Empathy Detection Using Machine Learning on Text, Audiovisual, Audio or Physiological Signals

Md Rakibul Hasan, *Member, ACL,* Md Zakir Hossain, *Member, IEEE,* Shreya Ghosh, Susannah Soon, and Tom Gedeon, *Senior Member, IEEE*

Abstract—Empathy is a social skill that 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. Empathy is a context-dependent term; thus, detecting or recognising empathy has potential applications in society, healthcare and education. Despite being a broad and overlapping topic, the avenue of empathy detection studies leveraging Machine Learning remains underexplored from a holistic literature perspective. To this end, we systematically collect and screen 801 papers from 10 well-known databases and analyse the selected 54 papers. We group the papers based on input modalities of empathy detection systems, i.e., text, audiovisual, audio and physiological signals. We examine modality-specific pre-processing and network architecture design protocols, popular dataset descriptions and availability details, and evaluation protocols. We further discuss the potential applications, deployment challenges and research gaps in the Affective Computing-based *empathy* domain, which can facilitate new avenues of exploration. We believe that our work is a stepping stone to developing a privacy-preserving and unbiased empathic system inclusive of culture, diversity and multilingualism that can be deployed in practice to enhance the overall well-being of human life.

Index Terms—Empathy, Deep Learning, Detection, Machine Learning, Recognition, Systematic Review

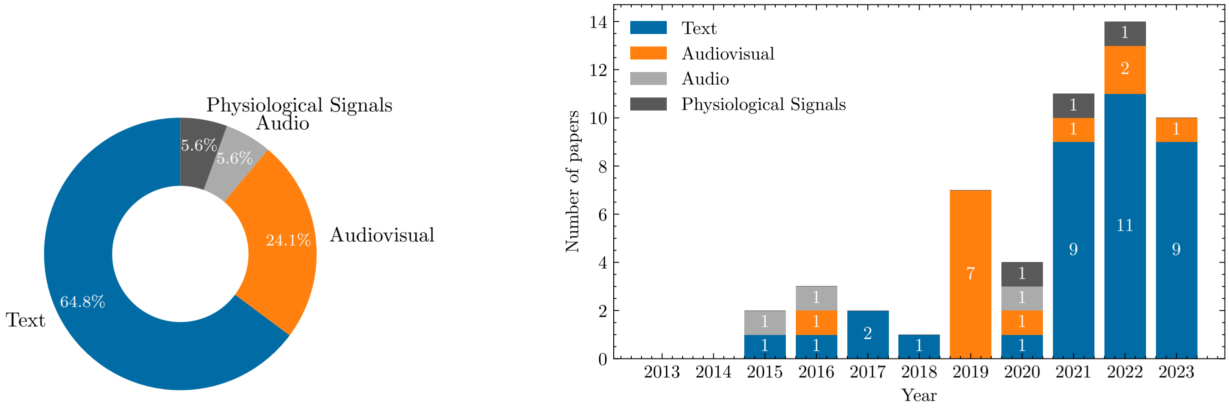
1 Introduction

Empathy can be defined as a multifaceted concept that involves perceiving, understanding and sharing emotional thoughts of others [1]. Understanding someone's thoughts and perspective is known as *cognitive empathy*, whereas sharing and experiencing the emotions of another person is known as *emotional empathy* [2]. Empathy is essential for effective communication in all aspects of human life, including social dynamics [3], healthcare [4] and education [5]. Research on empathy has been a major topic across a broad range of disciplines, including Social Science, Psychology, Neuroscience, Health and, most recently, Computer Science [6], [7]. With such broad usage, the definition of empathy sometimes varies. For example, empathy can also be defined as a multidimensional concept, such as four-dimensional empathy with perspective taking, fantasy, empathic concern and personal distress [8], and two-dimensional empathy with empathic concern and personal distress [9]. Despite varying definitions, all disciplines agree on its crucial role in human well-being [6]. This paper aims to review all works on empathy detection, and hence, papers are considered irrespective of the definition.

Machine Learning (ML) is a subdomain of artificial intelligence, which involves the development of algorithms to enable systems to learn from data. ML algorithms can be further classified into (1) classical ML, such as Decision Tree (DT) and Support Vector Machine (SVM), and (2) Deep Learning (DL), such as Multi Layer Perceptron (MLP) and Recurrent Neural Network (RNN). With the emergence of ML methodologies, the detection of emotional information has become a growing area of research in Affective Computing [10], [11]. To this end, emotion and facial expression recognition technologies have achieved maturity and widespread deployment, whereas empathy recognition lags in its development and practical implementation. Several reviews and surveys are available on various Affective Computing domains, such as facial affect recognition [12], [13] and emotion recognition [14], [15]. There are a few review papers [7], [16]–[18] available on empathy recognition, but all of these are specialised to specific use cases, such as artificial agent and social robot. Paiva *et al.* [7] and Yalcin and DiPaola [16] reviewed computational empathy in the context of artificial agents in 2017 and 2018, respectively. Park and Whang [17] systematically reviewed the empathy of various social robots in the human-robot interaction context in 2022. Published in the same year of 2022, Raamkumar and Yang [18] reviewed empathic conversational systems that primarily aim to generate empathic responses. Therefore, there has been a lack of holistic review on empathy, particularly in the context of *detecting* empathy using ML methodologies. A systematic literature review, to this end, facilitates methodically evaluating *all* published works against predefined criteria, thereby offering valuable insights into emerging trends, generating new research ideas, identifying gaps and shedding light on the existing body of work. We, therefore, present a systematic review of ML-based empathy detection in any human interaction.

Our method follows the standard practice of systematic literature review. We first devised search keywords and examined ten databases, including Scopus, Web of Science and IEEE Xplore. We screen the resulting 801 papers against five Exclusion Criteria (EC). Through rigorous title-and-

- *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, Susannah.Soon, 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.*



(a)

(b)

Fig. 1. (a) Statistics of the included papers according to data modalities and (b) growth of ML-based empathy detection literature from 2013 to 2023. In total, there are 35 text-based, 13 audiovisual-based, 3 audio-based and 3 physiological signal-based studies.

abstract and full-text screenings, we select 54 papers that are thoroughly reviewed in this paper.

Based on input modalities of empathy recognition, we group our analysis into four input modalities: text, audiovisual, audio and physiological signals. Figure 1 illustrates the number of papers from 2013 to 2023, separated into four categories. Text-based empathy detection comprises the major portion (n = 35), followed by audiovisual-based detection (n = 13). We find an equal number (n = 3) of audio- and physiological signal-based empathy detection studies. Surprisingly, no ML-based empathy detection works were reported in the years 2013 and 2014. Our major contributions include:

a. We provide a holistic review of all ML-based empathy recognition papers published from 2013 to 2023.
b. Based on our systematic review, we attempt to answer the following research questions:
   i. What are the datasets, how are they collected, and are the datasets publicly available?
   ii. What are the studies using the datasets, how do they analyse the data, and are the codes publicly available?
   iii. What ML methods are prominent with each input modality?
   iv. What are the opportunities of empathy detection systems, and where can they be applied?
   v. What are the major challenges in the computational empathy domain, such as in data collection and in building an ideal deployable system?

The paper is organised as follows. Section 2 describes the paper searching and screening process. Sections 3, 4, 5 and 6 present a comprehensive overview of the datasets and computational empathy studies using text, audiovisual, audio and physiological signals, respectively. In each category, we first summarise the datasets, followed by discussions on the studies involving the datasets. We report the details of the datasets, including their statistics, annotation protocol and their public availability, which refers to the availability of the whole annotated dataset used in corresponding studies. In the case of empathy detection works involving the datasets, we report the public availability of the software code, best-performing model and their performances. Section 7 provides the commonly used evaluation metrics in ML-based empathy detection studies. We discuss the opportunities, challenges and research gaps in Section 8 and conclude the paper in Section 9.

2 Paper Selection

To ensure reproducibility, we adhere to the PRISMA standard guidelines [19] when screening relevant papers for this systematic review. Our paper selection strategy is inclusive of the following 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. Not peer-reviewed
EC4. Published before 2013
EC5. Review, survey, meta-analysis, thesis or dissertation

Following the PRISMA standard, we report the paper search and screening results in the following subsections.

2.1 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, we searched ten databases (see Table 1 for more details) on 24 February 2023. Among the search engines, ACL Anthology does not support logical search. We, therefore, build 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 (EC4) and paper type (EC5). On those databases, we automatically filter out the search results. Table 1 presents the number of search results, search condition (e.g., title, abstract, full-paper, etc.),

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

TABLE 1
Search Results with Details in All 10 Databases.

