Original Paper

Framework for Ranking Machine Learning Predictions of Limited, Multimodal, and Longitudinal Behavioral Passive Sensing Data: Combining User-Agnostic and Personalized Modeling

Tahsin Mullick[1], MEng; Sam Shaaban[2], MBA; Ana Radovic[3], MD, MSc; Afsaneh Doryab[1], PhD

[1]Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, United States

[2]NuRelm, Pittsburgh, PA, United States

[3]Department of Pediatrics, University of Pittsburgh, Pittsburgh, PA, United States

Corresponding Author:
Tahsin Mullick, MEng
Department of Systems and Information Engineering
University of Virginia
Olsson Hall, 151 Engineer's Way
Charlottesville, VA, 22903
United States
Phone: 1 4349245393
Email: tum7q@virginia.edu

Abstract

Background: Passive mobile sensing provides opportunities for measuring and monitoring health status in the wild and outside of clinics. However, longitudinal, multimodal mobile sensor data can be small, noisy, and incomplete. This makes processing, modeling, and prediction of these data challenging. The small size of the data set restricts it from being modeled using complex deep learning networks. The current state of the art (SOTA) tackles small sensor data sets following a singular modeling paradigm based on traditional machine learning (ML) algorithms. These opt for either a user-agnostic modeling approach, making the model susceptible to a larger degree of noise, or a personalized approach, where training on individual data alludes to a more limited data set, giving rise to overfitting, therefore, ultimately, having to seek a trade-off by choosing 1 of the 2 modeling approaches to reach predictions.

Objective: The objective of this study was to filter, rank, and output the best predictions for small, multimodal, longitudinal sensor data using a framework that is designed to tackle data sets that are limited in size (particularly targeting health studies that use passive multimodal sensors) and that combines both user agnostic and personalized approaches, along with a combination of ranking strategies to filter predictions.

Methods: In this paper, we introduced a novel ranking framework for longitudinal multimodal sensors (FLMS) to address challenges encountered in health studies involving passive multimodal sensors. Using the FLMS, we (1) built a tensor-based aggregation and ranking strategy for final interpretation, (2) processed various combinations of sensor fusions, and (3) balanced user-agnostic and personalized modeling approaches with appropriate cross-validation strategies. The performance of the FLMS was validated with the help of a real data set of adolescents diagnosed with major depressive disorder for the prediction of change in depression in the adolescent participants.

Results: Predictions output by the proposed FLMS achieved a 7% increase in accuracy and a 13% increase in recall for the real data set. Experiments with existing SOTA ML algorithms showed an 11% increase in accuracy for the depression data set and how overfitting and sparsity were handled.

Conclusions: The FLMS aims to fill the gap that currently exists when modeling passive sensor data with a small number of data points. It achieves this through leveraging both user-agnostic and personalized modeling techniques in tandem with an effective ranking strategy to filter predictions.

(JMIR AI 2024;3:e47805) doi: 10.2196/47805

KEYWORDS

machine learning; AI; artificial intelligence; passive sensing; ranking framework; small health data set; ranking; algorithm; algorithms; sensor; multimodal; predict; prediction; agnostic; framework; validation; data set

Introduction

Background

Mobile and wearable sensing has garnered increasing interest in areas of physical health [1,2], mental health [3-5], and activity recognition [6,7]. Multimodal passive sensing accommodates data collection without disrupting the human routine, allowing it to be an important tool to understand human behavior. However, passive sensing, unlike other forms of data, encounters common fundamental challenges in mobile health studies pertaining to physical and mental health. These challenges include small data sets, noisy or sparse data, and sensor selection criteria. Next, we explain these challenges and discuss how our framework can help in alleviating them.

One of the primary challenges in passive sensing studies is small data sets. These arise due to limitations in the sample size of participants, the study duration, and ground truth restrictions. In this study, we explored this challenge from the viewpoint of studies conducted on passive sensing. Studies related to physical health (eg, [1,2]) have investigated dietary behavior with the help of passive sensing. Participant sample sizes in Rabbi et al [1,2] were 17 and 16, respectively, which is a limited participant count. This type of data limitation is even more prominent in mental health research that relies on passive sensing. Studies on depression [3] and schizophrenia [4], for example, had participant sample sizes of 28 and 5, respectively. The limited data sets in passive sensing research are also a factor of the study duration. To understand this, we can observe the duration of study. For example, the study duration in Rabbi et al [1,2] was 21 and 98 days, respectively, while the study by Canzian and Musolesi [3] lasted for 70 days and that by Difrancesco et al [4] was limited to only 5 days. The limitation in data led researchers away from using complex deep learning (DL) models, as demonstrated in previous studies [1-4]. This is because DL models have more hyperparameters and succumb to overfitting due to memorization of the data the models are trained on [8]. In this study, we took inspiration from the existing work and selected specific traditional machine learning (ML) algorithms that are less susceptible to overfitting in small-data scenarios. However, unlike previous studies [1-4,9-17], we also ensured that our predictions were ranked based on 2 different modeling paradigms that further helped circumvent overfitting and also assisted in noise removal, as explained later.

The second challenge commonly faced when tackling passive sensor data is that of sparsity or noise. This challenge arises due to signal inconsistencies and noise in sensor data collection because of software issues, data sync, or hardware problems. Discussions of sparsity and the negative effect it has on modeling have been previously documented [7,18-20]. These studies have presented an overview of the passive sensing landscape and highlighted the role signal inconsistencies can play in predictive modeling of passively sensed data. The fact that data are noisy, especially in the case of wearable sensors, was mentioned by Plötz [18]. Cornet and Holden [19] reported that a lack of sensor precision leads to sparsity, and Xu et al [20] documented the level of noise in data that prevents user-agnostic models from generalizing well. Our proposed framework attempts to reduce the effect of noise by forming a balance between predictions from user-agnostic modeling paradigms and personalized modeling paradigms. In addition, choosing specific ML algorithms, such as Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), elastic-net, and extra-tree, and ranking predictions from them help lessen the impact of sparsity [21-24].

Sensor selection is the third type of challenge that has not received significant attention in passive or mobile sensing literature. Studies have tested various feature combinations mainly in the light of performing feature selection or feature reduction [25]. Joshi and Boyd [26] and Altenbach et al [27], for example, used heuristic-based convex optimization to select sensors from an array of sensors. However, both these studies were purely from the perspective of sensor placement. They did not investigate which combination of sensors provided the best outcome for prediction-based modeling and were more in favor of wireless sensor network establishment. Mobile or wearable devices are laced with multiple sensors, and building and knowing which sensors create optimum models are vital particularly to mental and physical health–related studies. Through our framework, we present a way to test combinations of sensor data and derive and rank predictions from among those combinations, allowing investigators to understand which combinations of sensor data yield the best predictions for their passive sensing experimental setup.

All the aforementioned challenges are common to passive sensing data sets. However, they exhibit significant presence in mental and physical health–related studies [3,4]. Xu et al [20] talked of the general sequence of steps researchers take to build models and the struggles of working with passively sensed data. A strong framework to yield the best predictions can prove to be beneficial to the community at large and bring about greater insight from studies conducted with small data sets.

In this paper, we present our ML modeling and ranking framework to address these challenges. The framework is designed to induce improved predictions for multimodal sensing. It balances both user-agnostic and personalized modeling of small data sets encountered often in mental and physical health–based studies. Our framework makes the following contributions: (1) prediction filtering and ranking through tensor-based aggregation of small, multimodal sensing data sets, (2) sensor combination selection to derive the best predictions, and (3) a reduction in overfitting predictions due to limited data and noise through ensembling of user-agnostic and personalized modeling strategies.

Importantly, it should be noted that by the size of the data set, we refer to the final data sets where raw sensor readings are

aggregated into intervals to align with the sampling frequency of ground truth data. In this work, we defined small data sets as those comprising fewer than 1000 data points for training ML models. Sparse or noisy data sets were those that either consisted of many zero entries or data sets for which highly varying sensor values were observed among different participants in the study.

We evaluated the framework through its performance in the context of predicting changes in depression severity in a group of adolescent patients. The results showed the framework's ability to use multiple modeling approaches for providing robust predictions in critical cases, such as mental health.

Passive sensing data for human behavior modeling are different from other data formats, such as images, audio, or normal tabular data. Researchers in the field of passive sensing agree that passive sensing data have some common properties, such as they are time series data, multimodal, longitudinal, nonlinear, and noisy, as previously discussed [20]. Xu et al [20] also emphasized the researcher's need for tools that can help ease the time lost in traversing the common pitfalls of passively sensed data. Our work endeavors to resolve such pitfalls for cases where passive sensing data are limited. Next, we discuss the related work highlighting the state of the art (SOTA) in passively sensed small, multimodal data sets.

Related Work

Despite the growing body of work using multimodal passive sensing in physical and mental health applications [28-32], there exists scope for improvement in small-data scenarios.

In this section, we underline what exists in the current SOTA and why we need a ranking-based framework to address scenarios with small data sets. Keeping in line with our contribution, it will prove beneficial to present the current SOTA through understanding:

- How traditional ML algorithms are applied in the context of passive sensing
- Why complex DL models do not work well in limited data scenarios
- How ensemble modeling has been adapted in passive sensing studies
- What the role of data fusion is in modeling passive sensing data

Traditional Machine Learning Algorithms Applied in Passive Sensing

Traditional ML algorithms have been applied to passive sensing in the space of human activity recognition (HAR) [9-11], general health [12-15], and mental health [3,16,17]. A deeper dive into the studies reveals some common takeaways that include the following:

- All of them test multiple ML algorithms, followed by selecting predictions based on the overall chosen validation metric.
- They all follow a singular modeling strategy, resorting to either user-agnostic or personalized modeling.
- Cross-validation (CV) is either K-fold or leave-one-out CV.

This is a repetition of steps that authors in the field make independently and is discussed extensively in the highlighted literature presented in Table 1. Following a single modeling strategy is restricting as choosing to follow a user-agnostic approach exposes the model to a greater degree of noise due to the heterogeneity in sensor values among participants, while solely following a personalized approach reduces data availability further as the model learns from individuals' data rather than the general population data. Our endeavor through this ranking framework is to combine both the approaches, while using traditional ML algorithms.

Table 1. Summary of SOTA[a] literature using traditional ML[b] for passive sensing, with special focus on CV[c], the overall modeling strategy, and ML algorithms.

Study	Application	CV	Modeling strategy	ML algorithm
Kwapisz et al [9]	HAR[d]	10-fold	User agnostic	DT[e], LR[f], MLP[g]
Shukla et al [10]	HAR	5-fold	User agnostic	KNN[h], SVM[i]
Chen and Chen [11]	HAR	10-fold	User agnostic	RF[j], SVM, KNN
Huang et al [12]	Sleep	10-fold	User agnostic	SVM
Montanini et al [13]	Sleep	K-fold/leave 1 out	User agnostic/personalized	KNN, DT, RF, SVM
Teng et al [14]	Parkinson's tremors	5-fold	User agnostic	XGBoost[k], DT, RF
Azam et al [15]	Breath	K-fold	User agnostic	SVM
Canzian and Musolesi [3]	Depression	Leave 1 out	User agnostic	SVM
Grunerbl et al [16]	Bipolar disorder	K-fold	User agnostic/personalized	NB[l], KNN, DT
Saeb et al [17]	Depression/anxiety	10-fold	User agnostic	XGBoost, DT

[a]SOTA: state of the art.
[b]ML: machine learning.
[c]CV: cross-validation.
[d]HAR: human activity recognition.
[e]DT: decision tree.
[f]LR: linear regression.
[g]MLP: multilayer perceptron.
[h]KNN: K-nearest neighbor.
[i]SVM: support vector machine.
[j]RF: random forest.
[k]XGBoost: Extreme Gradient Boosting.
[l]NB: naive Bayes.

