# Case Studies in Multi Modal Learning for Emotion Recognition (and related applications in psychiatry)

Line H. Clemmensen, Professor,
lkhc@math.ku.dk, Dept. of Mathematical Sciences, University of Copenhagen
Technical University of Denmark

UNIVERSITY OF COPENHAGEN

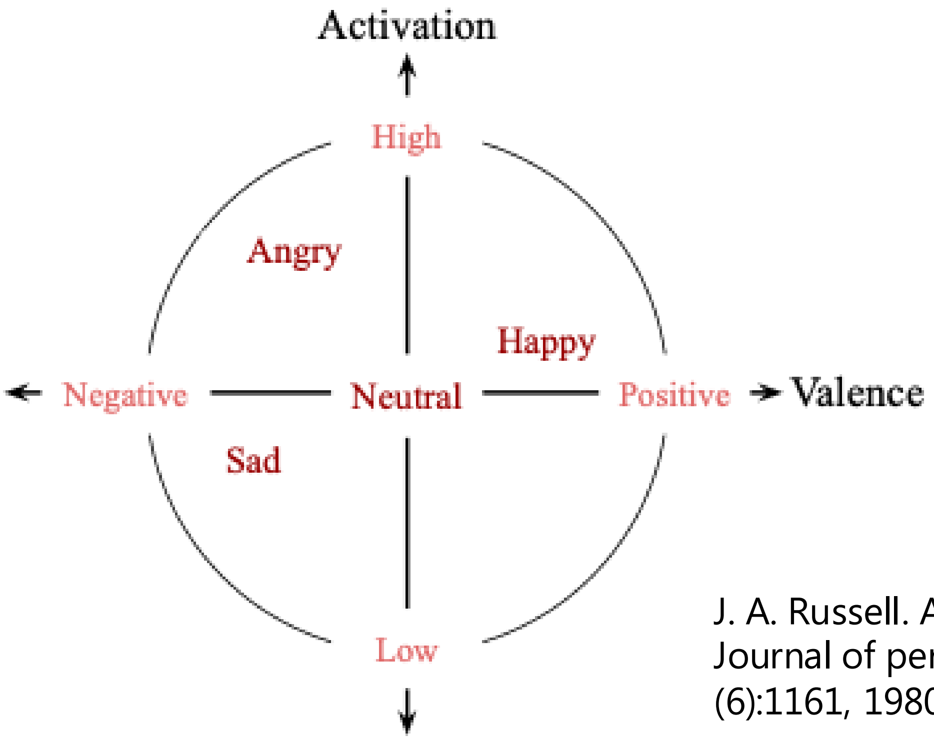
# Multimodal learning of emotions

- Ex1: Input modalities (video):
  - Visuals form video (sequence of images)
  - Speech (how are things said)
  - Text (speech-to-text; what is being said)
  - Motion capture data (not included here)

- Ex2: Input modalities (wearable):
  - 3-axis accelerometer (movement)
  - Photoplethysmography (PPG) sensor (heart rate, blood volume pulse)
  - Electrodermal activity (EDA) sensor (sweat)
  - Temperature sensor.

# Emotions

- Labels
  - Often categorical: Happy, sad, neutral, angry, disgust, etc.
  - Also: 2-dimensional, continuous constructs, like valence, arousal, etc.

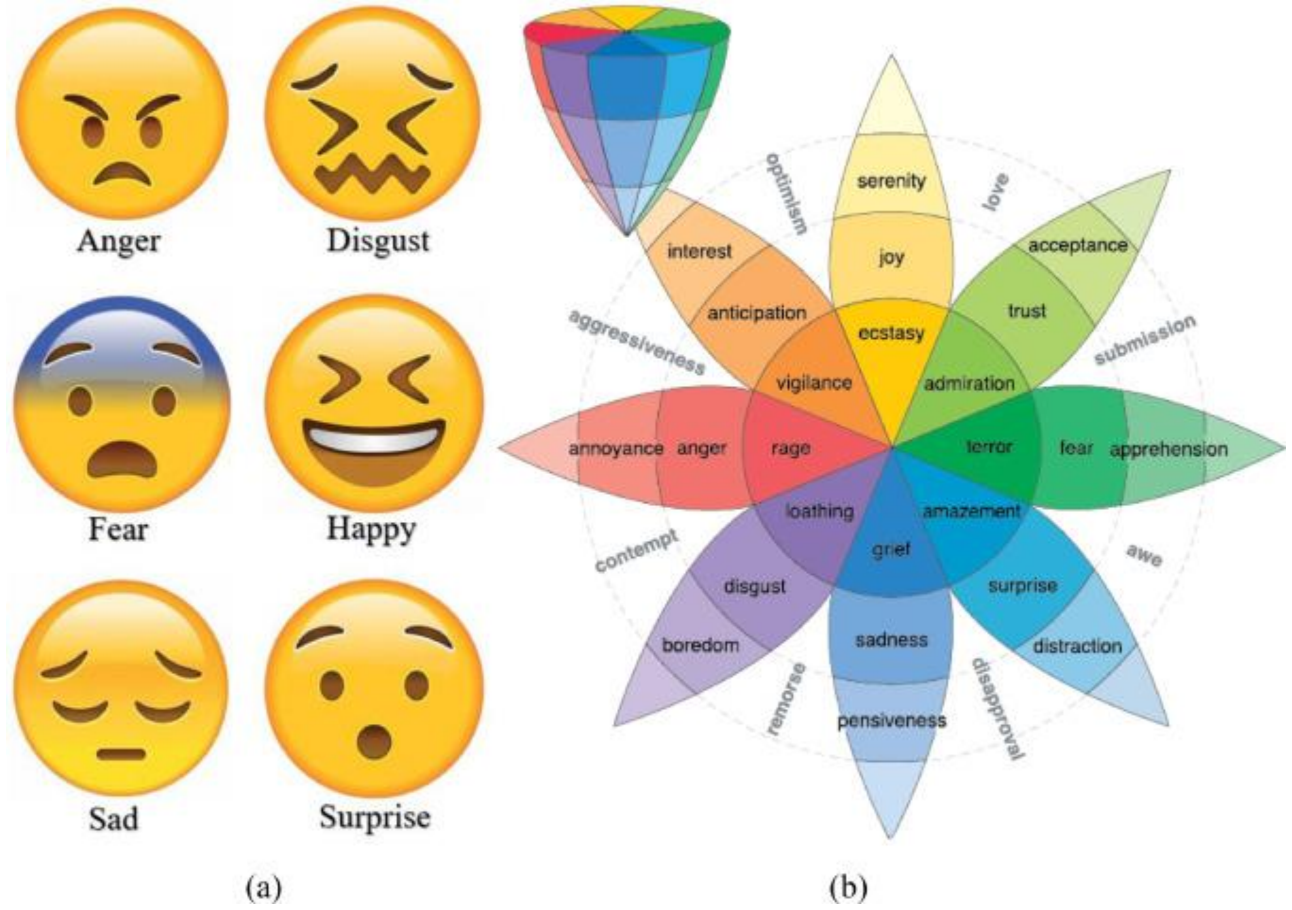



J. A. Russell. A circumplex model of affect. Journal of personality and social psychology, 39 (6):1161, 1980

# IEMOCAP dataset (Busso et al., 2008)- *Interactive Emotional Dyadic Motion Capture*

- Dyadic interactions between pairs of actors engaged in scripted dialogues and improvised scenarios
- 12 hours of interactions in five dyadic sessions, providing around 10,000 emotion-labeled utterances
- Categorical emotion labels (happy, sad, angry, neutral, disgust, fear, surprise) and dimensional attributes (valence, arousal, and dominance).
- Multiple evaluators, USC students
- 76% of utterances has 3 different evaluators, otherwise 4

# Plutchik's wheel of emotions (Robert Plutchik American Psychologist, Professor), 8 primary emotions, 1980




Wang et al, A systematic review on affective computing: emotion models, databases, and recent advances, 2022

# Pre-trained emotion recognition models

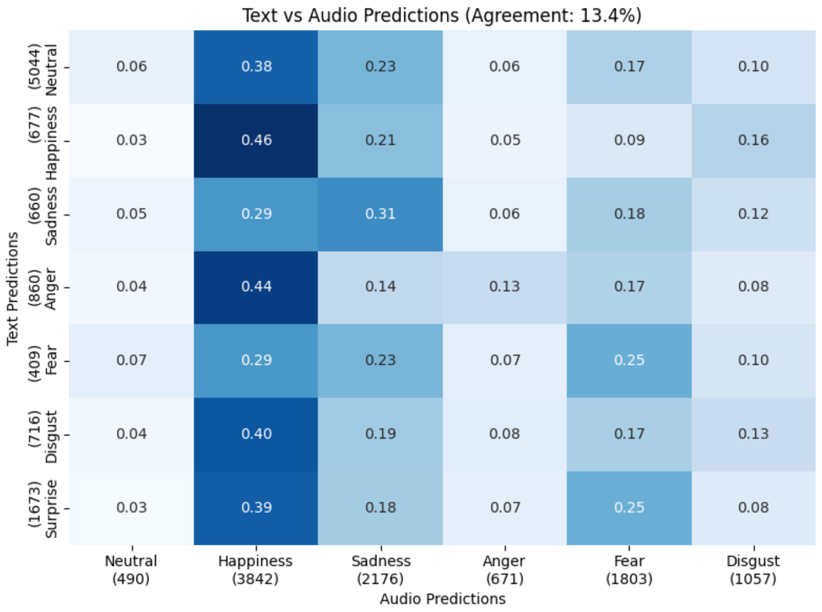
| Modality | Architecture | Model |
|---|---|---|
| Text | Transformer (DistilRoBERTa) | emotion-english-distilroberta-base; Hartmann [2022] |
| Audio | Transformer (Wav2Vec2) | w2v-speech-emotion-recognition; Khoa [2024] |
| Facial | CNN + LSTM (ResNet50 + LSTM) | EMO-AffectNetModel; Ryumina et al. [2022] |