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

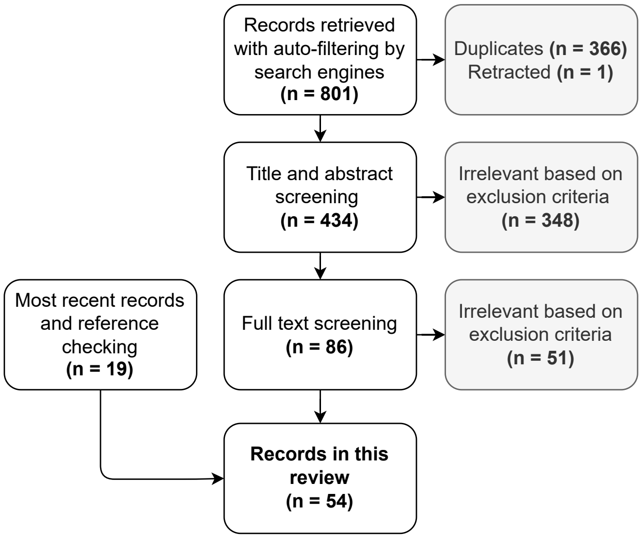

Records retrieved with auto-filtering by search engines (n = 801)
Duplicates (n = 366)
Retracted (n = 1)
Title and abstract screening (n = 434)
Irrelevant based on exclusion criteria (n = 348)
Most recent records and reference checking (n = 19)
Full text screening (n = 86)
Irrelevant based on exclusion criteria (n = 51)
Records in this review (n = 54)

Fig. 2. Number of papers at different stages in the screening process.

automatic filtering results and corresponding filtering criteria.

2.2 Paper Screening

Figure 2 illustrates step-by-step paper screening process. We have obtained 801 papers initially. After removing duplicates and retracted papers, we screen the remaining ones by reading titles and abstracts in Covidence systematic review management software [20]. In this stage, papers are excluded if and only if they clearly fall under any of the EC. We screen the remaining 86 papers by reading their full texts against the EC.

We screen another 19 recent papers, which we receive through notifications and reference checking. Several search engines, such as Scopus, Web of Science, IEEE Xplore, ACM and Google Scholar, offer email notification-based services on a predefined search string. In the case of reference checking, we have identified a few relevant papers by examining the reference lists of the papers we were reviewing. As the latest change, we have added eight recent papers [21]–[28] on 27 July 2023. These papers were recently published as part of the Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA) empathy detection shared task 2023 [21]. Finally, we have come up with 46 relevant papers that we examine in this systematic review. We categorise the analysis of the selected papers based on data modality: text, audiovisual, audio and physiological signals.

3 Empathy Detection from Text

In natural language processing research, empathy is detected from various textual contents, such as written essays, written conversations between patients and doctors or between customers and brands, etc.

3.1 Datasets

We identify 20 text-based empathy detection datasets. Eleven of them are publicly available. Table 4 presents the details of the datasets. We group the datasets based on their similarities, which are presented in the next few subsections.

3.1.1 People's Reaction towards Newspaper Articles

`NewsEmpathy` dataset consists of essays written by study participants who read news articles involving harm to individuals, groups or nature. In addition to the essays, the dataset consists of news articles and participants' demographic information. `NewsEmpathy` and its variants are annotated using Batson's empathy and distress scale [9] by the participants themselves. Batson's empathy and distress 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. The ground truth degree of empathy in these datasets is, therefore, range from 1 to 7.

Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA) has been organising empathy detection challenges since 2021. The aim of the WASSA 2021, 2022 and 2023 shared tasks is to create models that can detect the degree of empathy as a continuous value (regression task). WASSA 2021 and 2022 challenges use the same `NewsEmpathy v2` dataset [30], which extends from the `NewsEmpathy` dataset [29] by involving new participants in the data collection experiment.

In the 2023 challenge [21], WASSA released a new dataset similar to the `NewsEmpathy` dataset. The new `NewsEmpathy v3` dataset involves the top 100 negative news articles instead of all 418 articles and re-runs the data

TABLE 2
Text-Based Datasets and Their Details.

SL	Name	Data	Statistics	Output label[a]	Annotation	Public
People's Reaction towards Newspaper Articles						
1	NewsEmpathy [29]	300 – 800 character essay in response to 418 news articles	403 participants and 1,860 essays	$[1.0, 7.0]$ & {empathy, no-empathy}	Self	✓
2	NewsEmpathy v2 [30]	Extension of the NewsEmpathy dataset	564 participants and 2,655 essays	$[1.0, 7.0]$	Self	✓
3	NewsEmpathy v3 [21]	Similar to the NewsEmpathy dataset but based on 100 news articles and new conversation-level data	140 participants, 1,100 essays and 12,601 speech-turns	$[1.0, 7.0]$ (essay), $[0.0, 5.0]$ (speech-turn),	Self (essay), Third party (speech-turn)	✓
Social Media						
4	Brand-Customer [31]	Customer queries and brand response from Twitter	108 brands, 667,738 customers, and 2,013,577 tweets	{engaging, not-engaging}[b]	Third party	×
5	EPITOME [32]	Responses towards help-seeking posts in TalkLife and Reddit	8 million posts and 26 million interactions	{no, weak, strong}	Third party	✓
6	EPITOME v2 [33]	EPITOME, re-labelled into two classes	8 million posts and 26 million interactions	{positive, negative}	Third party	✓
7	PEC[c] [34], [35]	General conversations from Reddit	355K conversations	{empathic, non-empathic}	Third party	✓
8	Yelp Review [36]	People's reaction to customer reviews of financial providers	30,263 reviews	{negative, neutral, positive}	Self	×
9	Facebook Review [37]	Comments from 48 official public hospitals' Facebook pages	900 reviews	{tangibles, reliability, responsiveness, assurance, empathy}	Third party	×
10	Pathogenic Empathy [38]	Facebook posts and answers to a questionnaire	2,405 participants and 1,835,884 posts	$\mathbb{R}$	Self	×
11	TwittEmp [39]	Cancer and 200 high-rating empathy words-related tweets	3,000 tweets	{seeking-empathy, providing-empathy, none}	Third party	✓
12	iEmpathize [40]	Discussions from online cancer survivors network	5,007 sentences	{seeking-empathy, providing-empathy, none}	Third party	✓
13	CSN [41]	Discussions threads from online cancer survivor's network (lung and breast)	2,107 messages	{empathic, non-empathic}	Third party	×
Patient-Doctor Interaction						
14	MI [42]	Motivational interviews from six clinical studies between therapists and patients of drug or alcohol use	176 therapists and 348 sessions	{high, low}, $[1.0, 7.0]$	–	×
15	MI v2 [43]	Motivational interviews between therapists and patients of drug or alcohol use from six clinical studies	348 sessions	{high, low}	–	×
16	RolePlayMI[c] [34], [44]	Counselling conversations from online video sharing platforms	253 conversations	{empathic, non-empathic}	Third party	✓
17	MedicalCare [45]	Sentence-level annotation of essays on simulated patient-doctor interaction	774 essays	{empathic, non-empathic}	Third party	×
18	MedicalCare v2 [46]	Samples from the MedicalCare dataset re-annotated into four labels	440 essays	{cognitive, affective, prosocial, none}	Third party	×
General Conversations						
19	EmpatheticDialogues v1[c] [47]	Samples from a dialogue generation dataset [48], re-annotated into five labels	400 conversations	{not empathic, a little, somewhat, empathic, very much}	Third party	✓
20	EmpatheticDialogues v2[c] [34]	Collected from [48], the conversations between two people regarding a personal situation	810 participants and 24,850 conversations	{empathic, non-empathic}	Third party	✓

[a] Output labels in $[x, y]$ refer to continuous values between x and y
[a] $\mathbb{R}$ – real number, unspecified in the paper
[b] Empathy is classified into three categories as part of engagement estimation
[c] Was originally not an empathy detection dataset but repurposed in empathy detection

collection experiment. In addition to essay-level empathy, the new dataset involves conversation-level empathy annotation. In this case, participants converse with each other, and the speech turns from the conversation are annotated by independent annotators on a scale of 0 to 5.

3.1.2 Social Media

Several datasets consist of data from social media – primarily Twitter, Reddit and Facebook – which are annotated by trained annotators. For example, the `Brand-Customer` dataset [31] consists of Twitter threads about customer service-related queries and corresponding responses from brands. The authors [31] annotated the engagement between brands and customers into three categories – no, weak and strong empathy – in their primary goal of engagement estimation (engaging vs not engaging). Sharma *et al.* [32] proposed `EPITOME` framework, consisting 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: no, weak and strong, for each of the `EPITOME` mechanisms. `EPITOME` was relabelled by Hosseini and Caragea [33] into two classes: weak and strong communication as the positive class, and no communication as the negative class (`EPITOME v2`). `PEC` [34], [35] consists of general conversations from three subreddits (Reddit channels), which are annotated into empathic and non-empathic categories.

The `Yelp Review` dataset [36] consists of people's reactions to customer reviews related to financial providers. On the Yelp website (https://www.yelp.com/), potential customers' reactions to the existing customer reviews as either useful, cool or funny are considered empathy behaviour in this dataset. In `Facebook Review` dataset [37], comments from 48 official public hospitals' Facebook pages are annotated into four categories, including empathy. Abdul-Mageed *et al.* [38] 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 people who are empathic. The questionnaire used in the`Pathogenic Empathy` dataset consists of eight questions in total: three questions, on a scale of 1–7, are based on a previous study [49] and others, on a scale of 1–9, are prepared by the authors. The average of the responses is considered as the ground truth empathy score.

`TwittEmp` [39] dataset consists of cancer-related tweets, whereas `iEmpathize` [40] consists of discussion threads from online cancer survivor's network. `CSN` [41] is another dataset focusing only on lung and breast cancers from the online cancer survivor's network. `TwittEmp` and `iEmpathize` datasets are annotated into three categories (empathy-seeking, providing or none), whereas `CSN` are annotated into two categories (empathic versus non-empathic).

