Empathy Detection from Text, Audiovisual, Audio or Physiological Signals: A Systematic Review of Task Formulations and Machine Learning Methods

Md Rakibul Hasan, *Graduate Student Member, IEEE,* Md Zakir Hossain, *Member, IEEE,* Shreya Ghosh, Aneesh Krishna, and Tom Gedeon, *Senior Member, IEEE*

Abstract—Empathy indicates an individual's ability to understand others. Over the past few years, empathy has drawn attention from various disciplines, including but not limited to Affective Computing, Cognitive Science, and Psychology. Detecting empathy has potential applications in society, healthcare and education. Despite being a broad and overlapping topic, the avenue of empathy detection leveraging Machine Learning remains underexplored from a systematic literature review perspective. We collected 829 papers from 10 well-known databases, systematically screened them and analysed the final 62 papers. Our analyses reveal several prominent task formulations – including empathy on localised utterances or overall expressions, unidirectional or parallel empathy, and emotional contagion – in monadic, dyadic and group interactions. Empathy detection methods are summarised based on four input modalities – text, audiovisual, audio and physiological signals – thereby presenting modality-specific network architecture design protocols. We discuss challenges, research gaps and potential applications in the Affective Computing-based *empathy* domain, which can facilitate new avenues of exploration. We further enlist the public availability of datasets and codes. This paper, therefore, provides a structured overview of recent advancements and remaining challenges towards developing a robust empathy detection system that could meaningfully contribute to enhancing human well-being.

Index Terms—Empathy, Empathy Computing, Deep Learning, Machine Learning, Pattern Recognition, Systematic Review

1 Introduction

EMPATHY, the ability to understand others' point of view, is essential for effective communication in many aspects of human life, including social dynamics [1], healthcare [2] and education [3]. Empathy towards other individuals is essential for the survival of our species and contributes significantly to enhancing the quality of life and the depth of social interactions [4], [5]. Research on empathy has been a major topic across various disciplines, including Social Science, Psychology, Neuroscience, Health and, most recently, Computer Science [6], [7]. Although contexts of empathy research vary across disciplines, all agree on its crucial role in human well-being [6]. In Computer Science, a significant body of literature deals with operationalising empathy using Machine Learning (ML) tools.

ML-based empathy computing remains underdeveloped compared to the more mature fields of emotion and facial expression recognition in Affective Computing. While numerous reviews and surveys address tasks like facial affect recognition [8] and emotion recognition [9], reviews on empathy recognition [7], [10]–[14] are limited and further tailored to specific use cases. For example, Paiva *et al.* [7] and Yalcin and DiPaola [10] reviewed computational empathy in the context of artificial agents in the years 2017 and 2018, respectively. Park and Whang [11] systematically reviewed

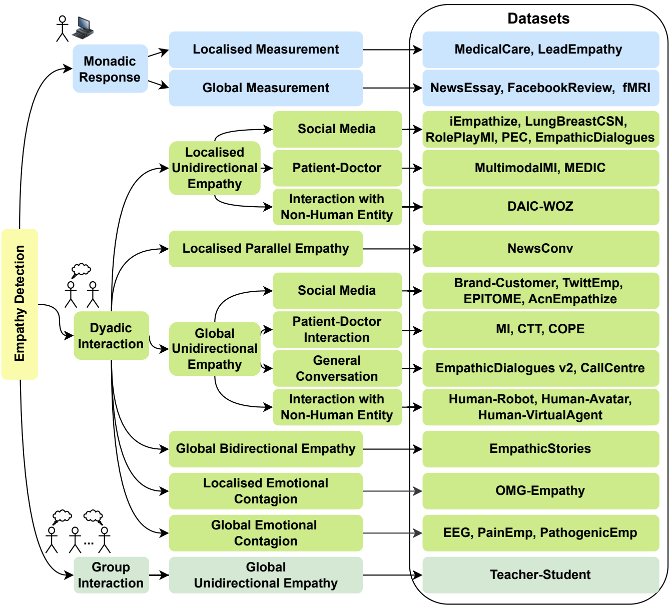

Empathy Detection
Monadic Response
Dyadic Interaction
Group Interaction
Localised Measurement
Global Measurement
Localised Unidirectional Empathy
Social Media
Patient-Doctor
Interaction with Non-Human Entity
Localised Parallel Empathy
Global Unidirectional Empathy
Social Media
Patient-Doctor Interaction
General Conversation
Interaction with Non-Human Entity
Global Bidirectional Empathy
Localised Emotional Contagion
Global Emotional Contagion
Global Unidirectional Empathy
Datasets
MedicalCare, LeadEmpathy
NewsEssay, FacebookReview, fMRI
iEmpathize, LungBreastCSN, RolePlayMI, PEC, EmpathicDialogues
MultimodalMI, MEDIC
DAIC-WOZ
NewsConv
Brand-Customer, TwittEmp, EPITOME, AcnEmpathize
MI, CTT, COPE
EmpathicDialogues v2, CallCentre
Human-Robot, Human-Avatar, Human-VirtualAgent
EmpathicStories
OMG-Empathy
EEG, PainEmp, PathogenicEmp
Teacher-Student

Fig. 1. Hierarchy of empathy detection task formulations and representative datasets.

- *All authors are with the School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Perth WA 6102, Australia.*
  E-mail: {Rakibul.Hasan, Zakir.Hossain1, Shreya.Ghosh, A.Krishna, Tom.Gedeon}@curtin.edu.au
- *M. R. Hasan is also with BRAC University, Bangladesh.*
- *M. Z. Hossain is also with The Australian National University, Australia.*
- *T. Gedeon is also with the University of Óbuda, Hungary.*

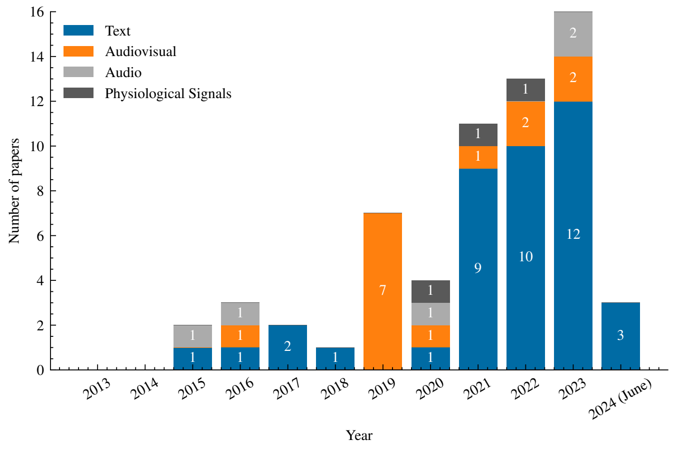


Fig. 2. Growth of ML-based empathy detection literature from 2013 to 2024 (June). There are 40 text-based, 14 audiovisual-based, 5 audio-based and 3 physiological signal-based studies.

empathy of various social robots in human-robot interaction contexts in 2022. Published in the same year, Raamkumar and Yang [12] reviewed empathic conversational systems that primarily aim to *generate* empathic responses. Lahnala *et al.* [13] and Shetty *et al.* [14] focused exclusively on Natural Language Processing (NLP)-based empathy computing. Therefore, there has been a lack of comprehensive review on empathy, particularly in the context of *detecting* empathy. A systematic literature review, in this case, facilitates a comprehensive evaluation of *all* published works screened through predefined criteria instead of cherry-picked studies. This paper presents a systematic review of ML-based empathy detection that covers any human interaction.

Our method adopts relevant aspects of the PRISMA 2020 guidelines [15], which are applicable to typical systematic reviews in Affective Computing. Refer to the Supplementary PRISMA 2020 Checklist for a mapping of how each relevant guideline has been considered in our study. We first devised search keywords and examined ten databases, including Scopus, Web of Science and IEEE Xplore. We screened the resulting 801 records against seven inclusion and exclusion criteria. Through rigorous title-and-abstract and full-text screenings, the final 62 papers are thoroughly reviewed in this paper. As shown in Figure 2, the distribution of papers reveals a predominant focus on text-based empathy detection (n = 40), followed by audiovisual (n = 14), audio (n = 5), and physiological signals (n = 3). Refer to the Supplementary Material for the details of our paper selection strategy.

We categorise datasets based on various task formulations found in the literature (Figure 1) and analyse ML methods across four input modalities: text, audiovisual, audio and physiological signals. The overarching objective of this study is to systematically review all ML-based empathy detection works published between 2013 and June 2024. Our major contributions include several key insights, such as:

a. Exploration of various task formulations in empathy computing
b. Detailed characteristics of available datasets, including their statistics and public availability
c. Overview of the studies, including chosen ML models and availability of the code
d. Identification of frequently used and high-performing ML methods applied to popular datasets
e. Potential application scenarios of empathy detection systems across diverse domains
f. Discussion of challenges and opportunities inherent in different task formulations and ML-based empathy detection approaches

The paper is organised as follows. Section 2 defines empathy and empathy detection based on several seminal works in Psychology. Section 3 introduces various task formulations with a comprehensive overview of representative datasets. Our dataset analysis includes their statistics, annotation protocol and public availability of the whole annotated dataset. Section 4 presents ML models specific to four input modalities – text sequences, audiovisual content, audio signals and physiological signals – where we discuss studies involving the datasets introduced in Section 3. In presenting empathy detection works, we report public availability of the software code, best-performing models and their performance. Both Section 3 and Section 4 end with a consolidating discussion on findings, challenges and opportunities. We present some applications of empathy detection in Section 5 and conclude the paper in Section 6.

2 Preliminaries

2.1 Empathy and Related Concepts

With broad usage across various disciplines, the definition of empathy varies. Cuff *et al.* [16] reviewed 43 discrete definitions of empathy and identified eight themes related to its nature. Themes include distinguishing empathy from similar concepts and determining whether it is cognitive or affective. The term 'empathy', as defined by Hoffman [17], is predominantly involuntary and vicarious reaction to emotional signals from another individual or their circumstances. In another work, Hoffman [4] defines empathy as 'an affective response more appropriate to another's situation than one's own'. Empathy is a multifaceted concept that involves perceiving, understanding and sharing the emotional thoughts of others [18]. It can also be defined as a multidimensional concept, such as four-dimensional empathy (perspective taking, fantasy, empathic concern and personal distress) [19], and two-dimensional empathy (empathic concern and personal distress) [20].

Numerous endeavours have been made to disentangle empathy from other similar concepts [16]. Some scholars (e.g., [20], [21]) conceptualise empathy as a comprehensive category including emotional contagion, sympathy and compassion. Empathy is defined as comprehending another's emotions through adopting their perspective; other related psychological states include compathy (feelings shared due to shared circumstances), mimpathy (copying another's emotions without personally experiencing them), sympathy (intentionally responding emotionally), transpathy (emotional contagion, where one becomes 'infected' by another's emotions) and unipathy (an elevated form of emotional contagion) [22], [23]. Despite the inherent ambiguity in defining empathy, scholars such as Ickes [22] and Blair [24] advocated separating these terms. The two most related terms are empathy and sympathy, which can be described as ‘feeling *as*’ versus ‘feeling *for*’, respectively [25]. Neuroscientific evidence supports the distinction between empathy and sympathy as they have distinct neural processes [26].

While distinctions in empathy are well-recognised in Psychology, empathy computing literature often overlooks these nuances during dataset collection and labelling. This is likely because the primary focus has been on detecting the *presence* of empathy in any form, typically categorised into levels such as ‘not empathic’, ‘somewhat empathic’, and ‘very empathic’ [27] or as binary classifications like ‘empathic’ and ‘not empathic’. Binary labelling, in particular, is the most prevalent approach in the literature [28]–[31]. Given that empathy computing literature does not always adopt a formal definition of empathy, we chose not to exclude any studies based on their operational definitions of the concept.

Within the domain of empathy, perhaps the two most common forms are cognitive empathy and emotional (also known as affective) empathy [32]. Understanding someone’s thoughts and perspective is known as cognitive empathy, whereas vicarious sharing of emotion is known as emotional empathy [32]. Cognitive empathy is closely related to the theory of mind, that is, understanding another person’s mental state, such as wants, beliefs and intentions [24]. In other words, cognitive empathy is ‘I *understand* what you feel’, whereas emotional empathy is ‘I *feel* what you feel’ [33]. Although most empathy computing studies treat empathy as a single, consolidated construct, Dey and Girju [34] explicitly distinguish between cognitive and affective empathy in a patient-doctor interaction setup.

Empathy detection differs from emotion detection, although both involve analysing human responses. Emotion detection focuses on recognising an *individual*’s emotional state, such as happiness or sadness [35]. In contrast, empathy detection goes into a deeper analysis of the interactions between *multiple* individuals. It considers the initial emotion expressed by one person and the emotional response of the other whose empathy is being measured. For example, if someone expresses sadness, empathy detection would analyse how the listener emotionally supports the speaker in response [36].

2.2 Empathy Measurement

In Psychology, questionnaires are widely employed to measure self-reported empathy levels [19]. These instruments typically present participants with statements or scenarios, prompting them to indicate their level of agreement or emotional response. Examples of widely used empathy questionnaires include the Interpersonal Reactivity Index (IRI) [37], the Empathy Quotient [38], Batson’s Empathy Scale [20] and the Toronto Empathy Questionnaire [39].

Empathic accuracy is another method of operationalising empathy in Psychology, which assesses how accurately an individual can infer another person’s thoughts and emotions. Experimentally, it is often determined by comparing one person’s reported thoughts and emotions with their partner’s in a dyadic interaction [40].

In Affective Computing, computational methods are developed to objectively measure empathy levels from verbal and non-verbal cues, such as facial expressions, tone of voice and body language. To achieve this, self-reported empathy levels through psychological questionnaires often provide necessary ground truths for training ML algorithms. Studies that use such self-reported annotations typically align with a specific definition of empathy as defined by the chosen questionnaire. For instance, Batson *et al.* [20]’s definition has been employed for measuring empathy in written essays [30], [41], [42], while the IRI questionnaire [19], [37] has been used in studies such as [43]. However, not all studies explicitly adhere to a specific definition of empathy. For example, datasets like `RolePlayMI` [44] and `PEC` adopt heuristic labelling approaches based on the origin of the data rather than examining each sample.

3 Task Formulations in Detecting Empathy

Let X be the input content on which empathy y will be measured. The content can be multimodal, i.e., $X \in \{X^s, X^a, X^v\}$, where X^s, X^a, X^v refer to text, audio, and video sequences, respectively. The empathy label y can be formulated as different levels of empathy, as in a classification problem, or a degree of empathy, as in a regression problem. The content X can consist of N segments X_i, where $i \in [1, N]$. Accordingly, measuring empathy on some segments of the content $X_i = [x_i, x_{i+1}, \cdots, x_{i+m}]$, where $0 \leq m < N$, can be interpreted as *localised* measurement, whereas measuring empathy on the whole content X can be interpreted as *global* measurement.

A range of experimental setups have been proposed in the literature to define and structure the specific goals for detecting empathy. Firstly, in *Monadic Response*, the focus is on measuring empathy on self-contained, individual responses of a person. Secondly, in *Dyadic Interaction*, empathy is measured from the interactions between two individuals. Expanding beyond dyads, in *Group Interaction*, empathy is measured on multiple individuals engaging in an interaction. Task formulations in each of these setups and corresponding datasets are summarised in the following subsections.

