

# UC San Diego

## UC San Diego Previously Published Works

### Title

Dynamic clustering via branched deep learning enhances personalization of stress prediction from mobile sensor data

### Permalink

<https://escholarship.org/uc/item/60r70375>

### Journal

Scientific Reports, 14(1)

### ISSN

2045-2322

### Authors

Luo, Yunfei

Deznabi, Iman

Shaw, Abhinav

et al.

### Publication Date

2024-03-19

### DOI

10.1038/s41598-024-56674-2

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

Peer reviewed



# OPEN Dynamic clustering via branched deep learning enhances personalization of stress prediction from mobile sensor data

Yunfei Luo<sup>1,2,4</sup>, Iman Deznabi<sup>1,4</sup>, Abhinav Shaw<sup>1,3</sup>, Natcha Simsiri<sup>1</sup>, Tauhidur Rahman<sup>2</sup> & Madalina Fiterau<sup>1</sup>

College students experience ever-increasing levels of stress, leading to a wide range of health problems. In this context, monitoring and predicting students' stress levels is crucial and, fortunately, made possible by the growing support for data collection via mobile devices. However, predicting stress levels from mobile phone data remains a challenging task, and off-the-shelf deep learning models are inapplicable or inefficient due to data irregularity, inter-subject variability, and the "cold start problem". To overcome these challenges, we developed a platform named Branched CALM-Net that aims to predict students' stress levels through dynamic clustering in a personalized manner. This is the *first platform that leverages the branching technique in a multitask setting to achieve personalization and continuous adaptation*. Our method achieves state-of-the-art performance in predicting student stress from mobile sensor data collected as part of the Dartmouth StudentLife study, with a ROCAUC 37% higher and a PRAUC surpassing that of the nearest baseline models. In the cold-start online learning setting, Branched CALM-Net outperforms other models, attaining an average F1 score of 87% with just 1 week of training data for a new student, which shows it is reliable and effective at predicting stress levels from mobile data.

**Keywords** Machine learning, Multitask learning, Stress prediction, Mobile computing

Stress is one of the most common contributors to widespread health problems, making its early identification and continuous monitoring essential for the development of effective treatments and interventions<sup>1</sup>. Chronic stress often causes anxiety disorders, which affect almost 30% of adults at some point in their lives<sup>2</sup>. In addition, prior research suggests that overwhelming stress can lead to deadly conditions such as cardiovascular diseases<sup>3,4</sup>, memory and cognition impairment<sup>5</sup>, and suppression of the immune system<sup>6</sup>. Furthermore, stress has been shown to aggravate metabolic dysfunctions, including insulin resistance as well as disruptions in glucose and lipid homeostasis<sup>7</sup>. For college students in particular, stress has a negative effect both academically and socially<sup>1</sup>.

Due to the serious impact of stress on college students, it would be extremely useful if their level of stress could be continuously monitored by the school's health staff and the students themselves, to facilitate interventions designed to preempt potential health problems. Since it is not feasible for the staff to perform clinical tests for all the students, such as measuring cortisol levels<sup>8</sup> at regular intervals throughout the day, using data collected continuously via cell phones and having a system to passively predict users' levels of stress is a cost-efficient and desired alternative. Studies focusing on this approach<sup>9,10</sup> have shown that machine learning has a tremendous potential applicability<sup>11</sup> to predict stress level, as well as other mental conditions<sup>10,12–14</sup> with recent attempts focusing on the use of deep learning methods<sup>11,15–17</sup>. Among deep learning methods, Long Short-Term Memory (LSTM) networks<sup>18</sup> and attention-based models<sup>19</sup> have been the top performers in tasks involving various types of time series data<sup>20–26</sup>. However, there are some major challenges specific to the stress prediction task, including intersubject variability, cold start problem, and data irregularity, which make prior methods ill-suited to this task. We will discuss each of the challenges in detail and present how our proposed method is designed to overcome each of them.

<sup>1</sup>Manning College of Information and Computer Science, University of Massachusetts Amherst, 140 Governors Drive, Amherst, MA 01003, USA. <sup>2</sup>Halicioğlu Data Science Institute, University of California San Diego, 9500 Gilman Dr, San Diego, CA 92093, USA. <sup>3</sup>Computer Science, Stanford University, 450 Jane Stanford Way, Stanford, CA 94305, USA. <sup>4</sup>These authors contributed equally: Yunfei Luo and Iman Deznabi. ✉email: yul268@ucsd.edu

An important problem is that most standard machine learning models assume homogeneity in the data, implying that all collected samples originate from the same distribution. However, the way people experience mental conditions and specifically *stress levels vary significantly from individual to individual*<sup>27</sup>. To address this challenge, personalized models have been introduced and explored<sup>17,28–30</sup>. Techniques involving personalized parameters have been used for a wide range of applications, from intelligent sensing and health status monitoring<sup>31</sup>, to next-word prediction on mobile devices<sup>32</sup>. For the prediction of mental state, recent work such as<sup>17</sup> and<sup>28</sup> have shown the importance of considering individual differences and constraints in modeling. We adopt *branched deep learning as a new approach to personalization* and experimentally show that our model attains the best performance at predicting student stress levels with limited data.

Despite promising performance, personalized models often suffer from the “cold-start” problem<sup>33</sup>. The cold start problem occurs when new subjects are introduced into a predictive model with personalized parameters. Because the model lacks data on the new subject, it cannot immediately provide a reliable prediction for that subject. Ideally, the new subject could be assigned to a group of subjects who share similar patterns and have well-trained parameters to obtain a reasonable prediction immediately. Previous works in this area<sup>11,17,28–30</sup> neither evaluate the performance of their proposed system in this cold start scenario nor explore to what extent the parameters should be shared between subjects. To achieve information sharing, we will use a technique called Learn-to-Branch, which was introduced in<sup>34,35</sup> for Computer Vision applications to find the best paths in neural networks for each task so that the information learned by some particular node(s) will only be shared with similar tasks. To the best of our knowledge, *we are the first to apply Learn-to-Branch to time-series data and use it for personalization*.

Another characteristic of time series data for this task is that they are usually *collected irregularly over time with different features collected with different frequencies*<sup>36–39</sup>. There are many different data modalities that are useful in student stress prediction, including audio, GPS, movement, phone usage statistics, and the duration of sleep and academic-dependent data which may include indicators for the exam period, and the deadline for the next assignment. All of these modalities have been shown to be useful<sup>40</sup> for student stress prediction. Che et al.<sup>39</sup> and Shukla et al.<sup>22</sup> provide an overview of the approaches on modeling data with irregularity using deep learning methods. Here, we adopt a feature engineering method to *regularize the sample rate* followed by dimensionality reduction<sup>41</sup>. This technique is also employed to *mitigate noise in the data* used to train the classification model for predicting student stress levels.

To assess the efficacy of our models and position them within the context of existing research, we used the StudentLife dataset<sup>40</sup> as our primary data source. Our selection of the StudentLife dataset stemmed from its distinct advantage: an extensive data collection period of approximately 3 months. This prolonged duration stands out in comparison to most depression monitoring studies, which typically gather data over only 1–2 weeks, as acknowledged in<sup>42</sup>. This extended timeframe allowed us to encompass data from students at various stages, including periods characterized by diverse course assignments and examination schedules.

Moreover, the dataset offered the opportunity to incorporate multimodal features into our modeling process. These encompassed physical activity, phone usage, sleep statistics, conversation records, and more. This marked a departure from previous studies that primarily focused on single features, such as location or sociability. By harnessing multiple features simultaneously to train our models, we demonstrated the superior performance of our multimodal learning approach compared to earlier state-of-the-art work, as exemplified by<sup>11</sup>, which primarily concentrated on location-based features.

In the literature on the StudentLife dataset, several significant contributions stand out as valuable benchmarks for comparison. Gatis et al.<sup>11</sup> introduced an approach named Multilayer Perceptron based on location features (Location MLP), leveraging novel location-based features extracted from the StudentLife dataset, to make stress level predictions. Similarly, Adler et al.<sup>43</sup> employed Gradient Boosted Decision Trees (GBDT) augmented with statistical features to predict stress levels within the same dataset.

Additionally, the work by<sup>44</sup> unveiled the PSP-IGR model, which combined a user-specific information-enriched multi-layered perception (MLP) for low-frequency sensor data with a CNN-LSTM model tailored for high-frequency sensor data. Alongside,<sup>45</sup> demonstrated the utility of XGBoost in stress prediction, enhancing feature potency through the application of Synthetic Minority Oversampling Technique (SMOTE) (Despite exerting diligent efforts, the source codes for the PSP-IGR model<sup>44</sup> and the work by<sup>45</sup> eluded our reach. Despite direct communication with the authors and exhaustive online searches, the unavailability of these codes prevented us from conducting a holistic comparison between our results and these two approaches.)

The method in<sup>43</sup> achieved state-of-the-art performance for binary and three-level stress prediction using five-fold cross-validation. Nevertheless, this work does not take inter-subject variability into account, and the evaluation under the cold-start and continuous learning settings is lacking.

In this work, we propose a method based on hierarchical multitask learning for predicting stress from smartphone data. Our approach addresses each of the identified problems through careful method selection and empirical validation. The standard version of our proposed method is called the Cross-personal Activity LSTM Multitask Network (CALM-Net), a system designed specifically to deal with inter-subject variability. CALM-Net learns personalized parameters by considering each subject as a task in a multitask learning structure. With shared and individual parts, CALM-Net can learn both the shared dynamics of the data between all subjects as well as personalized characteristics. We also propose Branched CALM-Net, a variant of CALM-Net designed to find and share parts of the network among similar subjects. Branched CALM-Net combines multitask learning with the Learn-to-Branch technique to efficiently model both personal and population-level characteristics. Similarly, we develop CATrans-Net and Branched CATrans-Net which use a Transformer model<sup>46</sup> instead of the LSTM model in the architecture.