3.1.3 Patient-Doctor Interaction

There are several datasets from counselling sessions between therapists and patients. For example, motivational interviewing-related `MI` [42] and `MI v2` [43] involves interview sessions from clinical interviews with patients of drug or alcohol use. `RolePlayMI` dataset [34] consists of counselling conversations from video-sharing platforms, such as YouTube and Vimeo, which were originally collected in a separate study [44]. Wu *et al.* [34] later annotated this dataset into utterance-level empathic and non-empathic categories.

`MedicalCare` [45] involves narrative essays about simulated patient-doctor interactions written by pre-med students, which were annotated into empathic and non-empathic categories. Dey and Girju [46] selected 440 essays from the whole 774 `MedicalCare` essays and re-annotated them into four labels: cognitive empathy, affective empathy, prosocial behaviour and no empathy (`MedicalCare v2`).

3.1.4 General Conversation

`EmpatheticDialogues`, consisting of conversations between two people regarding a personal situation, was originally collected by Rashkin *et al.* [48] for empathic dialogue generation studies. The dataset is later re-annotated into five categories (not empathic, a little, somewhat, empathic, very much empathic) by Montiel-Vázquez *et al.* [47] (`EmpatheticDialogues v1`) and two categories (empathic, non-empathic) by Wu *et al.* [34] (`EmpatheticDialogues v2`).

3.2 Studies and Methods

Text-based datasets are predominantly employed with DL algorithms and, to a lesser extent, with classical ML lexical-based algorithms. Figure 3a 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 Natural Language Processing (NLP) tasks [50], 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, Bidirectional Encoder Representations from Transformers (BERT)-based Robustly Optimized BERT Pretraining Approach (RoBERTa) is mostly used, followed by BERT itself.

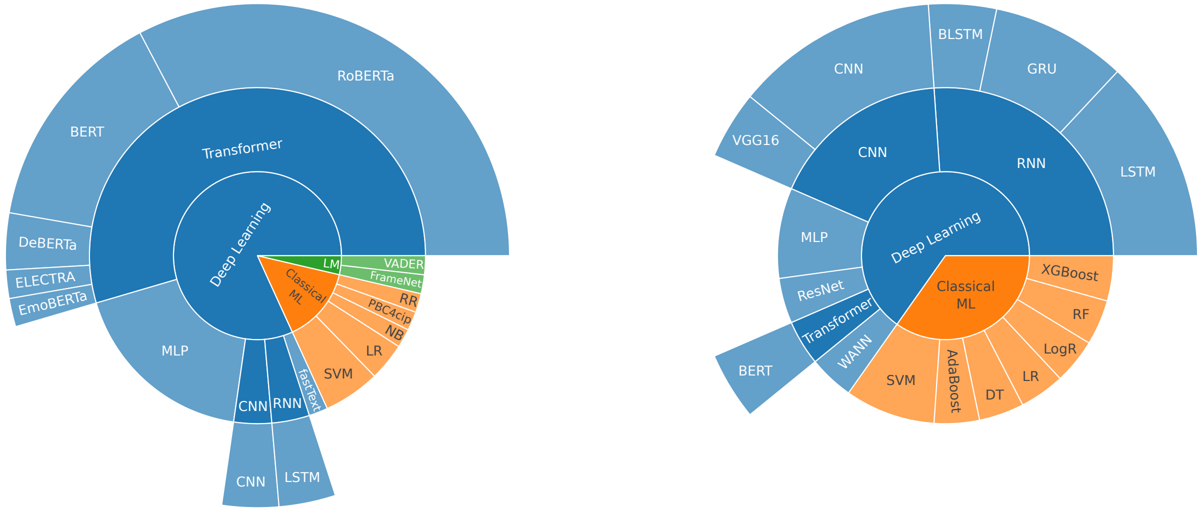
There are 21 studies where a continuous degree of empathy is detected (regression task) and 15 studies where a distinct level of empathy is detected (classification task). Table 3 summarises text-based empathy detection studies, and the following subsections discuss them, grouped into regression and classification tasks.

3.2.1 Regression Task (Degree of Empathy)

In detecting empathy as a continuous value, most works used `NewsEmpathy` and its variants. With `NewsEmpathy` dataset, Buechel *et al.* [29] leveraged fastText [51] for text embeddings, followed by a Convolutional Neural Network (CNN) regression model, achieving a Pearson correlation coefficient of 0.404. Mundra *et al.* [52] used an ensemble of ELECTRA and RoBERTa models and achieved a Pearson correlation coefficient of 0.558, outperforming all other works on `v2` dataset. The performance in detecting empathy in essays from the `v3` dataset is relatively lower than that observed in the `v1` dataset, which may be attributed to the smaller size of the `v3` dataset.

Interestingly, only one study [60] in `v2` dataset employed Linear Regression (LR) classical ML method and reported a Pearson correlation coefficient of 0.516. This performance is competitive to studies utilising DL-based language models such as BERT and RoBERTa, where the Pearson correlation coefficient ranged from 0.470 to 0.558 [52], [53]. Such an exceptional performance using classical ML [60] can be attributed to the incorporation of handcrafted features, including lexicon-based, n-gram, and demographic-based features. On the other hand, the highest performance by Mundra *et al.* [52] can be attributed to the ensemble ELECTRA and RoBERTa models.

Apart from `NewsEmpathy` datasets, other continuous degree of empathy detection works include therapists' empathy detection on `MI` dataset [42] and pathogenic empathy



(a) Text

(b) Audiovisual

Fig. 3. Usage of ML algorithms in (a) text-based and (b) audiovisual-based empathy detection studies. Deep learning-based models dominate in both cases, and transformer-based architectures are more utilised in text-based works.

detection on social media [38]. Both of them leveraged classical ML methods – LR and Ridge Regression (RR) – and reported Spearman's correlation coefficient of 0.6112 and Pearson correlation coefficient of 0.252, respectively.

3.2.2 Classification Task (Level of Empathy)

In the case of modelling empathy as a classification task, several authors used a diverse array of datasets and algorithms. Sharma *et al.* [32] used both unsupervised (domain adaptive pre-training) and supervised training on two datasets on their `EPITOME` framework, where they classified empathy into three categories using RoBERTa model. Hosseini and Caragea [33] used `EPITOME v2` as an in-domain dataset and `NewsEmpathy` as an out-of-domain dataset in the knowledge distillation strategy to transfer knowledge from a RoBERTa teacher model to a RoBERTa student model. In their study, the `NewsEmpathy` dataset was used in a binary classification setting as opposed to detecting the degree of empathy as a regression task. Binary classification in `NewsEmpathy` dataset is also utilised by Shi *et al.* [45] and Hosseini and Caragea [39] using SVM and BERT-MLP models, respectively. Shi *et al.* [45] used both `MedicalCare` dataset and `NewsEmpathy` dataset in binary classification setting: empathic versus non-empathic.

On detecting empathy in medical essays, Shi *et al.* [45] experimented with SVM and Naïve Bayes (NB) on `MedicalCare` dataset, yielding an F1 score of 78.4%. In `MedicalCare v2` dataset, Dey and Girju [46] 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%. Their experiment reveals the superior performance of BERT and RoBERTa as compared to other models. The higher performance of Dey and Girju [46] is also attributed to incorporating FrameNet pre-trained model [63].

Performance improvement from classical ML to DL models is also reported on empathy detection in counselling sessions. Gibson *et al.* [42] reported NB as the optimal model in `MI` dataset. In a later study, Gibson *et al.* [43] reported that a combination of MLP and LSTM are the optimal model in the closely related `MI v2` dataset, yielding a higher unweighted average recall from 75.28% to 79.6%.

Montiel-Vázquez *et al.* [47] experimented with several classical ML algorithms and reported Pattern-Based Classifier for Class Imbalance Problems (PBC4cip) – specially designed for imbalanced datasets – as the most effective classifier with an Area Under the receiver operating characteristics Curve (AUC) score of 62.48% on `EmpathicDialogues v1` dataset. Wu *et al.* [34] used `PEC` and `RolePlayMI` dataset, in addition to the `EmpathicDialogues v2` dataset and reported a maximum Matthews correlation coefficient of 0.85 using BERT model.

On detecting empathy in online discussion threads, Hosseini and Caragea [40], [62] detected empathy on cancer-related threads of `iEmpathize` dataset, leveraging BERT and RoBERTa models, respectively. Despite being the same dataset, Hosseini and Caragea [40], [62] reported classification performance using different evaluation metrics: a maximum F1 score of 85.88% in [40] and classification accuracy of 81.07% in [62]. Khanpour *et al.* [41] also detected empathy on cancer-related threads but using a combination of CNN and LSTM on another dataset (`CSN`).

Apart from cancer-related threads, Hossain and Rahman [36] and A. Rahim *et al.* [37] detected empathy on `Yelp Review` and `Facebook Review` datasets, respectively. Interestingly, Hossain and Rahman [36] used lexicon and rule-based Valence Aware Dictionary for Sentiment Reasoning (VADER) sentiment analyser, apart from widely-used classical ML or DL models. A. Rahim *et al.* [37] leveraged SVM model and reported a classification accuracy of 21.5% and an F1 score of 75.7%.

4 Empathy Detection from Audiovisual Data

The detection of empathy from audiovisual data utilises a combination of computer vision and natural language processing techniques.

TABLE 3
Summary of Empathy Detection Studies from Text Data.

