Concept-based Explainable Artificial Intelligence: A Survey

Concept-based Explainable Artificial Intelligence: A Survey

ELEONORA POETA*, GABRIELE CIRAVEGNA*, ELIANA PASTOR, TANIA CERQUITELLI, and ELENA BARALIS, Politecnico di Torino, Italy

The field of explainable artificial intelligence emerged in response to the growing need for more transparent and reliable models. However, using raw features to provide explanations has been discussed in several works lately, advocating for more user-understandable explanations. To address this issue, a wide range of papers proposing Concept-based eXplainable Artificial Intelligence (C-XAI) methods have been published in recent years. Nevertheless, a unified categorization and precise field definition are still missing. This paper fills the gap by offering a thorough review of C-XAI approaches. We identify and define different concepts and explanation types. We propose a taxonomy comprising nine categories and guidelines for selecting a suitable category based on the application context. Additionally, we discuss common evaluation strategies including metrics, human evaluations, and datasets employed, aiming to assist the development of future methods. Overall, we believe this survey will assist researchers, practitioners, and domain experts in enhancing their understanding and contributing to the progress of this innovative field.

CCS Concepts: • **Computing methodologies** → **Neural networks**; • **Information systems** → **Data analytics**.

Additional Key Words and Phrases: Explainable AI, XAI, Concept-based Explainability, Trustworthy AI

1 INTRODUCTION

In recent years, the importance of Artificial Intelligence (AI) has surged due to its transformative impact on various aspects of society and industry. This success is predominantly attributed to the advancement of Deep Learning models [57]. However, the inherent complexity and opaque nature of Deep Neural Networks (DNN) hinders the comprehension of the decision-making process underlying these models. This issue prevents the safe employment of these models in critical contexts, significantly affecting users. Consequently, decision-making systems based on Deep learning have faced constraints and limitations from regulatory institutions [36, 66], which increasingly demand transparency in AI models [48].

To address the challenge of building more trustworthy and transparent AI models, researchers have pursued eXplainable AI (XAI) methods [3]. Most XAI methods focus on a single prediction, highlighting which input features contributed the most to the prediction of a given class. They exploit gradient-based analysis (Vanilla Gradient [94], CAM [114], Grad-CAM [92]), local approximations with surrogate interpretable models (LIME [85]) or game-theoretic approaches (SHAP [65]). Other methods, instead, focus on node-oriented explanations to enhance network transparency [29, 94, 110]. A few methods also attempted to globally interpret a model behaviour via global feature importance [38], linear combination of non-interpretable ones [6], or global surrogate models [39, 85].

Standard XAI presents some issues: [4] showed that gradient-based explanations may not change even after randomly re-parametrizing the network or the data labels; [44] showed that methods based on surrogate models are unable to identify the most important features; [34, 53] demonstrated that feature importance methods can be misled by simple data modifications, that do not impact the model prediction. Even if these issues were addressed through the development of more robust methods, however, a fundamental concern remains: standard XAI techniques provide explanations at the feature level, that may lack a meaningful interpretation, especially for non-expert users [79].

*Equal contribution

Authors' address: Eleonora Poeta*, eleonora.poeta@polito.it; Gabriele Ciravegna*, gabriele.ciravegna@polito.it; Eliana Pastor, eliana.pastor@polito.it; Tania Cerquitelli, tania.cerquitelli@polito.it; Elena Baralis, elena.baralis@polito.it, Politecnico di Torino, Torino, Italy.

Fig. 1. Word cloud generated from the titles, abstracts, and keywords of the reviewed C-XAI papers. The trend towards human understanding and interpretability is noticeable, in contrast to standard XAI approaches.

Understanding *where* the network is looking may not be sufficient to explain *what* the network is seeing in a given input [2, 87]. To achieve this objective, there is a growing consensus that XAI techniques should provide explanations in terms of higher-level attributes [2, 31, 50], commonly referred to as *concepts*.

Concept-based explanations offer a compelling alternative by providing a more holistic view of the model's inner workings. By explaining model predictions in terms of human-understandable attributes or abstractions, concept-based explanations better resemble the way humans reason and explain [50, 52]. They also allow users to gain deeper insights into the underlying reasoning, helping them to detect model biases and improve the model itself [18, 46, 115]. Finally, they are more stable to perturbations than standard XAI techniques [9] and can create models inherently more robust to adversarial attacks [24].

This paper proposes the first systematic review of Concept-based eXplainable Artificial Intelligence (C-XAI) methods. While there exists an abundant literature of surveys in the realm of explainability [3, 10, 17, 26, 28, 40, 78], none of them is focused on the concept-based domain, and only a few [17, 78] identify C-XAI as a separate category. We analyzed papers from 2017 to July 2023 appearing in the proceedings of A* AI conferences (such as NeurIPS, ICLR, ICML, AAAI, CVPR) or in Q1 journals of the field (such as Nature Machine Intelligence, Transaction on Machine Learning Research, Artificial Intelligence), or also cited in the related work of these papers. The full details of our Systematic Literature Review (SLR) protocol are presented in the Section 2.

For a visual representation of the analyzed articles, please refer to Figure 1. This visualization depicts the 100 most frequent words in the titles, abstracts, or keywords of the examined articles. The figure highlights a noteworthy trend in the C-XAI field, shifting towards prioritizing interpretable models and explanations aligned with human understanding through the use of concepts. As an example, words like '*Concept*,' '*Human*' and '*Interpret*' are significantly more frequent than '*Feature*' or '*Post-hoc*' explanations.

To summarize, this work provides the following contributions.

- *Key term definition.* We properly define the key terms in the concept-based explanation field, including what a concept is and a concept-based explanation.
- *Taxonomy.* We present a taxonomy that outlines 9 different C-XAI categories based on using concepts during training, the specific type of concept used, and the required explanation.
- *Selection guidelines.* We propose guidelines for choosing a suitable approach, describing each category's applicability, advantages, and disadvantages.
- *Dimensions of analysis.* We identify 13 dimensions of analysis and categorize the C-XAI methods accordingly.

- *Method discussion.* We provide for each method a brief outline of its main aspects and how it differs from the other methods in the same category, including papers of critics and comparisons.
- *Evaluation.* We discuss metrics and datasets for assessing C-XAI methods.
- *Applications.* We highlight some C-XAI applications and promising future directions.

We hope this paper may serve as a reference to a wide spectrum of stakeholders, ranging from AI researchers to policymakers. The discussion of C-XAI methods may help researchers and practitioners to evaluate the different characteristics of explanation methods and models, while the suggested guidelines provide effective strategies for selecting the most appropriate method. Policymakers may also evaluate the accountability and trustworthiness of C-XAI methods for their deployment in safety-critical contexts.

The paper is structured as follows. In Section 2 we define the protocol employed to review the literature, in Section 3 we provide the definitions of the most important terms in the concept-based explanation field, while in Section 4, we define a high-level categorization of C-XAI approaches and provide readers with guidelines for selecting a suitable C-XAI method. Following the previously introduced categorization, in Section 5, we summarize the most important post-hoc concept-based explainability methods, while in Section 6, we review explainable-by-design concept-based models. In Section 7, we report valuable tools developed in the literature to develop and assess novel methods, including metrics, human evaluations, datasets, and resources. Section 8 introduces some applications of concept-based explainability methods and highlights the new emerging trend using and studying foundation models under the C-XAI lens. Finally Section 9 concludes the paper.

2 SYSTEMATIC LITERATURE REVIEW

This section presents the methodology adopted for the Systematic Literature Review (SLR) that forms the empirical foundation of this survey. The SLR protocol ensures transparency, replicability, and methodological rigor in the selection and analysis of literature, supporting our definitions, taxonomy, and practical guidelines for C-XAI. The review was structured around the following research questions, which serve as the main thread of the paper. These questions were designed to systematically explore the landscape of C-XAI and to help readers navigate the breadth of the literature with a clear and coherent structure. For each question, we indicate in parentheses the section where the corresponding answer can be found: (i) **RQ1:** What definitions and typologies of *concept* are proposed in the C-XAI literature? (Section 3.1); (ii) **RQ2:** What types of explanations are provided by C-XAI methods? (Section 3.2); (iii) **RQ3:** What are the methodological characteristics of existing C-XAI approaches (e.g., post-hoc vs. explainable-by-design)? (Section 3.3); (iv) **RQ4:** What taxonomies and dimensions can be used to classify C-XAI methods? (Section 4.1); (v) **RQ5:** What evaluation protocols, metrics, datasets, and resources are employed to assess C-XAI methods? (Section 7); (vi) **RQ6:** What are the main applications and emerging trends of C-XAI methods? (Section 8).

Review Protocol. The review inclusion criteria included peer-reviewed papers published between 2017 and July 2023 that focus on concept-based explainability methods for machine learning. Eligible sources comprise A* AI conferences (e.g., NeurIPS, ICLR, ICML, AAAI, CVPR), Q1 journals (e.g., *Nature Machine Intelligence*, *TMLR*, *Artificial Intelligence*), and influential works cited within these venues. Exclusion criteria ruled out papers that use the term “concept” without operationalizing it within an explainability framework, duplicates, non-English publications, and those without full-text availability. Relevant studies were identified via targeted searches across major digital libraries (IEEE Xplore, ACM DL, SpringerLink, ScienceDirect, Google Scholar), using domain-specific query strings (e.g., `("concept-based" OR`

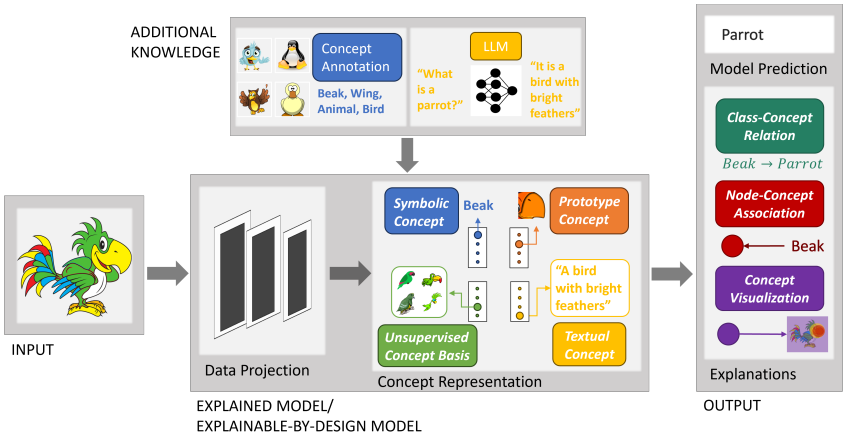

Fig. 2. A pipeline of C-XAI methods and models with commonly employed concepts and provided explanations.

"prototype") AND ("XAI" OR "explainable AI" OR "interpretability")), adapted to each platform. Forward and backward snowballing further extended the search by reviewing citations and references.

Study Selection Process. In order to define the final set of selected papers, we followed a three-phase selection pipeline: first, we conducted an initial screening based on titles and abstracts against the inclusion/exclusion criteria; secondly, we performed a full-test review of shortlisted articles to assess relevance and focus; final inclusion was determined based on consensus among authors. From each included study, we extracted metadata on publication year, venue, data types, task, model architecture, concept definition, explanation type, and evaluation strategy. These attributes were coded along 13 dimensions grouped into: (i) *Concept and Explanation Characteristics* (e.g., concept type, explanation scope), (ii) *Applicability and Implementation* (e.g., data modality, task type, model class), and (iii) *Evaluation and Resources* (e.g., evaluation strategy, metric type, dataset/code availability). We synthesized the data through a combination of qualitative and quantitative methods, along with datasets and resources available.

3 DEFINITIONS

Figure 2 shows a common pipeline of a C-XAI method. The input data is always projected into the latent space of the network, where the method extracts or directly represents a set of concepts For some concept types, additional knowledge may be needed: symbolic concepts may require concept annotations, while textual concepts may require an external Large Language Model (LLM). As an output of the pipeline, different explanations can be generated jointly with the model prediction. We will use Figure 2 and the task of bird classification as a case study throughout the paper. To signal the presence of this recurrent example, we will place a 💡 besides the text.

In the following, we address **RQ1–3**. We first introduce four types of concepts (Section 3.1), followed by three forms of concept-based explanations that offer different types of interpretability (Section 3.2). We then distinguish between post-hoc and explainable-by-design approaches within the C-XAI context (Section 3.3). A glossary of key terms and a list of acronyms used throughout the paper are provided in Appendix A.

Concept Type	Definition
Symbolic Concept	Human-defined attribute or abstraction
Unsupervised Concept Basis	Cluster of similar samples
Prototype	(Part-of) a training sample
Textual Concept	Part-of a textual description of a main class
Explanation Type	
Class-Concept Relation	Relation among a concept and an output class of a model
Node-Concept Association	Explicit association of a concept with a hidden node of the network
Concept-Visualization	Visualization of a learnt concept in terms of the input features

Table 1. Concept and explanation types.

3.1 What is a Concept?

In the concept-based explainability literature, the term *concept* has been defined in very different ways. Citing [73], "A concept can be any abstraction, such as a colour, an object, or even an idea". As summarized in the top part of Table 1, we propose a categorization of concepts into four typologies: *symbolic concepts*, *unsupervised concept bases*, *prototypes*, and *textual concepts*.

Symbolic concepts are human-defined symbols [24]. They can be high-level attributes of the task under consideration or interpretable abstractions, such as the color or the shape of the predicted object [13]. F or instance, Figure 2 shows that a *beak* is a suitable symbolic concept for a bird identification task. Since these concepts are pre-defined by humans, they generally require auxiliary data with concept annotation, particularly when dealing with non-symbolic features (e.g., image pixels). The representation of symbolic concepts in the network can be analyzed post-hoc or forced during training to create an explainable-by-design model.

Unsupervised concept bases are clusters of samples the network learns. Even if they are not built to resemble human-defined concepts, these unsupervised representations may still capture abstractions more understandable to humans than individual features or pixels [35]. A s shown in Figure 2, a network may learn a cluster of green birds when classifying a green-colored bird species. To extract this concept, a clustering algorithm may be employed, either post-hoc in the latent space of the model or during training.

Prototypes are representative examples of peculiar traits of the training samples. They can be training samples or only parts of a training sample. A s shown in Figure 2, the pattern of a hooked beak could be a useful part prototype to classify parrots. The set of prototypes is, in general, representative of the whole data set [61]. Prototypes are different from unsupervised concept bases because they have to be explicitly encoded in the network weights, and consequently, they can only be employed in explainable-by-design models. Following recent literature [68], we regard prototypes as concepts because they are still higher-level terms than input features.

Textual concepts are derived from textual descriptions of the classes. From an individual description, distinctive pieces are extracted, each of which embodies a characteristic of the corresponding class. A lways considering Figure 2 and our bird classification task, 'A bird with bright feathers' could be an important textual concept to classify a sample of a parrot. Textual concepts can be provided at training time by means of an external generative model [75], and they are employed inside a concept-based model in the form of a numerical embedding of the text. This type of concept is gaining prominence due to the recent development of Large-Language Models (LLMs [21], [32]). So far, it has only been employed in explainable-by-design models. For a more formal definition of Concept following category theory principles, please refer to Appendix B.

3.2 What is a Concept-based Explanation?

Even when considering a fixed concept type, the definition of a concept-based explanation can be elusive. Existing literature generally agrees that concept-based explanations should explain how DNNs make particular decisions using concepts [108] and adhere to specific criteria such as being meaningful, coherent, and relevant to the final class [35] and also explicit and faithful [9]. As we report in the bottom part of Table 1, we define three distinct categories of concept-based explanations: *class-concept relationship*, *node-concept association*, and *concept visualization*. Each method may offer one or more explanations, serving specific purposes in different scenarios.

Class-concept relationship considers the relationship between a specific concept and an output class of the model. This relationship can be expressed either as the importance of each concept for a particular class or through the use of a logic rule that involves multiple concepts and their connection to an output class. F igure 2 shows this type of explanation through a logical relation: *Beak* $\rightarrow$ *Parrot* since the input image is classified as a Parrot due to the presence of the symbolic concept *Beak*. This explanation can be extracted by analyzing post-hoc the correlations of concepts and classes in the latent space of a pre-trained network or employing an interpretable model from the concepts to the tasks.

Node-concept association explicitly assigns a concept to an internal unit (or a filter) of a neural network. This explanation enhances the transparency of deep learning models, highlighting what internal units see in a given sample. It can be defined post-hoc by considering the hidden units maximally activated on input samples representing a concept. Otherwise, it can also be forced during training by requiring a unit to predict a concept. R ecalling our example in Figure 2, a network node can automatically learn to recognize the beak since it is a crucial part of the prediction of a bird. Through a post-hoc analysis, we can identify it and associate the node with the concept. Otherwise, an expert can also define the concepts in advance and require their explicit representation in the network intermediate layers.

Finally, ***Concept visualization*** explanations highlight the input features that best represent a specific concept. When symbolic concepts are used, this explanation closely resembles the saliency map of standard XAI methods. However, when non-symbolic concepts are employed, the focus shifts towards understanding which unsupervised attributes or prototypes the network has learned. This form of explanation is often combined with one of the previous explanations, enabling the understanding of the concepts associated with a specific class or node. F or instance, consider the concept visualization explanation reported in Figure 2. A visualization of the concept, along with the parts identified as similar, is crucial for understanding which prototype the network has learned.

3.3 C-XAI Post-hoc vs Explainable-by-design Methods

In the XAI literature, the difference between *post-hoc explanation methods* and *explainable-by-design models* is well-defined: the former are tools applied post-training to explain complex models, such as DNNs; in contrast, the latter are models inherently designed for transparency, such as decision trees or linear models. On the contrary, in concept-based approaches, the difference becomes more subtle since C-XAI explainable-by-design models also work with DNNs but explicitly represent some concepts within the architecture. In the following, we try to clarify this difference.

