Stress prediction using micro-EMA and machine learning during COVID-19 social isolation

Huining Li [a],[*], Enhao Zheng [a], Zijian Zhong [a], Chenhan Xu [a], Nicole Roma [b], Steven Lamkin [b], Tania T. Von Visger [b], Yu-Ping Chang [b], Wenyao Xu [b]

[a] *Department of Computer Science and Engineering, University at Buffalo, United States*
[b] *School of Nursing, University at Buffalo, United States*

ARTICLE INFO

ABSTRACT

Accurately predicting users' perceived stress is beneficial to aid early intervention and prevent both mental illness and physical disease during the COVID-19 pandemic. However, the existing perceived stress predicting system needs to collect a large amount of previous data for training but has a limited prediction range (i.e., next 1–2 days). Therefore, we propose a perceived stress prediction system based on the history data of micro-EMA for identifying risks 7 days earlier. Specifically, we first select and deliver an optimal set of micro-EMA questions to users every Monday, Wednesday, and Friday for reducing the burden. Then, we extract time-series features from the past micro-EMA responses and apply an Elastic net regularization model to discard redundant features. After that, selected features are fed to an ensemble prediction model for forecasting fine-grained perceived stress in the next 7 days. Experiment results show that our proposed prediction system can achieve around 4.26 (10.65% of the scale) mean absolute error for predicting the next 7 day's PSS scores, and higher than 81% accuracy for predicting the next 7 day's stress labels.

1. Introduction

During the COVID-19 pandemic, college students may experience increased anxiety and stress levels due to social isolation and uncertainty regarding academic achievement brought by online education (Cao et al., 2020; Hwang, Rabheru, Peisah, Reichman, & Ikeda, 2020). The prediction of perceived stress is essential to aid early intervention and prevent potential harmful effects during the pandemic. Consequently, there is a critical need to investigate stress prediction.

Most prior works have focused on detecting the current stress state, rather than true prediction (Ferdous, Osmani, & Mayora, 2015; Jiang et al., 2019). Recent works started to explore a predictive model to forecast users' future mental health states based on individual histories. Jaques, Taylor, Sano, Picard, et al. (2017) leveraged personalized multitask learning to predict the next day's stress level from physiological signals, smartphone usage, location, behavioral survey data, and weather data in the past 15 days. Umematsu, Sano, Taylor, and Picard (2019) showed that using the previous 7 days of multi-modal data with an LSTM model can give acceptable results in the next day's stress label prediction. Yu, Klerman, Picard, and Sano (2019) compared the machine learning models to forecast the stress labels for the next few days, and found that the prediction accuracy for the next 3–7 days is significantly lower than the prediction for the next day. To summarize, these existing works mainly have two limitations: 1) *limited*
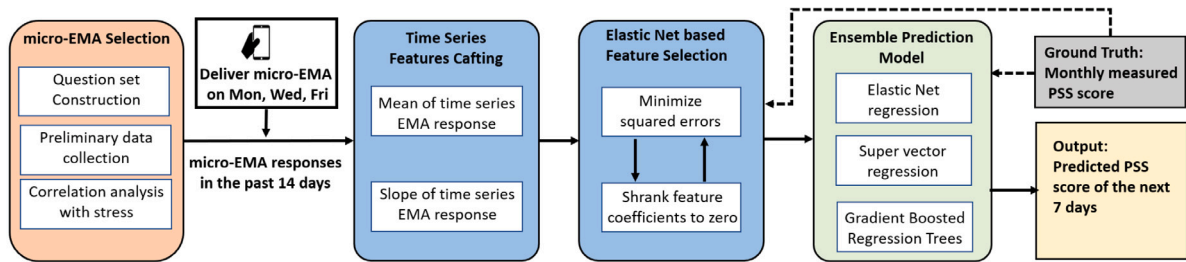
Fig. 1. The prediction system for forecasting the next 7 days' perceived stress includes optimal micro-EMA questions selection and delivery, time series features extraction and Elastic Net-based feature selection, and an ensemble prediction model.

prediction range, i.e., most predictive models cannot be used to forecast more than two days in advance; 2) *great efforts for training model*, i.e., most works need to continuously (24-hour-a-day) collect multi-modal data from physiological sensors, smartphone apps, and behavioral surveys for 7–15 days. Therefore, how to extend the prediction time span with small efforts is still an ongoing issue in stress prediction.