Limitation of Deep Learning in Small-Data Scenarios

A common replacement for traditional ML algorithms is DL. Here, we explain why DL models are not ideal solutions for the problem addressed in this study. DL models have gained immense popularity in the literature [33]. Their power lies in modeling the nonlinearity and noisy nature of passively sensed data. DL has a toolkit of strategies to handle small data that includes data augmentation [1], transfer learning [19], and ensembling [29]. However, the size of a small data set in DL studies ranges from 1000 to 10,000 training points [18]. This is unlike the ranking framework presented in this paper, which has been designed for data sets with fewer than 1000 data points. Therefore, despite their superiority in modeling larger passive sensing data sets, the performance of DL models suffers in cases where study data are limited and in the hundreds. The complexity of DL models results in overfitting to small data sets [14]. In this paper, we worked to solve the problem of limiting data by providing researchers with a reproducible way to run multiple models and select the best predictions from among them. By using traditional ML in conjunction with ranked predictions from user-agnostic and personalized models, the issue of overfitting due to model complexity is dealt with in the proposed work.

Ensemble Learning to Build Robust Models for Passive Sensing Data

Among the different ways of dealing with overfitting, ensemble learning has been instrumental. Ensemble ML is a widely used approach in passive sensing studies [14,17,34,35]. It mainly exists in the form of boosting [6,14,17,34], bagging [14,16], weighted ensembles [35], and max voting [36] ML algorithms. Ensemble learning presents better results in terms of evaluation metrics. Ensemble learners are trained using a single modeling strategy. Therefore, they are either personalized ensembles [35], which allows learners to derive interesting artifacts at personal levels, or user-agnostic ensembles [14,17,34,36-38], which only generate macrolevel information. Our contribution through the ranking framework is to provide a balance of both macrolevel patterns and user-specific patterns through a weighted ensemble of both approaches. Ensembling in this manner will allow us to reduce the noise that is picked up due to varying sensor values among users and account for user-specific patterns through the predictions on personalized data.

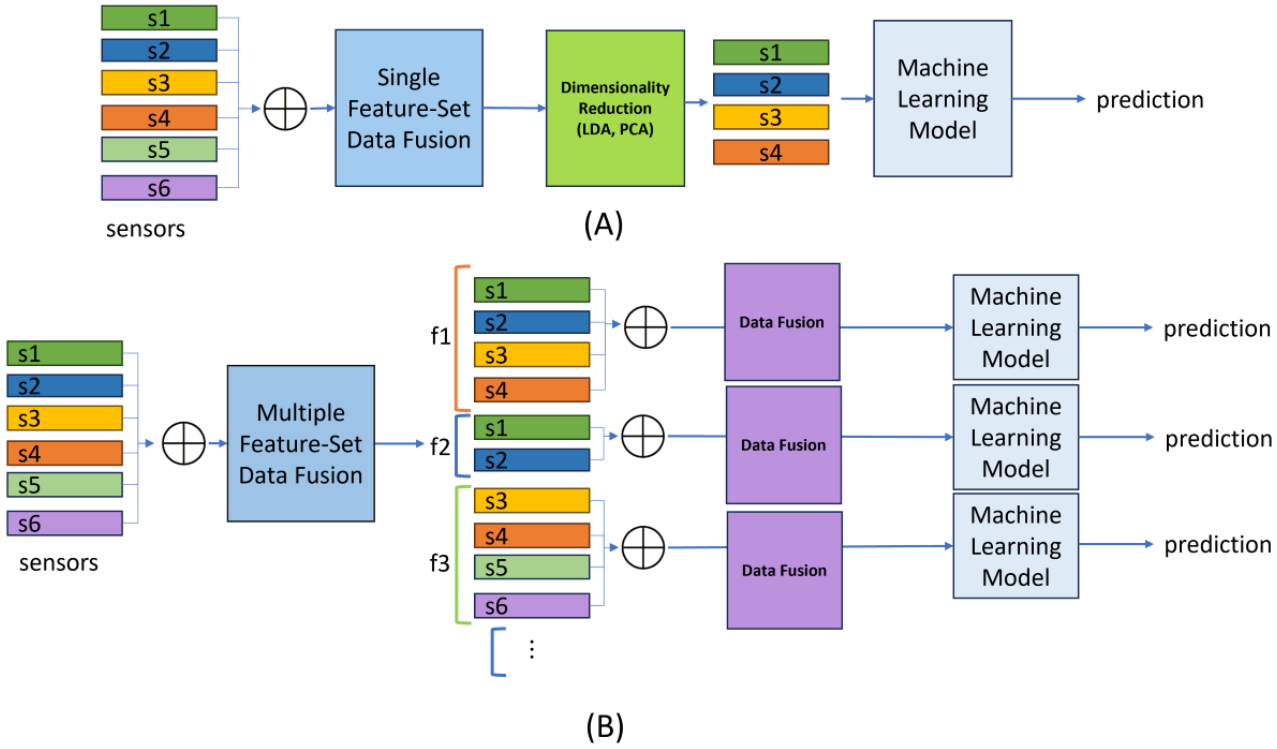
Role of Data Fusion in Passive Sensing Studies

The use of data fusion in passive sensing has seen a steady growth due to the use of multimodal sensors in passive sensing studies. Earlier studies were often restricted to single sensors, which were then manipulated to obtain a handful of features. For example, Canzian and Musolesi [3] primarily used GPS

sensor data, while Kwapisz et al [9] only opted for an accelerometer to base their predictive modeling. The way data fusion is approached has a common link among the surveyed studies in the current literature. The studies have applied feature-level fusion [10,39-43], where fusion takes place after feature extraction from raw signals. A single feature set is generated and then passed on to dimensionality reduction, such as linear discriminant analysis (LDA) [10] or principal component analysis (PCA) [40-42]. The focus in these papers tends to be a reduction in dimension, without trying to study the impact of multiple distinct feature combinations. In comparison, our contribution of feature selection focuses on studying the relationship between each group of sensors by creating multiple feature sets based on sensor availability. This will allow us to select the best set of features to work with for a specific type of study. An illustration of the difference in the existing literature and our feature fusion approach is shown in Figure 1 [10,39-43].

Overall, our ranking framework is motivated to aid researchers in situations in which data sets are small, sparse, or noisy and multimodal by taking advantage of its multiple model generation and the balanced outcome of the best predictions.

Figure 1. (A) Data fusion approach in the current literature and (B) proposed FLMS data fusion approach, where s1-s6 represent distinct sensors and f1-f3 represent feature set combinations, which were then fused prior to ML modeling. FLMS: framework for longitudinal multimodal sensors; LDA: linear discriminant analysis; ML: machine learning; PCA: principal component analysis.



Methods

Ethical Considerations

The data collection was approved by the Institutional Review Board of the University of Pittsburgh Human Research Protections Office (STUDY18120176).

Data Description

The study used passive sensing data and is presented through the lens of depression change prediction among adolescents. The data set comprised 55 adolescents from 12 to 17 years old, with an average age of 15.5 (SD 1.5) years. The AWARE app was used to collect the participants' smartphone and Fitbit data. The data completeness rate for AWARE and Fitbit was, on average, 65.11% and 30.36%, respectively. The levels of completeness echoed the difficulty in collecting passive sensing data. Smartphone and Fitbit data were collected from each participant over 24 weeks.

The 9-item Patient Health Questionnaire (PHQ-9) [44] was used to collect weekly self-reports of depression severity from the participants. The questionnaire consists of a set of 9 questions, which can be scored from 0 to 3, giving a score range of 0-27. We used PHQ-9 scores as the ground truth to compare the prediction accuracy of our models.

Relation of Sensor Data to Mental Health

Raw sensor data, including calls, location, conversation, screen usage, Wi-Fi, steps, sleep, and heart rate, were processed, and relevant features were extracted at daily intervals. We used RAPIDS [45] to extract 72 features from the sensors. The existing literature [3,46-51] shows how location [3,46,49,50,52], calls [48,53], screen usage [46,54,55], conversations [55-58], Wi-Fi [48,59], steps [60], and heart rate [61] can be effective in predicting mental health behavior. Studies [3,46,49,50] have used location sensors, such as the GPS, and shown a strong relation to depressive symptom severity. Clinical measures, such as the PHQ-9 [44], the PHQ-8 [62], the Hamilton Rating Scale for Depression (HAM-D) [63], and the Hamilton Rating

Scale for Anxiety (HAM-A) [64], have been used as target labels for prediction using sensor-based features, establishing a proof of association between sensor features and mental health predictions. Studies [47,48,51,54,60] have used multimodal sensors of smartphones that included the sensors we chose for this study: calls, location, conversation, screen usage, Wi-Fi, Fitbit steps, and Fitbit heart rate. In the *Results* section, we further elaborate on the feature engineering from each of the sensors. The validity of using the sensors to predict mental health, in particular the choice of sensors, was motivated by the aforementioned studies, which showed strong predictive capability of sensors in the area of mental health prediction.

Framework Design and Modeling

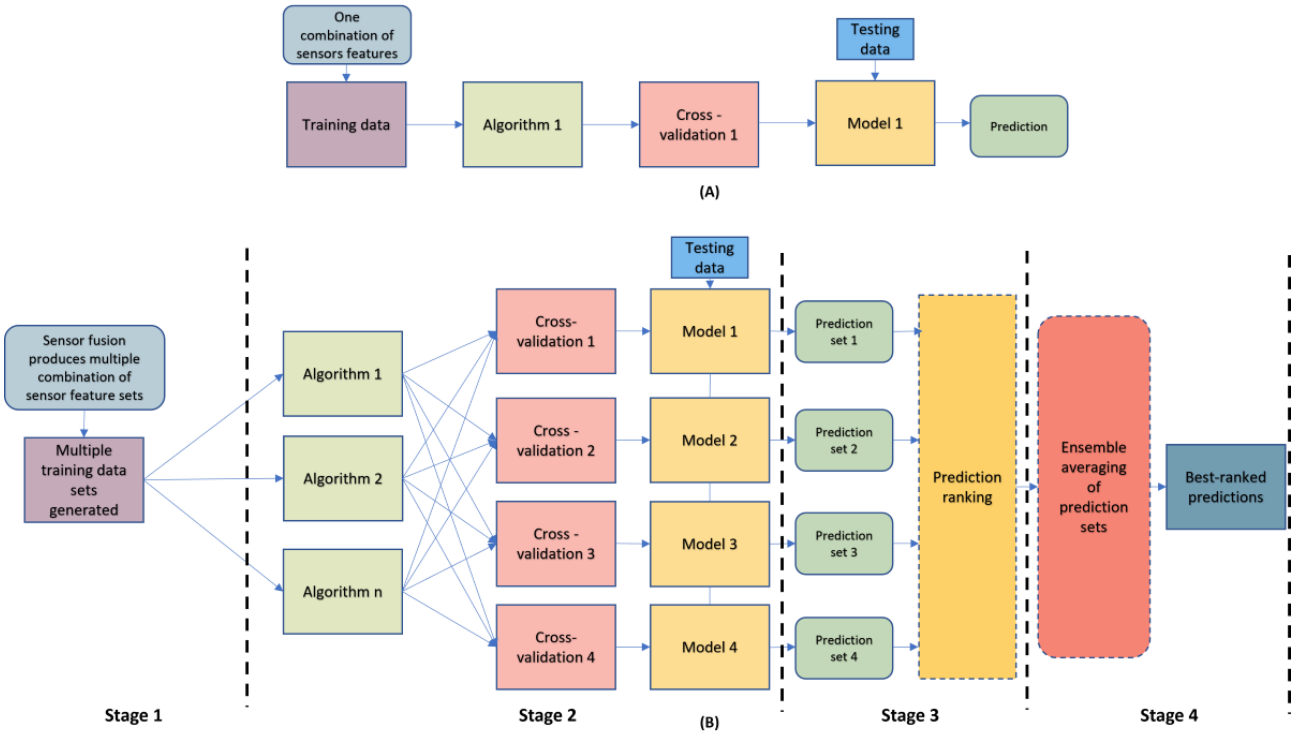
We proposed a framework for longitudinal multimodal sensors (FLMS) as a ranking framework to rigorously handle longitudinal, multimodal sensor data and incorporate different analysis and modeling strategies suited for small and sparse time series data sets to produce better results. The FLMS incorporates 4 stages to improve, rank, and filter data set predictions (see Figure 1):

- Stage 1: multimodal sensor fusion to explore the data set from multiple views and to identify the minimum number of sensors necessary to yield a good prediction. It also addresses sparsity.
- Stage 2: ML modeling with combined user-agnostic and personalized approach. This stage is designed to leverage user-agnostic and personalized predictions. The ML algorithms used in this stage were chosen due to their superior prediction capability in small-data scenarios and their ability to tackle sparse data sets.
- Stage 3: tensor-based aggregation and ranking leverage predictions from all fused combinations and modeling strategies to calculate more robust predictions.
- Stage 4: final prediction informed by the ensemble weighted average of both user-agnostic and personalized predictions to reduce the effect of overfitting in small data sets. This stage uses weights calculated via hamming distances to prevent any modeling approach from dominating the predictions.

