prediction for various types of datasets. In [95], a framework, called TreeExplainer, was introduced as an extension of SHAP for trees. The algorithm



(a) SP-LIME (global explanation) from LIME (local explanation)

(b) Combining different local explanations and from SHAP values for Global explanation

Fig. 6. Global explanation from Local interpretation SP-LIME and SHAP methods.

behind TreeExplainer finds the local interpretation for trees by computing the Shapley values, then structures them in such a way so that the model can explain the features at the global level. Without loss of consistency and accuracy, the extended version of SHAP, TreeExplainer, provides a quick local explanation for trees in polynomial times. The global interpretations include interaction and clustering values, summary and dependence plots, and feature importance (Fig. 6(b)). With SHAP, the global interpretation of the model becomes easier due to the fast computing ability of Shapley values.

4.2.2. Spectral Relevance Activation (SpRAy)

The SpRAy technique for explaining the AI models at the global level was introduced by Lapuschkin in 2019 [54]. This technique is based on the LRP method which explains the model for an instance. To view insight into the model and explain the decision-making process, a spectral clustering algorithm was applied to the local explanations of the model produced by the LRP method. The results produced by the LRP method help to spot and analyse those attributions which appear frequently with spatial structure. This spatial analysis would help the SpRAy method to identify any anomaly in the model.

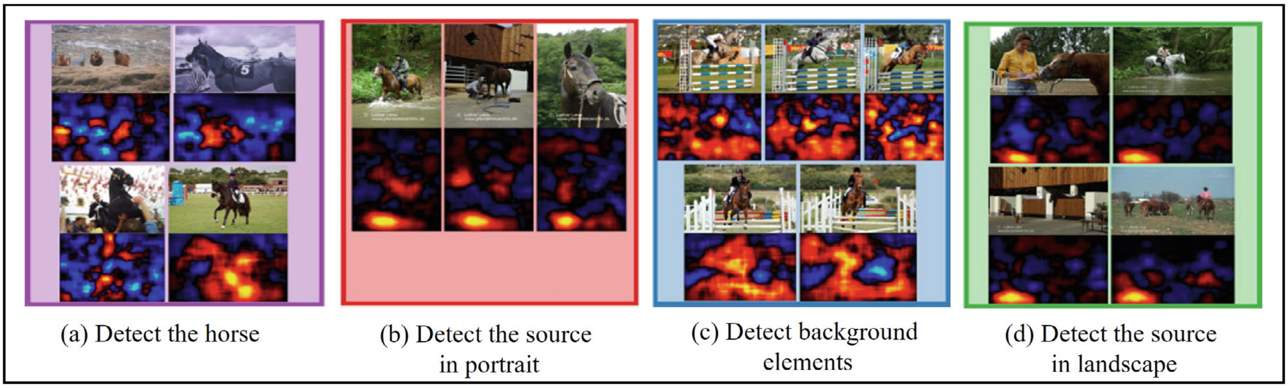
The algorithm of the SpRAy technique can be summarized as follows. (i) Firstly it uses the LRP method and finds local relevance maps that are used to explain every data instance. (ii) it then makes the definite and visible solution, scales down the relevance map to uniform size and shape (iii) it then evaluates the LRP relevance maps by using Spectral Cluster (SC) analysis to design clusters for the local explanations (iv) and finally it uses eigen maps analysis to compute eigen gap among two successive clusters and return relevant cluster to the user. Lastly, as an optional step, the user can visualise these clusters by using t-Stochastic Neighbour Embedding (t-SNE).

The paper [54] explains the SpRAy algorithm by implementing on the horse images of the PASCAL VOC dataset. Fig. 7 shows the following four different prediction strategies to classify horses. (i) spot the presence of rider and horse (ii) highlight the existence of source codes on portraits-based images (iii) identify some background elements (iv) and highlight tags on landscape-oriented images. On this large dataset, the SpRAy method acts as a semi-automated tool that can also detect any kind of anomaly such as misuses of source tags in horse images. Hence, without human intrusion, the combination of LRP and SpRAy methods enables the user to identify the strategies behind prediction and visualises them with the aid of heat maps.

4.2.3. Global Attribution Analysis (GAA)

Although the discussed global interpretation techniques explain the decisions by summarising local attributions or providing a set of rules, these methods failed to learn about the non-linear interactions of features across subpopulations during the training process. A technique called ‘Global Attribution Analysis (GAA)’ has been introduced in 2019 that produces explanation even for subpopulation by generating global attribution [96]. Each global attribution explains the specific part of the model that leads to the global explanation of the model. Fig. 8 shows the workflow of the GAA method. Firstly, the information about the local features is collected by employing some local interpretation methods such as LIME, DeepLIFT and, Integrated Gradient. At this stage, every local attribution highlights the significant features for a single prediction and treats these attributions as weighted conjoined rankings. To avoid anomalies, these local attribution vectors are normalised by

$$|\delta_w| \circ \frac{1}{(\sum_i |\delta_w(i)|} \tag{7}$$


(a) Detect the horse

(b) Detect the source in portrait

(c) Detect background elements

(d) Detect the source in landscape

Fig. 7. The four strategies of SpRAy method to classify horses [54].

