Nuances of Interrater Agreement on Automatic Affect Prediction from Physiological Signals

A Systematic Review of Datasets Presenting Various Agreement Measures and Affect Representation Schemes

1. Background

Affect Prediction Model - predict general emotional response to various stimuli.

Physiological signals – heart rate, cutaneous blood flow.

Interrater Agreement (IRA) – measure of similarity between labeling choices in supervised learning.

Problems:

- Human annotators are subjective with labeling
- Different signal and emotion interpretation [1]
- Interrater agreement measures are not standardized in all studies [1]
- Healthcare systems, gaming, automated driving [2]

2. Research Statement

To what extent does IRA influence the performance of automatic affect prediction systems in the context of physiological datasets?

3. Methodology

- **Two-Stage Systematic Literature Review** with focus on **Data Papers** (Fig. 1)[3]
- Reporting method: **PRISMA 2020 guidelines** (Fig. 2) [4]
- Search engines: Scopus, IEEE Xplore, Web of Science
- Query based on: **"rater" + "affect" + "database" + "physiological" + "predict"**
- Inclusion Criteria: with and without IRA as long as other subtopics are covered (affect prediction using physiological signals)
- Exclusion Criteria: non-English, non-human, annotation type not mentioned, no information on participants, published after April 2024
- Feasibilily Limitations: majority of non-IRA papers excluded, leaving model performance analysis for future work

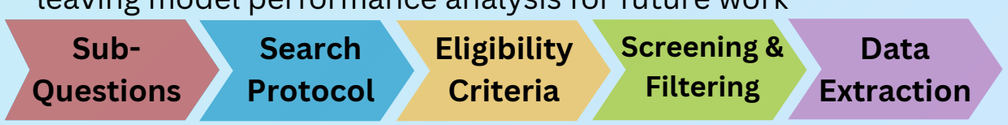

Fig. 1: Systematic Review Steps

4. Results & Discussion

a. Targetted Affect and Affect Representation Schemes (ARS)

- DEAP, MAHNOB-HCI, AMIGOS - emotions such as happiness, sadness, anger, and fear. EMOEEG, PhyMER - broader categories, including neutral states. AMIGOS, ASCERTAIN - represent moods (positive, negative, neutral) and link emotions to personality traits.
- ARS used include the circumplex model (valence and arousal), Six Basic Emotions model, and adapted models like K-EmoCon, chosen based on study goals and stimuli [5].

b. Trends in IRA Measures

- Common measures include Cohen's kappa, Krippendorff's alpha, Fleiss' kappa, Cronbach's alpha, and ANOVA. Raters vary from 3 to 346 across datasets, affecting reliability, with examples like PhyMER using 28 raters for robust estimates.
- Early 2010s: Predominantly used classical statistical methods like Cohen's kappa and Fleiss' kappa. Mid to Late 2010s: Shift towards Krippendorff's alpha and hybrid approaches using multiple measures like Cronbach's alpha and Fleiss' kappa.

c. Link Between ARS & IRA

- Datasets using the VA scheme, like EMOEEG and RECOLA, show substantial agreement due to the simplicity of evaluating only two dimensions. More detailed schemes (VAD, VADP) can reduce agreement due to increased complexity, but a higher number of raters, as seen in MAHNOB-HCI, can mitigate this effect..
- Datasets with discrete categories, such as PhyMER, achieve substantial agreement, though combining schemes (e.g., DREAMER) can lower consistency.

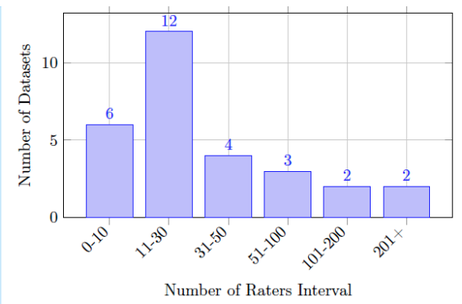

Fig. 3: No. Datasets per No. Raters

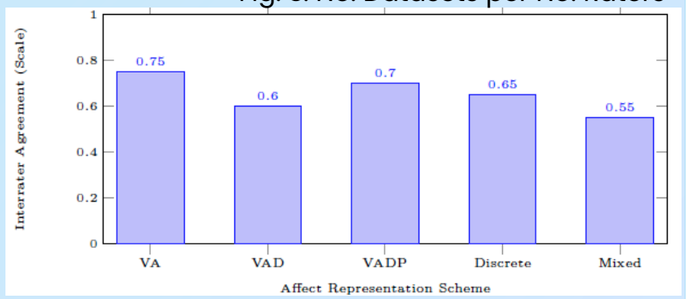

Fig. 4: IRA mean value per ARS. V : Valence, A: Arousal, D: Domination, P: Potency

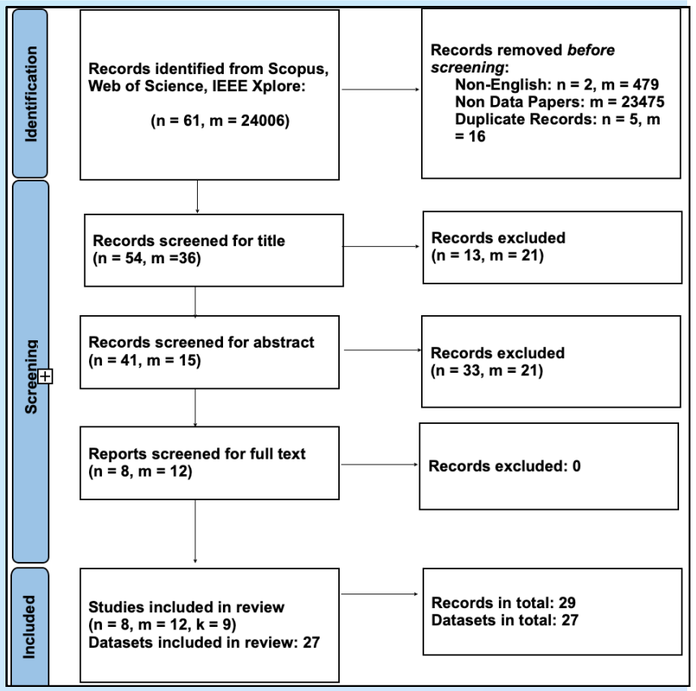

Fig. 2: Adapted PRISMA 2020 Flow Diagram

5. Conclusion & Future Work

- Simpler ARS Yield Higher IRA: Datasets with simpler affect representation schemes (ARS) like VA show higher interrater agreement (IRA), as observed in EMOEEG and RECOLA.
- Complex ARS and Raters: More complex schemes like VAD and VADP have varied agreement levels, which can be improved by increasing the number of raters, highlighting the significance of ARS selection.
- Impact on Model Performance: Future research should examine how ARS and IRA impact model performance and replicate studies without current limitations to provide more robust conclusions.

References

[1] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," Sensors, vol. 18, p. 2074, Jun 2018.

[2]Q. Meteier, M. Capallera, E. Salis, L. Angelini, S. Carrino, M. Widmer, O. Abou Khaled, E. Mugellini, and A. Sonderegger, "A dataset on the physiological state and behavior of drivers in conditionally automated driving," Data in Brief, vol. 47, p. 109027, 03 2023.

[3] M. G. Cherry, A. Boland, and R. Dickson, Doing a systematic review: A student's guide. SAGE Publications, 2024.

[4] C. Sohrabi, T. Franchi, G. Mathew, A. Kerwan, M. Nicola, M. Griffin, M. Agha, and R. Agha, "Prisma 2020 statement: What's new and the importance of reporting guidelines," International Journal of Surgery, vol. 88, p. 105918, Apr 2021.

[5] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," Scientific Data, vol. 7, Sept. 2020.

Author: Oana Madalina Fron (O.M.Fron-1@student.tudelft.nl)

TU Delft

Supervisor & Responsible Proffessor: Bernd Dudzik