To this end, we develop a prediction system based on the history data of microinstruction ecological momentary assessment (micro-EMA) to forecast an individual's perceived stress for the next 7 days. micro-EMA repeatedly prompts users to answer a small number of questions to capture an individual's mental health states in close proximity to the time that symptoms and behaviors happen, so it can mitigate recall bias and track users' stress with fine-grained resolution and less burden (King et al., 2019). To realize the prediction system, we first identify an optimal set of micro-EMA questions that yields strong correlations with the Perceived Stress Scale (PSS) (Cohen, Kamarck, Mermelstein, et al., 1994). These candidate micro-EMA questions are delivered to users every Monday, Wednesday, and Friday to enable high efficiency with less burden. Then, time-series features (e.g., mean and slope) are extracted from the previous 14 days of micro-EMA responses. To avoid overfitting issues, we adopt the Elastic net regularization model to automatically select more relevant features and discard redundant features. Finally, these selected features are fed to a no-bias ensemble prediction model that consists of Elastic Net regression, super vector regression, and gradient boosted regression for forecasting the next 7 days' perceived stress at a granular level.

Our contributions are summarized as three-fold:

- We develop a micro-EMA-based prediction system that consists of optimal micro-EMA selection and a machine learning-based prediction model to forecast users' perceived stress for the next 7 days.
- Our proposed stress prediction system can identify users' stress risks 7 days earlier with a low burden, which is beneficial to aid early intervention during the COVID-19 pandemic.
- We leverage mean absolute error and Pearson's r, and accuracy to evaluate the performance of predicting stress scores and labels, and investigate the prediction performance with different demographics.

2. Methodology

In this section, we demonstrate the prediction system that consists of optimal micro-EMA questions selection, time series features crafting, Elastic Net-based features selection, and ensemble prediction model, as shown in Fig. 1.

2.1. Optimal micro-EMA questions selection

The first step of our system is to explore a set of optimal micro-EMA questions for prediction. We first construct a question set in the aspect of social, physical, sleep, and cognitive, subjective feelings. These questions are derived from different clinical questionnaires, such as UCLA loneliness scale-8 (Roberts, Lewinsohn, & Seeley, 1993), FACIT fatigue scale (Tennant, 2015), generalized anxiety disorder 7-item (Mossman et al., 2017), cognitive and affective mindfulness scale-revised (Feldman, Hayes, Kumar, Greeson, & Laurenceau, 2007), and sleep hygiene index (Mastin, Bryson, & Corwyn, 2006), as shown in Table 1. Each question has four options ranging from never/rarely (0 points) to almost always (3 points).

Preliminary Data Collection: In the preliminary study, we deliver 13 independent questions and a perceived stress scale (PSS) survey to 20 subjects in the same while on three different days. These subjects are requested to answer all questions and the PSS survey carefully each time via a smartphone app. In total, we collect 60 responses for each question and the PSS survey.

Correlation Analysis: To identify a set of micro-EMA questions for prediction, we calculate the Pearson correlation between the PSS score and each independent response. Based on Pearson's r and *p*-value, we select the candidate questions that yields moderately to strongly positive correlation with perceived stress (*p*-value <0.0001): Q2: Tired ($r = 0.66$), Q5: Isolation ($r = 0.54$), Q11: Go to Bed ($r = 0.72$), Q12: Distracted ($r = 0.58$). Additionally, we further identify one negatively correlated question, i.e., Q9: Control Worry ($r = -0.62$), to investigate the ability of positive symptoms to predict stress. The inter-correlation between all independent questions and PSS is exhibited in Fig. 2.