Post-hoc concept-based explanation methods operate on concepts rather than input features. They explain a prediction, an entire class, or an internal network node, given a set of concepts. Typically, this involves projecting the samples representing the concepts in the model's latent space and analyzing their relationship to the prediction and the hidden node activations. Symbolic and unsupervised concept bases have been employed to provide all types of explanations. R ecalling the example in Figure 2, we may need to utilize a pre-existing bird classification model.

Nonetheless, we might wish to interpret this model using concepts, such as determining how a particular output class relates to the supervised concepts of beak shapes, wing colors, and feather types.

Explainable-by-design concept-based model, on the other hand, are neural models employing an explicit representation of a set of concepts as an intermediate layer. This way, the predicted concepts influence the task predictions. Since they generally provide node-concept association by-design, they can be regarded as inherently transparent models, at least to some extent. In this case, the whole spectrum of concepts can be used, either explicitly annotated, extracted in an unsupervised manner, encoding prototypes, or even generated by a separate model. R eferring once more to Figure 2, should a bird classification model be developed from scratch, transparency can be integrated into its decision process by requiring that a given layer extract concepts. For instance, a layer extracting bird prototypes that might be used to interpret the resultant species classification.

4 C-XAI TAXONOMY

In this section, we provide a taxonomy of concept-based methods together with guidelines for selecting an appropriate method considering development constraints, the application of interest, and the desired outcomes. In Section 4.1, we address **RQ4** by introducing the dimensions of analysis used to categorize and characterize C-XAI methods. Section 4.2 builds on this by presenting a strategy to select appropriate C-XAI methods based on specific requirements.

4.1 Dimensions of analysis

We identify 13 dimensions of analysis to characterize C-XAI methods. We categorize these dimensions into three groups: (i) concept and explanation characteristics, (ii) applicability of the method, and (iii) resources employed. The dimensions related to the first two groups will be analyzed in Sections 5 and 6, and reported in Tables 2 and 3, respectively. The third group of dimensions will be analyzed vertically (per resource/evaluation) in Section 7, but we also report a horizontal analysis (per method) in Section 7.4, and in Appendix C.3 in Tables 6 and 7.

(1) **Concept and Explanations Characteristics**. These dimensions capture key aspects of concepts and explanations, such as how concepts are integrated into models, concept annotation strategies, concept types, the form of explanations, and their scope. Specifically, this group includes:
   - *Concept Training*. Reviewed methods either employ concepts only to provide explanations of an existing model (Post-hoc methods) or while training the same model (Explainable-by-design models).
   - *Concept Annotation.* A method may employ an annotated set of concepts (Supervised), it may not employ it (Unsupervised), it may employ only a few annotated concepts (Hybrid), or it may generate annotations (Gen.).
   - *Concept Type.* It is described in Section 3.1. It can be either symbolic (Symb.), unsupervised concept bases (Uns. Basis), prototype-based (Proto.), or textual (Textual).
   - *Explanation Type.* It is defined in Section 3.2. It can be class-concept relationship (C-CR), node-concept association (N-CA), concept visualization (C-Viz), or a combination of them.
   - *Explanation Scope.* The explanation may be local (LO) (i.e., it explains an individual prediction), global (GL) (i.e., it provides insights into the overall model behavior), or both.

   We also characterize Concept-based Models with the following information.
   - *Concept Employment.* It describes how an explainable by-design supervised model employs the concepts during training, either through joint training (Joint) or concept instillation (Instill.).

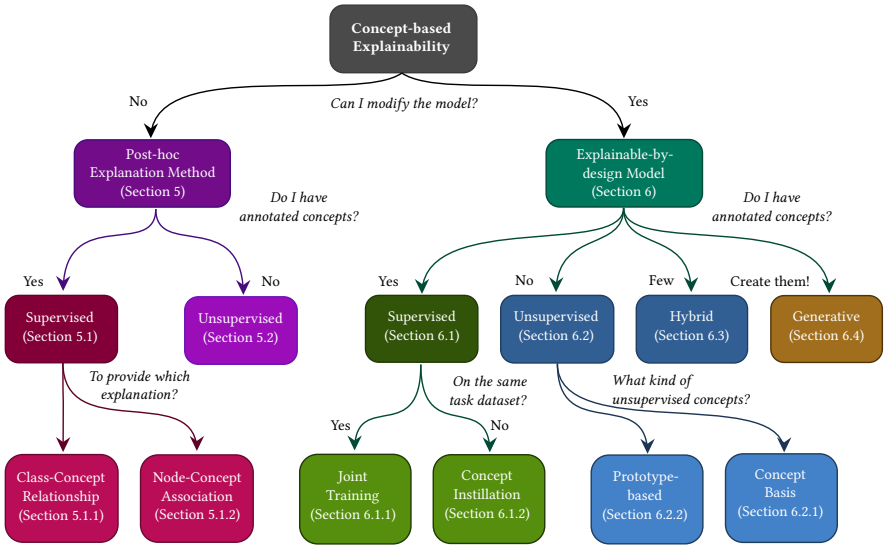

Fig. 3. Categorization of the C-XAI methods with guidelines for selecting a suitable approach.

- *Performance Loss.* It indicates if there is a performance loss (✓) or not (✗) compared to a black-box model.

(2) **Applicability of the approach**. These dimensions describe the approach's applicability by detailing the types of data it can handle, the primary tasks it is designed for, and the specific neural architectures utilized.

- *Data.* The data type the method can handle, generally images (IMG), but also text (TXT), graph (GRA), tabular data (TAB), videos (VID) and time series (TS).
- *Task.* The task the method is designed for, i.e., classification (CLF), regression (REG), or clustering (CLU).
- *Network.* It describes the type of neural network employed, including Convolutional Neural Network (CNN), 3D-CNN, Auto-Encoder (AE), Variational Auto Encoder (VAE), Graph Neural Network (GNN), Multi-Layer Perceptron (MLP), Large Language Model (LLM), and Transformer (TRANSF).

(3) **Resources and Evaluation conducted**. These dimensions focus on the resources provided (data releases and code/models availability) and the conducted evaluation.

- *Data Release.* It indicates if a new dataset annotated with concepts has been made available.
- *Code/Models Availability.* It indicates if the authors have released the code and/or the model.
- *New metric.* It indicates if the paper introduces a new evaluation metric.
- *Human evaluation.* It indicates if the paper includes a user study to evaluate the method.

4.2 Guidelines for selecting a suitable method

We provide the following guidelines employing Figure 3 as a visual reference. Let us consider a user interested in a C-XAI method. We propose a set of maximum three questions that he/she should answer to choose a category of methods for the application at hand. These questions regard the possibility of modifying the learning model, the availability of an annotated set of concepts, and, according to the subcategory, the desired outcome regarding explanation, concept type, or the training set employed.

Table 2. Post-hoc concept-based explainability methods. We characterize the approaches based on the employed concepts, explanations and their applicability. A full description of each category and of the acronyms is provided in Section 4.1.

Concept Annotation		Method	Concept Type	Expl. Type	Expl. Scope	Data Type	Task	Network Type
	Supervised	T-CAV [50]	Symb.	C-CR	GL	IMG	CLF	CNN
		CAR [25]		C-CR	GL	IMG + TS	CLF	CNN
		IBD [115]		C-CR, C-Viz	LO & GL	IMG	CLF	CNN
		CaCE [37]		C-CR	GL	IMG	CLF	CNN
		CPM [102]		C-CR	LO	TXT	CLF	TRANSF
		ND [14]	Uns. Basis	N-CA, C-Viz	GL	IMG	CLF	CNN
		Net2Vec [33]		N-CA, C-Viz	GL	IMG	CLF	CNN
		Comp. Exp. [74]		N-CA, C-Viz	GL	IMG, TXT	CLF	CNN, TRANSF
		GNN-CI [104]		N-CA, C-CR, C-Viz	GL	GRA	CLF	GNN
	Unsupervised	ACE [35]		C-CR, C-Viz	GL	IMG	CLF	CNN
		Compl. Aware [108]		C-CR	GL	IMG, TXT	CLF	CNN
		ICE [112]		C-CR, C-Viz	LO & GL	IMG	CLF	CNN
		MCD [99]		C-CR, C-Viz	LO & GL	IMG	CLF	CNN, TRANSF
		CRAFT [31]		C-CR	GL	IMG	CLF	CNN
		DMA & IMA [58]		-	-	IMG	CLF	CNN
		STCE [47]		C-CR, C-Viz	GL	VID	CLF	3D-CNN

Can I modify the model? This is the first coarse-grained question the user should answer. According to the possibility of intervening in the model's development, he/she can choose among two C-XAI macro-categories. The user should look at *Post-hoc* concept-based explanation methods (Section 5) if he/she needs to consider an already trained model that he/she cannot (or does not want to) modify. On the other hand, if the user can create a model from scratch, he/she can also employ an *Explainable-by-design* concept-based models (Section 6), which allows an explicit representation of the concepts within the model architecture.

Do I have annotated concepts? The user then should check whether he/she can find a dataset annotated with concepts related to the task at hand. In case of a positive response, a user could look at *supervised* methods. Otherwise, he/she should consider *unsupervised* approaches that can extract concepts from the same data automatically. For explainable-by-design approaches, a user can pick from two more categories: he/she can employ a *hybrid* solution that leverages a few supervised concepts and unsupervised (extracted) ones; he/she can also exploit a *generative* method to create concepts by means of an external model.

To provide which Explanation? If the user selects a post-hoc supervised approach, this class of methods further differs in the type of explanation, either class-concept relationship or node-concept association.

On the same task Dataset? Explainable-by-design supervised approaches may require concept annotation on the same dataset of the task at hand, or allow them on a separate one.

What kind of unsupervised concepts? Finally, unsupervised explainable-by-design approaches differ in the adopted representation of unsupervised concepts.

The methods in the remaining sub-categories share these main characteristics. Hence, we have not split them further.

5 *POST-HOC* CONCEPT-BASED EXPLANATION METHODS

Post-hoc concept-based explanation methods explain an existing model without modifying its internal architecture. Table 2 categorizes post-hoc concept-based explainability methods based on whether they rely on concept-annotated datasets (supervised, Section 5.1) or not (unsupervised, Section 5.2). Supervised methods use symbolic concepts

and either explain class-concept relationships (C-CR) or node-concept associations (N-CA), often also providing concept visualizations (C-Viz). Unsupervised methods extract latent concept bases and typically focus on class-level explanations and visualizations. Most methods offer global (GL) explanations, though some support both local and global (LO&GL), with one method providing only local (LO) explanations. All reviewed approaches address classification (CLF) tasks, predominantly for images (IMG) using CNNs, while a few extend to text (TXT), graphs (GRAPH), or videos (VID), employing transformers, GNNs, or 3D-CNNs. The table summarizes key strengths and limitations across these dimensions.

Advantages. Post-hoc explanation methods are the only viable option in scenarios where a pre-trained model exists or a predefined model is necessary. Also, concept-based explainability methods are the preferred choice when there can be no compromise on the predictive and generalization capabilities of the model. They allow for enhanced interpretability without affecting the model performance.

Disadvantages. The main drawback of these methods is that they do not guarantee that the model truly comprehends or employs the adopted concepts. This is because the model was not exposed to these concepts during its training process. Concept-based explainability methods can only be regarded as a better way of explaining the network behavior by employing terms that are more comprehensible to humans than raw input features. We will further discuss the issues and challenges post-hoc methods face at the end of this section (Section 5.3).

5.1 Post-hoc Supervised Concept-based Methods

Post-hoc supervised concept-based explanation methods analyze network behavior over samples annotated with *symbolic concepts*. The underlying idea is that a network automatically learns to recognize some concepts. These methods assess which concepts have been learned, where, and how they influence the model. This analysis is conducted differently according to the type of explanations delivered. In particular, to provide class-concepts relationships, some methods analyze how the current prediction or an entire class weight correlates with the projection of a concept in the latent space of the model (Section 5.1.1). On the contrary, methods extracting explanations in terms of node-concept associations study the activations of single hidden nodes to associate them to a given concept (Section 5.1.2).

5.1.1 Class-concept relationship methods. Supervised post-hoc methods examine the connection between output classes and a set of symbolic concepts. As illustrated in Figure 4, this is generally achieved by presenting samples representing concepts to the explained network and analyzing its behaviour. In the figure, we report the case of a model trained to recognize birds (e.g., a Parrot) and tested with samples representing a related concept (e.g., the Beak). The effect of the concepts on the final class is studied by either (i) analyzing the representation of concepts in the model's latent space and their relationship with the output classes (TCAV, CAR, IBD) or (ii) exploring the causal effect of inserting or removing a concept on the prediction of a particular class (CaCE, CPM). The samples representing the symbolic concepts are drawn from an external dataset, which is required for this class of methods. As previously mentioned, the concepts must be related to the output classes to extract meaningful explanations.

Testing with Concept Activation Vector **(T-CAV [50])** has been the first work in this category. For each user-defined concept, it requires a positive set of images representing the concept and a complementary one with negative samples. T-CAV (i) freezes the model's internal architecture and (ii) trains a linear layer (probe) in the latent space of the model to predict the concepts. The Concept Activation Vectors (CAVs) represent the orthogonal vectors to the weights of the linear probe, separating positive examples of a given concept from negative ones. Then, T-CAV computes the inner product between the directional derivatives of the class and the CAVs to quantify the models' conceptual sensitivity, i.e.,

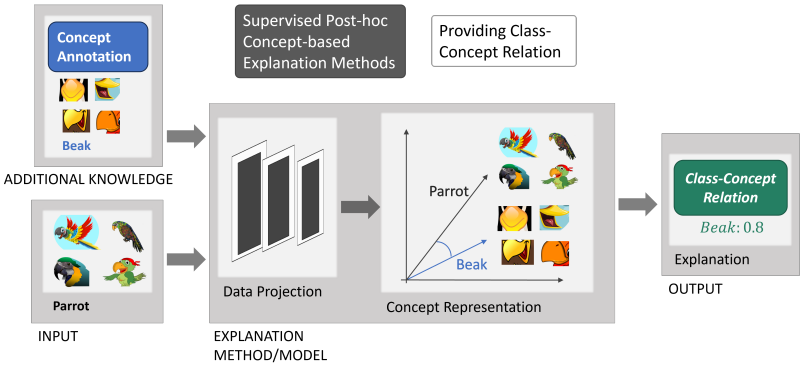

Fig. 4. Post-hoc Supervised Methods providing explanations in terms of class-concept relationships. These approaches provide a set of samples annotated with concepts to the explained network to determine their influence on the output class.

how much each concept positively influences the prediction of a given class. The T-CAV score is the fraction of the class's inputs with a positive conceptual sensitivity. The higher the score, the higher the influence of the concept. As noted by the authors, CAVs require concepts to be linearly separable. However, concepts may be semantically linked and mapped to latent space portions that are close to each other. In this case, the authors propose Relative CAVs, i.e., CAVs obtained when considering as negative samples the positive samples of other concepts. The authors showed that T-CAV better enables humans to identify the most important concepts for a network with respect to standard XAI methods.

Concept Activation Regions **(CAR [25])** relaxes the linear separability assumption of concepts imposed by T-CAV. It only requires concept examples to be grouped in clusters in the model's latent space such that they are identifiable by a non-linear probe (e.g., a kernel-based SVM). Each concept is thus represented by a *region* in the latent space and no longer by a single vector. A concept is important for a given example if its projection lies in the region of concept positive representations. Equivalently to the TCAV score, the authors propose the TCAR score, quantifying the proportion of instances belonging to a specific class whose representations fall within the region of a given concept. In some settings, TCAR returns high scores expressing significant associations between classes and concepts, while TCAV returns low scores. The authors claim that this is due to employing a more powerful concept classifier and, consequently, that TCAR scores reflect more accurately the relationships between classes and concepts.

Interpretable Basis Decomposition **(IBD [115])** provides class-concept relationships by decomposing the evidence supporting a prediction into the most semantically related components. As for T-CAV, to represent the concepts, IBD trains a set of linear classifiers over the sample projections in the latent space. IBD then performs an optimization to assess the relevance of each concept for a class. For all samples in the concept dataset, it tries to reconstruct the class prediction using a linear regression. In this, the variables are the inner product between each concept vector and the projection of the sample in the latent space. Instead, the coefficients represent each concept's associated importance for the given class. Interestingly, the *residual* of the optimization assesses the variability in the class predictions that cannot be represented by means of the given concepts. Each concept is accompanied by a heat map showing the features with the highest impact on the concept prediction. By means of human evaluation, the authors demonstrated that IBD explanations enhance standard XAI techniques (like CAM and Grad-CAM) regarding model assessment quality.

The Causal Concept Effect **(CaCE [37])** method investigates the causal effect made by the presence or absence of a symbolic concept on a network prediction. The CaCE measure is defined as the effect of the presence of the binary concept on the classifier's output. To measure how much a concept influences the classifier's output, they propose generating counterfactual samples by operating on the original input and removing the concept, either by (i) directly intervening in a controlled generation process to compute the exact CaCE, or (ii) training a generative model to approximate this process and produce samples with and without the concept. In the second scenario, a Variational Auto-Encoder (VAE) architecture can be used in two different variants, either generating images with a concept, or generating the counterfactual of an image with/without a certain concept. The authors compare CaCE with a non-causal baseline and with TCAV [50], and show that CaCE better understands and identifies when the concepts have a causal relationships with the class prediction.