We show that Branched CALM-Net achieves *state-of-the-art performance in student stress prediction from as little as 1 week of training data* for an individual subject. We also show the stability and robustness of our system

under cold start scenarios and online learning settings when a subject starts with no data and more data are added over time. Specifically, Branched CALM-Net achieves an F1 score of around 0.67 for 3-class stress prediction for a new subject who has only the first week of training data available. With more training data included for a subject, Branched CALM-Net remains the top performer among the tested methods. Thus, Branched CALM-Net enables the accurate detection of high levels of stress from passively collected smartphone data. This detection is a key prerequisite for enabling the deployment of early intervention systems to avoid the development of serious health problems and mitigate the influence of high stress on students' academic performance.

## Results

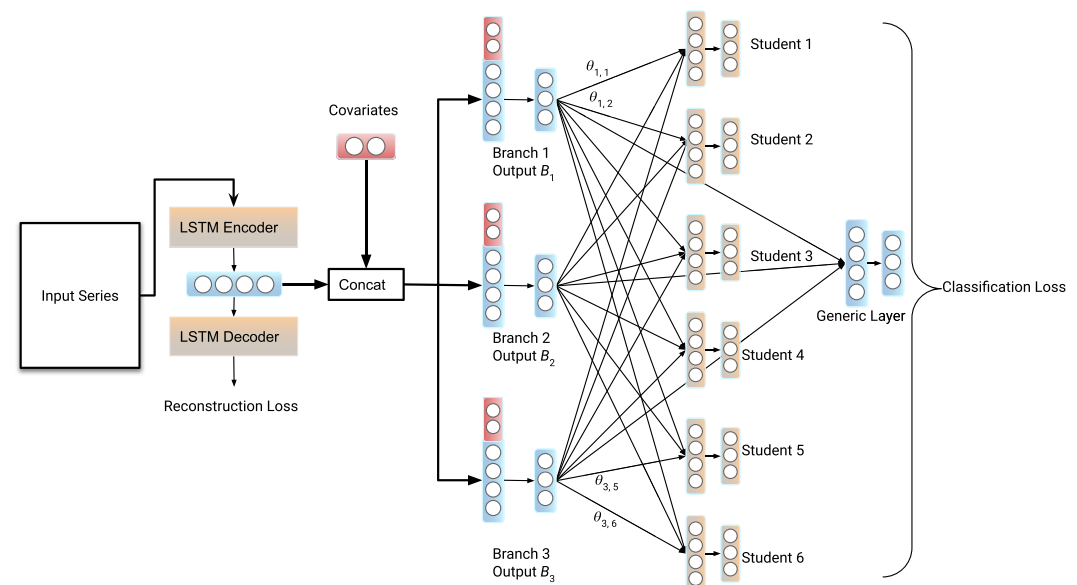
### Experimental setup

The StudentLife data set provides a benchmark for research on stress monitoring. Data have been collected from college students at Dartmouth College, over 9 weeks, containing psychological surveys that include stress level queries, GPS locations, phone usage statistics, audio, and physical activity inference, which are the indicators representing whether the motion is slight or intense based on the raw signal. Table 8 shows the features we choose along with their pre-processing methods. This selection strategy was implemented to create a balanced dataset, thereby minimizing data skewness. By doing so, we aimed to align the total number of samples with the figures reported in the Location MLP baseline study, ensuring consistency and reliability in our data analysis. We observed that many students had less than 35 days with self-reported stress level, which could potentially result in an under-representation of their stress patterns. Simultaneously, one student had more than 200 records of self-reported stress level with which is significantly more data than the others. To avoid data skewness and ensure a balanced analysis, we selected students with a data range that was neither too short nor excessively long – more than 35 days and less than 150 days with self-reported stress records.

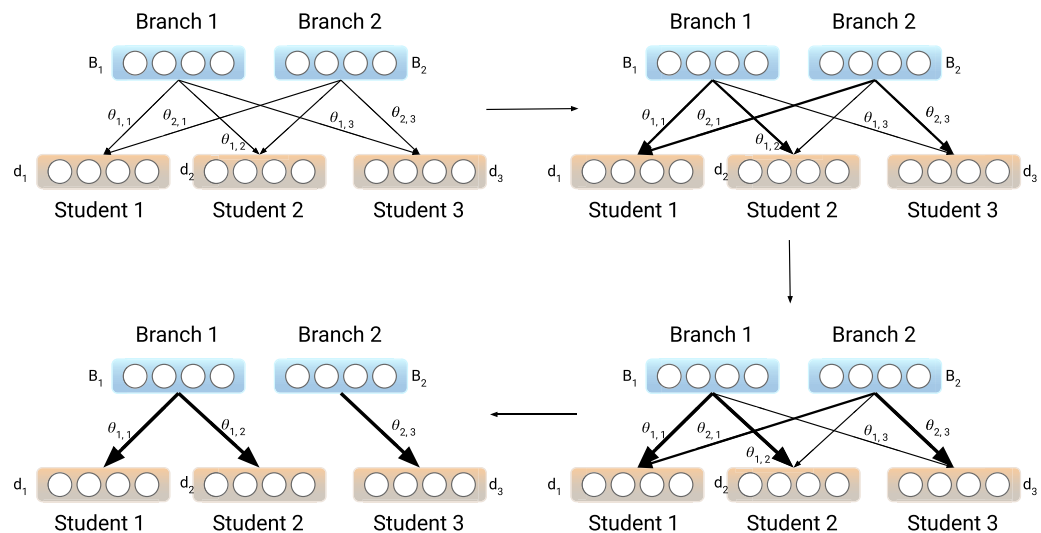
There are five stress labels in this data set: (1) no stress at all, (2) feeling well, (3) a little stressed, (4) stressed, (5) very stressed. Due to the high imbalance of the number of labels, we follow<sup>11</sup> to combine the first two labels as (1) *below median* or *no stress*, the third label as (2) *median stress*, and the last two labels as (3) *above median*.

We evaluate the baselines and our model, CALM-Net (Fig. 1) with Learn-to-Branch (Fig. 2), in two scenarios. First, we consider the case when there is sufficient data for all students to train the models. For this evaluation, we performed five-fold cross-validation of the data, for which the entire dataset is split into five subsets stratified by each student and each stress level. The temporal order of the samples is not important here because the prediction is made with data from a single day as input, meaning the prediction process is independent with respect to data from past and future days. Although the evaluation schema is theoretically sound, it does not account for the practical limitation that future data is unavailable in our training datasets. To address this and validate our model in more realistic scenarios, we performed an additional round of cross-validation. This involved segmenting the data chronologically, ensuring that our model's effectiveness is tested under conditions that closely resemble real-world applications.

The second scenario corresponds to the cold-start problem, where we conduct leave-one-subject-out cross-validation. In this evaluation, we are simulating the situation where we have a dataset for offline modeling which includes data from all but one of the students, then we conduct prediction on the new student, the left out one,



**Figure 1.** An overview of the structure of CALM-Net enhanced with Learn-to-Branch technique, forming the Branched CALM-Net model. The network consists of an LSTM autoencoder, followed by branches, and multi-layer outputs. The structure includes a generic layer set which models information for all subjects, and separate layers for each subject, obtained by adaptively combining information from multiple branches.



**Figure 2.** Pipeline of Learning to Branch. The probabilistic weights are uniformly initialized, and the sampling strategy follows the Gumbel-Softmax technique. The branching weights will finally converge to one-hot vectors by using Gumbel-Softmax trick. This branching mechanism can control the information shared between the layers for each student.

for whom we have no data. We also test the methods in the online learning setting, in which data from the new student is progressively added for training. This last evaluation simulates the scenario after the cold start, determining the amount of student monitoring time needed by each method to be able to predict the stress levels for a new student.

We use the Location MLP<sup>11</sup>, one of the previous state-of-the-art models, as a baseline to ascertain the competence of our model for stress prediction in the StudentLife data set. We also compare against GBDT with statistical features, as used in<sup>43</sup> for EMA prediction on StudentLife data.

In the cold-start-with-online-learning setting we also compare our results against another baseline: a neural network model that uses student clusters as tasks (named Clustered CALM-Net). To find these clusters, we use the survey scores to group the students, where the scores are computed from students' responses to psychological surveys: PHQ-9 for depression, Perceived stress scale (PSS) for stress, Longiness scale, Flourish scale, Positive and Negative Affect Schedule (PANAS), and Big Five Personality Traits.

## Evaluation

We use two different evaluation schemes to benchmark the performance of our models against the baselines. First, we use five-fold cross-validation on the time series of each student and report the resulting average F1 score and Area Under the Receiver Operating Characteristics Curve (AUC)<sup>47,48</sup>. Because the stress labels are very imbalanced (with portions of 22.1%, 43.4%, 34.5% for *below median*, *median*, *above median* respectively), both the F1 and AUC scores are appropriate for evaluation in such settings<sup>47,49</sup>. This approach of splitting each student's data into five-folds simulates the situation of warm-start when there is relatively sufficient data available to train personalized models for all the subjects.

We present evaluation results not only on the 3 classes stress prediction which test the models' ability of differentiating levels of stress, but also on binary detection that test the performance of models on identifying positive (combining labels of *median* and *above median*) and negative cases (*below median*). We validate our models under these two settings to demonstrate that our methods can meet a range of real-life requirements. Conducting binary stress detection is a basic requirement for the application. However, if the school's mental health staff or students themselves are interested in having different treatments for moderate and high levels of stress, a more fine-grained prediction could be provided with the three-class setting. In this scenario, our approach could provide a more detailed assessment of stress levels, allowing for tailored interventions and support for those in need.