A high-level view in Figure 2 illustrates how the FLMS is different from conventional ML approaches. Observing Figure 2A, we understand that the conventional modeling strategy uses a single algorithm with either a user-agnostic CV, where all users are included in the training and test sets, or a personalized CV strategy, where a single user's data are used to derive predictions. However, Figure 2B displays how the FLMS uses different combinations of sensors as input data, followed by multiple algorithms and a combination of user-agnostic and personalized modeling. The modeling stage is followed by a ranking of predictions and finally an ensemble of the predictions to yield the final output.

A detailed explanation of the stages of the FLMS and their utility is provided next.

Figure 2. (A) Conventional modeling approach and (B) proposed FLMS approach. FLMS: framework for longitudinal multimodal sensors.



Stage 1: Multimodal Sensor Fusion

Stage 1 was designed for the early fusion of sensors at a feature level. Sensor fusions followed a combinatorial approach using $\binom{Z}{x}$, where Z is the total number of modalities available and x is the number of sensors to fuse. Our case study had 6-sensor modalities that generated a set of 63 separate data sets calculated as $\binom{6}{1}, \binom{6}{2}, \binom{6}{3}, \binom{6}{4}, \binom{6}{5}, \binom{6}{6}$.

Data set preprocessing steps involved normalization and log transforms. Imputations to fill missing feature observations

were also conducted. The framework allowed for implementation of the K-nearest neighbor (KNN) algorithm for imputation, which is also the first level of defense against sparsity. The generated data sets were in 2D tabular data format. The sensor data were aggregated according to the granularity of the ground truth. Our case study collected PHQ-9 scores as an accepted depression measure. The total score range of the 9 questions was 0-27. This was collected on a weekly basis, and thus, our daily data were aggregated in weekly intervals.

Stage 2: ML Modeling With a Combined User-Agnostic and Personalized Approach

Stage 2 focused on modeling and predictions based on the data sets generated in stage 1. All stage 1 data sets were run through the modeling suite, which encompasses a series of ML algorithms and CV strategies to help build user-agnostic and personalized models.

The ML suite includes case-specific linear and nonlinear algorithms. For our case study on adolescent depression, we followed a regression-based approach, and therefore, we selected algorithms such as linear regression (LR), elastic-net, random forest (RF), AdaBoost, extra-tree, gradient boosting, and XGBoost. The algorithms were chosen based on (1) their performance in the existing literature when working with small data and robustness to sparsity, and (2) tree-based models, which were specifically chosen to provide added tractability for researchers to inspect which features mainly contributed to the models' predictive capability. The algorithms were used in each modeling strategy. The predictions of the ML algorithms for each time unit were stored in arrays for each participant and later used to select the best model for each participant. The best model selection strategy chose the model with the minimum error (in the case of regression) or the maximum accuracy (in the case of classification) among all algorithms. For example, among l number of regression algorithms, the best model was chosen as follows:

$$\text{argmin}_{\text{alg}}\left(\sum_{i=1}^{m} |\text{pred}_m(\text{alg}_t) - \text{true}_m|\right) \quad \textbf{(1)}$$

,where alg refers to the algorithm with the lowest absolute sum error and $\text{pred}_m(\text{alg}_t)$ is the prediction made by an algorithm l at unit time t. The array of prediction by the best model was retained for each respective participant.

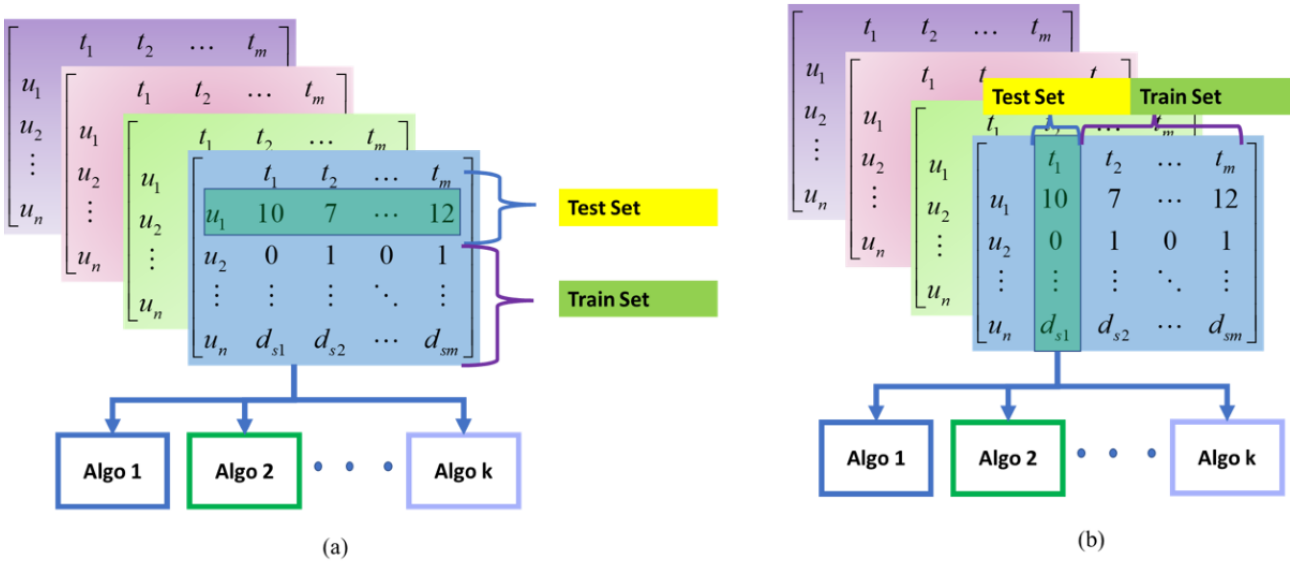
User-Agnostic Model Building

To leverage as much data as possible, we implemented the leave-one-participant-out (LOPO) and leave-time-unit-X-out (LTXO) strategies. This is illustrated in Figure 3A,B.

In LOPO, we held out all data from a single participant for validation and trained the model on other participants. This strategy reflected the cold start case where a new user started using the health app.

The LTXO is based on the unit of time for ground truth data (eg, a week). For training, we held out a given time unit of all participants and trained the model on the rest of the time units. This strategy evaluated the impact of time-specific segments of data on prediction. The training phase captures the similarity and variation of data during different time units to build user-agnostic models.

Figure 3. User-agnostic model building: (A) LOPO and (B) LTXO strategies. Algo: algorithm; LOPO: leave one participant out; LTXO: leave time unit X out.



Personalized Model Building

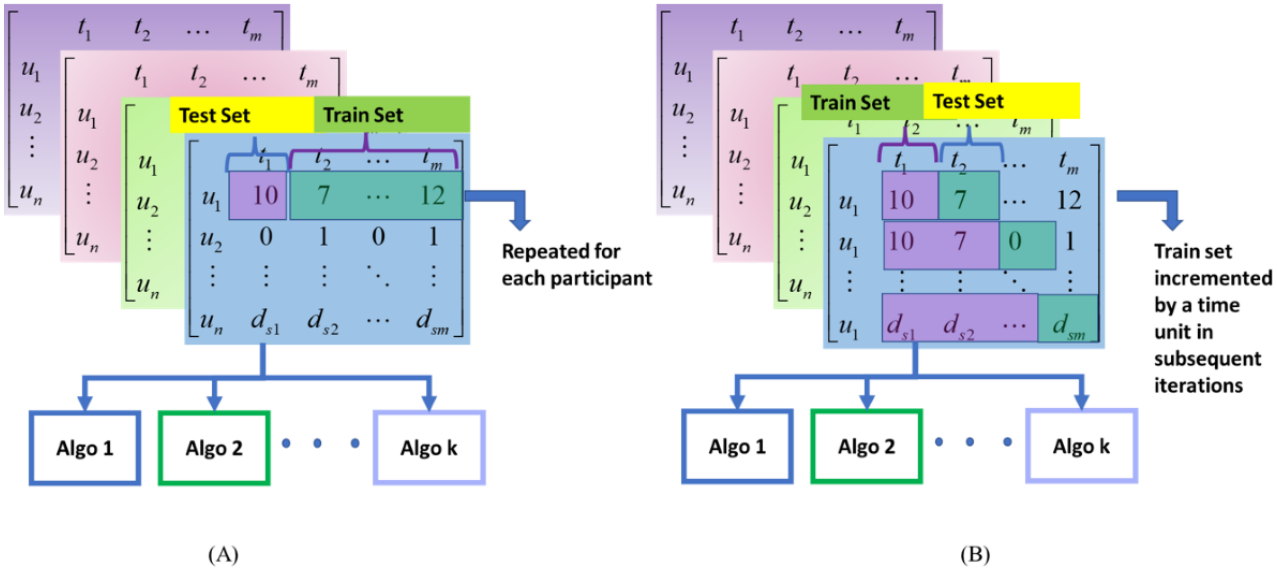
The personalized modeling strategy leverages each user's historical and cross-time data samples in a sliding window and the leave-one-time-unit-out approach.

For each participant, the accumulated-time-unit (ATU) strategy built a model from X_t time units of data to predict X_{t+1}. For example, the model built from weeks 1 and 2 predicted depression in week 3. In the next iteration, the sliding window was increased by T time units (eg, 2 weeks) to repeat the model-building process. This process continued until the

maximum number of time units was reached. This method examined the forecasting capability of the framework.

The leave-one-time-unit-one-participant-out (LOTPO) strategy trained the models on all time units of a participant across time to predict the target label for the current time unit. For example, for a participant with 10 weeks of data, we built a model from data in weeks 1-5 and weeks 7-10 to predict depression in week 6. This method evaluated the feasibility of past and future data for each participant to predict an outcome (Figure 4A,B).

Figure 4. Personalized model building: (A) LOTPO and (B) ATU strategies. Algo: algorithm; ATU: accumulated time unit; LOTPO: leave one time unit of participant out.



Stage 3: Tensor-Based Aggregation and Ranking

The output of stage 2 was a set of best prediction matrices for sensor fusion combinations, where each slot in the matrix represented prediction results for a participant in a particular time unit. We represented these predictions in the form of Z-dimensional tensors (Figure 5), where Z is the number of modalities being used. For example, a study with 6 modalities and 45 users over 24 weeks was represented in tensor form as (6, 45, 24). The tensor representation helped represent the high dimensionality of sensor combinations.

The predicted values for each slot across tensors were then aggregated using an aggregation function (eg, mean). This process took advantage of the stage 2 combinations to help reduce the error in prediction. For example, we aggregated predictions of 6 tensors (generated from 5-sensor fusion) into 1 tensor by calculating the mean of the predictions from the 6 combinations (see Figure 3). This was done for both user-agnostic and personalized models. The aggregated mean was calculated using the following equation:

$$M_{agg} = \frac{1}{k}\sum_{i,j=1}^{k,m} p_{ij}^{set_k}$$

(2)

,where M_{agg} is the aggregated mean, k is the total number of sensor combinations aggregated, i is the combination number, j is the corresponding time unit, and $p_{ij}^{set_k}$ is the prediction across each set of combinations. The data were now in a format where each 2D tensor represented a particular sensor fusion prediction set (Figure 6).

The predictions were next encoded into 0s and 1s to counter the large variance in the regression values from the original values. This logic can be set based on the type of ML problem the framework is being used to address. For example, in our case study, if the regressed change in depression score values was 0 or negative value, we classified it as 0, and if it was positive, we represented it as 1 (Figure 7).

