KAIST

Interactive Computing Laboratory

UBICOMP ISWC 2024 MELBOURNE, AUSTRALIA

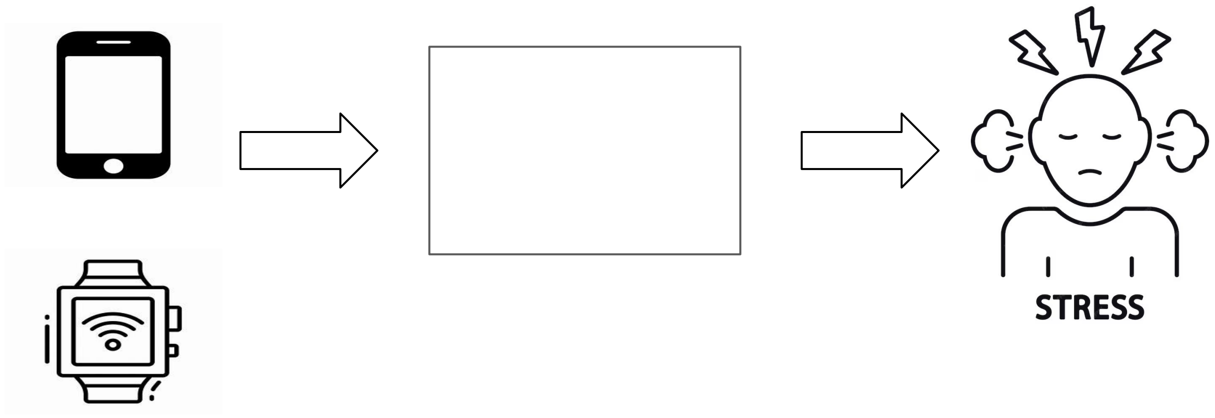
HumblebeeAI

A Reproducible Stress Prediction Pipeline Using Mobile Sensor Data

Panyu Zhang, Gyuwon Jung, Jumabek Alikhanov, Uzair Ahmed, Uichin Lee

Stress Prediction Using Mobile Sensor Data

Mobile & Wearable Data | Model "**X**" | Stress Level



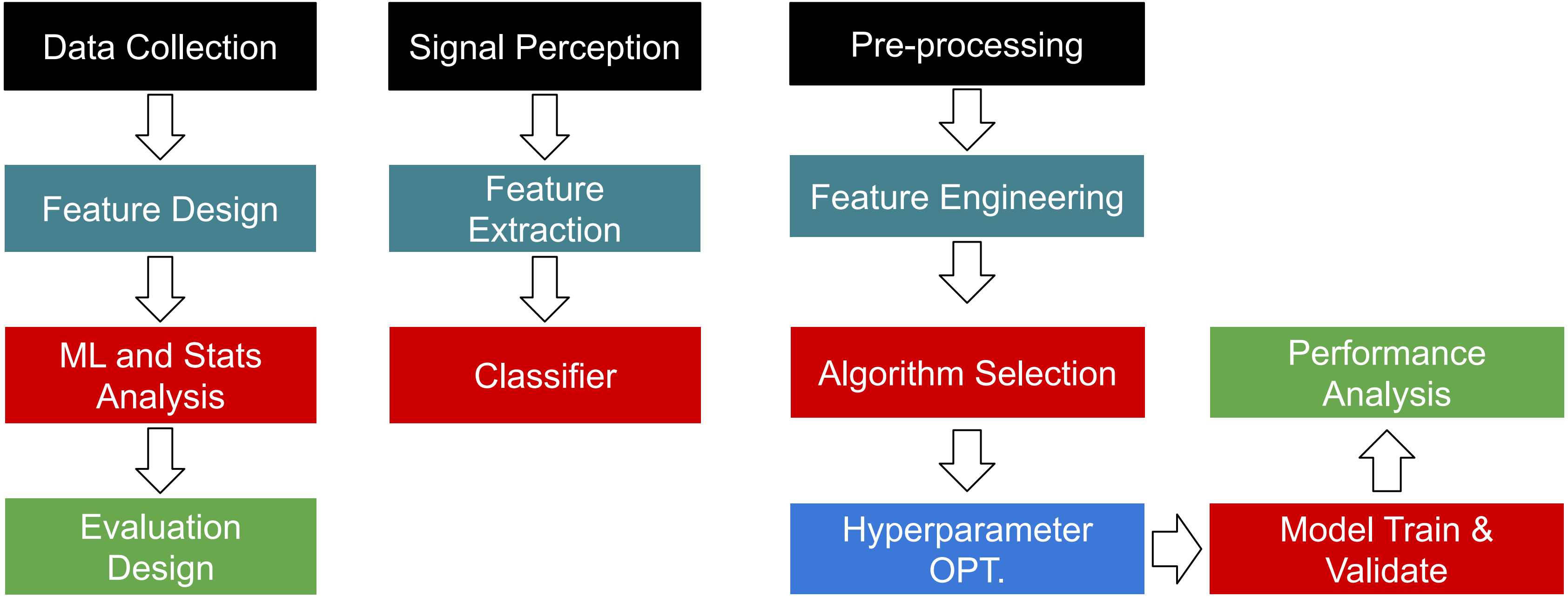
https://www.iconfinder.com/icons
https://creazilla.com/media/icon

Challenges

1. Lack of details in common pipeline for mobile stress prediction research

2. Difficulty in reproducing the results even on the same dataset

Challenge 1. Variety of Machine Learning Pipelines



Fukazawa et al., 2020 Yang et al., 2023 Vos et al., 2023

Y. Fukazawa, N. Yamamoto, T. Hamatani, K. Ochiai, A. Uchiyama, and K. Ohta. 2020. Smartphone-based Mental State Estimation: A Survey from a Machine Learning Perspective. Journal of Information Processing 28, 3 (2020), 650–669.
Kangning Yang, Benjamin Tag, Chaofan Wang, Yue Gu, Zhanna Sarsenbayeva, Tilman Dingler, Greg Wadley, and Jorge Goncalves. 2023. Survey on Emotion Sensing Using Mobile Devices. IEEE Transactions on Affective Computing 14, 4 (2023), 2678–2696.
Gideon Vos, Kelly Trinh, Zoltan Sarnyai, and Mostafa Rahimi Azghadi. 2023. Generalizable Machine Learning for Stress Monitoring from Wearable Devices: A Systematic Literature Review. International Journal of Medical Informatics 173 (May 2023)

Challenge 2. Reproducibility

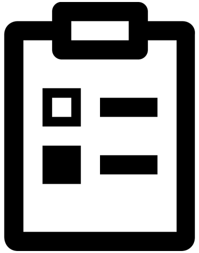
There are four types of reproducibility in this field. (Albertoni et al. 2023)

Code & Analysis		Dataset: Same	Dataset: Different
	Same	• Computational reproducibility • Method reproducibility • Experiment reproducibility • **Reproducibility**	• Replicability • **Generalizability**
	Different	• **Independent reproducibility** • Robustness • Data reproducible	• **Replicability** • Generalizable • Conceptual replicable

Albertoni et al. Reproducibility of Machine Learning: Terminology, Recommendations and Open Issues. arXiv preprint arXiv:2302.12691 (2023).