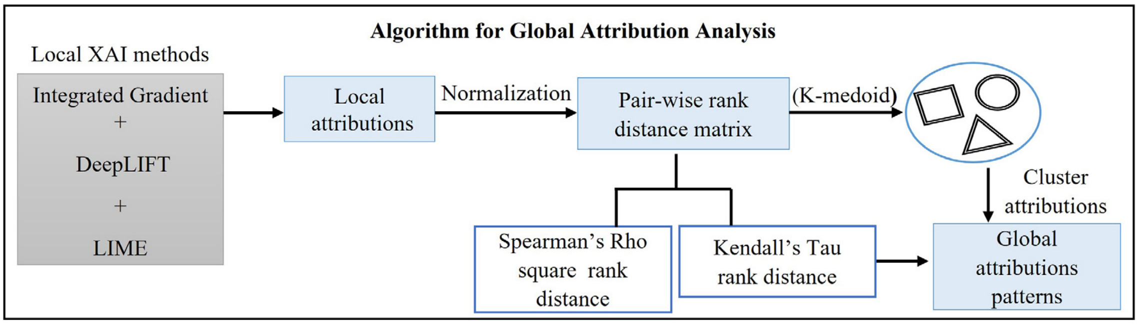

Algorithm for Global Attribution Analysis
Local XAI methods
Integrated Gradient
+
DeepLIFT
+
LIME
Local attributions
Normalization
Pair-wise rank distance matrix
(K-medoid)
Cluster attributions
Spearman's Rho square rank distance
Kendall's Tau rank distance
Global attributions patterns

Fig. 8. Algorithm for Global Attribution Analysis for global interpretation.

In Eq. 7, δ_w is a weighted attribution vector and $\circ$ is used to represent the Hadamard product. Next, the following two options are used to compare these normalised attributions and quantify similarities among them: (i) Kendall's Tau rank distance [97] (ii) Spearman's Rho square rank distance [98]. After this comparison, the GAA method uses a clustering algorithm, K-medoid [99], to make clusters of similar attributions and identify global attribution patterns. Hence, GAA allows us to look and find out differences among explanations of subpopulations. Fig. 8 depicts each step of GAA algorithms. In addition, GAA also offers a tuneable granularity to get information about the preferred number of subpopulations.

4.2.4. Automatic Concept-based Explanation (ACE)

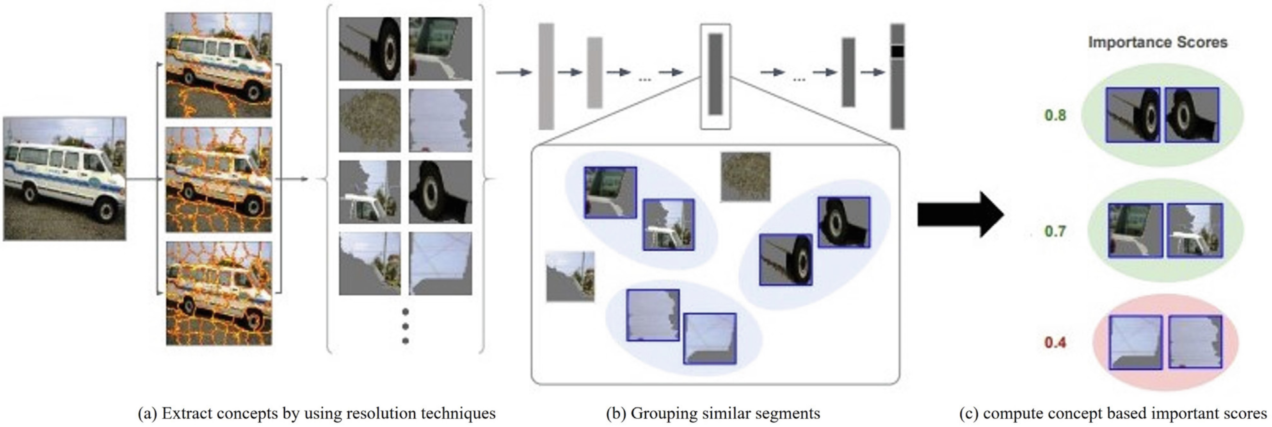
A concept-based method, called 'Automatic Concept-Based Explanation (ACE),' has been discussed in [100] that is used to globally explain the trained classification models such as CNN and Inception-V3. ACE suggests only those concepts for the explanations that are indispensable and coherent for a model's prediction as well as meaningful and understandable for humans.

The authors of the paper explained the ACE algorithm step by step (Fig. 9). They pick a trained classifier with a set of images as input data. In step one, the method extracts all concepts present in the images in the form of segments (groups of pixels). The method then applies different levels of resolution techniques to fully capture the hierarchy of concepts. Usually, three levels of resolution are considered enough to capture colours, texture, objects, or even their parts.

In the second step, ACE picks those segments that represent the same concepts and puts them in a group. To measure similarity among segments, the Euclidean distance (say δ) can be used, this distance also helps to remove the concepts with low similarity and maintain the coherency of the model. All these steps can be understood by the visual source provided in [101]. Lastly, the method uses any method like TCAV [102] to highlight important concepts by computing concept-based importance scores. This method is only performed on the image datasets as it is easy to group pixels in a meaningful way, this could be a big drawback of ACE.