3.1 Monadic Response

Several studies explored monadic response-based empathy computing in either the localised or global manner (Table 1).

3.1.1 Localised Measurement

In detecting empathy at a localised level of monadic responses, Shi *et al.* [28] proposed `MedicalCare` dataset, which consists of 774 narrative essays of simulated patient-doctor interactions written by pre-med students. Sentences of the essays were labelled as either ‘empathic’ or ‘non-empathic’ by six trained undergraduate students, followed by two meta-annotators. Samples were considered ‘empathic’ if they displayed cognitive or affective empathy. As an extension of this task formulation, Dey and Girju [34] selected 440 essays from the pool of 774 `MedicalCare` essays and re-annotated them into four labels: cognitive empathy, affective empathy, prosocial behaviour and non-empathy, hereinafter referred to as the `MedicalCare v2` dataset. Dey and Girju [34]’s task formulation, therefore, aims to measure different types of empathy rather than

TABLE 1
Task Formulations and Corresponding Datasets for Empathy Detection from **Monadic Responses**.

SL	Name	Data	Statistics	Output label[a]	Anno.[c]	Public
Localised Measurement						
1	MedicalCare [28]	Essays on simulated patient-doctor interaction	774 essays	$\{0, 1\}$	T	×
2	MedicalCare v2 [34]	Re-annotation of a subset of MedicalCare dataset	440 essays	{Cognitive, Affective, Prosocial, None}	T	×
3	MedicalCare v3 [29]	Re-annotation of MedicalCare v2 dataset	440 essays	$\{0, 1\}$	T	×
4	LeadEmpathy [45]	Leaders' email to their subordinates	770 emails, 385 participants	$\{-4, -3, \ldots, 6, 7\}$	T	✓
Global Measurement						
5	NewsEssay [30]	Written essays in response to news articles	403 participants, 1,860 essays, 418 articles	$[1.0, 7.0]$, $\{0, 1\}$	S	✓
6	NewsEssay v2 [41]	Extension of the NewsEssay dataset	564 participants, 2,655 essays, 418 articles	$[1.0, 7.0]$	S	✓
7	NewsEssay v3 [42]	New essays based on a subset of articles of NewsEssay dataset	140 participants, 1,100 essays, 100 news articles	$[1.0, 7.0]$	S	✓
8	FacebookReview [31]	Comments from Facebook pages	900 reviews, 48 hospitals	$\{0, 1\}$	T	×
9	fMRI [43]	Resting-state fMRI data from cocaine-dependent subjects	24 subjects	$\mathbb{R}$	S	✓

[a] Output labels in $[x, y]$ refer to continuous values between x and y
[a] Output label $\{0, 1\}$ refers to binary labelling to represent {No Empathy, Empathy}
[a] $\mathbb{R}$ – real number, unspecified in the paper
[c] Annotation: S – Self; T – Third party

different levels of generalised empathy like Shi *et al.* [28]. As a further extension of the MedicalCare v2 dataset, Dey and Girju [29] formulated an empathy versus non-empathy classification problem by collapsing cognitive, affective and prosocial labels into a single 'empathic' class, hereinafter referred to as the MedicalCare v3 dataset. In this updated dataset, the authors also identified four themes that a healthcare provider might express during the interactions: (1) empathic language, (2) medical procedural information, (3) both empathy and information and (4) neither empathy nor information. Such a thematic approach can help healthcare providers communicate effectively, given that they need to empathise as well as deliver procedural information.

LeadEmpathy is another dataset modelling localised measurement, which consists of emails from participants acting as leaders in a business organisation [45]. In the first phase of data collection, each participant wrote an email to their subordinate employees regarding a hypothetical concern that resulted in losing a customer. In the second phase of the experiment, the participants were asked to rewrite their earlier emails to increase empathy. Both emails were considered for empathy detection, and exploratory data analysis supports increased empathy in the second email. The annotation protocol considers empathy success and failure in both cognitive and affective empathy. Segments of emails are annotated into discreet empathy scores ranging from -4 to $+7$, which facilitates modelling it either as a classification task or a regression task. In addition, scores of 1 and below can be mapped to 'low' empathy and scores of 2 and higher to 'high' empathy in a binary classification setup [45].

3.1.2 Global Measurement

In predicting empathy on the whole content level, the NewsEssay dataset consists of essays written by Amazon Mechanical Turk participants about news articles involving harm to individuals, groups or nature. In addition to the essays, the dataset consists of participants' demographic information, such as age, gender, income and education level. This dataset has gone through a series of enhancements, serving empathy detection challenges in a conference workshop named Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA)[1]. The WASSA 2021 and 2022 challenges use the same NewsEssay v2 dataset [41], whereas WASSA 2023 challenge release the updated NewsEssay v3 dataset [42], both of which extend from the inaugural NewsEssay dataset [30] by involving new participants in the data collection experiment. Empathy labels in these datasets came from the essay writers themselves as they filled in Batson's empathy and distress scale [20]. This scale includes questions related to six empathy-related emotions (sympathetic, compassionate, tender, etc.) and eight personal distress-related emotions (alarmed, upset, worried, etc.). The responses were collected on a 7-point Likert scale, where a value of one and seven means the participant is not feeling the emotion at all and extremely feeling the emotion, respectively. After averaging the scores across questions, this dataset's ground truth degree of empathy ranges from 1 to 7 for each written essay.

[1] https://workshop-wassa.github.io/

People often leave reviews on products or services through online forums, including for hospitals. A. Rahim *et al.* [31] collected people's reviews on the official Facebook pages of 48 public hospitals. Two hospital quality managers labelled the reviews of this FacebookReview dataset into 'yes' or 'no' empathy. This task formulation aimed to analyse the service quality of the hospitals in addition to other characteristics such as assurance, responsiveness and reliability.

Apart from individuals' expressed contents like written essays, empathy can be measured from physiological signals since these indicate individuals' internal emotional states [46]. The fMRI dataset [43] consists of resting-state functional Magnetic Resonance Imaging (fMRI) data from 24 cocaine-addicted subjects. The subjects filled in the well-known IRI questionnaire [19], [37], which provides continuous empathy scores for empathy assessment.

3.2 Dyadic Interaction

Similar to monadic response-based empathy computing, dyadic interactions are explored both in localised and global measurements. Table 2 categorises datasets derived from such interactions into subcategories based on their task formulations.

3.2.1 Localised Unidirectional Empathy

Localised measurement of unidirectional empathy (i.e., empathy flowing in one direction) has been studied across various contexts, such as social media, healthcare and general conversations.

3.2.1.1 Social Media: In empathising by one person towards another, several studies leverage dyadic interaction from online forums. For example, the iEmpathize dataset [36] consists of discussion threads from a website named Cancer Survivors Network[2]. Collected from the same website, the LungBreastCSN dataset [47] focuses on lung and breast cancer data. While the iEmpathize dataset aims to detect empathy *direction* (seeking, providing or none), the LungBreastCSN dataset aims to detect the presence of empathy (empathic and non-empathic sentences).

The RolePlayMI dataset [44] consists of counselling conversations from video-sharing platforms, such as YouTube and Vimeo, which were originally collected in a separate study on counselling quality analysis [48]. Wu *et al.* [44] later annotated this dataset into utterance-level empathic and non-empathic categories.

Being a versatile platform, Reddit is often used for analysing data across different topics of interest. The PEC dataset [44], [49] consists of general conversations from three subreddits (forums dedicated to specific topics in Reddit), which are labelled heuristically. Utterances from the 'Happy' and 'OffMyChest' subreddits and the EmpathicDialogues corpus (dyadic conversations regarding any personal situation) [50] are considered empathic labels, whereas samples from the 'CasualConversation' subreddit are considered non-empathic labels.

3.2.1.2 Patient-Doctor Interaction: The MultimodalMI dataset [51], [52] consists of two real-world motivational interviewing sessions: (Dataset 1) students assigned to MI sessions due to alcohol-related matters, and (Dataset 2) volunteering heavy drinkers aged 17–20 years. The therapists' empathy (through a sequence of texts) is annotated primarily through a Likert scale, which is also converted to binary labels (low vs high empathy). This dataset, therefore, allows empathy detection as either a classification problem [52] or as a regression problem [51]. Annotation protocols utilise Motivational Interviewing Skill Code (MISC) 2.5 guidelines [72] for Dataset 1 and Motivational Interviewing Treatment Integrity (MITI) 3.1 code [73] for Dataset 2. The dataset consists of speech transcripts and audio, enabling it to model a multimodal empathy detection problem.

Another multimodal dataset is the MEDIC dataset [53], which consists of video, audio and text sequences of counselling case videos. It evaluates counsellors' empathy through three mechanisms: expression of experience, emotional reaction and cognitive reaction. The 'expression of experience' mechanism aims to measure a client's expression to trigger empathy from a counsellor. In contrast, the 'emotional reaction' and 'cognitive reaction' mechanisms aim to measure the empathy of counsellors. Five trained students annotated the speech turns into none, weak and strong expressions for each mechanism.

3.2.1.3 Interaction with Non-Human Entity: The DAIC-WOZ dataset comprises semi-structured interviews between human participants and a virtual agent, which aims to measure the empathy of the virtual agent towards the human participants. The conversations are segmented into small time windows and annotated by third-party annotators into three classes: negative empathy, positive empathy and no empathy [54], [55]. Responses such as 'That sounds really hard' are considered 'negative empathy', whereas no responses, expressed fillers or responses without sentiment are considered 'no empathy'.

3.2.2 Localised Parallel Empathy

In one study covering localised empathy measurements, both persons provide empathy to someone else (parallel empathy), such as two persons conversing and empathising with some disadvantaged people. This NewsConv dataset [42] consists of dyadic conversations regarding newspaper articles featuring harm to individual, entity or nature. Speech turns of the conversations are annotated by independent annotators on a scale of 0 to 5. It is worthwhile to note that the same news articles were used in the NewsEssay v3 dataset, which aims to measure the empathy of individual study participants towards others through written essays. In contrast, the NewsConv dataset aims to measure the empathy of two persons in the conversations.

3.2.3 Global Unidirectional Empathy

Similar to the localised measurement of unidirectional empathy, global measurement has been explored in various contexts, including social media and healthcare.

3.2.3.1 Social Media: Several studies assess empathy globally from social media data – primarily Twitter, Reddit and Facebook – which are annotated by trained

[2] https://csn.cancer.org/

TABLE 2
Task Formulations and Corresponding Datasets for Empathy Detection from **Dyadic Interactions**.

SL	Name	Data	Statistics	Output label[a]	Anno.[c]	Public
Localised Unidirectional Empathy						
1	iEmpathize [36]	Discussions from online cancer survivors network	5,007 sentences	{Seeking, Providing, None}	T	✓
2	LungBreastCSN [47]	Discussions from online cancer survivor's network (lung and breast)	2,107 messages	{0, 1}	T	×
3	RolePlayMI[c] [44], [48]	Counselling conversations from online video sharing platforms	253 conversations	{0, 1}	T	✓
4	PEC[c] [44], [49]	General conversations from Reddit	355K conversations	{0, 1}	T	✓
5	EmpathicDialogues [44], [50]	Dyadic conversations regarding any personal situation	810 participants and 24,850 conversations	{0, 1}	T	✓
6	MultimodalMI [51], [52]	Real-world motivational interviewing psychotherapy sessions	301 patients, 16 therapists and 301 sessions (each 50–60 minutes)	[0, 1], {High, Low}	T	×
7	MEDIC [53]	Psychological counselling	771 video clips (total 11 hours)	{None, Weak, Strong} Expression	T	✓
8	DAIC-WOZ [54], [55]	Semi-structured interviews with virtual agent	186 participants and 2,185 conversations	{Negative, Positive, None}	T	×
Localised Parallel Empathy						
9	NewsConv [42]	Conversation about news articles	140 participants and 12,601 speech-turns	$[0.0, 5.0]$	T	✓
Global Unidirectional Empathy						
10	Brand-Customer [56]	Customer queries and brand response from Twitter	108 brands, 667,738 customers, and 2,013,577 tweets	{None, Weak, Strong}	T	×
11	TwittEmp [57]	Cancer and 200 high-rating empathy words-related tweets	3,000 tweets	{Seeking, Providing, None}	T	✓
12	EPITOME [58]	Responses towards help-seeking posts in TalkLife and Reddit	8 million posts and 26 million interactions	{None, Weak, Strong}	T	✓
13	EPITOME v2 [59]	EPITOME, relabelled into two classes	8 million posts and 17 million interactions	{Positive, Negative}	T	✓
14	AcnEmpathize [60]	Posts, quotes and replies from an online acne support forum	12,212 samples	{0, 1}	T	✓
15	MI [61]	Motivational interviews between therapists and patients of drug or alcohol use	176 therapists and 348 sessions	{High, Low}, $[1.0, 7.0]$	T	×
16	MI v2 [62]	Same as the MI dataset	348 sessions	{High, Low}	–	×
17	CTT [63], [64]	Motivational interviewing sessions of drug and alcohol counselling	200 sessions	$[1, 7]$, {High, Low}	T	×
18	COPE [65], [66]	Conversations between cancer patient and healthcare provider(s)	425 sessions	{0, 1}	T	×
19	EmpathicDialogues v2 [27]	Samples from a dialogue generation dataset [50], re-annotated into five labels	400 conversations	{Not, A Little, Somewhat, Empathic, Very Much}	T	✓
20	CallCentre [67]	Human-human conversation in call centre	905 conversations	{0, 1}	T	×
21	Human-Robot [68]	Human participants listen to six scripted stories from a robot	46 participants and 6.9 hours audiovisual data	{Empathy, Less Empathy}	S	✓
22	Human-Avatar [69]	Interaction between avatar and normotypical (empathic), Down syndrome and intellectual disability people (non-empathic)	50 participants and 24,000 interactions	{0, 1}	O	×
23	Human-VirtualAgent [70]	Human participants watched a sad virtual character in virtual reality	28 participants and 56 surveys	$[0, 20]$	S	×
Global Bidirectional Empathy						
24	EmpathicStories [71]	Personal stories from social media sites, crowdsourcing and spoken narratives	2,000 similar story pairs	{1, 2, 3, 4}	T	✓

[a] Output labels in $[x, y]$ refer to continuous values between x and y
[a] Output label {0, 1} refers to binary labelling to represent {No Empathy, Empathy}
[c] Annotation: S – Self; T – Third party; O – Other

annotators. For example, the Brand-Customer dataset [56] consists of Twitter threads about customer service-related queries and corresponding brand responses. The authors [56] aimed to estimate engagement between brands and customers into three categories of empathy: none, weak and strong empathy (of brand agents).

The TwittEmp dataset [57] consists of cancer-related tweets labelled into three categories: seeking, providing and no empathy. In a binary classification setting, the 'seeking' and 'providing' samples are considered positive, and the no empathy samples are considered negative.