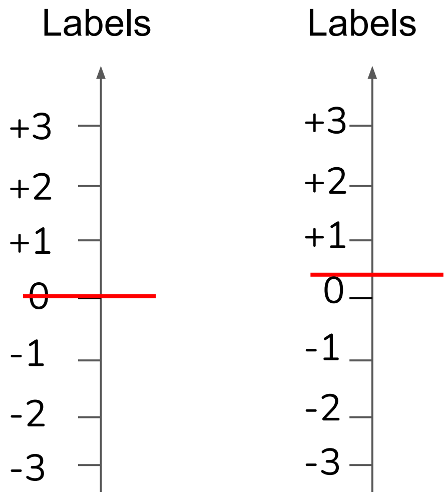
Challenge 2. Reproducibility

Even on the same dataset, many factors in the data analysis pipeline may influence the reproduced results.



My stress level right before doing this survey was
Q: not stressed at all (-3) ~ very stressed (+3)
(Kang et al.)

Self-reported Survey



Binarize Using **Mid Value**

Binarize Using **Mean Value**

Kang et al. 2023. K-EmoPhone: A Mobile and Wearable Dataset with In-Situ Emotion, Stress, and Attention Labels. Scientific Data 10 (2023).

Research Questions

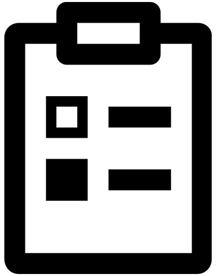
RQ1 What is the **common pipeline** for stress prediction using mobile sensor data?

RQ2 What is the **impact of each factor** in a stress prediction pipeline on the final performance using a public dataset? (***Independent Reproducibility***)

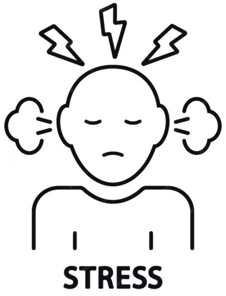
Scope of this Study

Despite a decade of efforts in this field, the performance of **in-the-wild, self-reported stress prediction in user-independent settings** remains limited.

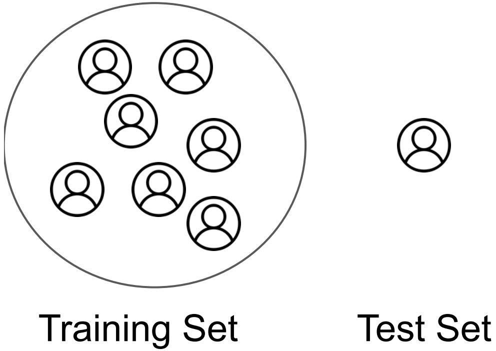
In-the-wild Study



Self-report Surveys



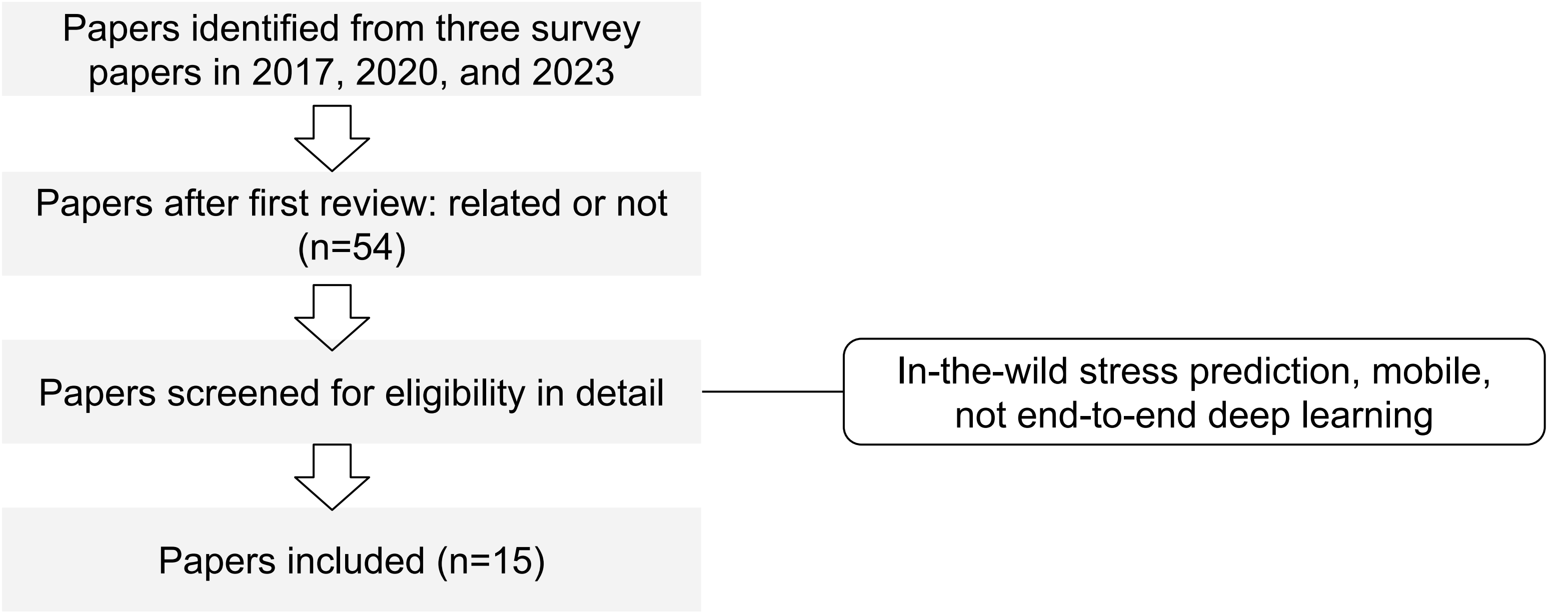
Stress Prediction



User-independent Settings (example: Leave-One-Subject-Out)

https://www.shutterstock.com/zh/search/daily-life-icon
https://www.freepik.com/premium-vector/stress-line-icon-symbol-anxiety-anxiety-headache-anger-sadness_46964960.htm

RQ1 Common Pipeline - Literature Review



Y. Fukazawa, N. Yamamoto, T. Hamatani, K. Ochiai, A. Uchiyama, and K. Ohta. 2020. Smartphone-based Mental State Estimation: A Survey from a Machine Learning Perspective. Journal of Information Processing 28, 3 (2020), 650–669.

