Explainable AI for Audio and Visual Affective Computing: A Scoping Review

David S. Johnson, Olya Hakobyan, Jonas Paletschek, and Hanna Drimalla

(Survey paper)

***Abstract*—Affective computing often relies on audiovisual data to identify affective states from non-verbal signals, such as facial expressions and vocal cues. Since automatic affect recognition can be used in sensitive applications, such as healthcare and education, it is crucial to understand how models arrive at their decisions. Interpretability of machine learning models is the goal of the emerging research area of Explainable AI (explainable AI (XAI)). This scoping review aims to survey the field of audiovisual affective machine learning to identify how XAI is applied in this domain. We first provide an overview of XAI concepts relevant to affective computing. Next, following the recommended PRISMA guidelines, we perform a literature search in the ACM, IEEE, Web of Science and PubMed databases. After systematically reviewing 1190 articles, a final set of 65 papers is included in our analysis. We quantitatively summarize the scope, methods and evaluation of the XAI techniques used in the identified papers. Our findings show encouraging developments for using XAI to explain models in audiovisual affective computing, yet only a limited set of methods are used in the reviewed works. Following a critical discussion, we provide recommendations for incorporating interpretability in future work for affective machine learning.**

***Index Terms*—Affective machine learning, audiovisual, interpretability, XAI.**

I. Introduction

HUMANS convey their affective states through various non-verbal signals, from facial expressions to subtle modulations in voice. In contrast to physiological signals, such as heart rate, audiovisual cues can be directly observed and help understand the affective states of others. The direct observability of audiovisual signals also means that they can be captured without intrusive sensors, typically only requiring a camera and microphone. This makes audiovisual data a valuable resource for affective computing, facilitating unobtrusive affect recognition in various domains. For instance, there are efforts towards enhanced medical diagnosis by detecting conditions like depression based on audio and visual cues [1], [2]. Furthermore, automatic affect detection is important for human-computer interactions as it would enable machines to understand human emotions and interact accordingly [3] - mirroring the way people naturally pick up on audiovisual cues in everyday interactions.

The automated detection of affect is often achieved by complex machine learning (ML) methods, such as deep learning. These methods can handle high-dimensional data in the audio and visual domain and achieve high performance in tasks such as facial expression recognition [4], [5], speech emotion recognition [6], [7] or sentiment analysis [8]. However, one problem associated with these methods is their black-box nature, i.e. a lack of transparency about how the models arrive at their decisions. Poor interpretability poses a challenge for identifying decisions that may lead to undesirable outcomes. For instance, it is known that machine learning models can be biased against different demographic groups in tasks directly related to affect detection, such as facial analysis [9]. As affective ML often deals with sensitive applications, such as healthcare and education [3], understanding model decisions is crucial for ensuring that everyone benefits from the technology and nobody suffers unfair disadvantages.

The concerns around black-box models in affective computing are not only ethical but also regulatory. The recently enacted Artificial Intelligence Act (AI Act) [10] also has implications for the field of affective ML, recognizing potential risks in emotion recognition systems. Specifically, it classifies emotion recognition systems that are able to “infer emotions or intentions on the basis of biometric data” [10] – including audiovisual behavioral data – as high-risk systems. Such systems are now governed by strict requirements in areas such as transparency and human oversight that emphasize the need for interpretability. Furthermore, emotion recognition systems have additional transparency and right-to-explanation regulations, to ensure that individuals exposed to such AI-systems are fully aware of the role AI plays in decisions that affect them. These regulations underscore the importance of developing AI-systems that are interpretable not only for ethical reasons, but as to comply with regulatory frameworks.

To address the call for more transparency in AI systems, the field of XAI has gained significant interest in the recent years [11], [12], [13]. This includes several approaches of “opening the black box” by providing insights into how models

Received 9 February 2024; revised 8 October 2024; accepted 11 November 2024. Date of publication 22 November 2024; date of current version 27 May 2025. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via the Transregional Collaborative Research Centre 318 “Constructing Explainability” under Grant TRR 318/1 2021 - 438445824. In addition, we acknowledge the Federal Ministry of Education and Research Germany (BMBF; 01IS20046). Recommended for acceptance by R. Subramanian. *(Corresponding author: David S. Johnson.)*

The authors are with the Center for Cognitive Interaction Technology (CITEC), Bielefeld University, 33615 Bielefeld, Germany (e-mail: djohnson@techfak.uni-bielefeld.de; ohakobyan@techfak.uni-bielefeld.de; jpaletschek@techfak.uni-bielefeld.de; drimalla@techfak.uni-bielefeld.de).

Digital Object Identifier 10.1109/TAFFC.2024.3505269

make their decisions and which aspects of the data have the highest impact on the model outcomes. Similar to other machine learning fields, affective computing is beginning to adopt XAI methods for improved model interpretability.

Affective computing presents unique challenges that set it apart from other domains, due to its focus on human behavior. Audiovisual data in affective ML is often unstructured, with complex and nuanced spatiotemporal and time-frequency relationships that can vary across individuals, cultures and contexts. These characteristics make XAI approaches from other fields, where model predictions are based on more straightforward, objective and localized causes, less suited to the explanatory needs of users of affective computing systems. Additionally, affective ML data is inherently more subjective, with ground truths that are less clear and harder to objectively validate. Given such challenges, research on the implications of applying XAI to affective computing data remains limited, prompting the need for a comprehensive review, especially in light of the regulatory demands introduced by the European AI Act.

Review Rationale and Objectives

This scoping review aims to survey the use of XAI methods within affective ML, with a specific emphasis on their application to audio and visual data. The focus on affective computing sets this work apart from existing XAI surveys [11], [12], [13] which do not detail the extent and specifics of how interpretability is applied in this domain. Within the field of affective computing, our choice to focus on audiovisual data is grounded in two factors. First, the implicit nature of non-verbal audiovisual cues highlights the need for interpretability due to subtle nuances embedded in these signals going beyond explicit words. Second, audiovisual cues are also explicit in the sense that they can be directly observed by unobtrusive sensors, such as cameras, in contrast to other non-verbal signals like heart rate. This property makes the use of audiovisual data more scalable, opening up possibilities for applications that may impact large populations. Model transparency is therefore needed to enable the responsible use of emerging affective computing applications.

To our knowledge, there are only two reviews discussing the use of XAI in affective ML. First, Cortiñas-Lorenzo and Lacey [14] surveyed several methods applied in affective computing without specifically targeting audiovisual data. Although they described a variety of methods in the field and provided examples, their approach did not offer a systematic analysis and *quantitative* summary of how often these methods were used and in which contexts. Furthermore, limited details were provided regarding the review protocol and methodology, posing a challenge for replication. Second, in a recent preprint [15], we offered an initial exploration into the XAI methods applied for affective tasks using audiovisual data. Here, we extend our previous work by offering a more comprehensive, systematic, and detailed examination of the subject.

We contribute to the field of affective computing by reviewing papers in affective ML that use audiovisual data and implement methods for enhancing model interpretability. Following a comprehensively recorded review protocol, we extract, quantitatively analyze and summarize the XAI methods identified in the selected papers. The goal is to provide a roadmap for researchers seeking to incorporate interpretability into their affective computing systems. The following research questions are targeted:

RQ1: What audiovisual affective ML tasks is XAI being applied to?
RQ2: How is XAI being applied to these tasks?
RQ3: Why is XAI being applied to the given tasks?
RQ4: How are the XAI approaches being evaluated?
RQ5: What are the gaps and challenges that should be addressed in the field?

In the next sections, we first introduce important XAI concepts relevant to the scope of this article (Section II). Next, we describe our review methodology, detailing the systematic approach used for literature analysis (Section III). In section Section IV we summarize our findings and finally, provide a critical discussion in Section V.

II. Explainable AI

Explainable AI (XAI) has emerged in recent years as a field of machine learning research that aims to make models and model decisions interpretable to human users. There are considerable variations within the field in the types of explanations and the methods to generate them. In this section, we provide a brief overview of the main concepts, taxonomies, and methods of XAI to give the reader a high-level view of the variety of approaches available. While not all of the methods described are currently applied in affective ML, we believe that covering the full range of XAI is important to advancing research in this domain. Readers that would like a deep dive into XAI are encouraged to read existing reviews and taxonomies [11], [12], [13], [16].

A. Interpretability Versus Explainability

With research in XAI still in its early stages, there are no universally accepted definitions of the commonly used terms interpretability and explainability, and they are often used interchangeably. However, in this work, similar to Graziani et al. [16], we refer to *interpretability* as a characteristic of models and methods indicating how understandable they are to human observers. *Explainability* in this regard, refers to approaches that take explicit steps to unveil the decision-making of an algorithm with the goal of making the model or prediction more understandable [12]. To this end, we consider interpretability as a broader goal to strive for when developing audiovisual affective ML methods and XAI as one of the methods used to reach this goal.

B. Explanation Stage

A common approach to categorize explanation methods is based on the *stage* in the machine learning process at which the explanation is generated [11]. Vilone and Longo [11] suggest there are two main categories, *ante hoc* approaches in which explanation generation is designed into the model itself and *post hoc* approaches which implement methods to generate explanations from existing black-box models (see Fig. 1).

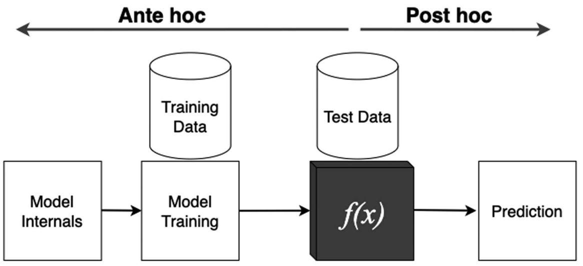
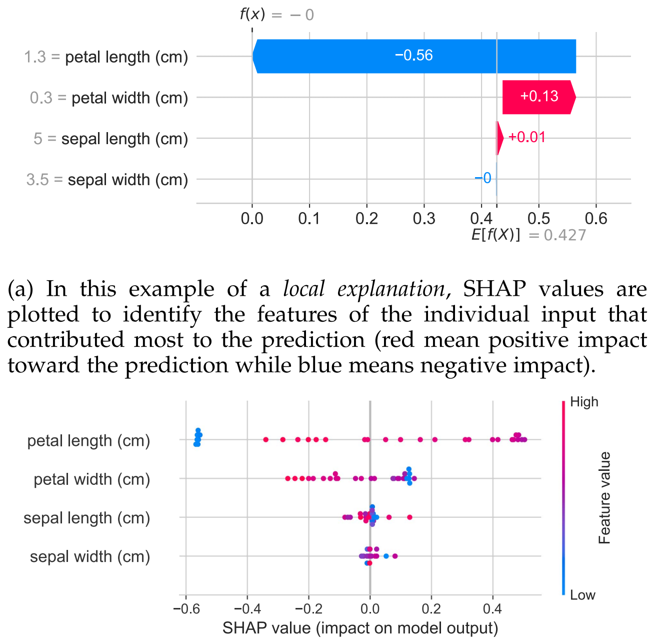


Fig. 1. Explanation Stage refers to the component in the machine learning process used to generate an explanation.



(a) In this example of a *local explanation*, SHAP values are plotted to identify the features of the individual input that contributed most to the prediction (red mean positive impact toward the prediction while blue means negative impact).

(b) In this example of a *global explanation*, SHAP values of individual predictions are aggregated into a single plot to show the overall importance of each feature for the model.

Fig. 2. Explanation scope refers to the level of the model that is being explained. *Global explanations* aim to illustrate general model behavior, while *local explanations* explain how the model generates an output for a specific input.

Ante hoc explainability approaches include so-called transparent models [12] which are inherently understandable by humans, including simple linear or tree-based models. Most transparent models, however, suffer from low capacity and may not be able to fully capture the complexity or non-linearity of the underlying data leading to low accuracy on complex tasks [12]. The recent emergence of attention mechanisms has led to an ante hoc approach that works with complex deep learning models by outputting attention weights activated by an input [17]. Yet, most other methods that are able to model this complexity, such as decision tree ensembles or deep learning methods, lack inherent interpretability.

Post hoc explainability methods generate model explanations based on an analysis of the input, the output and optionally, the trained model [11], [12]. Post hoc methods that consider the model during the analysis, known as *model-specific*, take advantage of a priori knowledge regarding the internal components of the model when generating explanations. Using the internal knowledge enables model-specific approaches to be more computationally efficient, but restricted to specific model architectures. *Model-agnostic* [18] approaches, on the other hand, work with any type of model, even when the model architecture is unknown, as is common with proprietary models. However, this comes at the cost of higher computational resources to compute the explanations.

C. Explanation Scope

Explanation scope refers to the level at which the explanations attempt to provide insights into the model [16]. Both ante and post hoc methods may generate explanations with either a *global* or a *local* scope [11], [16]. *Global explainability* approaches (Fig. 2(b)), aim to provide users with a holistic understanding of the model behavior and inference process. A complete global understanding of a model is hard to achieve, especially as the number of parameters increases [19]. By contrast, *local explainability* approaches (Fig. 2(a)) generate explanations for only a local region of the input space around a single prediction without trying to provide a holistic understanding of the model.