However, as mentioned, having sufficient data is not always the case. For instance, a new student who joined the system will have no or limited training data. The cold-start scenario is important for personalized models and has not been previously considered in prior research on stress prediction<sup>16,17,28</sup>. To evaluate our models in this scenario, we will use a leave-one-subject-out approach, which means that we use one student as our test data in each iteration of evaluation. We report the F1-score, averaged over all students. Furthermore, to simulate the situation where the model continuously collects data about a subject as the subject uses the mobile app, we will train the model on a growing portion of data from the test student in each iteration and test it on the rest of available data from the test student. In the rest of this section, we will discuss the results achieved by our models and baselines in these scenarios. A summary of the evaluation schemes is shown in Table 1.

| Scenarios   | Binary stress detection                                                                                                                                                                                             | Stress level prediction                                                                                                                                                                                                                | Stress level prediction with cold start                                                                                                                                                                     |
|-------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Labels      | Not stressed, stressed                                                                                                                                                                                              | Below median, median, above median                                                                                                                                                                                                     | Below median, median, above median                                                                                                                                                                          |
| Evaluations | Five-fold cross-validation                                                                                                                                                                                          | Five-fold cross-validation                                                                                                                                                                                                             | Leave-one-subject-out validation                                                                                                                                                                            |
| Sections    | "CALM-Net and branched CALM-Net attain AUC scores of more than 0.8 on student stress detection" section. On binary stress detection, CALM-Net attains precisions higher than 84% when recovering 90% positive cases | "Branched CALM-Net improves upon the SOTA on 3-class stress level prediction" section. Branched CALM-Net improves the state-of-the-art by introducing personalization and dynamic clustering on stress level prediction with 3 classes | "With 1 week of data, branched CALM-Net achieves 17.24% boost in performance over SOTA" section. Branched CALM-Net is the top performer with F1 score of 0.67 when training on 1 week data from new subject |

**Table 1.** Evaluation schemas. Our evaluation methodologies, along with their descriptions and main takeaways

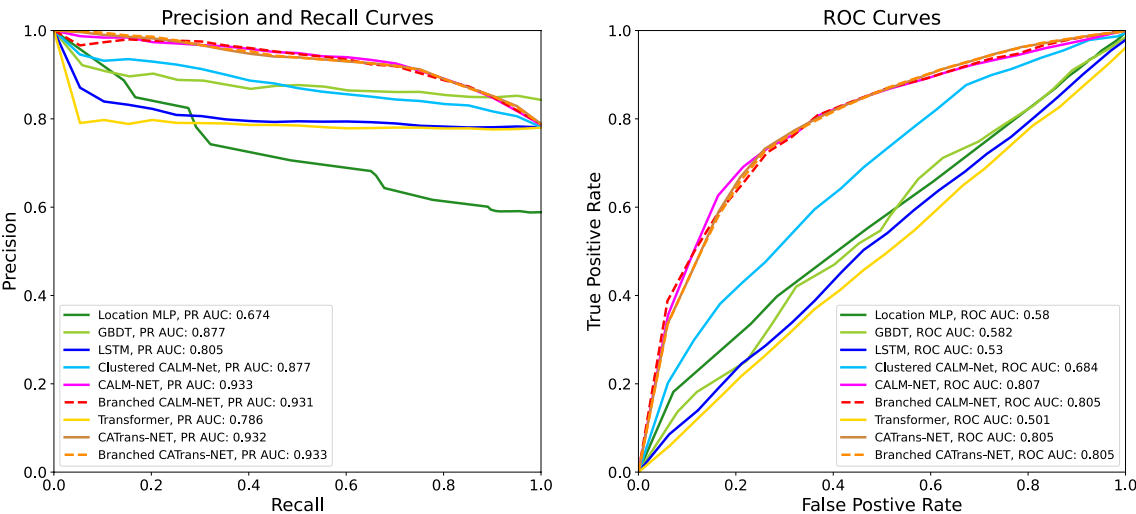
**CALM-Net and Branched CALM-Net attain AUC scores of more than 0.8 on student stress detection**

To evaluate our methods and compare them with baselines in the case that there is sufficient data for each student, we present the results achieved by each method in the warm-start setting.

We present the classification performance of our models versus the baselines in this scenario in Table 2. On the left-most column, we present the precision while fixing the recall at 90%. Table 2 also shows the performance of Branched CALM-Net, which is the version of our method that uses Learn-to-Branch. In Fig. 3 we show the ROC curve along with the AUC score for Location MLP, GBDT, LSTM, Transformer, CATrans-Net, and CALM-Net models. Empirical results, indicate that our proposed CALM-Net and CATrans-Net significantly outperforms Location MLP, GBDT, Transformer and LSTM, which do not have personalized parameters.

| Model                                | Precision (@Recall≈0.9) | PR AUC               | ROC AUC              |
|--------------------------------------|-------------------------|----------------------|----------------------|
| Location MLP <sup>11</sup>           | 0.588 ± 0.002           | 0.674 ± 0.012        | 0.580 ± 0.011        |
| GBDT <sup>43</sup>                   | <b>0.849 ± 0.026</b>    | 0.877 ± 0.031        | 0.582 ± 0.064        |
| LSTM <sup>48</sup>                   | 0.780 ± 0.004           | 0.805 ± 0.010        | 0.530 ± 0.020        |
| Clustered CALM-Net                   | 0.817 ± 0.003           | 0.877 ± 0.002        | 0.684 ± 0.003        |
| CALM-Net                             | <b>0.843 ± 0.003</b>    | <b>0.933 ± 0.001</b> | <b>0.807 ± 0.002</b> |
| Branched CALM-Net                    | <b>0.845 ± 0.005</b>    | <b>0.931 ± 0.001</b> | <b>0.805 ± 0.004</b> |
| Transformer (Trans) <sup>46,50</sup> | 0.780 ± 0.004           | 0.786 ± 0.008        | 0.501 ± 0.015        |
| CATrans-Net                          | <b>0.851 ± 0.006</b>    | <b>0.932 ± 0.002</b> | <b>0.805 ± 0.003</b> |
| Branched CATrans-Net                 | <b>0.851 ± 0.004</b>    | <b>0.933 ± 0.001</b> | <b>0.805 ± 0.002</b> |

**Table 2.** Performance evaluation in binary stress detection. The labels of *median stress* and *very stressed* are combined. We report the precision of different models for a recall of 0.9 and the Area Under the Curve (AUC) for the precision-recall curve and for the receiver operating characteristic (ROC). The PR and ROC curves are shown in Fig. 3. [bold] values indicate top performers in each category.



**Figure 3.** Precision and Recall curves (Left) and ROC curves (Right) of Location MLP, GBDT, LSTM, CALM-Net, and models with Transformer<sup>19</sup> as backbone (namely CATrans-NET) on the task of binary stress detection.

Branched CALM-Net improves upon the SOTA on 3-class stress level prediction

We additionally provide evaluation results for distinguishing between three stress levels (below median, median, and above median) in Table 3. In this context, we also compare our results to the baseline approach, referred to as Clustered CALM-Net, which explicitly forms student groups. We include this baseline to investigate whether introducing Learning-to-Branch<sup>34</sup> to dynamically discover implicit groups during the training process can yield performance improvements. According to Table 3, Branched CALM-Net achieves the best performance, which demonstrates the benefit of dynamic clustering.

To further validate our concept and align our evaluation with real-world applications, we conducted an additional round of cross-validation based on chronological data segmentation. This approach ensures that the data is split in a manner that mirrors actual temporal sequences, thereby enhancing the practical relevance of the analysis. The outcomes of this chronological cross-validation are detailed in Tables 4 and 5. Notably, the results corroborate our previous findings in Table 3, demonstrating consistent performance even under more realistic data segmentation scenarios. This consistency reinforces the robustness and applicability of our Branched CALM-Net approach in real-world settings.

With 1 week of data, Branched CALM-Net achieves 17.24% boost in performance over SOTA

To evaluate the performance of the models in the “cold-start” setting, we use leave-one-subject-out cross-validation. In this setting, the data set is split into subsets, referred to as folds, with numbers equal to the number of subjects, and each fold contains the data samples corresponding to one subject or student in this data set. For each fold, we train the model on the data from other students and then validate on the data that belongs to the left-out student. Furthermore, to assess the functionality of our models using continuously collected data from left-out students, we incrementally add the data of each student, 1 week at a time, and test the performance of the

| Model                      | F1-score             | ROC AUC              |
|----------------------------|----------------------|----------------------|
| Location MLP <sup>11</sup> | 0.388 ± 0.007        | 0.631 ± 0.003        |
| GBDT <sup>43</sup>         | 0.488 ± 0.033        | 0.603 ± 0.028        |
| LSTM                       | 0.479 ± 0.004        | 0.630 ± 0.002        |
| CALM-Net                   | <b>0.602 ± 0.004</b> | <b>0.778 ± 0.002</b> |
| Clustered CALM-Net         | 0.530 ± 0.003        | 0.690 ± 0.003        |
| Branched CALM-Net          | <b>0.605 ± 0.003</b> | <b>0.782 ± 0.002</b> |
| Transformer (Trans)        | 0.450 ± 0.006        | 0.616 ± 0.002        |
| CATrans-Net                | <b>0.593 ± 0.004</b> | <b>0.772 ± 0.002</b> |
| Branched CATrans-Net       | <b>0.591 ± 0.003</b> | <b>0.773 ± 0.001</b> |

