Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-17666-y

Navigating the landscape of concept-supported XAI: Challenges, innovations, and future directions

Zahra Shams Khoozani[1] · Aznul Qalid Md Sabri[1] · Woo Chaw Seng[1] · Manjeevan Seera[2] · Kah Yee Eg[3]

Received: 30 June 2023 / Revised: 25 October 2023 / Accepted: 14 November 2023
© The Author(s) 2024

Abstract
This comprehensive review of concept-supported interpretation methods in Explainable Artificial Intelligence (XAI) navigates the multifaceted landscape. As machine learning models become more complex, there is a greater need for interpretation methods that deconstruct their decision-making processes. Traditional interpretation techniques frequently emphasise lower-level attributes, resulting in a schism between complex algorithms and human cognition. To bridge this gap, our research focuses on concept-supported XAI, a new line of research in XAI that emphasises higher-level attributes or 'concepts' that are more aligned with end-user understanding and needs. We provide a thorough examination of over twenty-five seminal works, highlighting their respective strengths and weaknesses. A comprehensive list of available concept datasets, as opposed to training datasets, is presented, along with a discussion of sufficiency metrics and the importance of robust evaluation methods. In addition, we identify six key factors that influence the efficacy of concept-supported interpretation: network architecture, network settings, training protocols, concept datasets, the presence of confounding attributes, and standardised evaluation methodology. We also investigate the robustness of these concept-supported methods, emphasising their potential to significantly advance the field by addressing issues like misgeneralization, information overload, trustworthiness, effective human-AI communication, and ethical concerns. The paper concludes with an exploration of open challenges such as the development of automatic concept discovery methods, strategies for expert-AI integration, optimising primary and concept model settings, managing confounding attributes, and designing efficient evaluation processes.

Keywords Concept-Supported XAI · Explainable AI · Neural Networks · Interpretation Methods · Evaluation Methodology · Human-Centred XAI · Human-AI Interaction · Ethical AI

Extended author information available on the last page of the article

Published online: 22 January 2024

1 Introduction

Deep Neural Networks (DNN) have performed admirably in recent decades, resulting in significant advances in Artificial Intelligence (AI) fields such as computer vision [1, 2] and natural language modelling [3]. DNNs have effectively addressed critical limitations of traditional Machine Learning (ML) models, such as the requirement for extensive expert knowledge, the difficulty of feature vector selection, and the complex process of transforming raw pixel intensities into appropriate representations [4]. DNNs, on the other hand, are inherently a black-box approach due to their recursive functionality, complex architecture, and intricate decision-making logic, which can be difficult for humans to comprehend [5–7]. The general pipeline of black box neural nets is depicted in Fig. 1 with the output provided to the end user without any explanation.

DNNs' black box nature limits their full capability and application in sensitive and high-stakes domains such as healthcare [8, 9], criminal justice [10], self-driving [11], security [12], and finance [13]. Authors in [14], discovered a limit in the generalisation of convolutional neural network (CNN) models with diverse hospital datasets in medicine. They show that, while the trained CNN model performed well in image classification tasks for specific radiology datasets, the model can be misleading for new datasets with even slightly different scanner settings. Work in [15] highlighted the sensitivity of CNN models within the training dataset in Plant pathology. They imply that annotation errors in the original input data can harm the learning process and significantly reduce the performance of the CNN model. [16] describe the CNN model's poor performance and attempts to improve it through network architecture changes, data augmentation, and natural adversarial instances. As a result, a reasonable explanation of the black box model's decision to be compresence by humans is required.

The need for interpretable ML models has been heightened by the General Data Protection Regulation (GDPR) [17, 18], which requires audits and assessments of intelligent systems and requires explainability of ML decision-making processes. GDPR regulations establish mandatory audit and assessment for intelligent systems, as well as the capability of explaining ML decision tasks. In practise, the terms IML [19–21], XAI [22–25], and self-explainable models [26–29] are frequently used to refer to models or methods that aim to explain the learning and decision approach of ML or DNN models in a way that humans can understand. The current review focuses on concept-supported interpretation methods, which are a promising approach to bridging the gap between the complex logic of DNNs and human reasoning.

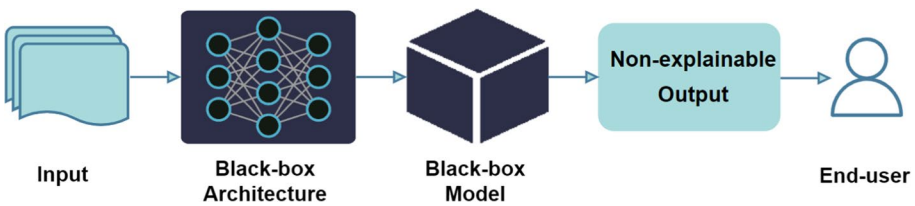

Fig. 1 General pipeline of black box neural network architecture (*source: Authors' own elaboration*)

1.1 General types of XAI methods

Current DNN models can be interpreted and comprehended from two perspectives: global and local. Global interpretation methods [30–35] place a premium on model performance and what the entire trained model learned based on a set of features, data points, or classes. These methods can be useful for tracking the logic behind overall results, comprehending training details such as weights and data structures, and revealing patterns in data to aid decision-making. However, due to the heavy computations, it is extremely difficult to interpret the entire complex DNN models using global interpretation approaches [36]. Local interpretation methods [37–41] on the other hand, attempt to explain specific input data in order to comprehend the specific attributes. They provide a thorough understanding of individual examples that are frequently overlooked in global explanations. Local interpretation methods, as opposed to global interpretation methods, have low computation and simple model implementation [41]. Saliency map [42, 43] are examples of local interpretation methods that assign an importance score to each image pixel in order to perform interpretation tasks locally.

Interpretation methods can also be classified based on the approach they use to solve a problem, which is known as the post-hoc and intrinsic method [44]. The primary and classical techniques used to explain DL models are post-hoc methods [45–48] also known as model-agnostic [49] or passive [50] interpretation methods. These methods employ an independent model to interpret existing DL models following the training and learning process. Post-hoc methods are frequently used to understand how perturbing input data affects model decision making. Post-hoc based methods include LIME [51], and SHAP [52]. In contrast, intrinsic models, also known as inherently interpretable or self-explainable models, are used during the training process to explain inner components of DL architecture [53]. These models, also known as active approaches, actively modify the network architecture to improve human-readable interpretation by utilising internal features [54]. The trade-off between model precision and explanation performance is one of the major differences between post-hoc and intrinsic approaches. Intrinsic models typically provide accurate interpretation due to their inherent approach, but they may cause model decision tasks such as prediction or classification to be reduced. Post-hoc models, on the other hand, retain model performance because they are not involved in the training process, but they are limited in their function of approximations [38]. Researchers regard post-hoc models as untrustworthy because they frequently fail to provide adequate information [55, 56]. Figures 2 and 3 show an overview of the post-hoc and intrinsic-based XAI pipelines, respectively.

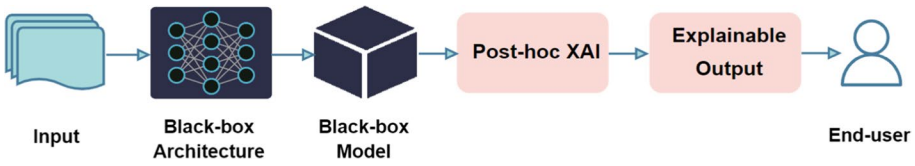

Fig. 2 Overview pipeline of *Post-hoc XAI* methods in convolutional neural network (*source: Authors' own elaboration*)

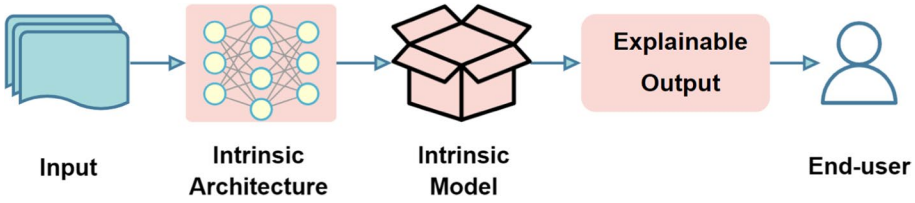

Fig. 3 Overview pipeline of *Intrinsic XAI* methods in convolutional neural network (*source: Authors' own elaboration*)

1.2 Gap between XAI's goals and traditional XAI

Explainable AI is attempting to address the issues associated with ambiguity and lack of interpretability in advanced complex machine learning models. XAI's goals are to make AI systems more transparent, understandable, and accountable. Several previous studies [44, 57–60] have well explained these objectives. Earlier attempts to provide explanation and address these goals included Saliency Map [42, 43], Grad-CAM [61] LIME [51], SHAP [52], DeepLIFT [62], and LRP [63]. They are typically based on how they map the input low level features such as pixels, weights, or vectors in order to estimate the significance of features used for the model's decision. As reported by [64], prediction tasks rely heavily on attributes. These features or pixel-based explanation methods, however, have significant limitations and have not fully achieved the key defined objectives of XAI methods. For example, [65] examined Saliency Map to determine its limits and discovered that the success ratio was very low at 60.7%. In another paper, [66] stated that existing explanation methods are not completely faithful to the primary computation functions of the black box model and cannot guarantee the true explanation and correct conclusion. Table 1 summarises some of the drawbacks of the existing feature-based XAI methods that discussed in the following.

Trustworthiness: One of the most important goals is to ensure that end-users can rely on model prediction. It refers to the dependability, credibility, and ethical soundness of an AI system's explanations. In other words, trustworthiness is a multifaceted concept that includes technical knowledge, transparency, and ethical considerations. Building and maintaining trust in AI systems is critical for their successful adoption and deployment in the real world. However, as many studies have shown, these methods are still

Table 1 Comparative analysis of XAI performance against stated goals in existing literature

XAI Goals	Reference	Traditional XAI Performance	Reference
Trustworthiness	[23, 60, 75]	Poor	[55, 67–69, 76]
Informativeness	[23, 60]	Poor	[70, 71]
Human-AI interaction	[57, 71, 77]	Poor	[69, 71]
Ethical AI	[23, 57]	Poor	[48, 73, 74]
End-user satisfaction	[44, 75]	Poor	[71, 78]
Understandability	[60, 75]	Poor	[76, 78]
Faithfulness	[60]	Poor	[58, 66]

ineffective at gaining end-user trust [55, 67, 68]. For example, [69] tested explanation methods with humans and discovered that they are very sensitive to human decision biases and do not improve human perception of trust in the model.

Informativeness: This goal refers to providing meaningful and relevant information about decisions or predictions in terms of both quality and quantity. Traditional methods, on the other hand, do not help to provide contextual information that has a significant impact on the model's decision. As an example, Rudin [70] examined the model using the Saliency Map method and discovered that the explanation is insufficient to provide adequate information and details to comprehend the black box learning mechanism. According to the author, the performance of these methods can be the same for multiple classes (either correct or incorrect) and uninformative in helping users understand why there is a misclassification in the model. Recently, authors in [71] comprehensively evaluated these methods using Saliency Map and Grad-CAM, as well as human experimentation, to understand end-user feedback. They notice that, while these methods provide users with some understanding of black box decisions, some participants dislike them and argue that the explanations are uninformative and rough.

Human-AI interaction: The core component targeted by explainable AI models is effective human-AI interaction, also known as collaboration [71]. Finally, we would like to gain trust and be understood by end users, whether they are experts, stakeholders, decision makers, or non-experts. This goal is related to fields where end users are extremely important, and their ability to interact with models is what ensures success. We must foster trust by ensuring that AI is consistent with human values and expertise. Furthermore, it is critical to incorporate user feedback and leverage human expertise in decision-making to improve model performance. Traditional XAI approaches, on the other hand, perform based on lower-level attributes such as pixel intensities, which do not correspond with human rational thinking, and they have significant limitations in communicating with domain experts [69].

Ethical AI: Ethical consideration, also known as fairness [23], refers to an AI system that makes decisions that are unbiased. Its goal is to address biases in AI decision-making and promote fairness, as well as to support ethical standards in AI development and deployment. Fair treatment of different groups and the avoidance of discriminatory outcomes contribute to the system's overall trustworthiness. Biassed results from the distribution of input data [72] or from misleading explanations [48]. However, a recent intensive evaluation on fairness performed by [48] revealed the limitations of these methods and reported that, while fairness is one of the key pillars in XAI, little research has been conducted in this area. In another effort, [73, 74] pointed out that methods that cannot provide a high importance score may result in unfair explanations.

1.3 Transitioning to human-centred XAI

Early AI applications were designed to mimic human cognition tasks [79]. Explainable AI was later developed to help end-users comprehend complex AI models [23, 60, 75]. However, as described in Sect. 1.2, existing feature-based XAI methods performed poorly in terms of providing explanations that end users could understand and satisfy their needs. According to a growing body of research, existing XAI systems are frequently developed without a thorough understanding of the end-users' requirements and characteristics [71,

80, 81]. Numerous XAI techniques, for example, developed to address explainability requirements during model development may encounter difficulties when applied to end-users with diverse needs [71, 82]. In other words, a human as an end-user willing to assess the success of XAI methods based on their own reasoning [76]. End-users frequently want to know 'why' a specific type of input made that prediction [23, 71, 83]. Existing feature-based XAI, on the other hand, focuses on 'what' image attributes or regions are more important to model decision function [71]. As a result, it is critical to use human-centred approaches in understanding the rational aspect of AI explanation users. Human reasoning can simply explain what an object such as a 'apple' is by listing its higher-level properties such as 'Colour': red, 'Shape': round, whereas DNN, despite its significant performance, cannot explain its learned knowledge in a human-comprehensible manner [23].

As a result, current feature-based explanation methods are still incapable of fully explaining the reason for model decision at a human-understandable level. They also do not correspond to high-level features such as concepts, which correspond to human thinking and reasoning. As a result, there is a need to provide human-readable interpretation methods that end-users, whether experts or decision-makers, can understand. In this way, we can ensure that we provide a trustworthy explanation that adheres to the primary model prediction function. Figure 4 depicts the general workflow of concept-supported XAI.

1.4 Contribution

Several surveys on feature-based explaining ML and DL models have been conducted [9, 41, 45, 49, 53, 84, 85]. In contrast, the work in [69] recently proposed the first attempt at concept-supported interpretation methods for computer vision-related tasks with a review paper [86] that focused on this area. Presently, a significant interest in these methods is shown by many researchers and leading publishers like NeurIPS, ICLR, ICML, AAAI, and IEEE, underlining the necessity to explore the recent intriguing advancements in this field. Therefore, this work aims to review concept-supported interpretation methods which align more with human understanding and reasoning, to furnish a solid grasp of the existing works as well as to spotlight the open questions and unresolved issues within this domain. To deepen our understanding and pinpoint the gaps in concept-supported interpretation methodologies, we have formulated the following research inquiries:

- Are there any strategies for incorporating human reasoning, denoted as 'Concept,' into Convolutional Neural Network (CNN) models?

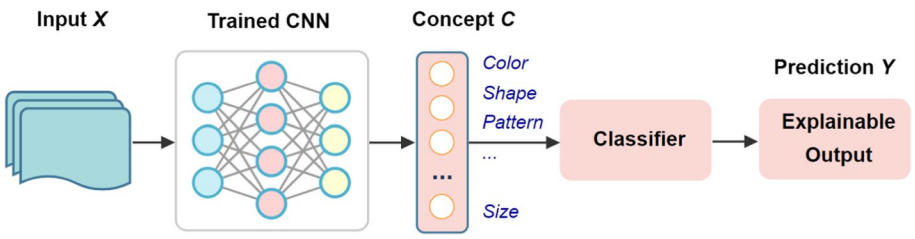

Fig. 4 General workflow of Concept-supported XAI methods. It takes an input image X, use trained CNN to predict pre-defined concepts C, and then use concepts C to provide the final prediction output Y. (*source: Authors' own elaboration*)

- Which visual 'Concept' attribute datasets are currently available for concept learning tasks?
- How can sufficiency metrics for concept-supported interpretation methodologies be defined and quantified?
- What factors have a significant impact on the efficiency of concept-supported interpretations?
- What are the main challenges and future research opportunities in the field of concept-supported interpretation?

1.5 Paper organisation

Figure 5 depicts a visual summary of the key points discussed in the literature. Section 2 offers a detailed examination of current methodologies, outlining their respective pros and cons, alongside a thorough comparison to highlight the strengths and weaknesses of existing methods. Section 3 delivers an extensive overview of available concept datasets, describing their unique features. Section 4 delves into sufficiency metrics, enriched with pertinent examples for better understanding. Section 5 explores the six crucial factors affecting the efficacy of concept-supported interpretation, along with the associated challenges and potential areas for future research. The importance and robustness of concept-supported interpretation methods are discussed in Sect. 6. Section 7 highlights several open research areas with the potential to drive various technological advancements. A summary of key findings is presented in Sect. 8. In the concluding sections of the paper, we encapsulate our insights and suggest areas for future research endeavours. A list of acronyms for methods and relevant contexts used in the literature provided in Table 2.