The next step in this stage measured the hamming distance between the 0-1-encoded tensor and the true labels tensor, as shown in Figure 8. These hamming distances were then aggregated (D_u) for the respective 2D tensor as follows:

$$D_u = \sum_{i=1}^{m} d(p_i, a_i)$$

(3)

,where $d(p_i, a_i)$ is the hamming distance between unit time predictions p_i and the true value a_i. Based on the measured distance, we ranked and chose the best set of predictions. This metric helped inform the choice of weightage to associate with a particular modeling strategy. The hamming distance helped further reduce errors after encoding and filtered down to the best set of predictions from each strategy.

Figure 5. An example of tensor representation of 6-sensor fusion predictions.

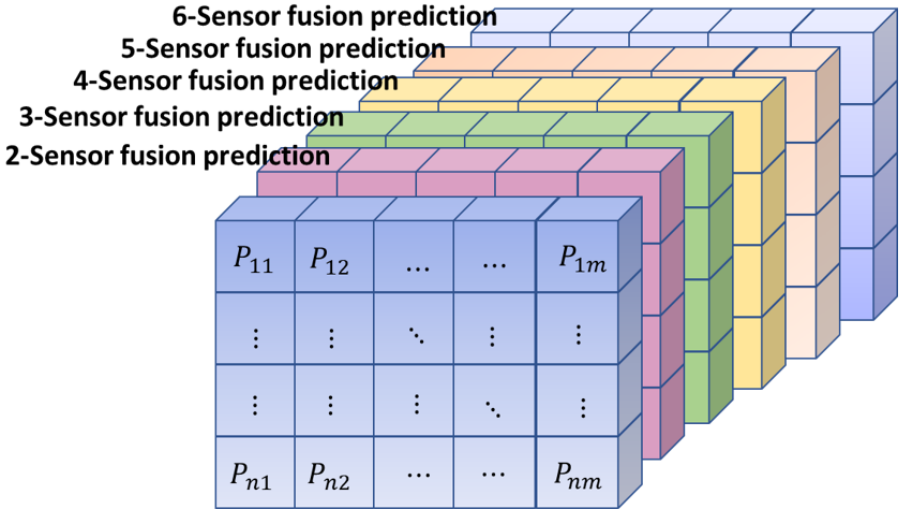


Figure 6. Instance of ATU where it shows how the mean aggregated prediction set is generated according to Equation (2). ATU: accumulated time unit; avg: average.

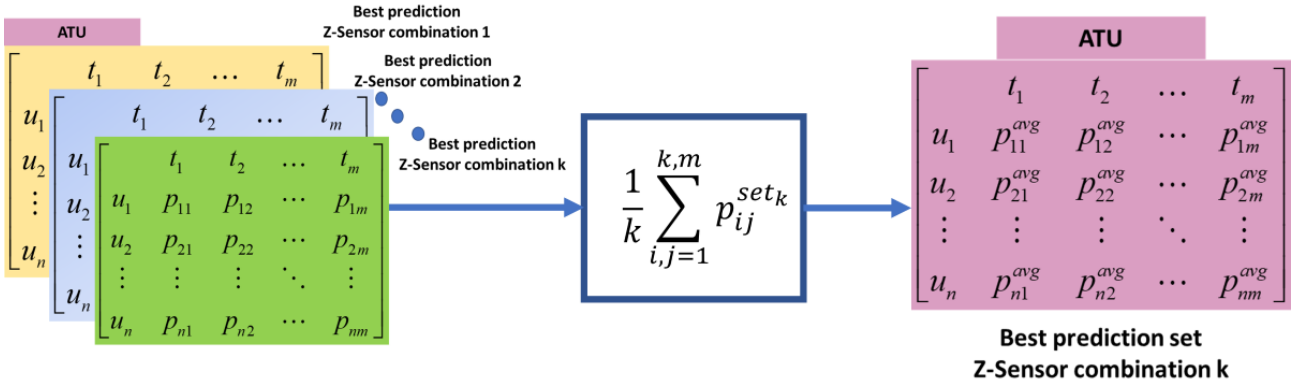


Figure 7. The 0-1 encoding process resolves dealing with large variances in regression values. ATU: accumulated time unit; LOPO: leave one participant out; LOTPO: leave one time unit of participant out; LTXO: leave time unit X out.

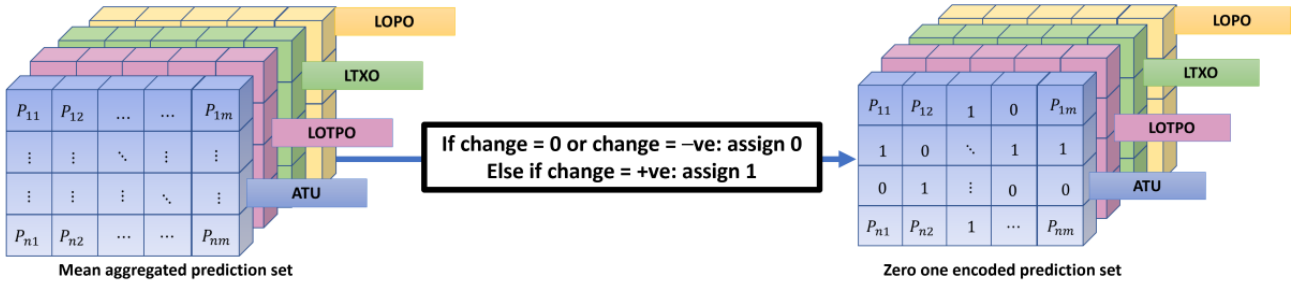
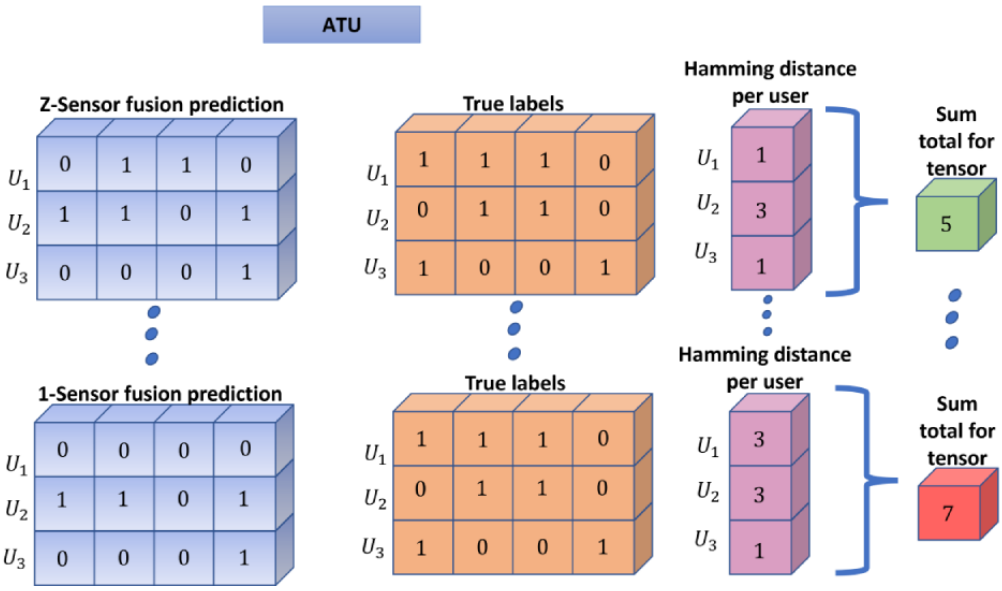


Figure 8. Hamming distance calculations reduce error and also determine the weight each of the 4 modeling approaches will contribute to stage 4's ensembled weighted average. ATU: accumulated time unit.



Stage 4: Weighted Ensemble

The final stage formed the most robust set of predictions via an ensemble weighted average approach, where weights were calculated based on the minimum hamming distances derived from each modeling strategy in stage 3 (Figure 9):

$$E_{avg} = \frac{\sum_{i,j,k=1}^{n,m,4} P_{i,j} \cdot w_k}{\sum_{k=1}^{4} w_k} \quad (4)$$

,where P_{ij} is the prediction tensor, w_k is the weight based on the minimum hamming distance, and i and j are the number of users and time units, respectively. The data were then encoded back to 0s and 1s. A complete version of the FLMS with all its stages is presented in Figure 10 (see Multimedia Appendix 1 for a higher quality image).

Figure 9. Ensemble average based on weights derived from the hamming distance to arrive at best-ranked predictions. ATU: accumulated time unit; LOPO: leave one participant out; LOTPO: leave one time unit of participant out; LTXO: leave time unit X out.

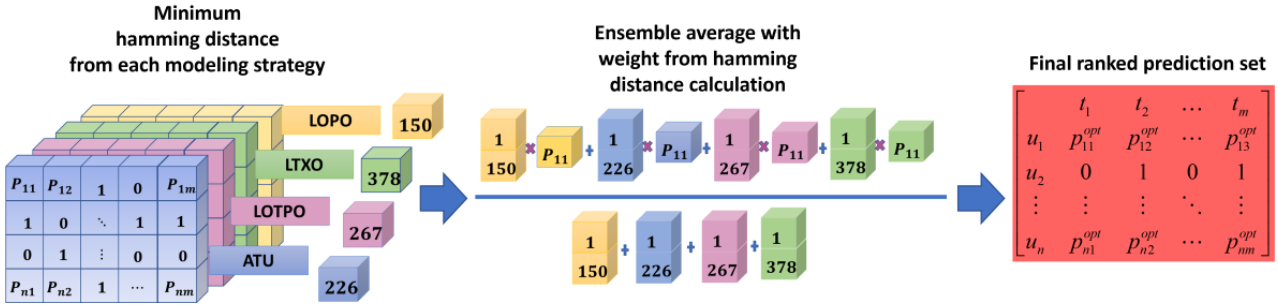
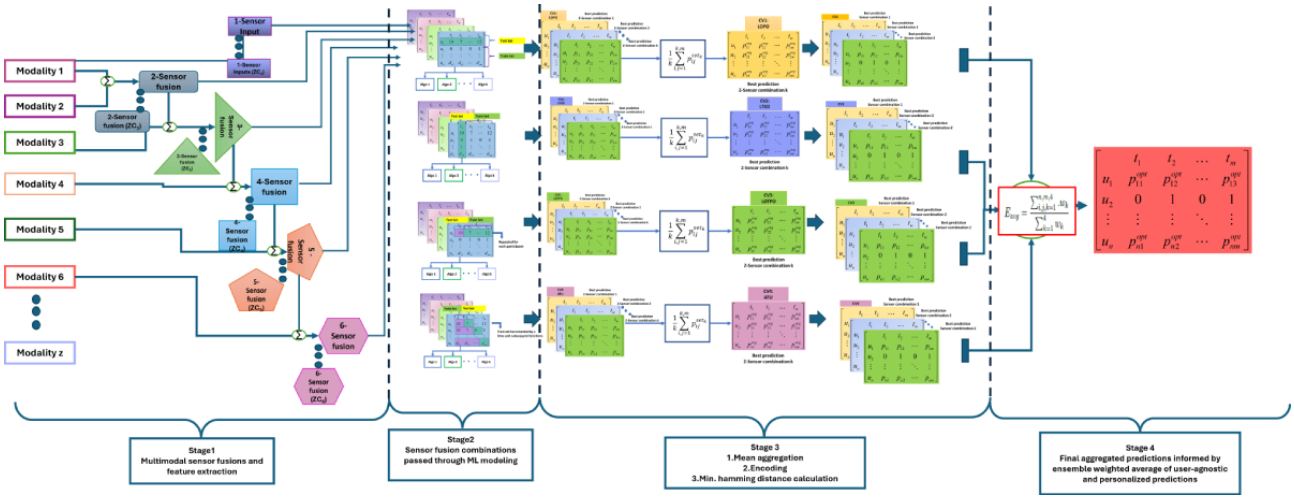


Figure 10. FLMS ranking overview. Algo: algorithm; ATU: accumulated time unit; avg: average; CV: cross-validation; FLMS: framework for longitudinal multimodal sensors; LOPO: leave one participant out; LOTPO: leave one time unit of participant out; LTXO: leave time unit X out; ML: machine learning.



Results

Stagewise Description of Framework Processing on an Adolescent Data Set

To evaluate the performance of the proposed FLMS, we used a depression data set of adolescents. This was a small data set, comprising noisy, multimodal sensor values from multiple participants—a suitable case study for our purpose of evaluating the performance of our proposed framework. Before presenting the experimental results, we first provide an understanding of how the adolescent data set was processed at each stage of the FLMS.