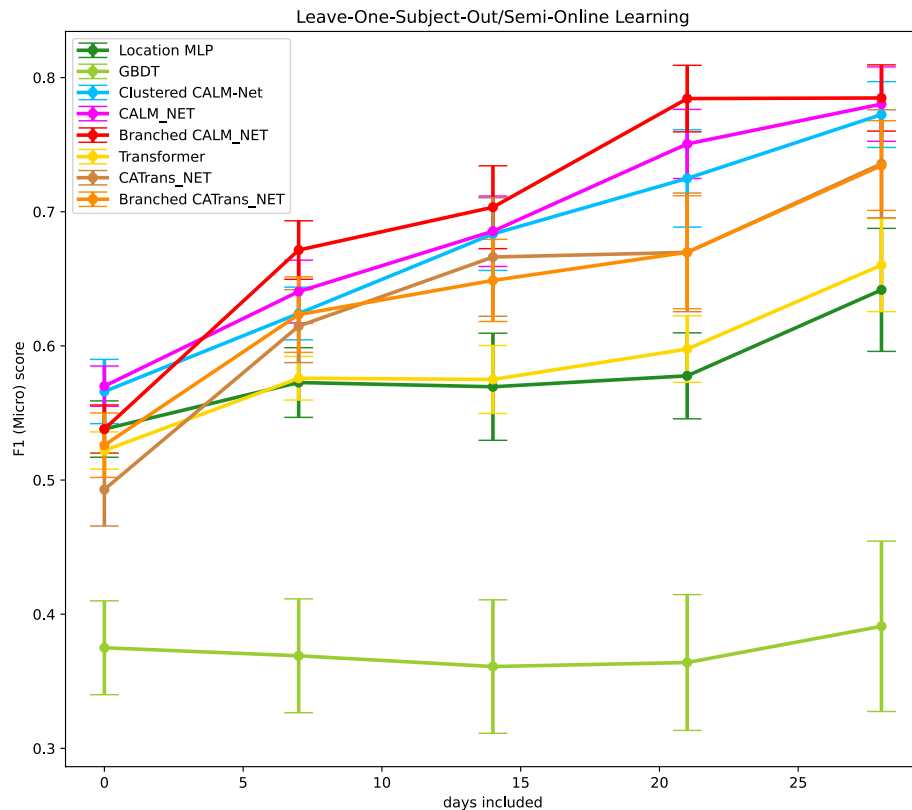
**Table 3.** The F1 scores for 3-class stress level prediction on the StudentLife dataset using five-fold cross-validation under the warm-start scenario. Performances on the task of stress level prediction with 3 levels (*below median, median, above median*). The top section contains the baseline models with the personalized models. The bottom section is models that have grouping methods introduced. Both F1 and AUC scores are calculated globally using micro-averaging by considering each element of the label indicator matrix as a positive label. [bold] values indicate top performers in each category.

| Model             | Precision (@Recall≈0.9) | PR AUC               | ROC AUC              |
|-------------------|-------------------------|----------------------|----------------------|
| LSTM              | 0.798 ± 0.004           | 0.832 ± 0.009        | 0.579 ± 0.013        |
| CALM-Net          | <b>0.835 ± 0.007</b>    | <b>0.923 ± 0.002</b> | 0.779 ± 0.005        |
| Branched CALM-Net | <b>0.836 ± 0.006</b>    | <b>0.925 ± 0.002</b> | <b>0.782 ± 0.004</b> |

**Table 4.** The performance on binary class stress level prediction on the StudentLife dataset using chronological five-fold cross-validation. [bold] values indicate top performers in each category.

| Model             | F1-score             | ROC AUC              |
|-------------------|----------------------|----------------------|
| LSTM              | 0.481 ± 0.006        | 0.796 ± 0.006        |
| CALM-Net          | <b>0.581 ± 0.006</b> | <b>0.858 ± 0.004</b> |
| Branched CALM-Net | <b>0.586 ± 0.005</b> | <b>0.863 ± 0.003</b> |

**Table 5.** The performance on 3-class stress level prediction on the StudentLife dataset using chronological fivefold cross-validation. [bold] values indicate top performers in each category.



**Figure 4.** Online learning in the cold-start setting: Performance plotted against the amount of data included in the training set from the left-out student. Error bars indicate the variance of the scores across students. For both CALM-Net and Branched CALM-Net, predictions with zero data included are generated by the generic layer.

models on the remaining data. This process is repeated for each student, and the mean and standard deviation of the results are reported across all students.

Figure 4 shows the F1 scores attained by different models. The first point (left most) on the plot represents the situation where there are no training data available for the left-out student. CALM-Net still outperforms Location MLP which is our baseline model. In this case, since there are no training data to train the personalized layers contained in CALM-Net and Branched CALM-Net, we will use the generic layer, which is trained with data from other students to make the predictions. Since branched CALM-Net has more trainable parameters, it would need more data points to achieve good performance, and it can be seen that with zero included data, CALM-Net and clustered CALM-Net outperform this model. When only 1 week of data is included from a student—meaning that the student has been using the application for at least 1 week—Branched CALM-Net outperforms other models.

We then inspect the performance of CALM-Net models on predicting binary stress labels (not stressed vs. stressed). The Branched CALM-Net achieves F1 scores of  $0.883 \pm 0.023$  and  $0.912 \pm 0.013$  with one and 2 weeks of data included, respectively, from left-out students. The findings we present not only demonstrate the effectiveness of our method when applied in a leave-one-subject-out setting but also align with our observations from the five-fold cross-validation experiments. Specifically, our results confirm that training the model with 3 levels of stress is adequately capable of handling the task of binary stress detection. This highlights the robustness and consistency of our approach.

### Results on WESAD dataset

To test the generalizability of our methodology, we apply the multi-heads personalization and dynamic branching methods to the Wearable Sensor and Affect Detection (WESAD)<sup>51</sup> dataset.

The WESAD dataset is a freely accessible collection of multimodal data used for identifying stress and emotional states. This dataset includes physiological readings from 15 participants who were involved in three different activities: a baseline task involving 20 minutes of neutral reading, an amusement phase where they watched humorous videos for 392 seconds, and a stress-inducing segment using the Trier Social Stress Test for 10 minutes. During these activities, a variety of physiological parameters were recorded with devices worn on the chest and wrist. These parameters include blood volume pulse (BVP), electrocardiogram (ECG), electrodermal activity (EDA), electromyogram (EMG), respiration (RESP), body temperature, and accelerometer readings. The chest device recorded data at 700 Hz, while the wrist device did so at 64 Hz, 32 Hz, and 4 Hz. For data processing, we adopted the approach outlined in the study by Dzieżyc, Maciej, et al.<sup>52</sup>, but we downsampled all signals to a uniform frequency of 4 Hz to ensure consistency. For each participant, several personal attributes

| Model                     | F1-score             | ROC AUC              |
|---------------------------|----------------------|----------------------|
| LSTM <sup>46</sup>        | 0.781 ± 0.023        | 0.883 ± 0.010        |
| CALM-Net                  | 0.794 ± 0.017        | <b>0.938 ± 0.016</b> |
| Branched CALM-Net         | <b>0.805 ± 0.014</b> | 0.929 ± 0.013        |
| Transformer <sup>46</sup> | 0.953 ± 0.008        | 0.995 ± 0.005        |
| CATran-Net                | <b>0.984 ± 0.012</b> | <b>0.999 ± 0.001</b> |
| Branched CATran-Net       | 0.967 ± 0.030        | 0.988 ± 0.013        |

**Table 6.** The performance for 3 classes (baseline, stress, amusement) prediction on the WESAD dataset under three-fold cross validation. Empirical results from three-fold cross-validation, where the splits are stratified by subjects. [bold] values indicate top performers in each category.

| Model                     | F1-score 20%         | F1-score 40%         | F1-score 60%         |
|---------------------------|----------------------|----------------------|----------------------|
| LSTM                      | 0.758 ± 0.152        | 0.797 ± 0.140        | 0.784 ± 0.116        |
| CALM-Net                  | <b>0.772 ± 0.113</b> | <b>0.832 ± 0.087</b> | 0.799 ± 0.116        |
| Branched CALM-Net         | 0.753 ± 0.137        | 0.726 ± 0.159        | <b>0.803 ± 0.109</b> |
| Transformer <sup>46</sup> | 0.761 ± 0.145        | 0.791 ± 0.176        | 0.944 ± 0.091        |
| CATran-Net                | <b>0.900 ± 0.100</b> | 0.890 ± 0.118        | 0.938 ± 0.089        |
| Branched CATran-Net       | 0.801 ± 0.166        | <b>0.935 ± 0.093</b> | <b>0.960 ± 0.061</b> |

**Table 7.** The performance for 3 classes (baseline, stress, amusement) prediction on the WESAD dataset under leave-one-subject-out validation. The validation is conducted under leave-one-subject-out, with data from the left out subject being added to training set gradually to simulate the real-world online-learning setting. The percentage represent the amount of total available data included. [bold] values indicate top performers in each category.

are used as covariates in the network to identify personalized differences. These attributes include age, height, weight, gender, handedness, whether they consumed coffee on the day of the experiment, their engagement in sports, smoking status, and whether they felt ill during the experiment.

Ultimately, the time series data is divided into one-minute segments, each labeled according to the specific task in which the participant was engaged during that minute. The objective is to predict the task being undertaken by the subject, using both their personal attributes and the data collected from the wearable sensors during that particular minute.

Table 6 shows the preliminary results under three-fold cross validation. Table 7 shows the leave-subject out validation with online learning setting. The empirical observation of models’ performance is consistent with the observation on the StudentLife<sup>40</sup> dataset.