Gideon Vos, Kelly Trinh, Zoltan Sarnyai, and Mostafa Rahimi Azghadi. 2023. Generalizable Machine Learning for Stress Monitoring from Wearable Devices: A Systematic Literature Review. International Journal of Medical Informatics 173 (May 2023)

Kangning Yang, Benjamin Tag, Chaofan Wang, Yue Gu, Zhanna Sarsenbayeva, Tilman Dingler, Greg Wadley, and Jorge Goncalves. 2023. Survey on Emotion Sensing Using Mobile Devices. IEEE Transactions on Affective Computing 14, 4 (2023), 2678–2696.

1. Preprocessing

a.Remove invalid survey samples
- **a.1** Remove expiratory
- **a.2** Removing neutral

b.Remove invalid users
- **b.1** Remove users with too few survey labels
- **b.2** Remove users with extreme label distribution

c.Label encoding
- **c.1** Theoretical threshold
- **c.2** Statistical threshold for all users
- **c.3** Statistical threshold for each user

2. Feature Extraction

a.Feature type
- **a.1** Sensor data
- **a.2** Survey data
  - **a.2.1** Participant information
  - **a.2.2** EMA context data
  - **a.2.3** Previous EMA labels

b.Time window
- **b.1** Current (last value before label)
- **b.2** Immediate past (fixed time window)
- **b.3** Extended past (daily)
  - **b.3.1** Epoch window
  - **b.3.2** Whole time window

3. Feature Preparation

a.Feature normalization
- **a.1** For all users (the statistics measure such as mean and std is calculated from training set)
- **a.2** For each user

b.Impute missing values

4. Feature Selection

a. Feature selection methods
- **a.1** Filter methods
- **a.2** Wrapper methods
- **a.3** Embedded methods

5. Data Splitting

a.User-independent cross validation
- **a.1** Leave one subject out
- **a.2** Group k-fold cross validation

b.User-dependent cross validation
- **b.1** K-fold cross validation
- **b.2** Time series k-fold

c. Partial personalization
- **c.1** Random
- **c.2** Stratified
- **c.3** Time series

6. Over/Undersampling

a.Oversample the minority class or undersample the majority class
- **a.1** Original Distribution
- **a.2** Random oversampling
- **a.3** Random undersampling
- **a.4** SMOTE/SMOTE-NC

7. Model Training

a.Personalized vs generalized
- **a.1** Fully personalized (only using single user's data)
- **a.2** Similar-user model (only using similar user group's data)
- **a.3** Multi-task learning
- **a.4** Generalized model

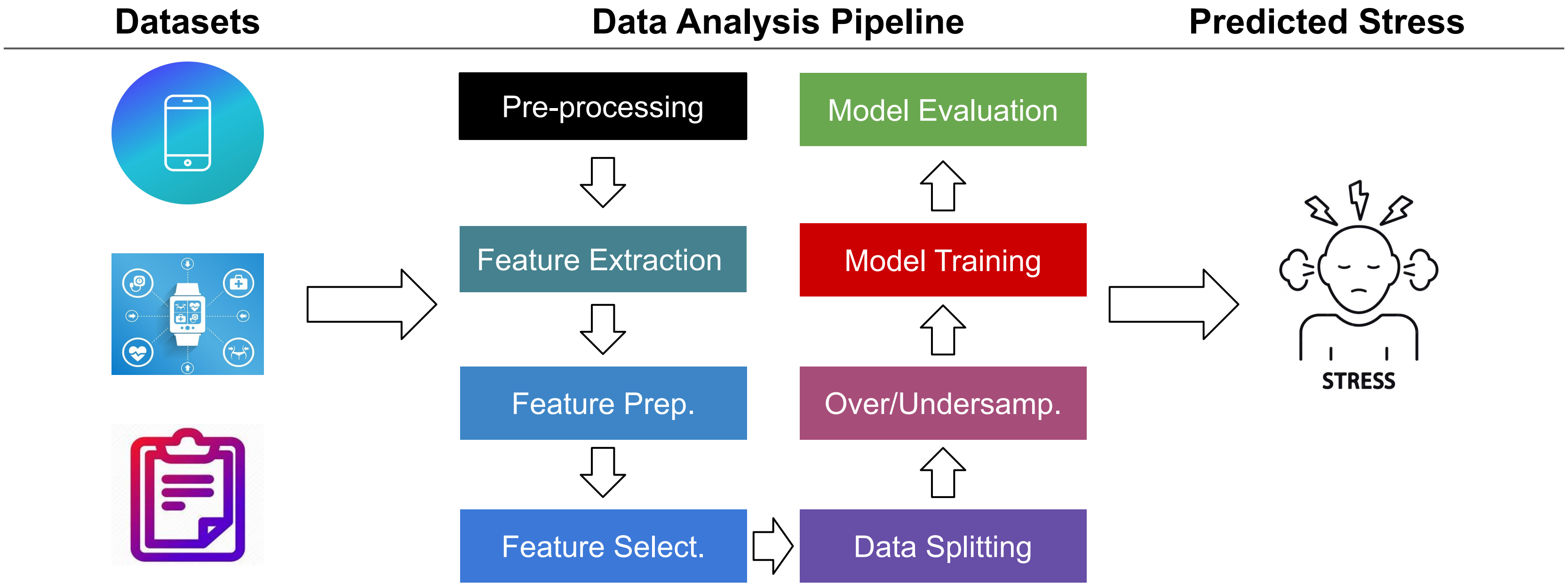
b.Model selection
- **b.1** Traditional machine learning models (**b.1.1** Gradient boosting, **b.1.2** RandomForest, **b.1.3** SVM, **b.1.4** logistic regression, **b.1.5** KNN, **b.1.6** decision tree, and **b.1.7** Naïve Bayes classifier)
- **b.2** Neural network models (i.e. MLP)

8. Model Evaluation

a.Metric selection
- **a.1** Accuracy
- **a.2** F1 score (positive)
- **a.3** macro F1 score
- **a.4** AUC-ROC
- **a.5** precision (PPV)
- **a.6** recall

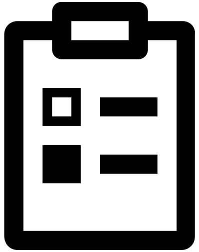
RQ2 Independent Reproducibility

What if we change one factor in the following pipeline?