Apart from these, various online forums facilitate consultations and mental health support. Sharma *et al.* [58] proposed a widely-recognised empathy detection framework, named EPITOME, which consists of three communication mechanisms: emotional reactions, interpretations and explorations. Mental health-related help-seeking posts were collected from Reddit and TalkLife (a dedicated mental health support network) and annotated into three categories – none, weak and strong – for each of the three mechanisms. EPITOME was relabelled by Hosseini and Caragea [59] into two classes: weak and strong communication as the positive samples, and no communication as the negative samples, hereinafter referred to as the EPITOME v2 dataset.

The AcnEmpathize dataset consists of discussions from an online acne-related forum[3]. Adopting the annotation principle of EPITOME [58], three annotators labelled each of the discussion components (posts, replies and quotes) as either 'empathic' or 'not empathic'. A discussion component is labelled as empathic if any part exhibits any of the three communication mechanisms (emotional reactions, interpretations and explorations) of the EPITOME framework.

3.2.3.2 Patient-Doctor Interaction: Global measurement has been formulated in several datasets of counselling sessions between therapists and patients. For example, the motivational interviewing dataset, named MI [61], comprises interview sessions from clinical interviews with patients of drug or alcohol use from six clinical studies. Another similar dataset (MI v2) [62] also evaluates session-level empathy in motivational interviewing. The CTT dataset [63], [64] includes 200 sessions between therapists and patients of drug and alcohol abuse. The annotation includes a continuous degree of empathy between 1 and 7 to model a regression problem and a low or high empathy level to model a classification problem.

The COPE dataset [65], [66] consists of 425 oncology encounters between cancer patients and healthcare providers. The task of this dataset is to detect empathic interactions and filter out non-empathic ones. Two trained annotators labelled this dataset into binary labels, where empathic interaction refers to when a patient expressed negative emotions and the oncologists responded empathically.

3.2.3.3 General Conversation: Apart from specialised peer support communities and patient-doctor interactions, some studies aim to measure empathy in general conversations. For example, the EmpathicDialogues dataset – consisting of conversations regarding any personal situation and earlier used in localised measurement [44], [50] – was relabelled into five levels of empathy (EmpathicDialogues v2) and used for global measurement [27].

The quality of support provided by call centre staff can be measured by measuring their empathy. Alam *et al.* [67] proposed CallCentre dataset consisting of human-to-human call-centre conversation, where conversations are labelled either empathic if the session contains at least one empathic segment or non-empathic otherwise.

3.2.3.4 Interaction with Non-Human Entity: Several global empathy detection datasets include interactions between humans and non-human entities, such as avatars and robots. The Human-Robot data collection includes a robot telling scripted stories to human participants [68]. Stories were told in either first-person or third-person point-of-view. To measure the participants' empathy towards the robot or story content, the participants answered a custom questionnaire of eight questions on a 5-point Likert scale. Thresholding based on median statistics is used to binarise the empathy scores into two labels ('empathic' and 'less empathic').

The Human-Avatar dataset aims to assess the empathy of human participants interacting with an avatar expressing six types of emotion [69]. Rather than self-reported annotation or third-party annotation, each interaction is labelled as empathic for normotypical participants and non-empathic for participants having social communication disorders such as Down syndrome and intellectual disability. Such a labelling approach was formulated to diagnose social communication disorders through empathy assessment.

In Human-VirtualAgent dataset [70], human participants watched a virtual character expressing sadness in a virtual reality environment. Participants fill in two questionnaires, including the Toronto empathy questionnaire [39], to reflect how much empathy they feel towards the agents. The questionnaire responses are leveraged as self-assessed ground truth empathy scores on a scale of 0 to 20.

3.2.4 Global Bidirectional Empathy

Bidirectional empathy can be defined as two individuals empathising with each other. Shen *et al.* [71] introduced EmpathicStories, which features bidirectional empathy in personal stories. Pairs of personal stories are labelled in terms of how two persons empathise with each other's experiences. The authors operationalise *empathic similarity* in terms of three key aspects: main event, emotion and moral of the story [71].

3.2.5 Localised Emotional Contagion

Emotional contagion – the process by which one person's emotions and behaviours trigger similar emotions and behaviours in others – is an element of empathy [20], [21]. Some studies exclusively aim to measure emotional contagion, and, as such, they are discussed in this separate category of task formulation.

One such dataset, named OMG-Empathy [74], consists of audiovisual conversations with semi-scripted stories in a speaker-listener setup. Following the conversations, the listeners answered how the story impacted their emotional state in terms of a valence score between -1 to $+1$. Using

[3] https://www.acne.org/

this dataset, an empathy detection challenge[4] was organised, and accordingly, several empathy detection models are proposed for this dataset. This dataset offers two detection protocols: personalised protocol, which detects the valence score of each listener across all conversations, and generalised protocol, which detects the valence score towards each story by all listeners.

3.2.6 *Global Emotional Contagion*

Three datasets aim to measure emotional contagion at the global level. These datasets are collected through passive dyadic interaction, where the subjects often look at some stimuli, for example, still images, video or text sequences (Table 3).

Some studies use subjects' physiological signals during a passive interaction. For example, the `EEG` dataset [75] contains Electroencephalogram (EEG) signals from 52 participants watching an emotional video (a young girl being abused as a domestic enslaved person) in virtual reality. The EEG signals were collected from the frontal, central and occipital regions of the brain before, during and after watching the video. Before the experiment, the participants filled in the Toronto empathy questionnaire [39]. Although the questionnaire allows empathy annotations to range from 0 to 96, participants' responses fell within a narrower range of 49 to 86, indicating a moderate to high level of empathy reported by the participants. Using a median split, samples are also grouped into high and low classes. Both regression and classification tasks in this dataset offer empathy detection at all three times when the EEG was collected: before, during and after.

The `PainEmp` dataset [76] comprises Electrocardiogram (ECG) and skin conductance data from 36 participants with different levels of autistic traits. After viewing pictures of individuals with different pain levels, the participants filled in a questionnaire regarding cognitive and affective empathy. Although it may sound a little frightening, the painful pictures (24 in total) were, in fact, collected from eight individuals going through different levels of electrical stimulation on the back of their hands. This dataset's task is to classify cognitive and affective empathy into high or low levels.

Finally, the `PathogenicEmp` dataset aims to measure emotional contagion from Facebook posts. Abdul-Mageed *et al.* [77] define 'pathogenic empathy' as the automatic contagion of negative emotions from others, which may lead to stress and burnout. The authors argued that this negative side of empathy is risky for the health and well-being of empathic people. In their data collection, participants answered a questionnaire, which was made of eight questions on a Likert scale, with 'not at all like me' on one end and 'very much like me' on the other end of the scale. The average of the responses is considered the ground truth pathogenic empathy score.

3.3 Group Interaction

There is only one dataset where empathy is measured from more than two persons as a group (Table 3). This dataset, hereinafter referred to as the `Teacher-Student` dataset, consists of 63 online audiovisual lectures in a one-to-many teaching setup [78]. Expert annotators label each lecture session on a scale of 0 to 10 (regression task), which are then thresholded to binarise into 'Excellent' and 'Good' categories (classification task). The broader aim of Pan *et al.* [78]'s work is to evaluate teaching quality through five characteristics of a good lecture: empathy, clarity, interaction, technical management and time management.

3.4 Discussion: Findings, Challenges and Research Gaps

The varieties in task formulations and trends across all datasets result in several key findings and opportunities, which are discussed in the following subsections.

3.4.1 *Prospective Task Formulations*

Although unidirectional empathy is well studied in localised and global measurements (Figure 1), there is a notable gap in exploring parallel empathy. We found only one study on parallel empathy in localised measurement. Parallel empathy is particularly relevant in understanding collective emotional dynamics, such as in team collaborations or group therapy sessions. Investigating these scenarios could reveal insights into how empathy propagates in multi-person interactions.

Similarly, studies are scarce in group interactions, with only one study addressing the global measurement of unidirectional empathy. Group settings capture naturalistic social environments, and therefore, empathy computing in group scenarios could advance effective collaboration and social cohesion. Overall, investigating these new task formulations could significantly enrich our understanding of the evolution and change in empathy during complex social interactions.

3.4.2 *Empathy from Observer's Physiological Signals*

Physiological signals contain essential affective cues in detecting internal states of people, which are often difficult to detect in other ways, such as classifying posed smiles [79] and pretended anger [80] from their real counterparts. Hossain *et al.* [79]'s observer-based smile veracity detection shows that it is possible to objectively measure subjective reactions. Considering the subjective nature of empathy, accurately assessing someone's *actual* (ground truth) empathy level can be challenging, making physiological signals potentially valuable.

Out of all the task formulations we review in this paper, three studies measured empathy from subjects' physiological signals, including ECG, EEG, fMRI and skin conductance. Firstly, other types of physiological signals that showed effectiveness in Affective Computing, such as pupillary response [80] and blood volume pulse [79] may be experimented with for empathy detection. Secondly, instead of physiological signals from one person, we could leverage signals from all parties involved, for example, both people in a dyadic interaction. Thirdly, detecting empathy from an observer's physiological signal could be an interesting avenue of exploration. This way, physiological signals can be collected from an observer observing the interaction on which

[4] https://www2.informatik.uni-hamburg.de/wtm/omgchallenges/omg_empathy_description_19.html

TABLE 3
Task Formulations and Corresponding Datasets for **Emotional Contagion from Dyadic Interaction** and Empathy from **Group Interactions**.

SL	Name	Data	Statistics	Output label[a]	Anno.[c]	Public
Emotional Contagion						
1	OMG-Empathy [74]	Speaker-listener conversations based on eight semi-scripted stories	4 speakers, 10 listeners and 80 audiovisual data (total 480 minutes)	$[-1, +1]$	S	✓
2	EEG [75]	Participants' EEG while watching an emotional video in virtual reality	52 participants and 52 EEG samples	$[0, 96]$, {High, Low}	S	×
3	PainEmp [76]	Participants viewing pictures of individuals with pain or no pain	36 participants, 36 ECG and skin conductance data	{High, Low}	S	×
4	PathogenicEmp [77]	Facebook posts and responses to a questionnaire	2,405 participants and 1,835,884 posts	$\mathbb{R}$	S	×
Group Interaction						
5	Teacher-Student [78]	Class lecture from one teacher to 5-10 students	10 teachers, 63 lectures, 338 audiovisual data	$[0.0, 10.0]$, {Excellent, Good}	T	×

[a] Output labels in $[x, y]$ refer to continuous values between x and y
[a] $\mathbb{R}$ – real number, unspecified in the paper
[c] Annotation: S – Self; T – Third party

we want to measure empathy. Then, we could investigate if the signals correlate with participants' empathy. Studying observer-based physiological signals could uncover how empathy is perceived and processed by third parties, a perspective that has implications for training empathy in professionals such as counsellors, educators and healthcare workers.

3.4.3 Annotation Protocol

The performance of supervised ML models is inherently tied to the quality of the labelled data. Empathy detection datasets are annotated primarily in two ways: self-annotation and third-party annotation. Self-rated annotation is a popular way to get the data labelled for Affective Computing tasks [81], In empathy computing, it refers to study participants filling in empathy-related questionnaires such as the IRI [19], [37] and the Toronto [39] questionnaires. In contrast, third-party annotation refers to annotations by third-party trained annotators instead of the study participants from whom the data is collected.

The choice between self-annotation and third-party annotators remains a debated topic in the literature. Of all the datasets we examine in this paper, 10 used self-annotation, and 25 used third-party annotation. Buechel *et al.* [30] argued that self-annotation provides a more appropriate measure of empathy than third-party annotators. Shi *et al.* [28] used `MedicalCare` dataset, annotated by trained third-party annotators, and `NewsEssay` dataset, annotated by study participants themselves. One interesting conclusion of their study is that third-party annotation could be more robust than self-rated annotation [28]. Using ensemble methods to combine the results of multiple third-party annotators is likely to be the most robust, which should be investigated thoroughly.

The `NewsEssay v3` and `NewsConv` datasets use the same participants but differ in annotation protocols, with self-assessment for essays and third-party annotation for conversations, respectively. As shown in Table 4, researchers have achieved higher empathy detection performance in the `NewsConv` dataset than in the `NewsEssay v3` dataset, potentially suggesting that third-party annotation provides greater consistency.

Judgement varies across individuals; for example, a certain empathic interaction can be felt as 'high' by someone, whereas the same can be felt as 'medium' by someone else. In this case, employing multiple third-party annotators to label many of the samples separately and subsequently testing their inter-rater reliability to reach a consensus for confounding samples should be preferred. However, a study has found that third-party annotators' conscious labelling of *subjective* reactions is worse than their non-conscious judgement [79]. Therefore, it can be argued that a third-party annotator may be unable to accurately assess the perceived empathy of the subject because empathy is subjective. To come to a conclusion, both self-annotation and third-party annotation while fixing the other aspects (such as dataset and model) would be a prospective research domain to understand more about annotation and simultaneously find an appropriate annotation scheme.

3.4.4 Public Availability of Datasets

Data availability facilitates reproducibility, comparative studies and benchmarking research. Among the 37 empathy detection datasets reviewed, 18 are publicly available. Challenges associated with making data public include privacy and ethical considerations. Ensuring the anonymisation of sensitive information and obtaining proper consent from participants are crucial steps, supposedly for which patient data such as `MI`, `CTT`, and `COPE` are unavailable. Additionally, there may be legal and institutional restrictions that prevent sharing of certain datasets. Addressing these challenges is essential during the early stage of planning to ensure data availability. We urge the authors of the 19 non-public datasets to take active steps towards making them public.

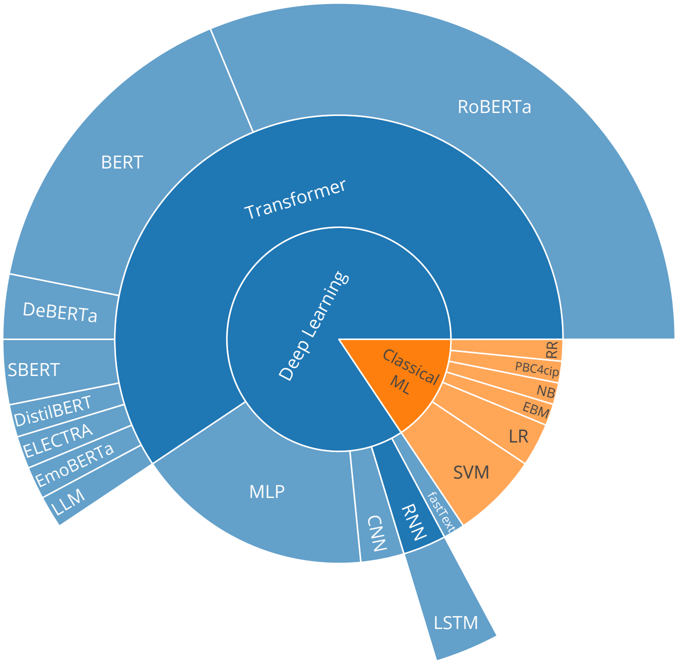


Fig. 3. Usage of ML algorithms in text-based empathy detection studies, demonstrating a substantial use of transformer-based architectures.