Dataset	Study	Best Model	Performance[b]	Code Availability[a]
Regression Task (Degree of Empathy)				
NewsEmpathy	[29]	fastText-CNN	PCC: 0.404	✓
NewsEmpathy v2	[53]	RoBERTa-MLP	PCC: 0.470	U
	[54]	BERT-MLP	PCC: 0.479	×
	[55]	RoBERTa	PCC: 0.504	U
	[56]	RoBERTa	PCC: 0.524	✓
	[57]	RoBERTa	PCC: 0.537	×
	[58]	RoBERTa	PCC: 0.541	×
	[59]	BERT-MLP	PCC: 0.473	✓
	[52]	ELECTRA + RoBERTa	PCC: 0.558	✓
	[60]	LR	PCC: 0.516	✓
	[61]	RoBERTa-MLP	PCC: 0.517	✓
NewsEmpathy v3	[21]	RoBERTa	PCC: 0.536 (essay), 0.660 (speech-turn)	×
	[22]	DeBERTa, RoBERTa	PCC: 0.331 (essay), 0.674 (speech-turn)	×
	[23]	BERT	PCC: 0.187 (essay), 0.573 (speech-turn)	✓
	[24]	RoBERTa-MLP	PCC: 0.270 (essay), 0.665 (speech-turn)	×
	[25]	{RoBERTa, EmoBERTa}-MLP	PCC: 0.415 (essay), 0.669 (speech-turn)	×
	[26]	RoBERTa	PCC: 0.348 (essay), 0.652 (speech-turn)	✓
	[27]	RoBERTa-SVM	PCC: 0.358 (essay)	×
	[28]	DeBERTa-MLP, RoBERTa-MLP	PCC: 0.329 (essay), 0.708 (speech-turn)	×
MI	[42]	LR	Spearman's correlation coefficient: 0.6112	×
Pathogenic Empathy	[38]	RR	PCC: 0.252	×
Classification Task (Level of Empathy)				
EPITOME	[32]	RoBERTa	Accuracy $\in [79.93\%, 87.50\%]$, F1 $\in [67.46\%, 74.29\%]$ (TalkLife); Accuracy $\in [79.43\%, 92.61\%]$, F1 $\in [62.60\%, 74.46\%]$ (Reddit)	✓
EPITOME v2 & NewsEmpathy	[33]	BERT, RoBERTa	Accuracy $\in [61.47\%, 71.80\%]$	×
NewsEmpathy, TwittEmp	[39]	BERT-MLP	F1: 68.41% (NewsEmpathy), F1: $\in [68.57\%, 85.71\%]$ (TwittEmp)	×
Brand-Customer	[31]	RoBERTa	F1: 73%	×
MedicalCare + NewsEmpathy	[45]	SVM	Accuracy: 89.4%, F1: 78.4%	×
MedicalCare v2	[46]	FrameNet-BERT	F1 $\in [75\%, 85\%]$	×
MI	[42]	NB	Unweighted average recall: 75.28%	×
MI v2	[43]	MLP-LSTM	Unweighted average recall: 79.6%	×
EmpatheticDialogues v1	[47]	PBC4cip	AUC: 62.48%	×
PEC, EmpatheticDialogues v2, RolePlayMI	[34]	BERT	Matthews correlation coefficient $\in [\approx 0.56, \approx 0.95]$	×
iEmpathize	[40]	BERT	F1: $\in [78.94\%, 85.88\%]$	×
	[62]	RoBERTa	Accuracy: 81.07%	×
CSN	[41]	CNN-LSTM	F1: 78.36%	×
Yelp Review	[36]	VADER	VADER score $\in [-0.2017, 0.8252]$	×
Facebook Review	[37]	SVM	Accuracy: 21.5%, F1: 75.7%	×

[a] U – Unofficially available on the Internet but not provided with the paper
[b] PCC – Pearson correlation coefficient
[b] Range of performance is reported when overall classification performance is unavailable

4.1 Datasets

Table 4 presents all six audiovisual empathy detection datasets and their details. The subsequent subsections discuss the datasets, grouping them based on their similarities.

4.1.1 General Conversation

The `OMG-Empathy` consists of audiovisual conversations with semi-scripted stories in a speaker-listener set-up. There were eight stories, four speakers and ten listeners in total. Following the conversations, the listener rated their valence

TABLE 4
Audiovisual Datasets and Their Details.

SL	Name	Data	Statistics	Output label[a]	Annotation	Public
General Conversations						
1	OMG-Empathy [64]	Speaker-listener conversations based on eight semi-scripted stories	4 speakers, 10 listeners and 80 audiovisual data (total 480 minutes)	$[-1, +1]$	Self	✓
Teacher-Student Interaction						
2	Teacher-Student [65]	Online lectures (one teacher and 5-10 adult students)	10 teachers and 338 audiovisual data (63 lectures)	$[0.0, 10.0]$, {excellent, good} lectures	Third party	×
Interaction with Non-Human Entity						
3	Human-Robot [66]	Human participants listen to six scripted stories from a robot	46 participants and 6.9 hours audiovisual data	{empathic, less-empathic}	Self	✓
4	Human-Avatar [67]	Interaction between avatar and normotypical, Down syndrome and intellectual disability users	50 participants and 24,000 interactions	{empathic, non-empathic}	Other[b]	×
5	DAIC-WOZ [68], [69]	Semi-structured interviews with virtual agent	186 participants and 2,185 conversations	{negative, positive, no} empathy	Third party	×
6	Human-Virtual Agent [70]	Human participants watched a sad virtual character in virtual reality	28 participants and 56 surveys	$[0, 20]$	Self	×

[a] Output labels in $[x, y]$ refer to continuous values between x and y

[b] Normotypical users' data are annotated as empathic; Down syndrome and intellectual disability users' data as non-empathic)

score on a scale from -1 to $+1$. Using the `OMG-Empathy` dataset, an empathy detection challenge[2] was organised as part of the IEEE International Conference on Automatic Face & Gesture Recognition in 2019. There are two types of empathy detection protocol in this dataset: personalised protocol, detecting empathy of each listener across all conversations and generalised protocol, detecting empathy towards each story by all listeners.

4.1.2 Teacher-Student Interaction

`Teacher-Student` dataset consists of online audiovisual lectures in one-to-many teaching set-up [65]. The lectures are annotated by third-party annotators in terms of teaching quality – as binary levels (classification) and as a scale of 0 to 10 (regression) – in five domains, including empathy.

4.1.3 Interaction with Non-Human Entity

Several audiovisual empathy detection dataset involves human and non-human entity. `Human-Robot` data collection involves a robot telling scripted stories to human participants [66]. Stories were told in either first-person or third-person point-of-view. As the ground truth empathy score, participants answered a questionnaire that assessed their level of elicited empathy towards the robot's story.

`Human-Avatar` dataset involves empathy of human participants interacting with an avatar expressing six types of emotion [67]. The ground truth is labelled as empathic for normotypical participants and non-empathic for participants having social communication disorders such as Down syndrome and intellectual disability.

`DAIC-WOZ` dataset involves interviews between human participants and a virtual agent, where the conversations are annotated by third-party annotators into three classes (negative, positive and no empathy) [68], [69].

Finally, in `Human-Virtual Agent` dataset, human participants watched a virtual character showing sadness in a virtual reality environment [70]. Participants fill in the Toronto empathy questionnaire [71] and another post-experiment questionnaire to reflect how much empathy they feel towards the agents. The responses to the questionnaires are leveraged as self-assessed ground truth empathy scores on a scale of 0 to 20.

4.2 Studies and Methods

Empathy detection studies from audiovisual datasets are designed mostly as a multimodal system, having inputs such as facial expressions, hand gestures and audio conversations. Figure 3b 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 RNN-based models are most frequently used, whereas in the classical ML category, SVM enjoys a higher level of usage.

Table 5 summarises these studies involving empathy detection from audiovisual datasets. There are nine studies of detecting continuous degrees of empathy (regression task) and five studies of detecting discrete levels of empathy (classification task). The following subsections describe these studies.

4.2.1 Regression Task (Degree of Empathy)

In predicting continuous degrees of empathy, seven out of 13 papers use the `OMG-Empathy` dataset. Barros *et al.* [64] provides baseline results with 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 and 0.23 correlation coefficients in

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

TABLE 5
Summary of Empathy Detection Studies from Audiovisual Data.

Dataset	Study	Best Model	Performance[a]	Code Availability
Regression Task (Degree of Empathy)				
OMG-Empathy	[64]	VGG16-LSTM-SVM	CCC 0.17 (P), 0.23 (G)	×
	[72]	GRU, LSTM, CNN, MLP	CCC: 0.17 (P, G)	✓
	[73]	LSTM	CCC: 0.14 (P, G)	✓
	[74]	BiLSTM	CCC: 0.11 (P), 0.06 (G)	✓
	[75]	SVM	CCC: 0.08 (P)	×
	[76]	CNN, RF	CCC: 0.02 (P), 0.04 (G)	×
	[77]	WANN	CCC: 0.25 (Validation set)	×
Human-Virtual Agent	[70]	LR	R^2: 0.485	×
Teacher-Student	[65]	AdaBoost	MSE: 0.374	✓
Classification Task (Level of Empathy)				
Teacher-Student	[65]	DT	Accuracy: 90.9%, F1: 90.1%	✓
Human-Robot	[66]	XGBoost	Accuracy: 69%, AUC: 72%	×
Human-Avatar	[67]	LogR	F1: $\in$ [72%, 78%]	×
DAIC-WOZ	[68]	ResNet, BERT, GRU, MLP	F1: 71%	×
Various online sources[b]	[78]	CNN	Accuracy: 98.98%, AUC: 99%, F1: 91%	×

[a] Performance refers to test-set performance unless otherwise stated
[a] P – Personalised protocol; G – Generalised protocol
[a] CCC – Concordance Correlation Coefficient
[b] Description of the dataset, such as the number of samples and ground truth label space, are unavailable on the paper

the personalised and generalised empathy protocols, respectively.