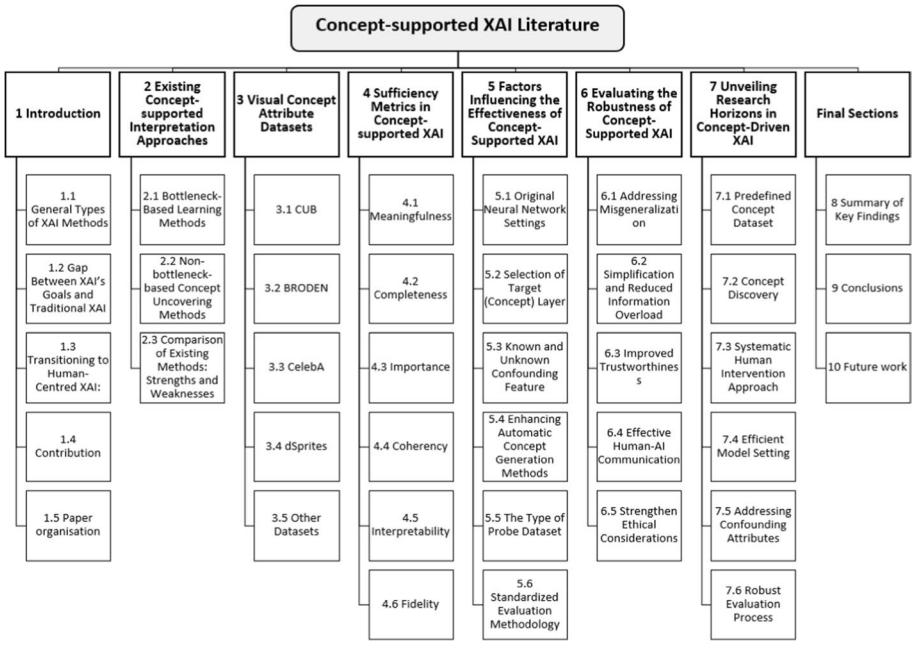

Fig. 5 Scope and organization of this comprehensive literature (*source: Authors' own elaboration*)

Table 2 List of acronyms used in the literature

Abbreviation	Definition
AAAI	Association for the Advancement of Artificial Intelligence
ACDTE	Automated Concept-based Decision Tree Explanations
ACE	Automatic Concept-based Explanations
AI	Artificial Intelligence
AUC	Additional Unsupervised Concepts
BBSD	black-box shift detection
BDD	Berkeley DeepDrive
BN	BottleNeck
BRODEN	Broadly and Densely Labelled Dataset
CAR	Concept Activation Region
CAV	Concept Activation Vectors
CB	Concept Bottleneck
CBM	Concept Bottleneck Model
CBNM	Concept Bottleneck-based Models
CBSD	Concept Bottleneck Shift Detection
CLIP	Contrastive Language-Image Pre-Training
CME	Concept-based Model Extraction
CNN	Convolutional Neural Network
COCO	Common Objects in Context
CNN2DT	Convolutional Neural Network to Decision Tree
CSS	Concept Shift Score
CUB	Caltech-UCSD Birds-200–2011
CVF	Computer Vision Foundation
DeepLIFT	Deep Learning Important FeaTures
DISSECT	Disentangled Simultaneous Explanations Via Concept Traversals
DL	Deep Leaning
DNN	Deep Neural Networks
DR	dimensionality reduction
DT	Decision Tree
DTD	Describable Textures Dataset
ECCV	European conference on computer vision
ECML	European Conference on Machine Learning
EDBT	Extending Database Technology
FCN	Fully Convolutional Network
GDPR	General Data Protection Regulation
Grad-CAM	Gradient-weighted Class Activation Mapping
ICE	Invertible Concept-based Explanations
ICLR	International Conference on Learning Representations
ICML	International Conference on Machine Learning
IEEE	Institute of Electrical and Electronics Engineers
IML	Interpretable Machine Learning
IN	Intrinsic
ISIC	International Skin Imaging Collaboration
LIME	Local Interpretable Model-Agnostic Explanations

Table 2 (continued)

Abbreviation	Definition
LRP	Layer-Wise Relevance Propagation
MACE	Model Agnostic Concept Extractor,
MCD	Multi-dimensional Concept Discovery
ML	Machine Learning
MLP	Multi-Layer Perceptron
MMD	Maximum Mean Discrepancy
NCAV	Non-negative Concepts Activation Vectors
NMF	Non-Negative Factorization Matrix
NN	Neural Network
OAI	Osteoarthritis Initiative
OIA	Object Induced Actions
PACE	Post-Hoc Architecture-Agnostic Concept Extractor
PASCAL	Pattern Analysis, Statistical Modelling and Computational Learning
PCA	Principal Component Analysis
PCBM	Post-hoc Concept Bottleneck models
PH	Post-Hoc
RCNN	Region-based Convolutional Neural Network
RGB	Red, Green and Blue
ROAR	RemOve And Retrain
SBN	Semantic Bottleneck Networks
SEAL	SEgmentation-Aware Loss
SENN	Self-Explaining Neural Network
SHAP	SHapley Additive exPlanations
SIIM	Society for Imaging Informatics in Medicine
SLIC	Simple Linear Iterative Clustering
SMS	Semi-supervised
SRP	Sparse Random Projection
SSN	Superpixel Sampling Networks
SVM	Support Vector Machine
TCAR	Testing with Concept Activation Region
TCAV	Testing with Concept Activation Vectors
UNS	Unsupervised
VAE	Variational AutoEncoder
XAI	Explainable Artificial Intelligence

2 Overview of existing concept-supported interpretation approaches

2.1 Bottleneck-based learning methods

Concept Bottleneck-based Models (CBNMs) have received a lot of attention in recent years as a way to develop interpretable Deep Neural Networks (DNNs) in the field of computer vision [87–93]. CBNMs' primary goal is to bridge the gap between raw input data (e.g., pixels) and high-level attributes, also known as 'concepts.' These concepts are then applied

to specific tasks. The underlying assumption of these methods is the hypothesis that interpretability can be achieved by incorporating a function that maps human-understandable concepts to input features, which then predict output labels. There are numerous variations of CBNMs that have recently been developed that merit a thorough examination. These variations are discussed in the following sections, and a summary is provided in Table 3.

2.1.1 Concept Bottleneck Models (CBM)

Authors in [87] made an early attempt to develop a supervised CBNM by embedding human-interpretive labels such as 'joint space narrowing' or 'undertail colour' into a supervised ResNet model. The goal of [87] was to address the shortcomings of the preliminary versions of CBM [94, 95] which demonstrated a trade-off between model accuracy and interpretability. The fundamental concept of the bottleneck (BN) architecture was first proposed by [1] as a strategy for shrinking the dimensions of the input and output layers, resulting in an efficient and pragmatic model. Work in [96] aptly demonstrated the use of the bottleneck-based architecture for language recognition tasks, as shown in Fig. 6. The BN layer is strategically positioned at the centre of the DNN in this architecture, serving as the point where concepts are integrated into the training process.

The Concept Bottleneck Model (CBM) proposed in [87] is based on the Bottleneck (BN) design and accepts input features such as pixels, denoted as (x). The intermediate layer of the DNN is then resized to match the number of labelled concepts, embedding the concept (c), which corresponds to the activation layer during training. Finally, the concept (c) is employed in order to forecast the output (y). During the training procedure, the CBM can be represented as $x \rightarrow c \rightarrow y$ while during the testing procedure, it takes input (x), predicts concept$\hat{c} = g(x)$, and then predicts $\hat{y} = f(g(x))$ from BN $\hat{c}$. [87] proposed three approaches to learning CBM$(\hat{f}, \hat{g})$: In *jointly* learning, the input is first trained to concept$c \rightarrow y$, and then the concept is trained to output$\hat{c} \rightarrow y$; in *independent* learning, the true (c) or ground-truth concepts are used to map concept to output $(c \rightarrow y)$; and in *sequential* learning, the predicted concepts $(\hat{c})$ are used to map concept to output $(\hat{c} \rightarrow y)$.

Authors in [87] attempted to meet three concept interpretation criteria in their work: interpretability (identifying the critical concepts for output y), predictability (predicting the output solely from concept (c)), and intervenability (altering the predicted concept with true concepts c or ground-truth concepts). However, subsequent research by [97] revealed that both independent and collaborative CBM learning approaches fail to meet these three criteria. They used a saliency map as a post-hoc explanation method that does not interfere with training and attempts to visualise neural network output and understand CBM performance. They investigate whether defined concepts used during training correctly correlate with input data. A model trained to recognise the concept 'wing pattern', for example, should logically focus on the wing patterns, but instead on the entire bird. Based on [97], as the joint BN model learns the output before the concept, independent BN learning may be the specific approach capable of meeting the three interpretation criteria.

CBM demonstrated impressive performance and provides numerous benefits, such as assisting in identifying incorrect predictions due to incorrect concept predictions by tracking weights in the model [88]. It supports interventions on concepts, which allow the user to modify the prediction functions by instantly changing the value of the concept [87]. They are more at ease with expert rational thinking because they used concept knowledge during the decision-making process [98]. For unseen datasets, it can mitigate dataset and covariate shift issues that are common in DNN models [87]. It can also provide causal

Table 3 Comprehensive summary of Supervised (S) and Semi-supervised (SMS) concept-supported interpretation methods. Demonstrate XAI category either Post-hoc (PH) or Intrinsic (IN), and Global (G) or Local (L) interpretation methods

Concept-Supported Method	Ref	Publisher	CNN Architecture	Learning Method	Training Dataset	Concept Dataset	Concept Usage	Concept Scope	Importance Scoring Method
PCBM	[88]	ICLR	CLIP-ResNet50, ResNet18, Inception	S	ImageNet	CUB, BRODEN, COCO-Stuff, Derm7pt	PH	L & G	N/A
Concept Transformer	[140]	ICLR	ResNet50	S	ImageNet, MNIST Even/Odd	CUB, aPY, MNIST Even/Odd	IN	L & G	attention scores
CBSD	[92]	ICLR	BB network	S	N/A	dSprites, 3D-Shapes	IN	G	N/A
Debiased CBM	[93]	ICLR	ResNet152	S	TorchVision	CUB	IN	L	Human defined annotation
CAV	[69]	ICML	GoogleNet, Inception V3	S	ImageNet	ImageNet abstract concept	PH	G	TCAV
CBM	[87]	ICML	ResNet-18	S	ImageNet	CUB, OAI	IN	L	N/A
CaCE	[141]	arXiv	ResNet-50, ResNet-100	S	Places365	CelebA, MNIST, COCO-Miniplaces	PH	G	VAE-CaCE
IBD	[83]	ECCV	AlexNet, VGG16, Resnet18, Resnet50	S	ImageNet Places365	BRODEN	PH & IN	G	TCAV
Hierarchical CBM	[90]	Elsevier	Resnet50	S	ImageNet	Fridges images	IN	G	N/A
CBM-AUC	[89]	IEEE	Inception-v3, RCNN network	SMS	ImageNet, COCO	CUB, BDD-OIA	IN	L & G	N/A
Concept Extract	[142]	IEEE	ResNet-101, VGG-16, FCN, DeepLabV3	SMS	ImageNet	ImageNet 10 classes	PH	L	TCAV
CME	[98]	arXiv	Inception-v3	SMS	ImageNet	CUB, dSprites	PH	G	coefficient magnitudes
Net2Vec	[110]	CVPR	AlexNet, VGG16, GoogLeNet	SMS	ImageNet, Places365	BRODEN	IN	G	IoU score

Table 3 (continued)

Concept-Supported Method	Ref	Publisher	CNN Architecture	Learning Method	Training Dataset	Concept Dataset	Concept Usage	Concept Scope	Importance Scoring Method
NetDissect	[143]	CVPR	AlexNet, VGG-16, GoogLeNet, ResNet-152	SMS	ImageNet, Places205, Places365,	BRODEN	PH	L & G	IoU score

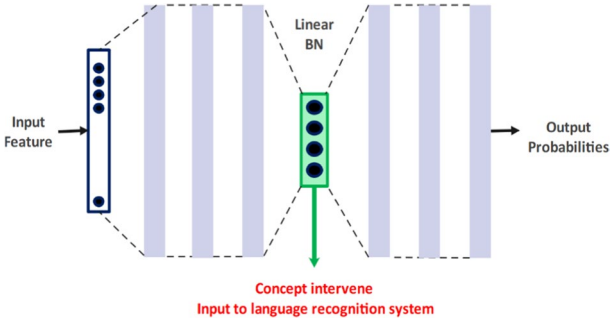

Fig. 6 Illustration of bottleneck-based architecture proposed by [96]

relationships between $x \rightarrow c \rightarrow y$ and they are adaptable and do not require c to cause y. As a result, it is useful for models where determining the causal $c \rightarrow y$ graph is difficult [87]. If large concept labels are available, a smaller training dataset may be required [87]. However, variant studies such as [98] pointed out that CBM models do not provide adequate information about which concept(s) are good enough and important to solve an assigned problem. The authors of [87, 88, 98] confirmed that they require a very large prepared concept annotation dataset to be used during the learning process to train the BN, whereas [89] revealed that proposing a high performance model with a small class of concepts is extremely difficult. The number of concept labels is limited by the DNN layer dimension. The authors of [88] observed that these types of models must complete all training tasks for all training datasets using concept-label datasets, which is a significant limitation. The work in [89] demonstrated that CBM is limited to supervised concepts and does not support unsupervised concepts, whereas the authors of [88] argued that they are incapable of meeting the precision of unrestricted NN models, reducing the motivation to deploy in real-world applications. Furthermore, it is unclear how to improve the model when appropriate concepts are unavailable. Researchers in [97] suggested that expecting access to concepts alone to fully capture the relationships between input data points and output labels is unrealistic.

2.1.2 CBM-AUC—Concept Bottleneck Model with Additional Unsupervised Concepts

The work in [89] presented a novel methodology for overcoming the performance limitations of CBM [87] when only a small set of concept-labeled datasets is available. The dimensions of the DNN's intermediate layer (the bottleneck), as shown in Fig. 6 must be reduced to the number of concept labels. When the number of concept labels is limited, this frequently results in suboptimal performance. The authors of [89] proposed a high-performance model using a combination of Self-Explaining Neural Network (SENN) as an auxiliary unsupervised concept and supervised CBM, as shown in Fig. 7. SENN is an unsupervised learning approach that uses a linear model to learn concepts automatically [29]. The target task of SENN is computed as Eq. 1, where $\theta(x)$ is the calculation of weights for each concept label and $c^{im}(.)$ is the encoder layer for the unsupervised concept.

$$f(x) = \theta(x)^T c^{im}(x) \tag{1}$$

To understand how factors affect concept model performance, Yoshihide and Keigo [99] used the Inception-v3 network pretrained with ImageNet and the Faster RCNN network [100] pretrained with the COCO [101] dataset. They tested both network architectures by

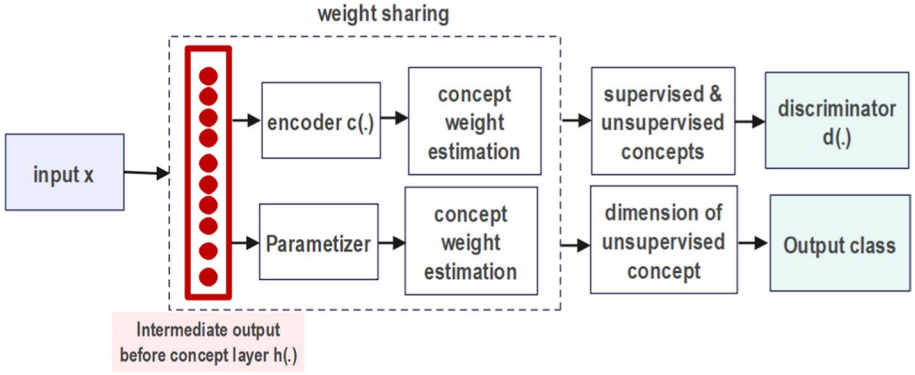

Fig. 7 CBM-AUC network overview proposed by [89]

replacing the intermediate layer of Faster RCNN with Inception-v3 and using the CUB [102] and BDD-OIA [103] datasets for concept learning. The experiments revealed that Inception-related networks are not a good choice for the BDD-OIA dataset, and that Faster RCNN performs better. They also compared a proposed concept-based model to multi-task models under three conditions: global feature [103], local feature [104] and global + local features [103]. Experiments revealed that the proposed CBM-AUC performs better for concept learning. The multi-task model with local features, on the other hand, outperforms the target task. Furthermore, by comparing the results, they discovered a low correlation between supervised and unsupervised concepts. As a result, they believe that unsupervised learning can aid in providing consistent correlations. Furthermore, they discovered a condition in which the concept output appears to be correct but does not provide a specific concept used during training.