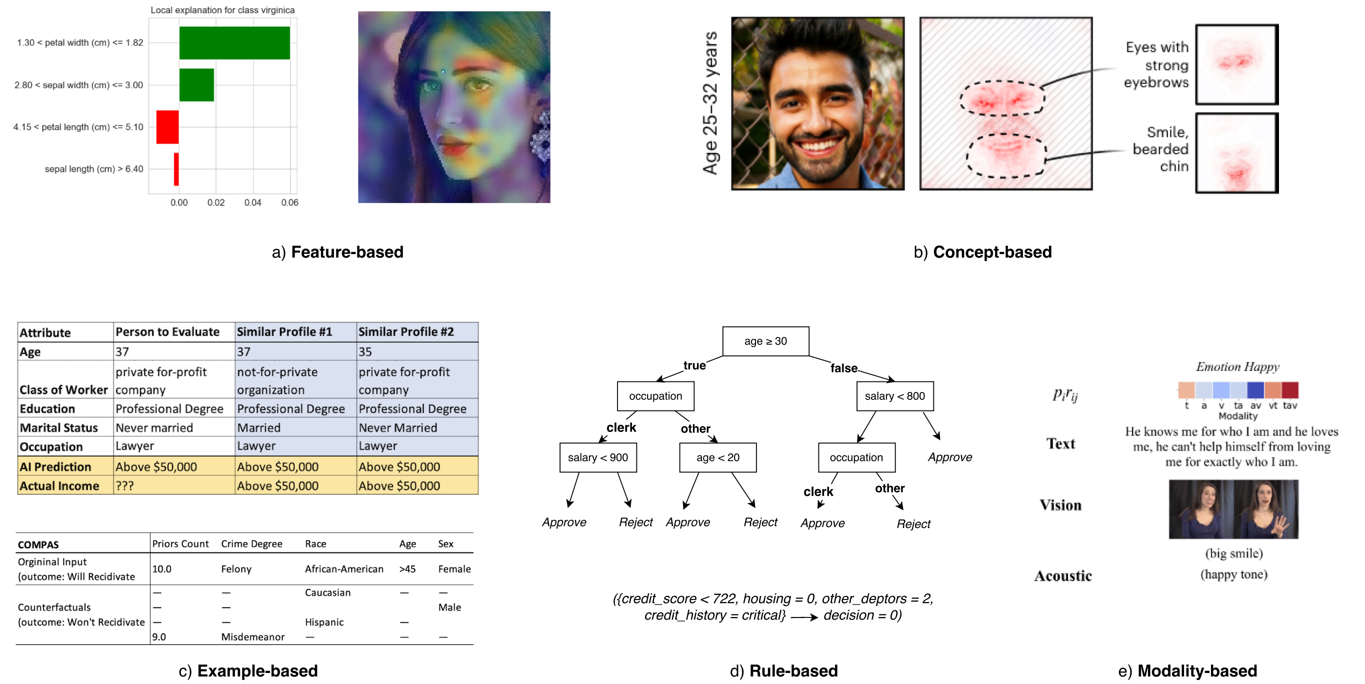
D. Modalities and Input

The methods used for generating explanations may significantly differ based on the modality and type of the data. Therefore, input modality is an important aspect that should be considered when applying XAI [11]. In this review, the focus is placed on the audio and visual modalities, which are typically either raw data input or extracted features.

Raw input is represented differently depending on the modality. For the visual modality, raw input is represented as matrices of pixel intensities, in the case of images, or sequences of image frames for videos. Regarding audio, raw input is represented as a 1-dimensional vector of audio signal amplitudes sampled at frequencies such as 16,000 and 44,100 Hz.

Extracted features, on the other hand, are pre-calculated representations of qualities important to the specified task as deemed by an expert. In the visual modality, feature representations include numerical features describing spatial information [20], or descriptive semantic features about the face [21]. Extracted features for the audio modality are typically numerical representations of important audio signal properties, such as attributes regarding a signal's frequency, energy, or temporal aspects [22]. Another method to extract features is the use of deep learning embedding models that have been pre-trained to learn a compressed representation of the input [23]. Performing feature extraction in the audiovisual domain, via expert knowledge or deep learning embeddings, often results in high-dimensional feature sets that have limited interpretability.

To enhance interpretability it is important that the features themselves are also interpretable. In the visual domain, facial action units (AUs) [21] provide an interpretable set of features by representing the intensity and presence of facial muscle movements that are intuitive even to lay users. Similarly, in


Local explanation for class virginica
1.30 < petal width (cm) <= 1.82
2.80 < sepal width (cm) <= 3.00
4.15 < petal length (cm) <= 5.10
sepal length (cm) > 6.40
0.00
0.02
0.04
0.06
Age 25–32 years
Eyes with strong eyebrows
Smile, bearded chin

a) **Feature-based**

b) **Concept-based**

Attribute	Person to Evaluate	Similar Profile #1	Similar Profile #2
Age	37	37	35
Class of Worker	private for-profit company	not-for-private organization	private for-profit company
Education	Professional Degree	Professional Degree	Professional Degree
Marital Status	Never married	Married	Never Married
Occupation	Lawyer	Lawyer	Lawyer
AI Prediction	Above $50,000	Above $50,000	Above $50,000
Actual Income	???	Above $50,000	Above $50,000

COMPAS	Priors Count	Crime Degree	Race	Age	Sex
Orginінal Input (outcome: Will Recidivate	10.0	Felony	African-American	>45	Female
Counterfactuals (outcome: Won't Recidivate	—	—	Caucasian	—	—
	—	—			Male
	—	—	Hispanic	—	
	9.0	Misdemeanor	—	—	—

c) **Example-based**

d) **Rule-based**

e) **Modality-based**

Fig. 3. The types of explanations according to our taxonomy are categorized into five different categories. a) *Feature-based* approaches explain models by identifying the most important features, e.g., using LIME [18] (left) or the most important pixels displayed as saliency maps via methods like GradCAM [24] (right). b) *Concept-based* approaches identify the relevance of certain high-level concepts found important for the task (image adapted from [25]). c) *Example-based* approaches explain predictions by identifying similar examples from the training data [26] (top; image adapted from [26]) or by identifying counterfactual examples that indicate how to change the input to obtain a more desirable outcome [27] (bottom; image adapted from [27]). d) *Rule-based* approaches generate explanations as rule sets that are easy for humans to understand and can be represented as a decision tree (top; image adapted from [28]) or as a set of conditional statements (bottom) [28] e) *Modality-based* approaches explain the contributions of individual modalities (in the image listed as *t*, *a*, or *v*) and modality interactions (in the image as *ta*, *av*, *vt*, *tav*) towards the final prediction (image adapted from [29]).

the audio domain, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [22] feature set aims for a minimalistic set for easier interpretation of the meaning of the extracted parameters.

E. Explanation Types

In this review, we propose a categorization of XAI methods based on the conceptual framing of the explanation. This includes whether the explanation is generated in terms of *features*, *concepts*, *rule-sets*, *examples* or *modalities* (see Fig. 3). We propose that this taxonomy is more intuitive for intra- and inter-type comparison and evaluation. Furthermore, this approach applies to both ante hoc and post hoc methods. Note that this is a different approach than taken by other taxonomies, which focus on the explanation output type (e.g. numerical, textual, visual) [11] or families of post hoc methods [12], [16].

1) Feature-Based Explanations: Feature-based explanation methods aim to make a model (global scope) or single prediction (local scope) more interpretable by ranking the contribution of input features, known as feature importance or feature attribution scores [30]. The main idea of these methods is to calculate scores that represent the importance of the input features for the overall model or a single prediction. Fig. 3(a) shows two examples of feature-based explanations, one for tabular data (left) and one for raw image data (right). There is a wide breadth of approaches for calculating these values.

Ante hoc approaches for feature-based explanations include the use of transparent models that are inherently interpretable. This inherent interpretability may be constrained to specific model scenarios. For instance, linear models are typically considered transparent but only with low-dimensional input spaces that afford human understanding of the model's feature weights. Other transparent models include fuzzy models that generate rule sets understandable to humans [11]. Transparent methods often result in global explanations since they describe the inner workings of the overall model. A more recent ante hoc approach towards interpretability is to implement deep learning models with attention mechanisms [31]. The attention weights activated during inference are considered as features to which the model pays the most attention, representing importance scores. These importance scores aim to provide insight into why the model made a prediction for a specific input by highlighting which features led to the prediction.

Post hoc explainability is the most common approach for feature-based explanations, especially for models that lack inherent transparency. It can be categorized into four approaches for calculating feature importance values: *model simplification* methods, *perturbation-based* methods *gradient-based* methods, and methods based on *Shapley values*.

Model simplification approaches first learn a transparent surrogate model from which feature importance scores are extracted [12], [16]. For example, the Local Interpretable Model-Agnostic Explanations (LIME) [18] method is a well-known model simplification approach. LIME generates local explanations from any type of model by training a sparse linear model on a local neighborhood of synthetic data points that are created by perturbing the original input instance.

Perturbation-based approaches apply small modifications to the input and observe how the output changes [32]. Analysis of the relation of the feature perturbations to the model output is used to calculate which features are most important to the model output.

Backpropagation-based methods calculate feature importance values by back-propagating an importance signal, such as gradients [33] or relevance values [34], from the output of a neural network back to the input [35], [36]. These methods are mostly used with image-based models to calculate the feature attribution values for each pixel of an input image [24], [34], [36].

Another approach for calculating feature importance values is to use *Shapley values* [37], a game-theoretic approach for transferable utility (TU) in cooperative games. Shapley values aim to calculate the contribution of an individual player in a cooperative game by calculating the average marginal contribution of that player to all coalitions that include the player. In terms of XAI, each feature is treated as a “player” and the prediction is treated as the “game output”. In this way, the contribution scores are considered as feature relevance values.

To present the calculated relevance scores to users, the values of feature relevance are visualized in various formats depending on the type of input data. In the case of audiovisual datasets, the visualizations are typically of two formats: sorted bar plots (or similar) and saliency maps as seen in Fig. 3(a). Bar plots are used if the input to the model is in a tabular format as with extracted features. For image-based methods, relevance values are represented via saliency maps where each pixel is assigned a color based on its relevance value.

2) Example-Based Explanations: Example-based explanations aim to address the complexity of other XAI methods by generating simple, easy-to-understand explanations that give users intuition about model decisions without needing to understand the inner workings of the model [38]. Examples are typically generated by selecting representative prototypes from the training data [26], [39], e.g. samples that are representative of a given class, that help users better understand the model judgments and identify mistakes made by the model(see Fig. 3(c) top).

Counterfactual explanations are another example-based approach, in which generated samples illustrate how to obtain a more desirable prediction by changing certain features of the input sample [27], [40] (see Fig. 3(c) bottom). Counterfactual explanations are typically generated through optimization approaches that perturb the original input to create a hypothetical example that is similar to the original input while changing the output to the desired prediction [40]. Due to the computational complexity, counterfactual methods are most commonly developed for tabular data, but there are also some emerging image based methods using generative AI approaches [41], [42].

3) Concept-Based Explanations: A major drawback to feature-based explanations is that highlighting where a model is looking is not the same as what a model is “seeing” [43]. *Concept-based* explanations aim to address this interpretability gap by providing explanations in terms of concepts, i.e., human-understandable abstractions of the input data [44] (see Fig. 3(b)). Supervised concept-based approaches require experts to define concepts in advance, with each concept having a set of pre-defined examples for the calculation of concept importance values [44]. However, important concepts are not always known beforehand and may be challenging to identify. Unsupervised approaches [45], [46] aim to identify concepts directly learned in a model’s latent space. The semantic meaning of the identified concepts, however, may not be entirely human-understandable or may fall prey to the same interpretability gap as feature attribution methods [47].

4) Rule-Based Explanations: Rule-based methods offer an alternative approach to gaining insight into a model by generating or extracting rule sets from the model. For example, decision trees are a transparent (ante hoc) model, in which decision rules can be extracted by following the decision path [28]. Rule sets are simpler compared to numerical feature attribution methods [13] and provide a logical and structured format, making it easy to simulate a model’s decision [48] (see Fig. 3(d)). However, the interpretability of the rule sets can suffer when the models become too large [12] or if high-dimensional feature spaces, such as those in audio and visual modalities, are used [49]. Fuzzy rule-based systems [50] aims to address this trade-off between interpretability and complexity using fuzzy logic to generate rule sets using induction and optimization. Many rule-based methods, however, work best on tabular data and are not suitable for raw audiovisual input.

5) Modality-Based Explanations: Existing XAI reviews and taxonomies have focused on methods for unimodal models, but in the field of audiovisual ML multimodality plays an important role. For this reason, we propose a category of *modality-based* explanations. We define them as explanations that help users understand the importance and contribution of modalities to model decisions (see Fig. 3(e)). For multimodal approaches, Liang et al. [51] describe three core principles: *heterogeneity*, *connections*, and *interactions*. *Heterogeneity* refers to the fact that each modality contains diverse representations and information. When modalities *interact* they provide new information to the model not present in a single modality, while *connection* between modalities refers to information that is shared by them [51].