While Barros *et al.* [64] provides the baseline results on `OMG-Empathy` dataset, other works [72]–[76] have failed to outperform the baseline results. The closest one [72] achieved a 0.17 correlation coefficient on both personalised and generalised protocols. They used separate models for separate modalities – 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 detections on 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.* [73] extracted multimodal features using VGG-Face [79] on faces, openSMILE [80] on audio and GloVe embedding [81] 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.* [74] reported correlation coefficients of 0.11 and 0.06 in personalised and generalised protocols, respectively, using openSMILE for extracting audio features and OpenFace [82] 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.* [75] 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.* [76] 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.* [77] 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-Virtual Agent` dataset and reported an R^2 score of 0.485. With the `Teacher-Student` dataset, Pan *et al.* [65] comprehensively experimented with a wide range of features from audio and video in an AdaBoost model for the regression task and reported a mean squared error score of 0.374.

4.2.2 Classification Task (Level of Empathy)

In classifying empathy levels, most studies leveraged a variety of classical ML algorithms. With the `Teacher-Student` dataset, Pan *et al.* [65] modelled a classification task (excellent vs good lectures) and reported an accuracy of 90.9% and F1 score of 90.1% using DT model. Mathur *et al.* [66] experimented with eight classical ML and two DL models and reported XGBoost as the best model on the `Human-Robot` dataset. Hervás *et al.* [67] leveraged LogR on the `Human-Avatar` dataset.

On the `DAIC-WOZ` dataset, Tavabi *et al.* [68] leveraged pre-trained BERT to calculate text embedding and pre-trained 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) [83] 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.

Lastly, Alanazi *et al.* [78] experimented with CNN and MLP and several classical ML algorithms, including DT, NB and SVM. Among these, CNN was the most effective model in their experimental set-up.

5 Empathy Detection from Audio

Audio-based empathy detection works include audio from conversations in various contexts, such as patient-doctor and customer-call centres.

5.1 Datasets

Table 6 reports all three datasets consisting exclusively of audio data. All of these datasets involve third-party annotations, and none of the datasets are publicly available. We are referring to audio-based datasets that exclusively consist

TABLE 6
Audio-Based Datasets and Their Details.

SL	Name	Data	Statistics	Output label[a]	Annotation	Public
General Conversations						
1	Call-centre [84]	Human-human conversation in call-centre	905 conversations	{empathic, non-empathic}	Third party	×
Patient-Doctor Interaction						
2	COPE [85], [86]	Conversations between cancer patient and healthcare provider(s)	425 sessions	{positive, negative}	Third party	×
3	CTT [87], [88]	Motivational interviewing sessions of drug and alcohol counselling	200 sessions	[1, 7], {low, high}	Third party	×

[a] Output labels in $[x, y]$ refer to continuous values between x and y

TABLE 7
Summary of Empathy Detection Studies from Audio.

Dataset	Study	Best Model	Performance	Code Availability
Regression Task (Degree of Empathy)				
CTT	[87]	LR	PCC: $\in [0.65, 0.71]$	×
Classification Task (Level of Empathy)				
COPE	[85]	SVM	Avg. Precision: 7.61%	×
Call-centre	[84]	SVM	Unweighted avg. recall: 65.1%	×
CTT	[87]	SVM	Accuracy $\in [80.5\%, 89.9\%]$, F1 $\in [85.3\%, 90.3\%]$	×

of audio, which are different from audiovisual datasets discussed in Section 4.

The Call-centre dataset [84] consists of 905 human-to-human conversations in call-centres annotated into empathic and non-empathic categories. The COPE dataset [85], [86] consists of 425 oncology encounters between cancer patients and healthcare providers. The task of this dataset is to detect empathic interactions – annotated as positive and negative – in oncology encounters. Lastly, the CTT dataset [87], [88] includes 200 sessions between therapists and patients of drug and alcohol abuse. The annotation includes both 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.

5.2 Studies and Methods

Detecting empathy from audio involves two main approaches: utilising audio as a signal directly or converting it into text and employing text-based methods. Table 7 summarises four studies and their methods for detecting empathy from audio datasets. None of the codes are officially available with the papers. All studies reported classical ML algorithms as the best in corresponding experiments: SVM in all three classification studies and LR in the one regression study. The CTT dataset has been used both in regression and classification studies [87].

Given that audio-based datasets involve conversations between two persons, the works of Chen *et al.* [85] and Xiao *et al.* [87] include voice activity detection (speech or no speech) and speaker diarisation (separate speakers) in the empathy detection workflow. All studies [84], [85], [87] converted the audio into text sequences, followed by extracting features from the text sequences. Chen *et al.* [85] and Alam *et al.* [84] 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.* [85] reported better performance of lexical features as compared to the acoustic features. Lastly, Xiao *et al.* [87]'s empathy detection model on audio-based CTT dataset is entirely text-based – leveraging uni-gram, bi-gram and tri-gram language models – without any audio-based features.

6 Empathy Detection from Physiological Signals

Physiological signals, such as galvanic skin response, have been shown to be indicative of emotional responses of individuals [92]. Empathy from physiological signals includes fMRI, ECG and EEG data.

6.1 Datasets

Table 8 summarises all three physiological signal-based datasets for empathy detection. One of them is publicly available. As for the annotation protocol, all of them include self-annotation by study participants.

The fMRI dataset [89] includes resting-state functional magnetic resonance imaging data from 24 cocaine-addicted subjects and 24 healthy controls matched on age, sex, employment and education information. The annotations were collected from cocaine-addicted subjects through the Interpersonal Reactivity Index (IRI) questionnaire [8], [93].

The PainEmpathy dataset [90] includes ECG and skin conductance from 36 participants with different levels of autistic traits. The participants filled in a questionnaire regarding cognitive and affective empathy after viewing pictures of individuals with different levels of pain. 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. The task with this dataset is the classification of both cognitive and affective empathy into high or low levels.

The EEG Cortical Asymmetry dataset [91] includes EEG signals from 52 participants watching an emotional

TABLE 8
Physiological Signal-Based Datasets and Their Details.

SL	Name	Data	Statistics	Output label[a]	Annotation	Public
1	fMRI [89]	Resting-state fMRI data from cocaine-dependent and healthy controls	48 participants	Fantasy empathy $\in \mathbb{R}$	Self	✓
2	PainEmpathy [90]	Participants viewing pictures of individuals with pain or no pain	36 participants, and 36 ECG and skin conductance data	{high, low}	Self	×
3	EEG Cortical Asymmetry [91]	Participants' EEG while watching an emotional video in virtual reality	52 participants and 52 EEG data	$[0, 96]$, {high, low}	Self	×

[a] $\mathbb{R}$ – real number, unspecified in the paper

TABLE 9
Summary of Empathy Detection Studies from Physiological Signals.

Dataset	Study	Best Model	Performance	Code Availability
Regression Task (Degree of Empathy)				
fMRI	[89]	LR	Pearson correlation: 0.54, MSE: 20.09	×
EEG Cortical Asymmetry	[91]	LR	MSE $\in$ 51.749, 150.556	×
Classification Task (Level of Empathy)				
PainEmpathy	[90]	SVM	Accuracy $\in [79\%, 84\%]$	×
EEG Cortical Asymmetry	[91]	SVM, DT	Accuracy $\in [61.8\%, 74.2\%]$, F1 $\in [61.5\%, 74.3\%]$	×

video (a young girl being abused as a domestic slave) 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 [71], which was utilised as self-annotation. Although the range of annotation could be 0 to 96 according to the questionnaire, the participant's actual responses varied from 49 to 86. Using a median split, annotation into high and low empathy groups is also available on this dataset. Both regression and classification tasks in this dataset involve empathy detection at all three time points the EEG were collected: before, during and after.

6.2 Studies and Methods

Research in physiological signal-based empathy detection typically follows feature extraction, followed by ML techniques. Table 9 reports the studies and methods of physiological signal-based empathy detection. None of the papers' code is publicly available with the paper. All of the physiological signal-based studies leveraged classical ML algorithms: LR and SVM in two studies each.

Wei *et al.* [89] detected fantasy empathy on the `fMRI` dataset using a LR model. With the `PainEmpathy` dataset, Golbabaei *et al.* [90] extracted ten features and leveraged a SVM with radial basis function kernel to detect cognitive and affective empathy. Lastly, Kuijt and Alimardani [91] extracted 15 features from the `EEG Cortical Asymmetry` data and leveraged multiple LR in the regression task and LR, SVM and DT in the classification task. In the case of 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 compared to 'after' and 'during' the experiment.

7 Evaluation Metrics

7.1 Regression Task (Degree of Empathy)

Correlation coefficients have been used as the evaluation metric in works where a continuous value of empathy is detected, i.e., regression task. Notably, the Pearson correlation, Spearman's correlation and concordance correlation coefficients are leveraged in works with `NewsEmpathy`, `MI` and `OMG-Empathy` datasets, respectively. Apart from that, mean squared error and R^2 are used in `Teacher-Student` and `Human-Virtual Agent` datasets, respectively.