4 Modality-Specific Empathy Detection Methods

Design protocols for empathy detection methods, including the choice of preprocessing techniques and specific ML models, are primarily influenced by the input data modality. This section outlines the methods based on four input modalities – text sequences, audiovisual contents, audio signals and physiological signals – observed across different task formulations.

4.1 Text Sequence

In NLP research, empathy is detected from various textual content, such as essays, conversations and social media discussions. Such text-based datasets are predominantly employed with transformer-based Deep Learning (DL) algorithms and, to a lesser extent, with classical ML algorithms. Figure 3 illustrates the usage of algorithms in text-based empathy detection studies. With the recent successes of fine-tuning pre-trained language models in a variety of NLP tasks [82], it comes as no surprise that pre-trained language models dominate the landscape of text-based empathy detection studies. Among different variants of pre-trained language models, the Bidirectional Encoder Representations from Transformers (BERT)-based Robustly Optimised BERT Pretraining Approach (RoBERTa) is mostly used, followed by the BERT base model itself.

Both a continuous degree of empathy (regression) and a distinct level of empathy (classification) detection tasks are described in the literature, which are discussed in the following subsections.

4.1.1 Regression Task (Degree of Empathy)

TABLE 4
Summary of Empathy Detection Studies Modelled as a *Regression* Task (Degree of Empathy) from Text Sequences.

Dataset	Study	Best Model	Performance[b]	Code Avail.[a]
NewsEssay	[30]	fastText-CNN	PCC: 0.404	✓
	[83]	**LLM-RoBERTa-MLP**	**PCC: 0.924**	✓
NewsEssay v2	[84]	BERT-MLP	PCC: 0.473	✓
	[85]	LR	PCC: 0.516	✓
	[86]	RoBERTa-MLP	PCC: 0.517	✓
	[87]	**ELECTRA + RoBERTa**	**PCC: 0.558**	✓
	[88]	RoBERTa-MLP	PCC: 0.470	U
	[89]	BERT-MLP	PCC: 0.479	×
	[90]	RoBERTa	PCC: 0.504	U
	[91]	RoBERTa	PCC: 0.524	✓
	[92]	RoBERTa	PCC: 0.537	×
	[93]	RoBERTa	PCC: 0.541	×
	[83]	LLM-RoBERTa-MLP	PCC: 0.505	✓
NewsEssay v3	[94]	BERT	PCC: 0.187	✓
	[95]	RoBERTa-MLP	PCC: 0.270	×
	[96]	RoBERTa-MLP	PCC: 0.329	×
	[97]	RoBERTa	PCC: 0.331	×
	[98]	RoBERTa	PCC: 0.348	✓
	[99]	RoBERTa-SVM	PCC: 0.358	×
	[100]	{RoBERTa, EmoBERTa}-MLP	PCC: 0.415	×
	[42]	RoBERTa	PCC: 0.536	×
	[83]	**LLM-RoBERTa-MLP**	**PCC: 0.563**	✓
NewsConv	[94]	BERT	PCC: 0.573	✓
	[98]	RoBERTa	PCC: 0.652	✓
	[42]	RoBERTa	PCC: 0.660	×
	[95]	RoBERTa-MLP	PCC: 0.665	×
	[100]	{RoBERTa, EmoBERTa}-MLP	PCC: 0.669	×
	[97]	DeBERTa	PCC: 0.674	×
	[96]	**DeBERTa-MLP**	**PCC: 0.708**	×
MI	[61]	LR	SCC: 0.611	×
PathogenicEmp	[77]	RR	PCC: 0.252	×
LeadEmpathy	[45]	BERT	PCC: 0.816	✓
EmpathicStories	[71]	SBERT	PCC: 0.309, SCC: 0.352	✓

Note: Studies using common datasets are sorted year-wise chronologically, followed by performance, where best results and methods are **bolded**.
[a] U – Unofficially available on the Internet but not provided with the paper
[b] PCC – Pearson correlation coefficient
[b] SCC – Spearman's correlation coefficient

Table 4 summarises all regression tasks using textual data, highlighting the best-performing models across commonly used datasets. Most works used NewsEssay and its variants to detect empathy as a continuous value. The average performance in detecting empathy in essays from the v3 dataset is relatively lower than that observed in the v1 and v2 datasets, which may be attributed to the smaller size of the v3 dataset.

Interestingly, only 2 out of 22 studies on NewsEssay datasets do not leverage any pre-trained language models. For example, Buechel *et al.* [30] leveraged fastText [101] for text embeddings, followed by a Convolutional Neural Network (CNN) regression model, achieving a Pearson correlation coefficient of 0.404 in the NewsEssay dataset. Vettigli and Sorgente [85] employed Linear Regression (LR) classical ML method on the v2 dataset and reported a Pearson correlation coefficient of 0.516. This performance is competitive with studies utilising transformer-based language models such as BERT and RoBERTa, where the Pearson correlation coefficient ranges from 0.470 to 0.558 [87], [88]. This exceptional performance using classical ML can be attributed to incorporating handcrafted features, such as lexicon-based, n-gram and demographic-based features

[85]. Handcrafted features combined with additional raw data could be experimented with transformer architectures, as this might yield even better performance.

Instead of traditional ML and DL models, Hasan *et al.* [83] introduces a novel system called Large Language Model (LLM)-Guided Empathy (LLM-GEm) that leverages Generative Pre-trained Transformer (GPT)-3.5 LLM for three distinct purposes: converting numerical demographic numbers into semantically meaningful text, augmenting text sequences and rectifying label noises. Experiments on `NewsEssay v1`, `v2` and `v3` datasets demonstrate that LLM-GEm achieves state-of-the-art performance on the `v1` and `v3` datasets using a RoBERTa-based pre-trained language model as the prediction model. On the `v2` dataset where LLM-GEm underperformed, Mundra *et al.* [87] reported the best result (Pearson correlation coefficient: 0.558) using an ensemble of ELECTRA and RoBERTa models. Such a higher performance can be attributed to the ensemble of two language models (ELECTRA and RoBERTa). Overall, RoBERTa appears to be the best method in detecting empathy within the `NewsEssay` datasets.

Performance on the `NewsConv` dataset is higher than `NewsEssay` datasets, with 0.708 as the highest Pearson correlation coefficient using a Decoding-Enhanced BERT with Disentangled Attention (DeBERTa) model [96]. Plausible reasons could be the annotation protocols (as discussed earlier, `NewsConv` uses third-party annotation, which is likely to be more consistent and reduce the noise in the labels) and the size of the datasets (12,601 samples in the `NewsConv` dataset compared to 1,100–2,655 samples in the `NewsEssay` datasets).

Other continuous degrees of empathy detection works include therapists' empathy detection on `MI` dataset [61] and pathogenic empathy detection on social media [77]. Both of them leveraged classical ML methods: LR and Ridge Regression (RR). Classical MLs require fewer computational resources but often underperform transformer-based DL algorithms, and as such, future research may explore recent algorithms, such as transformers, with these datasets.

4.1.2 Classification Task (Level of Empathy)

In the case of modelling empathy as a classification task, a diverse array of datasets and algorithms is used, as summarised in Table 5. Only two datasets – `EPITOME` and `iEmpathize` – are used in multiple studies to allow comparative analysis.

Sharma *et al.* [58] used both unsupervised learning (domain adaptive pre-training) and RoBERTa-based supervised learning on their `EPITOME` framework. In contrast, Lee *et al.* [102] used Micromodels [104] as an attempt towards explainable ML. Lee *et al.* [102]'s approach first calculates semantic similarity scores between the Sentence BERT (SBERT) representations of some fixed seed utterances and dataset samples. The similarity scores are then used as a feature in an Explainable Boosting Machine (EBM) model to classify empathy in each of the mechanisms of `EPITOME`. In terms of quantitative scores, Lee *et al.* [102]'s model provides better accuracy (a maximum of 95.3% vs 92.6%) but less F1 score (a maximum of 62.7% vs 74.5%) than Sharma *et al.* [58]'s models on the `EPITOME` dataset. However, one important insight from Lee *et al.* [102]'s study is that the current empathy detection models probably consider surface-level information rather than the whole conversation context.

In detecting empathy on the `iEmpathize` dataset, Hosseini and Caragea [36], [103] leveraged BERT and RoBERTa models, respectively. Despite being the same dataset, their reported classification performances are on different evaluation metrics: a maximum F1 score of 85.9% is reported in [36], and a classification accuracy of 81.1% is reported in [103]. The key contribution of Hosseini and Caragea [103]'s work is a data-agnostic technique for prompt-based few-shot learning to improve the performance of pre-trained language models on empathy and emotion classification tasks, especially when training data is limited and noisy.

As seen in Table 5, the remaining studies used separate datasets, which may not allow performance comparison between studies. However, a distinction can be made between classical ML and DL-based language model usage. For example, on detecting empathy in medical essays, Shi *et al.* [28] experimented with Support Vector Machine (SVM) and Naïve Bayes (NB) on `MedicalCare` dataset, yielding an F1 score of 78.4%. In `MedicalCare v2` dataset, Dey and Girju [34] experimented with BERT, RoBERTa, SVM, NB, Logistic Regression (LogR), Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) models and reported an F1 score of 85%. Similar experiments are also conducted with the `MedicalCare v3` dataset [29]. Both experiments reveal the superior performance of BERT-based models. Incorporation of FrameNet pre-trained model [105] boosted the baseline performance in the `v2` dataset [34]. In the `v3` dataset [29], various linguistic constructions – such as active or passive voice, static or energetic tone – enhanced binary empathy classification performance compared to the baseline BERT model.

Lee and Parde [60] experimented with two classical ML algorithms (NB and LogR) and four pre-trained language model (including BERT, RoBERTa and Distilled BERT (DistilBERT)) on the `AcnEmpathize` dataset. Among these, DistilBERT resulted in the best overall accuracy of 89.3%, while the accuracy of BERT and RoBERTa was also close: 89.1% and 88.5%, respectively. Due to the unbalanced nature of the dataset, the authors also reported class-wise precision, recall and F1 scores. Although NB provided better precision in the empathy class and better recall in the no-empathy class, it underperforms the BERT-based models in the precision and F1 scores. The best F1 scores of 77.4% in the empathy class and 93.1% in the no empathy class are achieved by RoBERTa and DistilBERT, respectively.

On the `LeadEmpathy` dataset, Sedefoglu *et al.* [45] employed SVM for binary classification and BERT for a 10-class classification task. The 10-class F1 score was notably lower at 45.7%, reflecting the challenge of fine-grained empathy classification compared to the more straightforward binary classification, which achieved a higher F1 score of 81.7%.

Instead of language models, several studies leveraged traditional DL models like LSTM and CNN. Gibson *et al.* [61] reported NB as the optimal model in `MI` dataset. In a later study, Gibson *et al.* [62] reported that a combination of Multi Layer Perceptron (MLP) and LSTM are the optimal model in the closely related `MI v2` dataset, yielding a higher unweighted average recall from 75.3% to 79.6%. Khanpour *et al.* [47] used a combination of CNN and LSTM

TABLE 5
Summary of Empathy Detection Studies Modelled as a *Classification* Task (Level of Empathy) from Text Data.

Dataset	Study	Best Model	Performance[a]	Code Avail.
EPITOME	[58]	RoBERTa	Acc ∈ [79.4%, 92.6%], **F1** ∈ [**62.6%, 74.5%**]	✓
	[102]	SBERT, EBM	**Acc** ∈ [**88.3%, 95.3%**], F1 ∈ [59.5%, 62.7%]	✓
iEmpathize	[36]	BERT	F1 ∈ [78.9%, 85.8%]	×
	[103]	RoBERTa	Acc: 81.1%	×
MedicalCare + NewsEssay	[28]	SVM	Acc: 89.4%, F1: 78.4%	×
MedicalCare v2	[34]	BERT	F1 ∈ [75%, 85%]	×
MedicalCare v3	[29]	BERT	F1 ∈ [66%, 75%]	×
AcnEmpathize	[60]	DistilBERT, RoBERTa	Acc 89.3%, F1 ∈ [77.4%, 93.1%]	
LeadEmpathy	[45]	SVM (Binary), BERT (Multi)	F1: 81.7% (Binary), 49.9% (Multi)	×
MI	[61]	NB	Unweighted average recall: 75.3%	×
MI v2	[62]	MLP-LSTM	Unweighted average recall: 79.6%	×
LungBreastCSN	[47]	CNN-LSTM	F1: 78.4%	×
PEC, EmpathicDialogues, RolePlayMI	[44]	BERT	Matthews correlation coefficient ∈ [≈ 0.56, ≈ 0.95]	×
EmpathicDialogues v2	[27]	PBC4cip	AUC: 62.5%	×
NewsEssay	[57]	BERT-MLP	F1: 68.4%	×
EPITOME v2, NewsEssay	[59]	BERT, RoBERTa	Acc ∈ [61.5%, 71.8%]	×
TwittEmp	[57]	BERT-MLP	F1 ∈ [68.6%, 85.7%]	×
Brand-Customer	[56]	RoBERTa	F1: 73%	×
FacebookReview	[31]	SVM	Acc: 21.5%, F1: 75.7%	×

Note: Studies using common datasets are sorted year-wise chronologically, followed by performance, where best results are **bolded**.
[a] Range of performance is reported when overall classification performance is unavailable

on the LungBreastCSN dataset. Other than commonly known models, Montiel-Vázquez *et al.* [27] reported Pattern-Based Classifier for Class Imbalance Problems (PBC4cip) – exclusively designed for imbalanced datasets – as the most effective classifier compared to several classical ML baselines on EmpathicDialogues v2 dataset.

Hosseini and Caragea [57], [59] used knowledge distillation, which refers to the process of transferring knowledge from a large, complex model (teacher) to a smaller, simpler model (student) to improve the latter's performance while maintaining efficiency. Hosseini and Caragea [59] used EPITOME v2 as an in-domain dataset and NewsEssay as an out-of-domain dataset to transfer knowledge from a RoBERTa teacher model to a RoBERTa student model. Their knowledge distillation framework boosted the performance compared to BERT and RoBERTa baselines. In their study, the NewsEssay dataset was used in a binary classification setting instead of the dataset's default usage as a regression task. Such binary classification setup is also utilised by Shi *et al.* [28] and Hosseini and Caragea [57] using SVM and BERT-MLP models, respectively.

4.2 Audiovisual Content

Empathy detection from audiovisual contents is designed mostly as a multimodal system combining computer vision and NLP techniques, with inputs such as facial expressions, hand gestures and audio conversations. This section, therefore, includes some multimodal approaches, which utilise audio, video and sometimes text sequences. Figure 4 illustrates the application of algorithms in audiovisual-based empathy detection works. As usual, DL models are the predominant choice, although classical ML models are also widely employed. Within the DL category, CNN and Recurrent Neural Network (RNN)-based models are most frequently used, whereas in the classical ML category, SVM enjoys a higher level of usage.