While the heterogeneity of the modalities can be explored by applying traditional unimodal approaches to the individual modalities, other methods are needed to quantify the interactions and connections between modalities. Ante hoc approaches for measuring cross-modal interactions use carefully designed networks to measure interactions between modalities [52]. One approach in this regard is the use of attention mechanisms in the fusion layer of the network [53]. By carefully designing the fusion layers, the attention weights can be used to indicate which modality or modality interactions are most important for the final output. Similarly, modality routing networks [29], inspired by capsule networks [54], are designed to model modality interactions in an interpretable way using a routing mechanism, the coefficients of which represent the cross-modal importance. Since ante hoc approaches require specific model designs, the networks are not typically generalizable to other modalities or ML tasks. Post hoc methods [52], [55], [56], on the other hand, offer a more general approach toward measuring modality importance and interactions that make no assumptions about the underlying modalities of the model.

F. Evaluating Interpretability

Due to the breadth of XAI types and methods, evaluating their interpretability is important for selecting the correct approach for a given task. Explanation evaluation is a challenging task as there is often no ground truth for quality assessment [57]. Hence, explanations should be evaluated with defined objectives that act as a proxy for interpretability, and with human-based approaches [57], [58]. Doshi-Velez and Kim [58] called for a more rigorous evaluation of XAI methods and proposed an evaluation taxonomy to help with this. The taxonomy proposes three categories of evaluation that inform each other: functionally-grounded, human-grounded, and application-grounded approaches.

Functionally-grounded approaches evaluate explanations using quantitative formalizations of explanations that act as proxies for explainability. These so-called proxy metrics offer a way to quantitatively verify that an explanation satisfies certain desirable properties [57], including metrics for model fidelity and metrics for important aspects of interpretability, such as simplicity, clarity, and broadness. Functionally-grounded approaches offer an inexpensive method for evaluating explanations without human involvement at a very general level.

Interpretability, however, is inherently linked to human understanding. For this reason, the taxonomy also includes approaches for human-centered evaluations: *human-grounded* and *application-grounded* evaluations [58]. Human-grounded evaluations aim to evaluate general notions of explainability by evaluating XAI methods on specific tasks with lay users. An example of this, known as forward simulation, asks a user to correctly simulate the model's output based on the explanation. Application-grounded evaluations, on the other hand, evaluate aspects of explainability that are specific to a given XAI application with real users of the system. Application-ground approaches directly evaluate the objectives of a real-world system but this comes with the cost of expensive studies with expert users.

Metrics in human-based evaluations can be either subjective or objective in nature [57]. Subjective metrics are typically evaluated via questionnaires to gather user opinions on aspects such as trust or preference. For systematic evaluation of qualitative factors Holzinger et al. [59] proposed the System Causability Scale (SCS), a questionnaire aimed to measure the quality of explanations for providing a causal understanding of the model decision. Objective metrics, on the other hand, are used to gather objective results about the effects of a method towards a specific property or goal, such as objective trust, i.e., how often a user follows a model's prediction, or task performance, i.e., how well the user is able to make the correct prediction [57].

III. Methods

A. Literature Search Methodology

To identify research in XAI for audio and visual affective ML, we employed a systematic search following the PRISMA extension for scoping reviews [60] and the updated review guidelines [61]. We employed the IEEE, ACM, PubMed, and Web of Science electronic databases for the search. We searched each database without explicit start and end dates, on Sept. 26, 2023, for key terms in the title, abstract, and keywords using the query:

```
("affective computing" OR (("emotion" OR
"facial expression" OR "affect") AND (recog*
OR detect* OR classif* OR predict* OR esti-
mat* OR "machine learning")) OR ("sentiment
analysis" AND multimodal))
AND (explainab* OR interpretab* OR "XAI").
```

The "*" indicates wildcard search, and the quotations are used for exact term search. As many papers on sentiment analysis rely on text rather than audiovisual input, we only considered this task in multimodal settings where audiovisual data was likely to be included.

The search resulted in a total of $N = 1702$ articles from the four databases, itemized per database in Fig. 4. After removing duplicates from the results, three of the authors independently screened all titles and abstracts for the inclusion and exclusion criteria. The selected papers were required to fulfill all of the following inclusion criteria:

INC1 Prediction of affect, emotion, facial expressions, sentiment, pain, stress or related psychological variables (e.g. personality traits, engagement, depression).

INC2 Use of unimodal or multimodal machine learning models with audio and/or visual data of facial behavior.

INC3 Proposing or applying XAI methods.

Works that met one of the following exclusion criteria were removed from the results:

EXC1 Assessment of the emotional impact of audio or visual stimuli rather than the emotional state of people (e.g. music emotion recognition, visual affect recognition, image generation).

EXC2 Models in which no audio or visual data was used (e.g. only text, EEG, etc) or if the data was collected using intrusive sensors (e.g. motion capture sensors).

EXC3 Explainability was claimed but not demonstrated, or if explainability was applied to model decisions not involving audio or visual data (e.g. sentiment from text).

EXC4 Theses or reviews.

After an independent review, the authors then met in person to resolve conflicts. $N = 1057$ articles were excluded in this round, many of the papers failing to fulfill the inclusion criteria. In the majority of cases, the word *affect* was used as a verb or the task was not in the domain of affective computing ($N = 914$). Other common reasons were that the articles did not use audiovisual data, were review papers or did not include an XAI component.

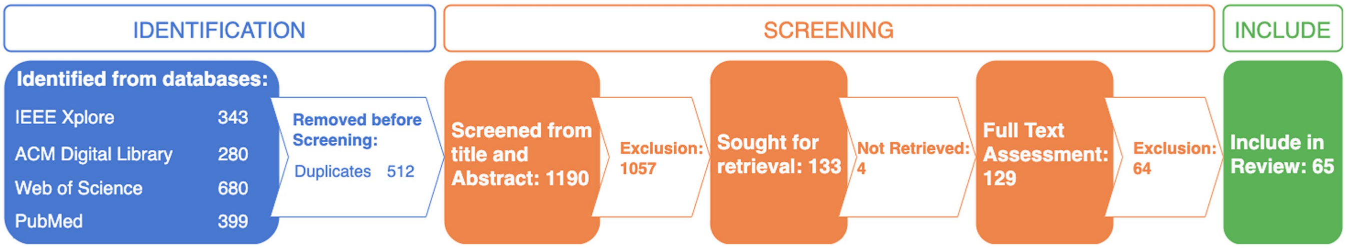


Fig. 4. Flowchart of PRISMA-based selection process.

Of the resulting $N = 133$ articles $N = 4$ were not accessible, leaving $N = 129$ articles for further full-text screening. For this round, each author independently reviewed a subset of two-thirds of the articles to ensure that each article had two reviewers assigned. After the screening, the authors again met in person to clear up the remaining conflicts. In this round, the main exclusion reason was that many of the papers merely mentioned interpretability but did not include any interpretability methods in the sense of XAI ($N = 59$). After this round of screening, $N = 65$ articles were included for the review. The full screening process is shown in Fig. 4.

B. Data Extraction and Analysis

To identify and extract relevant concepts from the included works, each of the three reviewing authors studied a set of either $N = 43$ or $N = 44$ papers. After independent data extraction, the authors met to discuss the results and to identify and correct any mismatches in data extraction.

Extracted information from papers included metadata, such as the publication year and authors of the paper. Next, to answer **RQ1**, information related to the machine learning implementation was extracted, such as the affective computing task (e.g. emotion recognition), used model (e.g. CNN), input type (e.g. raw or extracted features) and modalities (audio, visual, etc.). Another type of information extracted concerned the aspects of interpretability to address **RQ2**. These included the stage (posthoc or ante hoc) and scope (local or global) of the explanations as well as the exact XAI approach. Finally, addressing **RQ3** and **RQ4**, the objective of applying XAI (e.g. model validation, feature analysis) and the evaluation method were examined.

IV. Results

The literature search resulted in $N = 65$ articles that employed XAI methods in audio and visual affective ML up to September 2023, the date of the query. Fig. 5 shows that the research is relatively new with a focus on XAI starting in 2017 and growing since then, although one article approached interpretability as early as 2008. Table I and Fig. 6 provide an overview of the general results in terms of XAI concepts, such as explanation type, stage, scope for different input modalities or affective computing (AC) tasks. In the following sections, we review the implementation of the different XAI concepts and in Section IV-F we discuss the analysis of the purpose of applying XAI and how the methods were evaluated.

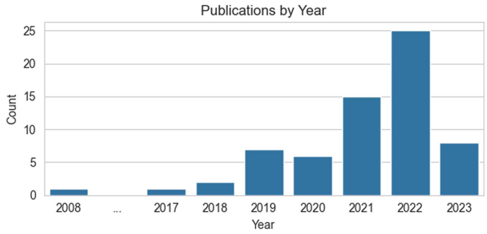


Fig. 5. Publication years of the selected works.

A. Research Task

To better understand where XAI has been applied within the field of audiovisual affective ML (**RQ1**), this section presents the main tasks identified from the selected articles. The papers employed a wide range of tasks, including well-known tasks such as emotion recognition and sentiment analysis and more niche tasks such as pain detection or job candidate screening.

By far the most common task was emotion recognition, typically via facial expressions ($N = 31$) but there were a few instances of speech emotion recognition ($N = 4$) and multimodal emotion recognition ($N = 2$). These works typically implemented one of the two commonly used emotion representation models, categorical emotional states (CES) or dimensional emotional space (DES) [125]. CES-based works aimed to classify emotion categories from basic emotion theory, often with an added "neutral" emotion. On the other hand, DES approaches focused on dimensional affective state recognition using the well-known valance-arousal dimensional approach [94], [100], [104]. Other works focused on categorical affective states such as engagement [98] or affective states derived from the Living in Familial Environments (LIFE) coding system [113]. Rather than detecting emotions or other affective states, two articles researched the detection of AUs (categorized as "expression recognition" in Fig. 6).

While emotion recognition comprised the large majority of the selected works, we also identified other affective computing tasks in which XAI was applied. $N = 6$ articles implemented XAI methods for multimodal sentiment analysis (SA). In these works, SA was performed via categorical sentiment analysis with 2-, 3- and 7-scaled sentiment [55], [106], [110], [114], [116], and one article proposed an SA method for sarcasm detection [112]. Other niche tasks that were identified include pain detection [64], [92], job candidate screening [118], [120],

TABLE I
List of All Papers Categorized by the Methods and Interpretability Concepts Extracted From the Selected Papers

Modality	Method Category	Stage	Scope	Input Type	Articles
Image	Feature-based	PostHoc	Local	Raw Input	[62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82]
				Extracted Features	[77, 79, 82]
				Graph Features	[83]
		Attention	Local	Raw Input	[84, 85]
			Global & Local	Raw Input	[86]
	Example-based	Data	Global	Raw Input	[87]
	Rule-based	Fuzzy	Global	Extracted Features	[88]
	Bayesian Network	Interpretable	Local	Extracted Features	[89]
	Uncertainty	PostHoc	Local	Raw Input	[90]
Video	Feature-based	PostHoc	Local	Raw Input	[91, 92, 93]
				Extracted Features	[91]
			Global	Extracted Features	[94]
				Graph Features	[95]
			Global & Local	Extracted Features	[96]
		Attention	Local	Raw Input	[97]
				Extracted Features	[98]
				Embeddings	[98]
		Interpretable	Global	Extracted Features	[99, 100]
			Global & Local	Extracted Features	[101]
		Sparse	Local	Raw Input	[102]
	Rule-based	Fuzzy	Local	Extracted Features	[103]
		Interpretable	Global & Local	Extracted Features	[101]
Audio	Feature-based	PostHoc	Global	Extracted Features	[104, 105, 106]
		Attention	Global	Extracted Features	[107]
		Interpretable	Global & Local	Raw Input	[108]
	Concept-based	PostHoc	Global & Local	Extracted Features	[109]
Audio, Video, & Text	Modality-based	Attention	Local	Extracted Features	[110, 111, 112]
			Global	Extracted Features	[113]
		Routing	Local	Extracted Features	[114]
			Global & Local	Extracted Features	[28, 115]
		PostHoc	Global & Local	Extracted Features	[55, 116]
	Feature-based	PostHoc	Global	Extracted Features	[117, 118, 119]
			Global & Local	Extracted Features	[55, 116]
		Attention	Global & Local	Extracted Features	[120]
	Concept-based	PostHoc	Global & Local	Extracted Features	[121]
Video & Text	Feature-based	Interpretable	Global	Extracted Features	[122]
Audio & Text	Modality-based	Attention	Local	Extracted Features	[123]
Video & EEG	Feature-based	PostHoc	Local	Extracted Features	[124]