7.2 Classification Task (Level of Empathy)

A wide variety of evaluation metrics are used in works where a discrete level of empathy is detected, i.e., classification task. Classification accuracy, F1 score, and Area Under the receiver operating characteristics Curve (AUC) are most commonly used. Apart from these, average precision, unweighted average recall and Matthews correlation coefficient are also utilised. Despite being named as a correlation coefficient, Matthews correlation coefficient is used to evaluate binary classification performance [94].

8 Opportunities, Challenges and Research Gaps

8.1 More Research on Empathy from Audiovisual, Audio and Physiological Signals

As illustrated earlier (Figure 1), there has been a rising trend in text-based works since 2020, but the number of papers from audiovisual datasets has not increased that much. The surge of audiovisual work in 2019 resulted from an empathy detection challenge at the end of 2018.

Empathy detection from physiological signals, such as EEG, ECG and skin conductance, is emerging, but no studies have been reported on audio-based empathy detection after 2020. Therefore, the research gap is evident in audiovisual-, audio- and physiological signal-based empathy detection.

It is important to note that spoken information from video and audio can be converted to text, and subsequently, a text-based empathy detection system may work for video and audio scenarios. However, video and audio have additional information, such as facial expressions and audio pitch, which would normally enhance the quality of empathy detection and thus necessitate separate research for video and audio.

8.2 Applications of Empathy Detection

Empathy holds significant importance across various real-life domains, including social life, healthcare, education and business [6]. The assessment of empathy in empathy-seeking scenarios allows us to identify the areas of improvement. Consequently, new strategies can be developed to improve empathic capabilities.

8.2.1 Society and Culture

Since empathy is a social skill, detecting it has a direct impact on society. The ability to empathise can vary across cultures and may be influenced by social norms and upbringing. Existing studies so far have not considered cultural aspects, which could be an exciting research direction.

Socially assistive robots could provide better support and care if they could detect and respond empathically to the emotional states of people, such as elderly individuals, stroke survivors, and patients with autism spectrum disorder or Alzheimer's disease [95], [96]. To this end, empathy detection between humans and other intelligent agents, such as robot [66], [78] and virtual agent [68], [70], could help assess the quality of the supports.

Empathy plays a key role in effective and supportive communication among people at different levels. In spousal relationships, empathy can build strong and healthy relationships, as partners who display empathy can understand and support each other's emotional needs [3]. In nonverbal communication with people with hearing disabilities, empathy would help them understand and communicate effectively. Apart from these, empathy is important for our leaders, such as politicians, community leaders and religious leaders. Empathy can help people reduce their anxiety and stress, such as in long-distance audiovisual communications of international students communicating with their families and online interviews for jobs or admission to universities.

People often seek mental support through social media platforms. To this end, several works have detected empathy in various social media, such as Reddit [32], Twitter [39] and cancer survivors network [40], [41], [62]. The use of empathy detection systems in social media platforms has the potential to foster empathic responses while discouraging non-empathic ones. Such a system can play a key role in cultivating a more compassionate online environment, thereby contributing to improved community well-being.

8.2.2 Healthcare

Empathic doctors are better equipped to understand their patients' concerns, leading to improved communication and patient outcomes [4]. A study on patient-doctor relationships showed that 85% of 563 patients changed their doctor or were thinking of changing, where one of the main reasons was a lack of effective communication related to empathy [97], [98].

Empathy detection systems can help diagnose diseases and cognitive disorders where a lack of empathy is a symptom, such as autism, psychopathy and alexithymia [99]. Several studies have already shown proof of concepts to this end. Hervás *et al.* [67] proposed to use an empathy detection system with affective avatars in diagnosing social communication disorders, such as Down syndrome and intellectual disability. They also mentioned that such a system could be used in diagnosing autism spectrum disorders. Golbabaei *et al.* [90] used physiological signals (EEG and skin conductance) and detected affective and cognitive empathy. The authors argued that such empathy detection can be correlated with different levels of autistic traits of the subjects.

The service quality of healthcare providers and hospitals can be assessed in terms of empathy, for example, patient-doctor consultation or hospital service quality [37]. Assessment of healthcare providers can be in various contexts, such as counselling sessions between therapists and patients [34], [42], [43], [87], oncology encounters [85] and other general patient-doctor interactions [45], [46]. In addition, telehealth has become popular since COVID-19, where empathy evaluation can be particularly useful because of its distance nature.

8.2.3 Education and Development

In teaching – especially with the shift towards online learning due to the COVID-19 pandemic – educators endowed with empathic capabilities are better positioned to understand their students' emotional states and create a positive learning environment [5]. Not only in teacher-student interactions but also in student-student interactions in team activities, empathy helps extract the most out of the learning experience when students can extend support to their peers. Among different disciplines, engineering students can lack empathy, which can lead to challenges when engaging in group projects later in their professional careers [100]. Quantitative assessment of empathy can create scopes to improve team dynamics through targeted interventions, such as empathy training programs.

In terms of assessing teaching quality, empathy evaluation can be used as a tool, as demonstrated by Pan *et al.* [65]. Such a teaching quality assessment system – acknowledging the importance of emotional intelligence and interpersonal skills – can aid the traditional evaluation method, leading to a better education system.

Empathy is an important aspect in design thinking [101] and user-centred design [102] to connect with users, clients and customers. In software development, empathy for end-users can help create user-friendly and accessible applications. Accordingly, understanding users' needs and experiences leads to more successful and widely adopted products. Inclusion of empathy is required in software engineering curricula to meet industry demands [103] and, to this end, an empathy detection system can assist in teaching empathy.

8.2.4 Economics and Business

Empathy plays a crucial role in economics, where understanding and accounting for others' emotional states is essential in making informed business decisions [104]. A lack of empathy in real-life business interactions can have detrimental effects on customer experience and overall business success. By incorporating empathy into business strategies and decision-making, businesses can provide better services, increase customer satisfaction, and ultimately drive growth and success.

Empathy can help businesses in analysing customer reviews and customer support. Nowadays, customers usually leave their product reviews on online platforms, such as Yelp (https://www.yelp.com/) and Product Review (https://www.productreview.com.au/), where empathy can be detected to understand customer satisfaction and identify areas of improvement [36].

Customer care representatives who display empathy in call-centre interactions can resolve customer issues more effectively, leading to higher levels of customer satisfaction [105]. The skilful identification and validation of customers' emotions through empathy can foster loyalty and trust [106] and enhance customer experiences [107]. To this end, call-centre conversations can be analysed to detect empathy, which could also benefit in training the agents [84]. Empathy can be detected as a measure of engagement in asynchronous customer service systems, where customers and agents are not necessarily active simultaneously [31].

Empathy is important in emotional intelligence, a crucial aspect of individuals' aptitude in business and workplace settings [108]. It can enhance organisational efficacy through constructive interpersonal relationships in employer-employee and employee-employee interactions. Empathetic interactions can, therefore, promote overall employee well-being, job satisfaction [109] and cohesive team dynamics [110] in contemporary business environments. It can further play a climactic role in high-level negotiations in management as it fosters a deeper understanding of stakeholders' perspectives, leading to more effective decision-making and conflict resolution. In such negotiations, where complex issues and diverse interests are at play, empathic leaders can bridge gaps and build trust among parties, ultimately enhancing collaboration and achieving mutually beneficial outcomes.

8.3 Data Collection

8.3.1 Self-Annotation vs Third-Party Annotation

Self-annotation in the literature includes study participants filling in empathy-related questionnaires such as the IRI questionnaire [8], [93] and Toronto empathy questionnaire [71]. Whereas third-party annotation includes annotation by third-party trained annotators apart from the study participants from whom the data is collected. Self-annotation versus third-party annotators remains debatable in the literature. Among the datasets we examine in this paper, 10 of them used self-annotation, 18 of them used third-party annotation and one of them used both self and third-party annotations on its two different tasks. Buechel *et al.* [29] argued that self-annotation provides a more appropriate measure of empathy compared to annotation by third-party annotators. Shi *et al.* [45] used both `MedicalCare` dataset, annotated by trained third-party annotators, and `NewsEmpathy` dataset, annotated by study participants themselves. One interesting finding of their study is the comparison between third-party annotation and self-annotation, which showed that third-party annotation could be more robust.

One particular dataset, where both self and third-party annotations are used in the `NewsEmpathy v3` dataset. In this dataset, detecting empathy on speech turns has higher performance than on essays (reported in Table 3). One difference between essay-level and speech-turn-level detection is the annotation protocol: self-assessment annotation in essays, whereas third-party annotation in speech-turns. Self-annotation versus 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.

8.3.2 Crowdsourcing: Pros and Cons

Data collection through crowdsourcing platforms, such as Amazon Mechanical Turk, is known to be a faster and easier way to collect large amounts of data in Affective Computing domains, such as empathy [30] and emotion [111]. However, crowdsourcing involves a major challenge: false information [112]. Montiel-Vázquez *et al.* [47], in their empathy detection work, doubted the validity of crowdsourcing annotation and re-annotated the `EmpathicDialogues` dataset. Erroneous data, which often result from crowd participants' multitasking and carelessness, threaten the validity of findings [113], [114]. Therefore, strict measures and quality control should be carried out with crowdsource data collection.