Finally, we obtain 5-item micro-EMA questions. Users are prompted to answer these micro-EMA questions every Monday, Wednesday, and Friday for prediction.
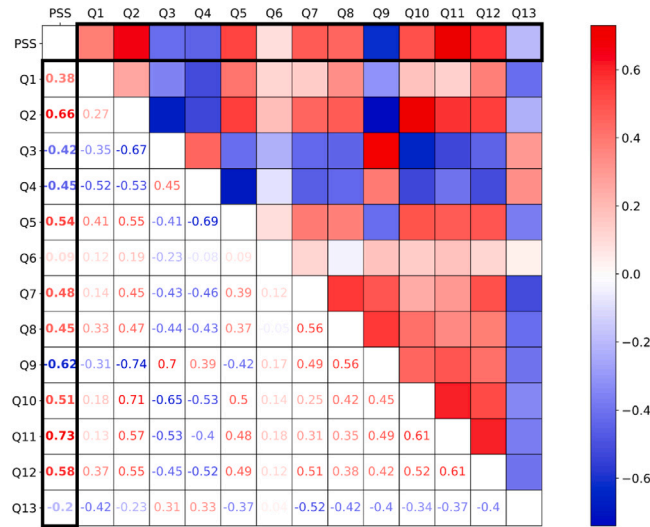
Fig. 2. Inter correlation between Q1–Q13 and PSS. The correlation between each question and PSS is outlined in black. Q2, Q5, Q9, Q11, Q12 are selected as micro-EMA questions.

Table 1

A set of questions derived from clinical surveys.

Q1: Did you have poor appetite or overeating?
Q2: Did you feel tired or have little energy?
Q3: Did you often do physical activities?
Q4: Did you often communicate with others?
Q5: Did you feel isolation from others?
Q6: Did you need sleep during the day?
Q7: Did you feel easily annoyed or irritable?
Q8: Did you have trouble relaxing?
Q9: Were you able to stop or control worrying?
Q10: Did you have trouble sleeping?
Q11: Did you go to bed feeling stressed, angry, upset, or nervous?
Q12: Did you feel easily distracted?
Q13: Did you be able to tolerate emotional pain?
Options: 0-rarely; 1-sometimes; 2-often; 3-almost always

2.2. Time series feature crafting

The second step is to extract time series EMA features. We first calculate the mean of each time series EMA response in the past 14 days. It characterizes the average level of the user's symptoms during the 14-day period. For example, the mean of tired feeling over the 14-day period represents the average intensity that the user feels tired daily.

To capture the change of symptoms, we further calculate the slope for each time series EMA response, which can characterize the direction and steepness of the change. Specifically, we fit each time series EMA response with a linear regression model and use the regression coefficient as the corresponding slope. The Bayesian information criterion (BIC) (Weakliem, 1999) is applied to select the best linear regression model for each time series EMA response, with a lower BIC value suggesting a better fitting model. A positive slope indicates the increase of a certain symptom, while a negative slope indicates the decrease of a certain symptom. For example, a positive slope in tired feeling represents that the user experiences more severe symptoms of feeling tired over time. The absolute value of the slope describes how fast the symptom changes over time, with a higher value indicating a faster change. We calculate three slopes for each time series EMA response: over the first 7 days, over the last 7 days, and over the whole 14 days.

2.3. Elastic net based feature selection

After achieving mean features and slope features, the next step is to select more relevant features and discard redundant features for avoiding overfitting issues. The Least Absolute Shrinkage and Selection Operator (LASSO) (Fonti & Belitser, 2017) is a conventional feature selection approach to enhance prediction accuracy for a small number of samples with a large number of features. However, the LASSO method fails to select a group of features that are highly correlated with each other. It tends to select one feature from a group and ignore the others.