As a result, the authors of [89] identified three key findings from the experimental results that merit further investigation: 1) The network architecture and number of intermediate layers used have a significant impact on the performance of concept-supported models. 2) Using a CBM completely unsupervised can improve model performance. 3) With a small concept annotation dataset, it is difficult to generate accurate output concepts, necessitating the creation of additional concept labels.

2.1.3 PCBM—Post-hoc Concept Bottleneck Models

The gripping post-hoc CBM proposed by [88] addressed three key limitations of the CBM built earlier by [87]. First, CBMs require concept classes during the training phase, which means that the training dataset must map to available concept classes. This is impractical for real-world problems because existing training datasets rarely have concept labels. Second, in the absence of appropriate concepts, it is unknown how to improve the model while maintaining the original network efficiency and model interpretability interest. Third, CBM allows only local intervention for a single input data set to improve model performance. However, it is unclear how to improve the overall model performance with human associations. Authors in [88] proposed a multimodal CBM to create concept representations using a text encoder and natural language form of concept descriptors in order to eliminate the concept annotation procedure. Furthermore, when efficient concept labels are unavailable and interpretation

performance is poor, they embed a residual module to restore the initial DNN model performance, which is referred to as the PCBM-h hybrid model. They also added global intervention capability to the proposed model in order to improve it through human evaluation and changing the model prediction score.

The PCBM [88] approach starts with learning concepts using CAV (Concept Activation Vectors) [69] to link input features and activation layer into the set of concepts. In this case, the concept dictionary can be created manually, as in the ConceptSHAP method [105], or automatically, as in the ACE method [106]. The linear SVM classifier was then trained with 50 positive and negative concept samples for each concept, as described in [69]. They then modified a multimodal CLIP (Contrastive Language-Image Pre-Training) model proposed by [107] to train concept vectors using both image and text encoders. They used ConceptNet [108] to generate relevant concepts for each class as subclasses with a few types of relations such as hasA (for example, Cat hasA 'sharp claws'). Following that, the interpretation component connects concept subclasses to the prediction, which can be described as $N_c \rightarrow y$.

The authors performed a thorough evaluation of the Post-hoc Concept Bottleneck Model (PCBM) using a variety of network architectures and datasets, demonstrating the robustness of the proposed methods. The authors used the Multimodal CLIP-ResNet50 network [107] trained on the CIFAR10 and CIFAR100 datasets [109] in the first experiment. The Concept Bottleneck (CB) model was trained for concept learning using 170 concept classes derived from the BRODEN visual dataset [110], as used in [111]. Following that, a similar network and concept annotation dataset were used in the second experiment, but with the addition of the COCO-Stuff dataset [112], which includes 20 classes of objects [113]. In the third experiment, the authors used the same methodology as in [87] to train the ResNet18 network [1]with 112 concept labels from the CUB concept dataset [102]. The fourth experiment used an existing Inception network trained on the HAM10000 dataset [114], as proposed by [115]. The HAM10000 dataset contains dermoscopic images of skin lesions that are used to determine whether they are benign or malignant. Eight concepts from the Derm7pt dataset [116] were used for malignancy identification in the concept learning task. Finally, in the fifth experiment, the previously mentioned Inception model trained with HAM10000 was tested under real-world conditions for Melanoma Classification using the SIIM-ISIC dataset [117].

With the exception of the CLIP-ResNet50 architecture trained with CIFAR100, the results from these five disparate settings encompassing various networks and datasets demonstrated the exceptional performance of the proposed PCBM. This accomplishment attests to PCBM's successful proposition of an interpretative methodology for black-box models while retaining the original model's performance. The findings also highlighted the importance of the concept learning dataset as a critical component in concept-based models, revealing that an inappropriate concept dataset can introduce biases [118]. Nonetheless, this study leaves some unanswered questions, particularly regarding the embedding of human-defined concept bottleneck models for large training datasets like ImageNet. The generation of concept subclasses using an unsupervised approach is still an open challenge that merits further investigation in order to improve the capabilities of concept bottleneck models. Furthermore, while the authors mention the possibility of incorporating human input in multimodal models, the methodology for effectively utilising expert commentary to improve concept bottleneck models is unknown.

2.1.4 Hierarchical concept bottleneck models

One of the limitations of Concept Bottleneck (CB) models, as previously stated, is the requirement for critical changes to the network architecture in order to reduce the dimensions of the Bottleneck (BN) layer to match the number of concept labels. This requirement may have an adverse effect on model precision by requiring a large number of concept annotations to improve model precision or by causing the omission of some input information. In light of this, authors in [90] attempted to build a Concept Bottleneck model that strikes a balanced correlation between model performance, the effort involved in data annotation, and the explanation task. The authors aimed to use a Concept Bottleneck-based model to improve multi-label and fine-grained image classification, with the goal of reducing mismatches. Furthermore, based on the mean Average Precision (mAP) formula, this study introduced new evaluation metrics specifically tailored for supervised Concept Bottleneck explanation models. Finally, the Concept Bottleneck model's attributes were used to facilitate an object tracking approach via the explanation task.

The first step in creating the proposed Hierarchical Concept Bottleneck model was to perform a multi-label classification task with hierarchically defined concepts that included both lower-level concepts (e.g., fruit, colour, shape) and higher-level concepts (e.g., apple). This hierarchical approach is associated with improved predictions and is capable of capturing valuable information, particularly when confronted with novel data distributions. For semantic segmentation and classification, a mask R-CNN network [119] is used. Lower-level concepts are used as input for the concept BN [87], which provides higher-level concepts. The basis loss function has been modified in this work as Eq. 2 to be applicable in concept explanation models, where K and N denote the number of training examples and high-level concept classes, respectively, and C^i denotes the number of subclasses in each concept class. The different objects in the image and colour attributes are extracted as a lower-level input concept for the MLP (Multi-Layer Perceptron) classifier to provide a higher-level concept or logical category during fine classification. Then, for each pair of higher level concepts i and j, a new offered concept evaluation metric mAP is modified as Eq. 3 to calculate the median of normalised Euclidean distance i and j. Here, $O^k_{i,f}$ denotes the one-time encoding of the lower-level concepts for item k in class i, and dimension F. Finally, the model feeds the Concept Bottleneck the lower-level attributes as well as colour information from specific regions in a pair of images. The probability associated with each pair of images is then calculated in order to determine the similarities between the objects they contain.

$$L_{class} = \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{1}{K} \left(-\log \frac{\exp\left(x^i_{k,y^i_k}\right)}{\sum_{c=1}^{C^i} \exp\left(x^i_{k,c}\right)} \right) \tag{2}$$

$$d_{i,j} = median_{k \in K} \left(\frac{1}{F} \sqrt{\sum_{f=1}^{F} \left(O^k_{i,f} - O^k_{j,f} \right)^2} \right) \tag{3}$$

The researchers chose the Resnet50 network [1], which was pre-trained on the ImageNet dataset, as the supervised deep learning model in the study, along with the concept BN model proposed by [87]. The authors used a proprietary smart fridges dataset for

the Hierarchical Concept annotation dataset, which contains images with various items, angles of view, and conditions. The annotations are divided into two categories: higher-level attributes and lower-level attributes. The dataset was created with five categories in mind for the lower-level concept attributes: 'Logical groups' (e.g. fruit, vegetable, Dairy), 'Consistency' (e.g. hard, soft, liquid), 'Shapes' (e.g. round, oval, cylinder), '3D Shapes' (e.g. cup, flat, tube), and 'Colour histogram' in L*a*b colour space. The fridge dataset was annotated based on human-understandable categories called 'Logical categories' such as apple, fish, meet, and so on for higher-level concept annotations.

To better understand the factors influencing the concept-supported interpretation task, the proposed Hierarchical Concept Bottleneck model was tested in two different scenarios. The lower-level concept attributes were used as input for the Concept Bottleneck model generated by the Mask R-CNN model in the first scenario. In the second scenario, the Concept Bottleneck model made use of ground-truth data. The authors discovered that when fed ground-truth annotations, the Concept Bottleneck model performed significantly better, whereas using features from the Mask R-CNN model resulted in lower performance due to misclassification of hierarchical labels. This comparison demonstrated the importance of accurate feature identification methods in concept-supported interpretation tasks. The research found that the Concept Bottleneck model could compete with Mask R-CNN models in classification tasks while avoiding confusion and generalising for unseen samples from similar data distributions. The performance of the Hierarchical Concept Bottleneck model for completely unknown datasets with new data distributions, on the other hand, remains unknown.

2.1.5 CBSD—Concept Bottleneck Shift Detection

The issue of dataset shift is one of the most significant barriers to achieving the best performance from DL models [120–122]. It primarily refers to unexpected results in a new dataset that are not seen by the model during the learning process and have different data distributions. This limitation can be problematic for real-world prediction models [14, 123]. To address this limitation, [92] proposed the CBSD (Concept Bottleneck Shift Detection) model, which uses concept-supported interpretation to identify and analyse concept attributes influenced by data shift in vision tasks. In comparison to common shift detection methods such as BBSD (black-box shift detection) [124, 125], the proposed idea is significantly useful for detecting the major cause of shift in a new dataset and improving the system, particularly in a human-readable format. Furthermore, a new statistical concept-based assessment metric called CSS (concept shift score) was introduced to recognise whether the model is suffering from shift problem, as shown in Eq. 4, to compute CSS for concept (t) of the i th concept (higher CSS specify higher shifts).

$$CSS(t, i) = \frac{t_i}{\sum_{l=1}^{k} t_l} \tag{4}$$

CBSD [92] is based on the decomposition idea of core CBM [87] and [98] to reduce the dimension of the intermediate layer. In which a pretrained network classifier decomposes input data to the concept $g : X \rightarrow C$ and concept attributes to the target $h : C \rightarrow Y$. A trained network architecture with three convolutional layers, two connected layers using RELUs, and a Softmax generated layer was deployed for the BBSD and DR components in both decomposed tasks. A sequential multi-task concept BN model was used for the CBSD component, with BBSD architecture used for the input-to-concept task and

logistic regression used for the output-to-concept task. In this study, dSprites [126, 127] and 3D-Shapes [128, 129] were used as concept datasets, with concepts such as 'Shape', 'Scale', and 'Rotation' used to detect a shift degree of CBSD in each concept. The input data was used to reduce the representation dimension, as shown in Fig. 8. The statistical scores are then used to detect shifts. The concept explanation concludes with the use of introduced CSS to demonstrate the degree of shift in each learned concept (i.e., among all concepts, 'Shape' achieved the highest CSS, indicating that it is highly affected by the shift problem and should be considered to improve the proposed DL model).

The authors of [92] used common techniques to generate artificial shifts such as image, knockout, Gaussian, and adversarial shifts. The dimensionality reduction component of the CBSD has been tested and compared to other DR techniques such as PCA (principal component analysis), SRP (sparse random projection), BBSDs that use softmax results, and BBSDh that use hard-thresholded results. The authors used MMD (maximum mean discrepancy) [130], KS (Kolmogorov–Smirnov) test and Bonferroni correction [131] to evaluate multi-dimensional representations such as proposed CBSDs, BBSDs, PCA, and SRP, and the Chi-squared test to assess CBSDh and BBSDh. The experiment demonstrated that CBSD outperformed common BBSD in detecting shift issue triggers. Furthermore, CBSD successfully detected shifts where BBSD, PCA, and SRP methods failed completely. However, it is unclear how expert comments can be incorporated into the CBSD model to modify concept learning and prediction functions.

2.1.6 Debiased concept bottleneck model

One of the major threats to DL model accuracy is confounding attributes. According to various studies, confounding attributes can cause a variety of problems, including creating unreliable training datasets [21], being a major cause of failed model generalisation tasks [14], being a significant degrade factor in Causal-based models [132, 133], and providing falsely accurate models [134]. Authors in [93] proposed a novel generative debiased concept BN Model to eliminate confounding attributes and biases and improve interpretation model to address the problem of confounding attributes. The authors' goal in this study is to find a correlation between concepts and confounding attributes in input data points using a novel causal prior graph shown in Eq. 5, where d denotes unconfounded concepts, u denotes a confounding attribute between concept c and input data x, and f_1,f_2,f_3, denotes identifying functions. Proposed debiased CBM can then be presented as $y \rightarrow d \rightarrow x$ which begins with label-to-concept and then moves to concept-to-input.

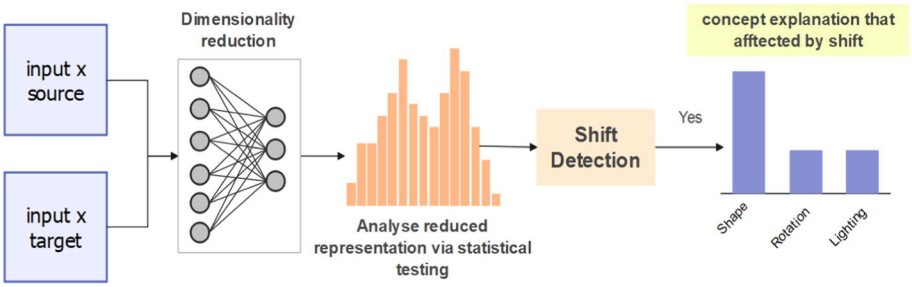

Fig. 8 Diagram pipeline of CBSD model proposed by [92]

$$\begin{aligned} d &= f_1(y) + \in_1, \\ c &= d + h(u), \\ x &= f_2(u, d) + f_3(y) + \in_2, \end{aligned} \tag{5}$$

The proposed generative model's hierarchy progresses from nodes with more general information, such as labels, to nodes with more specific information, such as pixel values. The ideal graph begins with target label y (first node), then proceeds to concept c (second node) and finally to features x (final node). In another setting, more realistic graph defined where both feature x and concept c are influenced by unknown confounding attributes$u$. Another graph setting for model completeness was provided, with the confounding connection of $u \rightarrow c$ removed. The Two-Stage Regression used in this work for CBM was inspired by [135] where the first stage regress input feature x in terms of Causal variable z to generate$\hat{x}$,, and the second stage regress label y in terms of $\hat{x}$ (i.e.$z \rightarrow x \rightarrow y$).

Authors in [93] used CBM proposed by [87, 91] and the TCAV (Testing with Concept Activation Vectors) method [69] to create the concept BN explanation model. For the task at hand, they used the ResNet152 network [1]. To evaluate concept explanation performance, the CUB dataset [102] with four different certainty scores (e.g., not visible, guessing, probably, and definitely) and average concept annotation scores was used.

The authors used the ROAR (RemOve And Retrain) framework [136], to compare the regular model to the debiased one to demonstrate the accuracy of the proposed debiased Concept Bottleneck Model (CBM) interpretation. The results of the experiment demonstrated robustness performance for both synthetic datasets (CUB) and real-world datasets. The concept is strongly associated with beneficial benefits such as improving high-level explanations, leveraging prediction task via human-understandable variables, increasing model generalisation task, and removing unwanted context during explanation task. However, how to fully achieve performance satisfaction in terms of completeness metric [105] of explanation models via debiased concept method remains a mystery. As a result, adding another two-stage regression method in future research may help to overcome this limitation and provide a more accurate model.

2.1.7 SBN—Semantic Bottleneck Networks

In semantic and meaningful segmentation applications, the lack of human-understandable methods is clearly visible. We are primarily looking for human understandable areas in visual materials in the semantic segmentation task. However, the common DNN architecture learns from data and produces representations as an output without involving human-readable features in the loop. As a result, the concept BN interpretation approach can be extremely beneficial in semantic and meaningful segmentation. Authors in [91] proposed the SBN (Semantic Bottleneck Networks) model to use human reasoning as a higher-level feature and improve the semantic segmentation task. The authors reduced the lower-level feature dimensions from thousands to ten meaningful concepts by employing the concept of BN intermediate layer. The method is not only useful for analysing the source of errors, but it also contributes to a high-level testing approach through direct manipulation of the BN layer. The representations in the backbone network architecture's intermediate layer (shown as the first traditional architecture) mapped to the added BN layer, and then the network was finetuned with new changes (shown as the second traditional architecture). The confidence prediction score was used as an evaluation criterion in this study, which takes pixels from all test samples and predicts confidence by computing the difference value between the top two largest softmax activation layers.