Discussion

Early detection and intervention of stress are crucial for preventing the onset of various health problems. Smartphones, which are widely used, provide an opportunity for continuous monitoring of stress levels. However, the detection of stress from smartphone data poses several challenges. The data collected from smartphones are often irregular, noisy, and heterogeneous. *In our work, we show that by converting smartphone-derived physiological time-series data into a sequence of histograms and leveraging an auto-encoder, it is possible to substantially decrease the noise present in this data. This approach has demonstrated improved classification performance in forecasting student stress levels.*

This paper also proposes a novel multitask learning model with dynamic clustering called Branched CALM-Net, which aims to provide accurate stress prediction using data collected from smartphones. The proposed approach leverages the personalized structure to address intersubject variability. *By defining each student as a different task in a multitask neural network, the empirical results demonstrate notable enhancements in the stress prediction classification performance for the StudentLife dataset. This approach effectively captures the commonalities between students, while also adeptly modeling individual variations.*

The experimental results presented in “Results” section demonstrate that CALM-Net and Branched CALM-Net outperform the state-of-the-art methods<sup>11,43</sup> in both binary diagnosis and stress level prediction tasks. This improvement can be attributed to the addition of personalized parameters that account for inter-subject variability. In addition, we observe that training the model with more fine-grained labels leads to better results. If we train CALM-Net directly on the binary stress labels, the precision in recall of 90% is around 0.782 ± 0.003 with an AUC of 0.532 ± 0.029, worse than the performance shown in Table 2 and Fig. 3 where the models are trained with 3 distinct stress labels. This empirical evidence justifies our choice that *it is worth training with more fine-grained labels*, with a relatively balanced number of samples, and to merge the output probability when we need a coarse level of prediction.

The experiments conducted in the cold-start online learning setting, the results of which are summarized in Fig. 4, indicate that CALM-Net and Branched CALM-Net can quickly adapt to new subjects, *with Branched CALM-Net outperforming the other methods as long as the subject has at least 1 week of training data.*

Notably, Branched CALM-Net employs a learning-to-branch method that dynamically clusters students during the training process. This approach captures more valuable patterns from the data compared to using preclustered groups, as evidenced by the empirical results presented in “CALM-Net and Branched CALM-Net attain AUC scores of more than 0.8 on student stress detection” and “Branched CALM-Net improves upon the SOTA on 3-class stress level prediction” sections. Overall, Branched CALM-Net is able to cluster new students into a closer group, providing accurate stress level predictions after the first week of data collection as shown in Fig. 4. These findings demonstrate the potential of personalized multitask learning models with dynamic clustering for stress prediction using smartphone data.

*Following the solid empirical evidence of the performance of CALM-Net and Branched CALM-Net under both the warm-start and cold-start settings, we summarize our contributions as follows:*

- (i) We proposed a novel platform for student stress prediction that models both population and personal characteristics with multitask learning and dynamic clustering. This model addresses the intersubject variability and cold-start problems.
- (ii) We developed a data preprocessing pipeline that effectively reduces data noise and handles data irregularity to improve the accuracy of stress prediction.
- (iii) We addressed an important but often ignored challenge in stress prediction by evaluating our models in the cold start setting. We added specific components to our model to handle this scenario, which enabled us to effectively predict stress levels even when we have little or no prior data for a given individual.

To sum up, our study demonstrates the ability of our system to effectively address the complex task of predicting student stress levels using mobile phone data. Our results suggest that it is possible to achieve automatic mental health monitoring through the use of deep neural networks. Our model's accuracy is improved through the introduction of personalized multitask learning and dynamic clustering architecture, enabling it to provide accurate stress level predictions in various real-world scenarios with promising performance.

## Generalization and limitations

### *Scalability to large-scale populations and data privacy*

While our model is designed with a unique neural network head for each student, facilitating personalization, this architectural choice inevitably leads to an increase in the model's size in proportion to the number of participants. This presents a significant scalability challenge, especially since we have not yet had the opportunity to evaluate the model in a large-scale setting, due to the absence of appropriate datasets.

Notwithstanding this, the architecture of our model is inherently conducive to a federated learning approach. In such a large-scale application scenario, each student's device would only maintain and update their individual neural network head. This modular design permits local updates and periodic synchronization with the overarching model framework, effectively mirroring a federated learning system<sup>53–55</sup>. Adopting this approach could provide a viable solution to the scalability issue by distributing the data processing and model training across devices. It is, however, crucial to note that both the implementation of this federated learning system and its validation in a large-scale context remain beyond the scope of our current research.

This proposed federated learning methodology also addresses key concerns regarding data privacy and security. The present model requires the sharing of student data with a server for periodic updates, which poses potential privacy risks. Yet, the modular structure of our model is ideally suited for a federated learning approach, wherein students would conduct updates on their individual network components locally. Consequently, only aggregated updates pertaining to the shared model components would need to be communicated to the server, significantly bolstering data privacy. Although the comprehensive implementation of this federated system is not covered in our existing research, it represents a promising avenue for future development, potentially ensuring scalability alongside enhanced privacy and security for practical applications.

### *Limitations and zero-shot performance*

A notable limitation of our approach is its dependency on initial data for effective personalization. The model's zero-shot performance, without prior data, is limited. This necessitates the collection of at least 1 week of labeled data from each student to achieve accurate stress predictions. To address this challenge, further integrating zero-shot learning techniques<sup>56–58</sup> could potentially enhance the model's initial performance for newcomers, thereby reducing the initial data dependency. Such advancements would be pivotal in making the model more robust and immediately useful for new participants.

### *Requirement of EMA stress labels and psychological surveys*

While our models can perform effectively without psychological surveys, the EMA (Ecological Momentary Assessment) stress labels are crucial for training. These labels provide the necessary data for the models to learn and adapt to each student's stress patterns. In cases where students cease to respond to EMAs, the model's predictive accuracy may gradually decline due to the temporal shift in the student's data. Addressing the challenge of declining student engagement with EMAs transcends the technical scope of our model and ventures into the realm of intervention strategies. It necessitates proactive outreach by school health staff to identify the causes behind the data gaps and evaluate the necessity for professional intervention. Furthermore, our framework holds the potential to be integrated with existing intervention tools, enhancing its practical utility. While the exploration of this integration is beyond the scope of the current paper, it represents a significant area for future research. Such studies could explore how our model can be symbiotically paired with intervention strategies to offer a more comprehensive solution for monitoring and addressing student stress.

**Adaptability to different demographics** Our current research primarily focuses on college students, a group with distinct stressors and challenges. However, the adaptability of our model to various demographics remains an area ripe for exploration. Different groups, such as high school students, students from diverse collegiate environments, or working professionals, encounter unique stressors that may not be fully encapsulated by our current model, which is tailored to the college student demographic. The performance and effectiveness of our model in these varied groups warrant thorough investigation. For instance, evaluating its applicability across different colleges, each with its unique demographic structure, is essential to understand the model's versatility and effectiveness in diverse educational settings. Such evaluations are crucial, especially if the model is to be implemented practically in various institutions. Unfortunately, our current research was limited by the availability of datasets. We did not have access to extensive datasets encompassing these varied demographics, which restricted our ability to test and adapt the model across a broader demographic spectrum. Future research should aim to bridge this gap, focusing on gathering and analyzing data from these diverse groups. This approach will not only validate the model's adaptability but also enhance its applicability and effectiveness in addressing the stress prediction needs of a wider population range.

## Methods

Since the survey data is taken from a publicly available dataset from a study conducted at Dartmouth College, no informed consent was required.

### Problem setup

We formulate the problem of predicting student stress levels as a supervised learning problem where the input to the model is the time series of the data collected from each student's phone during the day and the label is the student stress level indicated by the student for that day. We denote the data collected from the student  $i$  from time step 1 to time step  $T$  as  $X_{1:T}^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \dots, x_T^{(i)}\}$  and the stress label at time  $T$  as  $y_T$ . We want to predict  $y_T$  given  $X_{1:T}^{(i)}$ .

We will illustrate our solution to this problem on the StudentLife dataset. In this dataset, students select a stress level from 1 to 5 each day. However, due to the uneven distribution of stress level labels, we have followed the approach proposed by<sup>11</sup> and transformed the labels into a scale of 1 to 3. The new scale represents “below median stress”, “median stress”, and “above median stress”. Thus, the labels are now categorized as  $y \in \{1, 2, 3\}$ . The original stress levels of 1 and 2 have been mapped to “below median stress”, level 3 is “median stress”, and levels 4 and 5 have been mapped to “above median stress”. These three labels will be used for the supervised classification task.

### Baselines

We compared our models against the work by Gatis et al.<sup>11</sup> titled MLP based on location features, which they called Location MLP. To extract features, Gatis et al.<sup>11</sup> used GPS data and aggregated it on a daily basis. They extracted a total of eight location-based features, including the total distance covered, maximum displacement, distance entropy in 10-minute intervals, distance standard deviation, number of unique tiles visited, grid sections on the satellite map visited, the difference in tiles visited from the previous day, the approximate area of the GPS convex hull, and number of clusters in the GPS data. Additionally, they extracted four covariates based on the date, such as binary indicators of whether the day was the start of the term, the middle of the term, the end of the term, or a weekend.

We further evaluate our approach by contrasting it with the methodology outlined in<sup>43</sup>, which used a GBDT model for their predictions. In their study, they employed EMA data from the StudentLife dataset to extract a comprehensive set of 44 features. To mitigate class imbalance, they applied the SMOTE method. Their approach then determined the closest neighbors to a given sample using the Euclidean Distance metric, allowing for personalized modeling by training the model exclusively on the data of these nearest neighbors for each subject.