Similarly to CaCE, the Causal Proxy Model **(CPM [102])** provides causal concept-based explanations, this time tailored for Natural Language Processing (NLP) models. CPM builds upon the CEBaB benchmark [1], a dataset studying the effects of concept *presence/absence on textual data*. Starting from factual restaurant reviews, it provides counterfactual examples in which a concept (food, service, ambiance, etc.) has been modified. CPM is first initialized with the weights of the black-box model to explain. Then it is trained to mimic the behavior of the black-box model for both factual and counterfactual samples. Since the latter is not provided at test time, CPM is still prompted with the factual samples and information from the counterfactual ones for counterfactual training. CPM explanations are compared with TCAV [50], the baseline of CaCE [37], and others. They measure the distance between the model prediction variation when prompted with a counterfactual sample and the concept relative importance provided by the explainers. CPM results in the best method for assessing the causality of a concept.

5.1.2 Node-concept association methods. Supervised post-hoc methods may also associate symbolic concepts with internal nodes or filters of the neural model. In this case, the idea is to understand where the model has automatically learned some concepts. This allows the clarification of the decision process of a given model. As shown in Figure 5, these methods give node-concept association explanations by checking which nodes activate the most when the network is presented with input images representing the concepts (e.g., the beak) (ND, Net2Vec) or input graphs (GNN-CI). This can also be done through human annotation, where workers label neuron activation patterns with semantic concepts, as in early object detector studies [113]. The first case requires less effort, but its effectiveness relies on the quality of the underlying dataset. If a network unit corresponds to a concept understandable to humans but not present in the given dataset, it will not be detected.

In Network Dissection **(ND [14])**, the authors propose a new dataset, BRODEN (BROad and DENsely labeled dataset) [7], which includes concept labels from diverse data sources. To associate each node with a specific concept, the authors compute the Intersection over Unions (IoU) between the activation maps of the node and the concept bounding boxes reported in the BRODEN dataset. The node is then associated with the concept with the highest IoU. In output, ND reports the units with the highest average IoU and the samples with the highest overlap for each concept (we can have many units associated with the same concept). As a result, they show that individual units automatically learn to represent high-level semantic objects, concluding that the units of a deep representation may be more interpretable than expected.

The work Network to Vector **(Net2Vec [33])** argues that semantic representations of concepts might be distributed across filters. Hence, filters must be studied in conjunction (not individually as in ND). For each concept, Net2Vec first

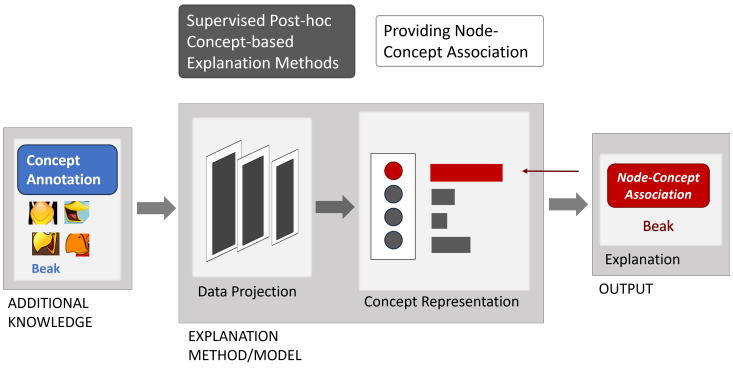

Fig. 5. Post-hoc supervised methods providing an explanation in terms of Node-Concept Associations. This class of approaches associates concepts with internal nodes or filters of the network to increase the transparency of the network decision-making process.

collects the filter activations of the network to explain. Then, Net2Vec performs an optimization process to classify or segment the same concept starting from these activations. The resulting weights tell us how much each node is relevant for predicting the given concept. The results show that combining filters allows for better concept identification than using a single filter in both tasks. Moreover, filters are not concept-specific, but they can encode multiple concepts. As a limitation, Net2Vec models the relationship between concepts and filters as linear, potentially failing to capture more complex and non-linear concept alignments.

While ND maps neurons to a single atomic concept, focusing on single-purpose neurons, Compositional Explanations of Neurons (**Comp. Exp.** [74]) seeks logical combinations of concepts to approximate neuron activations, enabling the description of *polysemantic neurons*—neurons responsive to multiple, potentially unrelated concepts. CompExp achieves this by composing single concepts using logical operators (AND, OR, NOT) via a beam search approach, offering more structured explanations and a higher mean IoU. Like ND, it operates at a global level, systematically analyzing neuron activations across large datasets. It applies to CNNs in image classification, utilizing BRODEN concepts, and Transformer-based architectures in NLP, where linguistic features (e.g., word categories, part-of-speech tags) serve as atomic concepts for describing hidden representations.

In Graph Neural Networks Concept-based Interpretability **(GNN-CI [104])**, the authors extend the idea of node-concept explanations to the graph classification scenario. They define concepts either as properties of a node in the graph (e.g., number of connected edges > 7) or as the class of the node (being an Oxygen atom). For each neuron, they check whether its activation resembles the presence of any concepts. The resemblance is computed for all samples as the IoU between the concept presence and the neuron activation. They allow each neuron to be described with a short composition of concepts. Furthermore, GNN-CI produces class-concept relations with a linear classifier trained only on explained nodes to mimic the original classifier. Finally, it also produces concept visualizations with the graph concept activation map technique. The experiments report that extracted concepts are meaningful for graph classification. For instance, they reveal the presence of functional groups known to be correlated with the graph class.

5.2 Post-hoc Unsupervised Concept-based methods

Relying on a given set of supervised concepts is not the only way to assess whether a network is learning higher-level semantics. Indeed, parts of the samples may represent practical explanations of a given prediction or a learned class. As

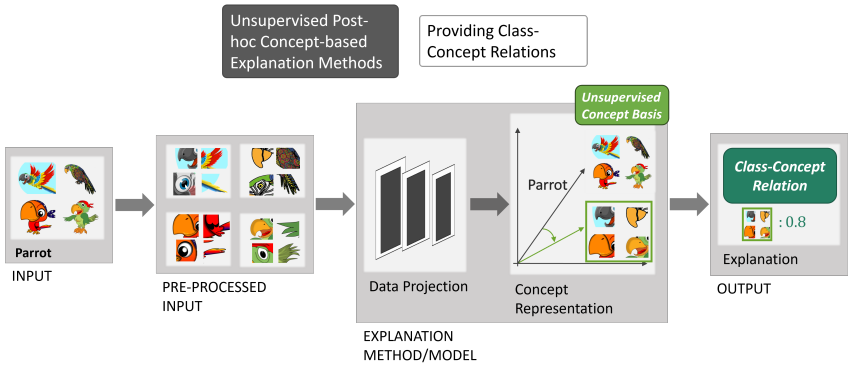

Fig. 6. Unsupervised post-hoc methods extract unsupervised concepts regarding parts of the same input samples. In particular, these methods provide class-concept relationships analyzing the representations of these parts with respect to the output class in the latent space of the network.

shown in Figure 6, input pre-processing is normally necessary for these methods. Afterward, unsupervised methods analyze sample representation in the latent space to identify clusters composed of parts of the samples (i.e., *unsupervised concept basis*). Also, they analyze how these clusters contribute to the final prediction, thereby providing explanations in terms of Class-concept relationships. How the clustering is conducted differentiates the methods: they either employ standard k-means (ACE), Non-Negative Matrix Factorization (ICE), recursion on different layers (CRAFT), subspace clustering (MCD), concept completeness maximization (Completeness-aware), or assuming concept independency, or, finally, working on voxels from videos. This category bears similarities to standard XAI methods. However, they differ in not selecting features relying solely on their saliency or internal similarity (as with superpixels) but rather on their capability to represent common patterns existing in other input samples.

The Automatic Concept-based Explanations **(ACE [35])** method explains a class within a trained classifier, employing a probing strategy that does not require human supervision. The process requires first segmenting images belonging to a specific class at multiple resolutions. Then, the segments are projected in the latent space of the network and clustered across different samples using K-means. The T-CAV importance score is calculated for each cluster (i.e., an unsupervised concept) to estimate the average influence of a cluster on the class predictions. In summary, to explain a prediction, ACE outputs the most pertinent concepts together with their T-CAV scores. With a human-based evaluation, the authors show that ACE coherently assigns samples to a concept and that, on average, the segments of image belonging to a same concept, can be labelled employing the same words.

While ACE allows the extraction of some concepts from a learned neural network, there is no guarantee that this set of concepts is sufficient to explain a prediction. In **Completeness-aware [108]**, the authors propose a method for extracting *complete* concept-based explanations. The authors define the completeness score as the extent to which a set of concepts can predict the model's output, measured by the accuracy when using only the concept scores. Their discovery algorithm is optimized to uncover maximally complete concepts. They also introduce a metric called ConceptSHAP, which adapts SHAP by replacing feature importance with concept sufficiency in representing a class. On a synthetic dataset, Completeness-aware outperforms ACE in identifying a complete set of concepts.

Invertible Concept-based Explanation **(ICE [112])** extends ACE by improving concept identification and importance assessment. ICE replaces K-means with Non-Negative Matrix Factorization (NMF) to extract concept vectors from feature

maps, associating each reduced dimension with a distinct concept. A classifier is then trained on these representations to approximate model outputs, with concept importance derived from the linear weights—replacing T-CAV scores. This NMF-based representation yields lower approximation error than PCA or K-means. Also, ICE concepts can be more consistently characterized by users, as attested by a user study.

Concept Recursive Activation FacTorization for Explainability **(CRAFT [31])** extends ACE and ICE for unsupervised concept discovery. It identifies input sub-regions via random crops. The hidden network activations are then factorized using NMF to extract concept vectors. Since concepts can emerge at different layers, CRAFT introduces a recursive mechanism to capture both concepts and sub-concepts across layers. Concept importance for the final prediction is quantified using *Sobol* scores from sensitivity analysis. Additionally, CRAFT visualize concepts by means of saliency maps of the concept representations. A Human evaluation reports CRAFT's higher utility than ACE and standard XAI methods like Saliency and GradCAM.

Multi-dimensional Concept Discovery **(MCD [99])** allow concepts to span across different convolutional channel directions. This is obtained using Sparse Subspace Clustering (SSC) over hidden representation and a subsequent PCA to derive concept activations. A linear layer is then trained (as in ICE) to approximate the model's output, reporting a per-sample concept completeness score. MCD offers two explanations: concept relevance, indicating a concept's importance for class predictions, and concept activation maps, for concept visualization. MCD improves concept faithfulness and conciseness over ICE and ACE, defined as prediction reduction when altering key concepts and the number of concepts needed for a completeness score, respectively.

DMA & IMA [58] aim to ensure *identifiability*, defined as the provable recovery of the underlying concept generating the data, at least in controlled settings. Disjoint Mechanism Analysis (DMA) ensures identifiability of generating concepts, even if ground truth concepts are correlated (unlike PCA). However, it requires concepts to affect disjoint parts of the input. Independent Mechanism Analysis (IMA) relaxes this constraint, requiring concept vectors to be only orthogonal but not disjoint. IMA and DMA yield results comparable to PCA or ICA on datasets without correlated generative components, and better otherwise. Also, discovered concepts represent ground truth concepts more often than ACE and Completeness-aware.

Spatial-temporal Concept-based Explanation **(STCE [47])** provides concept-based explanations for 3D ConvNets, a domain less explored due to the computational complexity of handling video data. STCE segments videos into supervoxels for each class, extracts features with a 3D ConvNet, and then clusters similar supervoxels to define unsupervised spatial-temporal concepts. Concept importance is determined using T-CAV scores, considering, as negative samples, random videos from non-related datasets. Similarly to previous works, the quality of the explanations is assessed in terms of faithfulness. Also, for qualitative analysis, the authors present video frames that feature the most and least important concepts for a given classification. The result highlights the consistency of STCE.

5.3 Discussion

In the following, we report the critics that have been moved to the concept-based methods reviewed in this section and how, in particular, they affect their performance.

Do standard networks learn concepts? Several works in the literature [23, 54] discuss whether standard models actually learn symbolic concepts when they are not explicitly trained to represent them. In particular, [54] tested the concept accuracy of a linear probe. They found that linear probes achieve a lower concept accuracy with respect to a model trained to predict the concepts (91% vs 97% on the CUB dataset [101] and 0.68 vs. 0.53 error on the OAI

dataset [77]). This result confirms that post-hoc concept-based explainability methods cannot ensure that the model actually learns symbolic concepts. However, they do help to understand what the network is learning: the still high accuracy obtained by the linear probes implies that the model has learnt many concepts.

Different probe datasets provide different explanations. In [82], the authors show that different probe datasets lead to different explanations even when explaining the same model with the same method. This issue arises for both methods providing node-concept associations (ND) and class-concept relationships (TCAV). They show ND may label neurons differently according to the dataset employed. While in some cases, the concepts are similar (e.g., *plant* vs *potted-plant*, *computer* vs *tv*), in other cases they are different (e.g., *chair* vs *horse*, *tent* vs *bus*). Also, they show that T-CAV vectors extracted from different datasets may point in different directions. To numerically assess this, they compute the cosine similarity between the vectors representing the same concept when extracted from different datasets. While for some concepts the similarity score is quite high (*ceiling* 0.27), for others, it is very low (*bag* 0.01, *rock* -0.02), with a similarity lower than 0.1 on average. As we also highlighted, they suggest carefully choosing the probe dataset and specifically recommend probe datasets to resemble the data distribution of the original training set closely.

Post-hoc concept-based explanation methods are vulnerable to adversarial attack. Concept-based post-hoc explainability methods, much like their standard XAI counterparts [5, 34, 53], are susceptible to adversarial perturbations [20, 71]. In [20], the authors show how a sample can be crafted maintaining the same classification while changing its interpretation. For example, an attack can make concept relevant for a class, such as making a *Corgis* important for the classification of a *Honeycomb*, or, conversely, can reduce the importance of an important concept, such as making *Stripes* not important for classifying a *Zebra*. These manipulations remain effective even in black-box settings, where attackers cannot directly access the model, but only a surrogate version. Instead, the authors in [71] analyze how adversarial attacks altering the final classification affect unsupervised concept representations extracted post-hoc from a neural networks. Comparing concept composition in clean sample representations to ones in adversarially attacked, they show that concept saliency maps, concept weights ranking and concept similarities get modified.

6 *EXPLAINABLE-BY-DESIGN* CONCEPT-BASED MODELS

Explainable-by-design concept-based models explicitly represent concepts within their neural network architecture. We define them as explainable-by-design because they inherently provide node-concept associations (N-CA) (except for a few cases), typically through a dedicated hidden layer predicting concept scores. These scores represent numerical values quantifying the relevance or presence of specific concepts in a given input sample and condition the model's output. The same concept scores can also be studied in relation to the output classes to define class-concept relationships (C-CR) and can be visualized through concept-visualization (C-Viz).

The way the intermediate representation is defined varies according to the concept annotations and, consequently, the concept types. As shown in Table 3, concept-based models either employ a dataset with (Supervised, Section 6.1), or without annotated concepts (Unsupervised, Section 6.2), with annotations for few concepts only (Hybrid, Section 6.3) or they generate the supervision by means of an external model (Generative, Section 6.4). Furthermore, if they employ annotated symbolic concepts (Symb.), they differ according to whether they employ them during training (Joint) or afterward to instill them in a trained network (Instill.). If they automatically extract concepts, they differ in the type of concepts extracted, either clusters (Uns. Basis) or prototypes (Proto.). If they adopt a hybrid approach, they generally employ both symbolic and unsupervised basis concepts. Finally, if they generate concept annotations, they are normally textual concepts. Concept-based models have been mostly developed to solve classification tasks (CLF), with one also

Table 3. Explainable by-Design Concept-based Models. We characterize the approaches based on the employed concepts, explanations and its applicability. A full description of each category and of the acronyms is provided in Section 4.1.

Concept Annotation		Method	Concept Employ.	Concept type	Scope	Expl Type	Data type	Loss	Task	Network Type
Supervised		CBM [54]	Joint	Symb.	N-CA, C-CR	GL	IMG	✓	CLF, REG	CNN
		LEN [13, 24]			N-CA, C-CR	LO & GL	TAB, IMG, TXT	✓	CLF, CLU	FCN, CNN
		CEM[30]			N-CA	GL	TAB, IMG	✗	CLF	FCN, CNN
		ProbCBM [51]			N-CA	GL	IMG	✗/≈	CLF	CNN
		DCR[12]			N-CA, C-CR	LO	TAB, IMG, GRA	✗	CLF	FCN, CNN, GNN
		SparseCBM [98]			N-CA, C-CR	LO	TXT	✓	CLF	TRANSF
		CW [23]	Instill.	Symb.	N-CA	GL	IMG	✓	CLF	CNN
		CME [49]			N-CA, C-CR	GL	IMG	✓	CLF	CNN
		PCBM [109]			N-CA, C-CR	GL	IMG	✓/≈	CLF	CNN
		CT [86]			N-CA, C-CR	LO & GL	IMG	✗	CLF	TRANSF
Unsupervised		Interp. CNN [111]	-	Uns. Basis	N-CA, C-Viz	GL	IMG	✓/≈	CLF	CNN
		SENN [9]			N-CA, C-CR	LO	IMG	✗/≈	CLF	CNN + AE
		BotCL [100]			N-CA, C-CR	LO	IMG	✗/≈	CLF	CNN + AE
		SelfExplain [81]			C-CR	LO & GL	TXT	✗	CLF	TRANSF
		PrototypeDL [61]	-	Proto.	N-CA, C-CR, C-Viz	GL	IMG	✗	CLF	AE
		ProtoPNet [22]			N-CA, C-CR, C-Viz	GL	IMG	✓	CLF	CNN
		ProtoPool [88]			N-CA, C-CR, C-Viz	GL	IMG	✗	CLF	CNN
		Def. ProtoPNet [27]			N-CA, C-CR, C-Viz	GL	IMG	✗	CLF	CNN
		HPNet [41]			N-CA, C-CR, C-Viz	GL	IMG	✗	CLF	CNN
		ProtoPShare [89]			N-CA, C-CR, C-Viz	GL	IMG	✗	CLF	CNN
		ProtoPDebug [18]			C-Viz	GL	IMG	-	CLF	CNN
Hybrid		CBM-AUC [91]	-	Uns. Basis, Symb.	N-CA, C-CR	GL	IMG + VID	✗/≈	CLF	CNN
		Ante-hoc [90]			N-CA	LO & GL	IMG	✓	CLF	CNN + AE
		GlanceNets [68]			N-CA, C-CR	GL	IMG	✗	CLF	CNN + VAE
Gen.		LaBO [107]	-	Textual	C-CR	GL	IMG	✗/≈	CLF	CNN+LLM
		Label-free CBM [75]			C-CR	LO & GL	IMG	✓	CLF	CNN+LLM

performing regression (REG) and another clustering (CLU). Similar to the post-hoc method, in this case, the network's backbone is mostly a CNN working on images (IMG), possibly reconstructing the data with an auto-encoder (AE). Some works employ a fully-connected network working on tabular data (TAB), GNN on graphs, and transformer (TRANS) working on both images and texts (TXT). To summarize the key aspects of concept-based models, we now discuss their main advantages and disadvantages compared to post-hoc methods.