5. Summary

We summarize the taxonomy that has been discussed in the previous sections and is visually depicted in Fig. 10. Among the explainable methods discussed above, not all kinds of data can be processed. There are methods that take into account numerical, binary, and categorical data, which are tabular, while others generate explanations by highlighting data that is comprised of pixels, which are images. Fig. 10 illustrates those objects that are explained within the development process to provide global explanations. Some methods focus on accessing the internal representation, such as layers, features, or vectors, and others explain how the model is trained. In this paper, we include only those explanators that certainly contribute to explaining the decision-route generated by the above-described explainable models. (i) *Saliency Map (SM):* is an efficient way to visually highlight and mask the causes of certain outcomes [103]. (ii) *Decision Tree (DT):* is easily understandable, also known as single tree approximation, and primarily used for the global explanation [104]. (iii) *Partial Dependence Plot (PDP):* plot the relationship between the outcome of the black box and the input [105]. (iv) *Decision Rule (DR):* is the most human understandable explanation technique that is used to transform the decision tree into a set of rules [96]. (v) *Prototype Selection (PS):* consists of returning the outcome with a set of similar



(a) Extract concepts by using resolution techniques (b) Grouping similar segments (c) compute concept based important scores

Fig. 9. Step by step Automatic Concept based Explanation algorithm. [101].

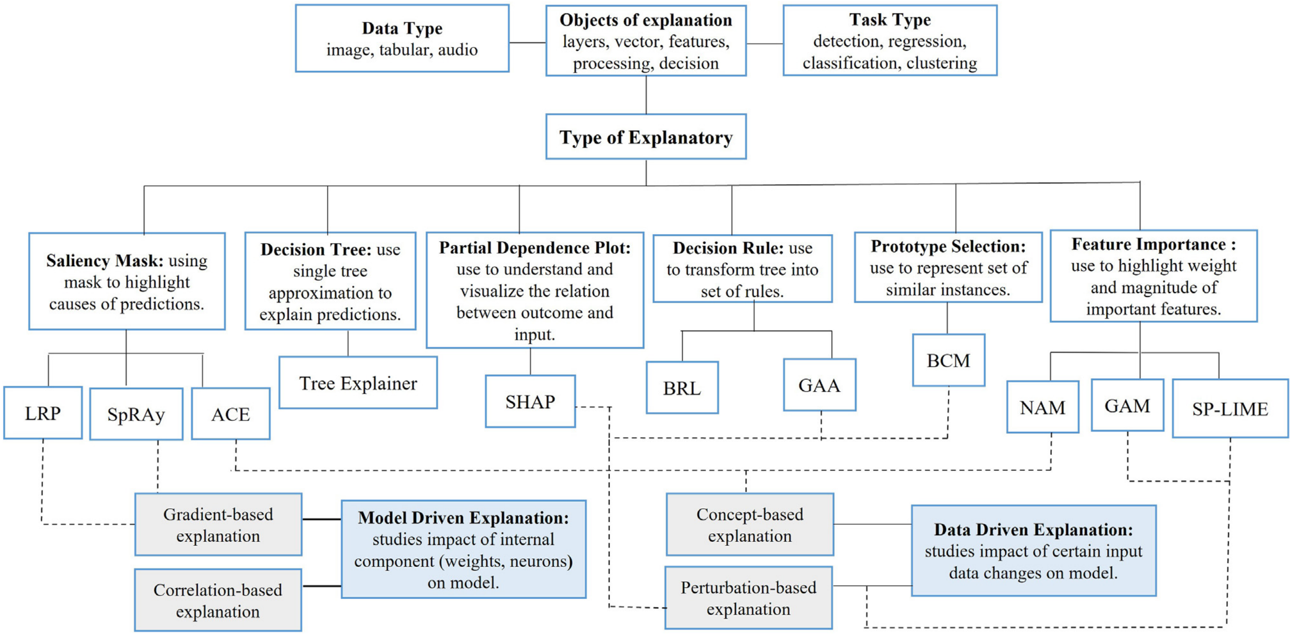

Data Type
image, tabular, audio
Objects of explanation
layers, vector, features, processing, decision
Task Type
detection, regression, classification, clustering
Type of Explanatory
Saliency Mask: using mask to highlight causes of predictions.
Decision Tree: use single tree approximation to explain predictions.
Partial Dependence Plot: use to understand and visualize the relation between outcome and input.
Decision Rule: use to transform tree into set of rules.
Prototype Selection: use to represent set of similar instances.
Feature Importance : use to highlight weight and magnitude of important features.
LRP
SpRAy
ACE
Tree Explainer
SHAP
BRL
GAA
BCM
NAM
GAM
SP-LIME
Gradient-based explanation
Correlation-based explanation
Model Driven Explanation: studies impact of internal component (weights, neurons) on model.
Concept-based explanation
Perturbation-based explanation
Data Driven Explanation: studies impact of certain input data changes on model.

Fig. 10. Explanatory Taxonomy of Data and Model Driven Global Explainable methods.

Table 7
Potential merits and demerits of interpretable Explanators.