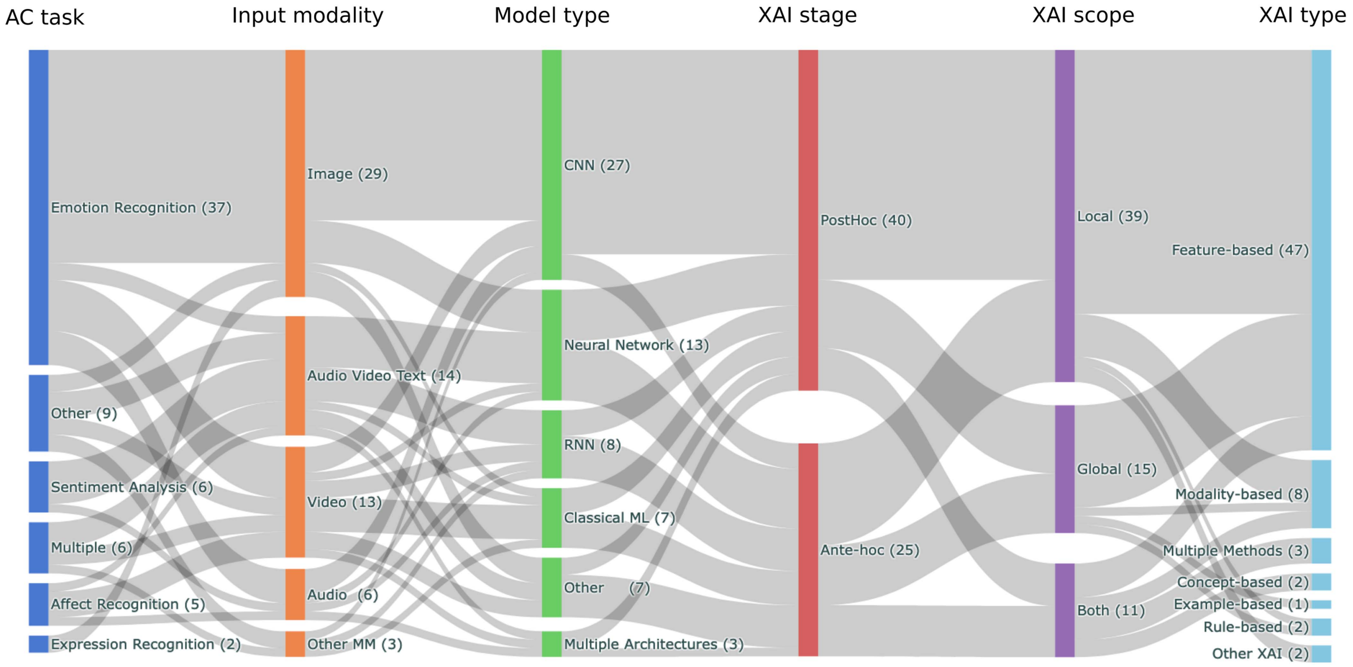


Fig. 6. Overview of the extracted interpretability concepts identified within the selected works.

depression recognition [81], [119], stutter detection [124], and personality trait recognition [122]. Finally, $N = 6$ works included multiple ML tasks, combining either SA and emotion recognition [29], [111], [115], [123] or emotion recognition and pain detection [96], [102].

B. Modalities & Input

In this section, we analyze the distribution of XAI for the different audiovisual modalities and various input representations (**RQ1**). A large portion of the articles worked on unimodal and visual input ($N = 29$ image, $N = 13$ video), with a small portion of the articles using only audio input ($N = 6$). The remaining papers used multimodal inputs, with most relying on the combination of audio, video and text ($N = 14$). A few outliers used only two modalities, such as audio and text [123], video and text [122], as well as video and EEG [124].

From the selected works raw input was used only in unimodal models ($N = 26$). The majority of these were images ($N = 21$) with some videos [92], [93], [97], [102], and one using only audio [108]. A few of the unimodal methods included extracted features in addition to the raw input ($N = 4$). Specifically, Zhu et al. [82] trained two models for generating complementary explanations using images and AUs. Ter Burg and Kaya [79] compared image-based explanations to those generated with handcrafted geometric features. Similarly, image-based explanations were compared with explanations from models trained using facial landmarks [77], [91].

The remaining papers, in both unimodal and multimodal cases, employed various forms of feature extraction as input for affective ML models ($N = 33$). For unimodal models, $N = 16$ works implemented extracted features, including $N = 4$ for image, $N = 7$ for video, and $N = 5$ for audio. The image-based methods employed a variety of feature extraction methods, such as the states of facial regions of interest [89], Garbor-wavelet coefficients [88], ResNet embeddings [90], and facial landmark graphs [83]. Video-based models, on the other hand, mainly used AU features [94], [96], [99], [100], [101], [103], or facial landmarks [95], [103]. Finally, audio-based methods used features such as 2-dimensional spectrograms [104], [106], [109], which can be processed similarly to images, or other expert-based features extracted via signal processing techniques [105], [106], [107]. In multimodal settings, all works used extracted features as input ($N = 17$). For the visual modality, all works used facial features like AUs and facial landmarks as input, except in two cases where different types of facial embeddings were implemented [111], [121]. For audio, classical audio features were used, such as COVAREP and eGeMAPS. For the text modality, word embeddings were used in all cases.

As extracted features may not always be interpretable for lay users, some of the selected works aimed for human-interpretable features as model input. A popular method for inducing interpretability in the image- and video-based approaches was the use of AUs ($N = 17$) or graph-based features from facial landmarks [83], [95]. In the audio domain, eGeMAPS [22] was proposed as a smaller, more interpretable feature set and was identified in $N = 4$ of the selected works [107], [113], [119], [120].

C. Explanation Stage

Here we explore at which stage explanations are generated for specific model types in the selected works (**RQ2**). Of the selected papers the majority apply post hoc explainability methods ($N = 40$), while the rest ($N = 25$) apply a variation of ante hoc interpretability in which explanations depend on specifics of the model architecture or the data.

In Table I, we categorize ante hoc approach based on the level at which the explanation is generated. The traditional approach for ante hoc interpretability is to implement models that are inherently interpretable. In the selected works $N = 6$ implemented explanation methods with models that are interpretable-by-design. Most of these employ classically interpretable models, such as linear models [99], [100], [101], [122], decision trees [101], or Bayesian networks [89]. Anand et al. [108] take a novel approach by integrating a differential audio filter layer into a convolutional neural network (CNN) architecture allowing the model to learn audio filter frequencies that are human-understandable. Alternatively, a large portion of the works ($N = 12$) implement attention mechanisms for ante hoc interpretability. This enables developers to take advantage of the power of deep networks, including standard feed-forward neural networks, CNNs, and recurrent neural networks (RNNs). The attention layer in these articles is used at either the feature- [84], [85], [86], [97], [107], [111], [120], [123], time-[98], or modality-level [110], [112], [113] to identify where the model pays most attention. Other ante hoc approaches identified ($N = 7$) include modality *routing* based on capsule networks [29], [114], [115], fuzzy models for developing rule-sets [88], [103], learning sparse representations [102], and selecting prototypical samples from the training data [87].

In the selected works, post hoc approaches were mostly applied to CNN architectures ($N = 28$). They were also applied to other neural network architectures, including standard feed-forward neural networks ($N = 5$) [79], [90], [93], [105], [118], RNNs ($N = 5$) [55], [106], [116], [117], [121], graph neural networks ($N = 2$) [83], [95], and in one case a transformer architecture [55]. Aside from neural networks, post hoc was also applied to non-interpretable, classical machine learning methods ($N = 4$), such as support vector machines (SVMs) [71], [119] and tree-based ensemble methods [94], [106], [119] among others [106], [119].

D. Explanation Scope

Analysis of the explanation scope (**RQ2**) found that most studies ($N = 39$) used local explanations, while $N = 15$ focused on global explanations and only $N = 11$ considered both. Local explanations frequently ($N = 19$) included information about all prediction classes, e.g. by providing visualizations of relevant input features for at least one instance from each emotion category (see [67], [71], [79] for some examples). However, in some papers, local explanations were limited to a single input instance across the entire dataset [69], [72], [110]. Explanations from different XAI [73], [76], [77], [79], [80] or machine learning [63], [83] methods were compared by examining the decisions made for the same set of inputs. In some papers, examples from each prediction class were drawn from different datasets to examine whether the model decisions aligned across diverse data sources [71], [74]. Global explanations also followed a similar pattern in that many papers separated the relevant features per prediction class (see [94], [100], [117] for some examples), XAI method [118], or dataset [105], [107]. In addition to reporting feature attribution values or attention weights, global explanations also employed ranking or selection of the top features relevant for model decisions [107], [117], [119]. It is worth noting that in some cases the relevant features identified by global explanations were contextualized in a local setting. For instance, Wicaksana and Liem [122] created human-understandable reports for model decisions by sharing which input attributes are *generally* relevant for the model (e.g. amount of spoken words) and then reporting the values of these attributes for a *specific example*.

E. Explanation Types

In this section, we analyze the types of XAI methods that are applied in the selected works (**RQ2**).

1) Feature-Based Explanations: Feature-based explanations were by far the most employed XAI method ($N = 50$, including $N = 3$ that implemented multiple methods), predominantly including post hoc explanations ($N = 37$).

Of post hoc methods, SHapley Additive exPlanations (SHAP) [126] and LIME [18] were two commonly used model-agnostic approaches for generating explanations. Although both are applicable to all feature representations, they were associated with different inputs in the selected papers. Specifically, we found that SHAP-based approaches were typically implemented to explain the importance of extracted audio [105], [106] or visual features [79], often in multimodal settings [55], [117], [118], [119], [124]. There were only a few cases where SHAP was used for explaining raw images [63], [74]. By contrast, LIME was typically used with models based on raw images [70], [72], [75], [80].

Another common approach for explaining raw images was the visualization of class activation maps (CAM) [127] through generalized extensions of the method, such as gradient weighted class activation maps (GradCAM) [24] and GradCAM++ [128]. In the selected papers activation maps were typically used to explain which parts of images were relevant towards predicting a specific class [71], [73], [76], [77], [81]. In some cases, layer-wise relevance propagation (LRP) was used to create saliency maps [78], [80], [82], [92]. Finally, the above-mentioned methods were sometimes used to identify important AUs or facial landmarks contributing towards a prediction when using LIME [82], CAM [91] or LRP [69].

There were also a number of ante hoc methods for generating feature-based explanations. The traditional method of using the feature-weights of linear models was identified in $N = 4$ of the selected works [99], [100], [101], [122]. We identified $N = 7$ works integrating attention mechanisms into the network architecture. The attention weights were used to generate attention maps from input images [84], [85], [86], [97], to calculate feature rankings of feature inputs [107], or to identify important time points of temporal inputs [98], [120]. One work claimed interpretability by proposing a model that learned sparse representations of the input [102]. The approach of Anand et al. [108] to learn audio filter parameters during training is another novel ante hoc method for feature-based explanations, since the learned filter parameters act as a human-understandable set of audio features.

2) Modality-Based Explanations: We found that most of the multimodal approaches implemented modality-based explanations ($N = 10$) to generate information about how modalities contribute to a given prediction. Two ante hoc methods represented the most common approach for this type of explanation. The first approach was to integrate an attention layer into the modality fusion layer [110], [111], [112], [113], [123] with the attention weights capturing modality importance. The second approach took advantage of the routing capabilities inspired by capsule networks [54] for multimodal fusion and calculated modality routing coefficients to represent modality importance scores [29], [114], [115]. For the two post hoc approaches, Shapley values were used as the basis for calculating modality contributions [55], [116].

In most of the works, the calculated scores were used to indicate modality importance or dominance, but other modality interactions are possible. Specifically, Wang et al. [55] calculated modality interactions by aggregating the SHAP values of all features for a given modality to calculate a modality importance score. These scores were then used to calculate if a modality was *dominating*, *complementary*, or *conflicting* with the others. Another approach going beyond modality dominance was employed by Wu et al. [112] who used the attention mechanism to model modality incongruity for improving sarcasm detection.

3) Other Explanation Types: In the selected works there were a few ($N = 8$) alternative approaches beyond feature- or modality-based explanations. These methods almost exclusively included ante hoc explanations with only one paper using a post hoc approach [90].

Rule-based approaches were implemented in three of the works, two of them proposed neuro-fuzzy models that learn human-understandable rules [88], [103] and one proposed using interpretable decision trees in which the learned rules were provided as the explanations [101]. Two papers implemented *concept-based* approaches to generate explanations [109], [121]. In particular, Asokan et al. [121] defined emotion recognition concepts for each modality and used concept activation vectors with statistical testing [44] to calculate the importance of each concept. Zhang and Lim [109], on the other hand, proposed a method that generated counterfactual explanations for speech emotion recognition in terms of how specific speech-related concepts, so-called “contrastive cues”, differ between the original input and a synthetically generated counterfactual speech sample.

Instead of explaining model predictions, Manresa-Yee et al. [87] proposed explaining the dataset through *example-based* explanations from prototypical examples for each class, identified from the dataset using Protodash [39]. Another approach quantified model uncertainty to dissociate biases that arise either due to the properties of the training data or annotator disagreements in the labels [90]. Finally, one work implemented interpretable facial emotion recognition using a Bayesian network [89]. The proposed system performs semantic facial segmentation and identifies states for each semantic region which were used to train a transparent Bayesian network.