The passively sensed depression data set was longitudinal, with a varying number of observations per participant. The goal was to predict changes in the depression score. This was achieved by passing the small set of observations through our ranking framework, which processed, modeled, ranked, and output the best set of overall predictions based on multiple modeling approaches. A prediction of change in depression is difficult and becomes even more challenging when the amount of data provided to the ML algorithms is limited.

Stage 1 Outcome

As part of stage 1, daily data were aggregated in weekly intervals to align with weekly ground truth values. Based on our extensive exploratory data analysis (EDA), we set thresholds for sparsity and adopted KNN as the imputation strategy.

Our final data set consisted of 507 data points with 72 features, with an average of 13 weekly data points per participant. A series of data sets were then produced from an early fusion of 6-sensor features. Each data set retained 45 (81.8%) of the 55 participants. We had to drop 11 (20%) participants as they were missing more than 60% of their sensor data. The true depression state of the participants was given by the PHQ-9 weekly survey. The change in participant depression scores was calculated as $W_m - W_{m-1}$, where W_m is the score for the m-th week; this served as the ground truth for our analysis.

Stage 2: ML Modeling Outcome

The ML algorithms in stage 2 regressed on the change in the depression score, with positive changes exhibiting a rise in the depression score in that week, negative changes representing a decrease, and 0 marking no change. The best predictive models of depression for each participant were built and selected following the steps in stage 2.

Stage 3: Encoding and Prediction Filtering Outcome

This led to stage 3, where after the mean aggregation, we encoded the regressed values as our goal was to predict whether the change in the depression score was positive, negative, or constant, rather than determining the exact value of the change. This step was followed by hamming distance calculations to further rank and filter the best set of predictions.

Stage 4: Final Prediction Ensembling of Adolescent Data

The predictions evaluated by the minimum hamming distances entered stage 4, where we calculated the final ensemble predictions. The predictions used weights determined by hamming distance calculations, which enabled us to balance between personalized and user-agnostic models. This step completed the offline training and prediction of change in depression in the adolescent data set.

Experiment Design and Results

In this section, we present the depression change prediction results of the FLMS. The experiments were designed to test the framework's claims of reducing overfitting on a small data set, reducing the impact of noise or sparsity, and identifying the best combination for sensor fusion.

We conducted 3 main experiments in support of our claims:

- Experiment 1 tested FLMS predictions against singular modeling strategies used in SOTA. This experiment evaluated our claim regarding the advantage of the overall framework that took steps to reduce noise and identify the best sensor combinations versus a singular modeling strategy.

- Experiment 2 was a SOTA comparison test conducted to evaluate how our prediction-ranking framework performed in comparison to existing ML and DL approaches used in the current literature. This comparison also substantiated the FLMS performance to overfitting versus the existing strategies in the literature from prediction in small-data scenarios.
- Experiment 3 was designed to compare the FLMS performance with that of commonly used ML algorithms that have been shown to perform well with sparse data. It is important to note that there is an overlap of ML algorithms used to tackle sparsity and those used in passive sensing studies for mental health, particularly for small data sets.

Evaluation Metrics

The task of the FLMS is to model, rank, and output the best set of predictions from multiple modeling approaches. The output of the FLMS are predictions encoded as 0s or 1s (ie, binary values). Therefore, our choice of evaluation metrics for the framework predictions was the average accuracy, average recall, and average F_1-scores amongst users.

Experiment Metadata

The metadata pertaining to each experiment is provided at the end of the experiments. The information included as metadata is based on the best practices used [65] to help with reproducibility of results. They include (1) feature preprocessing steps, (2) modeling CV strategy, (3) ML algorithms used, (4) random state, and (5) evaluation metrics specific to the experiments. They are presented in the form of tables following the corresponding results for each experiment.

Data Set Used in the Experiments

To standardize our experiments, we maintained a consistent data set, a combination of 6-sensor feature sets that included calls, location, screen usage, conversation, Fitbit, and Wi-Fi. After the stages of preprocessing, missing data imputation using the KNN strategy, and the removal of highly corelated features, the final data set comprised 61 features and 507 data points belonging to a total of 45 (81.8%) participants.

Feature Engineering in Experiments

Since we maintained a consistent data set for all our experiments, feature engineering for all the experiments was achieved through data collected from 6 sensors. As discussed earlier, the data were collected from participants' smartphones using the AWARE app [66] and then passed through the RAPIDS application programming interface (API). The features extracted using the API are discussed in detail next.

Call Sensor Features

The calls sensor features provide a context of how frequently the user has been in contact with someone else. Studies have revealed that higher degrees of depression are linked to reduced contact with social circles [48,53]. As part of call sensor features, we extracted the total number of missed calls; the counts of missed calls from distinct contacts, calls from the most frequent contacts for a time segment, incoming calls, and outgoing calls; the mean (SD), maximum, and minimum duration of both incoming and outgoing calls; and the entropy duration of outgoing and incoming calls, which provided an estimate of the Shannon entropy for the duration of all calls of a particular call type (ie, incoming, outgoing, or missed). All the extracted features were mean-aggregated over the period of 1 week to match the ground truth.

Location Sensor Features

Location sensor features provide a contextual idea of the amount of movement users of the sensors go through and show the correlation to mental health [3,46,49,50]. The location data are collected through the phones' GPS or the cellular towers around the phones. Location has been proven to be able to predict depressive states [3]. The features extracted from the location sensors included the location variance calculated through the sum of variance in longitude and latitude coordinates, the log of the location variance, the total distance covered, and the circadian movement [17] calculated using the Lomb-Scargle method that maps a person's location patterns following the 24-hour circadian cycle. The speed was also captured as a feature, and static labeled samples were clustered and K-means clustering was used to locate significant places visited by the participants. In addition, location entropy was also engineered to provide the proportion of time spent at each significant location visited during a day.

Screen Sensor Features

Screen sensor features are a strong indicator of how engaged users are with their phones. To capture this information, we extracted features that includes the minimum, maximum, sum, and mean (SD) of unlock episodes, along with the number of all unlock episodes and minutes until the first unlock episode. These features have been used in prior studies that proved their correlation to depressive symptom severity [46,54,55].

Conversation Sensor Features

Conversation is yet another interesting set of features that provide information pertaining to social interactions and has been used in a number of studies relating to mental health [55-58]. The computed features included the minimum, maximum, sum, and mean (SD) of the duration of all conversations. We also recorded the minutes of voice, silence, and noise. The energy associated with noise, which is the L2-norm and the sum of all energy values when noise or voice, was inferred.

Fitbit

Fitbit offers 2 features, which we extracted based on their application in previous studies relating to mental health [54,60,61], and included the maximum resting heart rate (average maximum heart rate over 1 week) and the maximum number of steps (average step count over 1 week). These features provided an idea of the physical movement and stress experienced by participants.

Wi-Fi

Wi-Fi can be a good indicator of social context. We extracted the Wi-Fi count scans that told us the number of scanned Wi-Fi access points connected to by the phone during a time segment and the number of unique connected devices during a time

segment based on the hardware address. In addition, we extracted the most scanned connected device. The use of Wi-Fi-based features in mental health prediction have been previously covered [48,59].

The data set used in our experiments had all the features discussed, which were part of the 61 features. Feature engineering helped provide a context to the data gathered from all the smartphones and Fitbit sensors and form predictions for ML models.

Results of Experiment 1

Experiment 1 showcased the overall performance of the FLMS in comparison with traditional user-agnostic and personalized models. The FLMS achieved a mean accuracy of 0.66 (SD 0.53) and a mean recall of 0.59 (SD 0.50), which are 7% and 13% higher than the best baseline performance achieved by ATU modeling. Among the singular modeling approaches, the ATU, a personalized strategy, performed best overall, with a mean accuracy of 0.59 (SD 0.50) and a mean recall of 0.46 (SD 0.66). The worst performances were shown by user-agnostic LOPO and LTXO approaches, both of which had a mean accuracy of 0.45 (SD 0.80) and 0.47 (SD 0.83), respectively. These results are presented in Table 2 and show that singular modeling approaches used in different studies [1-4,9-17] underperform when modeling involves small, noisy, multimodal sensor data in comparison to our FLMS. The FLMS uses a balance of these strategies to improve predictions.

Experiment 1 was also designed to show how the FLMS suggests the best feature combinations for the various modeling strategies it uses through the utility of hamming distance from stage 3. The lowest hamming distance in stage 3 for the various modeling approaches used is presented in Table 3. We observed that the ATU approach led to the lowest hamming distance of 226, followed by LOTPO, with a minimum hamming distance of 267. The highest hamming distances were those of LOPO at 350 and LTXO at 378. The lower the hamming distance, the closer the predictions to ground truth. Based on this, we saw that overall, 6-sensor fusion works best for this data set. The metadata of experiment 1 are shown in Table 4.

Table 2. Experiment 1 performance of the FLMS[a] in comparison to singular modeling strategies.

Modeling strategy	Type of modeling strategy	Test accuracy, mean (SD)	Test recall, mean (SD)	Test F_1-score, mean (SD)
FLMS	User agnostic + personalized	0.66 (0.53)	0.59 (0.50)	0.56 (0.55)
ATU[b]	Personalized	0.59 (0.60)	0.46 (0.66)	0.50 (0.57)
LOTPO[c]	Personalized	0.53 (0.65)	0.45 (0.70)	0.32 (0.73)
LOPO[d]	User agnostic	0.45 (0.80)	0.43 (0.72)	0.40 (0.87)
LTXO[e]	User agnostic	0.47 (0.83)	0.35 (0.81)	0.33 (0.86)

[a]FLMS: framework for longitudinal multimodal sensors.

[b]ATU: accumulated time unit.

[c]LOTPO: leave one time unit one participant out.

[d]LOPO: leave one participant out.

[e]LTXO: leave time unit X out.

Table 3. Experiment 1 minimum hamming distance for choosing the best sensor combination for the experiment.

Best sensor fusion	Modeling approach in the FLMS[a]	Hamming distance
6-sensor fusion (calls + location + screen usage + conversation + Fitbit + Wi-Fi)	ATU[b]	226
6-sensor fusion (calls + location + screen usage + conversation + Fitbit + Wi-Fi)	LOTPO[c]	267
1-sensor fusion (location)	LOPO[d]	350
2-sensor fusion (calls + location)	LTXO[e]	378

[a]FLMS: framework for longitudinal multimodal sensors.

[b]ATU: accumulated time unit.

[c]LOTPO: leave one time unit one participant out.

[d]LOPO: leave one participant out.

[e]LTXO: leave time unit X out.

Table 4. Experiment 1 metadata.

Metadata	Experiment 1
Feature preprocessing	KNN[a] imputation, dropping highly co-related columns, sklearn StandardScaler
Modeling CV[b] strategy	FLMS[c], ATU[d], LOTPO[e], LTXO[f], LOPO[g]
ML[h] algorithms used	import XGBoost[i] as xgb sklearn.linear_model import LinearRegression sklearn.ensemble import RandomForestRegressor sklearn.linear_model import ElasticNet sklearn.ensemble import GradientBoostingRegressor sklearn.ensemble import ExtraTreesRegressor sklearn.ensemble import AdaBoostRegressor
Random state	42
Evaluation metrics	Accuracy, recall, F_1-score

[a]KNN: K-nearest neighbor.
[b]CV: cross-validation.
[c]FLMS: framework for longitudinal multimodal sensors.
[d]ATU: accumulated time unit.
[e]LOTPO: leave one time unit one participant out.
[f]LTXO: leave time unit X out.
[g]LOPO: leave one participant out.
[h]ML: machine learning.
[i]XGBoost: Extreme Gradient Boosting.