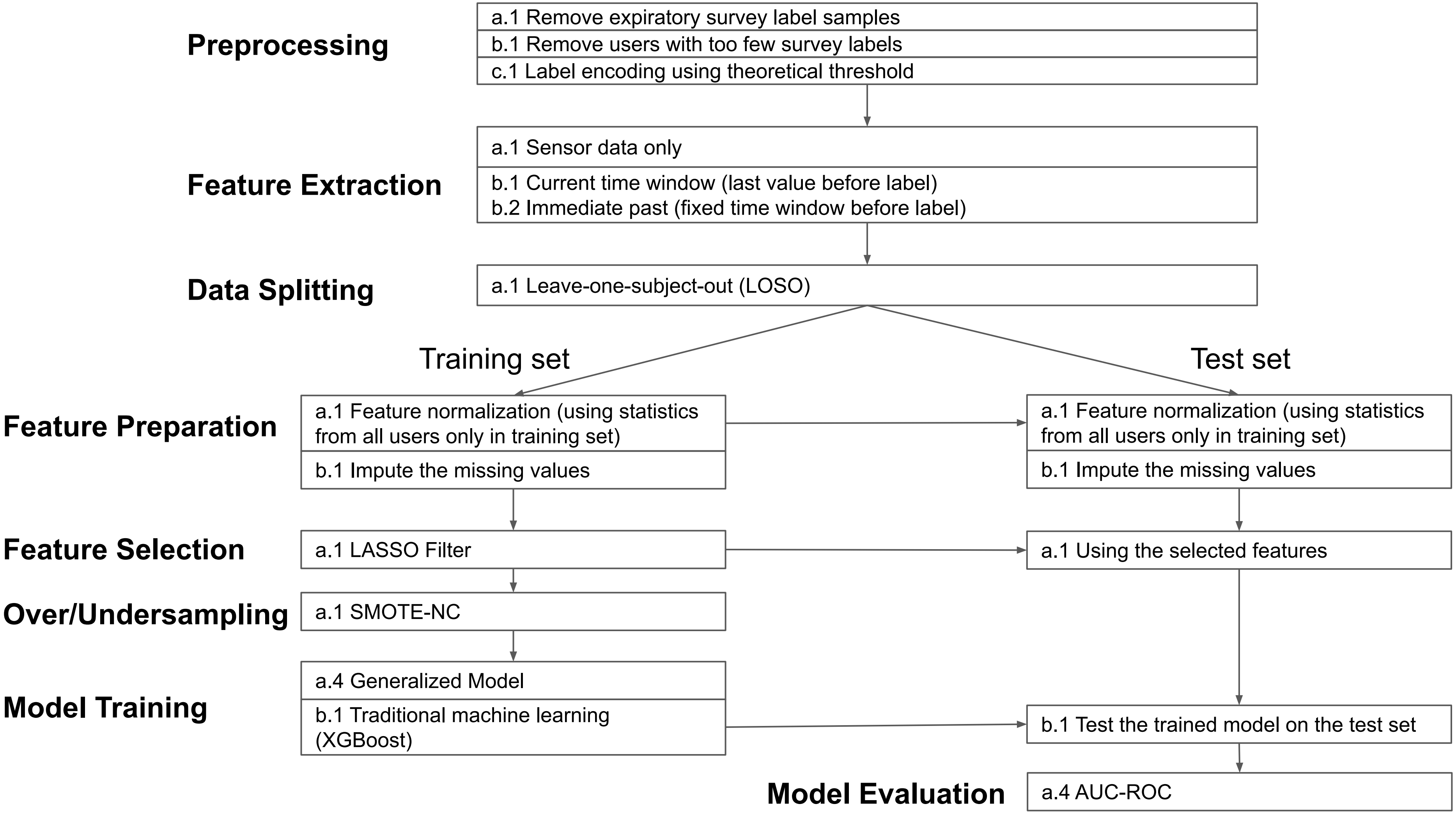


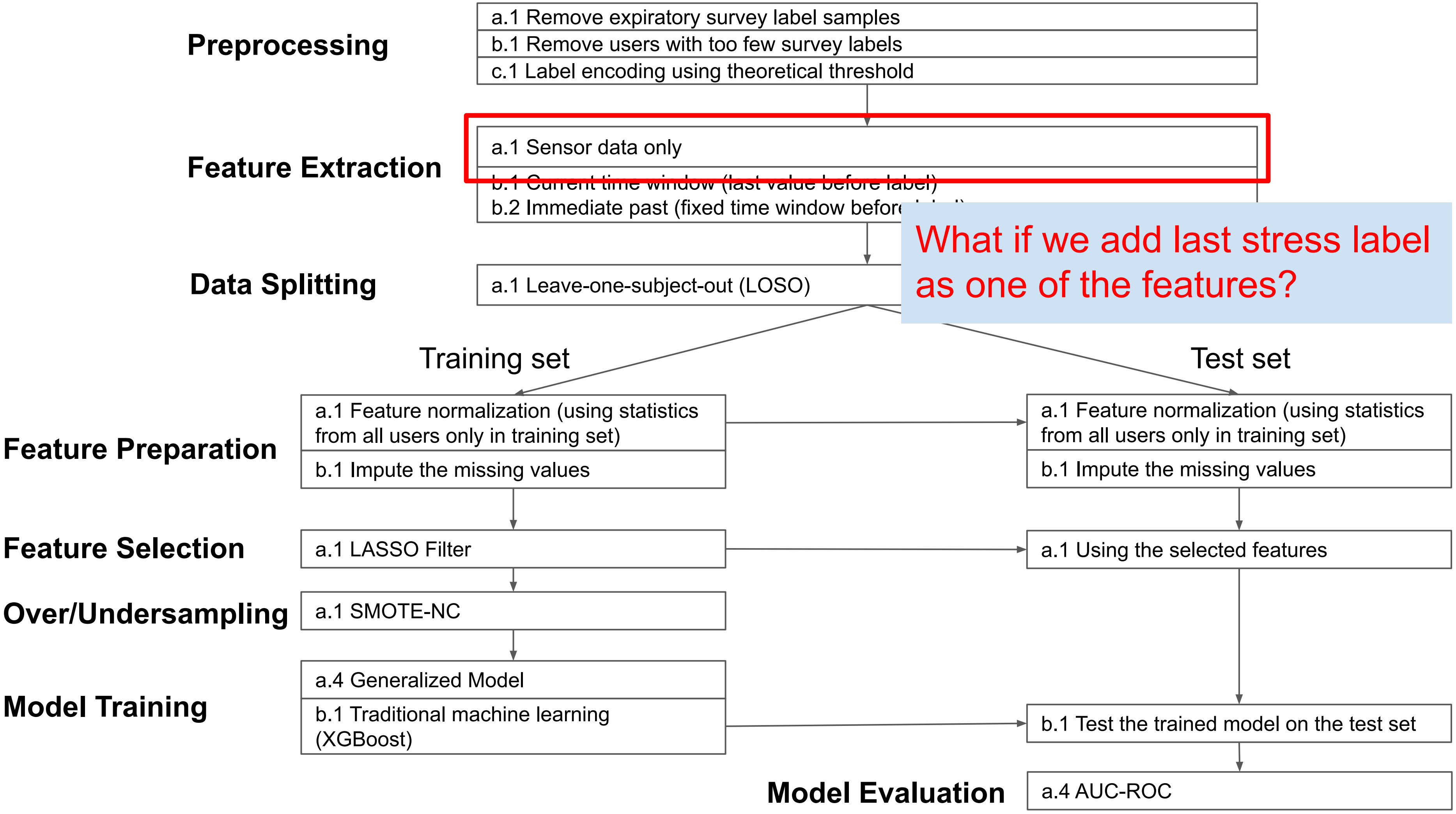
Datasets

Dataset	Duration	#Users	Feature Types	Freq. of Labels	Year
K-EmoPhone	1 week	77	Mobile and wearable sensor data, pre- and post- surveys	10 surveys per day	2023
DeepStress	6 weeks	24	Mobile sensor data, pre-survey	Avg. 4.9 surveys per day	2024

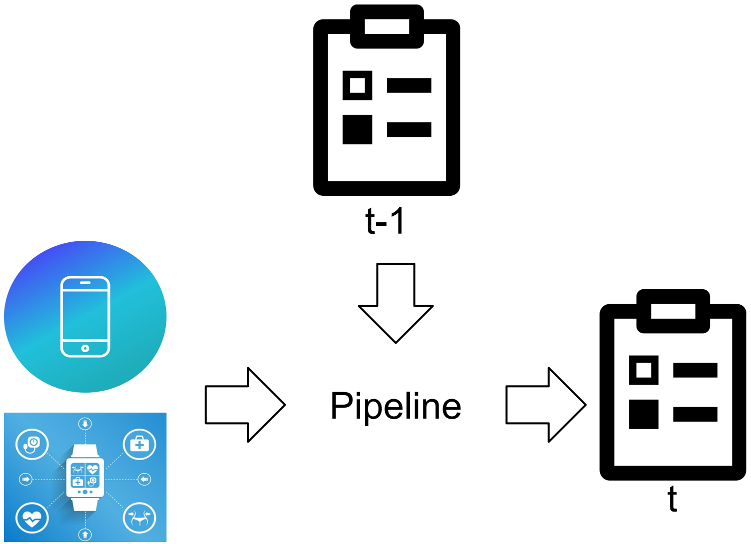


Gyuwon Jung, Sangjun Park, and Uichin Lee. 2024. DeepStress: Supporting Stressful Context Sensemaking in Personal Informatics Systems Using a Quasi-experimental Approach. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1000, 1–18. https://doi.org/10.1145/3613904.3642766


Preprocessing
a.1 Remove expiratory survey label samples
b.1 Remove users with too few survey labels
c.1 Label encoding using theoretical threshold
Feature Extraction
a.1 Sensor data only
b.1 Current time window (last value before label)
b.2 Immediate past (fixed time window before label)
Data Splitting
a.1 Leave-one-subject-out (LOSO)
Training set
Test set
Feature Preparation
a.1 Feature normalization (using statistics from all users only in training set)
b.1 Impute the missing values
a.1 Feature normalization (using statistics from all users only in training set)
b.1 Impute the missing values
Feature Selection
a.1 LASSO Filter
a.1 Using the selected features
Over/Undersampling
a.1 SMOTE-NC
Model Training
a.4 Generalized Model
b.1 Traditional machine learning (XGBoost)
b.1 Test the trained model on the test set
Model Evaluation
a.4 AUC-ROC


Preprocessing
a.1 Remove expiratory survey label samples
b.1 Remove users with too few survey labels
c.1 Label encoding using theoretical threshold
Feature Extraction
a.1 Sensor data only
b.1 Current time window (last value before label)
b.2 Immediate past (fixed time window before
What if we add last stress label as one of the features?
Data Splitting
a.1 Leave-one-subject-out (LOSO)
Training set
Test set
Feature Preparation
a.1 Feature normalization (using statistics from all users only in training set)
b.1 Impute the missing values
a.1 Feature normalization (using statistics from all users only in training set)
b.1 Impute the missing values
Feature Selection
a.1 LASSO Filter
a.1 Using the selected features
Over/Undersampling
a.1 SMOTE-NC
Model Training
a.4 Generalized Model
b.1 Traditional machine learning (XGBoost)
b.1 Test the trained model on the test set
Model Evaluation
a.4 AUC-ROC