To overcome this shortage, we use the Elastic net method (Zou & Hastie, 2005) to select groups of correlated features. The Elastic net-based feature selection solves the following optimization problem:

$$\hat{\beta}_{en}(\lambda) = \arg\min_{\beta}(\frac{\|\mathbf{Y} - \mathbf{X}\beta\|_2^2}{n} + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1), \tag{1}$$

where n is the number of samples, $\mathbf{Y}$ is $n\times 1$ response vector (i.e., ground truth), $\mathbf{X}$ is $n\times p$ feature matrix, β is $p\times 1$ coefficient vector. When minimizing the optimization problem, the regression coefficients are shrunk by combining L1-norm penalty and L2-norm penalty together. The L1-norm part of the penalty shrinks some coefficients to zero to produce a sparse model, and the L2-norm part of the penalty stabilizes the L1 regularization path using the LARS-EN algorithm (Reunanen, 2003) which encourages group effect. In this way, the features with coefficients equal to zero are discarded from the model.

2.4. Ensemble prediction model

The prediction of perceived stress scores is a regression problem. In predictive modeling, a single regression model that trains on a set of features may have biases or high variability. Therefore, we employ the following widely used regression model to improve the reliability for prediction:

- Elastic Net Regression: Considering some of the input features are highly correlated with each other, we adopt Elastic net regression (Zhang et al., 2017) to better fit these features. It is penalized with both the L1-norm and L2-norm to efficiently shrink the regression coefficients and set some to zero. Specifically, the constant that multiplies the penalty terms is set as 1, and the ElasticNet mixing parameter is set as 0.5 in our implementation.
- Super Vector Regression (SVR): SVR is a supervised learning method to find an appropriate line or hyperplane in higher dimensions to fit the features within an acceptable error. We adopt SVR for prediction as it has superior generalization ability for unseen data regression. Specifically, we use the Gaussian kernel function to find a maximum-margin hyperplane.
- Gradient Boosted Regression Trees (GBRT): GBRT is an ensemble of several weak regression trees (Friedman, 2001). It builds base regression trees (i.e., estimators) sequentially for mitigating the bias of the previously combined estimators. GBRT is applied because it is robust to overfitting and less sensitive to outliers. Specifically, the loss function is based on squared error, the learning rate is set as 0.1, and the number of boosting stages is set as 100 in our implementation.

Finally, we combine the output scores of these three prediction models by weighted sum. The weights are set as 0.5, 0.2, 0.3 for Elastic Net, SVR, and GBRT, respectively.

3. Evaluation

3.1. Experiment setup

We recruit college students who were taking online courses during the Fall term in 2020 to attend our 8-week study. Potential participants undergo an initial online screening to include the ones having stress issues caused by social isolation in the COVID-19 pandemic. This study is approved by the Institutional Review Board (IRB). Finally, 27 eligible subjects (12 female and 15 male) are enrolled with the age ranging from 18 to 37. They need to use our developed smartphone app for mental health assessment during the study. This app prompts participants to answer 5-item micro-EMA questions every Monday, Wednesday, and Friday, and delivers the Perceived Stress Scale (PSS) survey on day 1, day 28, and day 56. Participants are requested to respond to all pushed micro-EMA questions and the PSS survey on that day. PSS scores are in the range of 0–40 with 0–13 indicating low stress, 14–26 indicating moderate stress, and 27–40 indicating high stress, which is regarded as stress ground truth. We also monitor participants' PSS scores during the whole study and provide clinical care if in need. We use leave-one-record-out cross-validation to train and test the prediction model. The leave-one-record-out cross-validation leaves one sample out from the dataset for testing and uses all other samples for training, the process repeats until all samples are tested once, and thereby results in 2862 training samples and 54 test samples.