Advantages. Concept-based models offer the advantage of ensuring that the network learns a set of concepts explicitly. Furthermore, a domain expert can modify the predicted concepts and observe how the model's output changes in response. This process is referred to as *concept intervention* [54], enabling an interaction with the model and the generation of counterfactual explanations.

Disadvantages. On the other hand, these methods can only be employed when training a model from scratch, possibly tailoring it to the specific task. Furthermore, in simpler solutions, the predictive accuracy of concept-based models may be lower than standard black-box models (Loss column in Table 3). Other issues and challenges of concept-based models will be analyzed at the end of the section (Section 6.5).

6.1 Supervised Concept-based Models

Supervised concept-based models employ a dataset annotated with symbolic concepts to supervise an intermediate layer explicitly representing the concepts. When these annotations are readily available within the same training dataset, a *joint concept training* strategy can be employed (Section 6.1.1). Otherwise, they can also be embedded in the model

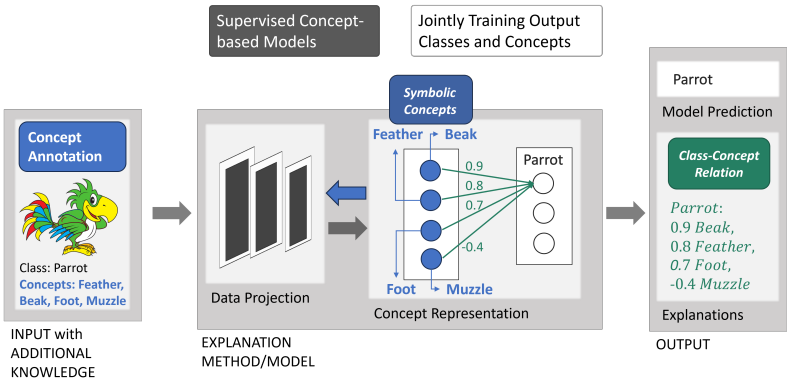

Fig. 7. Supervised Concept-based Models jointly training output classes and concept representation on a dataset where both class and concept annotation is provided.

through separate training on an external dataset dedicated solely to concept learning to perform a *concept instillation* (Section 6.1.2).

6.1.1 Joint concept training. As shown in Figure 7, supervised concept-based models employ concept annotations to supervise an intermediate representation and increase the transparency of the model. The training is performed *jointly* for the target task (e.g., image classification) and for concept representation. This requires employing a training dataset that includes annotations about both the main classes (e.g., parrot) and the related attributes (e.g., feather, beak, foot, muzzle). Concepts can be represented by means of a single neuron (CBM, LEN) or through an embedding (CEM, ProbCBM, DCR), which avoids performance loss with respect to a black-box model. As shown in the picture 7, these methods can also analyze the relationships between concepts and classes through concept importance (CBM, ProbCBM) or logic relationships (LEN, DCR).

The first work proposing a joint task-concept training strategy is Concept-Bottleneck Model **(CBM [54])**. CBM modifies the typical end-to-end neural architecture by introducing an intermediate *concept bottleneck* layer. Each neuron in this layer learns and predicts a human-specified concept. A task function predicts the final target, relying only on concept predictions (concept scores). A double loss is optimized, penalizing wrong predictions for both the tasks and the concepts. CBM also enables interaction with the model itself through *concept interventions*. A domain expert may intervene in the predicted concepts, allowing potential adjustments to the model's predictions and facilitating the extraction of counterfactual explanations. CBMs, however, are affected by two limitations. i) CBMs provide class-concept relationships only when employing a single layer to model the task function. In this case, the importance of a concept corresponds to the weight connecting the concept to the final class, as shown in Figure 7. (ii) Due to the bottleneck layer, the generalization capability of CBM is lower than standard end-to-end models. This limit becomes more severe in contexts with few concepts available and when employing a shallow task function (imposing a trade-off between generalization and interoperability).

Logic Explained Networks **(LENs [13, 24])** aims to enhance the explainability of CBM task predictions. LENs are trained to provide explanations in terms of logic formulas, associating task functions with the concept that has the highest influence on it. To achieve this goal, they impose architectural constraints enforcing sparsity in the weights of the task function. As an example, on a bird classification task, a LEN may provide explanations like

billShapeHooked ∧ mediumSize ∧ ¬throatColorWhite ⇔ blackFootedAlbatross. Further, LENs can be applied to tabular and textual data [46] without a concept extractor, and to solve clustering tasks, providing explanations about cluster assignment in terms of the most frequent concepts. The accuracy of LENs, however, is still lower than that of black-box networks since they are still based on a concept-bottleneck architecture.

Rather than representing a concept with a single neuron as in CBM, Concept Embedding Model **(CEM [30])** proposes using a set of neurons, thus creating a concept embedding. With this solution CEM overcome CBM representation bottleneck, obtaining the same classification accuracy as black-box neural networks even in contexts where the number of concept is very low. By assigning a dedicated embedding to each concept, CEM remain responsive to concept-level interventions, unlike hybrid solutions [67] that rely on unsupervised neurons to improve task accuracy. Furthermore, the authors show that CEM concept alignment – a measure of the quality of a concept representation – is close to the one provided by CBM.

The Probabilistic Concept Bottleneck Model **(ProbCBM [51])** employs probabilistic concepts to estimate uncertainty in both concept and class predictions. When concept presence is uncertain (e.g., concept non-visible), deterministic predictions may harm both the task accuracy and the concept-based explanation. ProbCBM instead outputs mean and variance for each concept, representing them as a multivariate normal distribution. Concept probabilities are obtained via Monte-Carlo sampling and computing the distance from the true and false concept vectors. Class predictions are also similarly computed. Uncertainty instead is quantified through the determinant of the covariance matrix. Empirically, ProbCBM outperforms CBM in task accuracy but falls short of CEM; for concept prediction, it matches CBM and surpasses CEM, suggesting that probabilistic embeddings improve concept reliability without compromising performance.

While concept embeddings improve predictive performance, they reduce explainability, since the dimensions of an embedding lack symbolic meaning and weaken task interpretability. To address this, Deep Concept Reasoner **(DCR [12])** is introduced as an *interpretable* concept embedding model. Starting from concept embeddings, DCR learns differentiable modules that output a fuzzy rule for each class and sample. The rules are then executed over the concept activations. This yields predictions that are locally interpretable with respect to the concepts. DCR task accuracy is on par with CEM and higher than CBM. Compared to LEN, DCR better supports counterfactual example generation and produces explanations less sensitive to perturbations. Finally, DCR can also work on top of a GNN on graphs data.

SparseCBM [98] extends CBMs to LLMs and text classification while introducing some novel features. It employs specialized subnetworks for concept prediction rather than a single backbone to improve concept prediction. It introduces a sparsity-guided optimization favouring task-interpretability that minimizes the number of important concept-task weights. The importance is assessed via the Hessian matrix. At inference, SparseCBM allows the tasks to use only a subset of concepts through masking. Additionally, SparseCBM supports automatic inference-time interventions, dynamically adjusting masks to discard low-importance or historically contradictory weights.

6.1.2 Concept Instillation. This class of approaches considers the case where concept information is not available in the training set, but it is available from an external data source, such as a separate supervised dataset (CW, CT), a semi-supervised one (CME), or a knowledge graph (PCBM). As shown in Figure 8, the process is composed of two steps: 1) a black-box network is standardly trained end-to-end; 2) through *concept instillation*, a given layer of the network is modified to represent concepts. At test time, this process still allows the employment of a concept-based model. Similarly to jointly trained models, concepts can be represented by means of single nodes (CW, CME, PCBM) or

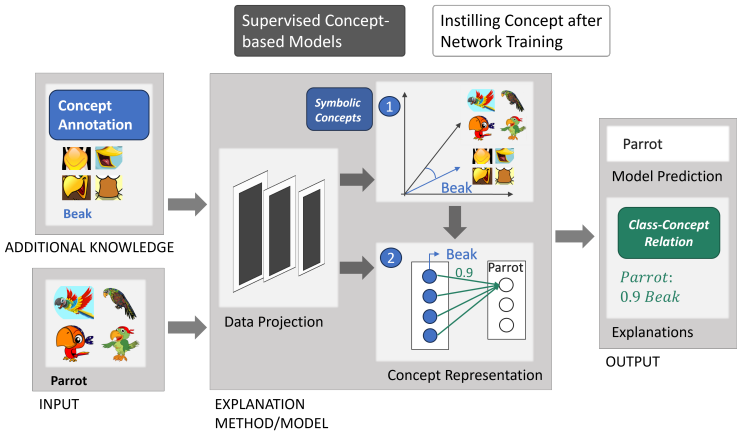

Fig. 8. Supervised Concept-based Models instilling concept in the network after a first end-to-end network training. Concept representation is either extracted from the network with post-hoc C-XAI techniques or from an external source (e.g., a knowledge graph).

through concept embeddings (CT). They also provide class-concept relationships if they replace the classifier head with an interpretable task predictor (CME, PCBM, CT).

Concept Whitening **(CW [23])** modifies a hidden layer of a neural network after it has been trained to predict for a given set of concepts. CW may replace the Batch Normalization (BN) layer as it combines batch whitening (decorrelating and normalizing each dimension) with a rotation matrix to align the concepts with the axes. Latent space axes thus align with concept axes, allowing each point to be interpreted through known concepts. Further training on the original classes is required to avoid a loss in task performance. The authors use as interpretability metric the *concept purity* computed as the 1-vs-all AUC score. A comparative analysis against post-hoc methods (TCAV, IBD) indicates that the concepts learned with CW exhibit higher purity.

Like CW, the Concept-based Model Extraction **(CME [49])** extracts an interpretable model from a black-box by training a concept extractor on latent representations with a small annotated dataset (semi-supervised). The extraction is repeated across layers, retaining the one with highest concept accuracy An interpretable task predictor (Decision Trees or Logistic Regression) is then trained on the concept scores to mimic the task predictions of the original model. CME's performance outperforms Net2Vec in task accuracy while remaining comparable to CBM. The authors also introduce a new metric, the MisPrediction Overlap (MPO), to measure the portion of samples with at least m mispredicted concepts; by MPO, CME performs similarly to Net2Vec but below CBM.

Post-hoc Concept Bottleneck Models **(PCBM [109])** extract concepts either (i) from external datasets or (ii) from a knowledge graph. In the first case, CAVs are employed to learn the concept representations in the latent space of a model trained on the final task. In the second, a knowledge graph provides concepts related to the queried classes, which are encoded into vectors via a text encoder. In both scenarios, the concepts are projected in the latent space of the pretrained model, and an interpretable task predictor is trained over concept activations, yielding class-concept relationships. Beyond concept interventions, PCBMs also support *global model edits*, enabling adaptation to novel scenarios such as data shifts. Similarly to [67], the authors propose an unsupervised set of neurons to improve network

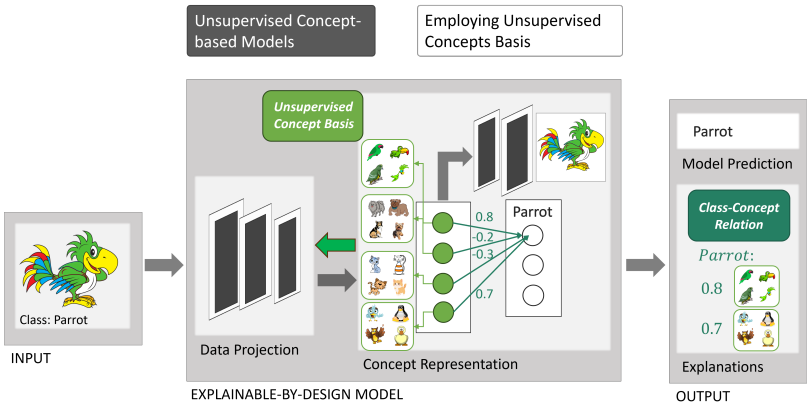

Fig. 9. Unsupervised concept-based model extracting unsupervised concept basis. How the concepts are extracted varies between the models. Here, we report the case in which a decoder branch is employed over the concept to reconstruct the input and extract meaningful concepts (SENN, BotCL).

generalization, but with residual fitting, so that predictions still depend on the interpretable pathway. Overall, PCBM performance is slightly below but comparable to the original model,.

ConceptTransformer **(CT [86])** replaces the task classifier of a DNN with an attention module enriched with concept embeddings. This time, concept embeddings come from an external transformer model trained to represent the concept scores. A linear layer on top of the attention module predict the final classes and provide class-concept relationship. To obtain more plausible explanations, they also propose supervising attention scores so that concept-output relationships align with domain knowledge. CT classification performance matches end-to-end black box models, and even improves them in contexts where concepts are well-defined, and the relationships with the final classes are known (as in the case of CUB-200 dataset [101]).

6.2 Unsupervised Concept-based Models

Unsupervised concept models modify the internal representation of a network to autonomously extract concepts without explicitly associating them with pre-defined symbols. In some cases, these models jointly train an intermediate layer with an unsupervised learning technique to extract an *unsupervised concept basis*, i.e., a clusterized representation identifying groups of similar samples (Section 6.2.1). Otherwise, they explicitly require the network to represent in the layer weights *prototypes-concepts*, (parts of) frequently seen input samples (Section 6.2.2).

6.2.1 Unsupervised Concept Basis. These approaches learn unsupervised disentangled representations in the model's latent space, grouping samples by underlying characteristics such as generative factors. This class of approaches takes inspiration and shares ideas with the disentangled representation learning field [15]. To extract unsupervised concepts, these models are jointly trained to reconstruct the input by means of a decoder branch (SENN, BotCL, and reported in Figure 9) to maximize the mutual information (Interpretable CNN) or via self-supervised loss over textual data (SelfExplain). When these models employ an interpretable task predictor over the concepts, they also provide class-concept relationships (SENN, BotCL).

In **Interpretable CNN [111]**, convolutional filters act as unsupervised concepts and are required to represent different object parts. To achieve this, the authors devise an unsupervised loss function, maximizing the mutual information between the image and the filter activations. Also, the loss promotes low entropy in both the inter-category activations and the spatial distributions of neural activations. The idea is thateach filter should encode a distinct concept associated with a single object category. In the evaluation, the authors measure the *IoU score* with ground-truth object bounding boxes, similar to ND, and the *part location stability*, assessing the consistency in how a filter represents the same object part across various objects. Experimental resultsindicate that Interpretable CNNs, exhibit higher interpretability and positional stability than standard CNNs, while their classification accuracy remain comparable.

Self-explanatory neural network **(SENN [9])** also adopts an unsupervised approach, relying on an auto-encoder architecture to derive concepts – the latent dimensions of the auto-encoder – from the input features. Second, it employs an input-dependent parametrizer that generates class-concept relevance scores. The task prediction is given by the linear combination of the extracted concepts and their relevance scores. Also, to visualize the identified concepts, SENN reports training example that maximally activates each concept. Compared to standard XAI techniques (i.e., LIME, SHAP), SENN explanations are i) more explicit and understandable to humans, ii) more faithful, as importance scores align with ground-truth relevance, iii) more stable, as similar input examples yield similar explanations. Finally, SENN's classification accuracy is comparable to or slightly lower than that of standard CNNs.

Bottleneck Concept Learner **(BotCL [100])** extends SENN by feeding the features from the convolutional encoder into an attention-based mechanism [60] to predict concept scores. Concept embeddings, learned through backpropagation, serves as keys of the attention mechanism. A linear layer working over concept scores is employed as a task classifier, and it captures class-concept relationships. To learn the unsupervised concepts, BotCL uses a contrastive loss along with two regularization terms that promote consistency and mutual distinctiveness in the learned concepts. Its task performance is, on average, higher than SENN, ProtoPNet, and Completeness-aware – particularly when employing contrastive loss. The extracted concepts, compared against ACE and Completeness-aware on a synthetic dataset annotated with concepts, show superior concept purity, efficiency, and reconstruction error.