Explanators	Merits	Demerits
Saliency Map	• Highlight important pixels. • Faster computation.	• Only qualitative evaluation is available. • Insensitive to model and data.
Decision Tree	• Easy to explain. • Need less effort for data preparation.	• Fail to deal with linear relationships. • Difficult and expensive to interpret deeper tree.
Partial Dependence Plot	• Easy to understand and interpret. • Easy to implement.	• Deal with maximum two features. • Hidden Heterogeneous effect.
Decision Rule	• Select only the relevant features. • Cost efficient.	• Difficult and tedious to list all the rules. • Fail to describe linear relationship.
Prototype Selection	• Easy detection of missing functionality. • Detect error at early stage.	• Expensive. • Higher number of features or clusters.
Feature Importance	• East interpretation. • Highly compressed and insight model globally.	• Expensive. • Time consuming.

instances [106]. (vi) *Feature Importance (FI):* is an effective but simple explanation solution that highlights and returns weights and features with a high magnitude [107]. In Table 7, these methods have been summarized with their advantages and disadvantages.

Perturbation-based Global explainable methods: The summary provided in Table 2 and 5 shows that most global explainable methods are based on the perturbation algorithm. Perturbation mainly focuses on perturbing the set of features (e.g. pixels) of the given input data by masking, occlusion, or filling operations. After finding a set of perturbations, a new set of predictions can be obtained by using the parameters of DNNs. To determine the significance of different features, these predictions are compared with the original data and an explanation is generated with the predefined set of explanation rules. Generally, Global methods such as BCM [77], GAM [87], SP-LIME [40], GAA [93], and, SHAP [52] use only forward pass to understand the neurons' activities and impact of each feature to demonstrate models' attributions. Perturbation methods provide a visual explanation (heatmaps, saliency maps) to explain the influencing features of images, videos, and natural language as input but only few of them have been evaluated at the qualitative human experiment level. Perturbation or Concept-based explanations are *Data Driven Explanations* that completely relies on input data for generating explanation. As only small change in input data can impact the Data Driven explanation so this approach does not need to understand the inner functionality of model [108].

Gradient-based Global explainable methods: On the contrary, gradient-based methods, understand the neurons' activities by doing more than one forward pass and use partial derivatives of activation to generate attribution representation during backpropagation. Naturally, gradient-based methods such as LRP [91] and SpRay [53] generates the human understandable visual explanation but there is no discussion and evaluation of these methods at the qualitative level to gain trust on the AI models especially for applications such as medical surgeries and autonomous vehicles. Correlation-based algorithms compute correlation scores rather than gradients by using the backpropagation technique. Under the set of constraints, correlation-based methods such as DeepLIFT [109] generate reasonable explanations. DeepLIFT calculates the scores based on a comparison between the values of activated neurons and the reference values. In some cases, DeepLIFT may consider both negative and positive values to observe the effect of each neuron. The Gradient-based explanation falls on the *Model Driven Explanation* category that analyse internal components such as weights and neurons to generate explanation [110]. Besides the above two methodologies, some global explainable methods are based on the defined rules [91], concepts [97], and clusters [72] approach. The major drawback of most of the discussed global explanation methods is the dearth of human subject experiments. Arguments, reasoning, and explanations are more effective if they help the end-user to build a true picture of the entire model process. There are two ways to keep humans in a loop while evaluating the explainable methods. First, a random selection of lay people (without technical or domain knowledge) interacting with explanation tools and providing their responses/feedback through the designed questionnaire by AI experts. Secondly, domain experts providing their opinions on the explanation tools and using domain knowledge to verify their consistency. Different free libraries and frameworks such as Python, R, Caffe, Pytorch, Keras and TensorFlow are available to generate textual, visual, audio, and video explanations depending on the type of dataset and the demand of the end-user.

6. Conclusion and Open Challenges with Future Directions

It can be observed that in the past decade, many researchers primarily paid attention to the development of global interpretation methods, even though they utilised many existing local or visual explanation methods for the complete model interpretation [111,112]. Although with sufficient accuracy, the local interpretation methods of AI can improve the user's trust, these methods never reveal the complete structure of the AI model. Therefore, it is considered as the biggest drawback of the local interpretation methods. This paper presents a brief history of global XAI methods from the mid of 20^{th} century to 21^{th} century, then a taxonomy of the global interpretation AI methods produced in the past ten years is presented. This, also provides answers to questions such as why the complete explainability and interpretability of AI models are so important and how the vague understanding of AI models and relating technologies would affect the human life? There is an inadequate illustration of some terminologies that are commonly but interchangeably used in the XAI field. The explanation of these terminologies in this paper will help the readers in understanding XAI methods. This survey will provide a detailed insight into the recent developments on global XAI methods, existing challenges, and the possible path towards trustworthy XAI methods that would be understandable by a human. The central focus of this survey paper is around answering the question: how XAI methods can be completely explained for their structure and decision routes? There is considerable work done in the past ten years on the global interpretation methods that have been highlighted and summarized in this paper and this area of research has been continuously evolved by introducing new and novel approaches. The key findings can be summarized as follows:

1. Mostly the existing interpretation methods explain the decision-making process of the DNNs by using local or visual approaches. However, these approaches are inadequate to explain the full architecture of the DNNs as the local methods generate an explanation just by following the decision route for one single instance at a time. While knowing the rationale for all possible outcomes, the global interpretation methods can explain the complete architecture of DNNs.
2. Existing global interpretation methods such as Global Attribution Analysis (GAA) explain the model at the global level by using some existing local interpretation methods such as LIME, integrated gradient, and DeepLIFT. These methods become computationally expensive as the number of features and parameters are quite high in the deep architectures of neural networks.
3. Some existing approaches such as surrogate models (LIME) approximate the black box model to explain the decisions. These interpretation models may be close to the black box model for one subset of a dataset but diverge widely for other subsets.
4. The existing global interpretation methods have been used to explain DNNs for the data types such as images and tabular, there is no global interpretation method for text datasets that can illuminate the decision rationale executed by DNNs.