PSPNet [137] was used as a network architecture, which was trained with the Cityscapes dataset [138]. Annotations in Cityscapes are coarse and fine-grained. For concept BN learning, the Broden + [139] dataset was used. The results of the experiments in this study demonstrated the obvious benefit of embedding concept BN into DNN architecture, which can surprisingly help to obtain meaningful partitioning in visual related tasks. However, the overall accuracy of concept-based output remains far from ideal. According to [91], only 75% of all samples could provide accurate results, which is insufficient for real-world case studies. As a result, the open research question in this study is how to improve SBN-based model precision by embedding multiple BN layers in different representation layers.

2.2 Non-bottleneck-based concept uncovering methods

Most CBNM interpretation methods, as discussed in Sect. 2.1, required predefined concept annotation datasets [87–93]. However, because concept labelling frequently requires experts for manual labelling and is impractical for real-world cases due to the high cost of creation [89, 98], the question of how to discover human-understandable concepts automatically and develop interpretation models based on these discovered concepts arises. This section examines concept-supported interpretation methods that address this question, as shown in Table 4. The terms 'concept extraction' and 'concept discovery' are frequently used in this field of study to refer to the unsupervised automatic discovery of concepts from datasets.

The authors in [106] proposed an earlier work in concept uncovering models called ACE (Automatic Concept-based Explanations). With three major conceptual validation properties, it provides a method for extracting visual meaningful units that are important for neural networks and understandable to humans. It used SLIC [144] to locate the desired region of interest in order to satisfy the 'Concept Meaningfulness' property, K-means Clustering [145], and Euclidean Distance as a similarity metric in order to satisfy the 'Concept Coherency' property, and finally the idea of linear measurable and TCAV [69] testing score in order to satisfy the 'Concept Importance' property and extract the important concepts. ACE performed well in extracting conceptual features without concept annotations, but it has several drawbacks: Inconsistent learning of weights for different samples, which is a common problem in linear interpretation methods [146]; ACE uses TCAV, which only provides concepts relevant to the target class, and it is unclear how to measure whether the provided concepts are fully important for the prediction output [86, 146]; During the process of inverting the features back to their initial dimensions, a large amount of input information can be overlooked [147]; The conceptual outputs of ACE may not always be faithful to the input information and may have less fidelity due to the use of K-mean as a dimension reduction method rather than PCA [146, 147]; The segmentation method employed is independent and may fail to extract semantic concepts, resulting in low fidelity to the original model [146].

Authors in [146] proposed Invertible Concept-based Explanations (ICE) to address some significant shortcomings in ACE [106] and improve the performance of the conceptual discovery model. It tries to provide: Consistent learning of feature weights; fidelity estimation by using NMF (non-negative factorization matrix) as dimension reduction method to provide NCAVs (Non-negative Concepts Activation Vectors) instead of K-mean matrix factorization; Accurate TCAV measurement to determine whether the provided concepts are fully significant and faithful to the primary CNN model. Through human

Table 4 Comprehensive summary of Unsupervised (UNS) concept-supported interpretation methods. Demonstrate XAI category either Post-hoc (PH) or Intrinsic (IN), and Global (G) or Local (L) interpretation

Concept-Supported Method	Ref	Publisher	CNN Architecture	Learning Method	Training Dataset	Concept Dataset	Concept Usage	Concept Scope	Importance Scoring Method
ACE	[106]	NeurIPS	Inception-V3	UNS	ImageNet	ImageNet 100 classes	PH	G	TCAV
ConceptSHAP	[105]	NeurIPS	Inception-V3	UNS	ImageNet	Shape, AwA	PH	G	ground truth
CAR	[160]	NeurIPS	ResNet-50 Inception-V3	UNS	MNIST	CUB, MNIST	PH	G	TCAR
ICE	[146]	AAAI	ResNet50, Inception-V3	UNS	ImageNet	CUB	PH	G & L	NCAV
CoCoX	[161]	AAAI	VGG-16	UNS	ImageNet	ImageNet 40 classes	PH	L	TCAV
DISSECT	[162]	ICLR	Inception-V3	UNS	ImageNet	3D Shapes, CelebA, SynthDerm	PH	G	DISSECT
MCD	[154]	arXiv	ResNet50, ResNet50v2, Swin-T	UNS	ImageNet	CIFAR10	IN	G & L	TCAV

interpretation experiments, ICE [146] demonstrated superior concept uncovering to ACE and PCA. ICE, on the other hand, only considers the concept weights of subspaces in a single dimension and thus in a single decision manner. It provides interpretation using a linear function of estimation, which may not be fully applicable to complex real-world problems.

The work in [147] proposed an innovative model design called TreeICE to take advantage of ICE outperformance while addressing its shortcomings using a Decision Tree (DT) classifier [148, 149]. Work in [150] proposed an earlier successful combination of CNN and DT model called CNN2DT to interpret model prediction in a human comprehensible way. However, CNN2DT required a labor-intensive pre-defined semantic label dataset and was incompatible with datasets with variations [151]. Another recent attempt in this area is ACDTE (Automated Concept-based Decision Tree Explanations) [152] which extracts higher-level features and provides a counterfactual interpretation. However, this model is limited by linear-based model constraints. TreeICE uses ICE as a baseline and DT to provide accurate information about important concepts and their importance for prediction output. It employs NMF rather than clustering to reduce the feature dimension, as demonstrated by [146], and then generates NCAV scores to provide accurate conceptual importance. The experiments were conducted using both human-experience and computational methods, with interpretation performance evaluated using five Satisfaction metrics proposed by [153]. The results revealed that decision tree interpretation outperformed linear ones, resulting in a significant improvement in fidelity over ICE. Furthermore, due to the tree-based meaningful structure, DT-based interpretation can strongly associate with expert understanding and can be used for complex CNN models. However, the proposed TreeICE framework as an automatic conceptual extraction model requires further development in order to be fully countable for some real-world problems and difficult concepts such as medical.

Authors in [154] proposed the most recent attempt at automatic concept uncovering as an MCD (Multi-dimensional Concept Discovery). It is a linear conceptual model comparable with [155] that does not require retraining and is very similar to ICE [146]. Unlike other methods such as ConceptSHAP [105], MACE [156], and PACE [157], which aim to learn concepts before mapping them to attribute space, MCD intends to shorten the concept extraction procedure by obtaining importance scores directly from the primary model. MCD's concept discovery method is similar to [158], but with the important advantage of using multiple directions of feature space, allowing for multi-dimensionality, whereas [158] is limited to a single orthogonal direction. In other words, MCD is an outstanding extension of previous works in terms of addressing important limitations and fully satisfying conceptual interpretation. 'Completeness' is one of the essential interpretation properties that previous studies have not fully achieved [153, 159]. Whereas [154] aims to achieve not only a massive improvement in interpretation models but also a high level of concept completeness in terms of actual model reasoning comprehension. MCD's main advantage over previous works such as ACE [106], ICE [146], CNN2DT [150], and ACDTE [152] is concept designation in multiple directions. The concepts are obtained from the hidden intermediate layer in various directions such as single, orthogonal, and arbitrary, resulting in the provision of a completeness correlate of conceptual contributions. This significant advancement in conceptual discovering attributes has the potential to be extremely beneficial in complex domains such as natural expertise severity degree in cancer detection areas.

We reviewed several concept-supported interpretation methods in this section, including supervised, semi-supervised, and unsupervised approaches, as summarised in Tables 3 and 4. Based on existing efforts, these methods can be compared from various perspectives, which we outline in Table 5 as comparison properties.

Table 5 Main comparison properties of existing concept-supported interpretation methods

Comparison Properties	Description of Comparable Properties
Predefined concept dataset is required	The concept annotated dataset is a key component of conceptual interpretation methods that aid in training CNN with human-readable attributes
Unsupervised concept generation	The removal of the manual concept annotation task, which is extremely expensive and impractical for real-world problems, is highly associated
The cost–benefit analysis of model and interpretation precision	The preferred interpretation method preserves the original model performance while providing interpretation to the end user
Interpretation can be either local or global	The global interpretation contributes to the overall trained CNN learned, whereas the local interpretation explains individual attributes that are frequently overlooked in the global interpretation. Thus, the method with both global and local interpretation has the advantage of allowing us to comprehend the entire structure
Interpretation based on intrinsic or post-hoc criteria	Intrinsically-based methods rely on and are embedded within CNN architecture and are used during the training phase. Post-hoc methods, on the other hand, use an independent model and are applied after the training process
Provide sufficient concept information	Interpretation methods must provide sufficient information about each concept(s) and whether or not they are important in explaining the assigned task
Intervention at the local or global level	The intervention capability can be very useful in allowing experts to modify the prediction function via the human evaluation process and thus improve the proposed CNN model

2.3 Comparison of Existing Methods: Strengths and Weaknesses

Various model characteristics indicate their respective weaknesses and strengths, according to a diverse body of literature on current concept-supported interpretation methods. Table 6 summarises these elements, which can be divided into three categories: concept data, model performance, and model editing via human intervention.

Concept Data: Models with only pre-defined concept classes perform well within those classes but struggle when concept labels are missing. Authors in [163] pointed out that meaningful attributes cannot be extracted completely unsupervised. As a result, high-level interpretation models that rely on predefined human-constructed annotations, as well as automatic concept discovery, are regarded as strengths. Furthermore, we found two studies that used innovative, data-efficient approaches to address the limitations of current conceptual data availability. PCBM [88] created a multimodal method for improving concept learning tasks by combining a text encoder and language structure with an image encoder. MCD [154] pioneered a novel multidimensional approach for extracting concepts directly from the primary model in multiple directions, a significant improvement over previous works such as ACE, ConceptSHAP, and ICE.

Table 6 Methods of existence comparison: strengths and weaknesses

Concept Method	Ref	Concept Data			Model Performance						Model Editing	
		Pre-defined Concept	Concept Discovery	Data Efficient Manner	Local Interpretation	Global Interpretation	Importance Score	Trade-Off	Multimodal Approach	Computational Cost	User Local Intervention	User Global Intervention
PCBM	[88]	✓	✗	✓	✓	✓	✗	Low	✓	Low	✗	✓
Concept Transformer	[140]	✓	✗	✗	✓	✓	✓	n/a	✗	n/a	✗	✗
CBSD	[92]	✓	✗	✗	✗	✓	✗	n/a	✗	n/a	✗	✗
Debiased CBM	[93]	✓	✗	✗	✓	✗	✓	Low	✗	n/a	✗	✗
CAV	[69]	✓	✗	✗	✗	✓	✓	✗	✗	n/a	✗	✗
CBM	[87]	✓	✗	✗	✓	✗	✗	low	✗	High	✓	✗
CaCE	[141]	✓	✗	✗	✗	✓	✓	✗	✗	High	✗	✓
IBD	[83]	✓	✗	✗	✗	✓	✓	n/a	✗	n/a	✗	✗
Hierarchical CBM	[90]	✓	✗	✗	✗	✓	✗	low	✗	low	✗	✗
CBM-AUC	[89]	✓	✓	✗	✓	✓	✗	Low	✗	Low	✗	✗
Concept Extract	[142]	✓	✓	✗	✓	✗	✓	✗	✗	Low	✗	✗
CME	[98]	✓	✓	✗	✗	✓	✓	✗	✗	Low	✗	✓
Net2Vec	[110]	✓	✓	✗	✗	✓	✓	✓	✗	n/a	✗	✗
NetDissect	[143]	✓	✓	✗	✓	✓	✓	✗	✗	n/a	✗	✗
ACE	[106]	✗	✓	✗	✗	✓	✓	✗	✗	Low	✗	✗
ConceptSHAP	[105]	✗	✓	✗	✗	✓	✓	✗	✗	Low	✗	✗
CAR	[160]	✗	✓	✗	✗	✓	✓	✗	✗	n/a	✗	✗
ICE	[146]	✗	✓	✗	✓	✓	✓	✗	✗	n/a	✗	✗
CoCoX	[161]	✗	✓	✗	✓	✗	✓	✗	✗	n/a	✗	✗
DISSECT	[162]	✗	✓	✗	✗	✓	✓	✗	✗	n/a	✗	✗
MCD	[154]	✗	✓	✓	✓	✓	✓	n/a	✗	n/a	✗	✗

Model Performance: A variety of factors contribute to the development of concept interpretation methods that perform optimally. Desired approaches frequently include both local and global explanations. While global explanations provide a broad understanding of the model's overall function, they are computationally expensive. Local explanations, on the other hand, focus on individual attributes, which can be computationally cheaper but may be overlooked in a global context. Due to their reliance on binarized concepts, concept bottleneck-based models generally require more computation [98]. Another critical factor is the trade-off between the primary convolutional model and the interpretation method; the ideal explanation method should provide insights while sacrificing prediction accuracy to a minimum. A method with a higher importance score provides more detailed information and more accurate interpretation.
Model Editing Through Human Intervention: Concept-supported interpretation methods that allow users to instantly change concept values and improve prediction functions provide significant benefits. Ideally, we want to use expert input to edit models in order to reduce computational complexity, simplify complex models, integrate expert feedback directly into the model, and optimise through active learning. Editing can be applied to individual instances or globally to improve overall model performance. As a result, we anticipate that both local and global human interventions will significantly strengthen the strengths of these model*s.*

3 Visual concept attribute datasets

Based on the literature, there are several datasets used for concept-supported explanations, which are often referred to as probe datasets and are distinct from training datasets. Instead of meaningless features like pixels, probe datasets contain concept labels associated with higher-level or human-understandable features. In a bird identification task, for example, the concepts can be defined as 'Beak,' which has attributes such as Shape, Colour, and Length that are equally meaningful to humans and black-box models. In contrast, training datasets contain only raw pixel input images. In this section, we look at available concept and training datasets that are relevant to the concept-supported research domain, as summarised in Tables 7 and 8.

3.1 CUB—Caltech-UCSD birds-200–2011

The CUB dataset [102] contains 11,788 samples from 200 different bird species (e.g., Acadian Flycatcher, American Crow, Cardinal), as shown in Fig. 9. Each sample was annotated with bird categories, 15 part location classes, 28 attribute-groupings, 312 concept attributes in binary format (Exists, Doesn't Exist), certainty level (Not Visible, Guessing, Probably, Definitely), and a single bird bounding box. In concept-supported modelling, CUB is the most useful probe dataset. To achieve more accurate results, some studies, such as [87–89, 93, 98, 160, 164] used 112 concept labels that appear for equal or greater than ten classes.

Some examples of concept part location and attribute-groupings in CUB datasets: 'Beak' has [billShape, billColor, BillLength], 'Belly' has [BellyPattern, BellyColor],

Table 7 Comprehensive list of Concept (Probe) datasets employed in concept learning models

Concept Dataset	Ref	Data Type	Explanator	Reference Link	Applied Publications
CUB	[102]	Image	Concept-based	https://www.vision.caltech.edu/datasets/cub_200_2011	[87–89, 93, 98, 140, 146, 147, 160]
BRODEN	[143]	Image	Concept-based	http://netdissect.csail.mit.edu/	[83, 88, 91, 110, 143, 164]
ADE20K	[1, 165]	Image	Concept-based	http://groups.csail.mit.edu/vision/datasets/ADE20K/	[91, 152, 178]
PASCAL	[166]	Image	Concept-based	http://host.robots.ox.ac.uk/pascal/VOC/	[91, 164]
OpenSurfaces	[167]	Image	Concept-based	http://opensurfaces.cs.cornell.edu/intrinsic	[91]
DTD	[168]	Image	Concept-based	https://www.robots.ox.ac.uk/~vgg/data/dtd/	[179]
3D-Shapes	[128, 129]	Image	Concept-based	https://github.com/deepmind/3d-shapes	[162]
dSprites	[127]	Image	Concept-based	https://github.com/deepmind/dsprites-dataset/	[92, 98, 171]
CelebA	[170]	Image	Concept-based	https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html	[141, 162, 171, 172]
Synthderm	[162]	Image	Concept-based	https://affect.media.mit.edu/dissect/synthderm	[162]
Derm7pt	[116]	Image	Concept-based	http://derm.cs.sfu.ca	[88, 177]
OAI	[173]	Image	Concept-based	N/A	[87, 97, 174, 175]
COCO-Miniplaces	[180]	Image	Concept-based	https://github.com/anonymous	[141]
colored-MNIST	[181]	Image	Concept-based	N/A	[141]
BARS	[181]	Image	Concept-based	N/A	[141]
AWA	[95]	Image	Concept-based	https://cvml.ista.ac.at/AwA/	[105]

Table 8 List of Training datasets employed in original neural network training process

Training Dataset	Ref	Data Type	Explanator	Reference Link	Applied Publications
ImageNet	[182]	Image	Concept-based	https://image-net.org/challenges/LSVRC/	[69, 83, 87–90, 98, 110, 140, 143, 154]
Cityscapes	[138]	Image	Concept-based	http://www.cityscapes-dataset.net/	[91]
CIFAR10, 100	[109]	Image	Concept-based	https://www.cs.toronto.edu/~kriz/cifar.html	[88]
MS-COCO	[101]	Image	Concept-based	https://cocodataset.org	[88]
Places365	[183]	Image	Concept-based	http://places2.csail.mit.edu/challenge.html	[83, 110, 141, 143, 152]
Places205	[184]	Image	Concept-based	http://places.csail.mit.edu/	[143]
MNIST Even/Odd	[185]	Image	Concept-based	https://github.com/pietrobarbiero/entropy-lens	[140]
HAM10000	[114]	Image	Concept-based	https://github.com/ptschandl/HAM10000_dataset	[88]

Fig. 9 Examples of CUB datasets. The birds' type from left to right: 'Cardinal', 'Common Yellowthroat', 'Gray Crowned Rosy Finch', 'Yellow Headed Blackbird', 'Winter Wren', 'Cedar Waxwing', 'Pine Warbler', 'Indigo Bunting', 'Painted Bunting'

'Throat' has [ThroatColor], 'Crown' has [CrownColor], 'Tail' has [UpperTailColor, UnderTailColor, TailPattern, TailShape].