SelfExplain [81] provides local explanations class-concept relationships in text classifiers. It extracts concepts in an unsupervised way from the input text as the non-terminal leaves of the semantic parsing tree. The model adds two layers to a standard text classifier: the Locally Interpretable Layer (LIL) which assigns relevance scores to the concepts extracted from the same input, and the Globally Interpretable Layer (GIL), which explains predictions by means of the concepts extracted from all the training data, enabling the analysis of how training concepts (possibly absent in the given sample) influence the classifier's decision. Integrated into transformer models, these layers yield consistently better classification performance across datasets, and produce explanations acknowledged as more understandable and plausible than those of standard XAI techniques.

6.2.2 Prototype-based Concepts. Within unsupervised concept-based models, another possibility is to require the network to explicitly encode *prototype-based concepts*, which represent specific (parts of) train examples. As shown in Figure 10, the prototypes are compared to input samples, their similarities are employed to provide the final prediction. How the prototypes are extracted is different among the methods, either by means of an autoencoder (Prototype DL) or per class by means of the same convolutional filters (ProtoPNets), sharing them among the classes (ProtoPShare, ProtoPool) with a hierarchy of prototypes (HPNet), with deformable filters (Deformable ProtoPNets) or through an interactive process with human experts (ProtoPDebug). The resulting models typically provide all explanation types:

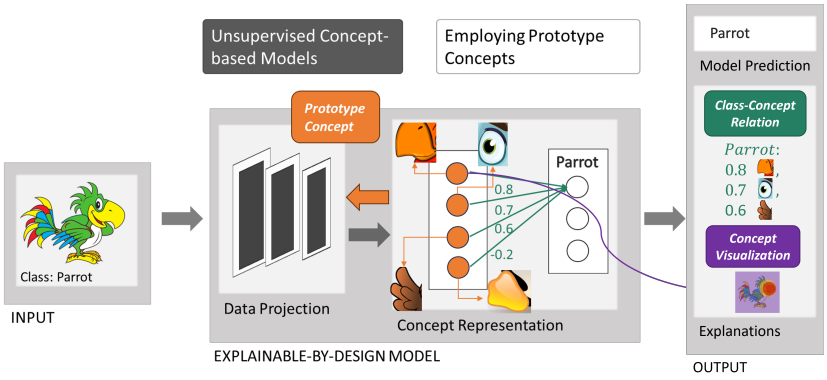

Fig. 10. Unsupervised concept-based models employing prototype concepts. The network explicitly encodes prototype concepts representing (part of) training examples. The explanations are typically provided as class-concept relationships and/or concept visualization.

class–concept relationships, as classification is often performed via a linear layer over the prototypes; node–concept associations, since each prototype corresponds to a network unit; concept visualizations, commonly used as reference for the learnt prototypes.

Prototype DL [61] introduces the use of prototypes in concept-based explainability. The model predictions are explained by means of similarity to prototypical observations within the dataset. Its architecture comprises an autoencoder and a prototype-based classification network. On top of the encoder, a prototype layer is employed, encoding a weight (prototype) vector mirroring a training input. The training objective has four terms. Other than the standard cross-entropy and the reconstruction loss for the autoencoder, a term requires each prototype vector to be as close as possible to a training sample, while the last term encourages every encoded input to be close to at least one prototype vector. The decoder, instead, enables the visualization of learned prototypes. A linear classifier on top of the prototype layer produces the predictions, and its weights identify the more representative prototypes of a class. The experimental results show that the proposed architecture does not compromise predictive ability compared to traditional non-interpretable architectures.

Prototype Parts Network **(ProtoPNet [22])**, rather than working on the whole picture, extract prototypes representing parts of the image. To achieve this, the prototype layer is inserted earlier in the network within the last convolutional filters. For each image patch, ProtoPNet computes the similarity between each prototype unit and all the image patches. This yields a similarity activation map, where the values indicate the presence of the prototype in each part of the image. The activation maps provide a visualization of the learned prototypes. The maps are then condensed into a single similarity score for each prototype through global max pooling. This unified score quantifies the similarity of a prototypical component to the input image. The final prediction is provided by a linear layer working on the unified scores and providing the importance of each prototype for each class. ProtoPNet achieves a slightly lower accuracy (up to −3.5%) than non-interpretable baselines. Deformable ProtoPNet [27] addresses ProtoPNet's limitation of using spatially rigid prototypes. Each prototype consists of adaptive prototypical parts that dynamically adjust their spatial positions based on the input image. The experimental results show that Deformable ProtoPNet achieves state-of-the-art accuracy.

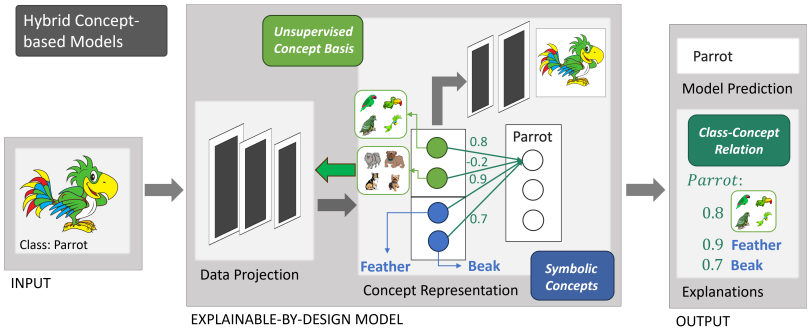

Fig. 11. Hybrid concept-based model employing both supervised and unsupervised concepts. This approach normally involves employing an unsupervised learning strategy (such as reconstructing the input image) to learn an unsupervised concept basis and a few supervisions over some neurons representing symbolic concepts.

The Hierarchical Prototype Network **(HPNet [41])**, the Prototypical Part Shared Network **(ProtoPShare [89])**, and **ProtoPool [88]** extend the ProtoPNet framework with different goals. HPNet organizes prototypes hierarchically, assigning them to different levels of a taxonomy and enabling multi-level explanations and novel class detection, at a slight cost in fine-grained accuracy. ProtoPShare reduces redundancy by merging similar prototypes across classes through a data-driven pruning phase, improving accuracy and consistency while reducing the prototype count. ProtoPool follows a similar goal but avoids the expensive pruning step by adopting soft, fully differentiable prototype assignments. It introduces a focal similarity function for more focused saliency maps and supports concept visualization via prototype projection. ProtoPool achieves competitive or superior accuracy with fewer prototypes, and a user study found its visualizations more interpretable than those of previous methods.

ProtoPDebug [18] introduces a human-in-the-loop concept-level debugger. ProtoPDebug assesses all prototypes learned during each iteration, retrieves maximally activated training samples, and prompts human experts to identify which prototypes are confounders. Confirmed confounders are added to the *forbidden concepts* set, while high-quality prototypes into the *valid concepts* set. For the fine-tuning step, ProtoPDebug proposes a *forgetting loss* penalizing forbidden concepts while a *remembering loss* encourages remembering valid ones. ProtoPDebug outperforms a predecessor debugger, and ProtoPNet, enhancing test classification accuracy

6.3 Hybrid Concept-based Models

Hybrid concept-based models propose the joint employment of supervised and unsupervised concepts, as shown in Figure 11. The strategy normally involves the employment of an unsupervised learning strategy, together with supervision over some concepts. The unsupervised strategy consists of either reconstructing the input image with an AE (Ante-hoc) or a VAE (GlanceNet) or optimizing an unsupervised objective such as mutual information (CBM-AUC). The goal of this integration is generally to allow the employment of concept-based models also in scenarios where few supervised concepts are available. This approach has been adopted to improve the performance of standard CBM (CBM-AUC, Ante-hoc Explainable Model) or its capability to represent the concepts (GlanceNet).

Concept Bottleneck Model with Additional Unsupervised Concepts **(CBM-AUC [91])** is the first approach in this area, merging CBM with SENN. CBM-AUC addresses the issue of CBM's restricted concept efficiency, enabling the model to leverage the representation capability associated with unsupervised concepts. To achieve this, they introduce

M-SENN, a variant of SENN, reducing the computational complexity and improving its scalability. It replaces the SENN decoder with a discriminator, which optimizes the mutual information between the input and the unsupervised concepts. Second, the M-SENN relevance predictor now shares the intermediate network with the encoder, reducing the computational complexity. For integrating CBM into SENN, they add supervision on some concept basis. CBM-AUC outperforms CBM's task accuracy while employing fewer parameters than SENN. Also, they visualize both supervised and unsupervised concepts through their saliency maps, showing that both types are semantically meaningful.

Similarly to CBM-AUC, Learning **Ante-hoc Explainable Models via Concepts [90]** integrates SENN and CBM, still employing SENN concept decoder to learn unsupervised concepts. The main difference is that the classifier is added on top of the basic encoder, not on the concepts. This entails that the final classification is not necessarily based on the concepts. To minimize this issue, they add a classifier on top of the concept encoder and minimize predictions' divergence via a fidelity loss. To learn supervised concepts, a loss function is employed to align the learned concepts with available annotations. Otherwise, self-supervision can also be integrated as an auxiliary task, such as predicting the rotation of the input images starting from the encoded concepts. The proposed framework demonstrates competitive performance across both annotated and unannotated datasets, improving the task accuracy of CBM and SENN, respectively.

Unlike previous work, **GlanceNets [68]** aims to improve the concept representations of SENN and CBM rather than the performance. It improves the interpretability of SENN by employing a VAE instead of a standard AE to learn disentangled concepts. Due to the variational approach, learned concepts can better represent sample generative factors and respond to their alterations. When supervised concepts are available, they force a mapping with the learned concepts to preserve known semantics. They also improve the representations of CBM by identifying a solution to the concept leakage issue (see Section 6.5 for a description). They claim it is a deficiency in the out-of-distribution generalization of concept-based models. Therefore, GlanceNets employs an open-set recognition technique at inference time to detect instances that do not belong to the training distribution. They show this solution substantially improves the concept alignment with respect to CBM. Task performances are on par with CBM and SENN.

6.4 Generative Concept-based Models

Rather than employing a set of symbolic supervised concepts or extracting unsupervised ones, generative concept-based models can create concept representations. As shown in Figure 12, this is generally performed by employing a Large Language Model (LLM) producing *textual Concepts*, short textual descriptions of the final class. The embeddings representing the textual concepts are aligned to the latent input representation to output concept scores. These scores are then employed to provide the final classification, possibly with an interpretable classifier providing class-concept relationships. At test time, these models predict both the final class and the most suitable descriptions among those learned. The effectiveness of these approaches is intrinsically linked to the representation of the classes and concepts within the adopted LLM. In the following, we outline two methods (LaBO [107] and Label-free CBM [75]) of this recent line of work, that are similar in their main components since they both employ an LLM in combination with a vision-language one and provide class-concept relationship explanations.

Language-Model-Guided Concept Bottleneck Model **(LaBO [107])** is the first generative concept-based model. In LaBO, first, the Generative Pre-trained Transformer 3 (GPT-3) model [21] is prompted with queries designed to generate a collection of *candidate concepts* for each class. An example of a prompt is: *"Describe what the <class_name> looks like."* Subsequently, a submodular optimization technique is applied to select a subset of the generated sentences to maximize the *discriminability* and *diversity* of the concepts. The generated textual concepts are then embedded with pretrained

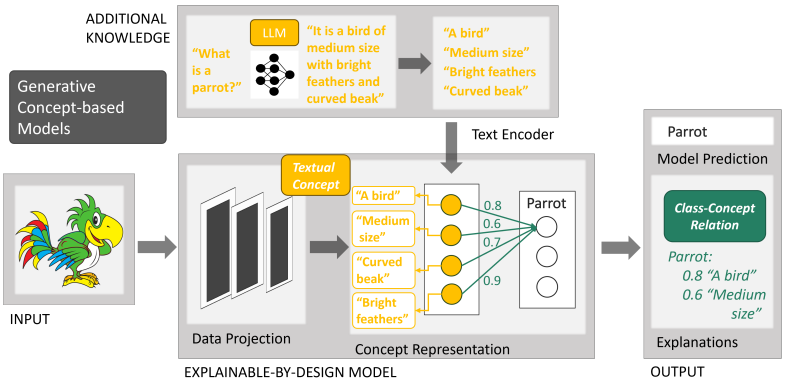

Fig. 12. Generative concept-based models employ an external LLM to generate textual concepts, i.e., short descriptions of the final classes without requiring a concept annotation. The embeddings representing the textual concepts are aligned with the input latent representation. The alignment produces the concept scores, which are used to provide the final classification.

CLIP [80] and aligned to image embeddings to form a concept bottleneck layer. Subsequently, a linear layer is applied to the similarity scores of concepts and images to learn a weight matrix whose values represent the importance of the concepts in the classification task. LaBo retains high accuracy in limited training data scenarios, but as data volume grows, its performance slightly decreases while remaining competitive compared to black-box models. Compared to PCBM [109] as an interpretable baseline, LaBo exhibits a substantial performance improvement.

As LaBO, Label-free Concept Bottleneck Model **(LabelFree-CBM [75])** generates the concepts using GPT-3 [21] and filters them to enhance the quality and reduce the concept set's size. Each matrix element is the product between the CLIP encoding of the images and the textual concepts. The framework then involves learning the concept weights to create the concept bottleneck layer. Lastly, the final predictor is learned by training a fully connected sparse layer. The authors assessed the interpretability of the model with a user study, while regarding task performance, LF-CBM demonstrates a small loss in task accuracy compared to the original neural backbones while it also outperforms PCBM [109].

6.5 Discussion

In the following, we analyze critical aspects of explainable-by-design models. We focus on three perspectives: (i) performance loss, (ii) information leakage in concept representation, and (iii) pitfalls of concept intervention.

Performance loss. Concept-based models explicitly represent a set of concepts within the model architecture and rely on these concepts for the final task predictions. However, the higher transparency of these architectures usually comes at the cost of a performance loss compared to traditional, black-box architectures. We analyze this aspect in Table 3. Still, we note that several approaches (that we report with ✗) match the performance of complex black-box prediction models, with the advantage of enabling decision explanations through concepts.

Information leakage in supervised concept representation. Recent works revealed that supervised concept-based models encode additional information in the concept space other than the concept information itself [42, 67, 69]. Concept representations may include representations of the task labels when trained jointly with the downstream task. This phenomenon is denoted as *information leakage in concept representation*, and compromises the model's

interpretability and intervenability (i.e., the possibility to perform concept intervention) [42, 67, 69]. Mahinpei et al. [67] argue this issue is shared among concept-based models that use soft or fuzzy representations to represent concepts and experimentally demonstrate it for CBM and CW. They also show that information leakage is not avoided even when applying mitigation strategies, including (i) training concept representations and the concept-to-task sequentially (rather than jointly), (ii) adding unsupervised concept dimensions to account for additional task-relevant information, or (iii) via concept whitening which explicitly decorrelates concept representations during training. Hence, they conclude that soft representations encode concept information and data distribution. Recent works specifically address leakage issues in concept-based models, such as [42] for CBMs and GlanceNets.

Pitfalls of concept intervention. Concept-based models enable human experts to *intervene* in the concept representations and modify them at test time. Intervening on mispredicted concepts can enhance task performance and enable the generation of counterfactual explanations. However, concept intervention techniques might be affected by major pitfalls. First, intervention could increase the task error, compromising the reliability of intervention techniques. This decrease in performance could happen when users nullify important concepts by setting unsure concepts to zero [93]. The second issue arises from the practice of using majority voting to associate concept values to classes to improve task performance (e.g., in [30, 42, 54]). Majority voting affects a fair representation of minority samples and makes task outcomes biased toward the majority, and intervention can exacerbate this unfair treatment, since it would increase minority performance only when changing the concept value to the majority one.

Concept-based models and adversarial attacks. Similarly to post-hoc methods, also C-XAI models are susceptible to adversarial attacks [95], but concepts can also be leveraged for creating defenses [24, 62, 84]. In particular, [95] shows that malicious perturbations in concept-based models (CBM, SENN) can alter concepts in different ways without affecting final predictions. To address this threat, they propose an adversarial training-based defense, demonstrating its effectiveness across datasets. Regarding attacks on final classes, [84] found that CBMs are slightly more robust than CNNs, especially with sequential training (first concept encoder, then task classifier). [24, 62], instead, show that concept-class relations can be exploited for defense. Specifically, [24] proposed using LEN logic explanations and checking test-time predictions' consistency with the explanations extracted at training time. [62] proposed enhancing CBMs with factor graphs, relying on existing knowledge instead of learned one. By encoding semantic concepts and final classes as nodes and logical relationships as edges of a factor graph, they can identify and rectify errors during inference and ensure logically consistent explanations even under perturbations.

7 EVALUATION

In this section, we address **RQ5** by reviewing how C-XAI methods are evaluated. We present quantitative metrics (Section 7.1), qualitative human-centered evaluations (Section 7.2), commonly used datasets (Section 7.3), and available resources 7.4. Section complements this with details on the availability of code.

7.1 Quantitative evaluation

We categorize the quantitative metrics proposed in the reviewed papers into two categories according to their final purpose and what they measure. The metrics of the first category assess how well concepts contribute to class predictions and task performance. The second category focuses on the intrinsic quality of concepts. For a more detailed explanation of each metric, please refer to Appendix C.1.

Concept effect on class prediction and task performance. These metrics assess how concepts contribute to (i) the final class prediction or (ii) the task performance.

Concept effect on class prediction. Concept-based methods often naturally provide the concept relevance (importance) by means of the concept weight connecting the concept to the final classes. Hence, these metrics have been mostly employed for post-hoc explanation methods, providing class-concept relationship to study the impact of the identified concepts on task prediction. Among these metrics we find the T-CAV score [35, 50], CaCE [37, 102],ConceptSHAP [108]. Although these metrics can explore the relationship between classes and concepts, they often assume relationships such as linearity or independence between concepts and predictions, which may not always hold. Moreover, they focus on model-side importance, which may not align with how humans perceive concept relevance.