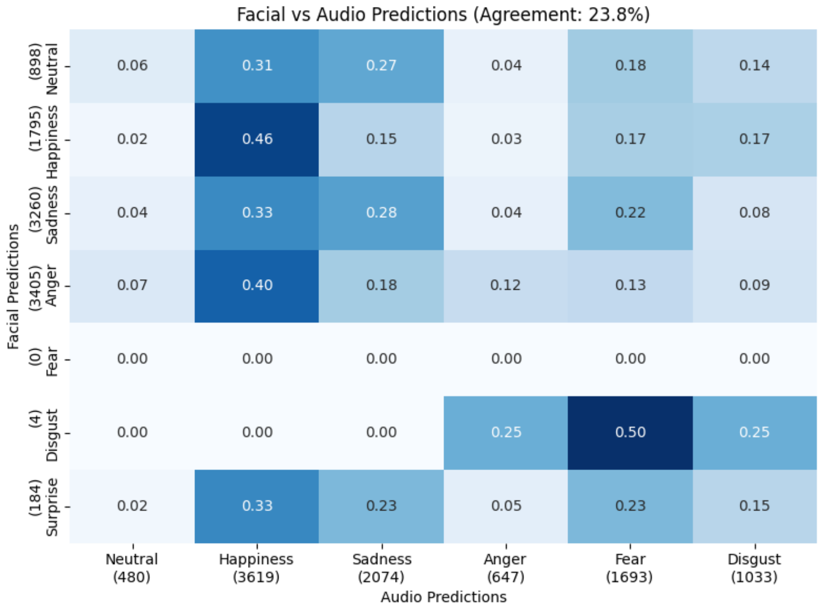
# Agreement Rate (intersection over union for two raters)

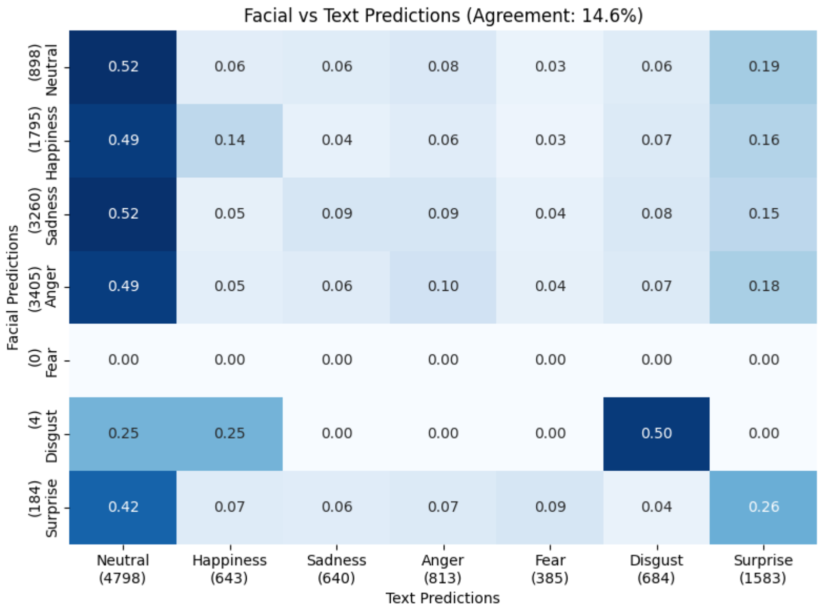
- *The proportion of utterances in which both evaluators independently labeled that same emotion*


- Agreement_*i* = $\frac{|i \cap i|}{|i \cup i|}$

# Agreement rates between modalities



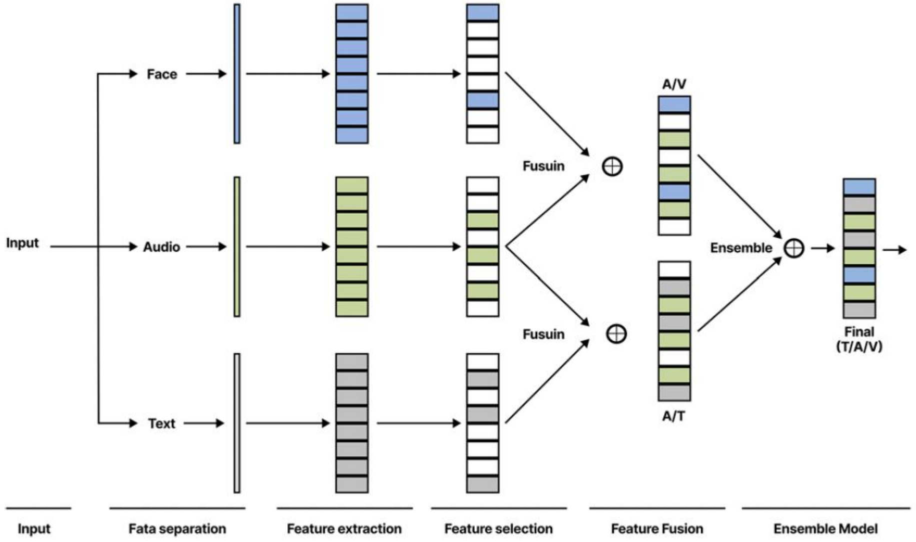

**Collaborators**: Anders Rolighed Larsen, Sneha Das, Paula Petcu, Nicole N Lønfeldt (in submission)

# Examples of ambiguity

| ID | Multimodal Information | Video frame | CEAs | VADs |
|---|---|---|---|---|
| Ses01F_script01_3_M010 | Text: "Am I embarrassing you? Are you - See, I didn't want to do it here with this yard, on this porch. I wanted it to be somewhere new, some place fresh for both of us."<br>Image: Smiling<br>Audio: Rising intonation |  | Fear<br>Excited<br>Neutral | val 5; act 5;<br>val 4; act 4;<br>val 4; act 3; |
| Ses02M_script03_1_M026 | Text: "Horrible thing, I hated it."<br>Image: Small smile<br>Audio: Falling intonation |  | Excited<br>Disgust<br>Anger | val 3; act 3;<br>val 3; act 4; |
| Ses04F_script01_3_F026 | Text: "And do you still feel that way?"<br>Image: Neutral<br>Audio: Rising intonation |  | Sadness<br>Excited<br>Neutral | val 3; act 2;<br>val 2; act 3; |

# What would you recommend now that we established the modalities (to some extend) give different predictions?

# One: Improving prediction accuracy (IEMOCAP)




| A&T | Angry | Happy | Neutral | Sad |
|---|---|---|---|---|
| *(A) Confusion matrix of audio and video fusion* | | | | |
| Angry | 74.92 | 3.58 | 3.83 | 16.81 |
| Happy | 1.92 | 73.25 | 2.08 | 22.03 |
| Neutral | 2.81 | 2.51 | 80.85 | 13.39 |
| Sad | 2.41 | 9.62 | 6.68 | 80.82 |

| A&V | Angry | Happy | Neutral | Sad |
|---|---|---|---|---|
| *(B) Confusion matrix of audio and text fusion* | | | | |
| Angry | 71.89 | 2.02 | 11.55 | 14.17 |
| Happy | 4.73 | 77.92 | 2.74 | 14.29 |
| Neutral | 4.48 | 2.22 | 80.93 | 12.05 |
| Sad | 2.91 | 9.87 | 12.16 | 74.62 |

| T&A&V | Angry | Happy | NEUTRAL | Sad |
|---|---|---|---|---|
| *(C) Confusion matrix of text, audio, and video fusion* | | | | |
| Angry | 79.62 | 1.26 | 2.1 | 16.51 |
| Happy | 1.62 | 82.8 | 0.3 | 14.9 |
| Neutral | 4.07 | 1.48 | 80.94 | 13.22 |
| Sad | 1.91 | 9.81 | 7.03 | 80.88 |

Hosseini, S.S., Yamaghani, M.R. & Poorzaker Arabani, S. Multimodal modelling of human emotion using sound, image and text fusion. *SIViP* **18**, 71–79 (2024)

# Two: Information in ambiguity?

- Think of applications where we can use differing predictions per modality to take actions
  - In an AI chatbot – ask a follow up question
  - In explainable AI – concepts like sarcasm could perhaps be revealed
  - Cultural differences – maybe different actions need to be taken in varying cultural circumstances