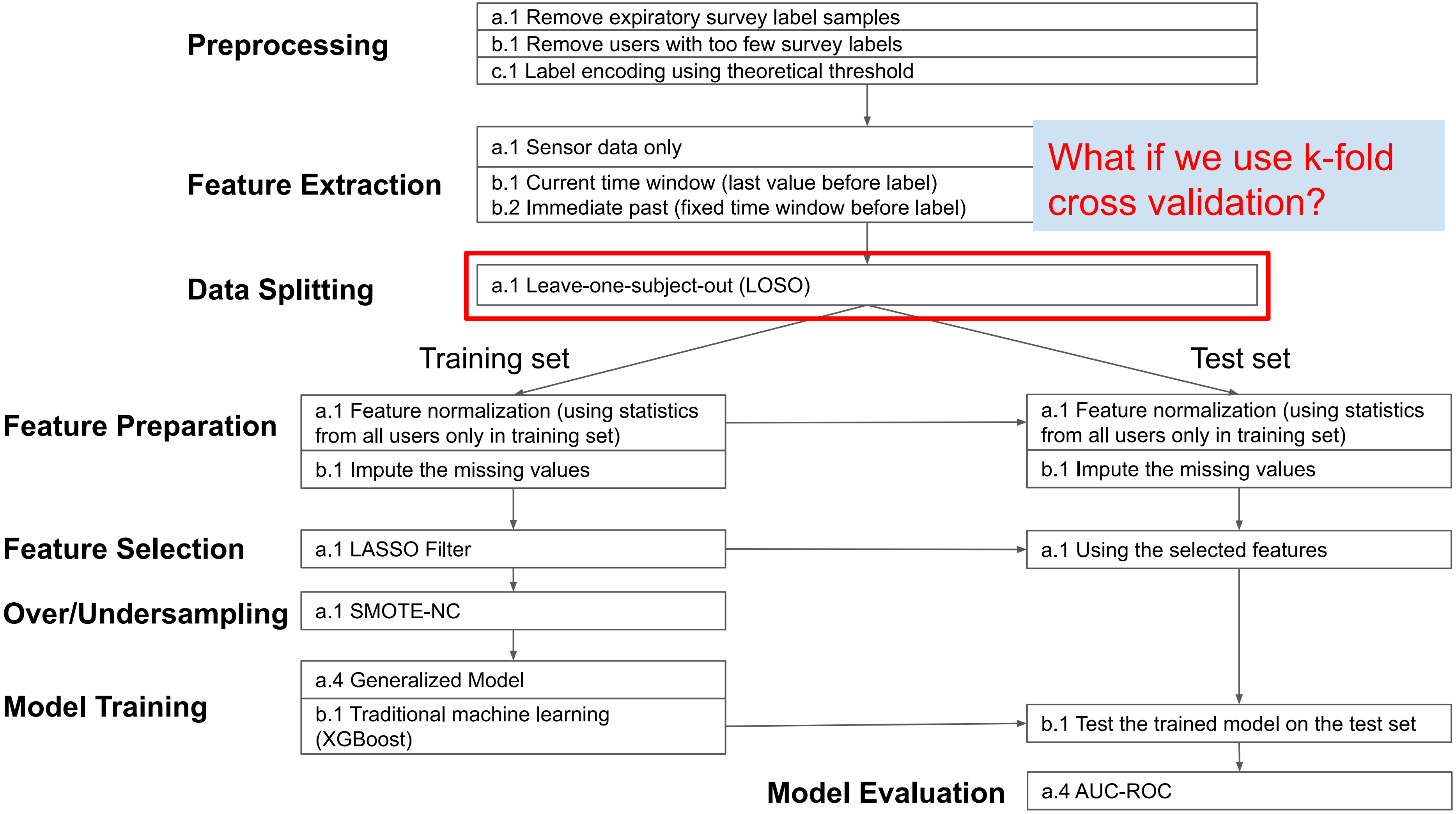
RQ2 Independent Reproducibility - Last Label

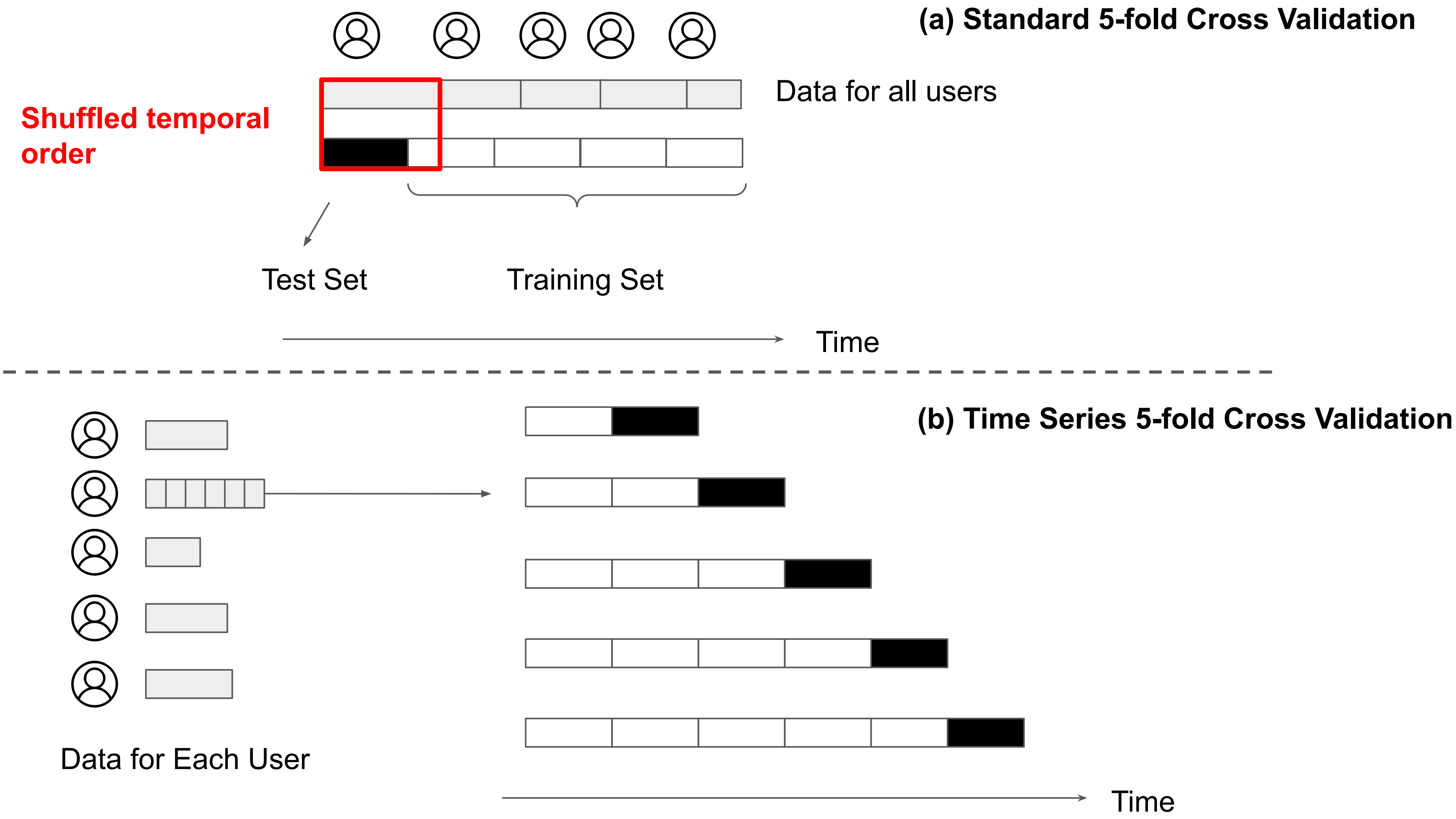


t denotes the timestamp of the label to predict
while t-1 denotes the timestamp of the last label