Concept effect on task performance. These metrics assess how well concepts contribute to the predictive capacity of the model. They are used both to evaluate concept quality in post-hoc explanations and to measure concept contribution in concept-based models. Among these metrics we find the Completeness Score [108]and ConceptEfficiency [30].These metrics typically focus on overall prediction capacity rather than concept-specific contributions, making it harder to understand the role of individual concepts.

Quality of concepts. This second category focuses on the quality of concepts, studying (i) their inherent properties, (ii) how they relate to internal network representations, and (iii) their impact on prediction errors. In the following, we review these metrics with respect to these three targets.

Concepts properties. This group of metrics centers on the intrinsic characteristics of the concepts themselves. They provide insights into the information content and coverage of concepts across several aspects, such as concept informativeness [30], distinctiveness [100], and purity [100]. Still, these measures can be sensitive to dataset bias and may not fully capture semantic relevance.

Relationships of concepts with internal network representation. This group of metrics quantifies the relationship between a node (or filter) of the model and the concept it represents. These metrics are adopted for methods that provide explanations as node-concept association. They assess the spatial overlap [14, 33, 104, 111], and location stability [111] of concepts within activation maps. While allowing for validation of the localization and consistency of concepts within the model, these metrics require well-defined concept localization.

Concept prediction error. This group of metrics measures the alignment between learned concepts and concept ground truth. Hence, they are applicable when concept labels are available. Among these metrics we find the Concept Error [54], and AUC score [23]. While offering an objective evaluation of concept quality, their adoption depends on the availability and quality of concept labels, limiting their applicability in settings where such annotations are scarce or noisy.

7.2 Qualitative evaluation

As is common in XAI assessment [17], several C-XAI works perform a human evaluation to assess a method's quality. We outline qualitative assessments across three main aspects.

Understandability of explanations. When evaluating explanations, the predominant focus involves assessing how understandable and reasonable the explanations are to humans. One relevant property of explanations is their *plausibility*, denoting the extent to which explanations align with human reasoning. This is often measured by asking users whether an explanation is justifiable and matches their expectations [14, 81]. To evaluate *human understanding* of concepts, users may select descriptive phrases from a predefined vocabulary that best describes the concept [100]. Other methods, such as [112], ask users to generate free-text labels and match them to a set of candidate explanations,

with accuracy measured by the percentage of correct matches. Similarly, the work in [113] focuses on quantifying the semantics learned by each network unit and evaluating the accuracy of human-assigned text annotations to concepts. The *trustability* of explanations is another critical aspect, where users provide feedback on their level of trust in the explanation [81]. The *reasonability* of explanations evaluates how effectively explanations communicate information. This can be tested by asking what information the explanation conveys [50], or by comparing explanations and having users choose which one seems more reasonable [115]. *Factuality* refers to the accuracy of concepts by having annotators judge their alignment with ground truth images [107].

Coherence of concepts. When concepts are extracted by a method and not supervised, a crucial aspect is their coherence with human-defined concepts. A method for measuring it is the intruder detection experiment. In this setup, users are asked to identify an image among a set that is conceptually different from the rest [31, 35]. Another setting involves examining the nearest neighbor images associated with each discovered concept vector and selecting the most prevalent and cohesive concept [108]. [107] defines the *groundability* to measure the consistency of the vision-language model adopted in their framework to ground concepts to images in a manner that aligns with human interpretations.

Utility of explanations. The *utility* of an explanation refers to its practical usefulness and effectiveness in providing insights into the model. For example, in [46], users have to identify the biased features that decrease classifier generalization in the explanations. In addition, users are asked to select the most general classifier based solely on the global explanations of various methods. Similarly, [31], evaluate how well explanations help users infer the reasoning behind classifications, assessing user accuracy in predicting the model's decisions on new images.

The strength of qualitative evaluations lies in directly assessing explanations from a user-centered perspective, capturing aspects like trust, coherence, and usefulness that are hard to quantify with automated metrics. However, these evaluations face several limitations. They are inherently subjective, influenced by individual and cognitive biases, prior knowledge, and familiarity with the task. They require significant time and resources to design and conduct, and often require a manual analysis, making them hard to scale. Their results can be difficult to reproduce or compare across studies due to variability in participants' interpretations, expertise, and recruitment, and due to the lack of standardized protocols. Furthermore, the results may also depend heavily on the phrasing of questions, the design of the user interface, or cultural and linguistic factors, which can further affect consistency and generalizability of results.

7.3 Datasets

C-XAI methods use datasets across multiple modalities. In the **image domain**, CUB [101] is widely adopted for its 100+ visual attributes across 200 bird species, while BRODEN [14] offers over 1,000 concepts spanning textures to scenes. OAI [77] uniquely supports both classification and regression tasks. In the **text domain**, CEBAB [1] is the only dataset with explicit concept annotations; others extract concepts from syntactic structures [81]. BDD-OIA [103] is the only **video** dataset with predefined concepts, while in **tabular** (e.g., COMPAS, MIMIC-II) and **graph** domains (e.g., MUTAG, PROTEINS), concepts are either predefined features or extracted during analysis. More details are in Appendix C.2.

7.4 Resources

We now report insights about the resources associated with the reviewed methods. Specifically, we consider whether the authors made their code and models publicly accessible or released a new dataset. Among the methods discussed, all but the following four have made their code publicly available on GitHub: two post-hoc methods (CaCE, Object Detection) and two concept-based models (SENN, CBM-AUC). Concerning data release, four methods released novel datasets,

equally distributed between post-hoc methods (ND, DMA & IMA) and concept-based models (CME, Glance-Nets). For additional information, please refer to Appendix C.3.

8 APPLICATION AND EMERGING TRENDS

In this section, we address **RQ6** by highlighting how C-XAI methods are being applied across domains and discussing emerging trends shaping their development. Applications illustrate the practical value of concept-based explanations in real-world scenarios, while recent trends—particularly those involving foundation and generative models—show how the field is evolving.

8.1 Applications

The recent emergence of C-XAI methods has led to some interesting applications. The medical sector is where we find most applications of C-XAI. Some of the previously discussed methods (CAR [25], CBM [54], PCBM [109], ProtoPDebug[18], LaBO [107]) are evaluated on medical data, as in [64], where the authors employ TCAV to the task of skin cancer diagnosis, showing that the concepts learned by the DNN align closely with those used by dermatologists. They suggest that explicitly listing the influential concepts can be sufficient to foster trust in automated medical decisions. [106] presents an explainable-by-design approach that uses concept to create a human-in-the-loop framework. In this, a concept bank is defined using both human-annotated data and automatically extracted confounding factors such as dark corners and dense hairs. Also, after training, it enables users to rewrite the model's decision using first-order logic rules. The approach ensures predictions are based on meaningful medical concepts rather than dataset biases.

The education domain also shows potential for applications. [11] uses a combination of a graph-based neural network approach for classifying student interaction time series and adapts the concept activation vectors TCAV [50] for interpreting a GNN internal state in terms of concepts. In this case, concepts are represented by six learning dimensions from educational scenarios defined a priori.

8.2 Emerging Trend: Foundation and Generative Models in C-XAI

Among recent trends in C-XAI, generative models, diffusion models, and LLMs are increasingly being adopted not only to support explanation tasks but also to enable concept-driven content generation.

In post-hoc methods, these models are increasingly used to explain concept-class relationships and neuron-level representations. For example, [72] leverages CLIP to derive Concept Activation Vectors from user-provided descriptions, removing the need for manual labels. Similarly, [55] uses VLM supervision to annotate concepts and perform test-time interventions. Node–concept associations are also addressed: [76] aligns neuron activations with textual embeddings via MLLMs, while [8] combines neuron–concept mapping with Shapley-based concept attribution. Moreover, some works aim to explain foundation models themselves: [16] decomposes CLIP embeddings into sparse, interpretable concepts. The growing field of *mechanistic interpretability* can be also considered under this lens: among other works, [19] uses sparse autoencoders to decompose polysemantic neurons into monosemantic features, and [70] extends this to identify causally relevant concept circuits.

Foundation and generative models also appear in explainable-by-design systems, with goals such as improving zero-shot concept prediction, enhancing concept representation interpretability, and directly embedding concepts. For example, [56] proposes regularization strategies to stabilize concept learning, while [83] and [43] use dictionary learning and CLIP to discover and align concepts without predefined vocabularies. Beyond CLIP, [96] employs Grounding DINO [63] for precise visual grounding, improving both concept prediction and task performance. In text classification,

[98] introduces sparsity to foundation model embeddings to support intervention. [105] replaces structured concepts with free-form textual explanations (e.g., "This is a zebra because it has stripes") using frozen decoders such as BLIP [59]. Finally, [45] and [97] propose inserting concept layers into foundation models (e.g., diffusion models or LLMs), enabling interpretable generation and concept-level control.

9 CONCLUSIONS

This survey provides a structured overview of Concept-based eXplainable AI (C-XAI), outlining concept types, explanation strategies, and a nine-class taxonomy. Our analysis supports practitioners in selecting suitable methods across different application contexts and highlights available datasets and metrics to facilitate future development.

While early applications of C-XAI—particularly in computer vision—have demonstrated potential in domains, such as healthcare, broader adoption remains constrained by challenges in concept annotation, explanation faithfulness, and integration complexity. Moreover, some modalities, including graph-structured and temporal data, remain underexplored. A notable emerging trend is the integration of generative models to discover or annotate concepts and to generate explanations in natural language or visual formats. Promising research directions include: (i) the development of robust and standardized evaluation protocols; (ii) the design of generalizable, multimodal concept representations; (iii) support for user-driven interventions in model behavior; and (iv) tighter integration with generative models to produce faithful and adaptive explanations.

Limitations. Despite our rigorous process, some limitations in our review may exist: *Publication bias*, as we focused on peer-reviewed literature, potentially excluding relevant unpublished work or negative results; *Terminological inconsistency*, since the notion of "concept" is variably defined across studies, affecting categorization; *Rapid evolution of the field*, which may have led to the omission of recent preprints or newly accepted papers.

ACKNOWLEDGMENTS

This work is partially supported by the FAIR - Future Artificial Intelligence Research (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013) and the spoke "FutureHPC & BigData" of the ICSC - Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing, both funded by the European Union - NextGenerationEU. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

REFERENCES

[1] Abraham, E. D., D'Oosterlinck, K., Feder, A., Gat, Y., Geiger, A., Potts, C., Reichart, R., and Wu, Z. Cebab: Estimating the causal effects of real-world concepts on nlp model behavior. *Advances in Neural Information Processing Systems 35* (2022), 17582–17596.

[2] Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., and Lapuschkin, S. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence 5*, 9 (2023), 1006–1019.

[3] Adadi, A., and Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access 6* (2018), 52138–52160.

[4] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. *Advances in neural information processing systems 31* (2018).

[5] Adebayo, J., Muelly, M., Abelson, H., and Kim, B. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *International conference on learning representations* (2021).

[6] Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., and Hinton, G. E. Neural additive models: Interpretable machine learning with neural nets. *Advances in neural information processing systems 34* (2021), 4699–4711.

[7] Agrawal, P., Girshick, R., and Malik, J. Analyzing the performance of multilayer neural networks for object recognition. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13* (2014), Springer, pp. 329–344.

[8] Ahn, Y. H., Kim, H. B., and Kim, S. T. Www: a unified framework for explaining what where and why of neural networks by interpretation of neuron concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 10968–10977.

[9] Alvarez Melis, D., and Jaakkola, T. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems 31* (2018).

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

[11] Asadi, M., Swamy, V., Frej, J., Vignoud, J., Marras, M., and Käser, T. Ripple: concept-based interpretation for raw time series models in education. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2023), vol. 37, pp. 15903–15911.

[12] Barbiero, P., Ciravegna, G., Giannini, F., Espinosa Zarlenga, M., Magister, L. C., Tonda, A., Lio, P., Precioso, F., Jamnik, M., and Marra, G. Interpretable neural-symbolic concept reasoning. In *Proceedings of the 40th International Conference on Machine Learning* (23–29 Jul 2023), vol. 202 of *Proceedings of Machine Learning Research*, PMLR, pp. 1801–1825.

[13] Barbiero, P., Ciravegna, G., Giannini, F., Lió, P., Gori, M., and Melacci, S. Entropy-based logic explanations of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2022), vol. 36, pp. 6046–6054.

[14] Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 6541–6549.

[15] Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence 35*, 8 (2013), 1798–1828.

[16] Bhalla, U., Oesterling, A., Srinivas, S., Calmon, F., and Lakkaraju, H. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems 37* (2024), 84298–84328.

[17] Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., and Rinzivillo, S. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery* (2023), 1–60.

[18] Bontempelli, A., Teso, S., Tentori, K., Giunchiglia, F., and Passerini, A. Concept-level debugging of part-prototype networks. In *The Eleventh International Conference on Learning Representations* (2023).

[19] Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread 2* (2023).

[20] Brown, D., and Kvinge, H. 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* (2023), pp. 620–627.

[21] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems 33* (2020), 1877–1901.

[22] Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems 32* (2019).

[23] Chen, Z., Bei, Y., and Rudin, C. Concept whitening for interpretable image recognition. *Nature Machine Intelligence 2*, 12 (2020), 772–782.

[24] Ciravegna, G., Barbiero, P., Giannini, F., Gori, M., Lió, P., Maggini, M., and Melacci, S. Logic explained networks. *Artificial Intelligence 314* (2023), 103822.

[25] Crabbé, J., and van der Schaar, M. Concept activation regions: A generalized framework for concept-based explanations. *Advances in Neural Information Processing Systems 35* (2022), 2590–2607.

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

[27] Donnelly, J., Barnett, A. J., and Chen, C. Deformable protopnet: An interpretable image classifier using deformable prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10265–10275.

[28] Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys 55*, 9 (2023), 1–33.

[29] Erhan, D., Bengio, Y., Courville, A., and Vincent, P. Visualizing higher-layer features of a deep network. *University of Montreal 1341*, 3 (2009), 1.

[30] Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., Lió, P., and Jamnik, M. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems* (2022), S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Curran Associates, Inc., pp. 21400–21413.

[31] Fel, T., Picard, A., Bethune, L., Boissin, T., Vigouroux, D., Colin, J., Cadène, R., and Serre, T. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 2711–2721.

[32] Floridi, L., and Chiriatti, M. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines 30* (2020), 681–694.

[33] Fong, R., and Vedaldi, A. 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* (2018), pp. 8730–8738.

[34] Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence* (2019), vol. 33, pp. 3681–3688.

[35] Ghorbani, A., Wexler, J., Zou, J. Y., and Kim, B. Towards automatic concept-based explanations. *Advances in neural information processing systems 32* (2019).

[36] Goodman, B., and Flaxman, S. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine 38*, 3

(2017), 50–57.

[37] Goyal, Y., Feder, A., Shalit, U., and Kim, B. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165* (2019).

[38] Greenwell, B. M., Boehmke, B. C., and McCarthy, A. J. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755* (2018).

[39] Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., and Giannotti, F. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820* (2018).

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

[41] Hase, P., Chen, C., Li, O., and Rudin, C. Interpretable image recognition with hierarchical prototypes. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (2019), vol. 7, pp. 32–40.

[42] Havasi, M., Parbhoo, S., and Doshi-Velez, F. Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems 35* (2022), 23386–23397.

[43] He, H., Zhu, L., Zhang, X., Zeng, S., Chen, Q., and Lu, Y. V2c-cbm: Building concept bottlenecks with vision-to-concept tokenizer. In *Proceedings of the AAAI conference on artificial intelligence* (2025), vol. 39.

[44] Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems 32* (2019).

[45] Ismail, A. A., Adebayo, J., Bravo, H. C., Ra, S., and Cho, K. Concept bottleneck generative models. In *The Twelfth International Conference on Learning Representations* (2024).

[46] Jain, R., Ciravegna, G., Barbiero, P., Giannini, F., Buffelli, D., and Lio, P. Extending logic explained networks to text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (2022), Association for Computational Linguistics, pp. 8838–8857.

[47] Ji, Y., Wang, Y., and Kato, J. Spatial-temporal concept based explanation of 3d convnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 15444–15453.

[48] Kaminski, M. E. The right to explanation, explained. *Berkeley Technology Law Journal 34*, 1 (2019), 189–218.

[49] Kazhdan, D., Dimanov, B., Jamnik, M., Liò, P., and Weller, A. Now you see me (cme): concept-based model extraction. *arXiv preprint arXiv:2010.13233* (2020).

[50] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning* (2018), PMLR, pp. 2668–2677.

[51] Kim, E., Jung, D., Park, S., Kim, S., and Yoon, S. Probabilistic concept bottleneck models. In *Proceedings of the 40th International Conference on Machine Learning* (2023), ICML'23, JMLR.org.

[52] Kim, S. S., Watkins, E. A., Russakovsky, O., Fong, R., and Monroy-Hernández, A. " 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* (2023), pp. 1–17.

[53] Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un) reliability of saliency methods. *Explainable AI: Interpreting, explaining and visualizing deep learning* (2019), 267–280.

[54] Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *International conference on machine learning* (2020), PMLR, pp. 5338–5348.

[55] Laguna, S., Marcinkevičs, R., Vandenhirtz, M., and Vogt, J. Beyond concept bottleneck models: How to make black boxes intervenable? *Advances in neural information processing systems 37* (2024), 85006–85044.

[56] Lai, S., Hu, L., Wang, J., Berti-Equille, L., and Wang, D. Faithful vision-language interpretation via concept bottleneck models. In *The Twelfth International Conference on Learning Representations* (2024).

[57] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature 521*, 7553 (2015), 436–444.