While deliberating the existing research gaps in the state of the art of global XAI methods, we have a significant opportunity to discuss and establish some future research goals and directions for academic researchers.

- The DNNs have model-free architecture and the existing global interpretation methods produced an explanation by approximating the black box model of DNNs. To make model-oriented architecture one needs to introduce a mathematical model that should explain each decision of the black box model of DNN deterministically and represent and govern the learning evolution happening in each iteration.
- The existing approaches are expensive in terms of computation complexity because of the stochastic behaviour and performance of many DNNs models. The deterministic explanation approach would explain the model with low computational cost and make the interpretation more accurate no matter how many times we execute this model.

The proposed future directions demand developing novel deterministic models that can highlight the influencing features and mathematically figure out the contribution of each part in the decision-making process. Consequently, the proposed deterministic explainable model will reduce the computational cost as it can provide desirable outcomes in fewer iterations. Moreover, the DNNs will become more trustworthy and reliable for risky applications as we would have a controlled learning process within DNNs. In the end, it is worth mentioning that interpretation methods should be built under some constraints such as data privacy and model confidentiality because explainability may lead to revealing some sensitive information about the model unless the experimentation and execution is carried out in a protected and compliant environment.

CRediT authorship contribution statement

Rabia Saleem: Conceptualization, Investigation, Writing – original draft, Writing – review & editing. **Bo Yuan:** Supervision, Validation, Writing – review & editing. **Fatih Kurugollu:** Supervision, Validation, Writing – review & editing. **Ashiq Anjum:** Supervision, Validation, Visualization, Writing – review & editing. **Lu Liu:** Validation, Visualization, Writing – review & editing.

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.

Acknowledgements

This work was carried out in the Data Science Research Center (DSRC) at the University of Derby.

References

[1] L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo, L. Zhang, A deep structured model with radius-margin bound for 3d human activity recognition, International Journal of Computer Vision 118 (2015) 256–273.

[2] N.D. Doulamis, A. Voulodimos, Fast-mdl: Fast adaptive supervised training of multi-layered deep learning models for consistent object tracking and classification, in: 2016 IEEE International Conference on Imaging Systems and Techniques (IST), 2016, pp. 318–323.

[3] N. Zeng, P. Wu, Z. Wang, H. Li, W. Liu, X. Liu, A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection, IEEE Transactions on Instrumentation and Measurement 71 (2022) 1–14, https://doi.org/10.1109/TIM.2022.3153997.

[4] X. Chen, B. Zhang, D. Gao, Bearing fault diagnosis base on multi-scale cnn and lstm model, Journal of Intelligent Manufacturing 32 (4) (2021) 971–987.

[5] C. You, J. Lu, D. Filev, P. Tsiotras, Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning, Robotics and Autonomous Systems 114 (2019) 1–18.

[6] S. Grigorescu, B. Trasnea, T. Cocias, G. Macesanu, A survey of deep learning techniques for autonomous driving, Journal of Field Robotics 37 (3) (2020) 362–386.

[7] A. Boles, P. Rad, Voice biometrics: Deep learning-based voiceprint authentication system, in: 2017 12th System of Systems Engineering Conference (SoSE), IEEE, 2017, pp. 1–6.

[8] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, K. Dietmayer, Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges, IEEE Transactions on Intelligent Transportation Systems 22 (3) (2020) 1341–1360.

[9] N.M. Rad, S.M. Kia, C. Zarbo, T. van Laarhoven, G. Jurman, P. Venuti, E. Marchiori, C. Furlanello, Deep learning for automatic stereotypical motor movement detection using wearable sensors in autism spectrum disorders, Signal Processing 144 (2018) 180–191.

[10] A.S. Heinsfeld, A.R. Franco, R.C. Craddock, A. Buchweitz, F. Meneguzzi, Identification of autism spectrum disorder using deep learning and the abide dataset, NeuroImage: Clinical 17 (2018) 16–23.

[11] S.H. Silva, A. Alaeddini, P. Najafirad, Temporal graph traversals using reinforcement learning with proximal policy optimization, IEEE Access 8 (2020) 63910–63922.

[12] Z. Wan, R. Yang, M. Huang, W. Liu, N. Zeng, Eeg fading data classification based on improved manifold learning with adaptive neighborhood selection, Neurocomputing 482 (2022) 186–196.

[13] Z. Wan, R. Yang, M. Huang, N. Zeng, X. Liu, A review on transfer learning in eeg signal analysis, Neurocomputing 421 (2021) 1–14.

[14] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, et al., Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy, Ophthalmology 126 (4) (2019) 552–564.

[15] A. Das, P. Rad, K.-K.R. Choo, B. Nouhi, J. Lish, J. Martel, Distributed machine learning cloud teleophthalmology iot for predicting amd disease progression, Future Generation Computer Systems 93 (2019) 486–498.

[16] J. Son, J.Y. Shin, H.D. Kim, K.-H. Jung, K.H. Park, S.J. Park, Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images, Ophthalmology 127 (1) (2020) 85–94.

[17] G.D.L.T. Parra, P. Rad, K.-K.R. Choo, N. Beebe, Detecting internet of things attacks using distributed deep learning, Journal of Network and Computer Applications 163 (2020) 102662.