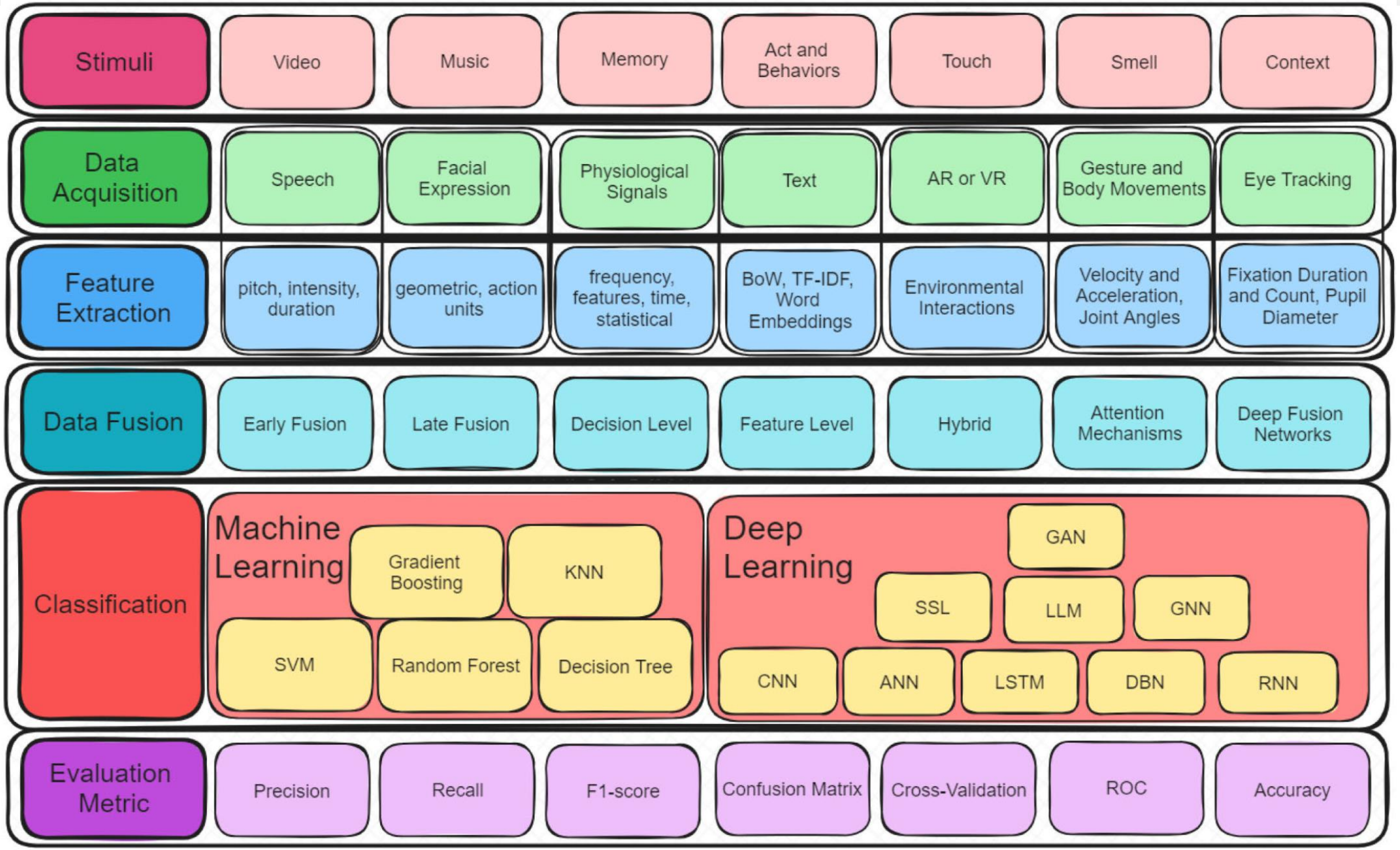
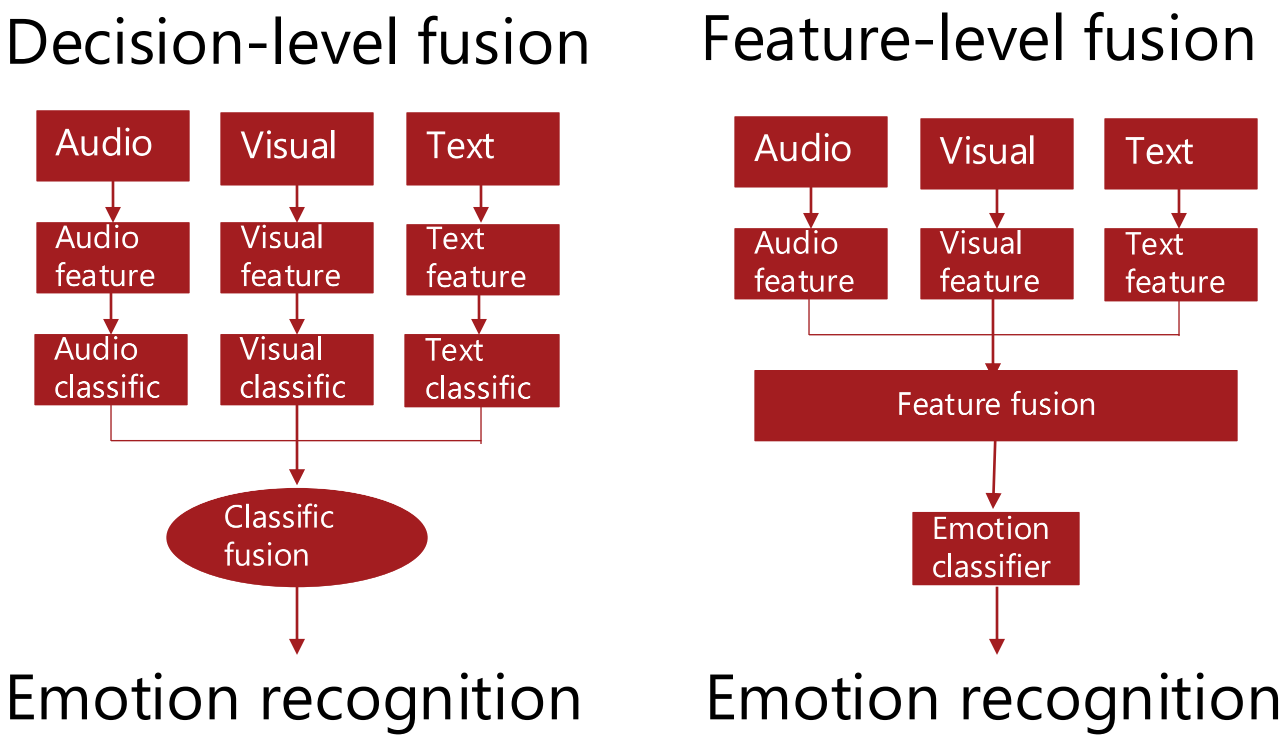



**FIGURE 15. Multimodal emotion recognition building blocks.**

Kalateh et al, A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges, IEEE Access, 2024

# Basic fusion strategies



# Review Fusion

**TABLE 8. Guidelines of the characteristics and strengths of each fusion technique.**

| Fusion technique | Suitable Modalities | Strengths | Limitations | Techniques |
|---|---|---|---|---|
| **Early Fusion (EF)** | Ideal for combining low-level features directly extracted from different modalities. | Captures detailed information from each modality early in the processing pipeline, allowing for comprehensive feature integration. | Can result in very high-dimensional feature vectors, leading to increased computational complexity and potential overfitting. Also, it might not handle missing or noisy data effectively. | Feature Concatenation, Shared Representation Learning, Tensor Fusion, Attentive Fusion. |
| **Late Fusion (LF)** | Best when each modality can be processed independently with separate classifiers. | Enables integration of decision outputs or confidence scores from individual classifiers trained on different modalities. | May miss out on capturing interactions between modalities since the integration occurs at a decision level. Performance depends heavily on the quality of individual classifiers. | Maximum Voting, Linear Weighting, D-S Evidence Theory. |
| **Mid-level Fusion** | Effective for integrating higher-level, semantically rich representations extracted from individual modalities. | Combines abstract features that capture deeper relationships between modalities. | Still can be computationally intensive and might require significant preprocessing to extract meaningful high-level features. | Extracts and combines features after initial processing stages to enhance interpretability and integration. |
| **Hybrid Fusion** | Combines multiple fusion techniques to leverage complementary strengths across different modalities. | Provides flexibility and robustness by integrating both low-level and high-level features effectively. | Complex to implement and optimize, potentially requiring more resources and sophisticated architectures. | Using EF for low-level feature integration and mid-level fusion for abstract representation integration. |
| **Feature-level Fusion** | Suitable when direct integration of raw or processed features is beneficial. | Facilitates comprehensive utilization of multimodal features at a fundamental level. | High-dimensional vectors can lead to increased computational cost and overfitting. Difficulty in handling missing data. | Feature concatenation, averaging, or applying more complex operations to merge the information from each modality. |
| **Decision-level Fusion** | Works well when each modality can provide an independent assessment of emotions. | Simplifies the fusion process by dealing with classifier outputs rather than raw features. | May lose nuanced interactions between modalities and rely heavily on individual classifier performance. | Voting schemes (average, majority vote), weighted averaging, or stacking to enhance decision-making. |
| **Attention-based Fusion** | Dynamically weights the contribution of different modalities based on their relevance to the task. | Effective for scenarios where modalities vary in importance or relevance over time or context. | Flexibility in focusing on informative parts while mitigating noise or irrelevant information. | Can be computationally intensive and requires careful tuning of attention mechanisms. |
| **Graph-based Fusion** | Uses GNNs to model relationships and interactions between modalities. | Beneficial when capturing complex dependencies and interactions between features from different modalities. | Enhances robustness and accuracy by integrating structured relationships in multimodal data. | Computationally expensive and requires expertise in GNNs. The performance can be sensitive to the graph structure and parameters. |
| **Transformers for Multimodal Fusion** | Utilizes transformer architectures to capture long-range dependencies and interactions across text, audio, and visual modalities. | Effective for integrating information across diverse and complex modalities. | Improves accuracy in capturing nuanced interactions and dependencies between modalities. | Very high computational requirements and complex architecture that requires large datasets and significant computational resources. |

Kalateh et al, A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges, IEEE Access, 2024

# Modalities

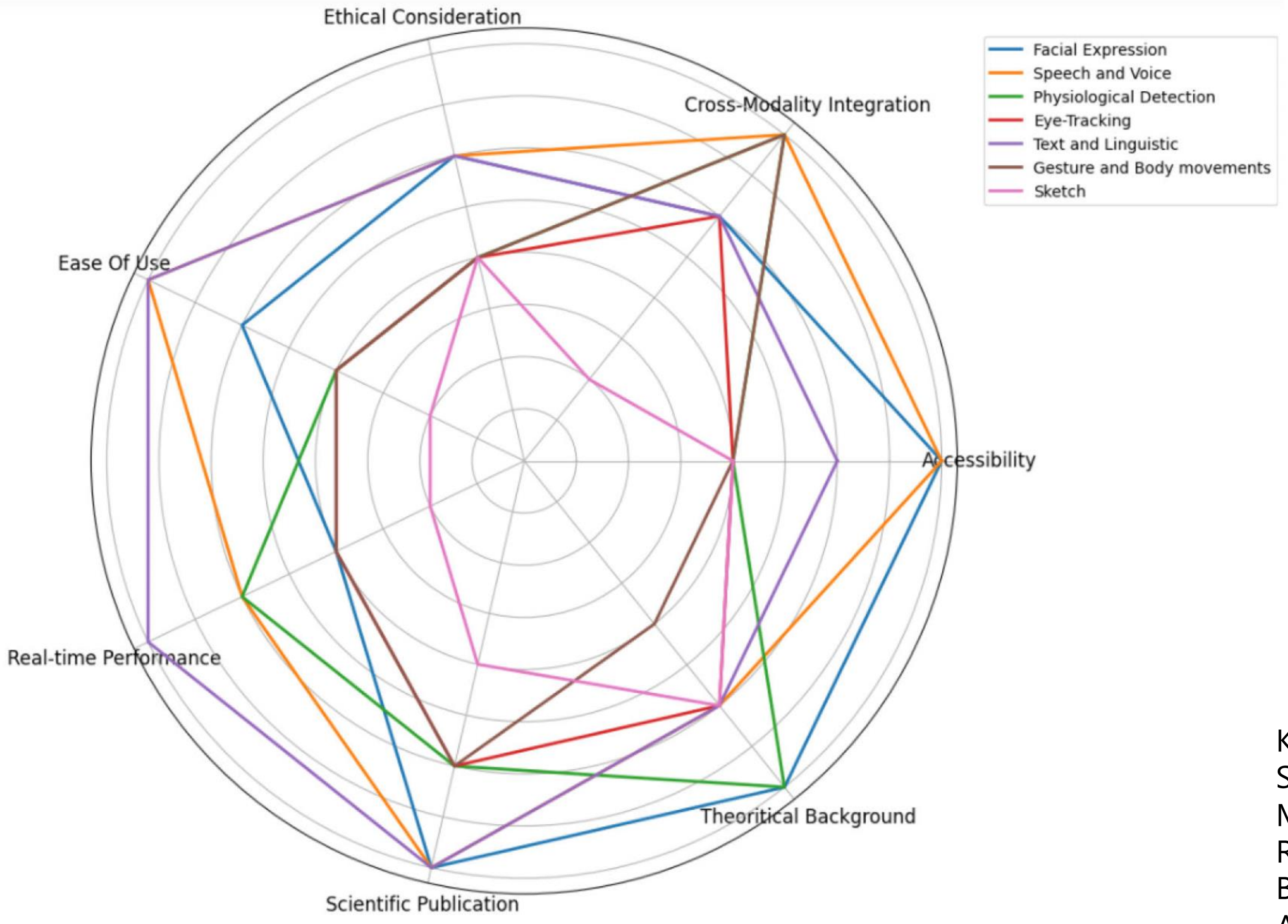
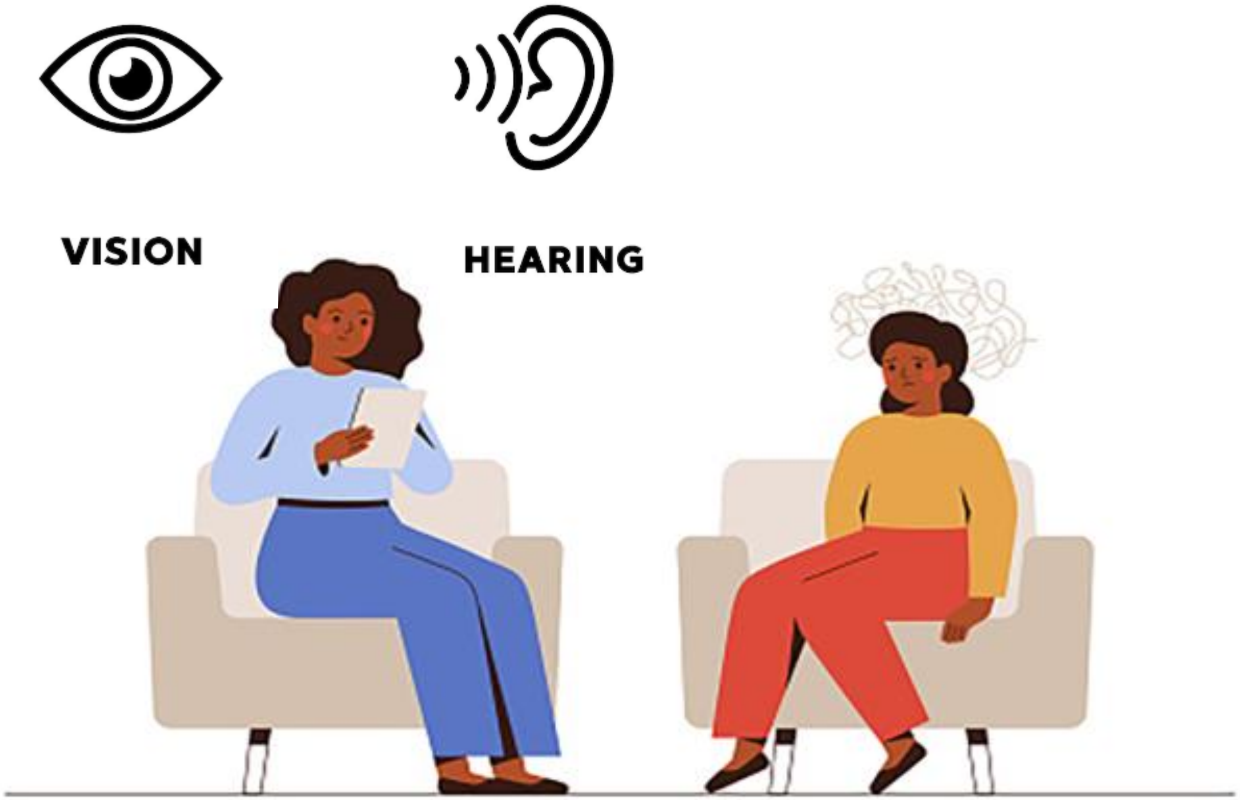



**FIGURE 16. Radar graph of emotion recognition modalities evaluation based on the selected criteria.**

Kalateh et al, A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges, IEEE Access, 2024

# Applications in Psychiatry

# Behavioral coding




**Applications**

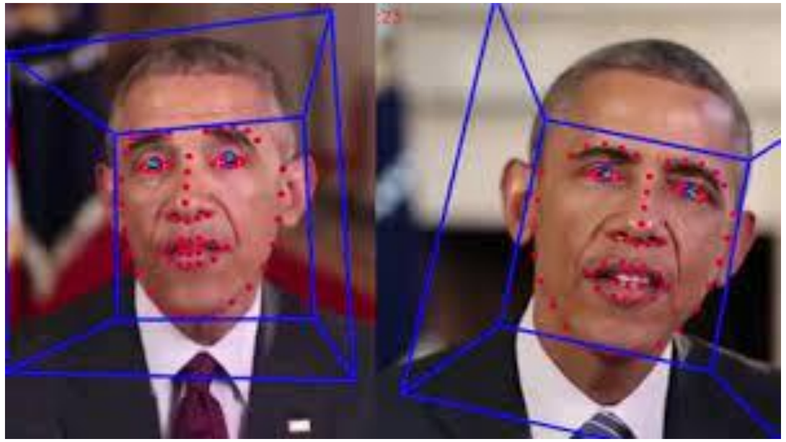
- Fidelity
- Therapy processess
- Parent & child behavior

**Limitations**

- Time-consuming
- Expensive
- Bias

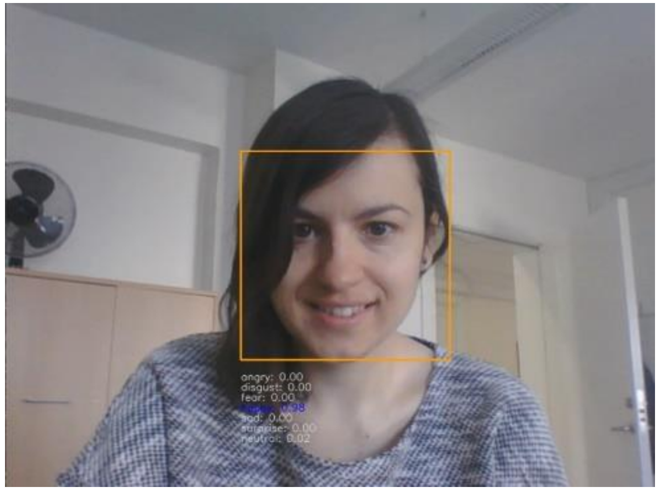
## OpenFace (Baltrusaitis et al., 2018)

- Gaze & Facial action units



## Facial Emotion Recognition (FER)

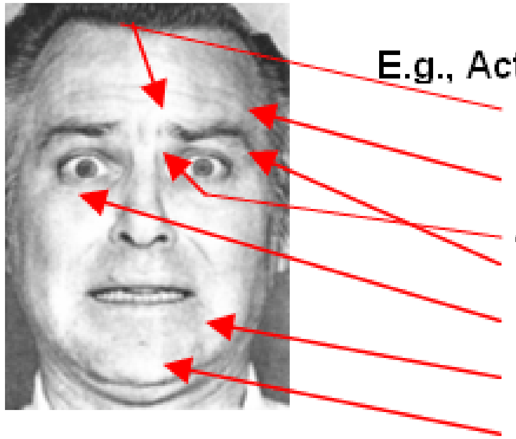
Python package *fer* (Zhang et al., 2016; Arriaga et al., 2017)



angry: 0.00
disgust:0.00
fear:0.00
happy:0.98
sad:0.00
surprise:0.00
neutral:0.02

## Facial Action Coding System (FACS)

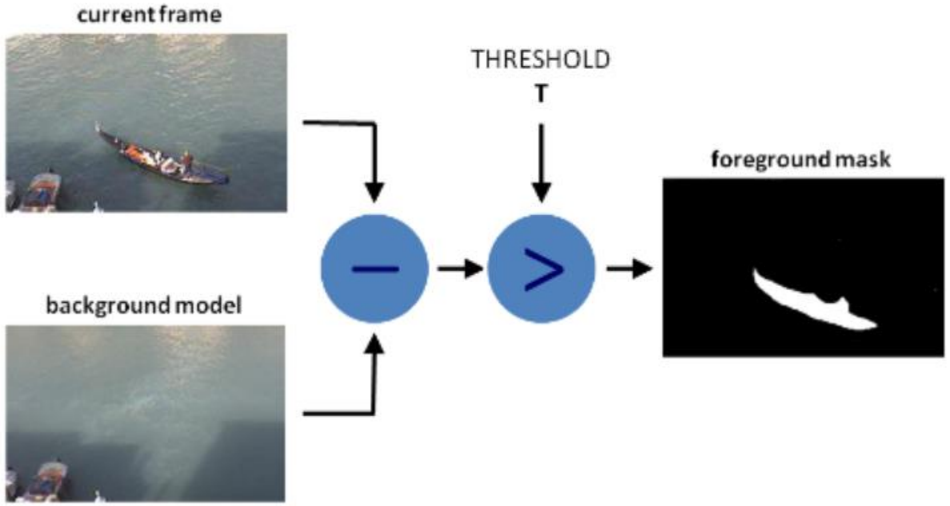
(www.paul.Ekman.com)



**E.g., Action code: 1, 2, 4, 5, 7, 20,**

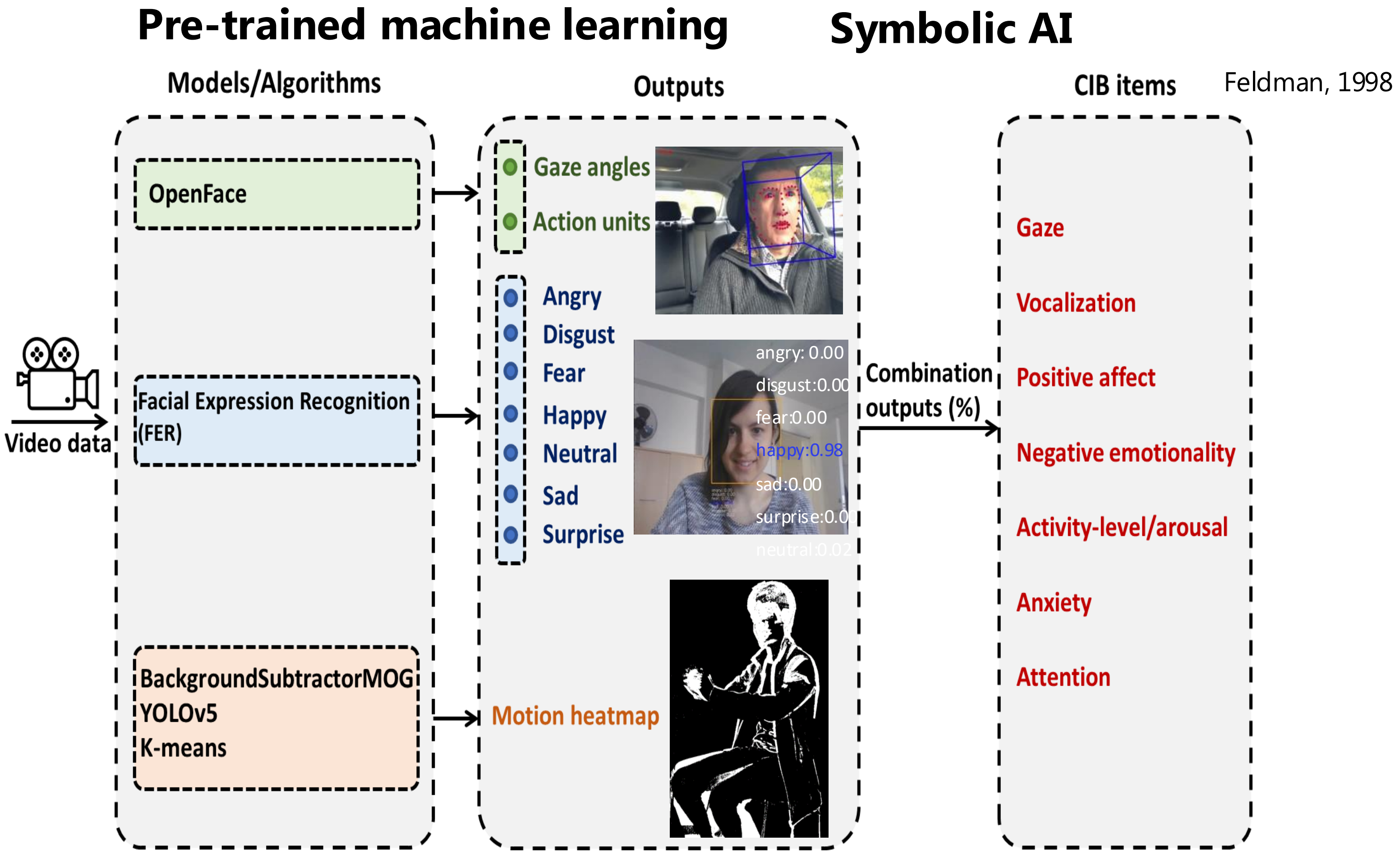
| | |
|---|---|
| 1C | Inner brow raise |
| 2C | Outer brow raise |
| 4B | Brow lower |
| 5D | Upper lid raise |
| 7B | Lower lid tighten |
| 20B | Lip stretch |
| 26B | Jaw drop |

## Motion Energy Analysis (MEA)



# Interpretability by design

**(Inspired by Concept bottleneck, Koh et al 2020)**

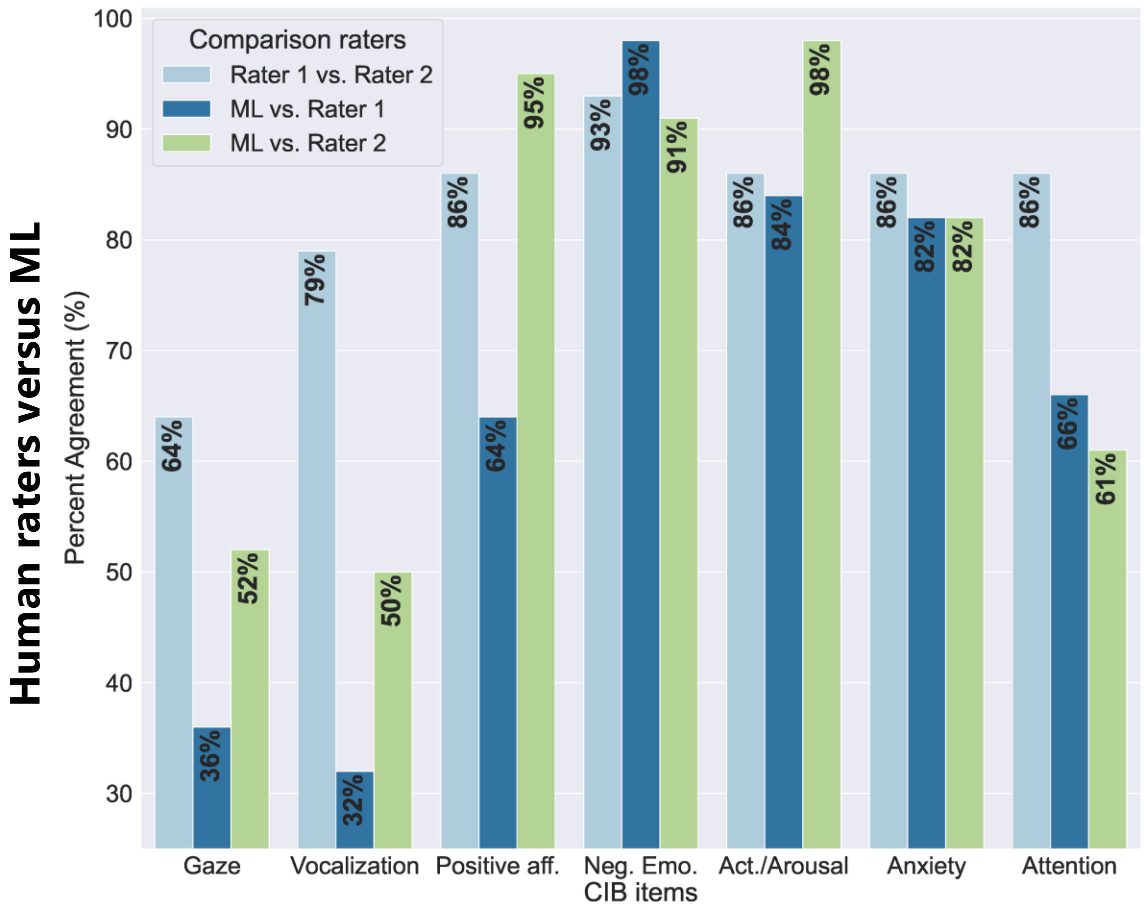



**Data:**

30-sec of mania & 30-sec of depression chapters of K-SADS screening videos. OCD = 50 videos, no-OCD = 24 videos.

Frumosu, Lønfeldt NN., Mora-Jensen., Das, S., Lund., Pagsberg, Clemmensen: *Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML), 2022.*

# Comparison to experts




$$\text{Percent agreement (\%)} = \frac{\text{number agreements}}{\text{total number items}} \times 100$$

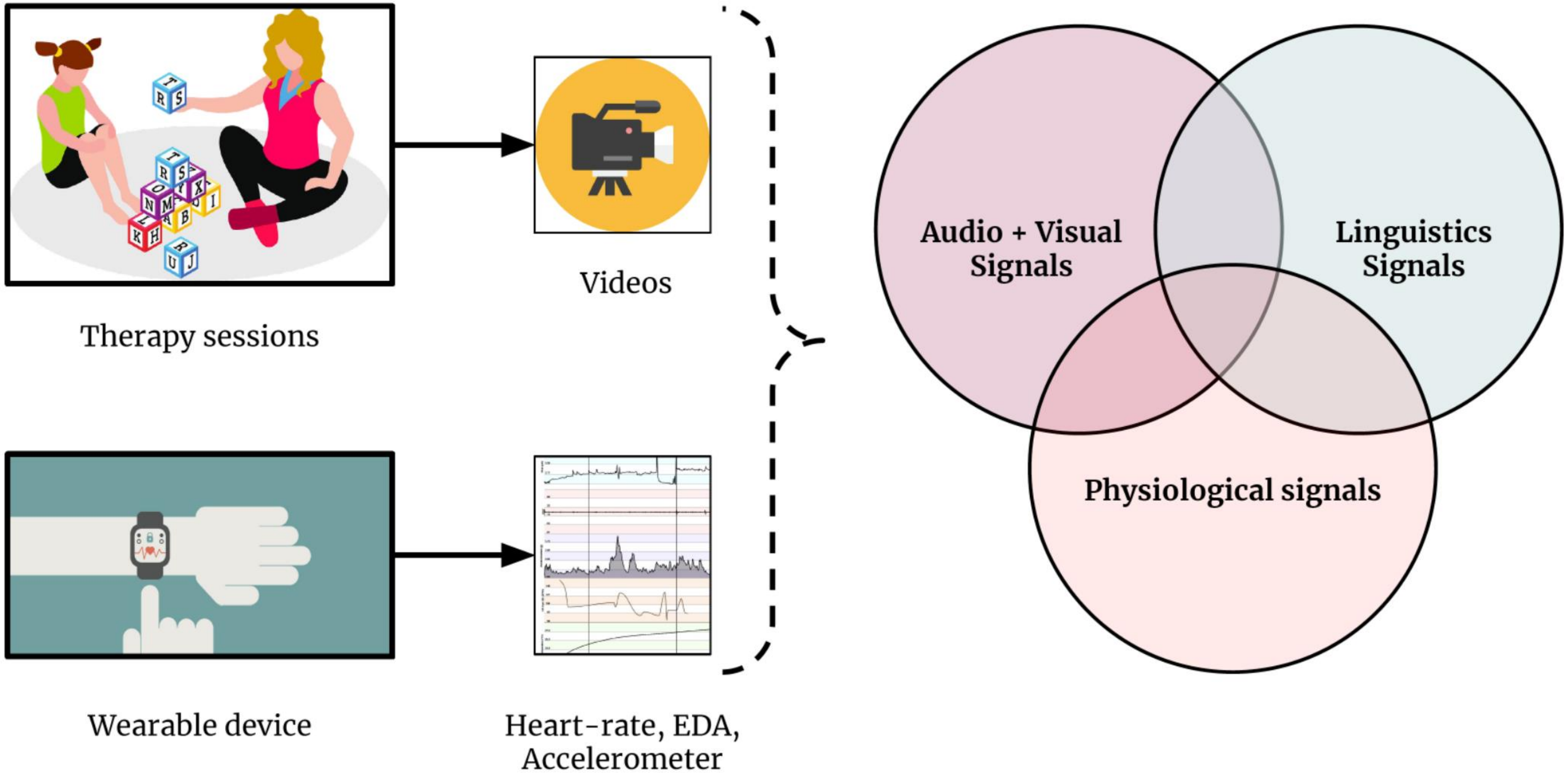
*Table 1.* Average percent agreement over the videos

| Comparison | Percentage agreement |
|---|---|
| Rater 1 vs. Rater 2 | 83% |
| ML vs. Rater 1 | 66% |
| ML vs. Rater 2 | 76% |