Results of Experiment 2

In experiment 2, we compared FLMS ranking results with ML algorithms that have been used in multiple studies on sensor-based assessment of mental health, as listed in Table 1. The ML algorithms XGBoost and KNN were chosen based on the popularity of their usage in the community, while the DL algorithm was chosen to be a basic multilayer perceptron (MLP) network and a long short-term memory (LSTM) network. These were also the best-performing algorithms compared to other ML algorithms in the literature on our data set. We initially tried using K-fold validation for the SOTA algorithms, but due to poor results, we switched to the leave-one-out strategy, which performed relatively better. This experiment first compared the overall performance of the FLMS with other SOTA algorithms based on the average test accuracy, recall, and F_1-score. Second, the experiment substantiated the claim that the FLMS is better in tackling overfitting, as shown by the mean training accuracy versus the mean test accuracy compared to the ML algorithms in Figure 11. The models with only the single ML algorithm performed no better than the majority baseline approach, with XGBoost showing a mean test accuracy 0.50 (SD 0.55) and the KNN showing around the same mean accuracy of 0.52 (SD 0.54), as shown in Table 5. The MLP achieved higher accuracy but a low test F_1-score, indicating the model's performance has high false-positive and false-negative rates. The LSTM was no different and showed a similar recall and F_1-score outcomes. The overfitting of the SOTA models is illustrated in Figure 11, where we compared the FLMS and the rest of the algorithms based on their respective performances using training and test accuracies. Figure 11 shows that the FLMS had a relatively consistent performance between a training accuracy of 68% and a test accuracy of 66%, while XGBoost, KNN, MLP, and LSTM models had high training accuracies but low test accuracies. The metadata of experiment 2 are shown in Table 6.

The experiments demonstrated support for the points highlighted in the contribution of this paper—that our ranking framework works well with small data sets in comparison to existing approaches and can reduce overfitting by using a balance-weighted ensembling of user-agnostic and personalized models.

Figure 11. Experiment 2 shows FLMS training and test accuracies in comparison to SOTA models. The FLMS is better at adapting to overfitting compared to the other algorithms. FLMS: framework for longitudinal multimodal sensors; KNN: K-nearest neighbor; LSTM: long short-term memory; ML: machine learning; MLP: multilayer perceptron; SOTA: state of the art; XGBoost: Extreme Gradient Boosting.

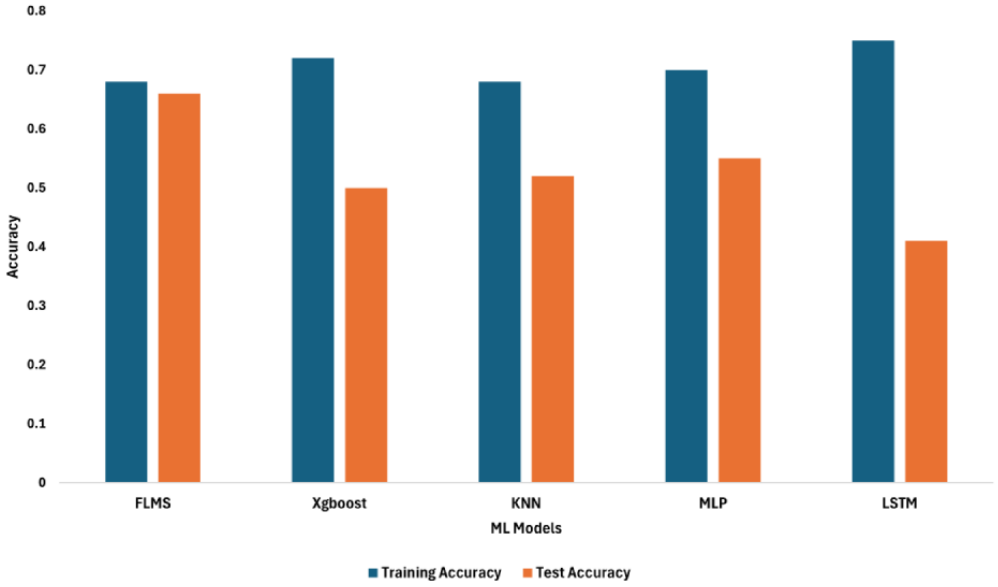


Table 5. Experiment 2 performance of the FLMS[a] compared to ML[b] and DL[c] algorithms used in the current literature on adolescent data.

Predictive learning approach	Modeling strategy	Test accuracy, mean (SD)	Test recall, mean (SD)	Test F_1-score, mean (SD)
FLMS	ATU[d] + LOTPO[e] + LOPO[f] + LTXO[g]	0.66 (0.53)	0.59 (0.50)	0.56 (0.55)
XGBoost[h] [14,17]	Leave 1 out	0.50 (0.55)	0.33 (0.52)	0.28 (0.57)
KNN[i] [10,11,13,16]	Leave 1 out	0.52 (0.54)	0.40 (0.61)	0.30 (0.73)
MLP[j] [9]	Leave 1 out	0.55 (0.70)	0.50 (0.71)	0.33 (0.70)
LSTM[k] [67]	Leave 1 out	0.41 (0.66)	0.25 (0.70)	0.35 (0.70)

[a]FLMS: framework for longitudinal multimodal sensors.
[b]ML: machine learning.
[c]DL: deep learning.
[d]ATU: accumulated time unit.
[e]LOTPO: leave one time unit one participant out.
[f]LOPO: leave one participant out.
[g]LTXO: leave time unit X out.
[h]XGBoost: Extreme Gradient Boosting.
[i]KNN: K-nearest neighbor.
[j]MLP: multilayer perceptron.
[k]LSTM: long short-term memory.

Table 6. Experiment 2 metadata.

Metadata	Experiment 2
Feature preprocessing	KNN[a] imputation, dropping highly co-related columns, sklearn StandardScaler
Modeling CV[b] strategy	FLMS[c], leave 1 out
ML[d] algorithms used	import XGBoost[e] as xgb sklearn.neural_network import MLPClassifier sklearn.neighbors import KNeighborsClassifier keras.layers import LSTM[f]
Random state	42
Evaluation metrics	Accuracy, recall, F_1-Score

[a]KNN: K-nearest neighbor.

[b]CV: cross-validation.

[c]FLMS: framework for longitudinal multimodal sensors.

[d]ML: machine learning.

[e]XGBoost: Extreme Gradient Boosting.

[f]LSTM: long short-term memory.

Results of Experiment 3

Sparsity is a challenge in dealing with small data sets. The large number of 0s or missing values can misdirect models and lead to overfitting [68]. Therefore, it is important to handle the problem of sparsity. Our experiment was designed specifically for small data sets, where sparsity proves to be a challenge. To tackle sparsity in small-data scenarios, the commonly used ML algorithms are KNN, MLP, support vector machine (SVM), decision tree (DT), random forest (RF), XGBoost, and AdaBoost [21-24,69-71].

In our experiment, we showcased a comparison of the FLMS with all the mentioned ML algorithms. We first calculated the sparsity of the adolescent data set that comprised all 6-sensor feature sets. The reason for continuing to use the 6-sensor feature sets as in the prior experiment was to test the algorithms with a data set that had a higher degree of sparsity compared to other feature combinations with lower number of sensors. The sparsity for this data set was calculated as the ratio of 0s to the total number of elements in the data set and is given as follows:

$$\text{Sparsity} = 1 - \frac{\text{Number of nonzero elements}}{\text{Total numberof elements}} \quad \textbf{(5)}$$

The sparsity of the data set used for this experiment was 35%. In a small data set, this is a significant amount of sparsity to negatively impact ML algorithms.

We performed the modeling and evaluated the performance based on F_1-scores as in the case of the prediction of mental health, the F_1-score is a good reflection of how sparsity affects the models' judgment in detecting positive and false cases. The models already shown in Table 4 remained, in addition to other models that have been mentioned in the literature to perform well on sparse data sets. Among the ML algorithms used in the literature, the best performance was shown by the RF, with an F_1-score of 0.35, while the FLMS showed an F_1-score 0.21 higher than that of the RF. Both MLP and AdaBoost performed close to the RF, with an F_1-score of 0.33. The algorithm that performed the worst in handling sparsity was the SVM, with an F_1-score of only 0.15. This experiment highlights the fact that due to the combination of modeling, the FLMS performs better when dealing with highly sparse small data sets (Table 7). The metadata of experiment 3 are shown in Table 8.

Table 7. Experiment 3 performance of the FLMS[a] compared to common ML[b] algorithms for tackling sparsity on the adolescent data set.

Predictive learning approach	Modeling strategy	Test F_1-score, mean (SD)
FLMS	ATU[c] + LOTPO[d] + LOPO[e] + LTXO[f]	0.56 (0.55)
XGBoost[g] [14,17]	Leave 1 out	0.28 (0.57)
KNN[h] [10,11,13,16]	Leave 1 out	0.30 (0.73)
MLP[i] [9]	Leave 1 out	0.33 (0.70)
SVM[j] [12]	Leave 1 out	0.15 (0.62)
DT[k] [13]	Leave 1 out	0.24 (0.70)
RF[l] [11,13]	Leave 1 out	0.35 (0.65)
AdaBoost[m] [14]	Leave 1 out	0.33 (0.60)

[a]FLMS: framework for longitudinal multimodal sensors.
[b]ML: machine learning.
[c]ATU: accumulated time unit.
[d]LOTPO: leave one time unit one participant out.
[e]LOPO: leave one participant out.
[f]LTXO: leave time unit X out.
[g]XGBoost: Extreme Gradient Boosting.
[h]KNN: K-nearest neighbor.
[i]MLP: multilayer perceptron.
[j]SVM: support vector machine.
[k]DT: decision tree.
[l]RF: random forest.
[m]AdaBoost: Adaptive Boosting.

Table 8. Experiment 3 metadata.

Metadata	Experiment 3
Feature preprocessing	KNN[a] imputation, dropping highly corelated columns, sklearn StandardScaler
Modeling CV[b] strategy	FLMS[c], leave 1 out
ML[d] algorithms used	import XGBoost[e] as xgb from sklearn.svm import SVM[f] sklearn.neural_network import MLPClassifier sklearn.neighbors import KNeighborsClassifier sklearn.tree import DecisionTreeClassifier sklearn.ensemble import RandomForestClassifier sklearn.ensemble import AdaBoostClassifier
Random state	42
Evaluation metrics	F_1-score

[a]KNN: K-nearest neighbor.
[b]CV: cross-validation.
[c]FLMS: framework for longitudinal multimodal sensors.
[d]ML: machine learning.
[e]XGBoost: Extreme Gradient Boosting.
[f]SVM: support vector machine.

Discussion

Principal Findings

Solving the problem of limited and sparse data sets is not a singular modeling-based endeavor. It requires flexibility and a combination of strategies to achieve predictions that can be trusted. In this section, we discuss our ranking framework's overarching aims, performance, and limitations based on our assessments.

In experiment 1, we tested the FLMS in comparison to baseline user-agnostic and personalized models. Our framework achieved a higher accuracy, recall, and F_1-score for the predictions when compared to singular modeling approaches, as seen in Table 2. We also demonstrated how we arrived at the sensor combination for the best set of predictions using hamming distances in stage 3 of the FLMS, as reflected in Table 3. In experiment 2, we compared the FLMS with SOTA algorithms used in the literature for predicting mental health states using sensors. The results from this experiment showed the FLMS to perform better than the existing algorithms in terms of accuracy, recall, and F_1-scores (Table 4). Experiment 2 also highlighted the FLMS's ability to reduce overfitting in comparison to the SOTA algorithms. The FLMS showed that the training accuracy and test accuracy did not diverge by large margins, indicating it had not been overfitting the models. Lastly, we compared the FLMS ranking with that of existing ML algorithms that perform well with sparse data in experiment 3. We saw that the data set we used in our experiments exhibited 35% sparsity, which is a significant amount in an already small data set. The FLMS had a higher F_1-score compared to the rest of the ML algorithms.

Comparison With Previous Research

The results of baseline modeling are consistent with previous studies [10,29] that showed superior performance when models were personalized. The increase in accuracy shows that our framework was able to narrow down the best set of predictions overall.

Hamming distance results showed that in LOPO and LTXO approaches, single-sensor deployment and a dual-sensor combination perform equally well as 6-sensor combinations and achieve a minimum hamming distance. This brings forth the advantage of our framework to prioritize sensor selection for yielding best predictions overall and for only the necessary number of feature sets.