[18] A.D. Torres, H. Yan, A.H. Aboutalebi, A. Das, L. Duan, P. Rad, Patient facial emotion recognition and sentiment analysis using secure cloud with hardware acceleration, in: Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications, Elsevier, 2018, pp. 61–89.

[19] S.M. Lee, J.B. Seo, J. Yun, Y.-H. Cho, J. Vogel-Claussen, M.L. Schiebler, W.B. Gefter, E.J. Van Beek, J.M. Goo, K.S. Lee, et al., Deep learning applications in chest radiography and computed tomography, Journal of thoracic imaging 34 (2) (2019) 75–85.

[20] R. Chen, L. Yang, S. Goodison, Y. Sun, Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data, Bioinformatics 36 (5) (2020) 1476–1483.

[21] D. Castelvecchi, Can we open the black box of ai?, Nature News 538 (7623) (2016) 20

[22] D. Dave, H. Naik, S. Singhal, P. Patel, Explainable ai meets healthcare: A study on heart disease dataset, 2020, arXiv preprint arXiv:2011.03195.

[23] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, arXiv preprint arXiv:2201.08164 (2022).

[24] H.S. Kapoor, K. Jain, S.K. Sharma, Generalized additive model for evaluation of premium for diabetic patients, Journal of Advances in Applied Mathematics 1 (3) (2016).

[25] M. Craven, J. Shavlik, Extracting tree-structured representations of trained networks, Advances in neural information processing systems 8 (1995).

[26] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning 2nd ed springer series in statistics (2009).

[27] A. Goldstein, A. Kapelner, J. Bleich, E. Pitkin, Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation, journal of Computational and Graphical Statistics 24 (1) (2015) 44–65.

[28] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232.

[29] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, K.-R. Müller, How to explain individual classification decisions, The, Journal of Machine Learning Research 11 (2010) 1803–1831.

[30] A.-M. Leventi-Peetz, T. Östreich, Deep learning reproducibility and explainable ai (xai), arXiv preprint arXiv:2202.11452 (2022).

[31] K. Weitz, T. Hassan, U. Schmid, J.-U. Garbas, Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable ai methods, tm-Technisches Messen 86 (7–8) (2019) 404–412.

[32] M.M. De Graaf, B.F. Malle, How people explain action (and autonomous intelligent systems should too), in: 2017 AAAI Fall Symposium Series, 2017.

[33] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (5) (2018) 1–42.

[34] M.T. Ribeiro, S. Singh, C. Guestrin, Model-agnostic interpretability of machine learning, arXiv preprint arXiv:1606.05386 (2016).

[35] Z.C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery, Queue 16 (3) (2018) 31–57.

[36] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, X. Yi, A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability, Computer Science Review 37 (2020) 100270.

[37] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115.

[38] Z. Lipton, The mythos of model interpretability, Queue 16 (3) (2018), 30: 31–30: 57.

[39] S.J. Oh, B. Schiele, M. Fritz, Towards reverse-engineering black-box neural networks, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, 2019, pp. 121–144.

[40] M.T. Ribeiro, S. Singh, C. Guestrin, , why should i trust you? explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[41] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (xai): A survey, arXiv preprint arXiv:2006.11371 (2020).

[42] D. Erhan, A. Courville, Y. Bengio, Understanding representations learned in deep architectures, Tech. rep., Technical Report 1355, Université de Montréal/DIRO (2010).

[43] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034 (2013).

[44] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, Smoothgrad: removing noise by adding noise, arXiv preprint arXiv:1706.03825 (2017).

[45] S.-K. Yeom, P. Seegerer, S. Lapuschkin, A. Binder, S. Wiedemann, K.-R. Müller, W. Samek, Pruning by explaining: A novel criterion for deep neural network pruning, Pattern Recognition 115 (2021) 107899.

[46] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.

[47] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[48] A. Miller, J. Panneerselvam, L. Liu, A review of regression and classification techniques for analysis of common and rare variants and gene-environmental factors, Neurocomputing (2021).

[49] S. Mishra, B.L. Sturm, S. Dixon, Local interpretable model-agnostic explanations for music content analysis., in: ISMIR, Vol. 53, 2017, pp. 537–543.

[50] T. Peltola, Local interpretable model-agnostic explanations of bayesian predictive models via kullback-leibler projections, arXiv preprint arXiv:1810.02678 (2018).

[51] S. Shi, X. Zhang, W. Fan, A modified perturbed sampling method for local interpretable model-agnostic explanation, arXiv preprint arXiv:2002.07434 (2020).

[52] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[53] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.

[54] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking clever hans predictors and assessing what machines really learn, Nature communications 10 (1) (2019) 1–8.

[55] C. Seifert, A. Aamir, A. Balagopalan, D. Jain, A. Sharma, S. Grottel, S. Gumhold, Visualizations of deep neural networks in computer vision: A survey, in: Transparent data mining for big and small data, Springer, 2017, pp. 123–144.

[56] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding neural networks through deep visualization, arXiv preprint arXiv:1506.06579 (2015).

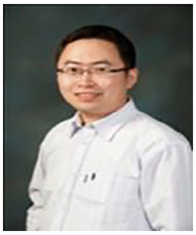
[57] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.