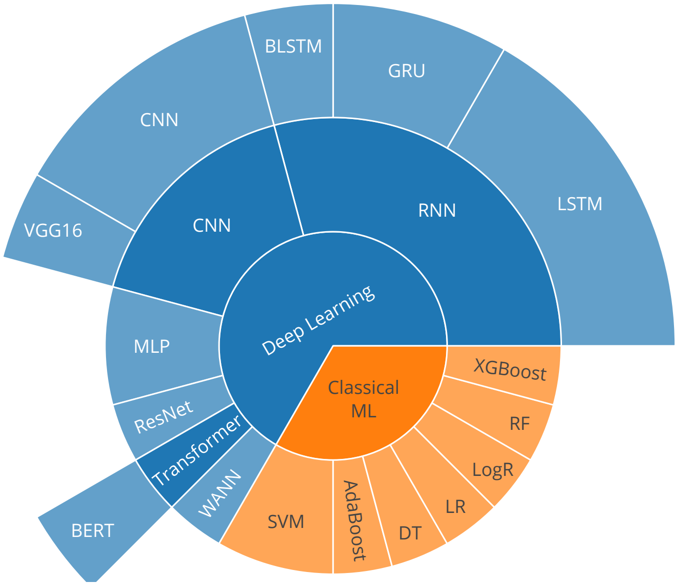


Fig. 4. Usage of ML algorithms in audiovisual content-based empathy detection studies.

Table 6 summarises the studies involving empathy detection from audiovisual datasets. There are nine studies detecting continuous degrees of empathy (regression task) and six studies detecting discrete levels of empathy (classification task). The following subsections describe these studies.

4.2.1 *Regression Task (Degree of Empathy)*

The baseline model in the OMG-Empathy challenge [74] used VGG16 architecture to process facial expression and LSTM to process spatial-temporal features. The outputs of these two networks were then concatenated and fed to a SVM for empathy detection, which resulted in 0.17

TABLE 6
Summary of Empathy Detection Studies from Audiovisual Content.

Dataset	Study	Best Model	Performance[a]	Code Avail.[b]
Regression				
OMG-Empathy	[106]	CNN, RF	CCC: 0.02 (P), 0.04 (G)	×
	[107]	SVM	CCC: 0.08 (P)	U
	[108]	BiLSTM	CCC: 0.11 (P), 0.06 (G)	✓
	[109]	LSTM	CCC: 0.14 (P, G)	✓
	[110]	GRU, LSTM, CNN, MLP	CCC: 0.17 (P, G)	✓
	[74]	**VGG16-LSTM-SVM**	**CCC 0.17 (P), 0.23 (G)**	×
	[111]	WANN	CCC: 0.25 (Validation set)	×
Human-VirtualAgent	[70]	LR	R^2: 0.485	×
Teacher-Student	[78]	AdaBoost	MSE: 0.374	✓
Classification				
Teacher-Student	[78]	DT	Acc: 90.9%, F1: 90.1%	✓
Human-Robot	[68]	XGBoost	Acc: 69%, AUC: 72%	×
Human-Avatar	[69]	LogR	F1 $\in$ [72%, 78%]	×
DAIC-WOZ	[54]	ResNet, BERT, GRU, MLP	F1: 71%	×
MEDIC	[53]	LSTM	Acc $\in$ [77.6%, 86.4%], F1 $\in$ [77.7%, 86.3%]	×
Various online sources[c]	[112]	CNN	Acc: 98.9%, AUC: 99%, F1: 91%	×

Note: Studies using common datasets are sorted year-wise chronologically, followed by performance, where best results and methods are **bolded**.

[a] Performance refers to test-set performance unless otherwise stated

[a] P – Personalised protocol; G – Generalised protocol

[a] CCC – Concordance Correlation Coefficient

[b] U – Unofficially available on the Internet but not provided with the paper

[c] Description of the dataset, such as the number of samples and ground truth label space, are unavailable on the paper

and 0.23 correlation coefficients in the personalised and generalised empathy protocols, respectively. Challenge participants [106]–[110] did not surpass these baseline results. Among them, Barbieri *et al.* [110] performed the best, achieving a 0.17 correlation coefficient for both protocols. They used separate models for each modality – Gated Recurrent Unit (GRU) on audio signals, LSTM on audio transcripts (text) and CNN on vision (face and body images) – followed by MLPs. To integrate the predictions across different modalities, they used a weighted average proportional to the validation score on each modality, followed by a Butterworth low-pass filter.

Tan *et al.* [109] extracted multimodal features using VGG-Face [113] on faces, openSMILE [114] on audio and GloVe embedding [115] on texts. Using a multimodal LSTM model, they reported a correlation coefficient of 0.14 on both personalised and generalised protocols. Mallol-Ragolta *et al.* [108] reported correlation coefficients of 0.11 and 0.06 in personalised and generalised protocols, respectively, using openSMILE for extracting audio features and OpenFace [116] for extracting video features, followed by a BiLSTM network.

In addition to verbal and non-verbal features from audio, image and text, Azari *et al.* [107] experimented with a different type of feature: mutual or contagious laughter as a measure of synchrony between the speaker and listener during the interaction. Hinduja *et al.* [106] used facial landmarks and spectrogram as hand-crafted features and CNN output as deep features in a Random Forest (RF) model. Lastly, Lusquino Filho *et al.* [111] leveraged a different type of model – Weightless Artificial Neural Network (WANN) – and reported a correlation coefficient of 0.25 on the validation set of the OMG-Empathy dataset.

Other works in empathy detection as regression tasks primarily utilised classical ML models. Kroes *et al.* [70] leveraged a LR model on the Human-VirtualAgent dataset. With the Teacher-Student dataset, Pan *et al.* [78] comprehensively experimented with a wide range of features from audio and video in an AdaBoost model. Their extracted features include mid-level behavioural features – such as facial expression, head pose and eye gaze – and high-level interpretable features, such as video length, frequency of speaker switch and total number of words. Such feature extraction often yields good results but may require substantial computational resources and careful tuning to optimise the model.

4.2.2 Classification Task (Level of Empathy)

In classifying empathy levels, most studies leveraged a variety of classical ML algorithms without using any common dataset across the studies. Mathur *et al.* [68] experimented with eight classical ML and two DL models and reported XGBoost as the best model on the Human-Robot dataset.

On the DAIC-WOZ dataset, Tavabi *et al.* [54] leveraged pre-trained BERT to calculate text embedding and pre-trained Residual Network (ResNet) to calculate visual features in addition to action units and head pose features from OpenFace. As audio features, they extracted extended Geneva minimalistic acoustic parameter set and Mel-Frequency Cepstral Coefficients (MFCC) [117] using OpenSMILE. With these features, they experimented with GRU and MLP in different fusion techniques, where GRU-based fusion of temporal audio and video sequences appeared to be the best fusion strategy in their setting, resulting an F1 score of 71%. Their ablation experiment shows that text modality is more effective than video and audio modalities: text alone resulted in an F1 score of 64%, whereas the video and audio individually provided F1 scores of 46% and 38%, respectively.

On the MEDIC dataset [53], the best result is achieved using SWAFN, a multimodal LSTM network proposed by

TABLE 7
Summary of Empathy Detection Studies from Audio Signals.

Dataset	Study	Best Model	Performance	Code Avail.
Regression				
CTT	[63]	LR	PCC $\in [0.65, 0.71]$	×
MultimodalMI	[51]	RoBERTa-GRU	CCC $\in [0.408, 0.596]$	✓
Classification				
CTT	[63]	SVM	Acc $\in$ [80.5%, 89.9%], F1 $\in$ [85.3%, 90.3%]	×
COPE	[65]	SVM	Avg. precision: 7.61%	×
CallCentre	[67]	SVM	Unweighted avg. recall: 65.1%	×
MultimodalMI	[52]	RoBERTa, HuBERT, GRU, MLP	F1 $\in$ [58.3%, 72.6%]	✓

Chen and Li [118]. The network uses three individual LSTMs to encode video, audio and textual modalities, followed by a novel aggregation strategy using a multi-task learning framework. Among the three mechanisms of the `MEDIC` dataset, the client's expression of experience was better classified than the counsellor's empathy [53], which supports the difficult nature of empathy detection compared to expression (i.e., emotion) recognition.

4.3 Audio Signals

Audio-based empathy detection works include audio from conversations in various contexts, such as healthcare and call centres. By audio-based empathy detection studies, we refer to studies that exclusively leverage audio, which differs from multimodal audiovisual studies presented earlier.

Processing audio includes two primary approaches: directly utilising audio as a signal or converting it into text and employing text-based methods. Table 7 summarises six studies and their methods for detecting empathy from audio datasets. Most of the studies reported classical ML algorithms as the best in corresponding experiments: SVM in empathy classification on the `CTT` [63], `COPE` [65] and `CallCentre` [67] datasets and LR in the regression study on the `CTT` dataset by Xiao *et al.* [63].

Given that audio-based datasets involve conversations between two persons, the work of Chen *et al.* [65] and Xiao *et al.* [63] include voice activity detection ('speech' or 'no speech') and speaker diarisation (i.e., speaker separation) in their empathy detection workflow. Xiao *et al.* [63], Chen *et al.* [65], and Alam *et al.* [67] converted the audio into text sequences, followed by extracting features from the text sequences. Chen *et al.* [65] and Alam *et al.* [67] extracted several lexical features, such as text embedding, from the audio transcripts and several acoustic features, such as MFCC, from the audio signal. Chen *et al.* [65] reported better performance of lexical features than acoustic features. Lastly, Xiao *et al.* [63]'s empathy detection model on audio-based `CTT` dataset is entirely text-based – leveraging uni-gram, bi-gram and tri-gram language models – without audio-based features.

On the recent `MultimodalMI` dataset, Tavabi *et al.* [51] used a distilled RoBERTa pre-trained language model, a bi-directional GRU layer followed by a two-head self-attention layer to predict continuous empathy score between 0 and 1 (regression task). Using the same dataset, Tran *et al.* [52] proposes a multimodal empathy classification system utilising both audio and text transcripts to predict high vs low empathy. Features from the audio and texts are extracted using Hidden-Unit BERT (HuBERT) [119] and distilled RoBERTa pre-trained models, respectively. The features are then passed through a bi-directional GRU model, followed by modality fusion. They experimented with early and late fusion through MLP layers. A wide range of experiments supports the effectiveness of late fusion in most experimental conditions, early fusion in some cases and text-only prediction in very few cases.

TABLE 8
Summary of Empathy Detection Studies from Physiological Signals.

Dataset	Study	Best Model	Performance	Code Avail.
Regression				
fMRI	[43]	LR	Pearson correlation: 0.54, MSE: 20.1	×
EEG	[75]	LR	MSE $\in 51.749, 150.556$	×
Classification				
PainEmp	[76]	SVM	Acc $\in$ [79%, 84%]	×
EEG	[75]	SVM, DT	Acc $\in$ [61.8%, 74.2%], F1 $\in$ [61.5%, 74.3%]	×

4.4 Physiological Signals

Research in physiological signal-based empathy detection typically adopts feature extraction, followed by ML algorithms. Table 8 reports the studies and methods of physiological signal-based empathy detection. All of these studies leveraged classical ML algorithms: LR and SVM each in two studies.

With the `PainEmp` dataset, Golbabaei *et al.* [76] extracted ten features and leveraged a SVM with radial basis function kernel to detect cognitive and affective empathy. Lastly, Kuijt and Alimardani [75] extracted 15 features from the `EEG` data and leveraged multiple LR in the regression task and LR, SVM and Decision Tree (DT) in the classification task. In the classification task, they only used five best-performing features. In both regression and classification settings, the participants' empathy before the experiment is better detected than 'after' and 'during' the experiment.

4.5 Discussion: Findings, Challenges and Research Gaps

4.5.1 Lack of Benchmarking

Benchmarking and comparative analysis of empathy detection models are hindered by a lack of unified dataset usage, particularly in the domains of audio and physiological signals. Code availability further impacts benchmarking and the broader adoption of methodologies, as minute details of proposed algorithms are often not fully captured in the papers. Among 50 text-based empathy detection models (Table 4 and Table 5), only 18 have shared their code. Similarly, for audiovisual, audio signal, and physiological signal-based models, the proportion of publicly accessible code is even lower, with only 6 out of 15 (Table 6), 2 out of 6 (Table 7), and none out of 4 (Table 8) releasing their implementations. Mandating the publication of code alongside research findings can ensure transparency and facilitate progress in the field.

A variety of evaluation metrics have been employed in the literature for empathy detection. For regression tasks predicting a continuous degree of empathy, metrics such as Pearson's correlation, Spearman's correlation, concordance, mean squared error, and R^2 are commonly used. In classification tasks predicting discrete empathy levels, metrics include accuracy, F1 score, Area Under the receiver operating characteristics Curve (AUC), average precision, unweighted average recall, and the Matthews correlation coefficient. However, inconsistent use of these metrics across studies can complicate cross-study comparisons. For instance, despite both using the `iEmpathize` dataset, the metrics reported in [36] and [103] differ. While it is unreasonable to expect universal adoption of a single metric, efforts toward greater standardisation in evaluation frameworks and metric reporting would enhance comparability.

4.5.2 Multimodal Empathy Detection

The growth of empathy detection modalities, as illustrated in Figure 2, shows a dominant rising trend in text-based empathy detection since 2020. However, the current body of research lacks equivalent development in audiovisual, audio, and physiological signals. Empathy detection systems based on these modalities can be particularly effective in scenarios where such signals are available and potentially provide a more comprehensive measure of empathy. While spoken information from video and audio can be converted to text for text-based empathy detection, video and audio contain additional information such as facial expressions and pitch. These elements can significantly enhance the accuracy and quality of empathy detection, which necessitates dedicated research in these areas.

A multimodal empathy detection system can effectively integrate these different types of data. Additionally, analysing the contributions of different modalities provides insights into the most important factors for an effective empathy detection system. Few studies [52]–[54], [78] have shown proof of concept towards multimodal empathy detection. Overall, the multimodal approach holds promise for creating a robust empathy detection system by leveraging the strengths of various input modalities.

4.5.3 LLM in Empathy Detection

The recent success of LLMs presents an opportunity to utilise them in empathy detection tasks. LLMs can serve as the primary prediction model or as a supportive tool to enhance predictions made by conventional models. While LLMs may excel in empathy detection due to their extensive language understanding capabilities, their training and deployment often require substantial resources, which may be impractical for low-resource settings. Smaller optimised models like BERT and RoBERTa can offer reasonable performance with better resource efficiency and may be better suited for certain applications, such as remote areas with limited healthcare access, community counselling centres, education settings in low-income schools, and humanitarian aid and crisis response.

Even when not utilised as the primary prediction model, LLMs can contribute to empathy prediction tasks, particularly in data preprocessing tasks such as text rephrasing [96] and empathy annotation [83]. A recent study [120] has shown that LLMs achieve human-level performance in theory of mind tasks. Drawing on the close relationship between cognitive empathy and theory of mind [24], this indicates that LLMs possess (or can mimic) empathic skills that could potentially assist in empathy detection.