The results of experiment 2 provide us with further evidence of the ranking frameworks' efficacy in balancing reliance between both user-agnostic and personalized approaches. Despite a higher accuracy, the recall of the FLMS does not overfit like that of other SOTA ML algorithms. The FLMS uses weights to balance out such effects, thus reducing the impact of overfitting in prediction performance. The test with popular existing ML algorithms showed that, despite the success of the models in previous studies [9-11,13-17], they struggle when the data set is small and noisy, as is the case of the depression data set presented in this work. This performance result is similar when we look at the capability of ML algorithms that are better at handling sparsity. We found the FLMS to perform better than those algorithms.

Overall, seeking a single user-agnostic model that fits all is an elusive problem as most existing works suggest better performance for specialized approaches. However, specialized modeling does not perform well on heterogeneous data sets. Therefore, neither user-agnostic nor personalized modeling alone can be applicable to a specific problem area. Our framework provides a practical way to balance the 2 approaches, particularly for dealing with limited data sets.

Limitations and Future Directions

We encountered a few limitations with this study that can be addressed in future work. The FLMS was tested on the case of depression in adolescents. As such, we have not been able to establish a lower bound on the data set size that our framework is capable of handling.

Another area that we could not elaborate on is the computing speed of such a framework that might be impacted if sensor numbers rise to higher levels. Lastly, the framework was equipped with lightweight and widely used ML algorithms. Methods such as the generalized linear mixed model (GLMM) for handling longitudinal data could not be tested.

Future work can address these limitations with exposure of the framework to more multimodal, longitudinal data sets and adapting and testing other ML algorithms. Interesting future directions for the framework include its online adaptation and a similarity-based cold-start solution.

Conclusion

In this study, we presented a novel prediction-ranking framework for modeling limited noisy or sparse, multimodal, longitudinal passive sensor data. We tested our framework on an adolescent depression data set consisting of 45 participants over a period of 24 weeks. The results showed that despite the complexity and limitations of the data set, our framework is able to provide better predictions compared to singular modeling approaches. In experiment 1, our model achieved a 7% increase in accuracy and a 13% increase in recall. In experiment 2 with synthetic data, our model achieved a 5% increase in accuracy and avoided overestimating the recall value through ensembling predictions. The framework also showed its ability to explore sensor combinations through feature fusion. Our tests with existing popular SOTA algorithms showed that the models struggle when data tend to be limited and noisy. We also tested the FLMS with algorithms that perform well with sparsity and found the FLMS to exhibit a better performance. In conclusion, the FLMS can be an effective tool for passive sensing studies.

Acknowledgments

This study was supported by a grant from the National Institute of Mental Health (NIMH)(1R44MH122067); the NIMH-funded "The Center for Enhancing Triage and Utilization for Depression and Emergent Suicidality (ETUDES) in Pediatric Primary Care"

(P50MH115838); the Center for Behavioral Health, Media, and Technology; and a career development award (NIMH 1K23MH111922-01A1). Research recruitment was supported by the Clinical and Translational Science Institute at the University of Pittsburgh by the National Institutes of Health Clinical and Translational Science Award (CTSA) program (grant UL1 TR001857).

Conflicts of Interest

None declared.

Multimedia Appendix 1

FLMS ranking overview. Algo: algorithm; ATU: accumulated time unit; avg: average; CV: cross-validation; FLMS: framework for longitudinal multimodal sensors; LOPO: leave one participant out; LOTPO: leave one time unit of participant out; LTXO: leave time unit X out; ML: machine learning.
[PNG File , 2313 KB-Multimedia Appendix 1]

References

1. Rabbi M, Pfammatter A, Zhang M, Spring B, Choudhury T. Automated personalized feedback for physical activity and dietary behavior change with mobile phones: a randomized controlled trial on adults. JMIR Mhealth Uhealth. May 14, 2015;3(2):e42. [FREE Full text] [doi: 10.2196/mhealth.4160] [Medline: 25977197]
2. Rabbi M, Aung MH, Zhang M, Choudhury T. MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. 2015. Presented at: UbiComp '15: 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing; September 9-11, 2015:707-718; Osaka, Japan. URL: https://doi.org/10.1145/2750858.2805840 [doi: 10.1145/2750858.2805840]
3. Canzian L, Musolesi M. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. 2015. Presented at: UbiComp '15: 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing; September 9-11, 2015:1293-1304; Osaka, Japan. [doi: 10.1145/2750858.2805845]
4. Difrancesco S, Fraccaro P, van der Veer SN, Alshoumr B, Ainsworth J, Bellazzi R. Out-of-home activity recognition from GPS data in schizophrenic patients. 2016. Presented at: CBMS 2016: IEEE 29th International Symposium on Computer-Based Medical Systems; June 20-24, 2016:324-328; Belfast and Dublin, Ireland. [doi: 10.1109/cbms.2016.54]
5. Sano A, Phillips AJ, Amy ZY, McHill AW, Taylor S, Jaques N, et al. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. 2015. Presented at: BSN 2015: 12th IEEE International Conference on Wearable and Implantable Body Sensor Networks; June 9-12, 2015:1-6; Cambridge, MA. [doi: 10.1109/bsn.2015.7299420]
6. Murahari VS, Plötz T. On attention models for human activity recognition. 2018. Presented at: ISWC '18: 2018 ACM International Symposium on Wearable Computers; October 8-12, 2018:100-103; Singapore. URL: https://doi.org/10.1145/3267242.3267287 [doi: 10.1145/3267242.3267287]
7. Allan S, Henrik B, Sourav B, Thor SP, Mikkel BK, Anind D, et al. Smart devices are different: assessing and mitigating mobile sensing heterogeneities for activity recognition. 2015. Presented at: SenSys '15: 13th ACM Conference on Embedded Networked Sensor Systems; November 1-4, 2015:127-140; Seoul, South Korea. URL: https://doi.org/10.1145/2809695.2809718 [doi: 10.1145/2809695.2809718]
8. Dietterich T. Overfitting and undercomputing in machine learning. ACM Comput Surv. Sep 1995;27(3):326-327. [doi: 10.1145/212094.212114]
9. Kwapisz JR, Weiss GM, Moore SA. Activity recognition using cell phone accelerometers. SIGKDD Explor Newsl. Mar 31, 2011;12(2):74-82. [FREE Full text] [doi: 10.1145/1964897.1964918]
10. Shukla PK, Vijayvargiya A, Kumar R. Human activity recognition using accelerometer and gyroscope data from smartphones. 2020. Presented at: ICONC3: 2020 IEEE International Conference on Emerging Trends in Communication, Control and Computing; February 21-22, 2020; Lakshmangarh, Sikar, India. [doi: 10.1109/iconc345789.2020.9117456]
11. Chen Y, Shen C. Performance analysis of smartphone-sensor behavior for human activity recognition. IEEE Access. 2017;5:3095-3110. [doi: 10.1109/access.2017.2676168]
12. Huang K, Ding X, Xu J, Guanling C, Ding W. Monitoring sleep and detecting irregular nights through unconstrained smartphone sensing. 2015. Presented at: 2015 IEEE UIC-ATC-ScalCom; August 10-14, 2015:10-14; Beijing, China. [doi: 10.1109/uic-atc-scalcom-cbdcom-iop.2015.30]
13. Montanini L, Sabino N, Spinsante S, Gambio E. Smartphone as unobtrusive sensor for real-time sleep recognition. 2018. Presented at: 2018 IEEE International Conference on Consumer Electronics (ICCE); January 12-14, 2018:12-14; Las Vegas. URL: https://doi.org/10.1109/ICCE.2018.8326220 [doi: 10.1109/icce.2018.8326220]
14. Teng F, Chen Y, Cheng Y, Ji X, Zhou B, Xu W. PDGes: an interpretable detection model for Parkinson's disease using smartphones. ACM Trans Sen Netw. Apr 20, 2023;19(4):1-21. [FREE Full text] [doi: 10.1145/3585314]

15. Azam M, Shahzadi A, Khalid A, Anwar S, Naeem U. Smartphone based human breath analysis from respiratory sounds. 2018. Presented at: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 17-21, 2018:445-448; Honolulu, HI. [doi: 10.1109/embc.2018.8512452]
16. Grunerbl A, Osmani V, Bahle G, Carrasco JC, Oehler S, Mayora O, et al. Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. 2014. Presented at: AH '14: 5th Augmented Human International Conference; March 7-9, 2014:1-8; Kobe, Japan. [doi: 10.1145/2582051.2582089]
17. Saeb S, Lattie EG, Kording KP, Mohr DC. Mobile phone detection of semantic location and its relationship to depression and anxiety. JMIR Mhealth Uhealth. Aug 10, 2017;5(8):e112. [FREE Full text] [doi: 10.2196/mhealth.7297] [Medline: 28798010]
18. Plötz T. If only we had more data!: sensor-based human activity recognition in challenging scenarios. 2023. Presented at: 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops); March 13-17, 2023:565-570; Atlanta, GA. [doi: 10.1109/percomworkshops56833.2023.10150267]
19. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. J Biomed Inform. Jan 2018;77:120-132. [FREE Full text] [doi: 10.1016/j.jbi.2017.12.008] [Medline: 29248628]
20. Xu X, Mankoff J, Dey A. Understanding practices and needs of researchers in human state modeling by passive mobile sensing. CCF Trans Pervasive Comp Interact. Jul 06, 2021;3(4):344-366. [FREE Full text] [doi: 10.1007/s42486-021-00072-4]
21. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016. Presented at: KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016:785-794; San Francisco, CA. [doi: 10.1145/2939672.2939785]
22. Xi Y, Xiang Z, Ramadge P, Schapire R. Speed and sparsity of regularized boosting. PMLR. 2009;5:615-622.
23. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc B: Stat Methodol. Apr 2005;67(2):301-320. [FREE Full text] [doi: 10.1111/j.1467-9868.2005.00503.x]
24. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. Mar 2, 2006;63(1):3-42. [doi: 10.1007/s10994-006-6226-1]
25. Muhammad G, Alshehri F, Karray F, Saddik AE, Alsulaiman M, Falk TH. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. Inf Fusion. Dec 2021;76:355-375. [FREE Full text] [doi: 10.1016/j.inffus.2021.06.007]
26. Joshi S, Boyd S. Sensor selection via convex optimization. IEEE Trans Signal Process. Feb 2009;57(2):451-462. [doi: 10.1109/TSP.2008.2007095]
27. Altenbach F, Corroy S, Böcherer G, Mathar R. Strategies for distributed sensor selection using convex optimization. 2012. Presented at: 2012 IEEE Global Communications Conference (GLOBECOM); December 3-7, 2012:2367-2372; Anaheim, CA. [doi: 10.1109/glocom.2012.6503470]
28. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. Jul 6, 2019;6(1):1-48. [FREE Full text] [doi: 10.1186/s40537-019-0197-0]
29. Feng SY, Gangal V, Wei J, Chandar S, Vosoughi S, Mitamura T, et al. A survey of data augmentation approaches for NLP. arXiv. Preprint posted online 2021. [doi: 10.48550/arXiv.2105.03075]. 2021. [FREE Full text] [doi: 10.48550/arXiv.2105.03075]
30. Florez AYC, Scabora L, Amer-Yahia S, Júnior JFR. Augmentation techniques for sequential clinical data to improve deep learning prediction technique. 2020. Presented at: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS); July 28-30, 2020:597-602; Rochester, MN. URL: https://doi.org/10.1109/CBMS49503.2020.00118 [doi: 10.1109/cbms49503.2020.00118]
31. Müller SR, Chen XL, Peters H, Chaintreau A, Matz SC. Depression predictions from GPS-based mobility do not generalize well to large demographically heterogeneous samples. Sci Rep. Jul 07, 2021;11(1):14007. [FREE Full text] [doi: 10.1038/s41598-021-93087-x] [Medline: 34234186]
32. Xu X, Chikersal P, Dutcher JM, Sefidgar YS, Seo W, Tumminia MJ, et al. Leveraging collaborative-filtering for personalized behavior modeling: a case study of depression detection among college students. Proc ACM Interact Mob Wearable Ubiquitous Technol. Mar 30, 2021;5(1):1-27. [FREE Full text] [doi: 10.1145/3448107]
33. Maxhuni A, Hernandez-Leal P, Sucar LE, Osmani V, Morales EF, Mayora O. Stress modelling and prediction in presence of scarce data. J Biomed Inform. Oct 2016;63:344-356. [FREE Full text] [doi: 10.1016/j.jbi.2016.08.023] [Medline: 27592309]
34. Jacobson N, Lekkas D, Huang R, Thomas N. Deep learning paired with wearable passive sensing data predicts deterioration in anxiety disorder symptoms across 17-18 years. J Affect Disord. Mar 01, 2021;282:104-111. [FREE Full text] [doi: 10.1016/j.jad.2020.12.086] [Medline: 33401123]
35. Ren B, Balkind EG, Pastro B, Israel ES, Pizzagalli DA, Rahimi-Eichi H, et al. Predicting states of elevated negative affect in adolescents from smartphone sensors: a novel personalized machine learning approach. Psychol Med. Jul 27, 2022;53(11):5146-5154. [doi: 10.1017/s0033291722002161]
36. Adhikary A, Majumder K, Chatterjee S, Shaw RN, Ghosh A. Human activity recognition for disease detection using machine learning techniques—a comparative study. In: Shaw RN, Das S, Piuri V, Bianchini M, editors. Advanced Computing and Intelligent Technologies. Lecture Notes in Electrical Engineering, Vol 914. Singapore. Springer; 2022.