3.2. Metrics

We use the following metrics to evaluate the performance of predicting perceived stress scores/labels in the next 7 days:

- Mean absolute errors (MAE): MAE measures the error between the predictive scores and ground truth, given by: $MAE = \frac{1}{n}\sum|y_i - x_i|$, where y_i is the ground truth, x_i is the predictive score. The Lower MAE indicates the prediction is closer to the ground truth.
- Pearson correlation coefficient (Pearson's r): Pearson's r measures the linear relations between the predictive scores and the ground truth, and $r > 0.5$ means linear relation.
- Accuracy: *Accuracy* measures the performance of predicting perceived stress labels (i.e., low, moderate, high), given by: $Accuracy = \frac{true\ classification\ samples}{total\ samples}$.
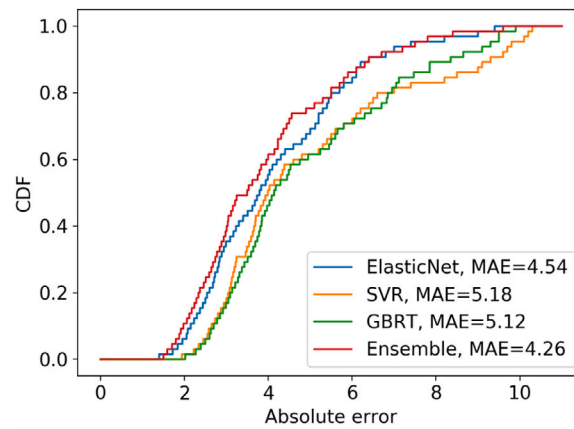
Fig. 3. The cumulative distribution function (CDF) of the absolute errors of predicted stress scores under leave-one-record-out cross validation.
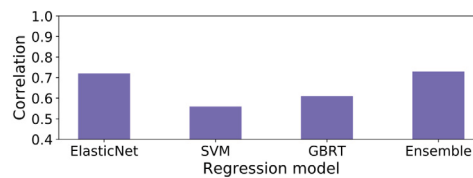

Fig. 4. The correlation between predicted stress scores and ground truth with different regression models under leave-one-record-out cross validation.
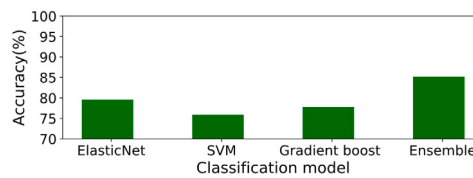

Fig. 5. The performance of predicting the perceived stress labels with different classification models under leave-one-record-out cross validation.

3.3. Overall performance

Regression Performance: We first show the proximity between the predictions and the ground truth. Fig. 3 shows the cumulative distribution function (CDF) (Drew, Glen, & Leemis, 2000) of the absolute error of the predicted stress scores in detail. CDF plot demonstrates the distribution of absolute error for each prediction. As observed, our proposed ensemble regression model has the lowest MAE (i.e., ±4.26 error, 10.65% of the scale) compared with Elastic Net regression, SVR, and GBRT. Our model also outperforms a recent stress prediction work that reports a MAE with 13.7% of the scale (Yu et al., 2019). For the ensemble regression model, around 75% of the prediction results are with absolute error less than 5 (i.e., 12.5% of the scale) when applying leave-one-record-out cross-validation.

Then, we evaluate the linear relations between the predictions and the ground truth. As shown in Fig. 4, the predicted PSS scores strongly correlate with the ground truth when using the ensemble regression model or Elastic Net regression model, with Pearson's $r>0.7$, $p<0.0001$. The correlation performance drops more than 10% when applying SVM and GBRT. To conclude, our proposed ensemble regression model performs better than the Elastic Net regression model, which in turn outperforms GBRT and SVM.

Classification Performance: To examine the system performance of predicting the stress labels (low, moderate, high), we further train some classification models, i.e., Elastic Net logistic regression (classification) model, SVM classifier, gradient boosted classifier. The ensemble classifier combines the outcome of these three classifiers by weighted sum and the fusion weights are optimized based on logistic regression. As observed in Fig. 5, the ensemble classifier achieves up to 85.2% accuracy. By contrast, the accuracy of the other three classifiers is less than 80%. In conclusion, the ensemble classifier employed in predicting stress labels can achieve superior performance.