Multimodal LLMs hold promise for empathy detection in real-life audiovisual interactions, as suggested by Hasan *et al.* [121]. This approach capitalises on LLMs' advanced language understanding and multimodal abilities to interpret the nuances of natural conversations across audio, visual and text modalities. The emergence of multimodal LLMs – such as OpenAI's GPT-4o (omni), GPT-4V (vision) and Google's Gemini – enables both zero-shot (i.e., no fine-tuning) and few-shot (i.e., fine-tuning with a few examples) applications in empathy detection. A recent study on GPT-4V [122] assessed its performance on 21 emotion recognition datasets across six tasks, including sentiment analysis, facial emotion recognition and multimodal emotion recognition, all in zero-shot settings. Although GPT-4V demonstrated strong visual processing capabilities, it struggled with micro-expression recognition. Since GPT-4V is designed primarily for general domains, future studies could focus on fine-tuning it in few-shot settings. Nonetheless, this proof of concept for multimodal LLMs in general emotion recognition indicates potential for empathy detection. With the rapid advancement of multimodal LLMs, they are likely to become a stronger candidate for empathy detection in the days to come.

5 Applications of Empathy Detection

Empathy detection has the potential to bring transformative changes across various domains, such as healthcare, education and social media, Focusing on a few key domains, this section discusses potential benefits, associated challenges and ethical considerations.

Assessment of Communication Quality

Empathy detection can be used to evaluate the quality of interpersonal communication, which can be used as feedback to improve interactions. Consider *healthcare* as an example. A study on patient-doctor interaction found that 85% of 563 patients either changed or were considering changing their doctors due to a lack of effective communication related to empathy being one of the main reasons [123], [124]. Empathic doctors would be better equipped to communicate medical information in a fashion that the patient will attend to [2]. Therefore, the service quality of healthcare providers could be assessed in terms of empathy if we could detect empathy in the first place. Assessment of healthcare providers can be in various contexts – such as counselling sessions [44], [61]–[63], oncology encounters [65], and other general patient-doctor interactions [28], [34] – either through telehealth or in-person. Measurement of empathy can also facilitate effective empathy training programs for healthcare professionals. In the same vein, empathy detection can be applied in educational interactions in teaching, customer services in businesses, and even in human-robot interactions.

Empathy detection can also improve communication quality in long-distance communication, such as those between international students and their families or during online interviews for jobs or university admissions. Video conferencing applications like Zoom and Microsoft Teams could incorporate live empathy feedback, similar to live transcripts, which is common nowadays. This way, people can see how their words and expressions are perceived in real-time, which may help people adjust their communication to be more empathic and responsive.

Disease Diagnosis

Empathy detection systems can help diagnose diseases and cognitive disorders where a lack of empathy is a symptom, such as autism, psychopathy and alexithymia [125]. Several studies have shown proof of concepts in this regard, such as diagnosing social communication disorders (Down syndrome, intellectual disability) [69] and autism spectrum disorders [76].

Social Media Moderation

People often seek mental support through social media platforms. Accordingly, several works have detected empathy in various social media interactions, such as Reddit [58], Twitter [57], and cancer survivors networks [36], [47], [103]. We can envision a peer support platform where non-empathic responses are filtered out through an empathy detection system. This way, social media platforms can foster empathic responses while discouraging non-empathic ones.

Ethical Considerations and Challenges

Visibility of empathy scores might encourage individuals to feign empathy, similar to how fake facial expressions are a concern in emotion detection [79]. Developing robust methods to differentiate genuine empathy from feigned responses will be crucial to ensure the reliability and effectiveness of empathy detection systems. Continuous feedback on empathic behaviour may also inadvertently create undue pressure on individuals, which may affect their mental health or authenticity in interactions. Safeguards should be in place to mitigate such unintended consequences.

What is considered a normal conversation and what is perceived as harsh can vary greatly across cultures. Ensuring fairness, therefore, becomes a significant challenge. Systems may inherit biases from training data that predominantly represent a specific demographic, resulting in unfair or inaccurate assessments towards other demographic or cultural groups. Addressing these biases is essential to ensure fairness and inclusivity. Before deploying an empathy detection system, it is imperative to rigorously evaluate the generalisability of the system across diverse populations. Apart from these, other ethical considerations and challenges, such as data privacy and consent, commonly associated with general Affective Computing tasks, must also be carefully addressed to ensure the responsible and ethical deployment of empathy detection systems.

6 Conclusion

Empathy, the capacity to comprehend and provide emotional support to others, has emerged as a promising research area across several disciplines. Empathy detection in Computer Science, particularly through ML methodologies, has grown substantially in recent years. In this research endeavour, this paper presents a rigorous systematic literature review following relevant Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines for reproducibility. Starting with an extensive search across ten scholarly databases, we select 62 papers after a thorough screening process, including abstract and full-text screening based on exclusion criteria. We discuss and group similar papers based on task formulations and ML methodologies. We present a task formulation hierarchy with representative datasets and their details, such as data collection, experiment detail, statistics, annotation protocol and public availability. To describe ML methodologies, we group our findings based on four input modalities: text sequences, audiovisual data, audio signals and physiological signals. In each modality, we enumerate the algorithms used, their performance and code availability.

This review uncovers several new insights into the computational empathy domain and identifies critical avenues for future research and development. Exploring novel task formulations such as parallel and bidirectional empathy, particularly in group settings and global-level measurements, can advance our understanding of empathy in complex social interactions. Further research comparing self-annotation and third-party annotation under controlled conditions is necessary to determine the most appropriate annotation scheme for empathy detection. The limited public availability of datasets and codes poses significant challenges for reproducibility and benchmarking in the field. At the same time, the diversity and lack of standardisation in evaluation metrics complicate consistent model comparison. Physiological signals offer promising avenues for more accurate empathy detection. Finally, the potential of LLMs and multimodal approaches to enhance empathy detection systems presents exciting opportunities for future research.

Acknowledgement

We thank Dr Susannah Soon, Senior Lecturer at Curtin University, for her comments on the initial version of this paper.

References

[1] L. Verhofstadt, I. Devoldre, A. Buysse, *et al.*, "The role of cognitive and affective empathy in spouses' support interactions: An observational study," *PloS one*, vol. 11, no. 2, e0149944, 2016.

[2] B. D. Jani, D. N. Blane, and S. W. Mercer, "The role of empathy in therapy and the physician-patient relationship," *Complementary Medicine Research*, vol. 19, no. 5, pp. 252–257, 2012.

[3] K. Aldrup, B. Carstensen, and U. Klusmann, "Is empathy the key to effective teaching? a systematic review of its association with teacher-student interactions and student outcomes," *Educational Psychology Review*, vol. 34, no. 3, pp. 1177–1216, 2022.

[4] M. L. Hoffman, *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge University Press, 2000.

[5] N. Eisenberg and A. S. Morris, "The origins and social significance of empathy-related responding. a review of empathy and moral development: Implications for caring and justice by M. L. Hoffman," *Social Justice Research*, no. 1, pp. 95–120, 2001.

[6] J. A. Hall and R. Schwartz, "Empathy present and future," *The Journal of social psychology*, vol. 159, no. 3, pp. 225–243, 2019.

[7] A. Paiva, I. Leite, H. Boukricha, and I. Wachsmuth, "Empathy in virtual agents and robots: A survey," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 7, no. 3, pp. 1–40, 2017.

[8] E. Sariyanidi, H. Gunes, and A. Cavallaro, “Automatic analysis of facial affect: A survey of registration, representation, and recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[9] E. A. Veltmeijer, C. Gerritsen, and K. V. Hindriks, “Automatic emotion recognition for groups: A review,” *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 89–107, 2023.

[10] Ö. N. Yalcin and S. DiPaola, “A computational model of empathy for interactive agents,” *Biologically inspired cognitive architectures*, vol. 26, pp. 20–25, 2018.

[11] S. Park and M. Whang, “Empathy in human–robot interaction: Designing for social robots,” *International journal of environmental research and public health*, vol. 19, no. 3, p. 1889, 2022.

[12] A. S. Raamkumar and Y. Yang, “Empathetic conversational systems: A review of current advances, gaps, and opportunities,” *IEEE Transactions on Affective Computing*, 2022.

[13] A. Lahnala, C. Welch, D. Jurgens, and L. Flek, “A critical reflection and forward perspective on empathy and natural language processing,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 2139–2158.

[14] V. A. Shetty, S. Durbin, M. S. Weyrich, A. D. Martínez, J. Qian, and D. L. Chin, “A scoping review of empathy recognition in text using natural language processing,” *Journal of the American Medical Informatics Association*, vol. 31, no. 3, pp. 762–775, Dec. 2023.

[15] M. J. Page, J. E. McKenzie, P. M. Bossuyt, *et al.*, “The PRISMA 2020 statement: An updated guideline for reporting systematic reviews,” *International journal of surgery*, vol. 88, p. 105 906, 2021.

[16] B. M. Cuff, S. J. Brown, L. Taylor, and D. J. Howat, “Empathy: A review of the concept,” *Emotion Review*, vol. 8, no. 2, pp. 144–153, 2016.

[17] M. L. Hoffman, “Toward a theory of empathic arousal and development,” in *The Development of Affect*, M. Lewis and L. A. Rosenblum, Eds. Boston, MA: Springer US, 1978, pp. 227–256.

[18] D. Goleman, *Emotional intelligence*. Bloomsbury Publishing, 2020.

[19] M. H. Davis *et al.*, “A multidimensional approach to individual differences in empathy,” 1980.

[20] C. D. Batson, J. Fultz, and P. A. Schoenrade, “Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences,” *Journal of Personality*, vol. 55, no. 1, pp. 19–39, 1987.

[21] S. D. Preston and F. B. M. de Waal, “Empathy: Its ultimate and proximate bases,” *Behavioral and Brain Sciences*, vol. 25, no. 1, pp. 1–20, 2002.

[22] W. Ickes, *Everyday mind reading: Understanding what other people think and feel*. Prometheus Books, 2003.

[23] H. P. Becker, “Some forms of sympathy: A phenomenological analysis.,” *The Journal of Abnormal and Social Psychology*, vol. 26, pp. 58–68, 1931.

[24] R. J. R. Blair, “Responding to the emotions of others: Dissociating forms of empathy through the study of typical and psychiatric populations,” *Consciousness and cognition*, vol. 14, no. 4, pp. 698–718, 2005.

[25] G. Hein and T. Singer, “I feel how you feel but not always: The empathic brain and its modulation,” *Current Opinion in Neurobiology*, vol. 18, no. 2, pp. 153–158, 2008, Cognitive neuroscience.

[26] J. Decety and K. J. Michalska, “Neurodevelopmental changes in the circuits underlying empathy and sympathy from childhood to adulthood,” *Developmental Science*, vol. 13, no. 6, pp. 886–899, 2010.

[27] E. C. Montiel-Vázquez, J. A. Ramírez Uresti, and O. Loyola-González, “An explainable artificial intelligence approach for detecting empathy in textual communication,” *Applied Sciences*, vol. 12, no. 19, p. 9407, 2022.

[28] S. Shi, Y. Sun, J. Zavala, J. Moore, and R. Girju, “Modeling clinical empathy in narrative essays,” in *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, 2021, pp. 215–220.

[29] P. Dey and R. Girju, “Investigating stylistic profiles for the task of empathy classification in medical narrative essays,” in *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, C. Bonial and H. Tayyar Madabushi, Eds., Washington, D.C.: Association for Computational Linguistics, Mar. 2023, pp. 63–74.

[30] S. Buechel, A. Buffone, B. Slaff, L. Ungar, and J. Sedoc, “Modeling empathy and distress in reaction to news stories,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 4758–4765.

[31] A. I. A. Rahim, M. I. Ibrahim, K. I. Musa, S.-L. Chua, and N. M. Yaacob, “Assessing patient-perceived hospital service quality and sentiment in malaysian public hospitals using machine learning and facebook reviews,” *International Journal of Environmental Research and Public Health*, vol. 18, no. 18, 2021.

[32] A. Smith, “Cognitive empathy and emotional empathy in human behavior and evolution,” *The Psychological Record*, vol. 56, no. 1, pp. 3–21, 2006.

[33] M. L. Healey and M. Grossman, “Cognitive and affective perspective-taking: Evidence for shared and dissociable anatomical substrates,” *Frontiers in neurology*, vol. 9, p. 372 314, 2018.

[34] P. Dey and R. Girju, “Enriching deep learning with frame semantics for empathy classification in medical narrative essays,” in *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 207–217.

[35] R. W. Picard, *Affective Computing*. MIT press, 2000.

[36] M. Hosseini and C. Caragea, “It takes two to empathize: One to seek and one to provide,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 13 018–13 026.

[37] M. H. Davis, “Measuring individual differences in empathy: Evidence for a multidimensional approach.,” *Journal of personality and social psychology*, vol. 44, no. 1, p. 113, 1983.

[38] S. Baron-Cohen and S. Wheelwright, “The empathy quotient: An investigation of adults with asperger syndrome or high functioning autism, and normal sex differences,” *Journal of autism and developmental disorders*, vol. 34, pp. 163–175, 2004.

[39] R. N. Spreng*, M. C. McKinnon*, R. A. Mar, and B. Levine, “The toronto empathy questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures,” *Journal of personality assessment*, vol. 91, no. 1, pp. 62–71, 2009.

[40] J. A. Hall and M. Schmid Mast, “Sources of accuracy in the empathic accuracy paradigm.,” *Emotion*, vol. 7, no. 2, pp. 438–446, 2007.

[41] S. Tafreshi, O. De Clercq, V. Barriere, S. Buechel, J. Sedoc, and A. Balahur, “WASSA 2021 shared task: Predicting empathy and emotion in reaction to news stories,” in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Online: Association for Computational Linguistics, Apr. 2021, pp. 92–104.

[42] V. Barriere, J. Sedoc, S. Tafreshi, and S. Giorgi, “Findings of WASSA 2023 shared task on empathy, emotion and personality detection in conversation and reactions to news articles,” in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 511–525.

[43] L. Wei, G.-R. Wu, M. Bi, and C. Baeken, “Effective connectivity predicts cognitive empathy in cocaine addiction: A spectral dynamic causal modeling study,” *Brain Imaging and Behavior*, vol. 15, pp. 1553–1561, 2021.

[44] Z. Wu, R. Helaoui, D. Reforgiato Recupero, and D. Riboni, “Towards low-resource real-time assessment of empathy in counselling,” in *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, Online: Association for Computational Linguistics, Jun. 2021, pp. 204–216.

[45] D. Sedefoglu, A. C. Lahnala, J. Wagner, L. Flek, and S. Ohly, “LeadEmpathy: An expert annotated German dataset of empathy in written leadership communication,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italy: ELRA and ICCL, May 2024, pp. 10 237–10 248.

[46] A. Tapus and M. J. Mataric, “Socially assistive robots: The link between personality, empathy, physiological signals, and task performance.,” in *AAAI spring symposium: emotion, personality, and social behavior*, 2008, pp. 133–140.

[47] H. Khanpour, C. Caragea, and P. Biyani, "Identifying empathetic messages in online health communities," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 246–251.

[48] V. Pérez-Rosas, X. Wu, K. Resnicow, and R. Mihalcea, "What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 926–935.

[49] P. Zhong, C. Zhang, H. Wang, Y. Liu, and C. Miao, "Towards persona-based empathetic conversational models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 6556–6566.

[50] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5370–5381.

[51] L. Tavabi, T. Tran, B. Borsari, *et al.*, "Therapist empathy assessment in motivational interviews," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2023, pp. 1–8.