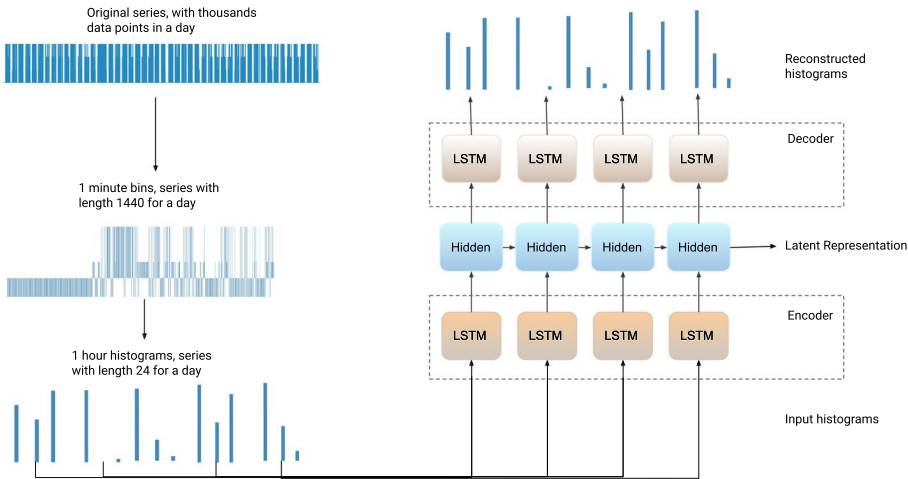
### Data pre-processing

As passive sensors continuously collect data, there are thousands of measurements collected throughout the day for each student. Since these collections are noisy and it is difficult to model such a long time series data, we use feature engineering to reduce the noise and the length of this data. The signals are first binned with 1-minute blocks. Each feature is then binned with a different aggregation rule, as demonstrated in Table 8. Most of the features are aggregated by summation, representing the sum of the values of the feature over 1-minute bins, e.g., the total amount of conversation, phone locking time, sleep duration, etc. For *Activity* and *Audio* we use the mode to indicate whether the individual is performing an activity or if events are happening during that minute. So, by aggregating the data in this way, we remove small changes in short amounts of time, which is likely due to noise. The aggregated value of each bin forms the new series, with 1440 sequences per day. This series represents what the subject was doing at that minute. This sequence is still shown to be too long for recurrent neural networks<sup>59</sup> so we further process the data by computing the histograms of 1-hour blocks. This step results in a sequence of 24 histograms per day. Figure 5 illustrates our feature engineering process. This conversion of the raw irregular time-series data to a series of histograms serves to reduce the noise and obtain regularly-sampled data.

Some of the features in the series exhibit missing values. Our imputation strategies for each feature are specified in Table 8. For most features, we adopt forward-fill imputation, where the missing time step is filled with that of the previous step. For the series with missing values at the beginning, we apply mean imputation after forward filling. For *Activity* and *Audio*, we use ‘0’ for the missing values and assume that nothing happened during those time steps.

| Feature type | Feature name     | Aggregation rule | Feature values | Imputation         |
|--------------|------------------|------------------|----------------|--------------------|
| Discrete     | Activity         | Mode             | [0, 3]         | Fill with zero     |
| Sequences    | Audio            | Mode             | [0, 3]         | Fill with zero     |
|              | Conversation     | Sum              | [0, 1]         | Forward fill       |
|              | Phone Charge     | Sum              | [0, 1]         | Forward fill       |
|              | Phone Lock       | Sum              | [0, 1]         | Forward fill, mean |
|              | Time to deadline | N/A              | [0, ∞+)        | N/A                |
| Covariates   | Day of the week  | N/A              | [0, 6]         | N/A                |
|              | Exam period      | N/A              | [0, 1]         | N/A                |
|              | Sleep rating     | Sum              | [0, ∞+)        | Forward fill, mean |
|              | Sleep duration   | Sum              | [0, ∞+)        | Forward fill, mean |

**Table 8.** Feature preprocessing information. During preprocessing, we first aggregate the time series data into bins of minutes, then compute the histogram within each hour. Here we list the aggregation rule, feature values, and modes in dataset for each feature. For the “mode” rule, we record whether events have happened. For the “sum” rule, we are extracting the number of times the events happened



**Figure 5.** CALM-Net overall structure. We first aggregate each feature into 1-minute bins, then we compute the histogram for each hour. This data is then inputted into an LSTM Autoencoder for denoising and dimension reduction. The last hidden state is used as a latent representation of the entire input series, which is provided to the classification layers.

This feature engineering technique actually models ‘how much conversation or activity a student has undergone in an hour which led to the stress label in a consistent manner’. More specifically, our time series for a student will become a set of the following shape  $X_{1:T}^{(i)} \in \mathbb{R}^{C \times T \times \hat{H}}$  as input to the model that contains the time series of hour histograms from one day.  $C$  is the number of channels such as activity, sound, and binary series;  $T$  is the length of the series, which is 24 in this case, and  $\hat{H}$  is the size of the histogram.

Psychological surveys are also collected from students before the beginning and end of the study. The surveys include PHQ-9 for depression, PSS for stress, Loneliness scale, Flourish scale, PANAS, and the Big Five Personality Traits. Each survey could generate a score based on the student’s responses, representing the extent of the corresponding mental states, such as stress, loneliness, conscientiousness, etc. As a result, we have survey scores denoted as  $S \in \mathbb{R}^{N \times \hat{S}}$  where  $N$  is the number of students and  $\hat{S}$  is the number of surveys collected from students.

Using auto-encoders to reduce data noise

Due to the nature of data collection from sensors on smartphones in the real world, there is a significant amount of noise present in the data. This noise can stem from factors such as the unreliability of the sensors, variations in phone placement, changes in environments, and more. While the data processing step mentioned earlier can remove some of this noise by aggregating the data into predefined bins, a substantial amount of noise will remain that could potentially harm the performance of models using this data. To address this issue,<sup>41</sup> has summarized several practical techniques for modeling time series. We have chosen to use LSTM auto-encoders, as briefly illustrated in Fig. 5, to reduce this noise and obtain higher quality embeddings of the data.

The input to the LSTM encoder is a sequence of histograms as described in “Data pre-processing” section. Prior to being fed to the LSTM unit, the set of histograms is flattened at each time step. The output of the LSTM encoder is considered the latent representation of the input sequence and will be merged with covariates for subsequent classification processing. The decoder is employed to reconstruct the input sequence from the latent representation. The autoencoder is trained simultaneously with the entire model. We use the Mean Absolute Error (MAE) as the Reconstruction Error, denoted as  $RE$ . With the Classification Error, denoted as  $CE$ , we have the integrated loss:

$$L_{integrated} = \alpha \cdot RE + \beta \cdot CE \quad (1)$$

where  $\alpha$  and  $\beta$  are the hyper-parameters ( $0 \leq \alpha, \beta \leq 1$ ) for weighting the two losses. The choice of these coefficients depends on the relative importance of each loss. Given that the classification error is the primary task, we set its coefficient to 1. To prevent reconstruction loss, which serves as a denoising and regularization mechanism, from dominating the gradient during backpropagation, we set its coefficient to  $1e - 4$ . We selected these values through hyperparameter optimization.

Both our LSTM encoder and decoder consist of a single LSTM layer. The last hidden state of the encoder with a size of 128 is used as the latent representation of the input data. The decoder will take the hidden states from the encoder at all time steps as input and predicts the original sequence.

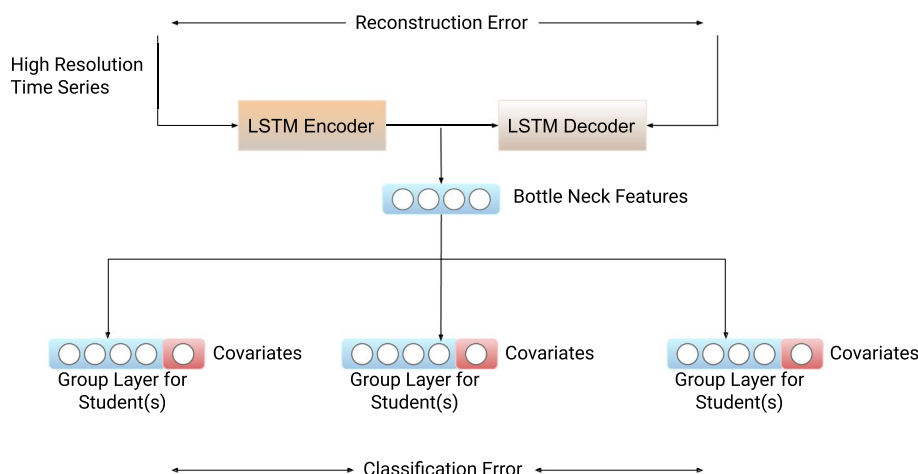
### Personalization

It is crucial to consider individual differences when modeling mental states such as stress, as these experiences can vary greatly between individuals<sup>27</sup>. This attribute should be taken into account when modeling data from mobile phones for the purpose of stress prediction. Busk et al.<sup>17</sup> and Jaques et al.<sup>16</sup> have emphasized the importance of personalized parameters when modeling individual moods. Kandemir et al.<sup>60</sup> approach the personalization in the prediction of affect (mood) by learning user-specific kernels, further strengthening the significance of modeling individual patterns. To learn personalized models for each student, we adopt a multitask approach that consists of an LSTM to model the sequence of histograms, followed by shared fully connected layers and a MLP for each student.

Specifically, our model has a 2-layer MLP for each student on top of a shared fully connected layer and the LSTM Auto-encoder. Each fully connected layer has a ReLU activation function except the last layer, which maps the input to the class probabilities. We named this proposed model the Cross-Personal Activity LSTM Multitask Auto-Encoder Network (CALM-NET). Figure 6 represents a CALM-Net model where each group layer contains exactly one student—thus becoming a personalization layer.

As indicated by our experiments in “Results” section, this approach can learn the differences between students and subsequently yield a significant improvement in performance. Furthermore, according to our ablation studies presented in the Appendix Section C, learning a single model for all the students is unsuited to this task. Multitask learning also acts as a heavy regularizer, preventing the model from overfitting to a single student or the most common label. The shared layers learn common features, while the personalization layers learn features that are relevant to the respective subject.

A drawback of this approach is that we need training data for every student. If a new student joins the system, the model will not be able to make predictions for the new student. This is called the “cold-start problem” in the literature. To solve this problem, we developed and tested different approaches, as outlined in the following subsections.



**Figure 6.** The structure of Multitask Learner, an LSTM autoencoder followed by multi-heads output. All the group layers have their own parameters, and they share the LSTM autoencoder along with the bottle-neck features. It is equivalent to CALM-NET if each group layer contains exactly one student.