3.4. Performance with demographic parameters

We are curious about whether the stress prediction performance will be affected by the demographics factors. Therefore, we evaluate the system using leave-one-record-out cross-validation under different demographic parameters as follows:
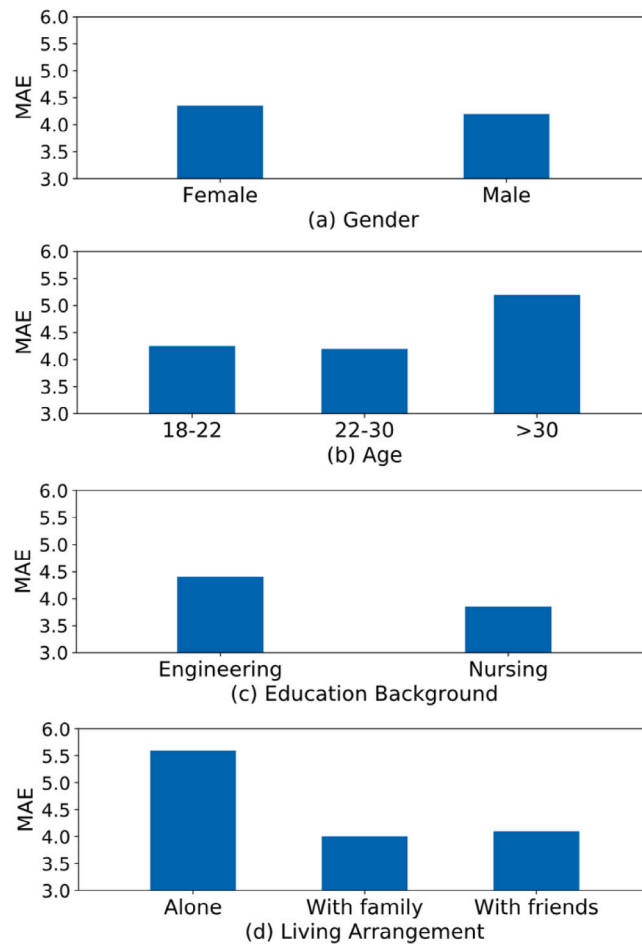
Fig. 6. The performance of predicting the perceived stress scores under different demographics.

Gender: 12 female and 15 male are enrolled in this evaluation. Fig. 6 (a) shows that the MAE of predicted stress scores is almost the same (i.e., around 4.2) for male subjects and female subjects. It indicates that our prediction model is insensitive to gender factors.
Age: In the age group, 15 subjects aging from 18–22, 10 subjects aging from 22–30, and 2 subjects aging above 30. As shown in Fig. 6 (b), the prediction MAE for the 18–22 age group and 22–30 age group are around 4.1, whereas the prediction MAE for users above 30 years old increases notably, reaching up to 5.6. It suggests that our proposed model is not good at studying the feature distribution of aging groups due to limited samples. In other words, the age factor might contribute to the prediction performance.
Education Background: As for the education background, there are 15 subjects major in engineering and 12 subjects major in nursing. Fig. 6 (c) shows a slight difference in prediction MAE for engineering group and nursing group, which are 4.4 and 3.85, respectively. The results indicate that we might need to consider the education parameter in the personalized prediction model.
Living Arrangement In the evaluation, 5 subjects live alone, 12 subjects live with family and 10 subjects live with friends. As shown in Fig. 6 (d), the prediction MAE for living with family and friends groups is around 4, whereas the prediction MAE for users who live alone is much higher, achieving up to 5.6. This is because our prediction model is trained on a limited number of college students who usually live with friends or family, which makes our model lack generality to some extent. It also indicates that the stress prediction model is sensitive to the life pattern. We plan to investigate demographic factors as predictors in the future.