F. XAI Application and Evaluation

To answer **RQ3** and **RQ4**, we extracted details regarding the purpose of applying XAI methods to the research and identified the methods for evaluating the application. The needs for evaluation differ depending on the reason for integrating XAI into the research. For example, research that develops a new XAI method may evaluate the explanations differently than research that uses explanations to analyze which features are the most important for a specific task. Therefore, we discuss the evaluation methods in the context of the reason for using XAI.

In the selected works we identified six categories for the application of XAI to audiovisual affective ML which are listed and described in Table II. Furthermore, we identified three main types of evaluations: *qualitative evaluation*, *quantitative evaluation*, and empirical evaluation via *user studies*. In the sections below, we describe details of the selected works for each type of XAI application and a summary of the types of evaluation.

1) XAI Methods & Interpretable Architectures: We found that one of the main reasons for applying XAI was to develop *novel XAI methods and ante hoc architectures* to enable interpretable affective ML ($N = 23$). Most of the papers ($N = 20$) in this category used ante hoc methods, accounting for the vast majority of the ante hoc approaches discussed in Section IV-C.

Of the post hoc approaches ($N = 3$), one proposed a novel implementation of SHAP for improved computational efficiency [74]. Another paper focused on measuring decision uncertainty [90], while Zhang and Lim [109] combined several methods for a diverse set of explanations from feature saliency to emotion concepts.

When developing new XAI methods and architectures, efforts were made to verify that the generated explanations provided effective insights into model predictions. To this end, the main approach was to perform a qualitative evaluation of the generated visualizations or explanations ($N = 15$). In these works, the authors generated and discussed a few example explanations or qualitatively compared them with other methods to show that generated explanations provided reasonable output. One work expanded on this idea by qualitatively comparing modality attention to the modalities focused on by human observers [113].

For a more objective evaluation of XAI effectiveness, some works implemented quantitative analysis ($N = 4$). These quantitative analysis approaches varied for each work, including a comparison of feature importance ranks to other methods [74], [107], performing statistical analysis of the output [90], [120], and implementing a metric that acted as a proxy for explanation effectiveness without human involvement [74], [120]. One work took a human-centric approach to evaluating the effectiveness of generated explanations through user studies [109].

2) Model Improvement & Validation: Another reason identified for the application of XAI to affective ML was to support

TABLE II
Listing of the Purposes and Evaluation Methods for the Selected Works

Purpose	Description	Evaluation	Articles
XAI Methods & Interpretable Architectures	The works in this category proposed novel XAI methods or novel ante hoc ML architectures for introducing interpretability into affective ML tasks.	Qualitative	[28, 83, 85, 86, 95, 97, 98, 103, 108, 110, 111, 113, 114, 115, 123]
		Quantitative	[28, 74, 90, 107, 120]
		User Study	[109]
		None	[88, 89, 102]
Model Improvement & Validation	These works use XAI to improve model performance or to ensure that models are working as expected.	Qualitative	[62, 63, 64, 65, 68, 71, 73, 77, 81, 84, 92, 93, 96, 100, 104, 105, 106, 118]
		Quantitative	[69, 78, 92, 96, 104, 117, 118]
XAI Evaluation	Works in this category evaluated the effectiveness of different XAI methods for specific affective ML tasks.	Qualitative	[66, 72, 80, 91]
		Quantitative	[76, 79]
		User Study	[79]
Application of XAI & Interpretable Systems	These works proposed new interpretable systems for affective ML tasks that included the use of existing XAI methods.	Qualitative	[67, 82]
		Quantitative	[121]
		User Study	[55, 70, 122]
Feature & Modality Analysis	The works in this category used XAI to analyze the role of the input features, or modalities, towards certain affective behaviors and tasks.	Qualitative	[101, 112, 116, 119]
		Quantitative	[94, 124]
Bias Analysis	These works used XAI to analyze and identify biases in affective ML models.	Qualitative	[75, 87]
		Quantitative	[87]

model improvement and validation ($N = 21$). These papers relied exclusively on feature-based explanations, mostly employing post hoc methods with two exceptions [84], [100]. For model *validation* ($N = 17$) explanations were used to verify that model predictions were based on reasonable features or image regions. Works focusing on model *improvement* ($N = 4$) used explanations to first select the top features and subsequently, retrain models with a reduced feature set.

In terms of evaluation, most of the time a qualitative approach was taken ($N = 18$). Normally, the explanations were simply reviewed and judged by the authors, but in some cases, the top features were compared to literature for validation [64], [93], [100] or compared qualitatively with other XAI methods [106]. On the other hand, some works took a quantitative approach to evaluate how the applied XAI approach improved their model ($N = 7$). For example, when using XAI to enhance the model or select better features, performance metrics were used to evaluate how well the proposed approach improved a model [78], [96], [117], [118]. Another quantitative approach was proposed to verify important concepts identified by two different models [92]. In this case, the feature embeddings of the two models were used to predict the presence of the identified concepts, and model performance was used to evaluate which model best learned the concepts. One model identified spectrogram profiles predictive of different emotion categories [104]. They then used this information to transform samples to a different class and evaluate how well the model predicted these transformed classes.

3) XAI Evaluation: XAI was also applied to affective ML tasks to *evaluate the effectiveness of explainability methods* towards the specified tasks ($N = 6$). These works typically evaluated and compared existing feature-based explainability approaches for different affective ML tasks [66], [79], [80]. One work [76] also compared saliency maps derived using XAI methods to human attention maps generated by analyzing which facial regions human annotators deemed important for their decisions. Kadakia et al. [72], on the other hand, compared LIME explanations with a custom algorithm for detecting important facial regions using non-ML-based techniques. Instead of comparing different methods of explanation, Gund et al. [91] compared explanations generated with different model architectures. Finally, an XAI evaluation framework using domain expertise to generate task-specific evaluation criteria that could be quantified as proxy metrics was proposed [69].

For this application of XAI, evaluation was often performed via a qualitative review ($N = 4$) by visually reviewing the explanations and making subjective judgments on their effectiveness. Quantitative metrics were used in $N = 2$ studies. Specifically, Park and Wallraven [76] calculated the overlap of the saliency maps with human attention maps. In another case, explanation fidelity was evaluated by measuring model performance when the top features were incrementally added as model input [79].

Finally, one paper [79] proposed an empirical evaluation with a user study, in which the participants evaluated the perceived quality of the model-generated explanations using the System Causability Scale (SCS) [59].

4) Interpretable Applications & Systems: Rather than proposing novel architectures or methods, some works proposed *interpretable systems* for affective ML tasks ($N = 4$) or the *novel application of existing XAI methods* to affective ML ($N = 2$). The four proposed *interpretable systems* vary in their main task. M^{2}Lens [55] implemented a SHAP-based information visualization system that afforded developers enhanced model debugging and understanding. NOVA [70] is a system for improving the annotation process for human-in-the-loop machine learning. The system integrates LIME explanations to help non-machine learning experts annotate selected data samples. Additionally, Zhu et al. [82] proposed the use of LIME and LRP to generate more transparent and understandable human-robot interactions. Another system was proposed to make job candidate screening more interpretable [122] by providing text-based explanations. These were generated from the importance values of facial action units. In terms of the *novel application of existing XAI methods*, one paper proposed an implementation of concept-based explanations via concept activation vectors as a novel method for explaining multimodal emotion recognition using emotion-specific concepts identified for each modality [121]. Another approach [67] proposed a training framework using LIME and facial action units to enhance facial expression interpretability.

Similar to *novel XAI Methods and interpretable architectures*, evaluation involves judgment on the effectiveness of the systems and applications towards XAI goals specified by their use cases. In two of the articles, this judgment was performed via a qualitative evaluation of the generated explanations with a subjective assessment by the authors. One article that implemented testing with concept activation vectors (TCAV) used a type of quantitative evaluation, typically employed with this approach, that evaluates the accuracy of the learned concept classification models [121]. Three of the works applied a human-centric approach and implemented empirical user studies to evaluate explanation effectiveness. Wang et al. [55] performed case studies with machine learning experts to evaluate the effectiveness of the visualization system for debugging and improving models. Wicaksana and Liem [122] had experts evaluate the explanations as part of the ChaLearn competition on explaining apparent personality [129]. A user study was also implemented to evaluate how effective LIME-based explanations were for non-expert users to improve model performance using the NOVA system for input annotation [70].

5) Feature & Modality Analysis: In the case of method and model validation explanations were analyzed to provide the developers insights into how and why their models are working as they do. On the other hand, some articles ($N = 7$) performed an analysis of the explanations to gain enhanced insights into human behavior for a specified affective ML task. The aim in these cases was to provide empirical insight into how different features [94], [99], [101], [119], [124] or modalities [112], [116] contributed towards different affective states, in order to gain a scientific understanding of the nonverbal behaviors leading to that state.

Qualitative evaluation of the explanations was typically performed to analyze the role of the features and modalities towards the model ($N = 4$). Two works on the other hand used quantitative approaches to gain objective insights from feature analysis. Das et al. [124] perform a statistical analysis of specific feature importance values to identify correlations between the features and model predictions. Additionally, Haines et al. [94] performed statistical analyses of global feature importance values from partial dependence plots to identify which features annotators place their focus on during annotation.

6) Bias Analysis: XAI was also applied as an approach for *bias analysis* in affective ML models ($N = 2$). Two approaches that applied different XAI methods were taken for the analysis of model bias. One case employed a feature-based approach with LIME for understanding how features affected model predictions per gender [75]. Another work used an example-based approach with Protodash to better understand if the training data was gender biased in certain classes [87]. Both cases performed a qualitative review of the explanations to evaluate for biases. However, Manresa-Yee et al. [87] expanded upon the qualitative analysis by performing a simple quantitative evaluation of the ratios of gender in the identified class prototypes.

V. Discussion

We reviewed how XAI approaches are implemented and used in the domain of audiovisual affective ML. Our findings show that research in XAI for affective ML has seen increasing growth in recent years following the trend of XAI as an important field in machine learning research [11]. XAI was predominantly employed in emotion recognition tasks mostly using post hoc, local and feature-based explanations. Nevertheless, the exact implementation and purpose of the employed methods varied from study to study. In the discussion that follows, we address **RQ5** by highlighting relevant gaps and challenges identified from the analysis of the reviewed works regarding interpretability.

In terms of the explanation stage, the prevalence of post hoc methods over ante hoc methods may be attributed to several factors. First, traditional ante hoc methods, which are interpretable-by-design (e.g. linear models), often do not reach the high level of performance achieved by more complex models [12]. This is particularly evident for high-dimensional data, such as images and video. Next, research on XAI for image-based machine learning has had a heavy focus on post hoc methods, such as saliency maps, leading researchers in affective ML towards these approaches. Furthermore, ante hoc methods often rely on the specifics of individual models and thus provide specialized solutions. Examples of such approaches in the reviewed papers included the use of sparse representations, fuzzy rules, or Bayesian networks. By contrast, post hoc methods afford the use of more model-agnostic methods, such as SHAP [126] and LIME [18] or more general model-specific methods that are not task-specific, such as backpropagation-based methods for feature importance [24], [33], [34], [35]. These are more flexible as they can be applied to various machine learning models and data types, including high-performance, complex networks and high-dimensional feature spaces. In addition, post hoc methods often come in different flavors, such as gradient-based SHAP implementations [126] or extensions of activation maps (Grad-CAM [24], GradCAM++ [128]).

An emerging addition to ante hoc methods that enable the use of high-performance models, similar to post hoc explainability, is the use of neural networks with attention mechanisms. While the architectural details vary from model to model, visualizations of attention scores were utilized to introduce feature- and modality-based interpretability to complex models.

In this review, we focused on audiovisual data and found that most explanations were provided for the visual domain, with only limited application of XAI to audio input. We expect this is due to a general lack of XAI approaches for audio data [130] and the additional cognitive demand of processing audio. While images have the added benefit that they are inherently visually interpretable, raw audio representations are not easily interpretable by humans. Therefore, understanding feature attribution values applied to them may prove challenging. Even in cases where 2-dimensional spectrograms are extracted, enabling audio to be processed similarly to an image, the spectrograms are not easily understood by non-expert users. For this reason, novel XAI methods are required for making audio models more explainable. One such approach would be to identify methods that provide the explanations aurally using sonification approaches [130].

When using audio modality most methods used features extracted via audio signal processing. Using expert-curated feature sets, such as eGeMAPS [22], was a common approach to introducing more interpretable features. However, it can be said that the interpretation of many audio features still requires audio expertise and is not feasible for lay users. For this reason, explanations in terms of more abstract high-level features using concept-based approaches would enable a non-expert understanding. This was behind the idea proposed by Zhang and Lim [109], in which counterfactual explanations were generated that enabled a user to compare values of pre-defined audio concepts to those of examples from another class.

In terms of explanation scope, many of the reviewed papers provided explanations for all of the prediction classes. Nevertheless, in the case of local explanations, it was not explained how the visualized examples were chosen. This is especially problematic when explanations for a small subset of images were used to compare XAI methods or machine learning models. In particular, it is not clear to what extent would the explanations for these few examples generalize to other instances in the dataset. Global explanations might be preferable in certain scenarios, such as for comparing different methods and evaluating features associated with different prediction classes. Nevertheless, global explanations pose other limitations.