8.3.3 Procedure and Equipment

The quality of data also depends on the collection procedure and equipment, such as the camera and microphone [87]. Occlusion-free high-resolution images and high-quality audio are some of the most desirable qualities. However, these are challenging, particularly in computational empathy, because of the simultaneous presence of multiple persons in the data collection experiment. For example, separate channels for separate persons in audio data collection can potentially benefit the detection but can be challenging [87]. In a similar vein, an exciting research avenue could be building a reliable empathy detection system from noisy data.

8.4 Towards an Ideal Empathy Detection System

8.4.1 Generalised System

The use of data from a broad context can sometimes be challenging. For example, Hossain and Rahman [36] analysed a subset from `Yelp Review` dataset and mentioned they were unable to analyse the whole dataset due to computing limitations. At times, data from a few specific sources are used. For example, A. Rahim *et al.* [37] analysed reviews from Facebook and left other social media platforms, such as Twitter and Instagram, for future work. To this end, future research can explore using data from a broad range of sources to build more generalised empathy detection systems.

8.4.2 Multilingual System

An ideal empathy detection system should be multilingual by including at least some widely used languages for wider adoption [36]. All existing studies detected empathy in the English language only. Therefore, there is a huge future research avenue towards making a multilingual empathy detection system.

8.4.3 Privacy and Bias

Given that empathy detection systems primarily detect empathy from human data such as facial expression, voice and physiological signals, maintaining privacy and mitigating biases is a major concern. Research design should consider concealing personal information and minimal use of personal information. The use of personal and demographic information may lead to a biased system, especially with existing biases of pre-trained models [115]. Deepfakes through generative artificial intelligence can be an exciting avenue to maintain privacy by generating fake images while maintaining the same facial expression of original subjects [116].

9 Conclusion

Empathy, defined as the capacity to comprehend and provide emotional support to others, has emerged as a promising research area across several disciplines. In Computer Science, empathy detection, particularly through ML methodologies, has witnessed substantial growth in recent years. In the pursuit of this research endeavour, we conducted an extensive search across ten scholarly databases, implementing a rigorous systematic review process based on PRISMA guidelines for reproducibility. Subsequently, we examine 54 selected papers, focusing on four primary input modalities: text, audiovisual data, audio and physiological signals. In each modality, we enumerate the details of the datasets, including data collection experiment detail, their statistics, annotation protocol and their public availability. We analyse the studies using these datasets, with a focus on ML algorithms, their performance and code availability.

Overall, this review reveals several new insights into the computational empathy domain. Firstly, there is an increasing amount of research in text-based empathy detection, but the amount of research on audiovisual, audio and physiological signals is lagging behind. Secondly, research and developments of empathy detection systems can be focused on various domains of our lives, such as society, healthcare, education and business. Thirdly, challenges and opportunities in data collection include annotation protocol (self vs third-party) and crowdsourcing. Finally, an ideal empathy detection system could be envisioned by building a generalised and multilingual system while maintaining privacy and mitigating biases.

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
DL Deep Learning
DT Decision Tree
EC Exclusion Criteria
GRU Gated Recurrent Unit
IRI Interpersonal Reactivity Index
LR Linear Regression
LogR Logistic Regression
LSTM Long Short-Term Memory
MFCC Mel-Frequency Cepstral Coefficients
ML Machine Learning
MLP Multi Layer Perceptron
NB Naïve Bayes
NLP Natural Language Processing
PBC4cip Pattern-Based Classifier for Class Imbalance Problems
ResNet Residual Network
RNN Recurrent Neural Network
RF Random Forest
RoBERTa Robustly Optimized BERT Pretraining Approach
RR Ridge Regression
SVM Support Vector Machine
VADER Valence Aware Dictionary for Sentiment Reasoning
WANN Weightless Artificial Neural Network
WASSA Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis

References

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

[2] A. Smith, "Cognitive empathy and emotional empathy in human behavior and evolution," *The Psychological Record*, vol. 56, no. 1, pp. 3–21, 2006. DOI: 10.1007/BF03395534.

[3] 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. DOI: 10.1371/journal.pone.0149944.

[4] 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. DOI: 10.1159/000342998.

[5] 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. DOI: 10.1007/s10648-021-09649-y.

[6] J. A. Hall and R. Schwartz, "Empathy present and future," *The Journal of social psychology*, vol. 159, no. 3, pp. 225–243, 2019. DOI: 10.1080/00224545.2018.1477442.

[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. DOI: 10.1145/2912150.

[8] M. H. Davis *et al.*, "A multidimensional approach to individual differences in empathy," 1980.

[9] 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. DOI: 10.1111/j.1467-6494.1987.tb00426.x.

[10] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, 2020. DOI: 10.1016/j.inffus.2020.01.011.

[11] S. D'Mello, A. Kappas, and J. Gratch, "The affective computing approach to affect measurement," *Emotion Review*, vol. 10, no. 2, pp. 174–183, 2018. DOI: 10.1177/175407391769658.

[12] 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. DOI: 10.1109/TPAMI.2014.2366127.

[13] Y. Liu, X. Zhang, Y. Li, J. Zhou, X. Li, and G. Zhao, "Graph-based facial affect analysis: A review," *IEEE Transactions on Affective Computing*, pp. 1–20, 2022. DOI: 10.1109/TAFFC.2022.3215918.

[14] 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. DOI: 10.1109/TAFFC.2021.3065726.

[15] A. Kaklauskas, A. Abraham, I. Ubarte, *et al.*, "A review of ai cloud and edge sensors, methods, and applications for the recognition of emotional, affective and physiological states," *Sensors*, vol. 22, no. 20, 2022, ISSN: 1424-8220. DOI: 10.3390/s22207824.

[16] Ö. N. Yalcin and S. DiPaola, "A computational model of empathy for interactive agents," *Biologically inspired cognitive architectures*, vol. 26, pp. 20–25, 2018. DOI: 10.1016/j.bica.2018.07.010.

[17] 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. DOI: 10.3390/ijerph19031889.

[18] A. S. Raamkumar and Y. Yang, "Empathetic conversational systems: A review of current advances, gaps, and opportunities," *IEEE Transactions on Affective Computing*, 2022. DOI: 10.1109/TAFFC.2022.3226693.

[19] 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.

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

[21] 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. DOI: 10.18653/v1/2023.wassa-1.44.

[22] 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. DOI: 10.18653/v1/2023.wassa-1.45.

[23] 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. DOI: 10.18653/v1/2023.wassa-1.47.

[24] 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. DOI: 10.18653/v1/2023.wassa-1.48.

[25] 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. DOI: 10.18653/v1/2023.wassa-1.49.

[26] 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. DOI: 10.18653/v1/2023.wassa-1.50.

[27] 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. DOI: 10.18653/v1/2023.wassa-1.52.

[28] 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 Lin-

guistics, Jul. 2023, pp. 574–580. DOI: 10.18653/v1/2023.wassa-1.54.

[29] 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. DOI: 10.18653/v1/D18-1507.

[30] 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. [Online]. Available: https://aclanthology.org/2021.wassa-1.10.

[31] 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. [Online]. Available: https://aclanthology.org/2022.nlpcss-1.12.

[32] 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. DOI: 10.18653/v1/2020.emnlp-main.425.

[33] 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. [Online]. Available: https://aclanthology.org/2022.emnlp-main.629.

[34] 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. DOI: 10.18653/v1/2021.clpsych-1.22.

[35] 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. DOI: 10.18653/v1/2020.emnlp-main.531.

[36] 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.

[37] 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, ISSN: 1660-4601. DOI: 10.3390/ijerph18189912.

[38] 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. DOI: 10.1609/icwsm.v11i1.14942.

[39] 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. DOI: 10.18653/v1/2021.findings-emnlp.314.

[40] 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. DOI: 10.1609/aaai.v35i14.17539.

[41] 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. [Online]. Available: https://aclanthology.org/I17-2042.

[42] 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.

[43] 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. DOI: 10.21437/Interspeech.2016-554.

[44] 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. DOI: 10.18653/v1/P19-1088.

[45] 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. DOI: 10.1109/ICSC50631.2021.00046.

[46] 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. [Online]. Available: https://aclanthology.org/2022.louhi-1.23.

[47] 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. DOI: 10.3390/app12199407.

[48] 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. DOI: 10.18653/v1/P19-1534.

[49] M. R. Jordan, D. Amir, and P. Bloom, "Are empathy and concern psychologically distinct?" *Emotion*, vol. 16, no. 8, p. 1107, 2016. DOI: 10.1037/emo0000228.

[50] 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. DOI: /10.3390/app12178805.

[51] 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, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00051.

[52] 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. [Online]. Available: https://aclanthology.org/2021.wassa-1.12.

[53] 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. DOI: 10.18653/v1/2022.wassa-1.27.

[54] 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. DOI: 10.18653/v1/2022.wassa-1.26.

[55] 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. DOI: 10.18653/v1/2022.wassa-1.29.

[56] 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. DOI: 10.18653/v1/2022.wassa-1.31.

[57] 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. DOI: 10.18653/v1/2022.wassa-1.21.

[58] 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. DOI: 10.18653/v1/2022.wassa-1.23.

[59] 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. [Online]. Available: https://aclanthology.org/2021.wassa-1.30.