### Clustering students

The first approach we used to address the cold start problem was to cluster students into groups that have similar characteristics. We then predict the stress level for a group of students instead of individuals. In this way, for a newly added student, we first assign them to a group and then use the predictive model corresponding to the group to obtain the individual prediction.

We use the student surveys to group them. In this way, students with similar characteristics who answered to surveys similarly will be clustered together. Then, we replace the personalized layers in the model architecture with group layers, where we use an MLP for each group of students.

Specifically, the features used for grouping are aggregated scores from the surveys, which are collected before the beginning of the study: Big Five, Flourishing Scale, Loneliness Scale, Positive and Negative Affect Schedule, Perceived Stress Scale, and PHQ-9. The clustering method that we applied is Density-Based Spatial Clustering of Applications with Noise (DBSCAN)<sup>61</sup>. The hyper-parameters were tuned during experiments. The empirical results show that models with 3 to 4 students in each cluster on average achieve the best possible performance. The model has the structure shown in Fig. 6, where the students are assigned to each group layer based on the survey scores.

### Branched CALM-NET

The clustering approach relies on static clustering using data collected during surveys at the beginning of data collection. However, this approach may be limited by the availability of data, and the inability of the survey data collected at the beginning of the study to capture changes in students' behaviors over time. To overcome this limitation, we developed a model that dynamically clusters students into groups while predicting their stress levels. Our approach is based on the "learning to branch" method originally proposed by<sup>34</sup>. Our model includes a shared layer followed by group nodes, with individual layers for each student connected to all group nodes. By training on the data, the model learns branches in the network that assign each individual to the best group, enabling clustering and classification. A diagram of the model architecture is provided in Fig. 1.

More formally, each personalized layer will receive the output from different group nodes with a likelihood distribution. An illustration of the pipeline is shown in Fig. 2. We will denote the output of the group node  $i$  as  $B_i$ , the likelihood of the personalized layer  $j$  receiving the output from the branch  $i$  as  $\theta_{ij}$ . Each personalized layer will take the output from the branch with the largest likelihood. Then the output, denoted as  $d_j$ , which will be received by the personalized layer  $j$  is

$$d_j = B^T \text{one\_hot}\{\arg\max_i(\theta_{ij})\}$$

However, as  $\frac{\partial d_j}{\partial \theta}$  is not directly differentiable, it will not work with the backpropagation algorithm in training the neural network. In order to force the probability close to the one-hot vector during training and still make the branching operation differentiable, we replace the one-hot operation with *Softmax*. We used the *Gumbel trick*, as stated in<sup>62</sup>, which is adopted as a solution provided in<sup>34</sup>. By applying the Gumbel softmax trick, a smooth version of the branching operation is given as:

$$d_j = B^T \frac{\exp((\log \theta_j + \epsilon)/\tau)}{\sum_k \exp((\log \theta_{k,j} + \epsilon_k)/\tau)}$$

The noise term  $\epsilon$  is generated from the Gumbel distribution, as described by<sup>62</sup>, to prevent the network from being overly sensitive to parameter initialization. This is crucial because if the network were to favor branches with higher likelihood at initialization, it could result in branches with lower likelihood suffering from gradient saturation during training. The temperature parameter  $\tau$  controls the smoothness of softmax operation, with smaller values making the output closer to a one-hot vector. During training, we initialize  $\tau$  with a relatively large value, then linearly decay it across epochs.

Therefore, at the end of the training, subjects that choose the same branch will be considered as being in the same group.

### CATrans-Net and branched CATrans-Net

Similar to CALM-Net and Branched CALM-Net we introduce CATrans-Net and Branched CATrans-Net which have similar structures to CALM-Net and Branched CALM-Net but replace the LSTM autoencoder with the Transformer model introduced by Foumani et al.<sup>46</sup> in the model architecture.

### Adding the generic layer

Since the CALM-Net and Learn-to-Branch models have personalized parameters, they cannot make predictions for a new incoming student without any prior training data available. To address this, we added a generic layer to both models, as shown in Fig. 1, which is trained using data from all students in the training set. By incorporating a multitask approach, the integrated loss function becomes:

$$L_{\text{integrated}} = \alpha \cdot RE + \beta \cdot CE_{\text{personal\_layer}} + \lambda \cdot CE_{\text{generic\_layer}}$$

The integrated loss function includes three components:  $RE$ , which is the reconstruction loss from the auto-encoder part of the model;  $CE_{\text{personal\_layer}}$ , which represents the classification loss from individual student layers; and  $CE_{\text{generic\_layer}}$ , which denotes the classification loss from the generic layer. To balance the optimization rate and account for the fact that the parameters of the generic layer are optimized in every iteration, we set the weight

assigned to the classification loss with output from the generic layer ( $\lambda$ ) to one over the number of students (excluding the new incoming student), which is 22 in this particular dataset.

## Conclusion

In this study, we introduced CALM-Net and its branched version, Branched CALM-Net, for stress level prediction in the StudentLife dataset. We used feature engineering and histogram categorical inference to address noise and align different signals in the data. We also employed LSTM-Autoencoder to further reduce data noise and overfitting, as well as to improve the quality of the embeddings. We implemented a Multi-Task network structure with personalized parameters for each subject to address inter-subject variability. We also enhanced our model through the Learn-to-Branch to identify similar groups of students and thus improve prediction performance.

Both CALM-Net and Branched CALM-Net achieved peak performance with an *F1-score* of 0.602 and 0.605, respectively, in the five-fold cross-validation setting in the 3-level stress classification task. Branched CALM-Net outperformed all other models with just 1 week of training data available under leave-one-subject-out. In the binary diagnosis task, Branched CALM-Net maintained a *precision* of over 84% and 82%, while *recovering* 90% and 95% of positive cases, respectively. These results suggest that the characteristics at both group-level and individual-level are significant in modeling the pattern of passive sensor data from various subjects.

In the future, we plan to explore our model's capabilities for forecasting tasks, where instead of only predicting the current stress level, the model is trained to forecast future stress levels. Additionally, we aim to use this model for the prediction of other mental well-being indicators such as mood. Ultimately, our platform is meant to assist as a guide for early intervention.

## Data availability

The datasets generated during and/or analysed during the current study are available at: <https://studentlife.cs.dartmouth.edu/dataset.html>.

## Code availability

The source code used for conducting experiments and generating results during the current study are available in a public repository on GitHub: <https://github.com/Information-Fusion-Lab-Umass/personalized-stress-prediction>.

Received: 11 October 2023; Accepted: 8 March 2024

Published online: 19 March 2024

## References

- Levitt, J. M., Saka, N., Hunter Romanelli, L. & Hoagwood, K. Early identification of mental health problems in schools: The status of instrumentation. *J. Sch. Psychol.* **45**(2), 163–191. <https://doi.org/10.1016/j.jsp.2006.11.005> (2007) (**Universal Screening for Enhanced Educational and Mental Health Outcomes**).
- Muskin, P. R. What are anxiety disorders? American Psychiatric Association. <https://www.psychiatry.org/patients-families/anxiety-disorders/what-are-anxiety-disorders> (2021).
- Rozanski, A., Blumenthal, J. & Kaplan, J. Impact of psychological factors on the pathogenesis of cardiovascular disease and implications for therapy. *Immunol. Today* **99**, 2192–2217 (1999).
- Kario, K., McEwen, B. & Pickering, T. Disasters and the heart: A review of the effects of earthquake-induced stress on cardiovascular disease. *Hypertens. Res.* **26**, 355–367 (2003).
- Lupien, S. J., McEwen, B. S., Gunnar, M. R. & Heim, C. Effects of stress throughout the lifespan on the brain, behaviour and cognition. *Nat. Rev. Neurosci.* **10**, 434–445 (2009).
- Khansari, D., Murgo, A. & Faith, R. Effects of stress on the immune system. *Immunol. Today* **11**, 170–175 (1990).
- Kivimäki, M., Bartolomucci, A. & Kawachi, I. The multiple roles of life stress in metabolic disorders. *Nat. Rev. Endocrinol.* <https://doi.org/10.1038/s41574-022-00746-8> (2022).
- Hellhammer, D. H., Wüst, S. & Kudielka, B. M. Salivary cortisol as a biomarker in stress research. *Psychoneuroendocrinology* **34**(2), 163–171. <https://doi.org/10.1016/j.psycheneu.2008.10.026> (2009).
- Pejovic, V., Mehrotra, A., & Musolesi, M. In *Anticipatory Mobile Digital Health: Towards Personalized Proactive Therapies and Prevention Strategies* (ed. Nadin, M.) 253–267 (Springer, 2017). [https://doi.org/10.1007/978-3-319-45142-8\\_15](https://doi.org/10.1007/978-3-319-45142-8_15).
- Robert, T. & Phillip, W. Predicting future mental illness from social media: A big-data approach. *Behav. Res. Methods* <https://doi.org/10.3758/s13428-019-01235-z> (2019).
- Mikelsons, G., Smith, M., Mehrotra, A., & Musolesi, M. Towards deep learning models for psychological state prediction using smartphone data: Challenges and opportunities. In *ML4H Workshop at 31st Conference on Neural Information Processing Systems (NIPS)* (2017). [arXiv:1711.06350](https://arxiv.org/abs/1711.06350)
- Thieme, A., Belgrave, D. & Doherty, G. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Trans. Comput. Hum. Interact.* <https://doi.org/10.1145/3398069> (2020).
- Shatte, A. B. R., Hutchinson, D. M. & Teague, S. J. Machine learning in mental health: A scoping review of methods and applications. *Psychol. Med.* **49**(9), 1426–1448. <https://doi.org/10.1017/S0033291719000151> (2019).
- Srividya, M., Mohanavalli, S. & Bhalaji, N. Behavioral modeling for mental health using machine learning algorithms. *J. Med. Syst.* **42**(5), 88. <https://doi.org/10.1007/s10916-018-0934-5> (2018).
- Gerych, W., Agu, E., & Rundensteiner, E. Classifying depression in imbalanced datasets using an autoencoder-based anomaly detection approach. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)* 124–127 (2019). <https://doi.org/10.1109/ICSC.2019.8665535>
- Jaques, N., Rudovic, O.O., Taylor, S., Sano, A., & Picard, R. Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation. In *Proceedings of IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing. Proceedings of Machine Learning Research* (eds Lawrence, N., Reid, M.) Vol. 66, 17–33 (PMLR, 2017). <http://proceedings.mlr.press/v66/jaques17a.html>.
- Busk, J. et al. Forecasting mood in bipolar disorder from smartphone self-assessments: Hierarchical Bayesian approach. *JMIR Mhealth Uhealth* **8**(4), 15028. <https://doi.org/10.2196/15028> (2020).
- Hochreiter, S. & Schmidhuber, J. Long-short-term memory. *Neural Comput.* <https://doi.org/10.1162/neco.1997.9.8.1735> (1997).