First, widely used feature attribution methods often work on a local level, especially in image-based approaches. Second, it is known that people differ in the expression of their affective states based on inter-personal factors, such as gender or culture [131] or psychological conditions, such as depression and autism [132]. Additionally, people may show intra-personal differences in their expressions based on their situation, e.g., if they are stressed or face a power imbalance [133]. Local explanations are therefore necessary to explain specific instances and inter-individual variations in the expression of affective states. To ensure that local explanations are representative of a given scenario, explanations across different instances can be aggregated [134] while paying special attention to intricacies associated with aggregating saliency maps [134], [135]. For example, to address the varying position of important facial regions among different samples, aggregation should be done within important regions of the face, identified via semantic segmentation.

Regarding the explanation types, most of the reviewed papers were limited to implementations of feature-based explanations, largely relying on different variations of saliency maps. At first glance, saliency maps look intuitive when superimposed on original input, but in actuality, they only show *which* regions of the input are important and not *what* is important about those regions [25], [43]. This results in an interpretability gap that requires some level of interpretation from the user, introducing the risk of subjective biases when deciphering model decisions. In addition, saliency maps have been criticized for generating visualizations that are not specific to the underlying model or dataset [136]. It has been shown that attention maps may suffer from similar issues and it is unclear how interpretable the feature weights are [17], [31].

Modality-based explanations mostly focused on identifying how modalities or modality combinations contributed to a prediction. Albeit limited, there were some encouraging attempts to examine complex interactions between modalities, such as dominance or conflict [55]. However, modality-based explanations typically did not delve into an analysis of heterogeneity [51] of the modalities by analyzing features *within* each modality. Including both inter- and intra-modality features might lead to more nuanced and robust modality-based explanations, as seen in the visualization system of M^{2}Lens [55].

The use of other types of explanations, such as rule-, concept- or example-based interpretability was very limited, even though these methods have been shown to be effective approaches for generating explanations [25], [26], [137]. These methods can be helpful for obtaining more trustworthy and intuitive explanations. For instance, example-based methods have been shown to reduce over-reliance on untrustworthy models [26], which is important when introducing models into clinical settings for AI-assisted decision-making. Concept-based explainability has been proposed as a potential approach to reduce the interpretability gap of feature-based methods [43]. Traditionally, this approach required an expert selection of concepts [44], some of which may be unknown a priori, but recent work has shown potential using so-called *glocal* explanations to identity and explain important concepts identified with concept relevance propagation (CRP) without the need for predefined concepts [25]. Another promising, yet under-explored category of explanations is the use of counterfactual explanations. Miller [138] suggests that explanations are contrastive in nature since they aim to address why one event happened instead of another (counterfactual) event. In this way, counterfactual explanations are intuitive to humans as they offer scenarios on how input alterations might influence model decisions and explain why one event happened instead of another [138]. This can be relevant for applications, in which the users wish to adapt their behavior for certain outcomes. For instance, users might aim to refine their speech patterns with the help of an AI assistant to project confidence and reduce nervousness.

Independent of the types of explanations implemented, we identified a breadth of objectives for applying XAI in the selected works. While it is promising to see XAI being included in the design and validation of machine learning methods, there were only a few applications of methods that were specifically geared towards affective computing tasks and contexts. Since explaining why an object was detected in an image is quite different than explaining why someone was diagnosed as being depressed, new methods that specifically focus on these types of tasks are needed. Furthermore, we found there was a lack of work that included XAI in real-world affective computing applications, such as AI-assisted mental health assessment or remote screening and treatment. Developing such systems with XAI will require careful consideration of the context in which the explanations are being provided. For example, different explanation methods are needed when the application is used by a mental health practitioner compared to an everyday user using a self-assessment application. Finally, it is still unclear which explanation types work best for the different contexts, and more research is needed evaluating the effectiveness of different explanation types in different affective computing contexts.

In terms of evaluation, these and other uses of interpretability were mostly limited to qualitative judgments, where the authors presented and discussed a few example explanations. This is a considerable limitation given that machine learning models and XAI methods can be selected or disregarded based on such evaluations. The use of a small subset of inputs, mostly for local explanations, further confounds this issue, as discussed above. Nevertheless, there were also encouraging examples towards more objective, quantitative evaluations. The most effective choice of evaluation depends strongly on the goal of interpretability and the target group, as identified in recent frameworks of human evaluation of XAI [139], [140]. While qualitative visualizations might be beneficial to developers when debugging a model at the initial stages, more quantitative approaches should be considered when using interpretability for model selection and improvement.

As affective computing becomes more ubiquitous through integration with mobile phones and remote diagnosis [141] applications are more likely to be used in cases where users are not ML or affective computing experts. Developing XAI approaches for these applications, therefore, requires a human-centric understanding of whom the explanation is for. However, XAI researchers often overlook the perspectives of these end-users [142]. In the reviewed papers we found it to be rare that the recipients of the explanations were considered when choosing or designing the XAI type. Furthermore, only five of the works performed user-based evaluations [55], [70], [79], [109], [122]. Following recent calls for more rigorous [58] and human-based [138] evaluation of explanations, researchers should take a more systematic and human-centered approach [143] to the selection and evaluation of XAI methods for affective computing tasks.

Rohlfing et al. [144] build on the human-centered approach by suggesting that explanations are a social practice and should be co-constructed together with the system and the user. Developing co-constructive explanation systems, however, requires knowing which explanations are most effective for a given situation and user, and using this knowledge to adapt the explanation in an interactive way. While this is still a lofty goal, in the reviewed articles there was a lack of any type of interactive system. Although Wang et al. [55] proposed a visualization system with interactive explanations, it was limited to specific XAI types and was not adaptive to users' needs. Situational context plays a critical role in explanation understanding. Therefore, knowledge of the types of explanations that work best for specific contexts is necessary for developing such adaptive approaches. Hence, when designing and evaluating affective XAI methods researchers should consider the real-world contexts in which the explanation will be situated. This becomes increasingly important as affective ML becomes more embedded in high-stakes decision-making, such as AI-assisted mental health assessment [1], [2].

VI. Conclusion

We performed a scoping review of XAI approaches in audiovisual affective ML. We found that affective computing interest in XAI has increased in recent years, but its application showed limited variation in terms of the used techniques, input types, and evaluation methods. While we acknowledge the promising developments in this field, there are still significant challenges and gaps that need to be addressed in future research to incorporate explainability in real-world AC applications. To this end, we suggest future directions to guide new and existing researchers in this field.

Future Directions

Enrichment of XAI Methods: To address the limited variation in XAI methods within AC, we recommend exploring explainability methods beyond post hoc feature-based techniques. Specifically, example-, rule- and concept-based explanations may be a promising avenue. Finally, explanation scope should be considered carefully to avoid cherry-picking, especially when dealing with local approaches.

Application of XAI for multimodal settings: Multimodal explanations should not be limited to identifying dominant modalities but explore all three principles of multimodality: heterogeneity, modality interactions, and connections. Furthermore, given the importance of speech in audiovisual affective computing, it is essential that researchers identify new methods that can be applied to currently limited audio-based models. Implementing sonification for audio explanations could be a promising avenue as it would present explanations in the same modality as we use for perceiving audio, instead of translating audio features into visual representations.

Real World Applications with human-centric emphasis: In addition to helping researchers debug their models, XAI approaches should keep the end goal of an affective computing task in mind. Specifically, considerations of what type of explanations would be the most helpful for the end user (e.g. a clinician) are relevant. Explicit statements about the goal of using XAI might help guide decisions on the choice and evaluation of the methods. In all cases, a quantitative evaluation approach is commendable for assessing the effectiveness of the XAI methods in audiovisual affective ML tasks.

Acknowledgments

The DFG and BMBF had no involvement in the study design, the collection, analysis and interpretation of data, the writing of the report and the decision to submit the article for publication.

References

[1] L. He et al., "Deep learning for depression recognition with audiovisual cues: A review," *Inf. Fusion*, vol. 80, pp. 56–86, May 2022.

[2] K. Min, J. Yoon, M. Kang, D. Lee, E. Park, and J. Han, "Detecting depression on video logs using audiovisual features," *Humanities Social Sci. Commun.*, vol. 10, no. 1, pp. 1–8, Nov. 2023.

[3] G. Pei, H. Li, Y. Lu, Y. Wang, S. Hua, and T. Li, "Affective computing: Recent advances, challenges, and future trends," *Intell. Comput.*, vol. 3, Jan. 2024, Art. no. 0076.

[4] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Third Quarter 2022.

[5] L. V. L. Pinto et al., "A systematic review of facial expression detection methods," *IEEE Access*, vol. 11, pp. 61 881–61 891, 2023.

[6] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[7] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, Jan. 2021, Art. no. 1249.

[8] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: A review," *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 4335–4385, Aug. 2020.

[9] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. 1st Conf. Fairness Accountability Transparency*, PMLR, 2018, pp. 77–91.

[10] European Parliament and Council of the European Union, Regulation (EU) 2024/1689 of the European parliament and of the council, 2024. [Online]. Available: https://data.europa.eu/eli/reg/2024/1689/oj

[11] G. Vilone and L. Longo, "Explainable artificial intelligence: A systematic review," 2020. [Online]. Available: https://arxiv.org/abs/2006.00093

[12] A. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[13] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, "A systematic review of explainable artificial intelligence in terms of different application domains and tasks," *Appl. Sci.*, vol. 12, no. 3, Jan. 2022, Art. no. 1353.

[14] K. Cortiñas-Lorenzo and G. Lacey, "Toward explainable affective computing: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 13101–13121, Oct. 2024.

[15] D. S. Johnson, O. Hakobyan, and H. Drimalla, "Towards interpretability in audio and visual affective machine learning: A review," Jun. 2023, *arXiv:2306.08933*.

[16] M. Graziani et al., "A global taxonomy of interpretable AI: Unifying the terminology for the technical and social sciences," *Artif. Intell. Rev.*, vol. 56, pp. 3473–3504, Sep. 2022.

[17] A. de Santana Correia and E. L. Colombini, "Attention, please! a survey of neural attention models in deep learning," *Artif. Intell. Rev.*, vol. 55, no. 8, pp. 6037–6124, Dec. 2022.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should i trust you?": Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144.

[19] C. Molnar, *Interpretable Machine Learning*, 2nd ed. Morrisville, NC, USA: Lulu.com, 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[21] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 59–66.

[22] F. Eyben et al., "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Second Quarter 2016.

[23] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., Curran Associates, Inc., 2014, pp. 3320–3328. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/375c71349b295fbe2dcdca9206f20a06-Paper.pdf

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[25] R. Achtibat et al., "From attribution maps to human-understandable explanations through Concept Relevance Propagation," *Nat. Mach. Intell.*, vol. 5, no. 9, pp. 1006–1019, Sep. 2023.

[26] V. Chen, Q. V. Liao, J. Wortman Vaughan, and G. Bansal, "Understanding the role of human intuition on reliance in human-AI decision-making with explanations," *Proc. ACM Hum.-Comput. Interact.*, vol. 7, no. CSCW2, pp. 1–32, Oct. 2023, doi: 10.1145/3610219.

[27] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. 2020 ACM Conf. Fairness Accountability Transparency*, 2020, pp. 607–617, doi: 10.1145/3351095.3372850.

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

[29] Y.-H. H. Tsai, M. Q. Ma, M. Yang, R. Salakhutdinov, and L.-P. Morency, "Multimodal routing: Improving local and global interpretability of multimodal language analysis," 2020, *arXiv: 2004.14198*.

[30] M. Flora, C. Potvin, A. McGovern, and S. Handler, "Comparing explanation methods for traditional machine learning models part 1: An overview of current methods and quantifying their disagreement," Nov. 2022, *arXiv:2211.08943*.

[31] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021.

[32] M. Ivanovs, R. Kadikis, and K. Ozols, "Perturbation-based methods for explaining deep neural networks: A survey," *Pattern Recognit. Lett.*, vol. 150, pp. 228–234, Oct. 2021.

[33] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, PMLR, 2017, pp. 3319–3328.

[34] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," Aug. 2017, *arXiv: 1708.08296*.

[35] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–16. [Online]. Available: https://openreview.net/forum?id=Sy21R9JAW

[36] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Mach. Learn.*, Sydney, NSW, Australia: JMLR.org, 2017, pp. 3145–3153.

[37] B. Rozemberczki et al., "The shapley value in machine learning," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Vienna, Austria, 2022, pp. 5572–5579.

[38] C. J. Cai, J. Jongejan, and J. Holbrook, "The effects of example-based explanations in a machine learning interface," in *Proc. 24th ACM Int. Conf. Intell. User Interfaces*, 2019, pp. 258–262, doi: 10.1145/3301275.3302289.

[39] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal, "Efficient data representation by selecting prototypes with importance weights," in *Proc. 2019 IEEE Int. Conf. Data Mining*, 2019, pp. 260–269. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICDM.2019.00036

[40] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard J. Law Technol.*, vol. 31, 2017, Art. no. 841.