37. Messalas A, Kanellopoulos Y, Makris C. Model-agnostic interpretability with Shapley values. 2019. Presented at: IISA 2019: 10th IEEE International Conference on Information, Intelligence, Systems and Applications; July 15-17, 2019:1-7; Patras, Greece. [doi: 10.1109/iisa.2019.8900669]
38. Li L, Qiao J, Yu G, Wang L, Li HY, Liao C, et al. Interpretable tree-based ensemble model for predicting beach water quality. Water Res. Mar 01, 2022;211:118078. [FREE Full text] [doi: 10.1016/j.watres.2022.118078] [Medline: 35066260]
39. Debie E, Fernandez Rojas R, Fidock J, Barlow M, Kasmarik K, Anavatti S, et al. Multimodal fusion for objective assessment of cognitive workload: a review. IEEE Trans Cybern. Mar 2021;51(3):1542-1555. [doi: 10.1109/tcyb.2019.2939399]
40. Pedrelli P, Fedor S, Ghandeharioun A, Howe E, Ionescu DF, Bhathena D, et al. Monitoring changes in depression severity using wearable and mobile sensors. Front Psychiatry. Dec 18, 2020;11:584711. [FREE Full text] [doi: 10.3389/fpsyt.2020.584711] [Medline: 33391050]
41. Wang R. On predicting relapse in schizophrenia using mobile sensing in a randomized control trial. 2020. Presented at: 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom); March 23-27, 2020:1-8; Austin, TX. [doi: 10.1109/percom45495.2020.9127365]
42. Sun S, Folarin AA, Zhang Y, Cummins N, Garcia-Dias R, Stewart C, et al. RADAR-CNS Consortium. Challenges in using mHealth data from smartphones and wearable devices to predict depression symptom severity: retrospective analysis. J Med Internet Res. Aug 14, 2023;25:e45233. [FREE Full text] [doi: 10.2196/45233] [Medline: 37578823]
43. Tlachac M, Toto E, Lovering J, Kayastha R, Taurich N, Rundensteiner E. EMU: early mental health uncovering framework and dataset. 2021. Presented at: ICMLA 2021: 20th IEEE International Conference on Machine Learning and Applications; December 13-16, 2021:1311-1318; Pasadena, CA. [doi: 10.1109/icmla52953.2021.00213]
44. Negeri ZF, Levis B, Sun Y, He C, Krishnan A, Wu Y, et al. Depression Screening Data (DEPRESSD) PHQ Group. Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: updated systematic review and individual participant data meta-analysis. BMJ. Oct 05, 2021;375:n2183. [FREE Full text] [doi: 10.1136/bmj.n2183] [Medline: 34610915]
45. Vega J, Li M, Aguillera K, Goel N, Joshi E, Khandekar K, et al. Reproducible analysis pipeline for data streams: open-source software to process data collected with mobile devices. Front Digit Health. Nov 18, 2021;3:769823. [FREE Full text] [doi: 10.3389/fdgth.2021.769823] [Medline: 34870271]
46. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res. Jul 15, 2015;17(7):e175. [FREE Full text] [doi: 10.2196/jmir.4273] [Medline: 26180009]
47. Wang R, Wang W, daSilva A, Huckins JF, Kelley WM, Heatherton TF, et al. Tracking depression dynamics in college students using mobile phone and wearable sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol. Mar 26, 2018;2(1):1-26. [doi: 10.1145/3191775]
48. Wahle F, Kowatsch T, Fleisch E, Rufer M, Weidt S. Mobile sensing and support for people with depression: a pilot trial in the wild. JMIR Mhealth Uhealth. Sep 21, 2016;4(3):e111,e5960. [FREE Full text] [doi: 10.2196/mhealth.5960] [Medline: 27655245]
49. Mehrotra A, Musolesi M. Using autoencoders to automatically extract mobility features for predicting depressive states. Proc ACM Interact Mob Wearable Ubiquitous Technol. Sep 18, 2018;2(3):1-20. [doi: 10.1145/3264937]
50. Farhan AA, Yue C, Morillo R, Ware S, Lu J, Bi J, et al. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. 2016. Presented at: 2016 IEEE Wireless Health; October 25-27, 2016:30-37; Bethesda, MD. [doi: 10.1109/wh.2016.7764553]
51. Chikersal P, Doryab A, Tumminia M, Villalba DK, Dutcher JM, Liu X, et al. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: a machine learning approach with robust feature selection. ACM Trans Comput-Hum Interact. Jan 20, 2021;28(1):1-41. [doi: 10.1145/3422821]
52. Lane ND, Lin M, Mohammod M, Yang X, Lu H, Cardone G, et al. BeWell: sensing sleep, physical activities and social interactions to promote wellbeing. Mobile Netw Appl. Jan 9, 2014;19(3):345-359. [FREE Full text] [doi: 10.1007/s11036-013-0484-5]
53. LiKamWa R, Liu Y, Lane N, Zhong L. MoodScope: building a mood sensor from smartphone usage patterns. 2013. Presented at: MobiSys'13: 11th Annual International Conference on Mobile Systems, Applications, and Services; June 25-28, 2013:25-28; Taipei, Taiwan. [doi: 10.1145/2462456.2464449]
54. Doryab A, Villalba DK, Chikersal P, Dutcher JM, Tumminia M, Liu X, et al. Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: statistical analysis, data mining and machine learning of smartphone and Fitbit data. JMIR Mhealth Uhealth. Jul 24, 2019;7(7):e13209. [FREE Full text] [doi: 10.2196/13209] [Medline: 31342903]
55. Wang R, Aung MSH, Abdullah S. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. 2016. Presented at: UbiComp '16: 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing; September 12-16, 2016; Heidelberg, Germany. [doi: 10.1145/2971648.2971740]
56. Lane N, Rabbi M, Lin M, Yang X. Bewell: a smartphone application to monitor, model and promote wellbeing. 2012. Presented at: 5th International ICST Conference on Pervasive Computing Technologies for Healthcare; May 23-26, 2011; Dublin, Ireland. [doi: 10.4108/icst.pervasivehealth.2011.246161]

57. Mashfiqui R, Ali S, Choudhury T, Berke E. Passive and in-situ assessment of mental and physical well-being using mobile sensors. 2011. Presented at: UbiComp '11: 13th International Conference on Ubiquitous Computing; September 17-21, 2011:385-394; Beijing, China. [doi: 10.1145/2030112.2030164]
58. Wang R, Chen F, Chen Z. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. 2014. Presented at: UbiComp '14: 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing; September 13-17, 2014:3-14; Seattle, WA. [doi: 10.1145/2632048.2632054]
59. Ware S, Yue C, Morillo R, Lu J, Shang C, Kamath J, et al. Large-scale automatic depression screening using meta-data from WiFi infrastructure. Proc ACM Interact Mob Wearable Ubiquitous Technol. Dec 27, 2018;2(4):1-27. [doi: 10.1145/3287073]
60. Dai R, Kannampallil T, Kim S. Detecting mental disorders with wearables: a large cohort study. 2023. Presented at: IoTDI '23: 8th ACM/IEEE Conference on Internet of Things Design and Implementation; May 9-12, 2023:39-51; San Antonio, TX. [doi: 10.1145/3576842.3582389]
61. Doryab A, Chikarsel P, Liu X, Dey AK. Extraction of behavioral features from smartphone and wearable data. arXiv. Preprint posted online 2018. [doi: 10.48550/arXiv.1812.10394]. 2021. [doi: 10.48550/arXiv.1812.10394]
62. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord. Apr 2009;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
63. Hamilton M. The Hamilton Rating Scale for depression. In: Assessment of Depression. Berlin, Heidelberg. Springer; 1986:143-152.
64. Thompson E. Hamilton Rating Scale for anxiety (HAM-A). Occup Med (Lond). Oct 13, 2015;65(7):601. [doi: 10.1093/occmed/kqv054] [Medline: 26370845]
65. Schelter S, Böse JH, Kirschnick J, Klein T, Seufert S. Automatically tracking metadata and provenance of machine learning experiments. Amazon Science. 2017. URL: https://assets.amazon.science/2f/39/4b32cf354e4c993b439d88258597/automatically-tracking-metadata-and-provenance-of-machine-learning-experiments.pdf [accessed 2024-05-01]
66. Ferreira D, Kostakos V, Dey AK. AWARE: mobile context instrumentation framework. Front ICT. Apr 20, 2015;2:6. [doi: 10.3389/fict.2015.00006]
67. Acikmese Y, Alptekin SE. Prediction of stress levels with LSTM and passive mobile sensors. Procedia Comput Sci. 2019;159:658-667. [doi: 10.1016/j.procs.2019.09.221]
68. Kucukozer-Cavdar S, Taskaya-Temizel T, Mehrotra A, Musolesi M, Tino P. Designing robust models for behaviour prediction using sparse data from mobile sensing: a case study of office workers' availability for well-being interventions. ACM Trans Comput Healthc. Jul 18, 2021;2(4):1-33. [FREE Full text] [doi: 10.1145/3458753]
69. Yin D, Li J, Wu G. Solving the data sparsity problem in predicting the success of the startups with machine learning methods. arXiv. Preprint posted online 2021. [doi: 10.48550/arXiv.2112.07985]. 2021;07985(2021). [doi: 10.48550/arXiv.2112.07985]
70. Zhang M, Sun Y, Liang F. Sparse deep learning for time series data: theory and applications. arXiv. Preprint posted online 2023. [doi: 10.48550/arXiv.2310.03243]. 2021. [doi: 10.48550/arXiv.2310.03243]
71. Rosidi N. Best machine learning model for sparse data. KD nuggets. Apr 7, 2023. URL: https://www.kdnuggets.com/2023/04/best-machine-learning-model-sparse-data.html [accessed 2024-05-01]

Abbreviations

AdaBoost: Adaptive Boosting
API: application programming interface
ATU: accumulated time unit
CV: cross-validation
DL: deep learning
DT: decision tree
FLMS: framework for longitudinal multimodal sensors
HAR: human activity recognition
KNN: K-nearest neighbor
LDA: linear discriminant analysis
LOPO: leave one participant out
LOTPO: leave one time unit one participant out
LR: linear regression
LSTM: long short-term memory
LTXO: leave time unit X out
ML: machine learning
MLP: multilayer perceptron
PCA: principal component analysis
PHQ-9: 9-item Patient Health Questionnaire
RF: random forest

SOTA: state of the art
SVM: support vector machine
XGBoost: Extreme Gradient Boosting

Edited by Y Huo; submitted 02.04.23; peer-reviewed by L Zheng, A Tomar; comments to author 02.07.23; revised version received 16.09.23; accepted 09.04.24; published 20.05.24

Please cite as:
Mullick T, Shaaban S, Radovic A, Doryab A
Framework for Ranking Machine Learning Predictions of Limited, Multimodal, and Longitudinal Behavioral Passive Sensing Data: Combining User-Agnostic and Personalized Modeling
JMIR AI 2024;3:e47805
URL: https://ai.jmir.org/2024/1/e47805
doi: 10.2196/47805
PMID: