Systematic Evaluation of Personalized Models for Affective Computing

Yunjo Han, Panyu Zhang, Minseo Park, Uichin Lee
Interactive Computing Lab, KAIST

UBICOMP ISWC 2025

https://github.com/Kaist-ICLab/Personalized_Affective_Computing

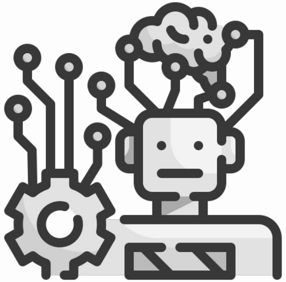
Affect Recognition via Physiological & Behavioral Signals

Affect

Valence, Stress

?

Affective Computing



Physiological & Behavioral Signal

→ Affect indicators

Machine Learning Model

→ Affect Recognition

[1] R. M. Nesse, "Evolutionary explanations of emotions," Human nature, vol. 1, pp. 261–289, 1990.
[2] R. A. Ferrer and W. B. Mendes, "Emotion, health decision making, and health behaviour," 2018.
[3] R. W. Picard, Affective computing. MIT press, 2000.
[4] J. Kim and E. Andr é, "Emotion recognition based on physiological changes in music listening," IEEE transactions on pattern analysis and machine intelligence, vol. 30, no. 12, pp. 2067–2083, 2008.
[5] M. Egger, M. Ley, and S. Hanke, "Emotion recognition from physiological signal analysis: A review," Electronic Notes in Theoretical Computer Science, vol. 343, pp. 35–55, 2019.

Personalized Affective Computing

Individual Differences

One-size-fits-all (generalized) model

→ Overlook individual differences

& Resulted in poor performance

↓

Developing **personalized models**

→ Enhanced model performance

[1] S. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," IEEE Transactions on Affective Computing, vol. 11, no. 2, pp. 200–213, 2017.
[2] J. Li, A. Waleed, and H. Salam, "A survey on personalized affective computing in human-machine interaction," arXiv preprint arXiv:2304.00377, 2023.
[3] Hamann and T. Canli, "Individual differences in emotion processing," Current opinion in neurobiology, vol. 14, no. 2, pp. 233–238, 2004.
[4] Y. S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, G. Riva, and C. Ersoy, "Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches," IEEE Access, vol. 8, pp. 38146–38163, 2020.
[5] L. K. Barr, J. H. Kahn, and W. J. Schneider, "Individual differences in emotion expression: Hierarchical structure and relations with psychological distress," Journal of Social and Clinical Psychology, vol. 27, no. 10, pp. 1045–1077, 2008.

Categories of Personalization Techniques

User-dependent	Creating separate models for each individual using only their own data
Hybrid	Creating separate models for using all users data, including each individual's data
Fine Tuning	Re-training generalized model with a small amount of individual data
Cluster-specific	Creating separate models for groups classified based on certain criteria (e.g., gender, personality)
Multi-task Learning	Learning multiple related tasks simultaneously and sharing representations

[1] J. Li, A. Waleed, and H. Salam, "A survey on personalized affective computing in human-machine interaction," arXiv preprint arXiv:2304.00377, 2023.

Prior Studies on Personalization Techniques

- **User-dependent** — Creating separate models for each individual using only their own data
- Creating separate models for using all users data, including each individual's data
- **Fine Tuning** — Re-training generalized model with a small amount of individual data
- **Cluster-specific** — Creating separate models for groups classified based on certain criteria (e.g., gender, personality)
- **Multi-task Learning** — Learning multiple related tasks simultaneously and sharing representations

No prior studies systematically evaluated the effectiveness of diverse personalization techniques using multiple open datasets

[1] J. Li, A. Waleed, and H. Salam, "A survey on personalized affective computing in human-machine interaction," arXiv preprint arXiv:2304.00377, 2023.

Research Goal

Systematically evaluating personalization techniques

in affective computing

- Understand the differences among various personalized models
- Determine whether they truly outperform the generalized models
- For reproducibility, publicly share evaluation process

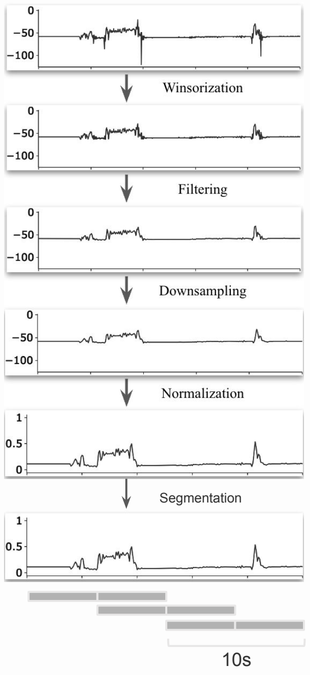
Used Open Datasets

Multimodal open dataset designed to explore affect responses under controlled conditions

Dataset	Signal	Label	# of Ps	Profile Survey
AMIGOS [1] *(2018)*	EEG, ECG, EDA, ACC	Self-report based (Arousal, Valence)	40	Big five inventory, gender, age
ASCERTAIN [2] (2016)	ECG, EDA, ACC	Self-report based (Arousal, Valence)	58	Big five inventory
WESAD [3] (2018)	RESP, ECG, EDA, EMG, TEMP, ACC	Stimulus based (Stress)	15	Gender, age
CASE [4] *(2019)*	ECG, RESP, BVP, EDA, TEMP, EMG	Self-report based (Arousal, Valence)	30	Gender, age
KEmoCon [5] *(2020)*	BVP, EDA, TEMP, ACC	Self-report based (Arousal, Valence)	21	Gender, age

[1] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," IEEE Transactions on Affective Computing, vol. 12, no. 2, pp. 479–493, 2018.
[2] R. Subramanian, J. Wache, M. K. Abadi, R.-L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," IEEE Transactions on Affective Computing, vol. 9, no. 2, pp. 147–160, Nov. 2016.
[3] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in Proceedings of the 20th ACM international conference on multimodal interaction, pp. 400–408, 2018.
[4] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, "A dataset of continuous affect annotations and physiological signals for emotion analysis," Scientific data, vol. 6, no. 1, p. 196, 2019.
[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, no. 1, p. 293, 2020.

Preprocessing: Signal

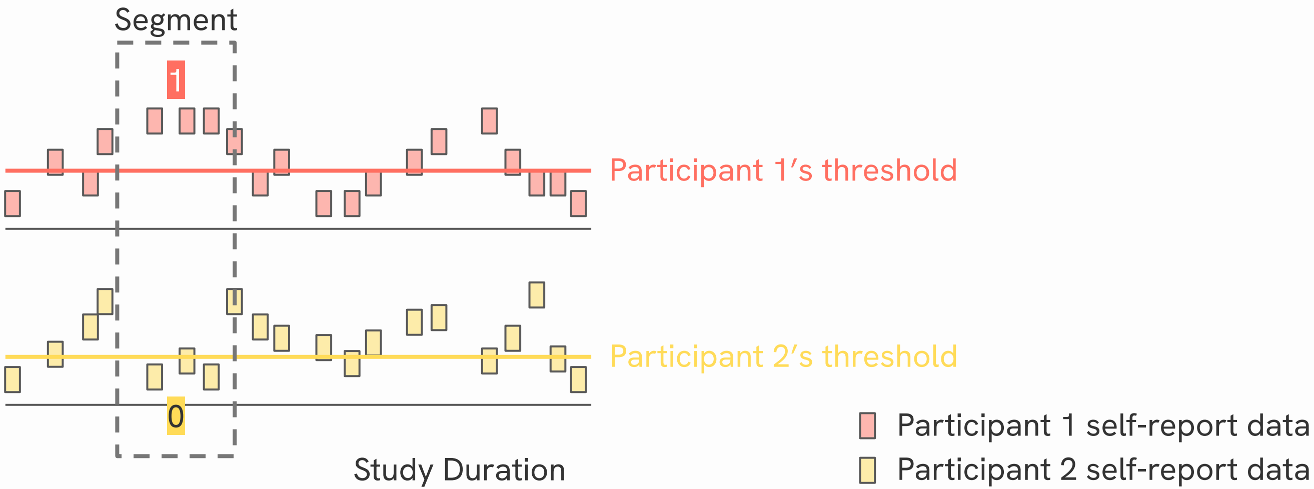


1. **Winsorization**
   - Outliers in the upper and lower 3% range removed
2. **Filtering**
   - Butterworth low-pass filter with a 10 Hz cut-off
3. **Downsampling**
4. **Normalization**
   - Min-max normalization
5. **Segmentation**
   - 10-second window with a 5-second sliding interval

[1] M. Dzie życ, M. Gjoreski, P. Kazienko, S. Saganowski, and M. Gams, "Can we ditch feature engineering? end-to-end deep learning for affect recognition from physiological sensor data," Sensors, vol. 20, no. 22, p. 6535, 2020.

Preprocessing: Labeling

1. WESAD (Stimulus-based labeling)
2. AMIGOS, ASCERTAIN, CASE, KEmoCon (Self-report based labeling)
    - **Participant-specific threshold** for binarization



[1] R. Dai, C. Lu, L. Yun, E. Lenze, M. Avidan, and T. Kannampallil, "Comparing stress prediction models using smartwatch physiological signals and participant self-reports," Computer Methods and Programs in Biomedicine, vol. 208, p. 106207, 2021.
[2] Z. D. King, J. Moskowitz, B. Egilmez, S. Zhang, L. Zhang, M. Bass, J. Rogers, R. Ghaffari, L. Wakschlag, and N. Alshurafa, "Micro-stress ema: A passive sensing framework for detecting in-the-wild stress in pregnant mothers," PACM IMWUT, vol. 3, no. 3, pp. 1–22, 2019.

Processed Signal Segment
Corresponding Affect Label → Affect Detection Models

Non-personalized Model

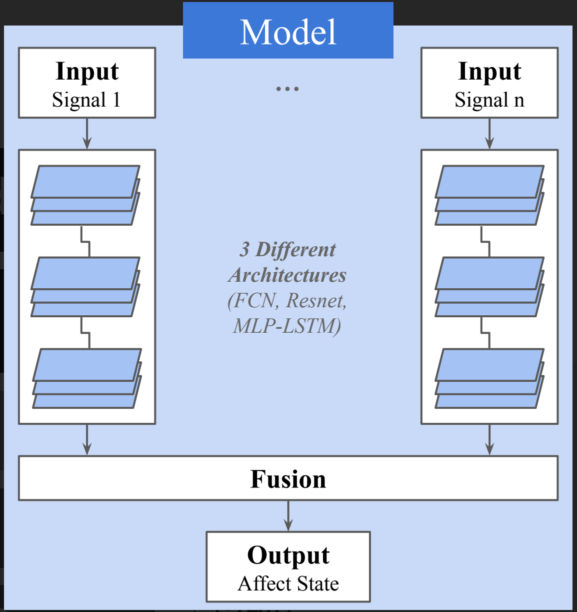
Personalized Models

1. Hybrid
2. Fine Tuning
3. Cluster-specific
4. Multi-task Learning

Main Focus

Three Popular Different DL Architectures

1. **Fully Convolutional Network (FCN)**
   a. n x [CL - CL - CL] - FC
2. **Residual Network (ResNet)**
   a. n x [ResBlock - . . . - ResBlock] - FC
3. **Multi-Layer Perceptron with LSTM (MLP-LSTM)**
   a. n x [FC - . . . - FC - LSTM] - FC



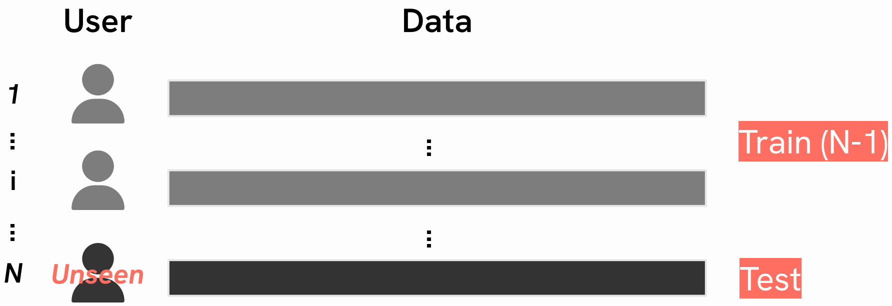
Late fusion: each signal is independently processed and later fused using fully connected layers to generate the final outcome

[1] M. Dzie życ, M. Gjoreski, P. Kazienko, S. Saganowski, and M. Gams, "Can we ditch feature engineering? end-to-end deep learning for affect recognition from physiological sensor data," Sensors, vol. 20, no. 22, p. 6535, 2020.
[2] M. Maithri, U. Raghavendra, A. Gudigar, J. Samanth, P. D. Barua, M. Murugappan, Y. Chakole and U. R. Acharya, "Automated emotion recognition: Current trends and future perspectives," Computer methods and programs in biomedicine, vol. 215, p. 106646, 2022.

n: number of signals, CL: convolutional layer, FC: fully connected layer, ResBlock: residual block with three CLs

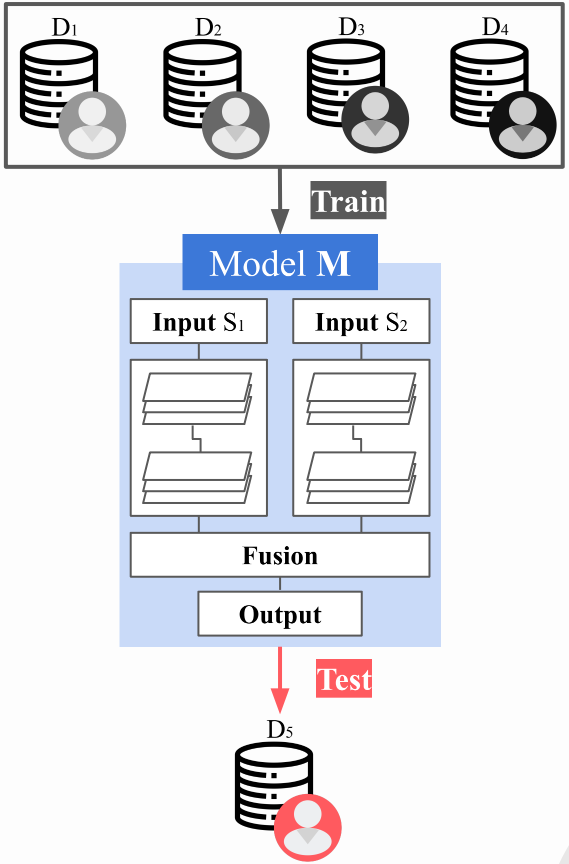
Non-Personalized Model

Leave-one-participant-out (LOPO) Evaluation



↺ iteratively hold out each individual

Example with N=5

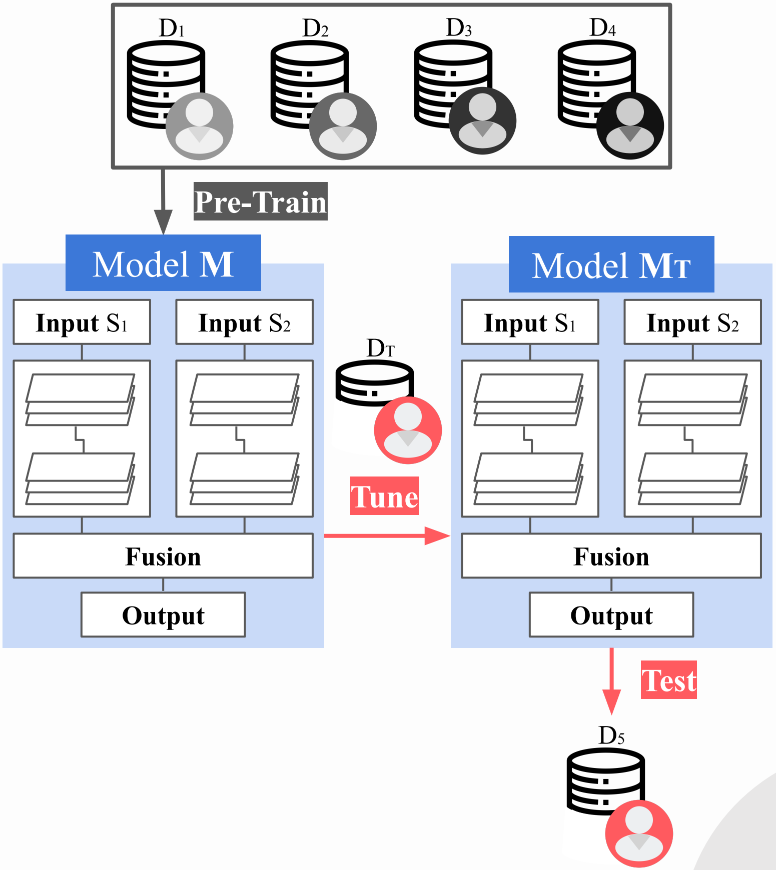


Example with N=5

Personalized Model: Fine Tuning

1. Pre-train network with **N-1** participants
2. Re-train network
   using a small number of target participant data
   - Layers tuned: Entire layers vs. Only the final layer
   - Specific number of data from each label
3. For testing,
   remaining data points of target is used

→ *Repeat for all participants being the target*



Example with N=5

Personalized Model: Fine Tuning

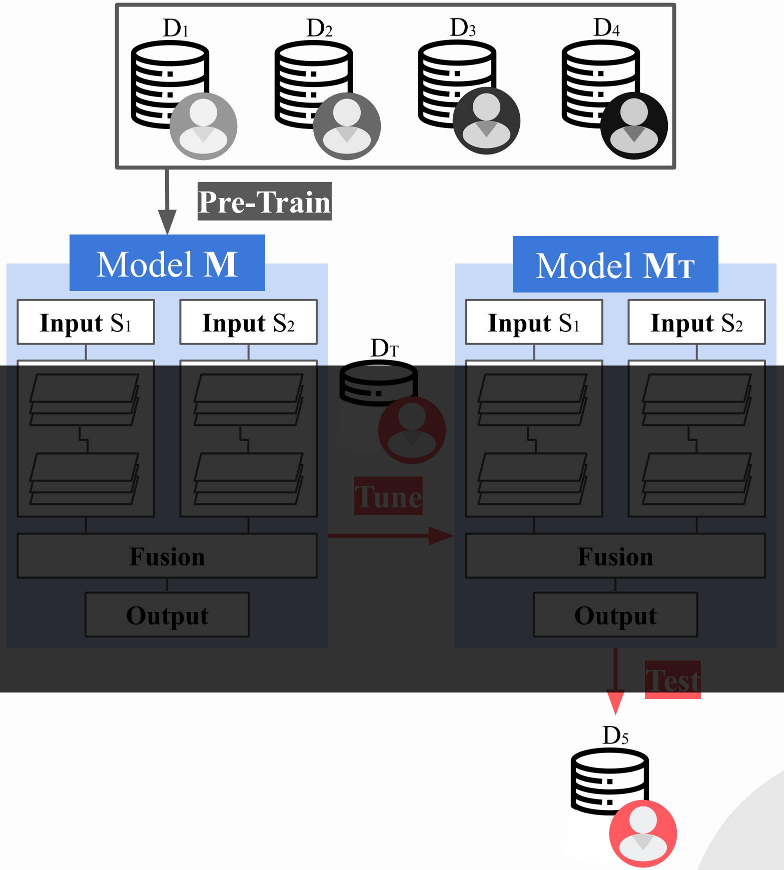
1. Pre-train network with **N-1** participants

2. <u>Re-train network using a small number of target participant data</u>
   - Layers tuned: Entire layers vs. Only the final layer
   - Specific number of data from each label

3. For testing, remaining data points of target is used

→ *Repeat for all participants being the target*

- Layers tuned
  - Entire layers (All) vs. Only the final layer (Last)
- Amount of target data for fine tuning
  - 20%, 30%, 40%, 50% of total data
  - Initial sequence of data points



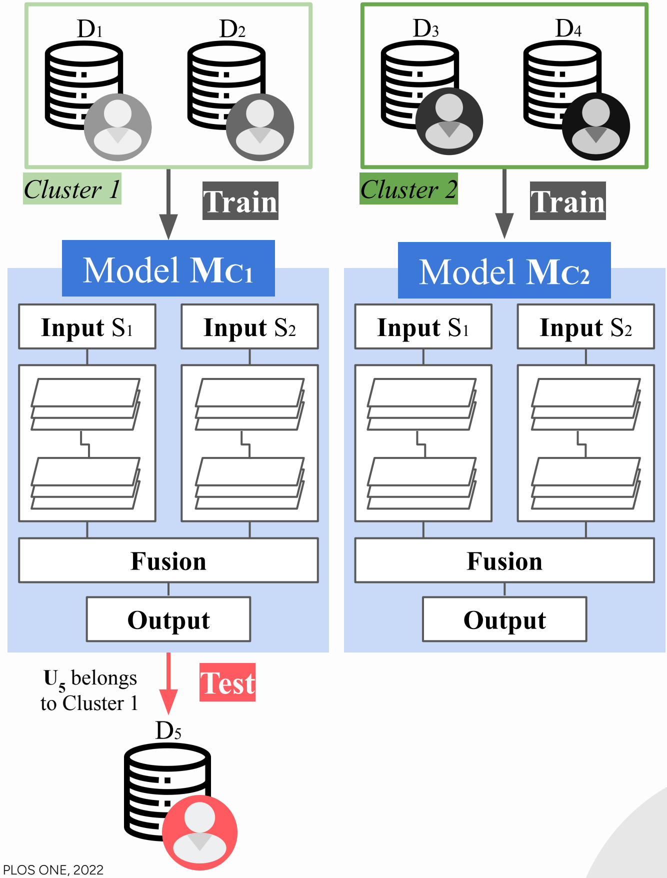
Personalized Model: Cluster Specific

Leveraging a model trained from users similar to the target

1. **K-means clustering**
   using trait information of **N-1 participants**
   - Trait info: Using the demographics or psychological info
2. **Forming distinct model for each cluster**
   - Only use participants within the same cluster to train their respective models
3. **Identifying the target participant's cluster**
   using his/her trait information
4. Corresponding cluster model is used for testing

→ *Repeat for all participants being the target*

Example with N=5



[1] D. A. Adler, F. Wang, D. C. Mohr, and T. Choudhury, "Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies," PLOS ONE, 2022
[2] Y. S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, G. Riva, and C. Ersoy, "Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches," IEEE Access, 2020.

Personalized Model: Cluster Specific

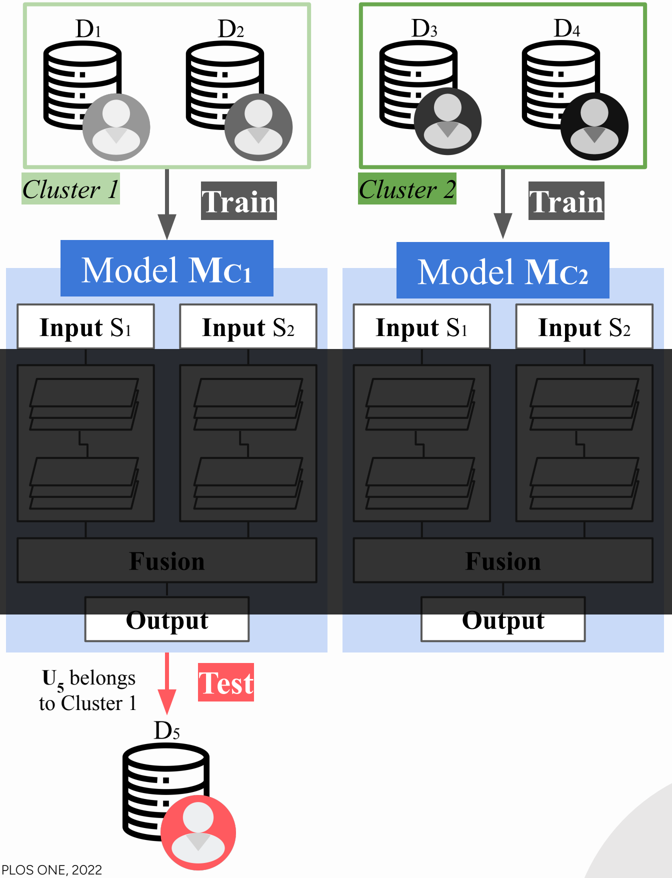
Leveraging a model trained from users similar to the target

1. K-means clustering using trait information of N-1 participants
    - Trait info: Using the demographics or psychological info
2. Forming distinct model for each cluster
    - to train their respective models
3. Identifying the target participant's cluster using his/her trait information
4. Corresponding cluster model is used for testing

→ *Repeat for all participants being the target*

- Impact of varying the number of clusters, K
    - **Fixed** K values: 2 to 5
    - **Dynamically** calculated K values using silhouette score

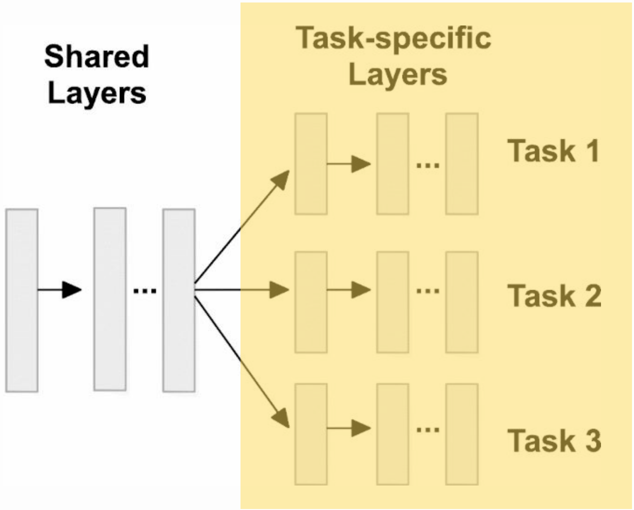
Example with N=5



[1] D. A. Adler, F. Wang, D. C. Mohr, and T. Choudhury, "Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies," PLOS ONE, 2022
[2] Y. S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, G. Riva, and C. Ersoy, "Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches," IEEE Access, 2020.

Personalized Model: Multi-task Learning (MTL)

MTL : simultaneously trains on multiple similar tasks by sharing information between them



General knowledge learning across tasks

Tailored learning for each task

Task definition for personalization:
User-as-task vs.
Cluster-as-task

[1] R. Caruana, "Multitask learning," Machine learning, vol. 28, pp. 41–75, 1997.
[2] B. Li and A. Sano, "Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 4, no. 2, pp. 1–26, 2020.
[3] A. Saeed, T. Ozcelebi, J. Lukkien, J. B. van Erp, and S. Trajanovski, "Model adaptation and personalization for physiological stress detection," in 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 209–216, IEEE, 2018.
[4] H. Yu, E. B. Klerman, R. W. Picard, and A. Sano, "Personalized wellbeing prediction using behavioral, physiological and weather data," in 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), pp. 1–4, IEEE, 2019.
[5] S. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," IEEE Transactions on Affective Computing, vol. 11, no. 2, pp. 200–213, 2017.

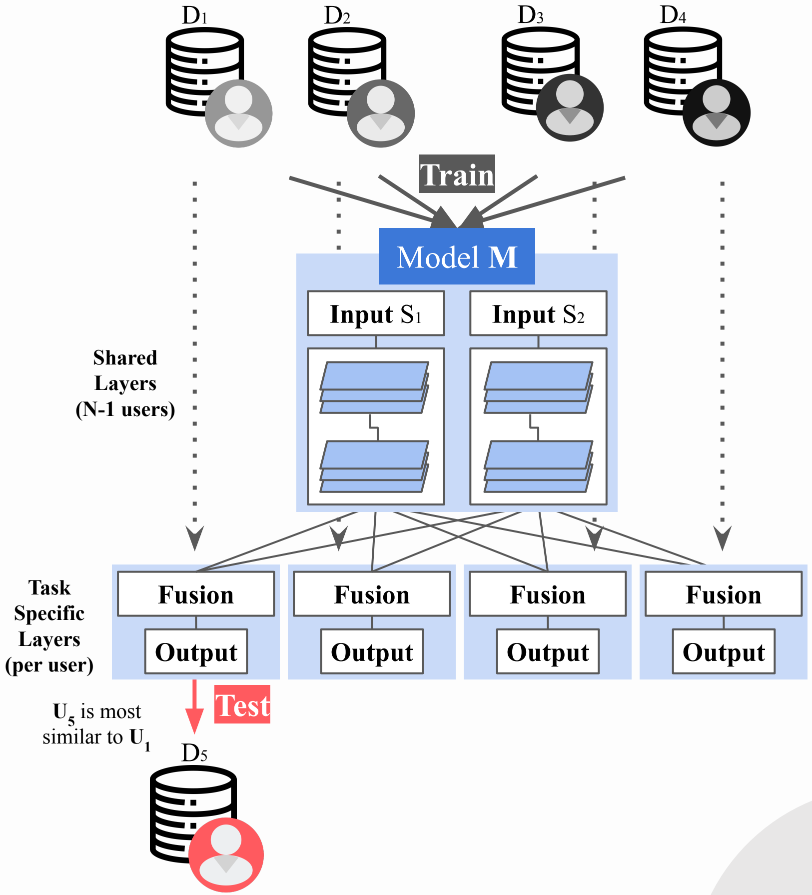
Personalized Model: MTL

User-as-task

1. Train all layers with N-1 participants
   except the last FC layer and the output layer
2. Train the **last FC layer and output layer**
   using each participant's data
3. Find the participant who is the most similar to target
   - Using the demographics or psychological information
4. Corresponding participant's weights
   are used for testing

→ *Repeat for all participants being the target*

Example with N=5

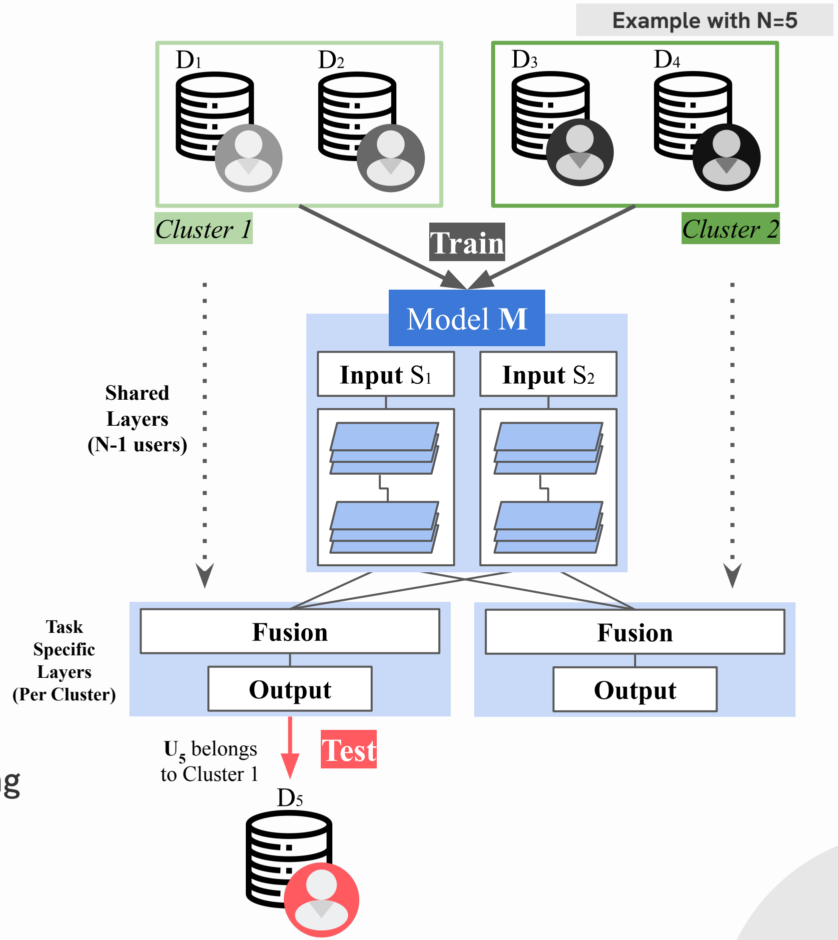


Personalized Model: MTL

Cluster-as-task

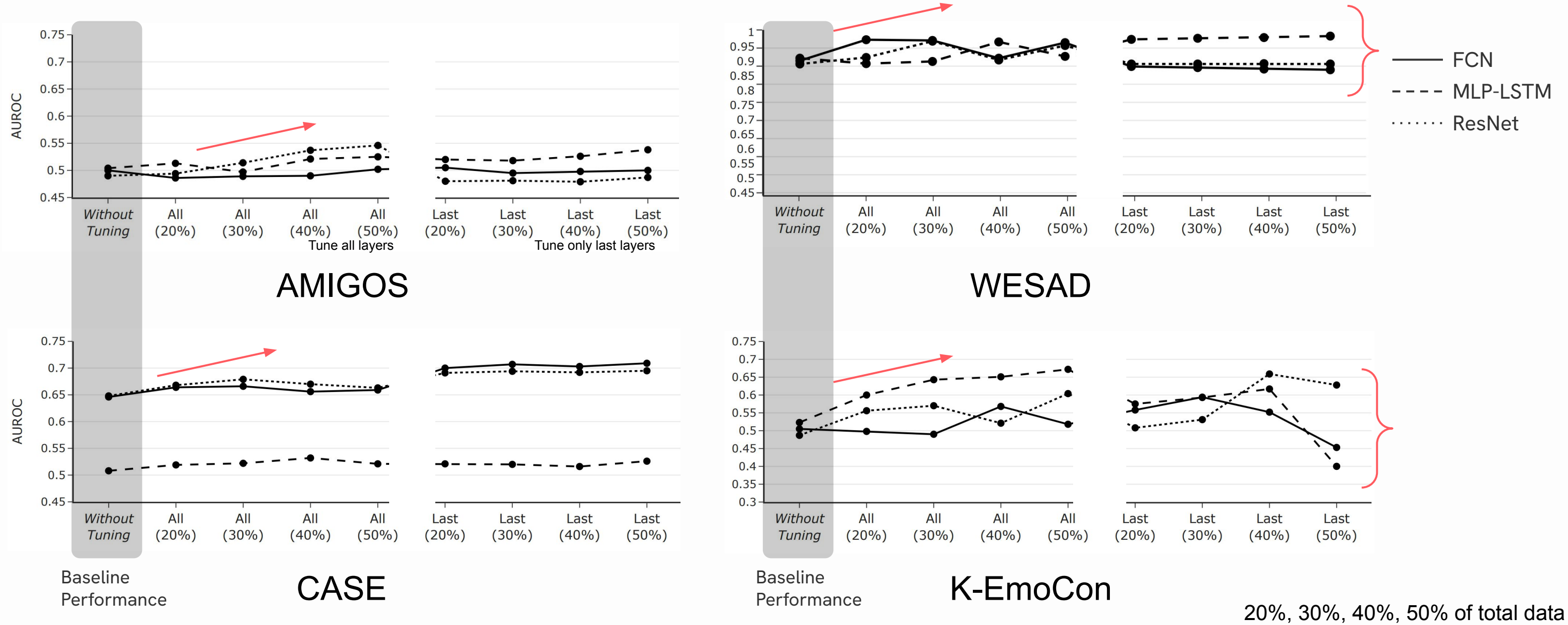
1. **K-means clustering**
   using trait information of N-1 participants
   - Trait info: the demographics or psychological info
   - Determine K using silhouette score
2. Train all layers with N-1 participants
   except the last FC layer and the output layer
3. Train **the last FC layer and output layer**
   using each cluster data
4. Identify the target participant's cluster
   - Using the demographics or psychological info
5. Corresponding cluster's weights are used for testing

→ *Repeat for all participants being the target*



Results - Personalized Model: Fine Tuning

For each dataset-architecture pair, we can find fine-tuned models with higher AUROC
No consistent performance patterns across different deep learning architectures

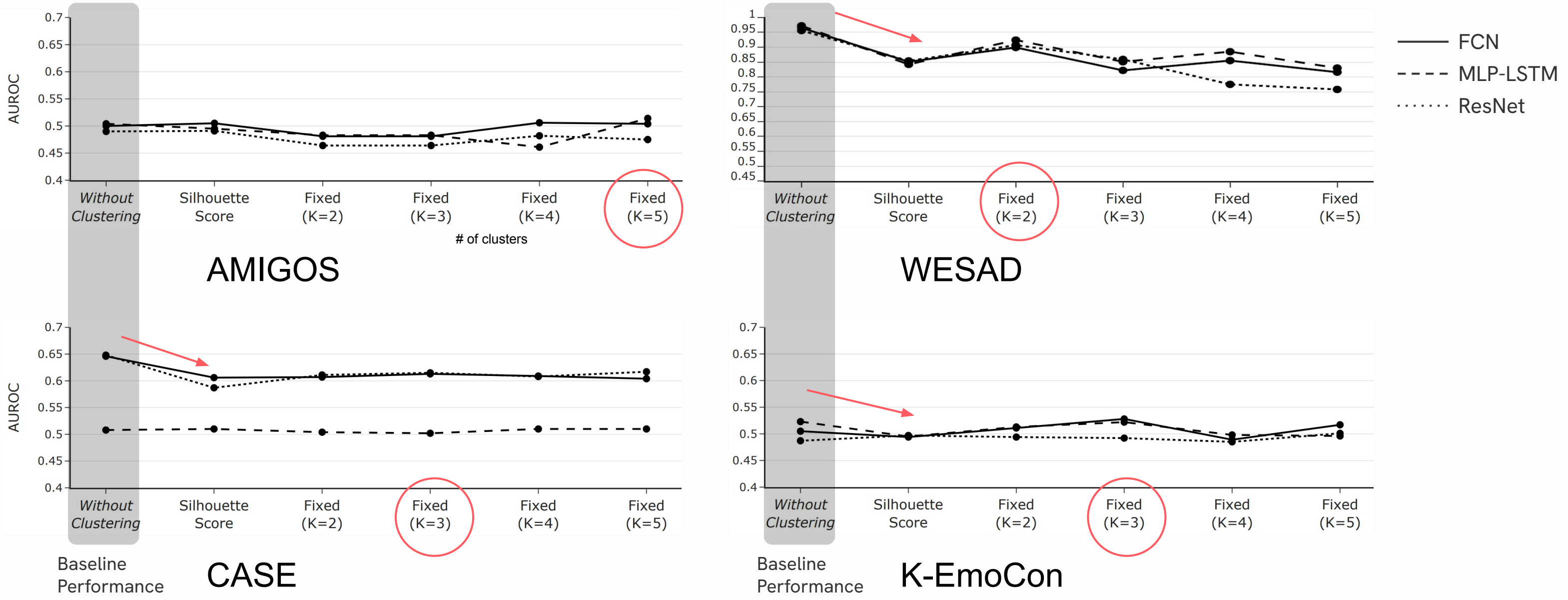


20%, 30%, 40%, 50% of total data
All layers vs. Last layers

Results - Personalized Model: Cluster-specific

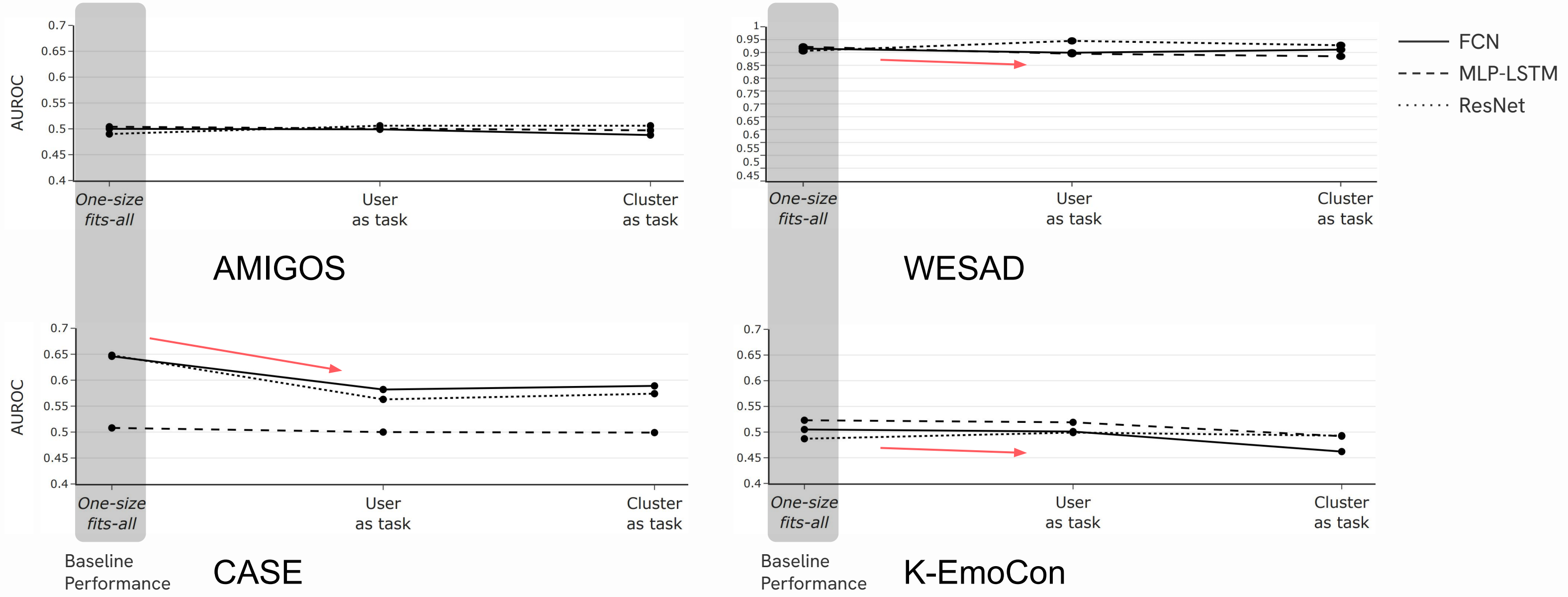
Cluster-specific models mostly show lower AUROC compared to non-personalized one

Cluster-specific models: optimal cluster number (K) varied



Results - Personalized Model: Multi-task Learning

Most cases, multi-task learning models (both user-as-task or cluster-as-task) show lower AUROC compared to non-personalized one



Results - Comparative Evaluation

Dataset	Architecture	Personalization Techniques: Non-Personalized	Fine Tuning	Hybrid	Cluster Specific	Multi-task Learning
AMIGOS (Arousal)	FCN	0.500 (0.100)	0.505 (0.159)	**0.512 (0.122)**	0.506 (0.122)	0.499 (0.038)
	MLP-LSTM	0.504 (0.100)	**0.538 (0.140)**	0.476 (0.107)	0.514 (0.129)	0.500 (0.000)
	ResNet	0.490 (0.115)	**0.546 (0.147)**	0.518 (0.136)	0.521 (0.078)	0.506 (0.044)
AMIGOS (Valence)	FCN	0.518 (0.131)	0.502 (0.159)	0.494 (0.145)	**0.531 (0.125)**	0.499 (0.021)
	MLP-LSTM	0.476 (0.109)	**0.528 (0.132)**†	0.511 (0.142)	0.515 (0.134)	0.500 (0.000)
	ResNet	0.493 (0.106)	**0.546 (0.147)**	0.515 (0.112)	0.513 (0.124)	0.489 (0.058)
ASCERTAIN (Arousal)	FCN	0.511 (0.071)	**0.521 (0.078)**	0.508 (0.067)	0.517 (0.071)	0.502 (0.026)
	MLP-LSTM	0.498 (0.035)	0.513 (0.073)	0.491 (0.056)	**0.517 (0.071)**†	0.500 (0.000)
	ResNet	0.506 (0.070)	0.511 (0.075)	0.505 (0.085)	**0.521 (0.078)**	0.505 (0.028)
ASCERTAIN (Valence)	FCN	0.514 (0.060)	0.515(0.075)	0.505 (0.075)	**0.520 (0.073)**	0.501 (0.009)
	MLP-LSTM	0.496 (0.047)	0.495 (0.060)	0.499 (0.035)	**0.507 (0.073)**	0.500 (0.000)
	ResNet	**0.520 (0.064)**	0.512(0.079)	0.515 (0.066)	0.518(0.082)	0.502 (0.029)
WESAD	FCN	0.915 (0.203)	0.973 (0.089)	**0.976 (0.074)**	0.849 (0.303)	0.911 (0.199)
	MLP-LSTM	0.922 (0.195)	**0.983 (0.053)**	0.913 (0.212)	0.874 (0.266)	0.895 (0.222)
	ResNet	0.906 (0.196)	0.969 (0.076)	**0.979 (0.066)**	0.857 (0.308)	0.945 (0.120)
CASE (Arousal)	FCN	0.646 (0.165)	**0.709 (0.173)**	0.655 (0.197)	0.613 (0.159)	0.589 (0.150)
	MLP-LSTM	0.508 (0.069)	**0.532 (0.105)**	0.520 (0.106)	0.510 (0.100)	0.500 (0.021)
	ResNet	0.648 (0.155)	**0.695 (0.162)**	0.646 (0.168)	0.617 (0.150)	0.574 (0.142)
CASE (Valence)	FCN	0.651 (0.159)	**0.688 (0.203)**	0.655 (0.217)	0.649 (0.132)	0.591 (0.139)
	MLP-LSTM	**0.548 (0.134)**	0.494 (0.038)	0.506 (0.089)	0.543 (0.134)	0.527 (0.094)
	ResNet	0.620 (0.169)	**0.676 (0.176)**	0.651 (0.200)	0.633 (0.154)	0.584 (0.159)
K-EmoCon (Arousal)	FCN	0.505 (0.176)	**0.594 (0.188)**	0.509 (0.358)	0.528 (0.152)	0.501 (0.146)
	MLP-LSTM	0.523 (0.173)	**0.672 (0.369)**	0.650 (0.373)	0.522 (0.172)	0.519 (0.139)
	ResNet	0.487 (0.188)	**0.659 (0.215)***	0.594 (0.312)	0.501 (0.136)	0.499 (0.142)
K-EmoCon (Valence)	FCN	0.507 (0.147)	**0.546 (0.229)**	0.443 (0.202)	0.519 (0.174)	0.534 (0.158)
	MLP-LSTM	0.520 (0.174)	0.619 (0.232)	**0.752 (0.255)***	0.526 (0.154)	0.513 (0.120)
	ResNet	0.508 (0.130)	0.602 (0.295)	**0.643 (0.349)**	0.528 (0.119)	0.523 (0.130)

Discussion - Personalized Model: Fine Tuning

Our Results Significant performance improvement in most cases

- Previous studies also showed improvements
  - Katahen et al. [1]
    - Tuning the last two layers led to an improvement in the performance of depression prediction and forecasting using contextual data
  - Yu et al. [2]
    - Tuning the last two layers required only 10% of data, while tuning the entire model required more than 30% of data to outperform non-personalized models
  - Behinaein et al. [3]
    - Using the WESAD dataset, tuning the entire model with 1%, 5%, and 10\% of individual data increases f1-score by 0.1%, 11.1%, 14.3%, respectively

[1] A. Kathan, M. Harrer, L. K ̈uster, A. Triantafyllopoulos, X. He, M. Milling, M. Gerczuk, T. Yan, S. T. Rajamani, E. Heber, et al., "Personalised depression forecasting using mobile sensor data and ecological momentary assessment," Frontiers in Digital Health, vol. 4, p. 964582, 2022.
[2] H. Yu and A. Sano, "Passive sensor data based future mood, health, and stress prediction: User adaptation using deep learning," in 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 5884–5887, IEEE, 2020.
[3] B. Behinaein, A. Bhatti, D. Rodenburg, P. Hungler, and A. Etemad, "A transformer architecture for stress detection from ecg," in Proceedings of the 2021 ACM International Symposium on Wearable Computers, pp. 132–134, 2021.

Discussion - Personalized Model: Cluster-specific

Our Results No significant performance improvement in most cases

- Previous studies showed mixed findings
  - Can et al. [1]
    - Cluster-specific models based on Perceived Stress Scale (PSS) scores led to an improvement in stress detection performance using physiological data
  - Kathan et al. [2]
    - Gender-based cluster-specific models slightly improved performance in depression prediction and forecasting using contextual data
  - Tervonen et al. [3]
    - Using the WESAD dataset, cluster-specific models showed slightly lower stress detection performance

[1] Y. S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, G. Riva, and C. Ersoy, “Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches,” IEEE Access, vol. 8, pp. 38146–38163, 2020.
[2] A. Kathan, M. Harrer, L. K ̈uster, A. Triantafyllopoulos, X. He, M. Milling, M. Gerczuk, T. Yan, S. T. Rajamani, E. Heber, et al., “Personalised depression forecasting using mobile sensor data and ecological momentary assessment,” Frontiers in Digital Health, vol. 4, p. 964582, 2022.
[3] J. Tervonen, S. Puttonen, M. J. Sillanp ̈a ̈a, L. Hopsu, Z. Homorodi, J. Ker ̈anen, J. Pajukanta, A. Tolonen, A. L ̈ams ̈a, and J. M ̈antyj ̈arvi, “Personalized mental stress detection with self-organizing map: From laboratory to the field,” Computers in Biology and Medicine, vol. 124, p. 103935, 2020.

Discussion - Personalized Model: Cluster Specific

Our Results No significant performance improvement in most cases

Cluster-specific personalization mostly failed to improve classification performance

Possible Explanations

1. Significant reduction of data amount used for training after clustering
2. Differences in finding 'similar' participants to the target
   - *Can et al. : stress scores → stress detection model*
   - *Ours : age and gender → arousal and stress detection model*

[1] Y. S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, G. Riva, and C. Ersoy, "Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches," IEEE Access, vol. 8, pp. 38146–38163, 2020.
[2] A. Kathan, M. Harrer, L. K üster, A. Triantafyllopoulos, X. He, M. Milling, M. Gerczuk, T. Yan, S. T. Rajamani, E. Heber, et al., "Personalised depression forecasting using mobile sensor data and ecological momentary assessment," Frontiers in Digital Health, vol. 4, p. 964582, 2022.
[3] J. Tervonen, S. Puttonen, M. J. Sillanp ä ä, L. Hopsu, Z. Homorodi, J. Ker änen, J. Pajukanta, A. Tolonen, A. L äms ä, and J. M äntyj ärvi, "Personalized mental stress detection with self-organizing map: From laboratory to the field," Computers in Biology and Medicine, vol. 124, p. 103935, 2020.

Discussion - Personalized Model: Multi-task Learning

Our Results No significant performance improvement in most cases

- But previous studies reported improvements
  - Saeed et al. [1]
    - Personalized stress detection model using physiological data and a user-as-task MTL models showed an average increase of 2.87% in AUROC
  - Yu et al. [2]
    - Personalized wellbeing detection using physiological, behavioral, and contextual data along with user-as-task and cluster-as-task MTL CNN and LSTM models increased f1-score with an average of 9.83%
  - Taylor et al. [3]
    - Cluster-as-task models on wellbeing detection showed an increase in AUROC values ranging from 11% to a maximum of 21%

[1] A. Saeed, T. Ozcelebi, J. Lukkien, J. B. van Erp, and S. Trajanovski, "Model adaptation and personalization for physiological stress detection," in 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 209–216, IEEE, 2018.
[2] H. Yu, E. B. Klerman, R. W. Picard, and A. Sano, "Personalized wellbeing prediction using behavioral, physiological and weather data," in 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), pp. 1–4, IEEE, 2019.
[3] S. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," IEEE Transactions on Affective Computing,vol. 11, no. 2, pp. 200–213, 2017.

Discussion - Personalized Model: Multi-task Learning

Our Results No significant performance improvement in most cases

- But previous studies reported improvements

Significant Difference

- Previous studies: User-dependent training & evaluation for MTL
  ↔ Ours: User-independent training & evaluation for MTL (target user's data were not used for training)

But Li & Sano (2020) showed significant improvements even in user-independent setting

- *Li & Sano (2020): wellbeing prediction (mood, health, stress) clustering based on gender and personality information, with a large number of of participants (N=239)*
- *With a larger dataset, it was possible to* find 'similar' participants to target

[1]
- Personalized stress detection model using physiological data and a user-as-task MTL models showed an average
- Personalized wellbeing detection using physiological, behavioral, and contextual data along with user-as-task and cluster-as-task MTL CNN and LSTM models increased f1-score with an average of 9.83%

Taylor et al. [3]
- Cluster-as-task models on wellbeing detection showed an increase in AUROC values ranging from 11% to a

[1] A. Saeed, T. Ozcelebi, J. Lukkien, J. B. van Erp, and S. Trajanovski, "Model adaptation and personalization for physiological stress detection," in 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 209–216, IEEE, 2018.
[2] H. Yu, E. B. Klerman, R. W. Picard, and A. Sano, "Personalized wellbeing prediction using behavioral, physiological and weather data," in 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), pp. 1–4, IEEE, 2019.
[1] B. Li and A. Sano, "Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress," PACM IMWUT vol. 4, no. 2, pp. 1–26, 2020.

Takeaways in Personalized Affective Computing

Result #1: Fine-tuning worked well *(but requiring **some use of unseen target users' labels**)*

* **How to adaptively find the optimal label amount necessary for effective personalization?**
* **How will other domain adaptation techniques (e.g., few-shot learning) work in general?**

Result #2: Cluster-specific or multi-task learning failed in user-independent setting

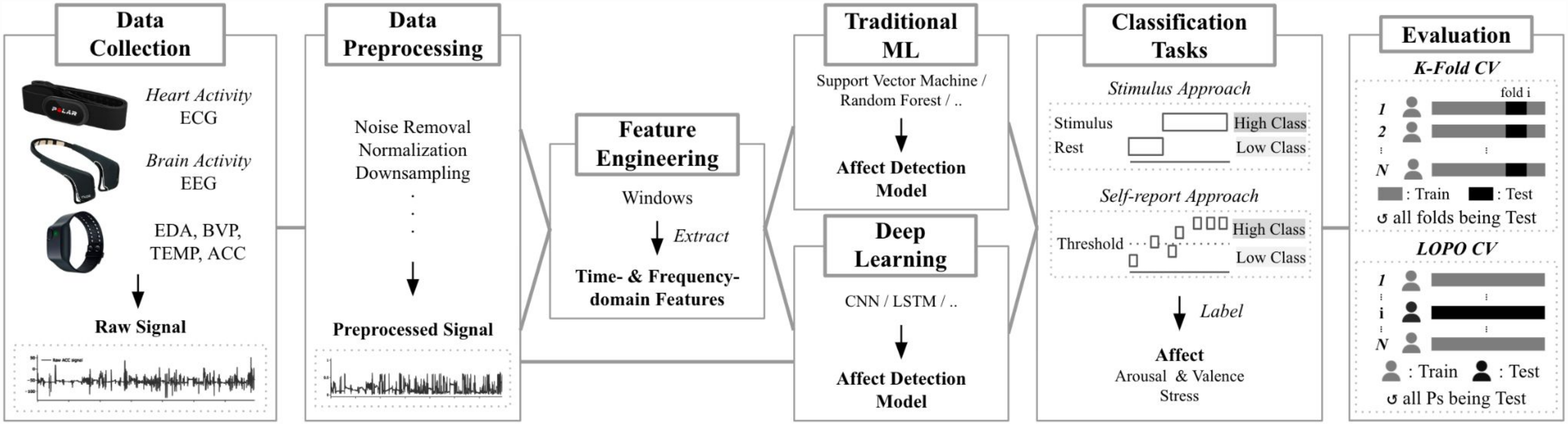
* **What are the better approaches to find "similar users" to the target users?**
  - **Trait-driven** (current): demographics or psychological traits
  - **Data-driven:** similarity in data, or **hybrid** (trait + data) towards **domain generalization**?
* **Will dataset scaling (increasing # participants) work?** (but requires large-scale open datasets)

Systematic Evaluation of Personalized Models for Affective Computing

Yunjo Han, Panyu Zhang, Minseo Park, Uichin Lee
Interactive Computing Lab, KAIST

https://github.com/Kaist-ICLab/Personalized_Affective_Computing

General Process of Affect Recognition Systems



Another Research Gap

Only 1 or 2 datasets for evaluation

- Unpublished

Lack of analysis code and detailed descriptions

→ Evaluation using Multiple Open Datasets

→ Openly sharing Evaluation Process

Reproducibility
- Releasing dataset and code
- Cross-dataset evaluation of models

[1] M. B. McDermott, S. Wang, N. Marinsek, R. Ranganath, L. Foschini, and M. Ghassemi, "Reproducibility in machine learning for health research: Still a ways to go," Science Translational Medicine, vol. 13, no. 586, p. eabb1655, 2021.

Research Direction

Open datasets

- Controlled setting
- Rich in physiological and behavioral signal data

1. Uniform data preprocessing
   a. End-to-end learning for deep learning models

2. Build non-personalized (i.e., one-size-fits-all) and personalized affect recognition models

3. Compare performances
   a. Evaluate the efficacy of each personalization technique across datasets

Publicly available

Personalized Model: Cluster Specific

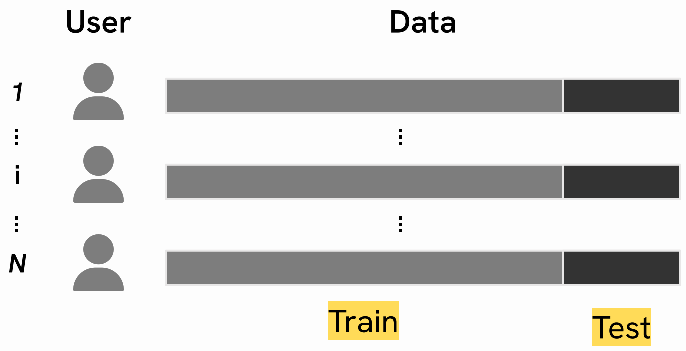
A group of 'similar' users that the target belongs to ⇒ leverage trained models from similar users

- Defining similar users based on **demographics or psychological information**
  - Age, gender, personality traits
- K-Means Clustering
  - Value of K (= # clusters)
    - Fixed value
    - Highest mean silhouette score

[1] D. A. Adler, F. Wang, D. C. Mohr, and T. Choudhury, "Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies," Plos one,vol. 17, no. 4, p. e0266516, 2022
[2] Y. S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, G. Riva, and C. Ersoy, "Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches," IEEE Access, vol. 8, pp. 38146–38163, 2020.
[3] A. Kathan, M. Harrer, L. K ̈uster, A. Triantafyllopoulos, X. He, M. Milling, M. Gerczuk, T. Yan, S. T. Rajamani, E. Heber, et al., "Personalised depression forecasting using mobile sensor data and ecological momentary assessment," Frontiers in Digital Health, vol. 4, p. 964582, 2022.
[4] J. Tervonen, S. Puttonen, M. J. Sillanp ̈a ̈a, L. Hopsu, Z. Homorodi, J. Ker ̈anen, J. Pajukanta, A. Tolonen, A. L ̈ams ̈a, and J. M ̈antyj ̈arvi, "Personalized mental stress detection with self-organizing map: From laboratory to the field," Computers in Biology and Medicine, vol. 124, p. 103935, 2020.
[5] B. Li and A. Sano, "Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 4, no. 2, pp. 1–26, 2020.

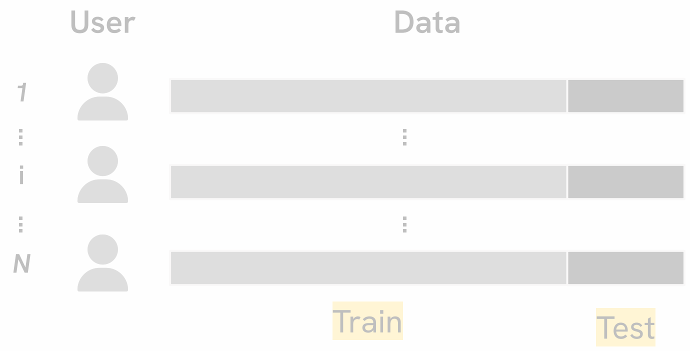
Research Direction

User-dependent approach

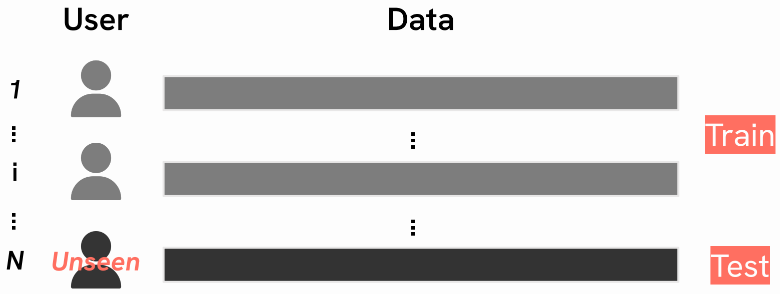


Research Direction

User-dependent approach



User-independent approach



↺ iteratively hold out each individual

Building User-independent Personalized Models

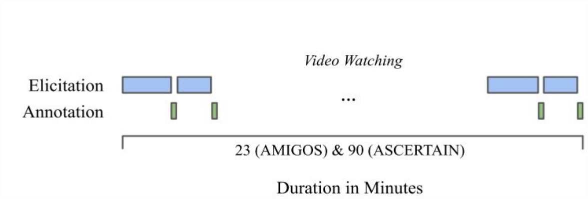
Assuming "similar people or groups"

Fine tuning	Cluster specific	Multi-task learning
Fine tuning using unseen user's data	Building separate models for each group (e.g., gender, personality)	Building a unified multi-task model (e.g., user/cluster as a task)

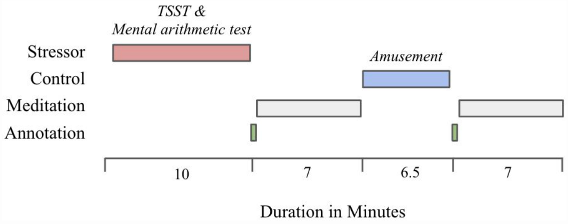
Used Open Datasets

Multimodal open dataset designed to explore affect responses under controlled conditions

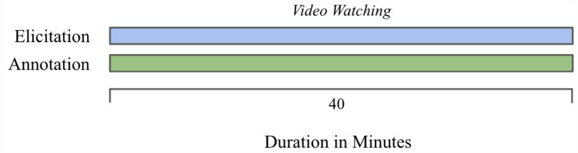
AMIGOS & ASCERTAIN



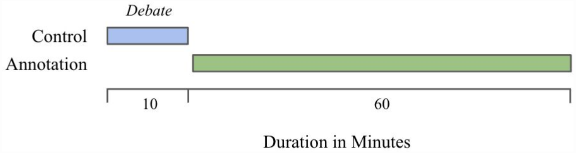
WESAD



CASE



KEmoCon



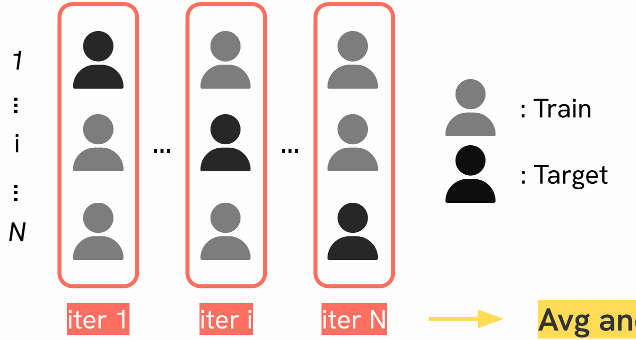
Evaluation

- Use of fixed hyperparameters
  - Referring to previous paper on DL for time series classification [1]
- Metrics
  - Accuracy
  - Macro f1-score
  - AUROC
    - Used mainly for comparing the performance [2]

[1] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," Data mining and knowledge discovery, vol. 33, no. 4, pp. 917–963, 2019.
[2] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," International journal of data mining & knowledge management process, vol. 5, no. 2, p. 1, 2015.

Results

- Iterative testing



→ Avg and Std calculated × 3 → Mean reported

- Repetition with different random seeds
  - Mean of results are reported

Overview

1. Non-personalized model
2. Each of 3 personalization techniques
3. Compare personalized models against non-personalized