[58] J.T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, arXiv preprint arXiv:1412.6806 (2014).
[59] A. Mahendran, A. Vedaldi, Understanding deep image representations by inverting them, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5188–5196.
[60] A. Mahendran, A. Vedaldi, Visualizing deep convolutional neural networks using natural pre-images, International Journal of Computer Vision 120 (3) (2016) 233–255.
[61] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, J. Clune, Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, Advances in neural information processing systems 29 (2016).
[62] G. Hooker, Discovering additive structure in black box functions, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 575–580.
[63] J.J. Thiagarajan, B. Kailkhura, P. Sattigeri, K.N. Ramamurthy, Treeview: Peeking into deep neural networks via feature-space partitioning, arXiv preprint arXiv:1611.07429 (2016).
[64] A. Datta, S. Sen, Y. Zick, Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems, in: 2016 IEEE symposium on security and privacy (SP), IEEE, 2016, pp. 598–617.
[65] J.D. Olden, D.A. Jackson, Illuminating the black box: a randomization approach for understanding variable contributions in artificial neural networks, Ecological modelling 154 (1–2) (2002) 135–150.
[66] P. Cortez, J. Teixeira, A. Cerdeira, F. Almeida, T. Matos, J. Reis, Using data mining for wine quality assessment, in: International Conference on Discovery Science, Springer, 2009, pp. 66–79.
[67] L. Breimann, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and regression trees, Wadsworth, Pacific Grove, 1984.
[68] S. Hara, K. Hayashi, Making tree ensembles interpretable, arXiv preprint arXiv:1606.05390 (2016).
[69] A.D. Arbatli, H.L. Akin, Rule extraction from trained neural networks using genetic algorithms, Nonlinear Analysis: Theory, Methods & Applications 30 (3) (1997) 1639–1648.
[70] Z.-H. Zhou, Y. Jiang, S.-F. Chen, Extracting symbolic rules from trained neural network ensembles, Ai Communications 16 (1) (2003) 3–15.
[71] Y. Lou, R. Caruana, J. Gehrke, Intelligible models for classification and regression, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp. 150–158.
[72] Y. Lou, R. Caruana, J. Gehrke, G. Hooker, Accurate intelligible models with pairwise interactions, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, pp. 623–631.
[73] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 1721–1730.
[74] A. Henelius, K. Puolamäki, H. Boström, L. Asker, P. Papapetrou, A peek into the black box: exploring classifiers by randomization, Data mining and knowledge discovery 28 (5) (2014) 1503–1529.
[75] A. Zien, N. Krämer, S. Sonnenburg, G. Rätsch, The feature importance ranking measure, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2009, pp. 694–709.
[76] S. Sonnenburg, A. Zien, P. Philips, G. Rätsch, Poims: positional oligomer importance matrices–understanding support vector machine-based signal detectors, Bioinformatics 24 (13) (2008) i6–i14.
[77] M.M.-C. Vidovic, N. Görnitz, K.-R. Müller, M. Kloft, Feature importance measure for non-linear learning algorithms, arXiv preprint arXiv:1611.07567 (2016).
[78] T. Calders, S. Verwer, Three naive bayes approaches for discrimination-free classification, Data mining and knowledge discovery 21 (2) (2010) 277–292.
[79] B. Ustun, C. Rudin, Supersparse linear integer models for optimized medical scoring systems, Machine Learning 102 (3) (2016) 349–391.
[80] D. Wei, S. Dash, T. Gao, O. Gunluk, Generalized linear rule models, International Conference on Machine Learning, PMLR (2019) 6687–6696.
[81] H. Elzayn, S. Jabbari, C. Jung, M. Kearns, S. Neel, A. Roth, Z. Schutzman, Fair algorithms for learning in allocation problems, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 170–179.
[82] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, K.R. Varshney, Optimized pre-processing for discrimination prevention, Advances in neural information processing systems 30 (2017).
[83] M.B. Zafar, I. Valera, M. Rodriguez, K. Gummadi, A. Weller, From parity to preference-based notions of fairness in classification, Advances in Neural Information Processing Systems 30 (2017).
[84] N. Grgić-Hlača, M.B. Zafar, K.P. Gummadi, A. Weller, Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.
[85] M. Kearns, S. Neel, A. Roth, Z.S. Wu, Preventing fairness gerrymandering: Auditing and learning for subgroup fairness, in: International Conference on Machine Learning, PMLR, 2018, pp. 2564–2572.
[86] B. Kim, C. Rudin, J.A. Shah, The bayesian case model: A generative approach for case-based reasoning and prototype classification, Advances in neural information processing systems 27 (2014).
[87] G.A. Klein, Do decision biases explain too much, Human Factors Society Bulletin 32 (5) (1989) 1–3.
[88] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 1721–1730.
[89] C. Xiao, T. Ma, A.B. Dieng, D.M. Blei, F. Wang, Readmission prediction via deep contextual embedding of clinical concepts, PloS one 13 (4) (2018) e0195024.
[90] Y. Ming, H. Qu, E. Bertini, Rulematrix: Visualizing and understanding classifiers with rules, IEEE transactions on visualization and computer graphics 25 (1) (2018) 342–352.
[91] B. Letham, C. Rudin, T.H. McCormick, D. Madigan, Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model, The Annals of Applied Statistics 9 (3) (2015) 1350–1371.
[92] H. Yang, C. Rudin, M. Seltzer, Scalable bayesian rule lists, in: International conference on machine learning, PMLR, 2017, pp. 3921–3930.
[93] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, G.E. Hinton, Neural additive models: Interpretable machine learning with neural nets, Advances in Neural Information Processing Systems 34 (2021).
[94] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (7) (2015) e0130140.
[95] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, Explainable ai for trees: From local explanations to global understanding, arXiv preprint arXiv:1905.04610 (2019).
[96] M. Ibrahim, M. Louie, C. Modarres, J. Paisley, Global explanations of neural networks: Mapping the landscape of predictions, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 279–287.
[97] P.H. Lee, L. Philip, Distance-based tree models for ranking data, Computational Statistics & Data Analysis 54 (6) (2010) 1672–1682.
[98] G.S. Shieh, Z. Bai, W.-Y. Tsai, Rank tests for independence–with a weighted contamination alternative, Statistica Sinica (2000) 577–593.
[99] H.-S. Park, C.-H. Jun, A simple and fast algorithm for k-medoids clustering, Expert systems with applications 36 (2) (2009) 3336–3341.
[100] A. Ghorbani, J. Wexler, J.Y. Zou, B. Kim, Towards automatic concept-based explanations, Advances in Neural Information Processing Systems 32 (2019).
[101] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.
[102] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), in: International conference on machine learning, PMLR, 2018, pp. 2668–2677.
[103] T.A. John, V.N. Balasubramanian, C.V. Jawahar, Canonical saliency maps: Decoding deep face models, ArXiv abs/2105.01386 (2021).
[104] N. Ranjbar, R. Safabakhsh, Using decision tree as local interpretable model in autoencoder-based lime, ArXiv abs/2204.03321 (2022).
[105] J. Moosbauer, J. Herbinger, G. Casalicchio, M.T. Lindauer, B. Bischl, Explaining hyperparameter optimization via partial dependence plots, NeurIPS (2021).
[106] D. Sisodia, D.S. Sisodia, Quad division prototype selection-based k-nearest neighbor classifier for click fraud detection from highly skewed user click dataset, Engineering Science and Technology, an, International Journal (2021).
[107] G.K. Rajbahadur, S. Wang, Y. Kamei, A.E. Hassan, The impact of feature importance methods on the interpretation of defect classifiers, ArXiv abs/2202.02389 (2021).
[108] J. Park, J. Kim, A data-driven exploration of the race between human labor and machines in the 21<sup>st</sup> century, Commun. ACM 65 (5) (2022) 79–87, https://doi.org/10.1145/3488376, URL:https://doi.org/10.1145/3488376.
[109] Y. Liang, S. Li, C. Yan, M. Li, C. Jiang, Explaining the black-box model: A survey of local interpretation methods for deep neural networks, Neurocomputing 419 (2021) 168–182.
[110] R. Wilming, C. Budding, K.-R. Müller, S. Haufe, Scrutinizing xai using linear ground-truth data with suppressor variables, ArXiv abs/2111.07473 (2022).
[111] E. Wang, P. Khosravi, G.V. d. Broeck, Probabilistic sufficient explanations, arXiv preprint arXiv:2105.10118 (2021).
[112] J. Gao, X. Wang, Y. Wang, Y. Yan, X. Xie, Learning groupwise explanations for black-box models, in: IJCAI, 2021.