[41] S. Mertes, T. Huber, K. Weitz, A. Heimerl, and E. André, “Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning,” *Front. Artif. Intell.*, vol. 5, 2022, Art. no. 825565. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2022.825565

[42] G. Jeanneret, L. Simon, and F. Jurie, “Diffusion models for counterfactual explanations,” in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 858–876.

[43] E. Poeta, G. Ciravegna, E. Pastor, T. Cerquitelli, and E. Baralis, “Concept-based explainable artificial intelligence: A survey,” 2023, *arXiv:2312.12936*.

[44] B. Kim et al., “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV),” in *Proc. 35th Int. Conf. Mach. Learn.*, PMLR, 2018, pp. 2668–2677.

[45] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, “Towards automatic concept-based explanations,” in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, pp. 9277–9286. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/77d2afcb31f6493e350fca61764efb9a-Paper.pdf

[46] R. Zhang, P. Madumal, T. Miller, K. A. Ehinger, and B. I. P. Rubinstein, “Invertible concept-based explanations for CNN models with non-negative concept activation vectors,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 11 682–11 690. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17389

[47] L. Sixt, M. Schuessler, O.-I. Popescu, P. Weiß, and T. Landgraf, “Do users benefit from interpretable vision? A user study, baseline, and dataset,” in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–23. [Online]. Available: https://openreview.net/forum?id=v6s3HVjPerv

[48] G. Vilone and L. Longo, “A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods,” *Front. Artif. Intell.*, vol. 4, 2021, Art. no. 717899. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2021.717899

[49] M. T. Ribeiro, S. Singh, and C. Guestrin, “Anchors: High-precision model-agnostic explanations,” in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1527–1535. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11491

[50] S. Guillaume, “Designing fuzzy inference systems from data: An interpretability-oriented review,” *IEEE Trans. Fuzzy Syst.*, vol. 9, no. 3, pp. 426–443, Jun. 2001.

[51] P. P. Liang, A. Zadeh, and L.-P. Morency, “Foundations and trends in multimodal machine learning: Principles, challenges, and open questions,” Feb. 2023. [Online]. Available: http://arxiv.org/abs/2209.03430

[52] Y. Lyu, P. P. Liang, Z. Deng, R. Salakhutdinov, and L.-P. Morency, “Dime: Fine-grained interpretations of multimodal models via disentangled local explanations,” in *Proc. 2022 AAAI/ACM Conf. AI Ethics Soc.*, 2022, pp. 455–467, doi: 10.1145/3514094.3534148.

[53] E. Aflalo et al., “VL-interpret: An interactive visualization tool for interpreting vision-language transformers,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21 406–21 415.

[54] G. E. Hinton, S. Sabour, and N. Frosst, “Matrix capsules with EM routing,” in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–15. [Online]. Available: https://openreview.net/forum?id=HJWLfGWRb

[55] X. Wang, J. He, Z. Jin, M. Yang, Y. Wang, and H. Qu, “M2Lens: Visualizing and explaining multimodal models for sentiment analysis,” Aug. 2021, *arXiv:2107.08264*.

[56] P. P. Liang et al., “Multiviz: Towards visualizing and understanding multimodal models,” 2023, *arXiv:2207.00056*.

[57] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, “Evaluating the quality of machine learning explanations: A survey on methods and metrics,” *Electronics*, vol. 10, no. 5, 2021, Art. no. 593. [Online]. Available: https://www.mdpi.com/2079-9292/10/5/593

[58] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” 2017, *arXiv: 1702.08608*.

[59] A. Holzinger, A. Carrington, and H. Müller, “Measuring the quality of explanations: The system causability scale (scs),” *KI - Künstliche Intelligenz*, vol. 34, no. 2, pp. 193–198, Jun. 2020.

[60] A. C. Tricco et al., “PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation,” *Ann. Intern. Med.*, vol. 169, no. 7, pp. 467–473, Oct. 2018.

[61] M. J. Page et al., “The PRISMA 2020 statement: An updated guideline for reporting systematic reviews,” *BMJ*, vol. 372, Mar. 2021, Art. no. n71.

[62] T. A. Araf, A. Siddika, S. Karimi, and M. G. R. Alam, “Real-time face emotion recognition and visualization using grad-CAM,” in *Proc. 2nd Int. Conf. Adv. Elect. Comput. Commun. Sustain. Technol.*, Bhilai, India, 2022, pp. 1–5.

[63] D. Boulanger, M. A. A. Dewan, V. S. Kumar, and F. Lin, “Lightweight and interpretable detection of affective engagement for online learners,” in *Proc. 2021 IEEE Int. Conf. Dependable Autonomic Secure Comput. Int. Conf. Pervasive Intell. Comput. Int. Conf. Cloud Big Data Comput. Int. Conf. Cyber Sci. Technol. Congr.*, 2021, pp. 176–184.

[64] L. P. Carlini et al., “A convolutional neural network-based mobile application to bedside neonatal pain assessment,” in *Proc. 34th SIBGRAPI Conf. Graph. Patterns Images*, Gramado, Rio Grande do Sul, Brazil, 2021, pp. 394–401.

[65] M. Cesarelli, F. Martinelli, F. Mercaldo, and A. Santone, “Emotion recognition from facial expression using explainable deep learning,” in *Proc. 2022 IEEE Int. Conf. Dependable Autonomic Secure Comput. Int. Conf. Pervasive Intell. Comput., Int. Conf. Cloud Big Data Comput. Int. Conf. Cyber Sci. Technol. Congr.*, Falerna, Italy, 2022, pp. 1–6.

[66] G. Del Castillo Torres, M. F. Roig-Maimó, M. Mascaró-Oliver, E. Amengual-Alcover, and R. Mas-Sansó, “Understanding how CNNs recognize facial expressions: A case study with LIME and CEM,” *Sensors*, vol. 23, no. 1, Dec. 2022, Art. no. 131.

[67] M. Deramgozin, S. Jovanovic, H. Rabah, and N. Ramzan, “A hybrid explainable AI framework applied to global and local facial expression recognition,” in *Proc. 2021 IEEE Int. Conf. Imag. Syst. Techn.*, Kaohsiung, Taiwan, 2021, pp. 1–5.

[68] M. M. Deramgozin, S. Jovanovic, M. Arevalillo-Herraez, and H. Rabah, “An explainable and reliable facial expression recognition system for remote health monitoring,” in *Proc. 29th IEEE Int. Conf. Electron. Circuits Syst.*, Glasgow, U.K., 2022, pp. 1–4.

[69] B. Finzel, I. Rieger, S. Kuhn, and U. Schmid, “Domain-specific evaluation of visual explanations for application-grounded facial expression recognition,” in *Machine Learning and Knowledge Extraction*, A. Holzinger, P. Kieseberg, F. Cabitza, A. Campagner, A. M. Tjoa, and E. Weippl, Eds, vol. 14065. Berlin, Germany: Springer Nature, 2023, pp. 31–44.

[70] A. Heimerl, K. Weitz, T. Baur, and E. André, “Unraveling ML models of emotion with NOVA: Multi-level explainable AI for NON-EXPERTs,” *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1155–1167, Third Quarter 2022.

[71] L. Ji, S. Wu, and X. Gu, “A facial expression recognition algorithm incorporating SVM and explainable residual neural network,” *Signal, Image Video Process.*, vol. 17, no. 8, pp. 4245–4254, Nov. 2023.

[72] R. Kadakia, P. Kalkotwar, P. Jhaveri, R. Patanwadia, and K. Srivastava, “Analysis of micro expressions using XAI,” in *Proc. IEEE 3rd Int. Conf. Comput. Analytics Netw.*, Rajpura, Punjab, India, Nov. 2022, pp. 1–7.

[73] A. A. Kandeel, H. M. Abbas, and H. S. Hassanein, “Explainable model selection of a convolutional neural network for driver’s facial emotion identification,” in *Proc. Pattern Recognit. ICPR Int. Workshops Challenges*, A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, and R. Vezzani, Eds. Cham: Springer International Publishing, 2021, pp. 699–713.

[74] S. Malik, P. Kumar, and B. Raman, “Towards interpretable facial emotion recognition,” in *Proc. 12th ACM Indian Conf. Comput. Vis. Graph. Image Process.*, Jodhpur India, 2021, pp. 1–9.

[75] C. Manresa-Yee and S. Ramis, “Assessing gender bias in predictive algorithms using eXplainable AI,” in *Proc. 21st ACM Int. Conf. Hum. Comput. Interact.*, Málaga Spain, 2021, pp. 1–8.

[76] S. Park and C. Wallraven, “Comparing facial expression recognition in humans and machines: Using CAM, GradCAM, and extremal perturbation,” in *Pattern Recognition*, C. Wallraven, Q. Liu, and H. Nagahara, Eds, vol. 13188. Berlin, Germany: Springer International Publishing, 2022, pp. 403–416.

[77] M. Rathod et al., “Kids’ emotion recognition using various deep-learning models with explainable AI,” *Sensors*, vol. 22, no. 20, Oct. 2022, Art. no. 8066.

[78] D. Schiller, T. Huber, M. Dietz, and E. André, “Relevance-based data masking: A model-agnostic transfer learning approach for facial expression recognition,” *Front. Comput. Sci.*, vol. 2, Mar. 2020, Art. no. 6.

[79] K. Ter Burg and H. Kaya, “Comparing approaches for explaining DNN-based facial expression classifications,” *Algorithms*, vol. 15, no. 10, Oct. 2022, Art. no. 367.

[80] K. Weitz, T. Hassan, U. Schmid, and J.-U. Garbas, “Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods,” *Technisches Messen*, vol. 86, no. 7–8, pp. 404–412, Jul. 2019.

[81] X. Zhou, K. Jin, Y. Shang, and G. Guo, “Visually interpretable representation learning for depression recognition from facial images,” *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 542–552, Third Quarter 2020.

[82] H. Zhu, C. Yu, and A. Cangelosi, "Affective human-robot interaction with multimodal explanations," in *Social Robotics*, F. Cavallo ed., vol. 13817. Berlin, Germany: Springer Nature, 2022, pp. 241–252.

[83] S. Liu, S. Huang, W. Fu, and J. C.-W. Lin, "A descriptive human visual cognitive strategy using graph neural network for facial expression recognition," *Int. J. Mach. Learn. Cybern.*, vol. 15, pp. 19–35, Oct. 2024.

[84] Y. Jiao, Y. Niu, Y. Zhang, F. Li, C. Zou, and G. Shi, "Facial attention based convolutional neural network for 2D+3D facial expression recognition," in *Proc. 2019 IEEE Vis. Commun. Image Process.*, Sydney, Australia, 2019, pp. 1–4.

[85] M. Sun et al., "Attention-rectified and texture-enhanced cross-attention transformer feature fusion network for facial expression recognition," *IEEE Trans. Ind. Inform.*, vol. 19, no. 12, pp. 11 823–11 832, Dec. 2023.

[86] J. Zhang and H. Yu, "Improving the facial expression recognition and its interpretability via generating expression pattern-map," *Pattern Recognit.*, vol. 129, Sep. 2022, Art. no. 108737.

[87] C. Manresa-Yee, S. Ramis Guarinos, and J. M. Buades Rubio, "Facial expression recognition: Impact of gender on fairness and expressions$_*$," in *Proc. 22nd ACM Int. Conf. Hum. Comput. Interact.*, Teruel Spain, 2022, pp. 1–8.

[88] H.-E. Lee, K.-H. Park, and Z. Bien, "Iterative fuzzy clustering algorithm with supervision to construct probabilistic fuzzy rule base from numerical data," *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 1, pp. 263–277, Feb. 2008.

[89] C. Wang, X. Gao, and X. Li, "An interpretable deep Bayesian model for facial micro-expression recognition," in *Proc. IEEE 8th Int. Conf. Control Robot. Eng.*, Niigata, Japan, Apr. 2023, pp. 91–94.

[90] A. Ghandeharioun, B. Eoff, B. Jou, and R. Picard, "Characterizing Sources of Uncertainty to Proxy Calibration and Disambiguate Annotator and Data Bias," in *Proc. 2019 IEEE/CVF Int. Conf. Comput. Vis. Workshop*, Seoul, South Korea, 2019, pp. 4202–4206.

[91] M. Gund, A. R. Bharadwaj, and I. Nwogu, "Interpretable emotion classification using temporal convolutional models," in *Proc. 25th Int. Conf. Pattern Recognit.*, Milan, Italy, 2021, pp. 6367–6374.

[92] P. Prajod, T. Huber, and E. André, "Using explainable AI to identify differences between clinical and experimental pain detection models based on facial expressions," in *MultiMedia Modeling*, B. þór Jónsson ed., vol. 13141. Berlin, Germany: Springer International Publishing, 2022, pp. 311–322.

[93] S. Zhao et al., "ME-PLAN: A deep prototypical learning with local attention network for dynamic micro-expression recognition," *Neural Netw.*, vol. 153, pp. 427–443, Sep. 2022.