*Table 2.* Average percentage agreement over the videos. Dropped CIB items : gaze and vocalization

| Comparison | Percent agreement |
|---|---|
| Rater 1 vs. Rater 2 | 87% |
| ML vs. Rater 1 | 79% |
| ML vs. Rater 2 | 85% |

# Wearables – Predicting OCD events from biosignals



# WristAngel - A Wearable AI Feedback Tool for OCD Treatment and Research

NNF Exploratory Synergy Grant

novo nordisk fonden

$N_{pt}$ = 9 + 9
$N_{ct}$ = 9 + 9

Lab visit

Parent questionnaires

Biosensor questionnaire

Lab visit

*In-the-wild*

Wear biosensor daily
Tag OCD/anxiety events

Week 0 — Start

Week 4

Week 7

Week 8 — End

# Summarizing

9 participants (Five girls, four boys)

- Ages of 10 and 16 years (mean age = 12.3, SD = 2.6)
- Diagnosed with OCD (F42.2 according to the International Statistical Classification of Diseases and Related Health Problems Organization, 1993)
- At enrolment, OCD severity scores were from mild to moderate severe (mean= 24.56, SD = 5.12).

- The Empatica E4 **wristband** measures:
  - Heart rate (HR),
  - Blood volume pulse (BVP),
  - External skin temperature (TEMP), and
  - Electrodermal activity (EDA).

# Pre-processing

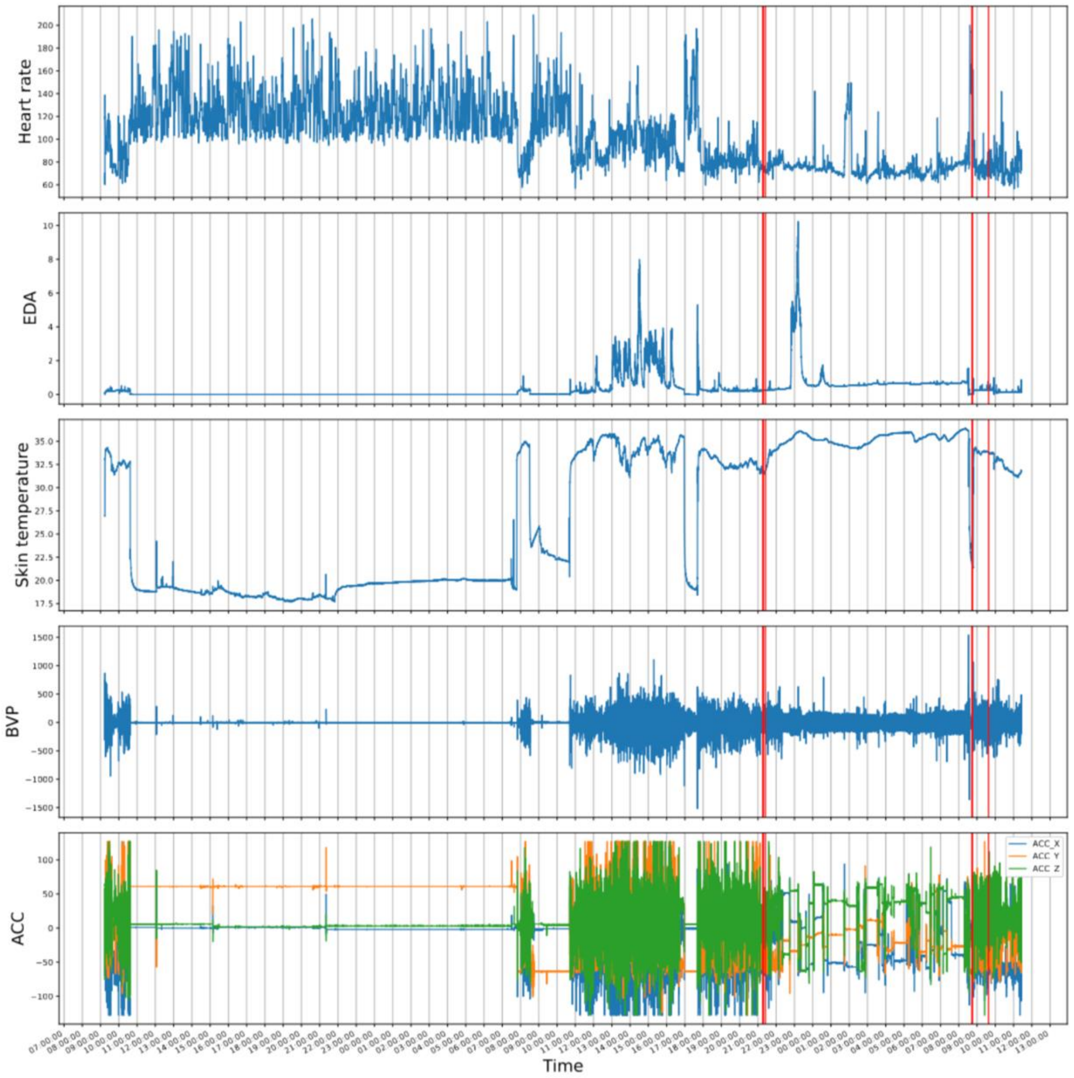
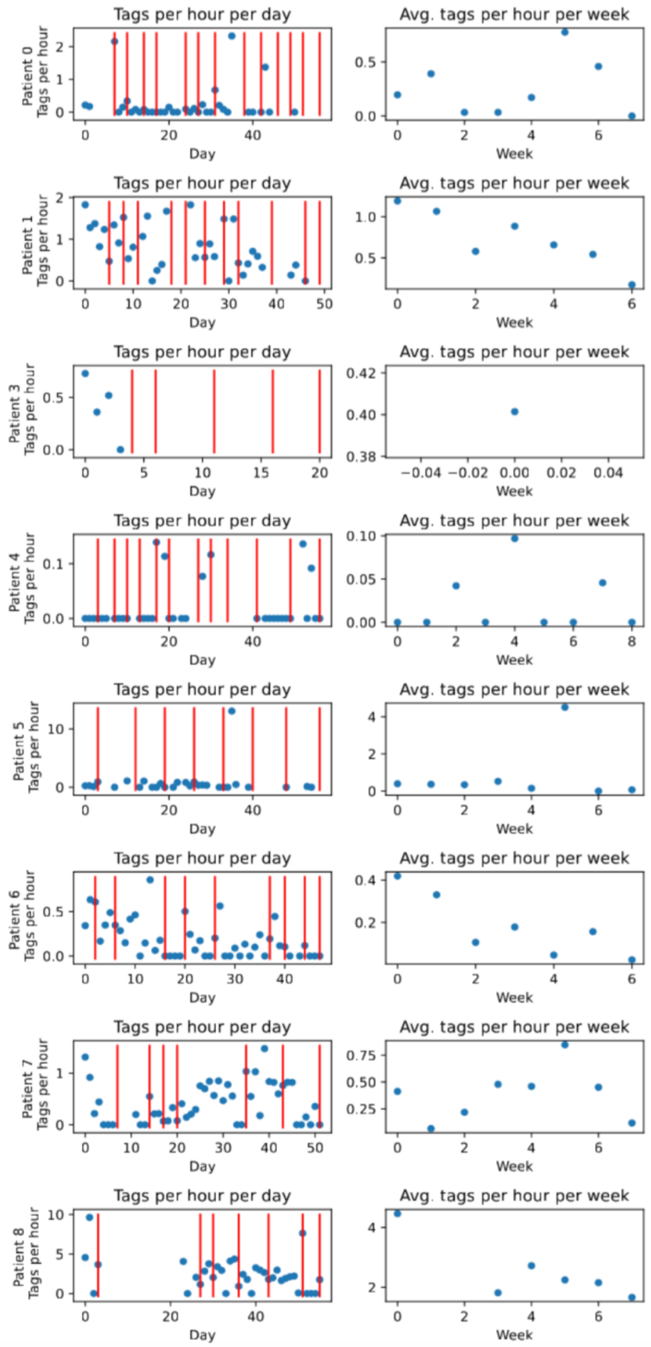
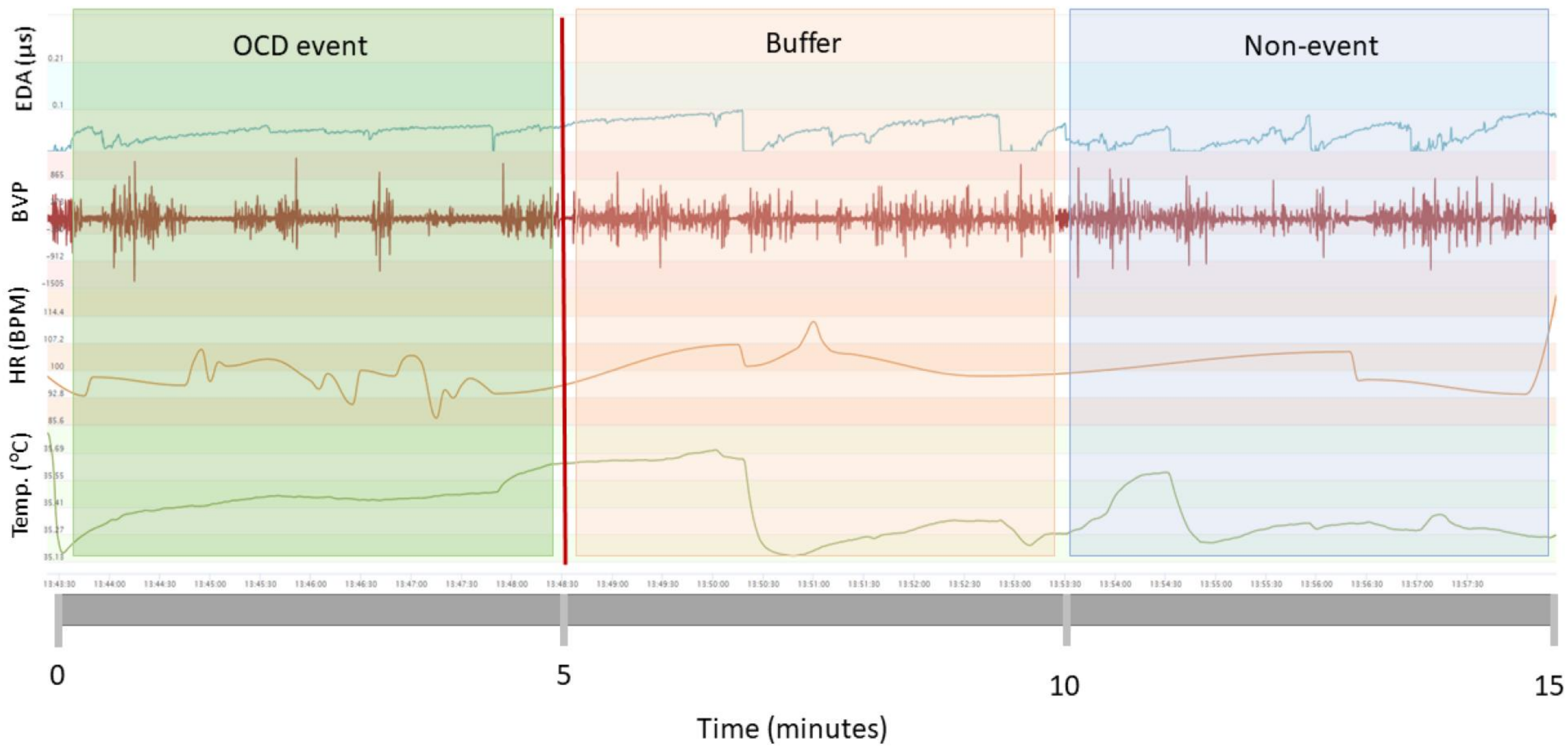



**Figure S1.** Recording from a wristband containing both a period of identified sleep and periods when the wristband was not worn. Red lines denote tagged OCD events.



# Sampling events and non-events



# Feature extraction

- We did quite a lot of this…
- Blood Volume Presssure (BVP)
  - Assess **noise** using skewness and kurtosis for windowed signals (5s)
  - Identify systolic peaks using the NeuroKit2
  - Extract: **Average** inter-beat interval and root mean square of successive differences (RMSSD) for **low-noise windows**.
  - Time-domain features: mean, standard deviation (SD), median, minimum, maximum, and slope.
  - Frequency-domain features: mean, SD, median, interquartile range, minimum, maximum, and sum of frequencies.
  - The frequency-domain features were split into real and imaginary components. All features were averaged across the low noise segments within each five-minute window for the final set of features.
  - Finally, we included the minimum and maximum slopes for a low-noise segment as features.

# Feature extraction

- Heart rate (HR)
  - Calculated directly in the E4 using a proprietary algorithm.
  - Five minute windows: mean, SD, minimum, 25% quantile, median, 75% quantile, maximum, interquartile range, and slope.

- Skin temperature
  - Pre-processed using a sixth-order Butterworth low-pass filter with a cut-off frequency of 1Hz.
  - Five minute windows: mean, standard deviation, minimum, maximum, and slope.

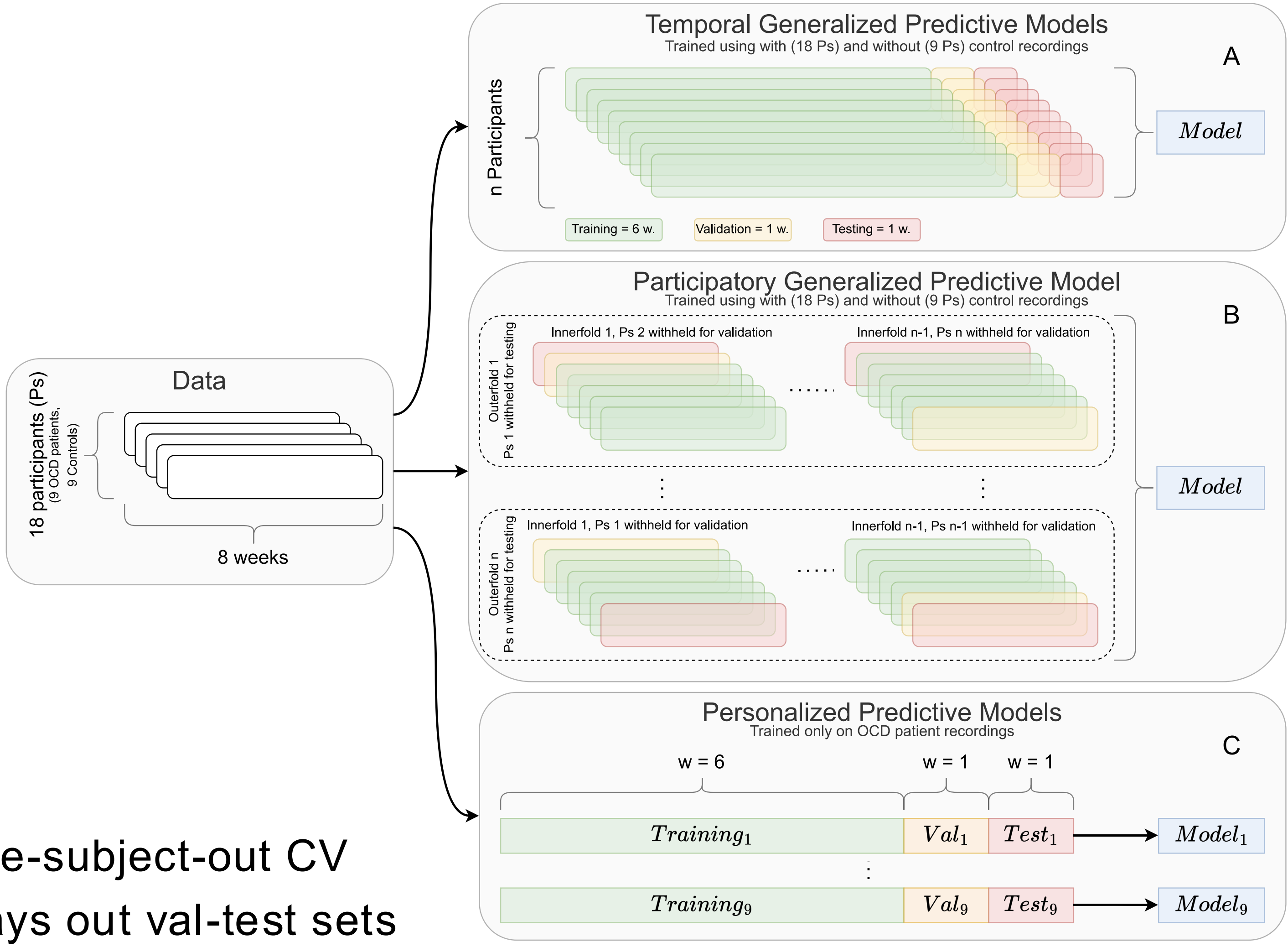
# Feature extraction

- Electrodermal Activity (EDA)
  - Pre-processed using a sixth-order Butterworth low-pass filter with a cut-off frequency of 1Hz, **Normalized** to [0, 1].
  - The normalized signal was decomposed into its **tonic** and **phasic** parts using the **NeuroKit2**.
  - Five minute windows: mean, standard deviation, minimum, maximum, and slope.
  - Tonic component, 5 min windows: minimum, 25% quantile, median, 75% quantile, maximum, interquartile range, and slope.
  - Phasic components: mean, standard deviation, number of peaks, average peak amplitude, average response time, and power in the frequency bands ultralow frequency (ULF: 0.01-0.04 Hz), low frequency (LF: 0.04-0.15 Hz), high frequency (HF:198 0.15-0.4 Hz), and ultra-high frequency (UHF: 0.4-1.0 Hz).
  - From the **unnormalized signal**: mean, standard deviation, minimum, 25% quantile, median, 75% quantile, maximum, interquartile range, and power in the frequency bands above.

# Methods

- Machine Learning models:
  - Logistic regression (LR),
  - Random forest (RF),
  - Feedforward neural networks (NN)
  - Mixed-Effect Random Forest (MERF)

- Cross-validation procedure:
  - 10-fold random CV
  - Generalized partipant based: leave-one-subject-out CV
  - Temporal generalized: leave-12.5%-days out val-test sets
  - Personalized: train and test on one person.




*Olesen, K. V., Lønfeldt, N. N., Das, S., Pagsberg, A. K., and Clemmensen, L. K. H. (2023). Feasibility of predicting obsessive-compulsive disorder events in children and adolescents from biosignals in-the-wild - a wrist angel analysis plan. JMIR Preprints 48571 doi:10.2196/preprints.48571*

# Predicting OCD events from biosignals

ROC validation, best possible (random CV) & across time

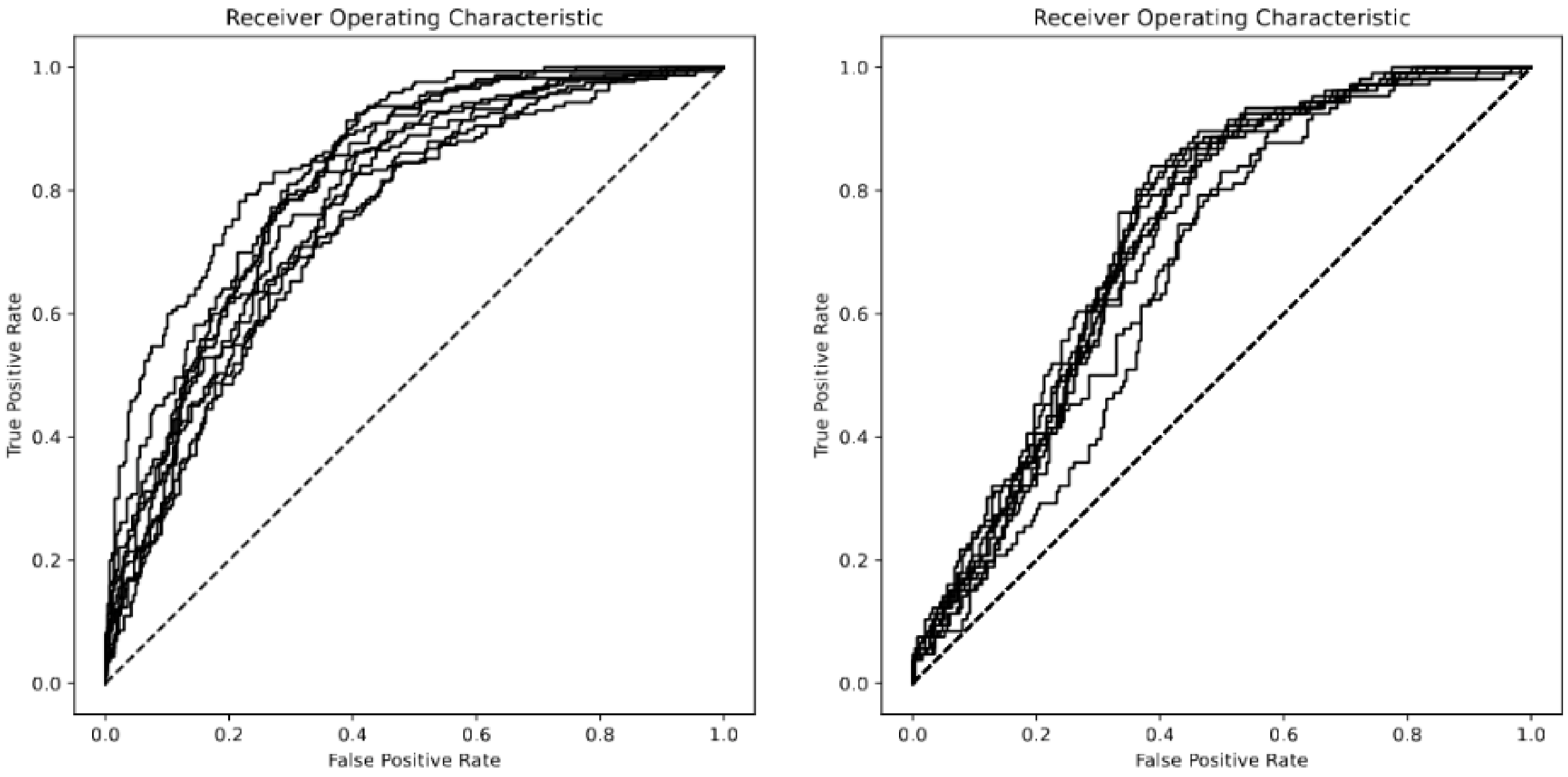
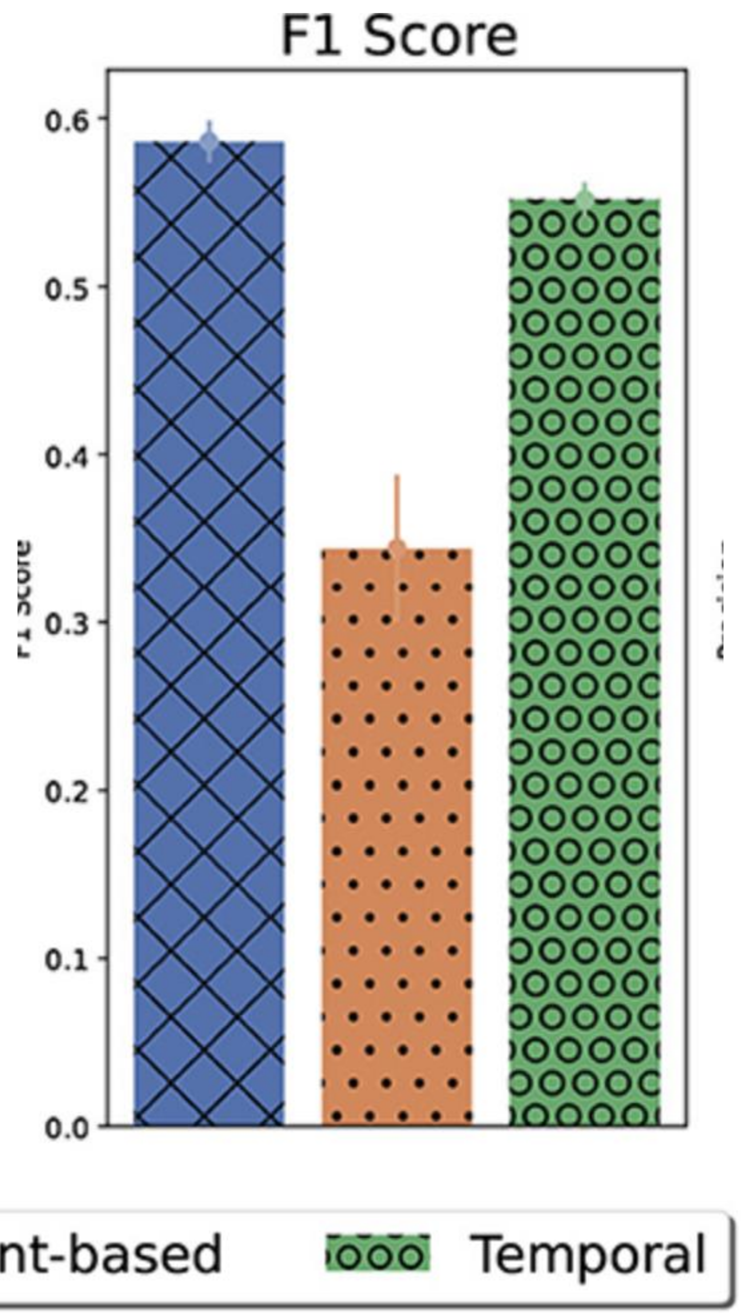



**Figure 3a.** Random cross-validation.

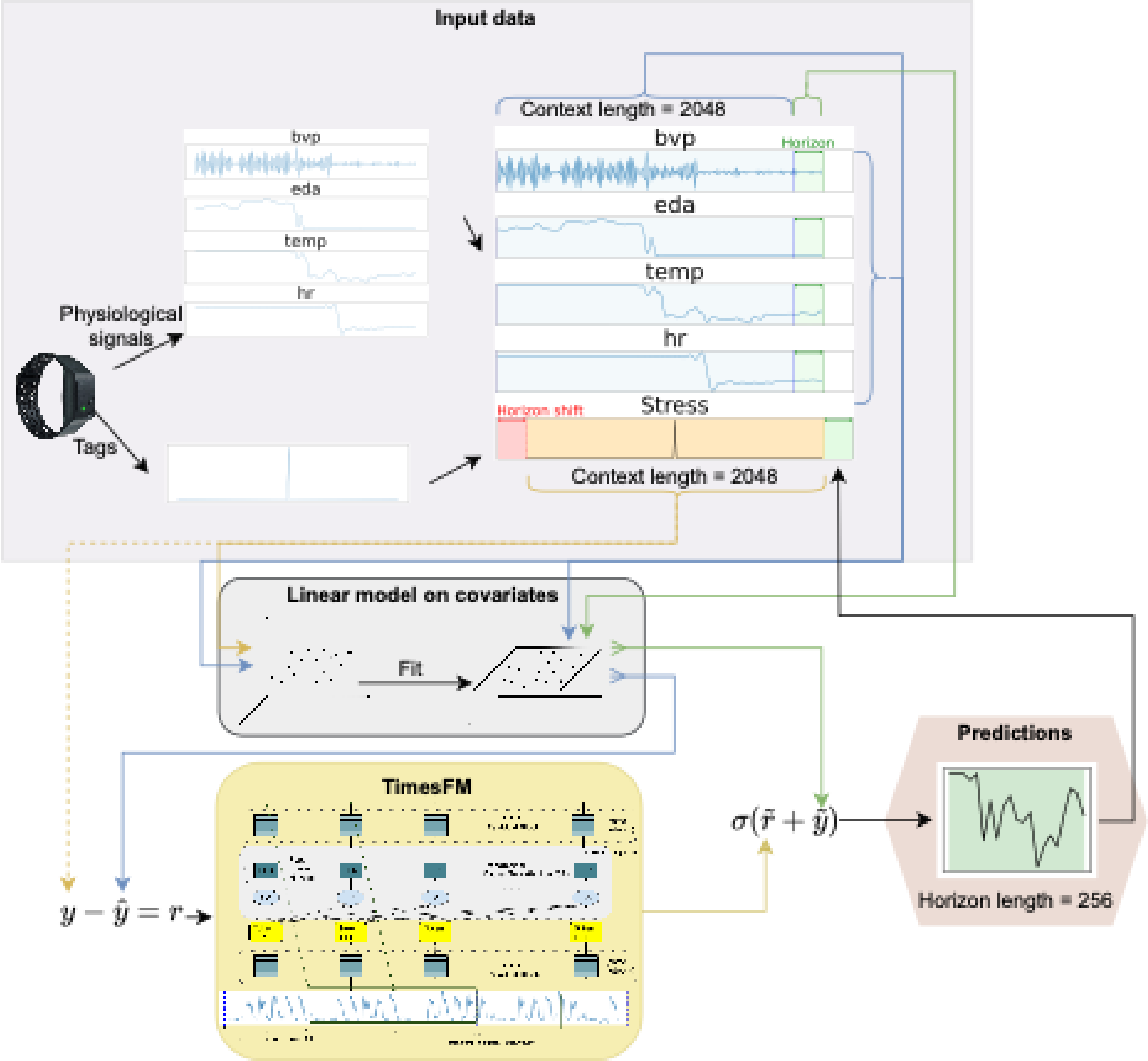
**Figure 3b.** Temporal cross-validation.




Lønfeldt, Olesen, Das, Mora-Jensen, Pagsberg, Clemmensen, Front. Psychiatry 2023.

# Multimodal learning using a foundation model (TimesFM)

- TimesFM, a newly proposed transformer, foundation model, for time series.
- F1_5min = 0.31




- Das, Abhimanyu, Weihao Kong, Rajat Sen, et al. (2024). A decoder-only foundation model for time-series forecasting. arXiv: 2310.10688 [cs.CL]. URL: https://arxiv.org/abs/2310.10688
- Collaboration: Harald Skat-Rørdam, Kathrine Sofie Rasmussen, Sneha Das

# Thank You – Keep Learning