Rabia Saleem Obtained her Masters Degree from the University of Engineering and Technology, Lahore, Pakistan. Ms. Rabia currently is a doctoral candidate at the University of Derby, UK. Her research is centered on Explainable Artificial Intelligence (XAI). Her research interests include Deep neural networks (DNNs), and mathematical modelling of DNNs.

Bo Yuan received the BEng and PhD degree in computer science from the Tongji University, Shanghai, China in 2011 and 2017, respectively. He is currently a Lecturer in Computer Science with the School of Computing and Mathematical Sciences, University of Leicester, UK. His research interests include Distributed Networks, Artificial Intelligence, Internet of Things, Federated Learning, and Edge Computing. His Email is b.yuan@leicester.ac.uk.

Fatih Kurugollu obtained BSc and MSc in Computer and Control Engineering degree from Istanbul Technical University, Turkey, in 1989 and 1994, respectively. He was awarded with a PhD degree in Computer Engineering from the same university in 2000. He joined University of Derby, UK, as a Professor of Cyber Security in 2016. He has recently been appointed as a full Professor at University of Sharjah, UAE.His current research interests are centred around Security and Privacy in Internet-of-Things, Cloud Security, Imaging for Forensics and Security, Security related Multimedia Content Analysis, Big Data in Cyber Security, Homeland Security, Security Issues in Healthcare Systems, Biometrics, Image and Video Analysis.



Ashiq Anjum is currently a Professor of distributed with the University of Leicester, Leicester, U.K. He was the Director of Data Science Research Centre, University of Derby, Derby, U.K. His research interests include data-intensive distributed systems and high-performance analytics platforms for continuous processing of streaming data.

Lu Liu is a Professor and Head of School of Computing and Mathematical Sciences at the University of Leicester, UK. Prof. Liu received his Ph.D. degree from the University of Surrey and M.Sc. degree from Brunel University. Prof. Liu's research interests are in the areas of data analytics, service computing, cloud computing, Artificial Intelligence and the Internet of Things. He is a Fellow of British Computer Society (BCS).