3.2 BRODEN—Broadly and Densely Labelled Dataset

BRODEN [143]] is the second most commonly used dataset in concept-supported interpretation research. BRODEN is made up of four distinct labelled and densely visual sample datasets: ADE20k [1, 165], PASCAL [166], OpenSurfaces [167], and DTD [168]. It contains 1376 concept labels and 62,000 image samples from various scene (468), object (584), material (32), object-part (234), texture (47), and colour (11) types. All concept classes are normalised and annotated with one of 11 colour concepts in [169]. Recent studies used the BRODEN dataset in part for higher-level model learning: [88] used 170 concepts, [83] used 660 concept labels with 30 K samples, [110] used 682 concepts of scenes and texture, and [91] used 377 material and object part concept labels. Figure 10 shows an image sample from the BRODEN dataset.

3.3 CelebA—CelebFaces Attributes Dataset

CelebA [170] is a large celebrity image face sample with over 200 K samples, annotated with 10,177 identities, 40 labels face attributes (e.g. Bald, Bangs, Big_Lips, Big_Nose, Black_Hair), and 5 landmark locations (e.g. lefteye_x, lefteye_y, nose_x, nose_y). All image samples in Fig. 11 were resized to 128 * 128 pixels aligned in the image centre. Many concept learning methods use it, including CaCE [141], DISSECT [162], GlanceNets [171], and [172].

Fig. 10 Samples of BRODEN dataset. Each pair images from left to right: concept 'Object' (flower), concept 'Color' (pink), concept 'Object Part' (headboard), concept 'Scene' (street), and concept 'Texture' (swirly)

Fig. 11 CelebA dataset concept samples. Each pair images from left to right: concept 'Eyeglasses', concept 'Bangs', concept 'Wavy Hair', concept 'Mustache', concept 'Smiling'

3.4 dSprites dataset

The work in [127] proposed well-created 2D shape samples, which are frequently used in unsupervised approaches. It consists of six concept labels with their defined values: 'Color' [White], 'Shape' [square, ellipse, heart], 'Scale' [6 values linearly spaced in [0.5, 1]], 'Orientation' [40 values in [0, 2 pi]], 'Position X' [32 values in [0, 1]], 'Position Y' [32 values in [0, 1]], which generate over 737 K image samples. Images samples in 64*64 pixel and black and white format, as shown in Fig. 12, are used in some recent conceptual explanation efforts such as CBSD [92], CME [98], GlanceNets [171].

3.5 Other datasets

Other datasets that have been used in concept studies include: The 3D-Shapes dataset [128, 129, 162] contains 480 K images of 3D shapes in 64*64 pixels and RGB format. It created based on six concepts and their values: 'Floor hue' [10 values linearly spaced in [0, 1]], 'Wall hue' [10 values linearly spaced in [0, 1]], 'Object hue' [10 values linearly spaced in [0, 1]], 'Scale' [8 values linearly spaced in [0, 1]], 'Shape' [4 values in [0, 1, 2, 3]], and 'Orientation' [15 values linearly spaced in [-30, 30]]. The OAI—Osteoarthritis Initiative dataset [173] contains knee X-ray samples from patients at risk of knee osteoarthritis that were used in CBM [87, 97, 174], Cause and Effect Conceptual Framework [175]. Authors in [162] created the SynthDerm dataset, which contains dermatology melanoma skin lesions samples developed from [176]. It includes over 2600 image samples labelled with skin colour, lesion shape, size, and location concepts. Another dermatology-related dataset used for concept explanation model of malignancy detection task [88, 177] is Derm7pt [116]. Figure 12 depicts sample images from the aforementioned datasets.

4 Sufficiency metrics in concept-supported XAI

The primary goal of interpretation is to make complex AI/ML/DL systems transparent to decision-makers, allowing them to determine whether the proposed model is rational enough to use. However, critical questions about the interpretation methods remain:

- What is the status of the explanation?
- Is the user happy?
- How well can end users comprehend model performance?

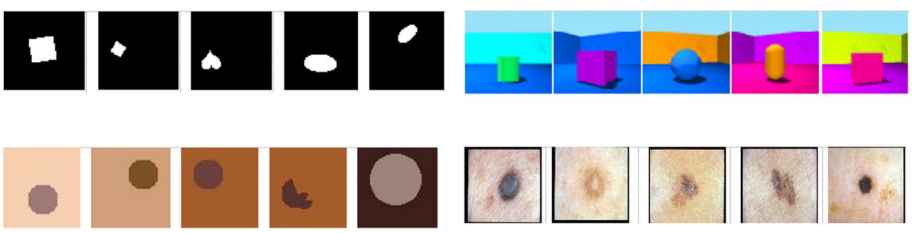

Fig. 12 Top-left set of images are the examples of **dSprites** dataset. Top-right set of images are the examples of **3D-Shapes** dataset. Bottom-left set of images are the samples of **SynthDerm** dataset. Bottom-right set of images are the examples of **Derm7p** dataset

- How reliable is the system?

This section examines key measurement scales of concept-supported interpretation methods used in various studies. As [159] points out, 'Fidelity' is an important evaluation metric that must be explicitly provided in post-hoc based models. However, when working on post-hoc based modelling, [88, 160] overlooked this sufficiency metric, but it was provided in some other studies [85, 146, 147, 152, 159]. Tables 9 and 10 show a summary of sufficiency metrics, their corresponding evaluation methods, and the equations that go with them.

4.1 Meaningfulness

Refers to each defined concept's semantic meaning. It is frequently applicable when a collection of single features create meaningful conceptual attributes such as colour, object, and texture that have meaning for human grasp [186]. In other words, a single input feature must be meaningfully connected to defined concepts. In imaging systems, for example, input pixels (X_i) must be linked to conceptual features (C_i) in order to identify different objects or concepts in image samples [106]. In the case of clinical data, for example, there is a meaningful relationship between 'insulin' and 'diabetes,' but not between 'insulin' and 'hypertension' [38]. As a result, meaningfulness can be directly defined or assessed using human cognitive knowledge. Meaningfulness can refer to semantical segmentation methods such as SLIC [144, 187], LI-SLIC [188], SEAL [189], SSN [190], and FCN [191] from a functional standpoint.

4.2 Completeness

Explanation sufficiency or completeness indicate whether the proposed explanation is sufficient to fully interpret the model or whether additional efforts are required [25, 105]. It collaborates to answer the following questions:

- Do end-users obtain a complete interpretation of the model?
- Is it possible to provide additional explanation?
- Have we defined or extracted enough conceptual features? [86, 159].

According to [192], the degree of interpretation completion can also be related to the lack of problem and target formalisation. Although the completeness metric was frequently overlooked in previous interpretation methods [25], some intriguing efforts have recently been made to measure it via concept completeness score [105, 154]. The quantification metric was proposed by ConceptSHAP [105] to identify the sufficiency and importance of high-level features. As a result, the low completeness score indicates the need for more concept-labels. ConceptSHAP calculates concept sufficiency based on final model prediction accuracy, whereas [154] calculates completeness based on inner parameters and concept weight vector. [154] proposed a multidimensional method for fully correlating concept features and addressing the sufficiency level of model reasoning.

Table 9 Summary of six sufficiency metrics in Concept-supported interpretation methods with their corresponding evaluation methods and ML tasks

Evaluation Metric	Other Terms	Reference	Explanator	Equivalent ML Task	Evaluation Method
Meaningfulness	Factual Correlation	[106, 186]	Concept-based	Segmentation	SLIC, LI-SLIC, SEAL, SSN, FCN
Completeness	Sufficiency	[105, 154]	Concept-based	Concept Activation Mapping	ConceptSHAP, CAR
Importance	Concepts Sensitivity	[69, 83, 106, 141, 146, 147, 160–162]	Concept-based	Feature Weight Learning	CAV, TCAV, NCAV, VAE-CaCE, DISSECT, CAR
Coherency	Similarity	[106, 186, 196]	Concept-based	Clustering	K-MEAN
Interpretability	Explainable, Understandable	[147, 153, 192]	Concept-based	Human Understandable	Human Experiment Assessment
Fidelity	Accuracy, Faithfulness	[85, 146, 147, 152, 159]	Concept-based	Accuracy Score	F1-Score, Recall

Table 10 Existing evaluation metrics equations/methods assessment

Evaluation Metric	Equations	Reference
Meaningfulness	Quantitative human experiment questionnaires E.g. to identify whether the obtained concepts are meaningful, the users were requested to choose the image segments related to a specific concept versus a random image segments and asked them to provide a term to the chosen concept	[106]
Completeness	$\eta f(c_1, ..., c_m) = \frac{sup_g \mathbb{P}_{x,y\sim V}[y=argmax_{y'} h_{y'}(g(v_c(X)))]-a_r}{\mathbb{P}_{x,y\sim V}[y=argmax_{y'} f_{y'}(X)]-a_r}$ With the proposed prediction function of $f(x) = h(\varnothing(x))$, and sample of concept vectors $c_1, \ldots, c_m$, where v refer to the group of verification samples, $sup_g \mathbb{P}_{x,y\sim V}[y = argmax_{y'} h_{y'}(g(v_c(X)))]$ indicate the best accuracy obtained by provided concept scores $v_c(X)$, and a_r is the random classification's accuracy	[105]
Importance	$TCAV_{Q_{C,k,l}} = \frac{\lvert\{x\in X_k : S_{C,k,l}(x)>0\}\rvert}{\lvert X_k\rvert}$ Where $S_{C,k,l}(x)$ refers to sensitivity of primary model prediction according to the concepts C in related neural network's layer l. k indicate given class and X_k related to all input sample belongs to that class	[69]
Coherency	Quantitative human experiment questionnaires E.g. Given six image samples (discovered or human-labelled concepts) to the human and asked them to choose an image sample out of these six sample that are different in concept	[106]
Interpretability	Quantitative human experiment questionnaires E.g. used statement "I understand how the AI system classifies inputs from this explanation model" to evaluate the quality of proposed explanation model	[147]
Fidelity	$Fid_{c\,F,\hat{F}}(I) = \frac{\#\left\{i \in I \lvert F(i) = \hat{F}(i)\right\}}{\#\{I\}}$ Where F refers to the primary pre-trained CNN model, $\hat{F}$ indicate approximation model and I is a set of input image samples	[146]

4.3 Importance

Denotes the level of contribution and significance of higher-level features for each model prediction class. The task of identifying important conceptual features is regarded as extremely difficult because ML models process lower-level features such as pixels while humans understand higher-level features such as objects. TCAV [69] was used in many studies [83, 106, 161] to convert an input lower-level feature and activation layer into human-understandable concept features and then quantify the sensitivity of concepts to model prediction. CAV, on the other hand, is subject to human bias due to the use of predefined concepts and a lack of fidelity to the original model. In other works, [146, 147] used NCAV to automatically extract significant concepts and provide an importance score. Authors in [141] proposed VAE-CaCE to overcome TCAV's limitations. TCAV has a significant difficulty distinguishing between confounding features such as 'car' and 'bicycle' that are very similar. The causal effect of higher-level features is considered in VAE-CaCE to quantify the importance of concept for model decision. Work in [162] proposed DISSECT, a robust generative method for determining the importance score and efficiently distinguishing concepts with high similarities. The authors applied the Concept Traversals Sequences method to a challenging malignant and benign skin dataset. Work in [160] recently proposed CAR (Concept Activation Region) to extend the CAV [69] idea and propose an accurate explanation of confounding concepts. The linear approach of CAV allows the model to determine that 'Stripe' is an important concept for identifying 'Zebra,' but what if we have other classes with 'Stripes,' such as 'Tiger, Lion'? CAR, try to generalise CAV to nonlinear decisions rather than linear ones and provide an accurate explanation.

4.4 Coherency

Indicates that samples from one class of concepts must be similar while being dissimilar to samples from another class of concepts. For example, in terms of bird species identification data [102], there is a concept of 'wing' that contains various samples of wings that differ from the concept of 'leg'. Clustering methods such as K-Means [145], U-K-Means [193], Contrastive Clustering [194], and DeepClustering [195] can achieve coherency. In some studies [106, 186, 196], coherency was used to demonstrate the quality of extracted visual concepts. However, it frequently evaluates based on human experts' comprehension rather than numerical scores [196].

4.5 Interpretability

Defined as the ability to provide explanation, preferably in the form of logical rules that are understandable by human knowledge [50]. The most difficult evaluation metric is interpretability, and how it can be explicitly measured remains an open area of research [50, 85, 197]. For decision trees models, the size (depth of tree structure) of model often used as measurement criteria [198]. The size (depth of tree structure) of the model is frequently used as a measurement criterion for decision tree models [198]. In some studies, the human experiment assessment was used to determine the degree of interpretability [147, 153, 192].

4.6 Fidelity

Indicates the level of precision with which the interpretation method can explain the black-box model. It measures and approximates the level of faithfulness of proposed human understandable explanation for decision making function logic. To put it another way, how well it can mimic the primary model [85, 147, 159]. Due to the structure that embeds into primary model architecture, fidelity score may not require a thorough examination for intrinsic interpretation methods. However, it is required for post-hoc methods because their function of interpretation frequently differs from the primary model architecture [159]. Fidelity score is calculated by dividing the number of samples in the dataset by the ratio of correct explanation c of black-box b for the dataset $c(x) = b(x)$, for $(x, \hat{y}) \in D_{test}$. It can also be measured using common accuracy scores like F1-score and recall [85, 152].

5 Factors influencing the effectiveness of concept-supported XAI

Based to the literature, conceptual interpretation methods can significantly contribute to closing the gap between advanced black-box models and human rational thinking. Current methods, however, are still incapable of providing full interpretation matched to neural network approximation functions. Throughout our extensive research, we discovered six important factors that lead to efficient interpretation models, which we discussed in this section and summarised in Fig. 13. These factors provide useful guidance for current challenges and future research in this field.

5.1 Original neural network settings

The primary neural network model's configuration has a significant impact on concept interpretation. As described in [110, 160], network depth is an important factor in improving concept interpretation accuracy, just as model prediction accuracy is explained in [199, 200]. For example, [154] generated concept interpretation using three different networks and obtained very different results for $n_c = 5$ (number of concepts): ResNet50 obtained a completeness score of 0.89, ResNet50(v2) obtained a completeness score of 0.49, and Swin-T obtained a completeness score of 0.84. In another case, [88] tested the explanation

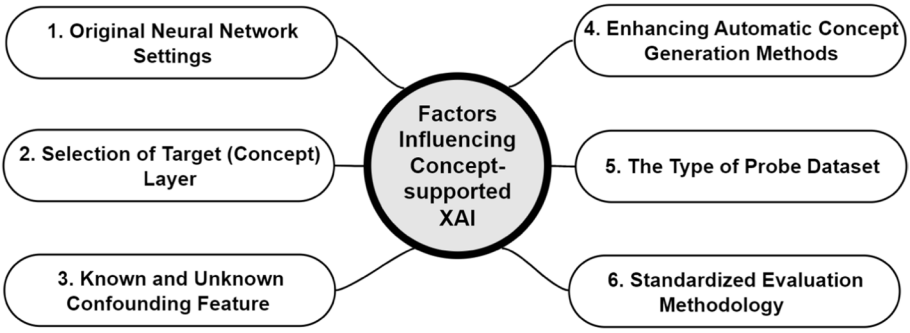

Fig. 13 Summarized factors influencing concept-supported interpretation efficiency (*source: Authors' own elaboration*)

model with CLIP-ResNet50, ResNet18, and Inception networks, as well as various pre-training datasets, and found that the PCBM model performed well but had varying explanation accuracy. In another study, [164] used three different interpretation methods: simple Baseline one [201], NetDissect [143], and TCAV [69] and discovered that model setting was more important than interpretation methods. Layer width or the number of nodes in each layer can be considered as another effective component that has received less attention than layer depth. As a result, it is critical to determine a maximum limit for layer depth and width, which will necessitate additional research.