[52] T. Tran, Y. Yin, L. Tavabi, *et al.*, "Multimodal analysis and assessment of therapist empathy in motivational interviews," in *Proceedings of the 25th International Conference on Multimodal Interaction*, ser. ICMI '23, Paris, France: Association for Computing Machinery, 2023, pp. 406–415.

[53] Z. Zhu, C. Li, J. Pan, *et al.*, "MEDIC: A multimodal empathy dataset in counseling," in *Proceedings of the 31st ACM International Conference on Multimedia*, New York, NY, USA: Association for Computing Machinery, 2023, pp. 6054–6062.

[54] L. Tavabi, K. Stefanov, S. Nasihati Gilani, D. Traum, and M. Soleymani, "Multimodal learning for identifying opportunities for empathetic responses," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 95–104.

[55] J. Gratch, R. Artstein, G. Lucas, *et al.*, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3123–3128.

[56] S. Singh and A. Rios, "Linguistic elements of engaging customer service discourse on social media," in *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, Abu Dhabi, UAE: Association for Computational Linguistics, Nov. 2022, pp. 105–117.

[57] M. Hosseini and C. Caragea, "Distilling knowledge for empathy detection," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3713–3724.

[58] A. Sharma, A. Miner, D. Atkins, and T. Althoff, "A computational approach to understanding empathy expressed in text-based mental health support," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 5263–5276.

[59] M. Hosseini and C. Caragea, "Calibrating student models for emotion-related tasks," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 9266–9278.

[60] G. Lee and N. Parde, "AcnEmpathize: A dataset for understanding empathy in dermatology conversations," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 143–153.

[61] J. Gibson, N. Malandrakis, F. Romero, D. C. Atkins, and S. S. Narayanan, "Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms," in *Sixteenth annual conference of the international speech communication association*, 2015, pp. 1947–1951.

[62] J. Gibson, D. Can, B. Xiao, *et al.*, "A deep learning approach to modeling empathy in addiction counseling," *Commitment*, vol. 111, no. 21, pp. 2016–554, 2016.

[63] B. Xiao, Z. E. Imel, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, ""Rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing," *PloS One*, vol. 10, no. 12, e0143055, 2015.

[64] J. S. Baer, E. A. Wells, D. B. Rosengren, B. Hartzler, B. Beadnell, and C. Dunn, "Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors," *Journal of substance abuse treatment*, vol. 37, no. 2, pp. 191–202, 2009.

[65] Z. Chen, J. Gibson, M.-C. Chiu, *et al.*, "Automated empathy detection for oncology encounters," in *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, 2020, pp. 1–8.

[66] J. A. Tulsky, R. M. Arnold, S. C. Alexander, *et al.*, "Enhancing communication between oncologists and patients with a computer-based training program: A randomized trial," *Annals of internal medicine*, vol. 155, no. 9, pp. 593–601, 2011.

[67] F. Alam, M. Danieli, and G. Riccardi, "Can we detect speakers' empathy?: A real-life case study," in *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, IEEE, 2016, pp. 000 059–000 064.

[68] L. Mathur, M. Spitale, H. Xi, J. Li, and M. J. Matarić, "Modeling user empathy elicited by a robot storyteller," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2021, pp. 1–8.

[69] R. Hervás, E. Johnson, C. G. L. de la Franca, J. Bravo, and T. Mondéjar, "A learning system to support social and empathy disorders diagnosis through affective avatars," in *2016 15th International Conference on Ubiquitous Computing and Communications and 2016 International Symposium on Cyberspace and Security (IUCC-CSS)*, IEEE, 2016, pp. 93–100.

[70] K. Kroes, I. Saccardi, and J. Masthoff, "Empathizing with virtual agents: The effect of personification and general empathic tendencies," in *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, IEEE, 2022, pp. 73–81.

[71] J. Shen, M. Sap, P. Colon-Hernandez, H. Park, and C. Breazeal, "Modeling empathic similarity in personal narratives," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6237–6252.

[72] W. R. Miller, T. B. Moyers, D. Ernst, and P. Amrhein, "Manual for the motivational interviewing skill code (misc)," *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*, 2003.

[73] T. Moyers, T. Martin, J. Manuel, W. Miller, and D. Ernst, "Revised global scales: Motivational interviewing treatment integrity 3.1. 1 (miti 3.1. 1)," *Unpublished manuscript, University of New Mexico, Albuquerque, NM*, 2010.

[74] P. Barros, N. Churamani, A. Lim, and S. Wermter, "The omg-empathy dataset: Evaluating the impact of affective behavior in storytelling," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2019, pp. 1–7.

[75] A. Kuijt and M. Alimardani, "Prediction of human empathy based on eeg cortical asymmetry," in *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, 2020, pp. 1–5.

[76] S. Golbabaei, N. SammakNejad, and K. Borhani, "Physiological indicators of the relation between autistic traits and empathy: Evidence from electrocardiogram and skin conductance signals," in *2022 29th National and 7th International Iranian Conference on Biomedical Engineering (ICBME)*, 2022, pp. 177–183.

[77] M. Abdul-Mageed, A. Buffone, H. Peng, J. Eichstaedt, and L. Ungar, "Recognizing pathogenic empathy in social media," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, 2017, pp. 448–451.

[78] Y. Pan, J. Wu, R. Ju, *et al.*, "A multimodal framework for automated teaching quality assessment of one-to-many online instruction videos," in *2022 26th International Conference on Pattern Recognition (ICPR)*, IEEE, 2022, pp. 1777–1783.

[79] M. Z. Hossain, T. Gedeon, and R. Sankaranarayana, "Using temporal features of observers' physiological measures to distinguish between genuine and fake smiles," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 163–173, 2020.

[80] L. Chen, T. Gedeon, M. Z. Hossain, and S. Caldwell, "Are you really angry? detecting emotion veracity as a proposed tool for interaction," in *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, New York, NY, USA: Association for Computing Machinery, 2017, pp. 412–416.

[81] S. Afzal and P. Robinson, "Natural affect data: Collection and annotation," in *New Perspectives on Affect and Learning Technologies*, R. A. Calvo and S. K. D'Mello, Eds., New York, NY: Springer New York, 2011, pp. 55–70.

[82] M. Mars, "From word embeddings to pre-trained language models: A state-of-the-art walkthrough," *Applied Sciences*, vol. 12, no. 17, p. 8805, 2022.

[83] M. R. Hasan, M. Z. Hossain, T. Gedeon, and S. Rahman, "LLM-GEm: Large language model-guided prediction of people's empathy levels towards newspaper article," in *Findings of the Association for Computational Linguistics: EACL 2024*, Y. Graham and M. Purver, Eds., St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 2215–2231.

[84] Y. Butala, K. Singh, A. Kumar, and S. Shrivastava, "Team Phoenix at WASSA 2021: Emotion analysis on news stories with pre-trained language models," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Online: Association for Computational Linguistics, Apr. 2021, pp. 274–280.

[85] G. Vettigli and A. Sorgente, "EmpNa at WASSA 2021: A lightweight model for the prediction of empathy, distress and emotions from reactions to news stories," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Online: Association for Computational Linguistics, Apr. 2021, pp. 264–268.

[86] A. Kulkarni, S. Somwase, S. Rajput, and M. Marathe, "PVG at WASSA 2021: A multi-input, multi-task, transformer-based architecture for empathy and distress prediction," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Online: Association for Computational Linguistics, Apr. 2021, pp. 105–111.

[87] J. Mundra, R. Gupta, and S. Mukherjee, "WASSA@IITK at WASSA 2021: Multi-task learning and transformer finetuning for emotion classification and empathy prediction," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Online: Association for Computational Linguistics, Apr. 2021, pp. 112–116.

[88] H. Vasava, P. Uikey, G. Wasnik, and R. Sharma, "Transformer-based architecture for empathy prediction and emotion classification," in *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 2022, pp. 261–264.

[89] S. Ghosh, D. Maurya, A. Ekbal, and P. Bhattacharyya, "Team IITP-AINLPML at WASSA 2022: Empathy detection, emotion classification and personality detection," in *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 2022, pp. 255–260.

[90] S. Qian, C. Orašan, D. Kanojia, H. Saadany, and F. Do Carmo, "SURREY-CTS-NLP at WASSA2022: An experiment of discourse and sentiment analysis for the prediction of empathy, distress and emotion," in *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 2022, pp. 271–275.

[91] A. Lahnala, C. Welch, and L. Flek, "CAISA at WASSA 2022: Adapter-tuning for empathy prediction," in *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 2022, pp. 280–285.

[92] Y. Chen, Y. Ju, and S. Kübler, "IUCL at WASSA 2022 shared task: A text-only approach to empathy and emotion detection," in *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 2022.

[93] F. M. Plaza-del-Arco, J. Collado-Montañez, L. A. Ureña, and M.-T. Martín-Valdivia, "Empathy and distress prediction using transformer multi-output regression and emotion analysis with an ensemble of supervised and zero-shot learning models," in *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 239–244.

[94] M. R. Hasan, M. Z. Hossain, T. Gedeon, S. Soon, and S. Rahman, "Curtin OCAI at WASSA 2023 empathy, emotion and personality shared task: Demographic-aware prediction using multiple transformers," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 536–541.

[95] A. S. Srinivas, N. Barua, and S. Pal, "Team_Hawk at WASSA 2023 empathy, emotion, and personality shared task: Multi-tasking multi-encoder based transformers for empathy and emotion prediction in conversations," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 542–547.

[96] X. Lu, Z. Li, Y. Tong, Y. Zhao, and B. Qin, "HIT-SCIR at WASSA 2023: Empathy and emotion analysis at the utterance-level and the essay-level," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 574–580.

[97] Y. Wang, J. Wang, and X. Zhang, "YNU-HPCC at WASSA-2023 shared task 1: Large-scale language model with LoRA fine-tuning for empathy detection and emotion classification," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 526–530.

[98] F. Gruschka, A. Lahnala, C. Welch, and L. Flek, "Caisa at wassa 2023 shared task: Domain transfer for empathy, distress, and personality prediction," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 553–557.

[99] T. Chavan, K. Deshpande, and S. Sonawane, "PICT-CLRL at WASSA 2023 empathy, emotion and personality shared task: Empathy and distress detection using ensembles of transformer models," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 564–568.

[100] T.-M. Lin, J.-Y. Chang, and L.-H. Lee, "NCUEE-NLP at WASSA 2023 shared task 1: Empathy and emotion prediction using sentiment-enhanced RoBERTa transformers," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 548–552.

[101] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[102] A. Lee, J. K. Kummerfeld, L. An, and R. Mihalcea, "Empathy identification systems are not accurately accounting for context," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds., Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1686–1695.

[103] M. Hosseini and C. Caragea, "Feature normalization and cartography-based demonstrations for prompt-based fine-tuning on emotion-related tasks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 12 881–12 889.

[104] A. Lee, J. K. Kummerfeld, L. An, and R. Mihalcea, "Micromodels for efficient, explainable, and reusable systems: A case study on mental health," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds., Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4257–4272.

[105] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet project," in *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, Montreal, Quebec, Canada: Association for Computational Linguistics, Aug. 1998, pp. 86–90.

[106] S. Hinduja, M. T. Uddin, S. R. Jannat, A. Sharma, and S. Canavan, "Fusion of hand-crafted and deep features for empathy prediction," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, IEEE, 2019, pp. 1–4.

[107] B. Azari, Z. Zhang, and A. Lim, "Towards an emocog model for multimodal empathy prediction," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019, pp. 1–4.

[108] A. Mallol-Ragolta, M. Schmitt, A. Baird, N. Cummins, and B. Schuller, "Performance analysis of unimodal and multimodal models in valence-based empathy recognition," in *2019 14th*

IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), 2019, pp. 1–5.

[109] Z.-X. Tan, A. Goel, T.-S. Nguyen, and D. C. Ong, "A multimodal lstm for predicting listener empathic responses over time," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019, pp. 1–4.

[110] F. Barbieri, E. Guizzo, F. Lucchesi, G. Maffei, F. M. del Prado Martín, and T. Weyde, "Towards a multimodal time-based empathy prediction system," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019, pp. 1–5.

[111] L. A. D. Lusquino Filho, L. F. R. Oliveira, H. C. C. Carneiro, *et al.*, "A weightless regression system for predicting multi-modal empathy," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020, pp. 657–661.

[112] S. A. Alanazi, M. Shabbir, N. Alshammari, M. Alruwaili, I. Hussain, and F. Ahmad, "Prediction of emotional empathy in intelligent agents to facilitate precise social interaction," *Applied Sciences*, vol. 13, no. 2, p. 1163, 2023.

[113] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*, British Machine Vision Association, 2015.

[114] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[115] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[116] B. Amos, B. Ludwiczuk, M. Satyanarayanan, *et al.*, "OpenFace: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, vol. 6, no. 2, p. 20, 2016.

[117] M. R. Hasan, M. M. Hasan, and M. Z. Hossain, "How many Mel-frequency cepstral coefficients to be utilized in speech recognition? a study with the Bengali language," *The Journal of Engineering*, vol. 2021, no. 12, pp. 817–827, 2021.

[118] M. Chen and X. Li, "SWAFN: Sentimental words aware fusion network for multimodal sentiment analysis," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds., Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1067–1077.

[119] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[120] W. Street, J. O. Siy, G. Keeling, *et al.*, *Llms achieve adult human performance on higher-order theory of mind tasks*, 2024.

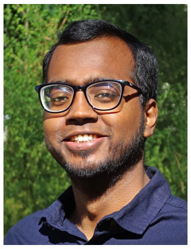
[121] M. R. Hasan, M. Z. Hossain, A. Krishna, S. Rahman, and T. Gedeon, "Thesis proposal: Detecting empathy using multimodal language model," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 338–349.

[122] Z. Lian, L. Sun, H. Sun, *et al.*, "Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition," *Information Fusion*, vol. 108, p. 102 367, 2024.

[123] N. Cousins, "How patients appraise physicians," *New England Journal of Medicine*, vol. 313, no. 22, pp. 1422–1424, 1985.

[124] P. S. Bellet and M. J. Maloney, "The importance of empathy as an interviewing skill in medicine," *JAMA*, vol. 266, no. 13, pp. 1831–1832, Oct. 1991.

[125] C. Lamm, H. Bukowski, and G. Silani, "From shared to distinct self–other representations in empathy: Evidence from neurotypical function and socio-cognitive disorders," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 371, no. 1686, p. 20 150 083, 2016.



Md Rakibul Hasan received his BSc (Hons) (2019) and MSc (2021) degrees from Khulna University of Engineering & Technology, Bangladesh. Currently, he is a PhD candidate in Computing at Curtin University, Western Australia, where he builds deep learning models to detect empathy from multimodal data, including text, video and audio signals. His overarching research interest includes advancing Affective Computing and multimodal systems using deep learning algorithms.