19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. Attention is all you need. *CoRR* (2017) [arXiv:1706.03762](https://arxiv.org/abs/1706.03762)
20. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L. & Muller, P.-A. Deep learning for time series classification: A review. *Data Min. Knowl. Discov.* **33**(4), 917–963 (2019).
21. Shukla, S. N., & Marlin, B. M. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318* (2021)
22. Shukla, S. N., & Marlin, B. M. A survey on principles models and methods for learning from irregularly sampled time series: From discretization to attention and invariance. *arXiv preprint* (2020)
23. Kumar, P., Chandra, R., Bansal, C., Kalyanaraman, S., Ganu, T., & Grant, M. Micro-climate prediction-multi scale encoder-decoder based deep learning framework. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* 3128–3138 (2021)
24. Zhao, B., Lu, H., Chen, S., Liu, J. & Wu, D. Convolutional neural networks for time series classification. *J. Syst. Eng. Electron.* **28**(1), 162–169. <https://doi.org/10.21629/JSEE.2017.01.18> (2017).
25. Zhang, C. *et al.* A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. *Proc. AAAI Conf. Artif. Intell.* **33**(01), 1409–1416. <https://doi.org/10.1609/aaai.v33i01.33011409> (2019).
26. Sen, R., Yu, H.-F. & Dhillon, I. S. Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. In *Advances in Neural Information Processing Systems* Vol. 32 (eds Wallach, H. *et al.*) (Curran Associates, Inc., 2019).
27. Lacey, O. M., Jeronimus, B. F., van Aken, M. A. G. & Ormel, J. Why not everyone gets their fair share of stress: Adolescent's perceived relationship affection mediates associations between temperament and subsequent stressful social events. *Eur. J. Personal.* **29**(2), 125–137. <https://doi.org/10.1002/per.1989> (2015).
28. Wang, X., Zhang, H., Cao, L., & Feng, L. Leverage social media for personalized stress detection. In *Proceedings of the 28th ACM International Conference on Multimedia*. MM '20 2710–2718 (Association for Computing Machinery, 2020). <https://doi.org/10.1145/3394171.3413596>.
29. Kulkarni, V., Kulkarni, M., & Pant, A. Survey of personalization techniques for federated learning. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)* 794–797 (2020). IEEE.
30. Palmius, N. *et al.* Group-personalized regression models for predicting mental health scores from objective mobile phone data streams: Observational study. *J. Med. Internet Res.* **20**(10), 10194. <https://doi.org/10.2196/10194> (2018).
31. Chen, J. *et al.* Development, implementation, and evaluation of a personalized machine learning algorithm for clinical decision support: case study with shingles vaccination. *J. Med. Internet Res.* **22**(4), 16848 (2020).
32. Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., & Ramage, D. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604* (2018).
33. Yuan, J., Shalaby, W., Korayem, M., Lin, D., Aljadda, K., & Luo, J. Solving cold-start problem in large-scale recommendation engines: A deep learning approach. In *2016 IEEE International Conference on Big Data (Big Data)* 1901–1910 (2016). <https://doi.org/10.1109/BigData.2016.7840810>
34. Guo, P., Lee, C.-Y., & Ulbricht, D. *Learning to Branch for Multi-task Learning* (2020).
35. Vandenheide, S., Brabandere, B.D., & Gool, L.V. Branched multi-task networks: Deciding what layers to share. *CoRR arXiv:1904.02920* (2019).
36. Hu, F., Smeaton, A. F., & Newman, E. Periodicity detection in lifelog data with missing and irregularly sampled data. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 16–23 (2014). <https://doi.org/10.1109/BIBM.2014.6999284>
37. Zhong, Y., & Deng, Y. Sensor orientation invariant mobile gait biometrics. In *IEEE International Joint Conference on Biometrics* 1–8 (2014). <https://doi.org/10.1109/BTAS.2014.6996246>.
38. Kostikis, N., Hristu-Varsakelis, D., Arnaoutoglou, M., Kotsavasiloglou, C., & Baloyiannis, S. Towards remote evaluation of movement disorders via smartphones. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 5240–5243 (2011). <https://doi.org/10.1109/IEMBS.2011.6091296>.
39. Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.* **8**(1), 1–12 (2018).
40. Wang, R., Wang, W., Dasilva, A., Huckins, J. F., Kelley, W. M., Heatherton, T. F., & Chambell, A. T. Tracking depression dynamics in college students using mobile phone and wearable sensing. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 2 (2018).
41. Långkvist, M., Karlsson, L. & Loutfi, A. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognit. Lett.* <https://doi.org/10.1016/j.patrec.2014.01.008> (2014).
42. De Angel, V. *et al.* Digital health tools for the passive monitoring of depression: A systematic review of methods. *npj Digit. Med.* **5**(1), 3. <https://doi.org/10.1038/s41746-021-00548-8> (2022).
43. Adler, D. A., Wang, F., Mohr, D. C. & Choudhury, T. Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *PLoS One* **17**(4), 1–20. <https://doi.org/10.1371/journal.pone.0266516> (2022).
44. Oh, B., Hwang, J., Seo, S., Chun, S. & Lee, K.-H. Inductive Gaussian representation of user-specific information for personalized stress-level prediction. *Expert Syst. Appl.* **178**, 114912 (2021).
45. Du, W. Application of improved smote and xgboost algorithm in the analysis of psychological stress test for college students. *J. Electr. Comput. Eng.* **2022** (2022)
46. Foumani, N. M., Tan, C. W., Webb, G. I., & Salehi, M. *Improving Position Encoding of Transformers for Multivariate Time Series Classification* (2023).
47. Powers, D. M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* (2020).
48. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **30**(7), 1145–1159. [https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2) (1997).
49. Hand, D. J. & Till, R. J. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach. Learn.* **45**(2), 171–186. <https://doi.org/10.1023/A:1010920819831> (2001).
50. Mohammadi Foumani, S. N., Wei Tan, C., & Salehi, M. Disjoint-CNN for multivariate time series classification. In *2021 International Conference on Data Mining Workshops (ICDMW)* 760–769 (2021). <https://doi.org/10.1109/ICDMW53433.2021.00099>.
51. Philip Schmidt, A., Reiss, R. D., & Kristof Van Laerhoven, I. W. A multimodal dataset for wearable stress and affect detection. In *Proceedings of the International Conference on Multimodal Interaction* (2018).
52. Dzieżyc, M., Gjoreski, M., Kazienko, P., Saganowski, S. & Gams, M. Can we ditch feature engineering? End-to-end deep learning for affect recognition from physiological sensor data. *Sensors* <https://doi.org/10.3390/s20226535> (2020).
53. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics* 1273–1282 (2017). PMLR.
54. Li, T., Sahu, A. K., Talwalkar, A. & Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.* **37**(3), 50–60 (2020).
55. Yang, Q., Liu, Y., Chen, T. & Tong, Y. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol. (TIST)* **10**(2), 1–19 (2019).
56. Larochelle, H., Erhan, D., & Bengio, Y. Zero-data learning of new tasks. In *AAAI*, Vol. 1, 3 (2008).

57. Rohrbach, M., Stark, M., & Schiele, B. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR 2011* 1641–1648 (IEEE, 2011).
58. Wang, W., Zheng, V. W., Yu, H. & Miao, C. A survey of zero-shot learning: Settings, methods, and applications. *ACM Trans. Intell. Syst. Technol. (TIST)* **10**(2), 1–37 (2019).
59. Bengio, Y., Frasconi, P., & Simard, P. The problem of learning long-term dependencies in recurrent networks. In *IEEE International Conference on Neural Networks* 1183–1188 (IEEE, 1993)
60. Kandemir, M., Vetek, A., Gönen, M., Klami, A. & Kaski, S. Multi-task and multi-view learning of user state. *Neurocomputing* **139**, 97–106. <https://doi.org/10.1016/j.neucom.2014.02.057> (2014).
61. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. Density-based spatial clustering of applications with noise. In *International Conference on Knowledge Discovery and Data Mining* Vol. 240, 6 (1996).
62. Jang, E., Gu, S., & Poole, B. *Categorical Reparameterization with Gumbel-Softmax* (2017)

## Author contributions

M.F. and T.R. coordinated the research project and supervised the project. A.S., N.S., and I.D. implemented the baseline models and the CALM-Net framework, conducted experiments in 5-fold cross-validation, and discussed and analyzed the results. Y.L., I.D. implemented the models in the Clustered and Branched CALM-Net, conducted experiments in 5-fold cross-validation and leave-one-subject-out validation, and discussed and analyzed the results. Y.L. and I.D. contributed to the writing of the manuscript with assistance from M.F. and T.R. All authors contributed to the discussion and design of the Branched CALM-Net framework, evaluation schema, and proofreading of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-56674-2>.

**Correspondence** and requests for materials should be addressed to Y.L.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024