	AUC-ROC (K-EmoPhone)	AUC-ROC (DeepStress)
Baseline	0.518	0.522
Include Last Label as Feature	0.568	0.616

Including the last stress label as feature improves the model performance even on new users

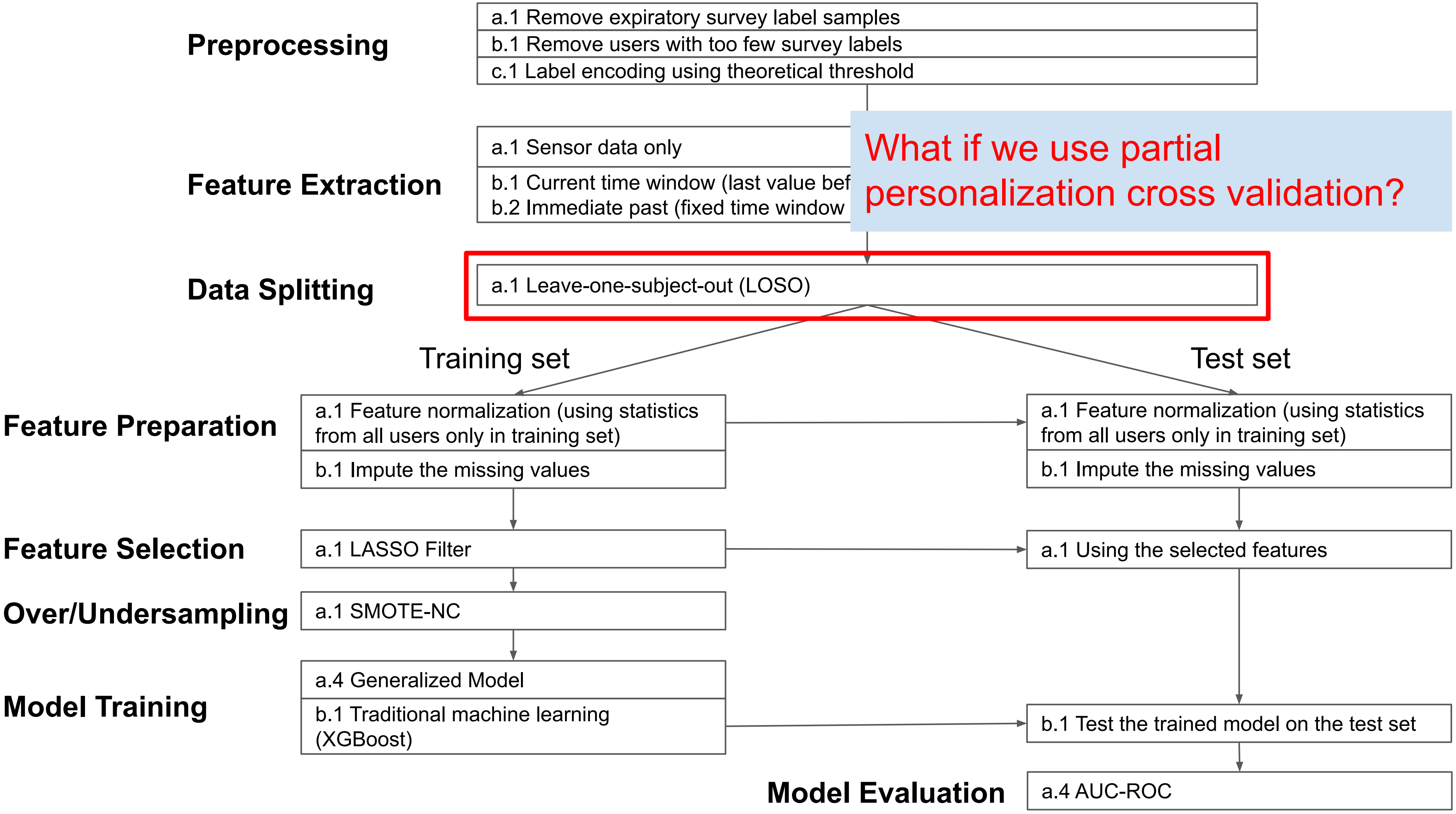

Preprocessing
a.1 Remove expiratory survey label samples
b.1 Remove users with too few survey labels
c.1 Label encoding using theoretical threshold
Feature Extraction
a.1 Sensor data only
b.1 Current time window (last value before label)
b.2 Immediate past (fixed time window before label)
What if we use k-fold cross validation?
Data Splitting
a.1 Leave-one-subject-out (LOSO)
Training set
Test set
Feature Preparation
a.1 Feature normalization (using statistics from all users only in training set)
b.1 Impute the missing values
a.1 Feature normalization (using statistics from all users only in training set)
b.1 Impute the missing values
Feature Selection
a.1 LASSO Filter
a.1 Using the selected features
Over/Undersampling
a.1 SMOTE-NC
Model Training
a.4 Generalized Model
b.1 Traditional machine learning (XGBoost)
b.1 Test the trained model on the test set
Model Evaluation
a.4 AUC-ROC


(a) Standard 5-fold Cross Validation
Data for all users
Shuffled temporal order
Test Set
Training Set
Time
(b) Time Series 5-fold Cross Validation
Data for Each User
Time

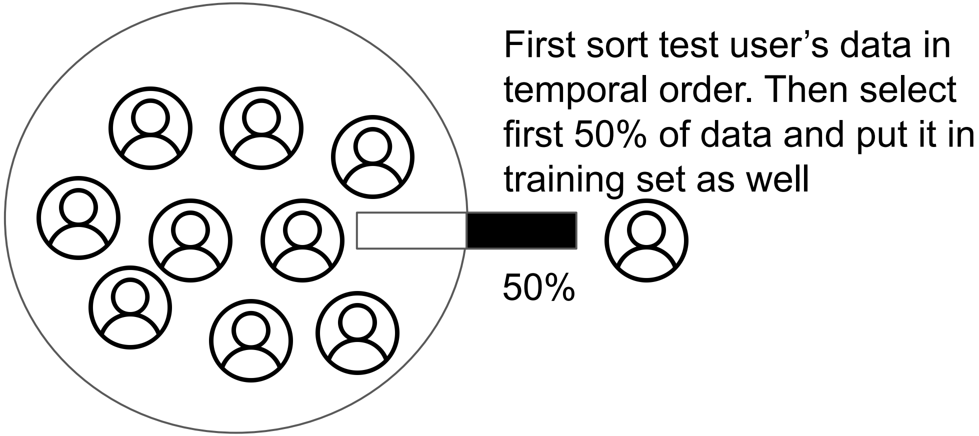
RQ2 Independent Reproducibility - K-fold Cross-Val.