[94] N. Haines, M. W. Southward, J. S. Cheavens, T. Beauchaine, and W.-Y. Ahn, "Using computer-vision and machine learning to automate facial coding of positive and negative affect intensity," *PLoS One*, vol. 14, no. 2, Feb. 2019, Art. no. e0211735.

[95] J. Zhou, X. Zhang, Y. Liu, and X. Lan, "Facial expression recognition using spatial-temporal semantic graph network," in *Proc. 2020 IEEE Int. Conf. Image Process.*, Abu Dhabi, UAE, 2020, pp. 1961–1965.

[96] V. Pandit, M. Schmitt, N. Cummins, and B. Schuller, "I see it in your eyes: Training the shallowest-possible CNN to recognise emotions and pain from muted web-assisted in-the-wild video-chats in real-time," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102347.

[97] Z. Yang, Y. Zhang, and J. Luo, "Human-centered emotion recognition in animated GIFs," in *Proc. 2019 IEEE Int. Conf. Multimedia Expo*, Shanghai, China, 2019, pp. 1090–1095.

[98] X. Chen, L. Niu, A. Veeraraghavan, and A. Sabharwal, "FaceEngage: Robust estimation of gameplay engagement from user-contributed (YouTube) videos," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 651–665, Second Quarter 2022.

[99] V. Lin, J. M. Girard, M. A. Sayette, and L.-P. Morency, "Toward Multimodal modeling of emotional expressiveness," in *Proc. 2020 ACM Int. Conf. Multimodal Interact.*, 2020, pp. 548–557.

[100] D. Seuss et al., "Automatic estimation of action unit intensities and inference of emotional appraisals," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1188–1200, Second Quarter 2023.

[101] M. Cardaioli et al., "Face the truth: Interpretable emotion genuineness detection," in *Proc. 2022 IEEE Int. Joint Conf. Neural Netw.*, Padua, Italy, 2022, pp. 01–08.

[102] R. Zhi and M. Wan, "Dynamic facial expression feature learning based on sparse RNN," in *Proc. IEEE 8th Joint Int. Inf. Technol. Artif. Intell. Conf.*, Chongqing, China, 2019, pp. 1373–1377.

[103] E. Morales-Vargas, C. A. Reyes-García, and H. Peregrina-Barreto, "On the use of action units and fuzzy explanatory models for facial expression recognition," *PLoS One*, vol. 14, no. 10, Oct. 2019, Art. no. e0223563.

[104] M. De Velasco, R. Justo, A. López Zorrilla, and M. I. Torres, "Analysis of deep learning-based decision-making in an emotional spontaneous speech task," *Appl. Sci.*, vol. 13, no. 2, Jan. 2023, Art. no. 980.

[105] N. T. Pham, S. D. Nguyen, V. S. T. Nguyen, B. N. H. Pham, and D. N. M. Dang, "Speech emotion recognition using overlapping sliding window and Shapley additive explainable deep neural network," *J. Inf. Telecommun.*, vol. 7, no. 3, pp. 317–335, Jul. 2023.

[106] A. C. Shruti, R. H. Rifat, M. Kamal, and M. G. R. Alam, "A comparative study on bengali speech sentiment analysis based on audio data," in *Proc. 2023 IEEE Int. Conf. Big Data Smart Comput.*, Jeju, South Korea, Feb. 2023, pp. 219–226.

[107] Y. Ma and W. Wang, "MSFL: Explainable multitask-based shared feature learning for multilingual speech emotion recognition," *Appl. Sci.*, vol. 12, no. 24, Dec. 2022, Art. no. 12805.

[108] A. Anand, S. Negi, and N. Narendra, "Filters know how you feel: Explaining intermediate speech emotion classification representations," in *Proc. 2021 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2021, pp. 756–761.

[109] W. Zhang and B. Y. Lim, "Towards relatable explainable AI with the perceptual process," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–24.

[110] J. Guo, J. Tang, W. Dai, Y. Ding, and W. Kong, "Dynamically adjust word representations using unaligned multimodal information," in *Proc. 30th ACM Int. Conf. Multimedia*, Lisboa, Portugal, 2022, pp. 3394–3402.

[111] Y. Gu et al., "Human conversation analysis using attentive multimodal networks with hierarchical encoder-decoder," in *Proc. 26th ACM Int. Conf. Multimedia*, Seoul, South Korea, 2018, pp. 537–545.

[112] Y. Wu et al., "Modeling incongruity between modalities for multimodal sarcasm detection," *IEEE MultiMedia*, vol. 28, no. 2, pp. 86–95, Second Quarter 2021.

[113] T. Wörtwein, L. B. Sheeber, N. Allen, J. F. Cohn, and L.-P. Morency, "Human-guided modality informativeness for affective states," in *Proc. 2021 ACM Int. Conf. Multimodal Interact.*, Montréal, QC, Canada, 2021, pp. 728–734.

[114] J. Wu, S. Mai, and H. Hu, "Graph capsule aggregation for unaligned multimodal sequences," in *Proc. 2021 ACM Int. Conf. Multimodal Interact.*, Montréal, QC, Canada, 2021, pp. 521–529.

[115] J. Wu, S. Mai, and H. Hu, "Interpretable multimodal capsule fusion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1815–1826, 2022.

[116] D. K. Jain, A. Rahate, G. Joshi, R. Walambe, and K. Kotecha, "Employing co-learning to evaluate the explainability of multimodal sentiment analysis," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4673–4680, Aug. 2024.

[117] A. Khalane, R. Makwana, T. Shaikh, and A. Ullah, "Evaluating significant features in context-aware multimodal emotion recognition with XAI methods," *Expert Syst.*, Aug. 2023, Art. no. e13403. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.13403

[118] W. Rahman, S. Mahbub, A. Salekin, M. K. Hasan, and E. Hoque, "HirePreter: A framework for providing fine-grained interpretation for automated job interview analysis," in *Proc. IEEE 9th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos*, Nara, Japan, Sep. 2021, pp. 1–5.

[119] Y. Zhou, X. Yao, W. Han, Y. Wang, Z. Li, and Y. Li, "Distinguishing apathy and depression in older adults with mild cognitive impairment using text, audio, and video based on multiclass classification and shapely additive explanations," *Int. J. Geriatr. Psychiatry*, vol. 37, no. 11, Nov. 2022, Art. no. gps.5827.

[120] L. Hemamou, A. Guillon, J.-C. Martin, and C. Clavel, "Multimodal hierarchical attention neural network: Looking for candidates behaviour which impact recruiter's decision," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 969–985, Second Quarter 2023.

[121] A. R. Asokan, N. Kumar, A. V. Ragam, and S. S. Shylaja, "Interpretability for multimodal emotion recognition using concept activation vectors," in *Proc. 2022 IEEE Int. Joint Conf. Neural Netw.*, Padua, Italy, 2022, pp. 01–08.

[122] A. S. Wicaksana and C. C. S. Liem, "Human-explainable features for job candidate screening prediction," in *Proc. 2017 IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Honolulu, HI, USA, 2017, pp. 1664–1669.

[123] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Volume 1*, Melbourne, Australia, 2018, pp. 2225–2235.

[124] A. Das, J. Mock, F. Irani, Y. Huang, P. Najafirad, and E. Golob, "Multimodal explainable AI predicts upcoming speech behavior in adults who stutter," *Front. Neurosci.*, vol. 16, Aug. 2022, Art. no. 912798.

[125] S. Zhao et al., “Affective image content analysis: Two decades review and new perspectives,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10_Part_2, pp. 6729–6751, Oct. 2022, doi: 10.1109/TPAMI.2021.3094362.

[126] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA, Curran Associates Inc., 2017, pp. 4768–4777.

[127] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[128] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks,” in *Proc. 2018 IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 839–847, doi: 10.1109/WACV.2018.00097.

[129] H. J. Escalante et al., “Modeling, recognizing, and explaining apparent personality from videos,” *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 894–911, Second Quarter 2022.

[130] B. W. Schuller et al., “Towards sonification in multimodal and user-friendlyexplainable artificial intelligence,” in *Proc. 2021 ACM Int. Conf. Multimodal Interact.*, New York, NY, USA, 2021, pp. 788–792, doi: 10.1145/3462244.3479879.

[131] J. A. Hall, T. G. Horgan, and N. A. Murphy, “Nonverbal communication,” *Annu. Rev. Psychol.*, vol. 70, no. 1, pp. 271–294, 2019.

[132] K. Grabowski et al., “Emotional expression in psychiatric conditions: New technology for clinicians,” *Psychiatry Clin. Neurosci.*, vol. 73, no. 2, pp. 50–62, Feb. 2019.

[133] E. W. Carr, P. Winkielman, and C. Oveis, “Transforming the mirror: Power fundamentally changes facial responding to emotional expressions,” *J. Exp. Psychol., Gen.*, vol. 143, no. 3, pp. 997–1003, 2014.

[134] I. van der Linden, H. Haned, and E. Kanoulas, “Global aggregations of local explanations for black box models,” Jul. 2019, *arXiv: 1907.03039*.

[135] J. Pfau, A. T. Young, M. L. Wei, and M. J. Keiser, “Global saliency: Aggregating saliency maps to assess dataset artefact bias,” 2019, *arXiv: 1910.07604*.

[136] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, “Sanity checks for saliency maps,” *Adv Neural Inf. Process. Syst.*, vol. 31, 2018.

[137] H. Lakkaraju, S. H. Bach, and J. Leskovec, “Interpretable decision sets: A joint framework for description and prediction,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, New York, NY, USA, 2016, pp. 1675–1684, doi: 10.1145/2939672.2939874.

[138] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artif. Intell.*, vol. 267, pp. 1–38, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370218305988

[139] M. Chromik and M. Schuessler, “A taxonomy for human subject evaluation of black-box explanations in XAI,” in *ExSS-ATEC@IUI*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:214730454

[140] S. Mohseni, N. Zarei, and E. D. Ragan, “A multidisciplinary survey and framework for design and evaluation of explainable AI systems,” *ACM Trans. Interact. Intell. Syst.*, vol. 11, no. 3-4, pp. 1–45, Sep. 2021, doi: 10.1145/3387166.

[141] J. Han et al., “Deep learning for mobile mental health: Challenges and recent advances,” *IEEE Signal Process. Mag.*, vol. 38, no. 6, pp. 96–105, Nov. 2021.

[142] T. Miller, P. Howe, and L. Sonenberg, “Explainable AI: Beware of inmates running the asylum Or: How I learnt to stop worrying and love the social and behavioural sciences,” 2017, *arXiv: 1712.00547*.

[143] Y. Rong et al., “Towards human-centered explainable AI: A survey of user studies for model explanations,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2104–2122, Apr. 2023.

[144] K. J. Rohlfing et al., “Explanation as a social practice: Toward a conceptual framework for the social design of AI systems,” *IEEE Trans. Cogn. Devel. Syst.*, vol. 13, no. 3, pp. 717–728, Sep. 2021.

David S. Johnson received the PhD degree from the University of Victoria in 2019, where his research integrated machine learning, mixed reality, and human-computer interaction reality for sound and music computing. He currently leads the junior independent research group, Human-Centric Explainable AI (HCXAI) at the Center for Cognitive Interaction Technology (CITEC), Bielefeld University. His research focuses on human-centered computing, explainable AI, and affective computing, aiming to enhance AI-assisted decision-making systems that leverage multimodal behavioral data. After his PhD, he joined the Fraunhofer Institute for Digital Media Technology (IDMT), where he conducted research on industrial sound analysis and sound event detection. Prior to establishing the HCXAI group, he was a postdoctoral researcher with the Multimodal Behavior Processing group, Bielefeld University.

Olya Hakobyan received the PhD degree in computational neuroscience from the University of Bochum, Bochum, Germany, in 2022. She is a postdoctoral researcher with the Center for Cognitive Interaction Technology (CITEC), Bielefeld University. Her current research centers on digital social interactions, focusing on the ethical use of data-driven technologies. In the latter domain, she has contributed to a review of representational bias in audiovisual datasets and developed a participant-focused data donation tool.

Jonas Paletschek received the master’s degree in physics from Regensburg University, where he studied Temporal Graph Neural Networks. He is currently working toward the PhD degree with the Center for Cognitive Interaction Technology (CITEC), Bielefeld University. His current research is on the influence of context, such as stress and social interaction conditions, on signals of understanding. He aims to integrate signals of understanding to augment explainable AI and ensure inclusivity in machine learning models by developing diverse datasets.

Hanna Drimalla received the PhD degree from Humboldt-Universität zu Berlin, in 2019, focusing on mimicry and empathy analyzed with methods from psychophysiology and machine learning. She is a professor of Human-Centered Artificial Intelligence with the Center for Cognitive Interaction Technology (CITEC), Bielefeld University. Combining methods from computer science and psychology, her research focuses on automatic affect recognition, computer-based stress measurement, and the analysis of social interaction patterns. As a postdoc, she worked on digital mental health with the Hasso-Plattner-Institute, University of Potsdam. Until 2024, she led a junior research group on empathic artificial intelligence and the Multimodal Behavior Processing group, Bielefeld University