[58] Leemann, T., Kirchhof, M., Rong, Y., Kasneci, E., and Kasneci, G. When are post-hoc conceptual explanations identifiable? In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence* (31 Jul–04 Aug 2023), R. J. Evans and I. Shpitser, Eds., vol. 216 of *Proceedings of Machine Learning Research*, PMLR, pp. 1207–1218.

[59] Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning* (2022), PMLR, pp. 12888–12900.

[60] Li, L., Wang, B., Verma, M., Nakashima, Y., Kawasaki, R., and Nagahara, H. Scouter: Slot attention-based classifier for explainable image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 1046–1055.

[61] Li, O., Liu, H., Chen, C., and Rudin, C. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2018), vol. 32.

[62] Li, Y., Zhou, K., Yu, S., Zhang, Q., Luo, R., Li, X., and Xia, F. Factor graph-based interpretable neural networks. In *The Thirteenth International Conference on Learning Representations* (2025).

[63] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision* (2024), Springer, pp. 38–55.

[64] Lucieri, A., Bajwa, M. N., Braun, S. A., Malik, M. I., Dengel, A., and Ahmed, S. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In *2020 international joint conference on neural networks (IJCNN)* (2020), IEEE, pp. 1–10.

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

[66] MacCarthy, M. An examination of the algorithmic accountability act of 2019. *Algorithms* (2020).

[67] Mahinpei, A., Clark, J., Lage, I., Doshi-Velez, F., and Pan, W. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314* (2021).

[68] Marconato, E., Passerini, A., and Teso, S. Glancenets: Interpretable, leak-proof concept-based models. *Advances in Neural Information Processing Systems 35* (2022), 21212–21227.

[69] Margeloiu, A., Ashman, M., Bhatt, U., Chen, Y., Jamnik, M., and Weller, A. Do concept bottleneck models learn as intended? *CoRR abs/2105.04289* (2021).

[70] Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations* (2025).

[71] Mikriukov, G., Schwalbe, G., Motzkus, F., and Bade, K. Unveiling the anatomy of adversarial attacks: Concept-based xai dissection of cnns. In *World Conference on Explainable Artificial Intelligence* (2024), Springer, pp. 92–116.

[72] Moayeri, M., Rezaei, K., Sanjabi, M., and Feizi, S. Text2concept: Concept activation vectors directly from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 3744–3749.

[73] Molnar, C. *Interpretable machine learning*. Independently published, 2020.

[74] Mu, J., and Andreas, J. Compositional explanations of neurons. *Advances in Neural Information Processing Systems 33* (2020), 17153–17163.

[75] Oikarinen, T., Das, S., Nguyen, L. M., and Weng, T.-W. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations* (2023).

[76] Oikarinen, T., and Weng, T.-W. CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *The Eleventh International Conference on Learning Representations* (2023).

[77] Osteoarthritis Initiative. Osteoarthritis Initiative (OAI) Data. https://nda.nih.gov/oai/. Accessed: October 23, 2025.

[78] Poché, A., Hervier, L., and Bakkay, M.-C. Natural example-based explainability: a survey.

[79] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (2021), pp. 1–52.

[80] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning* (2021), PMLR, pp. 8748–8763.

[81] Rajagopal, D., Balachandran, V., Hovy, E. H., and Tsvetkov, Y. Selfexplain: A self-explaining architecture for neural text classifiers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (2021), pp. 836–850.

[82] Ramaswamy, V. V., Kim, S. S., Fong, R., and Russakovsky, O. Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 10932–10941.

[83] Rao, S., Mahajan, S., Böhle, M., and Schiele, B. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision* (2024), Springer, pp. 444–461.

[84] Rasheed, B., Abdelhamid, M., Khan, A., Menezes, I., and Khatak, A. M. Exploring the impact of conceptual bottlenecks on adversarial robustness of deep neural networks. *IEEE Access* (2024).

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

[86] Rigotti, M., Miksovic, C., Giurgiu, I., Gschwind, T., and Scotton, P. Attention-based interpretability with concept transformers. In *International Conference on Learning Representations* (2022).

[87] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence 1*, 5 (2019), 206–215.

[88] Rymarczyk, D., Struski, Ł., Górszczak, M., Lewandowska, K., Tabor, J., and Zieliński, B. Interpretable image classification with differentiable prototypes assignment. In *European Conference on Computer Vision* (2022), Springer, pp. 351–368.

[89] Rymarczyk, D., Struski, Ł., Tabor, J., and Zieliński, B. Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021), pp. 1420–1430.

[90] Sarkar, A., Vijaykeerthy, D., Sarkar, A., and Balasubramanian, V. N. A framework for learning ante-hoc explainable models via concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10286–10295.

[91] Sawada, Y., and Nakamura, K. Concept bottleneck model with additional unsupervised concepts. *IEEE Access 10* (2022), 41758–41765.

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

[93] Shin, S., Jo, Y., Ahn, S., and Lee, N. A closer look at the intervention procedure of concept bottleneck models. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022* (2022).

[94] Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations (ICLR)* (2014), ICLR.

[95] Sinha, S., Huai, M., Sun, J., and Zhang, A. Understanding and enhancing robustness of concept-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2023), vol. 37, pp. 15127–15135.

[96] Srivastava, D., Yan, G., and Weng, L. Vlg-cbm: Training concept bottleneck models with vision-language guidance. *Advances in Neural*

Information Processing Systems 37 (2024), 79057–79094.

[97] Sun, C.-E., Oikarinen, T., Ustun, B., and Weng, T.-W. Concept bottleneck large language models. In *The Thirteenth International Conference on Learning Representations* (2025).

[98] Tan, Z., Chen, T., Zhang, Z., and Liu, H. Sparsity-guided holistic explanation for llms with interpretable inference-time intervention. In *Proceedings of the AAAI conference on artificial intelligence* (2024), vol. 38, pp. 21619–21627.

[99] Vielhaben, J., Bluecher, S., and Strodthoff, N. Multi-dimensional concept discovery (mcd): A unifying framework with completeness guarantees. *Transactions on Machine Learning Research* (2023).

[100] Wang, B., Li, L., Nakashima, Y., and Nagahara, H. Learning bottleneck concepts in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 10962–10971.

[101] Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. Caltech-ucsd birds 200.

[102] Wu, Z., D'Oosterlinck, K., Geiger, A., Zur, A., and Potts, C. Causal proxy models for concept-based model explanations. In *International Conference on Machine Learning* (2023), PMLR, pp. 37313–37334.

[103] Xu, Y., Yang, X., Gong, L., Lin, H.-C., Wu, T.-Y., Li, Y., and Vasconcelos, N. Explainable object-induced action decision for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 9523–9532.

[104] Xuanyuan, H., Barbiero, P., Georgiev, D., Magister, L. C., and Liò, P. Global concept-based interpretability for graph neural networks via neuron analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2023), vol. 37, pp. 10675–10683.

[105] Yamaguchi, S., and Nishida, K. Explanation bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2025), vol. 39.

[106] Yan, S., Yu, Z., Zhang, X., Mahapatra, D., Chandra, S. S., Janda, M., Soyer, P., and Ge, Z. Towards trustable skin cancer diagnosis via rewriting model's decision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 11568–11577.

[107] Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., and Yatskar, M. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023).

[108] Yeh, C.-K., Kim, B., Arik, S., Li, C.-L., Pfister, T., and Ravikumar, P. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems 33* (2020), 20554–20565.

[109] Yuksekgonul, M., Wang, M., and Zou, J. Post-hoc concept bottleneck models. In *ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data* (2022).

[110] Zeiler, M. D., and Fergus, R. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13* (2014), Springer, pp. 818–833.

[111] Zhang, Q., Wu, Y. N., and Zhu, S.-C. Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 8827–8836.

[112] Zhang, R., Madumal, P., Miller, T., Ehinger, K. A., and Rubinstein, B. I. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2021), vol. 35, pp. 11682–11690.

[113] Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., and Torralba, A. Object detectors emerge in deep scene cnns. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015), Y. Bengio and Y. LeCun, Eds.

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

[115] Zhou, B., Sun, Y., Bau, D., and Torralba, A. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 119–134.

A GLOSSARY OF TERMS

In the following, we provide a glossary of the most important term in the field of Concept-based XAI. We hope this glossary will serve readers as a unique reference source, allowing future literature to use the different terms consistently. For the sake of readability, we always reintroduced these terms upon their first mention. In Table 4, instead, we provide a list of the acronyms and abbreviations.

Explainability	The capacity of a method to provide intelligible explanations for model predictions or decisions. It should unveil the inner workings and factors that contribute to the model's outcomes in a manner that is understandable to humans.
Transparency	The characteristics of a model whose inner processes, mechanisms, and logic are clear and comprehensible to users. A transparent model should reveal how it reaches specific predictions, enhancing user trust and aiding in identifying potential biases or errors.
Interpretability	It is the degree to which the outcomes and decision-making of a machine learning model can be grasped and reasoned about by humans. An interpretable model reveals the relationships between input data and predictions, facilitating a deeper understanding of its functioning.
Local Expl.	Explanation focusing on the behavior of a model for a specific prediction. It provides insights into why the model made a particular decision for a given input instance.
Node/Filter Expl.	Explanation focusing on individual nodes, or filters, within a neural network. It helps decipher the patterns or concepts that targeted nodes consider during predictions.
Global Expl.	It provides insights into a model's behavior in general. It identifies patterns, or overall rules holistically governing the model decision process.
Post-hoc Expl.	The outcome of a reverse engineering procedure trying to understand the patterns or the pathways that led to a specific prediction. It does not interfere with the model standard working.
Model Agnostic	Technique that is not dependent on the specific algorithm or model architecture used to provide explanations. It generally just analyses the model input and output, hence It can be applied to a variety of ML models.
Counterfactual Expl.	Explanations showing how altered input data could change model predictions. They create modified scenarios to show why a model made a specific prediction and how input changes would result in different outcomes.
Feature Importance	It measures the contribution of each input feature to a model's predictions. It helps identify which features have the most significant influence on the model's outcomes (e.g., pixel images).
Saliency Map	A saliency map is a visual representation highlighting the important regions or features of an input that significantly affect a model or a node's behavior.
Concept	High-level, human-understandable abstraction that captures key features or patterns within data (e.g., object parts, colors, textures, etc.). Concepts are used to simplify complex information, enabling easier interpretation and communication of AI model behavior.
Concept-based Expl.	Type of explanations aiming to elucidate model predictions by decomposing them into human-understandable concepts. Rather than providing explanations regarding feature importance, these methods provide explanations in terms of high-level entity attributes (i.e., concepts).

	These explanations aim to bridge the gap between complex model decisions and intuitive human reasoning.
Prototype	A representative example or part of an example that captures essential features of a group of samples (e.g., belonging to the same class). Prototypes can be regarded as concepts since they capture higher-level abstraction than raw input features.
Concept-based Model	Type of machine learning model designed to make predictions while employing high-level concepts or abstractions rather than working directly with raw data. These models enhance interpretability by relying on human-understandable concepts.
Concept Bottleneck	Design principle of certain concept-based models. It involves incorporating an intermediate layer within the model's architecture to represent a given set of concepts. This layer constrains the flow of information by forcing the model to represent its predictions using these predefined concepts, possibly decreasing the model classification performance.
Concept Score	Numerical value that quantifies the relevance or presence of a specific concept within the context of a machine learning model's decision or prediction.
Concept Embedding	Representation of concepts into a numerical form. Concept embeddings map complex human-understandable concepts into numerical representations, providing a model of more information than what can be described by a single score.
Logic Expl.	An explanation outlining the step-by-step logical reasoning process leading to a particular decision or prediction. It presents a sequence of logical conditions (i.e., a rule) used by the model to reach its conclusion, e.g., $paws(x) \wedge tail(x) \wedge muzzle(x) \rightarrow dog(x)$.
Probing Method	Technique to assess how the latent representations captured by a given model can discriminate a set of concepts. This method involves training an auxiliary concept-specific model on top of the neural network's latent representation. Probing methods enable researchers to uncover the underlying concepts learned by deep models.
Concept Intervention	Technique to assess and modify a machine learning prediction based on modification of the predicted concepts. They enable domain experts to test hypothetical scenarios to improve the model's prediction and provide concept-based counterfactual explanations.
Generative Model	Type of machine learning model that learns the underlying patterns and structures of a dataset to generate new, realistic instances similar to the training data. Used in concept-explainability to generate counterfactual examples for causality assessment and for defining concepts.

B FORMAL CONCEPT ANALYSIS

We now formally define the notion of concepts, using as reference the category theory literature [17, 19, 71]. This area has examined the topic from various perspectives, providing a robust base for our definition. A concept is a set of attributes that agrees on the intention of its extension [71], where the extension of a concept consists of all objects belonging to the concept and the intention of a concept is the set of attributes that are common to all these objects.

More precisely, consider a context defined as the triple (O, A, I), where we indicate with O a set of objects, with A a set of attributes, and with I the satisfaction relation (also called "incidence") which is a subset of $A \times O$. We also define a subset of objects $\hat{O} \subset O$ and a subset of attributes $\hat{A} \subset A$, a derivation operator $\hat{A}' = \{o \in O | (o, a) \in I, \forall a \in \hat{A}\}$ and dually $\hat{O}' = \{a \in A | (o, a) \in I, \forall o \in \hat{O}\}$. We define $(\hat{A}, \hat{O})$ as a Formal Concept if and only if $\hat{B}' = \hat{A}$ and $\hat{A}' = \hat{B}$, or,

Acronym	Full name
ACE [18]	Automatic Concept-based Explanations
AE	Auto Encoder
AI	Artificial Intelligence
BotCL [61]	Bottleneck-Concept Learner
CaCE [21]	Causal Concept Effect
CAM [75]	Class Activation Map
CAV [30]	Concept Activation Vector
CAR [11]	Concept Activation Region
CBM [33]	Concept Bottleneck Models
CBM-AUC [54]	CBM with Additional Unsupervised Concept
CEM [14]	Concept Embedding Models
CME [28]	Concept-based Model Extraction
CNN	Convolution Neural Network
CPM [62]	Causal Proxy Model
CRAFT [15]	Concept Recursive Activation FacTorization
CT [49]	Concept Transformer
CW [9]	Concept Whitening
C-XAI	Concept-based XAI
DMA [35]	Disjoint Mechanism Analysis
DCR [3]	Deep Concept Reasoning
DeconvNet [70]	Deconvolutional Network
DL	Deep Learning
DNN	Deep Neural Network
GDPR	General Data Protection Regulation
Grad-CAM [56]	Gradient-weighted Class Activation Mapping
GNN	Graph Neural Network
GNN-CI [65]	GNN with Concept Interpetability
HPNET [22]	Hierarchical Prototype Network
IBD [77]	Interpretable Basis Decomposition
ICE [73]	Invertible Concept-based Explanation
IMA [35]	Independent Mechanism Analysis
LaBO [67]	Language in a bottle
LEN [10]	Logic Explained Networks
LIME [48]	Local Interpretable Model-agnostic Explanation
LLM	Large Language Model
MCD [59]	Multi-dimensional Concept Discovery
ML	Machine Learning
MLP	Multi-Layer Perceptron
ND [5]	Network Dissection
Net2Vec [16]	Network to Vector
PCA	Principal Component Analysis
PCBM [69]	Post-hoc Concept Bottleneck Models
ProtoPNets [8]	Prototype Parts Networks
SENN [2]	Self-Explaining Neural Network
SHAP [39]	SHAPley additive explanation
STCE [27]	Spatial Temporal Concept Explanation
TCAV [30]	Testing with CAV
VAE	Variational Auto Encoder
XAI	EXplainable AI

Table 4. Table of acronyms and abbreviations.

also, if $\hat{A}'' = \hat{A}$ and $\hat{O}'' = \hat{O}$. In other words, a concept is a set of attributes that are shared among a set of objects and

for which no other attribute exists that is shared among all the objects, even though other attributes can characterize some objects. According to [19], this is the broadest possible concept definition. Indeed, it allows us to consider all the previously identified typologies of concepts according to what we consider as the set of attributes A.

C EVALUATION SUPPLEMENTARY MATERIAL

C.1 Metrics Description

We now deepen the analysis provided in Section 7.1 with a short description of each metric within each category outlined.

C.1.1 Concept effect on class prediction and task performance.

(i) Concept effect on class prediction.

- T-CAV score: it is the fraction of the class's inputs with a positive conceptual sensitivity [18, 30], where the conceptual sensitivity is how much each concept influences the prediction of a given class. So, it measures how many samples have concepts whose effects lie positively on the final prediction.
- CaCE: it measures the causal effect of the presence of concepts on the final class prediction [21, 62].
- ConceptSHAP: it assesses the effect of a concept on the final completeness score [68]. The completeness is defined as the amount to which concept scores are sufficient for predicting model outcomes.
- Smallest Sufficient Concepts (SSC): it computes the class accuracy when employing only a subset of the concepts, looking for the smallest set of concepts [18].
- Smallest Destroying Concepts (SDC): it looks for the smallest set of concepts deleting, which causes prediction accuracy to fall the most [18, 59].
- STCE Concept Importance: it computes the importance rank of each concept for a given class [27].
- T-CAR score: it quantifies the proportion of instances belonging to a specific class whose representations fall within the positive concept region [11].
- Sobol score: it measures the contribution of a concept and its interactions of any order with any other concepts to the model output variance [15].
- Concept Weight: in concept-based models employing a linear layer from concept to task, the same weight represents the importance of a concept for a given class [67, 69].
- Concept Relevance: the predicted importance of a concept for a final class in a concept-based model, determined differently for each sample [2, 3, 49].

(ii) Concept effect on task performance.

- Completeness score: it measures the amount to which concept scores are sufficient for predicting model outcomes [68]. This is based on the assumption that the concept scores of *complete* concepts are sufficient statistics of the model prediction.
- Fidelity: it counts the number of times in which the original and approximate model predictions are equal over the total number of images [15, 53, 73].
- Faithfulness: this measures the predictive capacity of the generated concepts [53]. It represents the capability of the overall concept vector to predict the ground truth task label.
- Concept Efficiency: in concept-based models, the concept capacity to predict the task depends on the number of concepts employed [14]. Richer concept representations mitigate this problem.

- Conciseness: similar to the concept efficiency but for post-hoc methods, it represents the number of concepts required to reach a certain grade of completeness and so a desired level of final accuracy [59].

C.1.2 Quality of Concepts.

i) Concepts properties.

- Mutual Information: it quantifies the amount of information retained in a concept representation with respect to the input data [14].
- Distinctiveness: it quantifies the distinction between different concepts by considering the extent of their coverage [61].
- Completeness: Different from the previous definition of completeness given in [68], it measures how well a concept covers a specific associated class in the dataset [61].
- Purity: it quantifies the ability to discover concepts covering only a single class [61].

(ii) Relation of Concepts with internal network representation.

- Intersection over Union (IoU): it measures the degree of overlapping of the bounding boxes representing the concepts and the activation maps of the single node [5, 16, 65] (and adopted [72] with the name of *Part Interpretability*).
- Location stability: it assesses the consistency in how a filter represents the same object part across various objects [72].

(iii) Concept prediction error.

- Concept Error: it measures how close the concepts learned are to the concept ground truth [33] (and used in [53] with the name Explanation Error). It involves computing the L2 distance between the concepts learned and the ground truth concepts to measure the alignment.
- AUC score: this score measures whether the samples belonging to a concept are ranked higher than others [9]. That is, the AUC score indicates the purity of the concepts.
- Misprediction-overlap metric (MPO): it computes the sample fraction in the test set with at least m relevant concepts predicted incorrectly [28]. A larger MPO score implies a bigger proportion of incorrectly predicted concepts.

C.2 Dataset Table

As introduced in Section 7, we report in Table 5 the most used datasets to evaluate concept-based explanation methods. We considered only those where the authors either explicitly describe the concept annotations or the procedure followed to extract them (Extracted). We classify the datasets based on the data type: images (IMG), text (TXT), videos (VID), tabular (TAB), time series (TS), and graph data (Graph).

C.3 Methods Resources

We report here the analysis of the resources. Rather than describing the singular resources, here we report tables 6, 7 that contain, for each method, the type of resources employed in terms of metrics, dataset, and human evaluation, following the categorization provided in Section 4.1.

Table 5. Datasets employed by the reviewed methods in terms of type of data, size, task label and concepts. We report only the dataset in which the concepts are explicitly defined. All datasets have been defined for classification tasks. We report with *, the datasets suitable for regression too.

Type of Data	Dataset	Used in	Size	Label	Concepts
IMG	MNIST [34]	[11] [2] [61] [36]	60000	Digits	Geometrical attributes
	MNIST even/odd [4]	[49]	60000	Digit'parity	Digit value
	Fashion MNIST [63]	[36]	60000	Clothes	Details of clothes
	colored-MNIST [29]	[21]	60000	Digits	Digit's colours
	Caltech-UCSD Birds-200 (CUB) [60]	[11] [73] [35] [33] [31] [28] [72] [61] [8] [51] [50] [13] [54] [53] [67] [42] [14]	11788	Bird species	Visual attributes of the birds
	BRODEN [5]	[5] [77] [16]	63305	Concepts	1197 visual concepts.
	CelebA [38]	[21] [40] [14] [3]	202599	Person's visual attributes	Non-overlapping person visual attributes
	OsteoArthritis Initiative (OAI)* [43]	[33]	36369	Osteoarthritis levels	Knee conditions relative to the osteoarthritis
	Places365 [76]	[74] [9] [42]	10million	Scenes	Objects landmarks
	Animals with Attribute (AwA and AwA2) [25]	[68] [31] [53]	37322	Animals	85 numeric attribute
	HAM10000 [58]	[69] [67]	10000	Malignant/Benignant	Characteristics of malignant skin lesions
	SIIM-ISIC [24]	[69]	33126	Malignant/Benignant	Characteristics of malignant melanoma
	MPI3D [20]	[40]	>1million	Objects	Object attributes and robot's sensory measurements
TXT	CEBaB [1]	[62]	>15000	Restaurant reviews' sentiment	Aspect-level sentiment
	IMBD [23]	[68]	50000	Movie reviews' sentiment (pos/neg)	Extracted
	SST-2 [57]	[47]	>70000	Movie reviews' sentiment (pos/neg)	Extracted
	SST-5 [57]	[47]	>11000	Movie reviews' sentiment (5 classes)	Extracted
	TREC-6/50 [37]	[47]	>5000	Question types (6 classes/50 classes)	Extracted
	SUBJ [44]	[47]	9000	Subjective/objective.	Extracted
VID	Kinetics-770 [7]	[27]	65000	Human actions (700 classes)	Extracted
	KTH Action [55]	[27]	>2000	Human actions (6 classes)	Extracted
	BDD-OIA [64]	[54]	>22000	Human actions (4 classes)	21 concepts reated to driving conditions
TAB	COMPAS [46]	[2]	>60000	Recidivism (yes/no)	Table attributes
	V-Dem [45]	[4, 10]	202	State Democracy level	Table attributes
	MIMIC-II [52]	[4, 10]	>40000	Patient survival (dead/alive)	Table attributes
TS	MIT-BIH Electrocardiogram (ECG) [41]	[11]	47	Heartbeat (normal/abnormal)	Heart issues
GRAPH	MUTAG [12]	[65]	188	Node labels (7 classes)	Extracted
	Reddit-binary [32]	[65]	2000	Type of graphs (question/answer-based community or a discussion-based community)	Extracted
	PROTEINS [6]	[65]	>1000	Type of protein (enzymes/not enzymes)	Extracted
	IMBD-Binary [66]	[65]	1000	Movies	Extracted

The release of new datasets is infrequent due to the resource-intensive nature of creating concept-annotated datasets. It requires up to $C \times N$ extra annotations, where C is the number of concepts and N the number of samples. Sometimes, a class-level concept annotation is used to reduce this effort, applying the same concept annotation to all the samples in a class, thus reducing the number of annotations to $N \times Y$, Y being the number of classes.

Table 6. Post-hoc Concept-based Explainability methods. We characterize the approaches based on the following dimensions. The *Code/Models availability, Data release*: (✓), no (✗). *Human evaluation* (Human eval), if a user study is conducted for evaluation purposes: (✓), no (✗). Full category description in Section 4.

	Method	Code/Mod. Avail.	Data Release	New metric	Human eval
Supervised	T-CAV [30]	✓	✗	✓	✓
	CAR [11]	✓	✗	✓	✗
	IBD [77]	✓	✗	✗	✓
	CaCE [21]	✗	✗	✓	✗
	CPM [62]	✓	✗	✗	✗
	Obj. Det [74]	✗	✗	✗	✓
	ND [5]	✓	✓	✓	✓
	Net2Vec [16]	✓	✗	✗	✗
	GNN-CI [65]	✓	✗	✗	✗
Unsupervised	ACE [18]	✓	✗	✗	✓
	Compl. Aware [68]	✓	✗	✓	✓
	ICE [73]	✓	✗	✗	✓
	CRAFT [15]	✓	✗	✓	✓
	MCD [59]	✓	✗	✗	✗
	DMA & IMA [35]	✓	✓	✗	✗
	STCE [27]	✓	✗	✗	✗

REFERENCES

[1] Abraham, E. D., D'Oosterlinck, K., Feder, A., Gat, Y., Geiger, A., Potts, C., Reichart, R., and Wu, Z. Cebab: Estimating the causal effects of real-world concepts on nlp model behavior. *Advances in Neural Information Processing Systems 35* (2022), 17582–17596.

[2] Alvarez Melis, D., and Jaakkola, T. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems 31* (2018).

[3] Barbiero, P., Ciravegna, G., Giannini, F., Espinosa Zarlenga, M., Magister, L. C., Tonda, A., Lio, P., Precioso, F., Jamnik, M., and Marra, G. Interpretable neural-symbolic concept reasoning. In *Proceedings of the 40th International Conference on Machine Learning* (23–29 Jul 2023), vol. 202 of *Proceedings of Machine Learning Research*, PMLR, pp. 1801–1825.

[4] Barbiero, P., Ciravegna, G., Giannini, F., Lió, P., Gori, M., and Melacci, S. Entropy-based logic explanations of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2022), vol. 36, pp. 6046–6054.

[5] Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 6541–6549.

[6] Borgwardt, K. M., Ong, C. S., Schönauer, S., Vishwanathan, S., Smola, A. J., and Kriegel, H.-P. Protein function prediction via graph kernels. *Bioinformatics 21*, suppl_1 (2005), i47–i56.

[7] Carreira, J., Noland, E., Hillier, C., and Zisserman, A. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987* (2019).

[8] Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems 32* (2019).

[9] Chen, Z., Bei, Y., and Rudin, C. Concept whitening for interpretable image recognition. *Nature Machine Intelligence 2*, 12 (2020), 772–782.

[10] Ciravegna, G., Barbiero, P., Giannini, F., Gori, M., Lió, P., Maggini, M., and Melacci, S. Logic explained networks. *Artificial Intelligence 314* (2023), 103822.

[11] Crabbé, J., and van der Schaar, M. Concept activation regions: A generalized framework for concept-based explanations. *Advances in Neural Information Processing Systems 35* (2022), 2590–2607.

[12] Debnath, A. K., Lopez de Compadre, R. L., Debnath, G., Shusterman, A. J., and Hansch, C. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry 34*, 2 (1991), 786–797.

[13] Donnelly, J., Barnett, A. J., and Chen, C. Deformable protopnet: An interpretable image classifier using deformable prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10265–10275.

[14] Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., Lió, P., and Jamnik, M. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing*

Table 7. Explainable By-Design Concept-based Models. We characterize the approaches based on the following dimensions. The *Code/Models availability, Data release*: (✓), no (✗). *Human evaluation* (Human eval), if a user study is conducted for evaluation purposes: (✓), no (✗).: (✓), no (✗). *Human evaluation* (Human eval), if a user study is conducted for evaluation purposes: (✓), no (✗). Full category description in Section 4.

	Method	Code/Mod Avail.	Data Release	New Metric	Human eval
Supervised	CBM [33]	✓	✗	✗	✗
	LEN [4, 10, 26]	✓	✗	✗	✓
	CEM[14]	✓	✗	✓	✗
	ProbCBM [31]	✓	✗	✗	✗
	DCR[3]	✓	✗	✗	✗
	CW [9]	✓	✗	✓	✗
	CME [28]	✓	✓	✓	✗
	PCBM [69]	✓	✗	✗	✗
	CT [49]	✓	✗	✗	✗
Unsupervised	Interp. CNN [72]	✓	✗	✗	✗
	SENN [2]	✗	✗	✗	✗
	SelfExplain [47]	✓	✗	✗	✓
	BotCL [61]	✓	✗	✗	✓
	PrototypeDL [36]	✓	✗	✗	✗
	ProtoPNet [8]	✓	✗	✗	✗
	ProtoPool [50]	✓	✗	✗	✓
	DeformableProtoPNet [13]	✓	✗	✗	✗
	HPnet [22]	✓	✗	✗	✗
	ProtoPShare [51]	✓	✗	✗	✓
Hybrid	CBM-AUC [54]	✗	✗	✗	✗
	Ante-hoc expl. [53]	✓	✗	✗	✗
	GlanceNets [40]	✓	✓	✗	✗
Generative	LaBO [67]	✓	✗	✗	✓
	Label-free CBM [42]	✓	✗	✗	✗

Systems (2022), S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Curran Associates, Inc., pp. 21400–21413.

[15] Fel, T., Picard, A., Bethune, L., Boissin, T., Vigouroux, D., Colin, J., Cadène, R., and Serre, T. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 2711–2721.

[16] Fong, R., and Vedaldi, A. 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* (2018), pp. 8730–8738.

[17] Ganter, B., and Wille, R. *Formal concept analysis: mathematical foundations*. Springer Science & Business Media, 2012.

[18] Ghorbani, A., Wexler, J., Zou, J. Y., and Kim, B. Towards automatic concept-based explanations. *Advances in neural information processing systems 32* (2019).

[19] Goguen, J. What is a concept? In *International Conference on Conceptual Structures* (2005), Springer, pp. 52–77.

[20] Gondal, M. W., Wuthrich, M., Miladinovic, D., Locatello, F., Breidt, M., Volchkov, V., Akpo, J., Bachem, O., Schölkopf, B., and Bauer, S. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *Advances in Neural Information Processing Systems 32* (2019).

[21] Goyal, Y., Feder, A., Shalit, U., and Kim, B. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165* (2019).

[22] Hase, P., Chen, C., Li, O., and Rudin, C. Interpretable image recognition with hierarchical prototypes. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (2019), vol. 7, pp. 32–40.

[23] IMDb. IMDb Non-Commercial Datasets. https://developer.imdb.com/non-commercial-datasets/. Accessed: October 23, 2025.

[24] International Skin Imaging Collaboration (ISIC). ISIC Challenge 2020. https://challenge2020.isic-archive.com/. Accessed: October 23, 2025.

[25] IST Austria. AwA2: An Attribute Database for Animal Behavior Analysis. https://cvml.ist.ac.at/AwA2/. Accessed: October 23, 2025.

[26] Jain, R., Ciravegna, G., Barbiero, P., Giannini, F., Buffelli, D., and Lio, P. Extending logic explained networks to text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (2022), Association for Computational Linguistics, pp. 8838–8857.

[27] Ji, Y., Wang, Y., and Kato, J. Spatial-temporal concept based explanation of 3d convnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 15444–15453.

[28] Kazhdan, D., Dimanov, B., Jamnik, M., Liò, P., and Weller, A. Now you see me (cme): concept-based model extraction. *arXiv preprint arXiv:2010.13233* (2020).

[29] Kim, B., Kim, H., Kim, K., Kim, S., and Kim, J. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 9012–9020.

[30] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning* (2018), PMLR, pp. 2668–2677.

[31] Kim, E., Jung, D., Park, S., Kim, S., and Yoon, S. Probabilistic concept bottleneck models. In *Proceedings of the 40th International Conference on Machine Learning* (2023), ICML'23, JMLR.org.

[32] Kipf, T. N., and Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[33] Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *International conference on machine learning* (2020), PMLR, pp. 5338–5348.

[34] LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/* (1998).

[35] Leemann, T., Kirchhof, M., Rong, Y., Kasneci, E., and Kasneci, G. When are post-hoc conceptual explanations identifiable? In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence* (31 Jul–04 Aug 2023), R. J. Evans and I. Shpitser, Eds., vol. 216 of *Proceedings of Machine Learning Research*, PMLR, pp. 1207–1218.

[36] Li, O., Liu, H., Chen, C., and Rudin, C. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2018), vol. 32.

[37] Li, X., and Roth, D. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics* (2002).

[38] Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (December 2015).

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

[40] Marconato, E., Passerini, A., and Teso, S. Glancenets: Interpretable, leak-proof concept-based models. *Advances in Neural Information Processing Systems 35* (2022), 21212–21227.

[41] Moody, G. B., and Mark, R. G. The impact of the mit-bih arrhythmia database. *IEEE engineering in medicine and biology magazine 20*, 3 (2001), 45–50.

[42] Oikarinen, T., Das, S., Nguyen, L. M., and Weng, T.-W. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations* (2023).

[43] Osteoarthritis Initiative. Osteoarthritis Initiative (OAI) Data. https://nda.nih.gov/oai/. Accessed: October 23, 2025.

[44] Pang, B., and Lee, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075* (2005).

[45] Pemstein, D., Marquardt, K. L., Tzelgov, E., Wang, Y.-t., Krusell, J., and Miri, F. The v-dem measurement model: latent variable analysis for cross-national and cross-temporal expert-coded data. *V-Dem working paper 21* (2018).

[46] ProPublica. COMPAS Analysis Repository. https://github.com/propublica/compas-analysis, Year Accessed. Accessed: October 23, 2025.

[47] Rajagopal, D., Balachandran, V., Hovy, E. H., and Tsvetkov, Y. Selfexplain: A self-explaining architecture for neural text classifiers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (2021), pp. 836–850.

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

[49] Rigotti, M., Miksovic, C., Giurgiu, I., Gschwind, T., and Scotton, P. Attention-based interpretability with concept transformers. In *International Conference on Learning Representations* (2022).

[50] Rymarczyk, D., Struski, Ł., Górszczak, M., Lewandowska, K., Tabor, J., and Zieliński, B. Interpretable image classification with differentiable prototypes assignment. In *European Conference on Computer Vision* (2022), Springer, pp. 351–368.

[51] Rymarczyk, D., Struski, Ł., Tabor, J., and Zieliński, B. Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021), pp. 1420–1430.

[52] Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., and Mark, R. G. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine 39*, 5 (2011), 952.

[53] Sarkar, A., Vijaykeerthy, D., Sarkar, A., and Balasubramanian, V. N. A framework for learning ante-hoc explainable models via concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10286–10295.

[54] Sawada, Y., and Nakamura, K. Concept bottleneck model with additional unsupervised concepts. *IEEE Access 10* (2022), 41758–41765.

[55] Schuldt, C., Laptev, I., and Caputo, B. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* (2004), vol. 3, IEEE, pp. 32–36.

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

[57] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. Recursive deep models for semantic compositionality over