5.2 Selection of target (Concept) layer

The task of concept learning is directly dependent on the choice of the target layer in a neural network. The feature learning structure of a neural network is hierarchical, which means that layers adjacent to the output (top layers) provide a higher-level feature rather than input layers (low layers) [6, 175, 202]. Many studies [146, 147, 161, 185] assumed that the final layer of a convolutional neural network is the best layer for feature mapping and concept learning. However, [98] pointed out that using a single layer causes a trade-off between lower and higher level conceptual attributes, and that multiple layers are required to address this problem. As a result, the target layer and the number of layers, whether single or multiple, are significant factors on conceptual interpretation performance that must be discovered in the future [147].

5.3 Known and unknown confounding feature

According to [86, 93, 141, 160, 162], confounding attributes have a significant impact on interpretation accuracy. They frequently refer to an incorrect relationship between concept attributes and input data. According to [141], for example, the concepts 'car' and 'bicycle' have strong similar attributes and correlation in input information, which can confuse the interpretation task. To reduce the impact of known confounding factors, [141] proposed the VAE-CaCE score, which can accurately demonstrate the importance of a specific concept for its related label. Work in, [160] provided TCAR and performed a comparable analysis between TCAR and TCAV in another work. They discovered that the TCAR extracted much better correct correlation between concepts and classes, demonstrating a strong concept interpretation performance with strong corelation of concepts and labels. Authors in [93] proposed a method for debiasing and removing confounder vectors by employing a graph-based causality method to discover the relationship between an input feature, a concept, and a target class. Although interpretation methods can be used to identify confounded attributes [203, 204], it is unclear how unobserved confounded attributes can be detected [86].

5.4 Enhancing automatic concept generation methods

Recent interesting efforts for supervised concept interpretation tasks include Hierarchical CBM [90], PCBM [88], CBM-AUC [89], CBM [87], CAV [69], CBSD [92]. However, because of their reliance on predefined concept annotation datasets, these methods are impractical for real-world problems. Creating concept annotations is an extremely expensive operation that frequently results in expert bias [89, 98]. While some efforts have been

made to automatically discover concepts, such as MCD [154], DISSECT [162], ConceptSHAP [105], CoCoX [161], TreeICE [147], ACDTE [152], ICE [146], ACE [106], they are still incapable of performing on large-scale systems with extensive data variations, such as ImageNet [88]. Although [163] argued that meaningful attributes cannot be extracted from fully unsupervised methods, semi-supervised methods can significantly reduce the cost of labelling predefined and human-constructed annotations. As a result, there is currently an active research line in concept discovering methods in an unsupervised manner that can have a significant impact on human-supported interpretation tasks.

5.5 The type of probe dataset

The effectiveness of conceptual interpretation is influenced not only by neural network settings, but also by the type of probe dataset that is typically used for concept learning and differs from the training dataset. [164] highlighted this limitation by employing a simple classifier and interpreting a prediction task using two different probe datasets ADE20k [1, 165] PASCAL [166], both of which contain scene concept labels. They discovered, however, that the interpretation performance varies greatly with different probe datasets. For example, when using the PASCAL dataset, the explanation provides the incorrect concept of 'dog' for the classification of maize farm, whereas the ADE20k provided the correct concept interpretation. Due to insufficient concept labels, [88] raised the same issue and obtained varying interpretation performance using different datasets CIFAR10, CIFAR100 [109], and COCO-Stuff [112]. They proposed using multimodal and textual concept descriptions to improve explanation with lower performance. There are currently only a few pre-defined probe datasets CUB [102], BRODEN [143], CelebA [170] that are limited to specific concepts but do not cover critical tasks such as complex cancer subtype detection or in general medical imaging that require a broad range of concepts to be understandable by experts and regulators. As a result, creating sufficient probe and concept datasets is an active research area that has a significant impact on interpretation effectiveness.

5.6 Standardized evaluation methodology

According to the literature, the evaluation methodology must be standardised in three major ways: 1) *Sufficiency Metrics* such as Meaningfulness [106, 186], Completeness [105, 154], Importance [160–162]; 2) *Experimental parameters* such as neural network settings such as layer depth [110, 160], Layer width, and target layer [98, 146], and 3) *test dataset selection* [88, 164]. For example, as pointed out by [159], Fidelity scores are not always required for intrinsically based interpretation methods like MCD [154], Hierarchical CBM [90], CBM-AUC [89], but they are for post-hoc methods like PCBM [88], DISSECT [162], CAR [160]. Furthermore, the test samples should include both the best and worst cases to demonstrate the model's performance in different scenarios, and they should not be cherry-picked to include only the best samples [53]. However, current studies compared the proposed methods to standard evaluation metrics using different efficiency criteria, so we don't know how close or far they are. This standardisation greatly aids in understanding: a) the true current state of concept-supported methods, b) the areas that require improvement in order to create fully adequate concept-supported interpretation models, and c) how far we are from meeting fully human understandable models for real-world systems and GDPR regulations.

6 Evaluating the robustness of concept-supported XAI methods

According to recent research, using concept-supported XAI improves vision-based complex neural networks significantly. Using concept-supported interpretation techniques such as PCBM, CBSD, CAV, CBM, CAR, ICE, and MCD rather than traditional pixel-based methods such as Saliency Map, Grad-CAM, LIME, and SHAP provides significant benefits for the next generation of vision-based deep and complex models. The following are key strengths of these concept-supported methods in addressing current CNN limitations, as illustrated in Fig. 14.

6.1 Addressing misgeneralization

Convolutional neural nets struggle with extrapolation and generalizing learned information to new conditions with previously unseen and unknown attributes [14, 205, 206]. However, due to conceptualization abilities, human reasoning can successfully generalize learned knowledge and apply it in new environments [207]. In contrast to traditional feature-based explainable methods, which rely on pixels and weights that are meaningless to humans, the new concept-supported approaches are highly aligned with human conceptualization thinking. As a result, these methods significantly associate with interpreting black-box models through higher-level attributes and human-understandable manners, which can significantly aid in addressing the misgeneralization problem.

6.2 Simplification and reduced information overload

Deep convolutional nets, by definition, are trained with millions of parameters, resulting in simplification and reduced information overload. As an illustration, [208] proposed a comparison study for different CNN architectures with different number of layers and parameters, such as AlexNet: 60 M, VGGNet-16: 138 M, Inception-V1: 7 M, and ResNet-152: 50 M (M refers to the number of parameters in Millions). Although deep models performed admirably in recent studies, the large number of trainable parameters in CNN models is

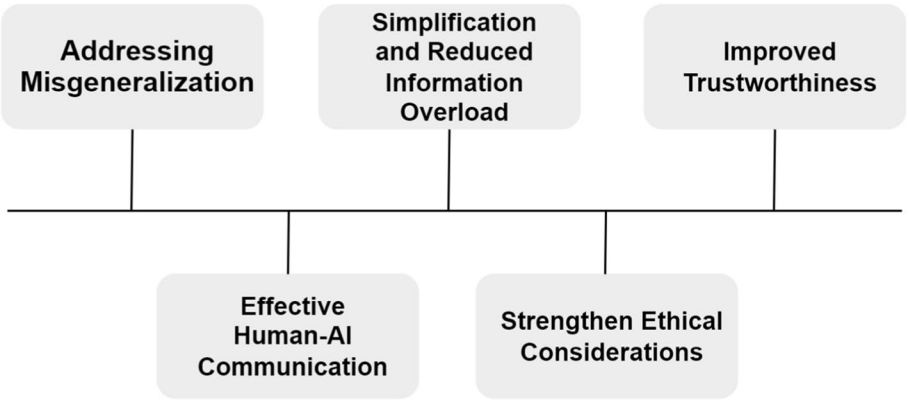

Fig. 14 Significance and robustness of concept-supported Interpretation methods (source: Authors' own elaboration)

one of the key reasons for increasing complexity. However, the issue of model complexity is a significant impediment to optimization and regularization tasks [209]. Princeton University scholars [164] recently conducted the core experiments to better understand the relationship between simplicity and correct recognition at the human level. They discovered that humans perform better with *simplicity* and a smaller number of examples provided. In contrast, given more examples, the recognition result is inaccurate. As a result, using human-level approaches can help to simplify complex model behaviour. Concept-supported interpretations aggregate information into more manageable and comprehensible chunks rather than providing explanations at the feature level.

6.3 Improved trustworthiness

The primary goal of XAI methods is to enable human experts to properly understand and trust machine learning systems [23]. Several studies [55, 67, 68]. have identified feature-based interpretation methods as an untrustworthy approach. As a result of explaining black-box decisions in terms of concept-supported reasoning, experts or decision makers are more likely to trust the model's predictions because the reasoning aligns with their domain knowledge and expectations.

6.4 Effective human-AI communication

AI applications that mimic human cognition tasks such as learning and decision making to assist humans in the same way that team assistants do [79]. This type of interaction assists both parties in achieving common goals [71]. However, feature-based explanation methods do not correspond to human cognition and are difficult to communicate with domain experts. Concept-supported interpretations, on the other hand, facilitate efficient communication between human experts and complex models. These effective human-AI collaborations enable human intervention to provide appropriate and straightforward feedback to aid system enhancement at a higher level. As a result of concept-supported methods, we can achieve Active Learning [210], in which domain experts are actively involved in the learning process.

6.5 Strengthen ethical considerations

Ethical artificial intelligence is essential for deploying complex models in real-world applications [211, 212]. Concept-supported XAI methods, as opposed to feature-based approaches, perfectly align with decision makers' ethical concerns by making it easier to identify and correct undesirable model behaviour. It paves the way for complex models to become more understandable and aligned with human values and expert judgements.

7 Unveiling research horizons in concept-driven XAI

Recent efforts to incorporate human cognition into vision-based complex models have revealed a plethora of unexplored research areas. Exploring opportunities in concept-supported XAI opens the door to a variety of technical innovations in this field. In essence, the distinction between human cognition, which operates at a higher level of comprehension, and complex convolutional neural networks with lower-level learning processes represents

a watershed moment in the vision-based domain [213]. We consolidate the active research areas previously detailed in relevant sections and illustrated in Fig. 15.

7.1 Predefined concept dataset

One of the primary requirements in a human-understandable learning procedure is the creation of a predefined concept-supported dataset. Creating concept datasets, on the other hand, is currently one of the most significant challenges in the concept-supported interpretation task. This is because the significant progress made to date in generating large training datasets, such as ImageNet, CIFAR10, and MNIST, is based on pixel and lower-level learning procedures, rather than the human-understandable approach. Although there is still a lack of field-specific conceptual datasets, the main advantage of these datasets over lower-learning ones is that they require concept classes with fewer examples. For example, IBD [83] uses the BRODEN concept dataset with approximately 10 training samples for each class.

7.2 Concept discovery

As properly described in Sec 5.4, pre-defined concept annotation tasks are expensive operations; the automatic concept discovery task will strongly associate with unsupervised annotation, avoid human errors, and speed up dataset creation. Furthermore, in the complex cancer detection domain, an unsupervised concept discovery approach is extremely useful for generating subclasses such as severity degree.

7.3 Systematic human intervention approach

The key critical advantage of concept-supported XAI methods over traditional methods is proper integration between human experts and complex AI models. As discussed in Sec 6.4, the intervention capability is critical for incorporating human cognition into complex models. Effective human-AI communication is not only important and applicable to improving the explainable AI domain, but it can also greatly improve another key domain. Another type of explanation in mathematical symbol formats is mathematical expression [214, 215]. However, meaningful and semantical recognition are heavily used in this area to represent the meaning of the expression in a way that captures its semantics or intended interpretation [216–219]. As another type of explanation task, concept-supported methods can help to go beyond syntactic parsing and capture the deeper meaning of mathematical

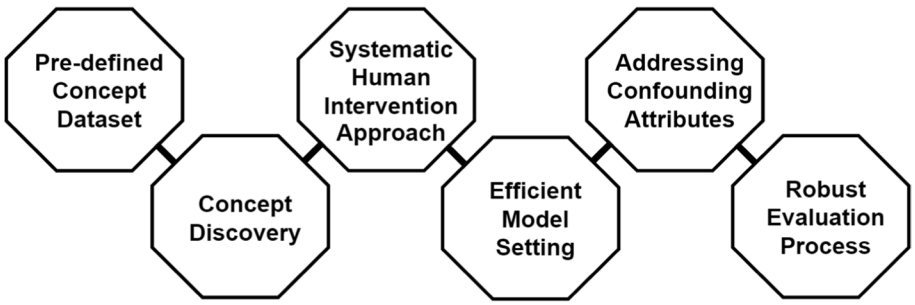

Fig. 15 Opportunities for New Research in Concept-supported XAI (*source: Authors' own elaboration*)

expressions. However, the systematic approach with appropriate criteria requires additional research. For example, how experts can intervene during the training and testing phases, how many examples may be required to avoid human bias, and how the regulatory perspective of this active collaboration between experts and AI systems requires further investigation are all questions that need to be addressed.

7.4 Efficient model setting

This refers to both the primary neural network architecture and the concept or target layer. Layer depth and layer width in the primary model are critical components of model performance, as described in Sects. 5.1, 5.2, and 2.3. Furthermore, the position of the concept layer, whether embedded in the middle of the CNN model or at the tail end, and the number of layers, whether single or multiple, are important active research areas that require further investigation. Another open question in this domain is the optimal number of concept layers, whether single or multiple. We would also recommend appropriate consideration for standardised model setting for high-stack and sensitive domains, such as complex cancer detection, as well as non-sensitive areas.

7.5 Addressing confounding attributes

Controlling the effect of confounding higher-level attributes is one of the current open challenges in concept-supported methods, as stated in Secs. 2.1.6 and 5.3. Few studies have proposed a method for reducing the impact of known confounding conceptual attributes like VAE-CaCE, TCAR, and TCAV. However, eliminating the effect of unknown confounding factors remains an open research area. As a result, detecting unobserved higher-level confounded attributes that need to be discovered greatly aids in improving model performance.

7.6 Robust evaluation process

Another current challenge in the concept-supported domain is the development of an effective evaluation methodology. As detailed in Sect. 5.6 and Table 9, various studies proposed various evaluation metrics. Some evaluation metrics, such as meaningfulness, are strongly related to the human thinking approach, which may lead to biased assessment. The variation in evaluation metric and method used in the literature indicates the need for appropriate evaluation design. The design can be divided at the highest level into qualitative human experiment approaches such as questionaries and quantitative computational approaches such as the TCAV score. The proper evaluation design, which corresponds to the specific type of concept-supported method, such as intrinsic or post-hoc, supervised or unsupervised, has a significant correlation with model enhancement and evaluation methodology standardisation.

8 Summary of key findings

In this comprehensive review, we combed through a vast amount of literature on Machine Learning (ML), Deep Learning (DL), and their applications in a variety of domains. This journey shed light on several novel methodologies and frameworks, demonstrating the immense potential as well as the inherent challenges of this thriving field of study. Among the most important findings:

- ***Concept-based Explanations:*** The review goes into great detail about the emergence of concept-based explanations as crucial in improving model interpretability. Frameworks such as Concept Activation Regions and DISSECT are highlighted in papers for providing disentangled explanations. Overlooked factors that influence the effectiveness of concept-based explanations include dataset selection, concept salience, and human capability.
- ***Deep Learning in Object Recognition and Segmentation:*** A wide range of methodologies for object detection, segmentation, and recognition were investigated, with advances in real-time object detection frameworks such as Faster R-CNN and Mask R-CNN receiving special attention. The effectiveness of superpixel segmentation methods in image segmentation tasks was highlighted.
- ***Datasets and their Impact:*** Various datasets such as ADE20K, Pascal VOC, and COCO were discussed, demonstrating the wide range of real-world problems that ML and DL technologies can address. The paper emphasised how dataset selection influences concept-based explanations and model performance.
- ***Medical Applications of ML and DL:*** The review highlighted several applications, including early detection of diseases such as melanoma and image segmentation. It investigated deep learning for skin lesion classification, indicating a significant potential in medical imaging and diagnostics.
- ***Ethical Considerations in ML and DL*:** Ethical discussions centred on trust in medical AI, as well as the alignment problem, demonstrating the growing awareness of the societal implications of ML and AI technologies.
- ***Unsupervised Learning and Clustering:*** Progress in unsupervised learning and deep clustering for learning visual features was discussed, demonstrating progress in learning representations without labelled data.
- ***Model Complexity and Scalability:*** The trade-off between model depth and width, challenges in scaling ML models to larger datasets, and complex tasks were discussed, shedding light on model complexity nuances.
- ***Mathematical Expression Recognition Exploration:*** The review delves into machine learning models for mathematical symbol recognition, highlighting efforts in recognising handwritten mathematical expressions and symbols.
- ***Interactive Machine Learning:*** The discussion on human-in-the-loop machine learning and leveraging explanations in interactive ML revealed how human interaction can augment ML models.
- ***New Challenges and Future Directions:*** The review also alluded to several challenges and future directions, such as the need for more robust models, ethical frameworks for AI, and more research into unsupervised learning techniques.

This comprehensive review weaves a rich tapestry of the current state of ML and DL, shedding light on the numerous methodologies, applications, and challenges that characterise this dynamic field of study. The review illuminates the path forward through a meticulous exploration of literature, beckoning the scholarly community to delve deeper into the realms of ML and DL, and to continue pushing the boundaries of what is possible with these transformative technologies.

9 Conclusions

Concept-supported Explainable Artificial Intelligence (XAI) is a burgeoning field that seeks to bridge the gap between complex machine learning models and human interpretability, a venture that is critical in crucial domains such as healthcare and autonomous systems. This comprehensive review has focused on concept-supported interpretation methods, which are useful in reducing the opacity of sophisticated deep learning models. By leveraging higher-level attributes or 'Concepts', these methods elucidate the underlying mechanisms of models, bringing a semblance of human reasoning into the realm of machine intelligence.

This review, one of the pioneering efforts, sheds light on the substantial potential and current challenges confronting concept-supported interpretation methods as graphically summarized in Fig. 16. It dissects the anatomy of these methods and investigates the landscape of factors that have a significant impact on their effectiveness. Among these are the original neural network settings, the prudent selection of target or concept layers, the presence of confounding attributes, the advent of automatic concept generation methods, the selection of concept or probe dataset, and the requirement for a standardised evaluation methodology.

The discussion presented here not only emphasises the promise of concept-supported interpretation methods in demystifying deep learning models, but also emphasises the importance of ongoing research and innovation in this domain. The review outlines the trajectory of progress thus far while also highlighting the remaining roadblocks.

The increasing integration of deep learning models in critical decision-making systems, where the stakes are high and the margin for error is small, emphasises the need for such methods. Concept-supported interpretation methods are poised to play a pivotal role in the broader adoption and responsible deployment of artificial intelligence technologies by fostering a deeper understanding and engendering trust.

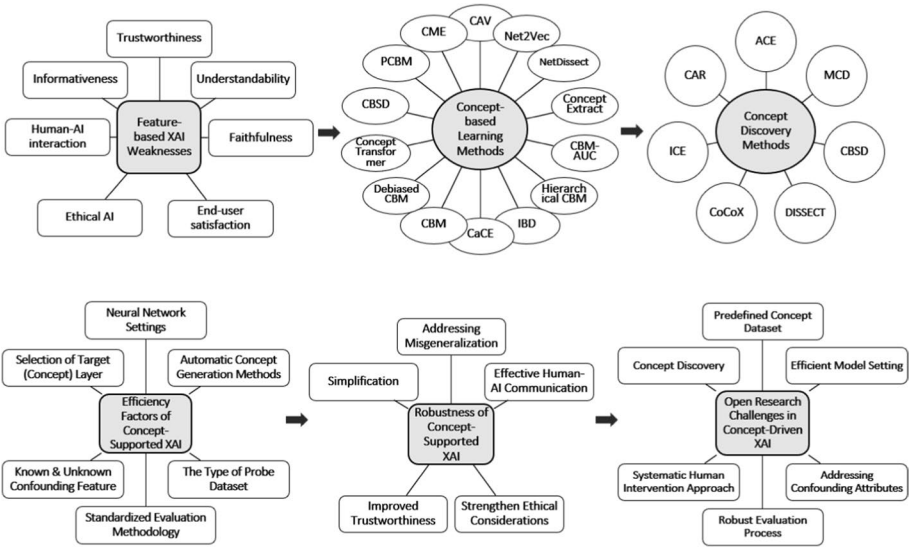

Fig. 16 Summary of concept-supported XAI challenges, innovations, and future directions discussed in this review (*source: Authors' own elaboration*)

In retrospect, the path to fully interpretable AI is fraught with difficulties, but it is a venture that has the potential to significantly improve the synergy between human intelligence and artificial cognition. As a result, this review not only adds to the academic discussion of XAI, but it also provides a vantage point from which practitioners and policymakers can navigate the complex terrain of machine interpretability.

10 Future work

When it comes to automatic concept generation methods, the horizon is brimming with opportunities and avenues for further research and development. The importance of lowering both the financial and time costs of creating concept-labeled datasets cannot be overstated, and it necessitates a collaborative effort from the scholarly community. Similarly, the development of a solid and standardised evaluation methodology emerges as a critical endeavour for meticulously assessing the current state of existing methods and identifying areas ripe for refinement and innovation. The manuscript raises the issue of confounding attributes, a territory that, if explored, could reveal solutions to mitigate misleading outputs, thereby improving the reliability and trustworthiness of the interpretation methods.

Furthermore, a more in-depth investigation into the selection of the target or concept layer for assimilation of concept information has the potential to yield novel insights and methodologies, potentially leading to more accurate and interpretable models. The selection of concept or probe dataset, a critical determinant of the concepts that can be learned, beckons more research attention to ensure that the models can address a broad range of real-world problems. Similarly, the impact of the original neural network settings on the efficacy of concept-supported interpretation methods is ripe for further investigation, potentially leading to more optimised and effective interpretative models.

As the narrative shifts to discussing broader challenges and future directions, it opens up new avenues for the scholarly community to delve deeper into unsupervised learning techniques, ethical frameworks for AI, and model robustness. The manuscript foreshadows an exciting future in which machine learning models and human interpretability are refined to the point where complex models are more accessible, understandable, and trustworthy. The search for new and improved methods to provide meaningful insights into machine learning models' decision-making processes is far from over. It is a continuous effort that promises to not only improve our understanding of these models but also significantly increase their transparency and trustworthiness, especially in high-stakes applications like medical diagnostics and safety–critical systems. The manuscript establishes a solid foundation and points to a future in which the synergy between machine learning models and human interpretability is more than a pipe dream but a tangible reality.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Data availability The datasets mentioned in this review paper are publicly and openly available through their related references.

Declarations

Conflicts of interests / Competing interests The authors whose names are listed in this manuscript certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultan-

cies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. He K et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition
2. Shafiq M, Gu Z (2022) Deep residual learning for image recognition: A survey. Appl Sci 12(18):8972
3. Radford A et al (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9
4. Onyema EM et al (2023) Remote monitoring system using slow-fast deep convolution neural network model for identifying anti-social activities in surveillance applications. Meas: Sens 27:100718
5. Azodi CB, Tang J, Shiu S-H (2020) Opening the black box: Interpretable machine learning for geneticists. Trends Genet 36(6):442–455
6. Buhrmester V, Münch D, Arens M (2021) Analysis of explainers of black box deep neural networks for computer vision: A survey. Mach Learn Knowl Extraction 3(4):966–989
7. Petch J, Di S, Nelson W (2022) Opening the black box: The promise and limitations of explainable machine learning in cardiology. Can J Cardiol 38(2):204–213
8. Van der Velden BH et al (2022) Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis 79:102470
9. Salahuddin Z et al (2022) Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Comput Biol Med 140:105111
10. Zhang X et al (2022) Interpretable machine learning models for crime prediction. Comput Environ Urban Syst 94:101789
11. Zablocki É et al (2022) Explainability of deep vision-based autonomous driving systems: Review and challenges. Int J Comput Vis 130(10):2425–2452
12. Mahbooba B et al (2021) Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. Complexity 2021:1–11
13. Bussmann N et al (2021) Explainable machine learning in credit risk management. Comput Econ 57:203–216
14. Zech JR et al (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 15(11):e1002683
15. Barbedo JG (2018) Factors influencing the use of deep learning for plant disease recognition. Biosys Eng 172:84–91
16. Hendrycks D et al (2021) Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
17. Goodman B, Flaxman S (2017) European Union regulations on algorithmic decision-making and a "right to explanation." AI Mag 38(3):50–57
18. Chattopadhayay A, Rijal I (2023) Towards inclusive privacy consenting for GDPR compliance in visual surveillance: a survey study. In: 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC). IEEE
19. Molnar C, Casalicchio G, Bischl B (2020) Interpretable machine learning–a brief history, state-of-the-art and challenges. In: Joint European conference on machine learning and knowledge discovery in databases. Springer
20. Agarwal R et al (2021) Neural additive models: Interpretable machine learning with neural nets. Adv Neural Inf Process Syst 34:4699–4711
21. Wang C et al (2023) In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. J Quant Criminol 39(2):519–581

22. Angelov PP et al (2021) Explainable artificial intelligence: an analytical review. Wiley Interdiscip Rev: Data Min Knowl Discov 11(5):e1424
23. Arrieta AB et al (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115
24. Doran D, Schulz S, Besold TR (2017) What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv:1710.00794
25. Gilpin LH et al (2018) Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE
26. Alvarez Melis D, Jaakkola T (2018) Towards robust interpretability with self-explaining neural networks. Adv Neural Inf Proces Syst 31
27. Elton DC (2020) Self-explaining AI as an alternative to interpretable AI. In: Artificial General Intelligence: 13th International Conference, AGI 2020, St. Petersburg, Russia, September 16–19, 2020, Proceedings 13. Springer
28. Kumar S et al (2022) Self-explaining neural network with concept-based explanations for ICU mortality prediction. In: Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics
29. Sawada Y, Nakamura K (2022) C-SENN: Contrastive Self-Explaining Neural Network. arXiv preprint arXiv:2206.09575
30. Saleem R et al (2022) Explaining deep neural networks: a survey on the global interpretation methods. Neurocomputing
31. Clough JR et al (2019) Global and local interpretability for cardiac MRI classification. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22. Springer
32. Koo PK et al (2021) Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. PLoS Comput Biol 17(5):e1008925
33. Setzu M et al (2021) Glocalx-from local to global explanations of black box ai models. Artif Intell 294:103457
34. Tan H (2023) Visualizing Global Explanations of Point Cloud DNNs. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision
35. Li X et al (2023) G-LIME: Statistical learning for local interpretations of deep neural networks using global priors. Artif Intell 314:103823
36. Molnar C (2020) Interpretable machine learning. Lulu.com
37. Huang Q et al (2022) Graphlime: local interpretable model explanations for graph neural networks. In: IEEE Transactions on Knowledge and Data Engineering
38. Shawi RE, Al-Mallah MH (2022) Interpretable local concept-based explanation with human feedback to predict all-cause mortality. J Artif Intell Res 75:833–855
39. Mollas I, Bassiliades N, Tsoumakas G (2023) LioNets: A neural-specific local interpretation technique exploiting penultimate layer information. Appl Intell 53(3):2538–2563
40. Li Z (2022) Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. Comput Environ Urban Syst 96:101845
41. Liang Y et al (2021) Explaining the black-box model: A survey of local interpretation methods for deep neural networks. Neurocomputing 419:168–182
42. Petsiuk V et al (2021) Black-box explanation of object detectors via saliency maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
43. Mundhenk TN, Chen BY, Friedland G (2019) Efficient saliency maps for explainable AI. arXiv preprint arXiv:1911.11293
44. Mohseni S, Zarei N, Ragan ED (2021) A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans Interact Intell Syst (TiiS) 11(3–4):1–45
45. Moradi M, Samwald M (2021) Post-hoc explanation of black-box classifiers using confident itemsets. Expert Syst Appl 165:113941
46. Jin W et al (2023) Generating post-hoc explanation from deep neural networks for multi-modal medical image analysis tasks. MethodsX 10:102009
47. Han T, Srinivas S, Lakkaraju H (2022) Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. arXiv preprint arXiv:2206.01254
48. Dai J et al (2022) Fairness via explanation quality: evaluating disparities in the quality of post hoc explanations. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society
49. Chou Y-L et al (2022) Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. Inf Fusion 81:59–83
50. Zhang Y et al (2021) A survey on neural network interpretability. IEEE Trans Emerg Top Comput Intell 5(5):726–742

51. Ribeiro MT, Singh S, Guestrin C (2016) "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
52. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Adv Neural Inf Proces Syst 30
53. Räukur T et al (2022) Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. arXiv preprint arXiv:2207.13243
54. Patrício C, Neves JC, Teixeira LF (2022) Explainable deep learning methods in medical diagnosis: A survey. arXiv preprint arXiv:2205.04766
55. Adebayo J et al (2018) Sanity checks for saliency maps. Adv Neural Inf Proces Syst 31
56. Zhang Z et al (2022) Protgnn: Towards self-explaining graph neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence
57. Ali S et al (2023) Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion 99:101805
58. Groen AM (2022) A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: limited use of explainable AI? Eur J Radiol 110592
59. Fiok K et al (2022) Explainable artificial intelligence for education and training. J Defense Model Simul 19(2):133–144
60. Rawal A et al (2021) Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives. IEEE Trans Artif Intell 3(6):852–866
61. Selvaraju RR et al (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision
62. Shrikumar A, Greenside P, Kundaje A (2017) Learning important features through propagating activation differences. In: International conference on machine learning. PMLR
63. Bach S et al (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 10(7):e0130140
64. Kharya S et al (2022) Weighted Bayesian belief network: a computational intelligence approach for predictive modeling in clinical datasets. Comput Intell Neurosci 2022
65. Alqaraawi A et al (2020) Evaluating saliency map explanations for convolutional neural networks: a user study. In: Proceedings of the 25th International Conference on Intelligent User Interfaces
66. Lipton ZC (2018) The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16(3):31–57
67. Ghorbani A, Abid A, Zou J (2019) Interpretation of neural networks is fragile. In: Proceedings of the AAAI conference on artificial intelligence
68. Gimenez JR, Ghorbani A, Zou J (2019) Knockoffs for the mass: new feature importance statistics with false discovery guarantees. In: The 22nd International Conference on Artificial Intelligence and Statistics. PMLR
69. Kim B et al (2018) nterpretability beyond feature attribution: quantitative testing with concept activation vectors (tcav). In: International conference on machine learnin. PMLR
70. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215
71. Kim SS et al (2023) "Help me help the AI": Understanding how explainability can support human-AI interaction. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems
72. Saeed W, Omlin C (2023) Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowl-Based Syst 263:110273
73. Lakkaraju H, Bastani O (2020) "How do I fool you?" Manipulating user trust via misleading black box explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society
74. Slack D et al (2020) Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.
75. Gunning D, Aha D (2019) DARPA'S explainable artificial intelligence (XAI) program. AI Mag 40(2):44–58
76. Wang X, Yin M (2021) Are explanations helpful? A comparative study of the effects of explanations in ai-assisted decision-making. In: 26th International Conference on Intelligent User Interfaces
77. Langley P et al (2017) Explainable agency for intelligent autonomous systems. In: Proceedings of the AAAI Conference on Artificial Intelligence
78. Bove C et al (2022) Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In: 27th international conference on intelligent user interfaces.
79. Berretta S et al (2023) Defining human-AI teaming the human-centered way: a scoping review and network analysis. Front Artif Intell 6

80. Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. Artif Intell 267:1–38
81. Ehsan U, Riedl MO (2020) Human-centered explainable ai: Towards a reflective sociotechnical approach. In: HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22. Springer.
82. Liao QV, Gruen D, Miller S (2020) Questioning the AI: Informing design practices for explainable AI user experiences. In Proceedings of the 2020 CHI conference on human factors in computing systems.
83. Zhou B et al (2018) Interpretable basis decomposition for visual explanation. In: Proceedings of the European Conference on Computer Vision (ECCV)
84. Alicioglu G, Sun B (2022) A survey of visual analytics for explainable artificial intelligence methods. Comput Graph 102:502–520
85. Guidotti R et al (2018) A survey of methods for explaining black box models. ACM Comput Surv (CSUR) 51(5):1–42
86. Hitzler P, Sarker M (2022) Human-centered concept explanations for neural networks. Neuro-Symb Artif Intell: State Art 342(337):2
87. Koh PW, et al (2020) Concept bottleneck models. In: International conference on machine learning. PMLR
88. Yuksekgonul M, Wang M, Zou J (2022) Post-hoc concept bottleneck models. arXiv preprint arXiv:2205.15480
89. Sawada Y, Nakamura K (2022) Concept bottleneck model with additional unsupervised concepts. IEEE Access 10:41758–41765
90. Pittino F, Dimitrievska V, Heer R (2023) Hierarchical concept bottleneck models for vision and their application to explainable fine classification and tracking. Eng Appl Artif Intell 118:105674
91. Losch M, Fritz M, Schiele B (2019) Interpretability beyond classification output: Semantic bottleneck networks. arXiv preprint arXiv:1907.10882
92. Wijaya MA et al (2021) Failing conceptually: Concept-based explanations of dataset shift. arXiv preprint arXiv:2104.08952
93. Bahadori MT, Heckerman DE (2020) Debiasing concept-based explanations with causal analysis. arXiv preprint arXiv:2007.11500
94. Kumar N et al (2009) Attribute and simile classifiers for face verification. In: 2009 IEEE 12th International Conference on Computer Vision. IEEE
95. Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE
96. Lozano-Diez A et al (2017) An analysis of the influence of deep neural network (DNN) topology in bottleneck feature based language recognition. PLoS One 12(8):e0182580
97. Margeloiu A et al. (2021) Do concept bottleneck models learn as intended? arXiv preprint arXiv:2105.04289
98. Kazhdan D et al (2020) Now you see me (CME): concept-based model extraction. arXiv preprint arXiv:2010.13233
99. Szegedy C et al (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition
100. Ren S et al (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Adv Neural Inf Proces Syst 28
101. Lin T-Y et al (2014) Microsoft coco: common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. 2014. Springer
102. Wah C et al (2011) The caltech-ucsd birds-200-2011 dataset
103. Xu Y et al (2020) Explainable object-induced action decision for autonomous vehicles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
104. Wang D et al (2019) Deep object-centric policies for autonomous driving. In: 2019 International Conference on Robotics and Automation (ICRA). IEEE
105. Yeh C-K et al (2020) On completeness-aware concept-based explanations in deep neural networks. Adv Neural Inf Process Syst 33:20554–20565
106. Ghorbani A et al (2019) Towards automatic concept-based explanations. Adv Neural Inf Proces Syst 32
107. Radford A et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR
108. Speer R, Chin J, Havasi C (2017) Conceptnet 5.5: An open multilingual graph of general knowledge. In: Proceedings of the AAAI conference on artificial intelligence

109. Krizhevsky A, Hinton G (2009) Learning multiple layers of features from tiny images
110. Fong R, Vedaldi A (2018) Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition
111. Abid A, Yuksekgonul M, Zou J (2022) Meaningfully debugging model mistakes using conceptual counterfactual explanations. In: International Conference on Machine Learning. PMLR
112. Caesar H, Uijlings J, Ferrari V (2018) Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
113. Singh KK, et al (2020) Don't judge an object by its context: learning to overcome contextual bias. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
114. Tschandl P, Rosendahl C, Kittler H (2018) The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 5(1):1–9
115. Daneshjou R et al (2021) Disparities in dermatology AI: assessments using diverse clinical images. arXiv preprint arXiv:2111.08006
116. Kawahara J et al (2018) Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J Biomed Health Inform 23(2):538–546
117. Rotemberg V et al (2021) A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Sci Data 8(1):34
118. Bontempelli A et al (2021) Toward a Unified Framework for Debugging Concept-based Models. arXiv preprint arXiv:2109.11160
119. He K et al (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision
120. Wu J et al (2022) The power and limitation of pretraining-finetuning for linear regression under covariate shift. Adv Neural Inf Process Syst 35:33041–33053
121. Shao J-J, Yang X-W, Guo L-Z (2022) Open-set learning under covariate shift. Mach Learn 1–17
122. Chapaneri S, Jayaswal D (2022) Covariate shift in machine learning. In: Handbook of Research on Machine Learning. Apple Academic Press, pp 87–119
123. Schulam P, S Saria (2017) Reliable decision support using counterfactual models. Adv Neural Inf Proces Syst 30
124. Lipton Z, Wang Y-X, Smola A (2018) Detecting and correcting for label shift with black box predictors. In: International conference on machine learning. PMLR
125. Rabanser S, Günnemann S, Lipton Z (2019) Failing loudly: an empirical study of methods for detecting dataset shift. Adv Neural Inf Proces Syst 32
126. Higgins I et al (2016) beta-vae: Learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations
127. Matthey L et al (2017) dsprites: disentanglement testing sprites dataset
128. Burgess C, Kim H (2018) 3d shapes dataset. [Online]. Available: https://github.com/deepmind/3d-shapes. Accessed 01/05/2023
129. Kim H, Mnih A (2018) Disentangling by factorising. In: International Conference on Machine Learning. PMLR
130. Gretton A et al (2012) A kernel two-sample test. J Mach Learn Res 13(1):723–773
131. Weisstein EW (2004) Bonferroni correction. https://mathworld.wolfram.com/. Accessed 01/05/2023
132. He X et al (2023) Addressing confounding feature issue for causal recommendation. ACM Trans Inf Syst 41(3):1–23
133. Li C, Shen X, Pan W (2023) Nonlinear causal discovery with confounders. J Am Stat Assoc 1–10
134. Brisk R et al (2021) The effect of confounding data features on a deep learning algorithm to predict complete coronary occlusion in a retrospective observational setting. Eur Heart J-Digit Health 2(1):127–134
135. Stock JH (2015) Instrumental variables in statistics and econometrics
136. Hooker S et al (2019) A benchmark for interpretability methods in deep neural networks. Adv Neural Inf Proces Syst 32
137. Zhao H et al (2017) Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition
138. Cordts M et al (2016) The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition
139. Xiao T et al (2018) Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV)
140. Rigotti M et al (2021) Attention-based interpretability with concept transformers. In: International conference on learning representations
141. Goyal Y et al (2019) Explaining classifiers with causal concept effect (cace). arXiv preprint arXiv:1907.07165
142. Zhao Z et al (2021) Human-in-the-loop extraction of interpretable concepts in deep learning models. IEEE Trans Vis Comput Graph 28(1):780–790

143. Bau D et al (2017) Network dissection: quantifying interpretability of deep visual representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition
144. Wang X et al (2022) Revisiting SLIC: Fast superpixel segmentation of marine SAR images using density features. IEEE Trans Geosci Remote Sens 60:1–18
145. Basar S et al (2020) Unsupervised color image segmentation: A case of RGB histogram based K-means clustering initialization. PLoS One 15(10):e0240015
146. Zhang R et al (2021) Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In: Proceedings of the AAAI Conference on Artificial Intelligence
147. Mutahar G, Miller T (2022) Concept-based Explanations using Non-negative Concept Activation Vectors and Decision Tree for CNN Models. arXiv preprint arXiv:2211.10807.
148. Chatzimparmpas A, Martins RM, Kerren A (2023) VisRuler: Visual analytics for extracting decision rules from bagged and boosted decision trees. Inf Vis 22(2):115–139
149. Zhang Q et al (2019) Interpreting cnns via decision trees. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
150. Jia S et al (2020) Visualizing surrogate decision trees of convolutional neural networks. J Vis 23:141–156
151. Chatzimparmpas A et al (2023) DeforestVis: Behavior analysis of machine learning models with surrogate decision stumps. arXiv preprint arXiv:2304.00133.
152. El Shawi R, Sherif Y, Sakr S (2021) Towards Automated Concept-based Decision TreeExplanations for CNNs. In: EDBT
153. Hoffman RR et al (2018) Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.
154. Vielhaben, J., S. Blücher, and N. Strodthoff, *Multi-dimensional concept discovery (MCD): A unifying framework with completeness guarantees.* arXiv preprint arXiv:2301.11911, 2023.
155. Chen C et al. (2019) This looks like that: deep learning for interpretable image recognition. Adv Neural Inf Process Syst 32
156. Kumar A et al (2021) MACE: Model agnostic concept extractor for explaining image classification networks. IEEE Trans Artif Intell 2(6):574–583
157. Kamakshi V, Gupta U, Krishnan NC (2021) Pace: Posthoc architecture-agnostic concept extractor for explaining cnns. In: 2021 International Joint Conference on Neural Networks (IJCNN). IEEE
158. Chormai P et al (2022) Disentangled explanations of neural network predictions by finding relevant subspaces. arXiv preprint arXiv:2212.14855
159. Yang F, Du M, Hu X (2019) Evaluating explanation without ground truth in interpretable machine learning. arXiv preprint arXiv:1907.06831
160. Crabbé J, van der Schaar M (2022) Concept activation regions: A generalized framework for concept-based explanations. arXiv preprint arXiv:2209.11222
161. Akula A, Wang S, Zhu S-C (2020) *Cocox:* Generating conceptual and counterfactual explanations via fault-lines. In: Proceedings of the AAAI Conference on Artificial Intelligence.
162. Ghandeharioun A et al (2021) Dissect: Disentangled simultaneous explanations via concept traversals. arXiv preprint arXiv:2105.15164
163. Locatello F et al (2019) Challenging common assumptions in the unsupervised learning of disentangled representations. In: international conference on machine learning. PMLR
164. Ramaswamy VV et al (2022) Overlooked factors in concept-based explanations: Dataset choice, concept salience, and human capability. arXiv preprint arXiv:2207.09615
165. Zhou B et al (2019) Semantic understanding of scenes through the ade20k dataset. Int J Comput Vis 127:302–321
166. Everingham M et al (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88:303–338
167. Bell S, Bala K, Snavely N (2014) Intrinsic images in the wild. ACM Trans Graph (TOG) 33(4):1–12
168. Cimpoi M et al (2014) Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition
169. Van De Weijer J et al (2009) Learning color names for real-world applications. IEEE Trans Image Process 18(7):1512–1523
170. Liu Z et al (2015) Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision
171. Marconato E, Passerini A, Teso S (2022) GlanceNets: Interpretabile, leak-proof concept-based models. arXiv preprint arXiv:2205.15612
172. Brocki L, Chung NC (2019) Concept saliency maps to visualize relevant features in deep generative models. In: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). IEEE

173. Nevitt M, Felson D, Lester G (2006) The osteoarthritis initiative. Protocol for the Cohort Study 1
174. Sinha S et al (2022) Understanding and enhancing robustness of concept-based models. arXiv preprint arXiv:2211.16080
175. Zaeem MN, Komeili M (2021) Cause and effect: Hierarchical concept-based explanation of neural networks. arXiv preprint arXiv:2105.07033
176. Tsao H et al (2015) Early detection of melanoma: Reviewing the ABCDEs. J Am Acad Dermatol 72(4):717–723
177. Lucieri A et al (2020) On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE
178. Zhou B et al (2017) Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition
179. Brown D, Kvinge H (2023) Making corgis important for honeycomb classification: adversarial attacks on concept-based explainability tools. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
180. Yang M, Kim B (2019) Benchmarking attribution methods with relative feature importance. arXiv preprint arXiv:1907.09701
181. Kim B et al (2019) Learning not to learn: training deep neural networks with biased data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
182. Russakovsky O et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vision 115:211–252
183. Zhou B et al (2016) Places: An image database for deep scene understanding. arXiv preprint arXiv:1610.02055
184. Zhou B et al (2014) Learning deep features for scene recognition using places database. Adv Neural Inf Proces Syst 27
185. Barbiero P et al (2022) Entropy-based logic explanations of neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence
186. Fang Z et al (2020) Concept-based explanation for fine-grained images and its application in infectious keratitis classification. In: Proceedings of the 28th ACM international conference on Multimedia
187. Giraud R, Ta V-T, Papadakis N (2018) Robust superpixels using color and contour features along linear path. Comput Vis Image Underst 170:1–13
188. Di S et al (2021) Image superpixel segmentation based on hierarchical multi-level LI-SLIC. Opt Laser Technol 135:106703
189. Tu W-C et al (2018) Learning superpixels with segmentation-aware affinity loss. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
190. Jampani V et al (2018) Superpixel sampling networks. In: Proceedings of the European Conference on Computer Vision (ECCV).
191. Yang F et al (2020) Superpixel segmentation with fully convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.
192. Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608
193. Sinaga KP, Yang M-S (2020) Unsupervised K-means clustering algorithm. IEEE Access **8**:80716-80727
194. Li Y et al (2021) Contrastive clustering. In: Proceedings of the AAAI conference on artificial intelligence
195. Caron M et al (2018) Deep clustering for unsupervised learning of visual features. In: Proceedings of the European conference on computer vision (ECCV)
196. Sauter D et al (2022) Validating automatic concept-based explanations for AI-based digital histopathology. Sensors 22(14):5346
197. Montavon G, Samek W, Müller K-R (2018) Methods for interpreting and understanding deep neural networks. Digit Signal Process 73:1–15
198. Wu M et al (2018) Beyond sparsity: tree regularization of deep models for interpretability. In: Proceedings of the AAAI conference on artificial intelligence
199. Eldan R, Shamir O (2016) The power of depth for feedforward neural networks. In: Conference on learning theory. PMLR
200. Nguyen T, Raghu M, Kornblith S (2020) Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. arXiv preprint arXiv:2010.15327
201. Ramaswamy VV et al (2022) ELUDE: Generating interpretable explanations via a decomposition into labelled and unlabelled features. arXiv preprint arXiv:2206.07690
202. Liu Y et al (2019) DeepCrack: A deep hierarchical feature learning architecture for crack segmentation. Neurocomputing 338:139–153

203. Schramowski P et al (2020) Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nat Mach Intell 2(8):476–486
204. Teso S et al (2022) Leveraging explanations in interactive machine learning: An overview. arXiv preprint arXiv:2207.14526
205. Zhang L et al (2020) Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans Med Imaging 39(7):2531–2540
206. Ngo R, Chan L, Mindermann S (2022) The alignment problem from a deep learning perspective. arXiv preprint arXiv:2209.00626
207. Saxe A, Nelli S, Summerfield C (2021) If deep learning is the answer, what is the question? Nat Rev Neurosci 22(1):55–67
208. Hassan SM et al (2021) Identification of plant-leaf diseases using CNN and transfer-learning approach. Electronics 10(12):1388
209. Hu X et al (2021) Model complexity of deep learning: A survey. Knowl Inf Syst 63:2585–2619
210. Mosqueira-Rey E et al (2023) Human-in-the-loop machine learning: A state of the art. Artif Intell Rev 56(4):3005–3054
211. Durán JM, Jongsma KR (2021) Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics 47(5):329–335
212. Lo Piano S (2020) Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. Human Soc Sci Commun 7(1):1–7
213. Goyal A, Bengio Y (2022) Inductive biases for deep learning of higher-level cognition. Proc R Soc A 478(2266):20210068
214. Sakshi, Kukreja V (2023) Image segmentation techniques: Statistical, comprehensive, semi-automated analysis and an application perspective analysis of mathematical expressions. Arch Comput Methods Eng 30(1):457–495
215. Kukreja V, Lodhi S (2023) Impact of varying strokes on recognition rate: A case study on handwritten mathematical expressions. Int J Comput Digit Syst
216. Kukreja V (2021) A retrospective study on handwritten mathematical symbols and expressions: Classification and recognition. Eng Appl Artif Intell 103:104292
217. Kukreja V, Sakshi (2022) Machine learning models for mathematical symbol recognition: A stem to stern literature analysis. Multimed Tools Applic 81(20):28651–28687
218. Kukreja V (2023) Recent trends in mathematical expressions recognition: An LDA-based analysis. Expert Syst Appl 213:119028
219. Sakshi, Kukreja V (2023) A dive in white and grey shades of ML and non-ML literature: A multivocal analysis of mathematical expressions. Artif Intell Rev 56(7):7047–7135

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Zahra Shams Khoozani[1] · Aznul Qalid Md Sabri[1] · Woo Chaw Seng[1] · Manjeevan Seera[2] · Kah Yee Eg[3]

✉ Aznul Qalid Md Sabri
aznulqalid@um.edu.my

✉ Manjeevan Seera
ManjeevanSingh.Seera@monash.edu

Zahra Shams Khoozani
nazshams23@gmail.com

[1] Department of Artificial Intelligence, Faculty of Computer Science & Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia

[2] Econometrics and Business Statistics, School of Business, Monash University Malaysia, Selangor, Malaysia

[3] Key ASIC Bhd (707082-M), Selangor, Malaysia