Md Zakir Hossain completed the BSc (2011) and MSc (2014) from Khulna University of Engineering & Technology (KUET), Bangladesh, and PhD (2019) from the Australian National University (ANU). He is a Senior Research Fellow at the Curtin University. His research direction leads to the development of advanced technologies for health-related prediction, including emotion / facial expression recognition, human computing, and diagnosing and managing diseases.

Shreya Ghosh received a BTech degree in CSE from the Govt. College of Engineering and Textile Technology, India, and the MS(R) degree in computer science and engineering from the Indian Institute of Technology Ropar, India, and the PhD degree from the Monash University, Australia, in 2022. She is currently a research academic at Curtin University. Her research interests include affective computing, computer vision, and deep learning.

Aneesh Krishna is currently an Associate Professor and Discipline Lead of Computing within the School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Australia. He holds a PhD in computer science from the University of Wollongong, Australia. His research interests include AI for software engineering, model-driven development/evolution, data mining, computer vision, and machine learning. His research is (or has been) funded by the Australian Research Council (ARC), and various Australian government agencies as well as companies.

Tom Gedeon is the Human-Centric Advancements Chair in AI and was recently the Optus Chair in AI at Curtin University. Prior to this, he was a Professor of Computer Science and former Deputy Dean of the College of Engineering and Computer Science at ANU. He gained his BSc (Hons) and PhD from the University of Western Australia. He has over 400 publications and has run multiple international conferences. He is a former president of the Asia-Pacific Neural Network Assembly and former President of the Computing Research and Education Association of Australasia. He was a member of the Australian Research Council's College of Experts from 2018-2021 and continues from 2024-2026.

Empathy Detection from Text, Audiovisual, Audio or Physiological Signals: A Systematic Review of Task Formulations and Machine Learning Methods

Md Rakibul Hasan, *Graduate Student Member, IEEE,* Md Zakir Hossain, *Member, IEEE,* Shreya Ghosh, Aneesh Krishna, and Tom Gedeon, *Senior Member, IEEE*

Supplementary Material

This document provides supplementary details for the reported systematic review. Section 1 outlines the paper selection process, including inclusion and exclusion criteria, search keywords, search results, and the number of papers at each stage of screening. Section 2 lists the acronyms used throughout the manuscript.

1 Paper Selection

The systematic nature of this review ensures reproducibility and transparency in the selection and analysis of relevant papers. We adhered to the relevant recommendations from the PRISMA 2020 guidelines [1] as well as published systematic reviews in Affective Computing [2]–[7]. For example, our paper selection process considered the following inclusion and exclusion criteria.

Inclusion Criteria (IC)

IC1. Detect empathy using any Machine Learning (ML) algorithms
IC2. Peer-reviewed full-length research paper
IC3. Published between January 2013 and June 2024

Exclusion Criteria (EC)

EC1. Not a full-length research paper (e.g., conference abstracts and conference proceeding books)
EC2. No use of artificial intelligence, machine learning or deep learning
EC3. Review, survey, meta-analysis, thesis or dissertation
EC4. Not in English

We included papers from 2013 onward to capture the period when ML, and particularly Deep Learning (DL), became widely feasible. Key developments, such as the first modern CNN (AlexNet) in 2012 [8] and optimisation algorithms like Adam [9] in 2014-2015, lead DL's success on a variety of tasks [10], [11]. Our systematic search validated this time frame, as we found no relevant empathy detection studies published in 2013 or 2014.

1.1 Paper Search

We formulate a search string using logical operators (AND and OR) among synonymous terms of empathy, detection and artificial intelligence: empath* AND (detect* OR recog*) AND ("deep learning" OR "machine learning" OR "artificial intelligence" OR AI). The asterisk (*) is a wildcard character that facilitates the inclusion of any number of characters in place of the asterisk.

With the search string, one researcher (MRH) searched for relevant records across ten databases (see Supplementary Table 1 for more details) on 24 February 2023. Among the search engines, ACL Anthology does not support logical search. We, therefore, built a program[1] to search in the ACL database using the available bibliography document. Several search engines, such as Scopus and Web of Science, support filtering based on publication year (IC3) and paper type (EC1 and EC3), so we automatically filtered out the search results. Supplementary Table 1 presents the number of search results, search condition (e.g., title, abstract, full-paper, etc.), automatic-filtering results and corresponding filtering criteria.

1.2 Paper Screening

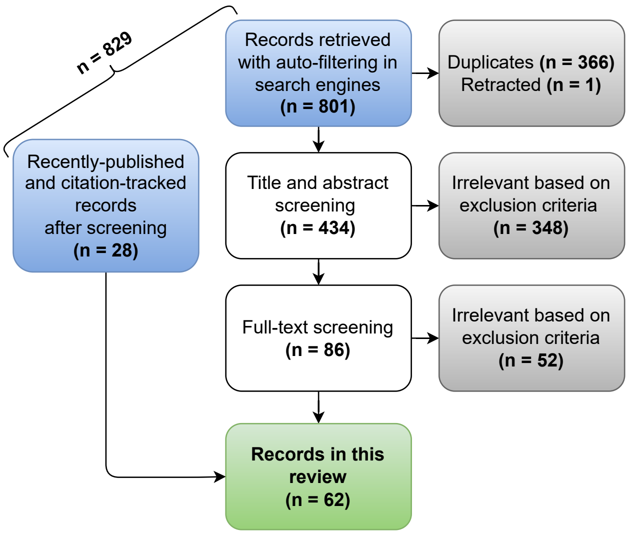
Supplementary Figure 1 illustrates our step-by-step paper screening strategy. We initially gathered 801 records from the databases. After removing duplicated and retracted records, one researcher (MRH) screened the remaining 434

- *All authors are with the School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Perth WA 6102, Australia. E-mail: {Rakibul.Hasan, Zakir.Hossain1, Shreya.Ghosh, A.Krishna, Tom.Gedeon}@curtin.edu.au*
- *M. R. Hasan is also with BRAC University, Bangladesh.*
- *M. Z. Hossain is also with The Australian National University, Australia.*
- *T. Gedeon is also with the University of ÓBuda, Hungary.*

[1] https://github.com/hasan-rakibul/boolean-search-bib-abstract

SUPPLEMENTARY TABLE 1
Initial Search Results with Details in All 10 Databases.

SL	Database	Search condition	Items	Auto-filtering criteria	Post-filter items
1	Scopus	Searched in title, abstract and keywords	233	EC4, EC5	198
2	Web of Science	Searched in title, abstract, keywords on all databases	227	EC4, EC5	183
3	ScienceDirect	Search engine did not support wildcard	27	EC4, EC5	16
4	IEEE Xplore	Searched in all metadata	93	EC4	84
5	ACM Guide to Computing Literature	Searched in abstracts	25	EC4	22
6	dblp	Combined dblp search; search string: empath (detect | recog)	37	–	37
7	Google Scholar	Sorted by relevance	18,100	EC4, First 100	100
8	PubMed	Searched in all fields	55	EC4	51
9	ProQuest	Searched in abstracts	93	EC4	88
10	ACL Anthology	Searched in title and abstract	22	–	22


n = 829
Records retrieved with auto-filtering in search engines (n = 801)
Duplicates (n = 366) Retracted (n = 1)
Recently-published and citation-tracked records after screening (n = 28)
Title and abstract screening (n = 434)
Irrelevant based on exclusion criteria (n = 348)
Full-text screening (n = 86)
Irrelevant based on exclusion criteria (n = 52)
Records in this review (n = 62)

SUPPLEMENTARY FIGURE 1. Number of records at different stages in our screening process. Refer to Section 1 of the Supplementary Material for more details.

records by reading titles and abstracts using the Covidence systematic review management software [12]. At this stage, records were excluded only if they clearly met one or more Exclusion Criteria (EC). Records that could not be conclusively evaluated based on the title and abstract proceeded to the full-text screening stage. At this stage, three researchers (MRH, MZH and SG) screened the 86 remaining records. Some records were deemed ambiguous for inclusion or exclusion, and these were discussed among all researchers in regular weekly meetings during the full-text screening period until a consensus was reached. For instance, while Hossain and Rahman [13] and Hinduja [14] initially appeared to meet the inclusion criteria, further examination revealed that Hossain and Rahman [13] perform sentiment analysis on how people react toward online reviews, and Hinduja [14] analyses potential biases in empathy detection without detecting any types of empathy.

We screened another 27 recent papers, which we received through notifications and 'snowballing'. Several search engines, such as Scopus, Web of Science, IEEE Xplore, ACM and Google Scholar, offer email notification services based on a predefined search string. Our 'snowball' search involves examining reference lists of the included papers to identify potential new records. In the latest update on 06 June 2024, we incorporated eight newly published papers [15]–[22], which made the total number of relevant papers in this systematic review to 62. MRH, in consultation with other researchers (co-authors) in this study, extracted relevant information from these selected papers. We categorise the analysis of the selected papers based on task formulations and data modality: text, audiovisual, audio and physiological signals.

Overall, the selection and analysis of papers were systematic and transparent, involving automated searches across multiple databases, clearly defined inclusion and exclusion criteria, and collaborative decision-making. This definitive and collaborative approach minimised the likelihood of individual bias. However, potential biases may still arise from limitations in the selected databases, which might not capture all relevant studies, as well as from the interpretation of ambiguous records by the researchers.

2 List of Acronyms

AUC Area Under the receiver operating characteristics Curve
BERT Bidirectional Encoder Representations from Transformers
BiLSTM Bidirectional LSTM
CNN Convolutional Neural Network
DeBERTa Decoding-Enhanced BERT with Disentangled Attention
DistilBERT Distilled BERT
DL Deep Learning
DT Decision Tree
EBM Explainable Boosting Machine
EC Exclusion Criteria
ECG Electrocardiogram
EEG Electroencephalogram
fMRI functional Magnetic Resonance Imaging
GPT Generative Pre-trained Transformer
GRU Gated Recurrent Unit
HuBERT Hidden-Unit BERT
IC Inclusion Criteria
IRI Interpersonal Reactivity Index
LLM Large Language Model
LR Linear Regression
LogR Logistic Regression
LSTM Long Short-Term Memory

MFCC Mel-Frequency Cepstral Coefficients

MISC Motivational Interviewing Skill Code

MITI Motivational Interviewing Treatment Integrity

ML Machine Learning

MLP Multi Layer Perceptron

NB Naïve Bayes

NLP Natural Language Processing

PBC4cip Pattern-Based Classifier for Class Imbalance Problems

PRISMA Preferred Reporting Items for Systematic reviews and Meta-Analyses

ResNet Residual Network

RNN Recurrent Neural Network

RF Random Forest

RoBERTa Robustly Optimised BERT Pretraining Approach

RR Ridge Regression

SVM Support Vector Machine

SBERT Sentence BERT

WANN Weightless Artificial Neural Network

WASSA Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis

REFERENCES

[1] M. J. Page, J. E. McKenzie, P. M. Bossuyt, *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *International journal of surgery*, vol. 88, p. 105 906, 2021. DOI: 10.1016/j.ijsu.2021.105906.

[2] A. Pampouchidou, P. G. Simos, K. Marias, *et al.*, "Automatic assessment of depression based on visual cues: A systematic review," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 445–470, 2019. DOI: 10.1109/TAFFC.2017.2724035.

[3] R. V. Aranha, C. G. Corrêa, and F. L. S. Nunes, "Adapting software with affective computing: A systematic review," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 883–899, 2021. DOI: 10.1109/TAFFC.2019.2902379.

[4] H. Ma and S. Yarosh, "A review of affective computing research based on function-component-representation framework," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1655–1674, 2023. DOI: 10.1109/TAFFC.2021.3104512.

[5] L. Pepa, L. Spalazzi, M. Capecci, and M. G. Ceravolo, "Automatic emotion recognition in clinical scenario: A systematic review of methods," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1675–1695, 2023. DOI: 10.1109/TAFFC.2021.3128787.

[6] S. Saganowski, B. Perz, A. G. Polak, and P. Kazienko, "Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1876–1897, 2023. DOI: 10.1109/TAFFC.2022.3176135.

[7] E. M. Jacobs, F. Deligianni, and F. Pollick, "Threat perception captured by emotion, motor and empathetic system responses: A systematic review," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 1116–1135, 2024. DOI: 10.1109/TAFFC.2023.3323043.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012.

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980.

[10] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into Deep Learning*. Cambridge University Press, 2023, https://D2L.ai.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015. DOI: https://doi.org/10.1038/nature14539.

[12] Veritas Health Innovation, *Covidence systematic review software*, Melbourne, Australia. [Online]. Available: www.covidence.org.

[13] M. S. Hossain and M. F. Rahman, "Detection of potential customers' empathy behavior towards customers' reviews," *Journal of retailing and consumer services*, vol. 65, p. 102 881, 2022. DOI: 10.1016/j.jretconser.2021.102881.

[14] S. Hinduja, "Mitigating the bias in empathy detection," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, IEEE, 2019, pp. 60–64. DOI: 10.1109/ACIIW.2019.8925035.

[15] D. Sedefoglu, A. C. Lahnala, J. Wagner, L. Flek, and S. Ohly, "LeadEmpathy: An expert annotated German dataset of empathy in written leadership communication," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italy: ELRA and ICCL, May 2024, pp. 10 237–10 248.

[16] M. R. Hasan, M. Z. Hossain, T. Gedeon, and S. Rahman, "LLM-GEm: Large language model-guided prediction of people's empathy levels towards newspaper article," in *Findings of the Association for Computational Linguistics: EACL 2024*, Y. Graham and M. Purver, Eds., St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 2215–2231. [Online]. Available: https://aclanthology.org/2024.findings-eacl.147.

[17] L. Tavabi, T. Tran, B. Borsari, *et al.*, "Therapist empathy assessment in motivational interviews," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2023, pp. 1–8. DOI: 10.1109/ACII59096.2023.10388176.

[18] T. Tran, Y. Yin, L. Tavabi, *et al.*, "Multimodal analysis and assessment of therapist empathy in motivational interviews," in *Proceedings of the 25th International Conference on Multimodal Interaction*, ser. ICMI '23, Paris, France: Association for Computing Machinery, 2023, pp. 406–415. DOI: 10.1145/3577190.3614105.

[19] Z. Zhu, C. Li, J. Pan, *et al.*, "MEDIC: A multimodal empathy dataset in counseling," in *Proceedings of the 31st ACM International Conference on Multimedia*, New York, NY, USA: Association for Computing Machinery, 2023, pp. 6054–6062. DOI: 10.1145/3581783.3612346.

[20] G. Lee and N. Parde, "AcnEmpathize: A dataset for understanding empathy in dermatology conversations," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 143–153. [Online]. Available: https://aclanthology.org/2024.lrec-main.13.

[21] P. Dey and R. Girju, "Investigating stylistic profiles for the task of empathy classification in medical narrative essays," in *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, C. Bonial and H. Tayyar Madabushi, Eds., Washington, D.C.: Association for Computational Linguistics, Mar. 2023, pp. 63–74. [Online]. Available: https://aclanthology.org/2023.cxgsnlp-1.8.

[22] A. Lee, J. K. Kummerfeld, L. An, and R. Mihalcea, "Empathy identification systems are not accurately accounting for context," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds., Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1686–1695. DOI: 10.18653/v1/2023.eacl-main.123.