[60] 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. [Online]. Available: https://aclanthology.org/2021.wassa-1.28.

[61] 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. [Online]. Available: https://aclanthology.org/2021.wassa-1.11.

[62] 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. DOI: 10.1609/aaai.v37i11.26514.

[63] 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. DOI: 10.3115/980845.980860. [Online]. Available: https://aclanthology.org/P98-1013.

[64] 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. DOI: 10.1109/ACII.2019.8925530.

[65] 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. DOI: 10.1109/icpr56361.2022.9956185.

[66] 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. DOI: 10.1109/ACII52823.2021.9597416.

[67] 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. DOI: 10.1109/IUCC-CSS.2016.021.

[68] 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. DOI: 10.1145/3340555.3353750.

[69] 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.

[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. DOI: 10.1109/AIVR56993.2022.00017.

[71] 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. DOI: 10.1080/00223890802484381.

[72] 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. DOI: 10.1109/FG.2019.8756532.

[73] 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. DOI: 10.1109/FG.2019.8756577.

[74] 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. DOI: 10.1109/FG.2019.8756517.

[75] 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. DOI: 10.1109/FG.2019.8756612.

[76] 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. DOI: 10.1109/FG.2019.8756522.

[77] 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. DOI: 10.1109/FG47880.2020.00086.

[78] 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. DOI: 10.3390/app13021163.

[79] 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.

[80] 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. DOI: 10.1145/2502081.2502224.

[81] 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. DOI: 10.3115/v1/D14-1162.

[82] 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.

[83] 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. DOI: 10.1049/tje2.12082.

[84] 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. DOI: 10.1109/CogInfoCom.2016.7804525.

[85] 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. DOI: 10.1109/ICHI48887.2020.9374402.

[86] 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. DOI: 10.7326/0003-4819-155-9-201111010-00007.

[87] 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. DOI: 10.1371/journal.pone.0143055.

[88] 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. DOI: 10.1016/j.jsat.2009.01.003.

[89] 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. DOI: 10.1007/s11682-020-00354-y.

[90] 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. DOI: 10.1109/ICBME57741.2022.10053068.

[91] 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. DOI: 10.1109/ICHMS49158.2020.9209561.

[92] 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.

[93] 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. DOI: 10.1037/0022-3514.44.1.113.

[94] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020. DOI: 10.1186/s12864-019-6413-7.

[95] H. Abdollahi, M. H. Mahoor, R. Zandie, J. Siewierski, and S. H. Qualls, "Artificial emotional intelligence in socially assistive robots for older adults: A pilot study," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2020–2032, 2023. DOI: 10.1109/TAFFC.2022.3143803.

[96] M. Spitale, S. Okamoto, M. Gupta, H. Xi, and M. J. Matarić, "Socially assistive robots as storytellers that elicit empathy," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 4, pp. 1–29, 2022. DOI: 10.1145/3538409.

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

[98] 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, ISSN: 0098-7484. DOI: 10.1001/jama.1991.03470130111039.

[99] 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. DOI: 10.1098/rstb.2015.0083.

[100] C. Rasoal, H. Danielsson, and T. Jungert, "Empathy among students in engineering programmes," *European Journal of Engineering Education*, vol. 37, no. 5, pp. 427–435, 2012. DOI: 10.1080/03043797.2012.708720.

[101] T. Kelley and D. Kelley, *Creative confidence: Unleashing the creative potential within us all*. Currency, 2013.

[102] L. Crossley, "Building emotions in design," *The Design Journal*, vol. 6, no. 3, pp. 35–45, 2003. DOI: 10.2752/146069203789355264.

[103] W. Groeneveld *et al.*, "Software engineering education beyond the technical: A systematic literature review," in *Proceedings of the 47th SEFI Conference 2019*, SEFI-European Society for Engineering Education, vol. 47, 2019, pp. 1607–1622.

[104] A. Kirman and M. Teschl, "Selfish or selfless? the role of empathy in economics," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 365, no. 1538, pp. 303–317, 2010. DOI: 10.1098/rstb.2009.0192.

[105] C. M. Clark, U. M. Murfett, P. S. Rogers, and S. Ang, "Is empathy effective for customer service? evidence from call center interactions," *Journal of Business and Technical Communication*, vol. 27, no. 2, pp. 123–153, 2013. DOI: 10.1177/10506519124688.

[106] T. Hennig-Thurau, K. P. Gwinner, and D. D. Gremler, "Understanding relationship marketing outcomes: An integration of relational benefits and relationship quality," *Journal of service research*, vol. 4, no. 3, pp. 230–247, 2002. DOI: 10.1177/1094670502004003006.

[107] V. Kumar, I. Dalla Pozza, and J. Ganesh, "Revisiting the satisfaction–loyalty relationship: Empirical generalizations and directions for future research," *Journal of retailing*, vol. 89, no. 3, pp. 246–262, 2013. DOI: 10.1016/j.jretai.2013.02.001.

[108] P. E. Salovey and D. J. Sluyter, *Emotional development and emotional intelligence: Educational implications.* Basic Books, 1997.

[109] P. N. Lopes, D. Grewal, J. Kadis, M. Gall, and P. Salovey, "Evidence that emotional intelligence is related to job performance and affect and attitudes at work," *Psicothema*, pp. 132–138, 2006.

[110] K. Cameron and J. Dutton, *Positive organizational scholarship: Foundations of a new discipline*. Berrett-Koehler Publishers, 2003.

[111] S. Mohammad and P. Turney, “Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon,” in *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, 2010, pp. 26–34.

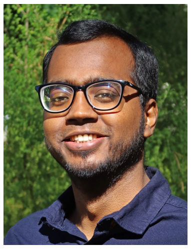
[112] K. B. Sheehan, “Crowdsourcing research: Data collection with amazon’s mechanical turk,” *Communication Monographs*, vol. 85, no. 1, pp. 140–156, 2018. DOI: 10.1080/03637751.2017.1342043.

[113] R. Jia, Z. R. Steelman, and B. H. Reich, “Using mechanical turk data in is research: Risks, rewards, and recommendations,” *Communications of the Association for Information Systems*, vol. 41, no. 1, p. 14, 2017.

[114] J. L. Huang, P. G. Curran, J. Keeney, E. M. Poposki, and R. P. DeShon, “Detecting and deterring insufficient effort responding to surveys,” *Journal of Business and Psychology*, vol. 27, pp. 99–114, 2012.

[115] A. Wang and O. Russakovsky, *Overwriting pretrained bias with finetuning data*, 2023. arXiv: 2303.06167 [cs.CV].

[116] H.-C. Yang, A. R. Rahmanti, C.-W. Huang, and Y.-C. J. Li, “How can research on artificial empathy be enhanced by applying deepfakes?” *Journal of Medical Internet Research*, vol. 24, no. 3, e29506, 2022. DOI: 10.2196/29506.



Md Rakibul Hasan received his BSc (2019) and MSc (2021) degrees from Khulna University of Engineering & Technology, Bangladesh. Currently, he is a PhD candidate in Computing at the ‘Human-Centric Group in AI’ at Curtin University, Western Australia, where he builds deep learning models to detect empathy from multimodal data, including video, audio, text and physiological signals. His overall research interest includes advancing Affective Computing and multimodal signal processing using deep learning algorithms. As a young academician, he has several publications on deep learning applications, including Affective Computing.

Md Zakir Hossain completed the BSc (2011) and MSc (2014) from Khulna University of Engineering & Technology (KUET), Khulna, Bangladesh in Electrical and Electronic Engineering, and PhD (2019) from Australian National University (ANU), Canberra, Australia in Computer Science. He is a Senior Research Fellow at the Curtin University. He has working experience with several universities and organisations, including KUET, ANU, University of Canberra, Macquarie University, and CSIRO. His research direction leads to the development of advanced technologies for health-related prediction, including emotion recognition, human computing, and diagnosing and managing diseases. He has published a number of articles in the field of affective computing and computer vision.

Shreya Ghosh received the BTech degree in CSE from the Govt. College of Engineering and Textile Technology, Serampore, 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, Melbourne, Australia, in 2022. She is currently a research academic at Curtin University. Prior to this, she was a postdoctoral fellow with Monash University, funded by DARPA. Her research interests include affective computing, computer vision, and deep learning.

Susannah Soon holds a Bachelor of Engineering from the University of Western Australia and a PhD in Artificial Intelligence from the University of Melbourne. She is currently a Senior Lecturer at Curtin University in the Discipline of Computing. Susannah’s research interests have explored the ways in which humans interact with AI and technology. Susannah’s work informs the design of effective and usable intelligent systems. She is an award-winning lecturer passionate about teaching the next generation of Computing students human-centred design.

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. Professor Gedeon’s main research area is Responsive and Responsible AI. His focus is on the development of automated systems for information extraction, from eye gaze and physiological data, as well as textual and other data, and for the synthesis of the extracted information into humanly useful information resources, primarily using neural/deep networks and fuzzy logic methods, and delivered in real, augmented and virtual environments.

Professor Gedeon 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 has been General Chair for the International Conference on Neural Information Processing (ICONIP) three times. He has been nominated for VC’s awards for postgraduate supervision at three Universities. He was a member of the Australian Research Council’s College of Experts from 2018-2021 and continues from 2024-2026. He is an associate editor of the IEEE Transactions on Fuzzy Systems and the INNS/Elsevier journal Neural Networks.