	AUC-ROC (K-EmoPhone)	AUC-ROC (DeepStress)
Standard k-fold	0.650	0.764
Time-series k-fold	0.588	0.636

Standard k-fold works much better than time-series k-fold. Either because of more data in training set or data leakage due to shuffled time order.


Preprocessing
a.1 Remove expiratory survey label samples
b.1 Remove users with too few survey labels
c.1 Label encoding using theoretical threshold
Feature Extraction
a.1 Sensor data only
b.1 Current time window (last value bef
b.2 Immediate past (fixed time window
What if we use partial personalization cross validation?
Data Splitting
a.1 Leave-one-subject-out (LOSO)
Training set
Test set
Feature Preparation
a.1 Feature normalization (using statistics from all users only in training set)
b.1 Impute the missing values
a.1 Feature normalization (using statistics from all users only in training set)
b.1 Impute the missing values
Feature Selection
a.1 LASSO Filter
a.1 Using the selected features
Over/Undersampling
a.1 SMOTE-NC
Model Training
a.4 Generalized Model
b.1 Traditional machine learning (XGBoost)
b.1 Test the trained model on the test set
Model Evaluation
a.4 AUC-ROC

RQ2 Independent Reproducibility - Partial Personal.



Training Set Test Set

Partial Personalization Cross Validation
(Test on Single User)

L. Meegahapola et al. 2023. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT). ACM. https://dl.acm.org/doi/abs/10.1145/3569483

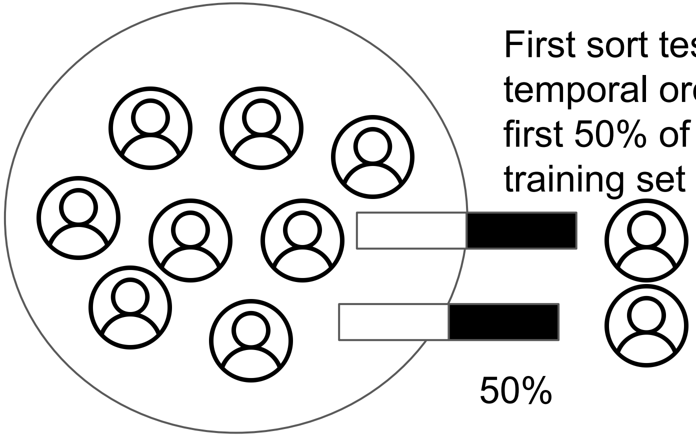
RQ2 Independent Reproducibility - Partial Personal.

	AUC-ROC (K-EmoPhone)	AUC-ROC (DeepStress)
Baseline	0.511	0.524
Partial Personalization	0.534	0.573

Partial personalization did help improve model performance.

RQ2 Independent Reproducibility - Partial Personal.

First sort test users’ data in temporal order. Then select first 50% of data and put it in training set as well



Training Set Test Set

Partial Personalization Cross Validation
(Test on Multiple Users)

L. Meegahapola et al. 2023. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT). ACM. https://dl.acm.org/doi/abs/10.1145/3569483

RQ2 Independent Reproducibility - Partial Personal.

	AUC-ROC (K-EmoPhone)	AUC-ROC (DeepStress)
w/o Partial Personalization	0.575	0.505
Partial Personalization	0.613	0.676

Testing on Multiple Users

Partial personalization works much better when testing on a group of users instead of single user.

Longer duration of data collection also helps success of partial personalization.

Summary

Importance of Labels from New Users

Both adding last label in feature set and partial personalization improve the model performance.

Labeled data from target users is important for adapting the model to unseen users.

Summary

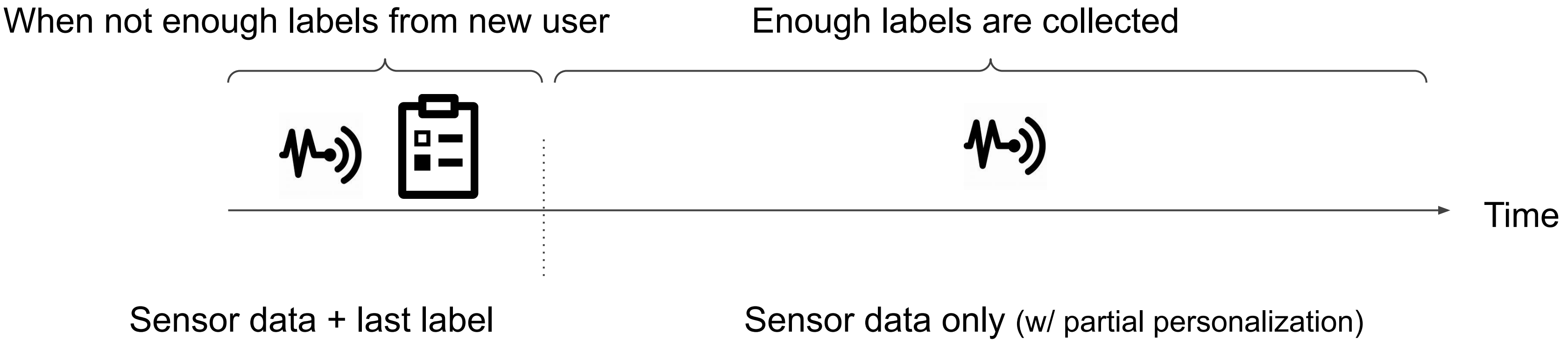
Temporal Order in Evaluation Design

Previous success of k-fold cross validation could be due to either more data in training set or **potential data leakage in time domain**.

It is more recommended to consider time order when designing evaluation settings.

Discussion

Improving Prediction Performance via User-in-the-loop Strategies



KAIST

Interactive Computing Laboratory

UBICOMP ISWC 2024 MELBOURNE, AUSTRALIA

HumblebeeAI

A Reproducible Stress Prediction Pipeline Using Mobile Sensor Data

Email: panyu@kaist.ac.kr
Website: steinpanyu.github.io

Panyu Zhang, Gyuwon Jung, Jumabek Alikhanov, Uzair Ahmed, Uichin Lee