3.5. Performance with new users

In real practice, we hope the stress prediction model can be easily applied to new users without consuming much effort for training. Thereby, we use leave-one-subject-out cross-validation to evaluate the performance of predicting a new participant whose history data are not available in the training phase. Fig. 7 shows an overall tendency of better prediction performance with the increasing number of subjects for training, under both user-dependent and independent settings. The best performance of predicting existing users is superior to the best of predicting new users, which is expected. To be specific, the prediction MAE on new subjects steadily decreases as the number of subjects for training increases. Eventually, the MAE drops to 6.5 (i.e., 16.25% of the scale) when
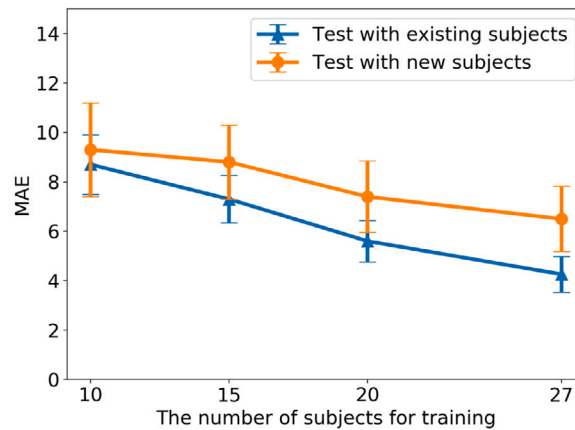
Fig. 7. The stress prediction performance for new subjects and existing subjects.
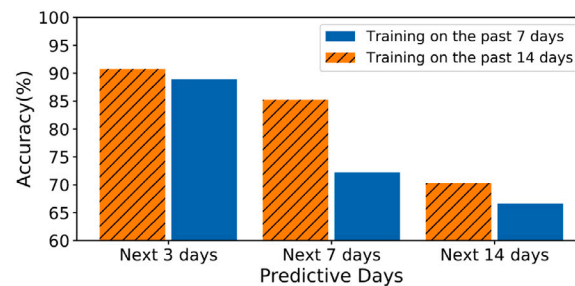

Fig. 8. The performance of predictive days under two training settings.

we have 27 subjects for training. In other words, it will be easy to find the closest existing user for a new user and "pretend" that they are the same person when the training set is large. To improve the prediction performance on new users, we can 1) expand the training dataset to involve different distributions of features for the generalization of the model; 2) develop a domain adaptation model to adapt the training weights to the existing model.

3.6. Performance on predictive days

It is interesting to investigate how many days we need to look back for training the model and how many days this model can predict. Therefore, we evaluate the prediction performance of the next 3–14 days' stress labels under two training settings. As shown in Fig. 8, the model can achieve above 85% accuracy for predicting the next 3–7 day's stress labels using the past 14 days history data. However, the accuracy drops to 70% or even lower when using the past 7 days for predicting the next 7–14 days. The difficulties of long-term prediction originate from the time-sensitivity of the predictors. Therefore, on the one hand, we reduce the number of predictors and select the most optimal predictors by performing correlation analysis. On the other hand, we investigate the changes of predictors over time and extract the changing slope of the predictors for feeding to the model. Considering both the training efforts and prediction time span, our model is efficient to forecast the next 7 days' stress by using the past 14 days information.

4. Conclusion and future work

In this paper, we develop a prediction system based on micro-EMA questions for forecasting the next 7 day's perceived stress. We first select an optimal set of micro-EMA questions that yields a strong correlation with PSS for reducing users' burden. After achieving users' responses to micro-EMA questions in the past 14 days, we extract time-series features and adopt an Elastic net regularization model for selecting more efficient features, which are then fed to an ensemble prediction model. Experiment results show that the prediction system can forecast stress scores with 6.5 MAE for a new participant. In future work, we plan to enlarge our data size to hundreds of subjects and further extend our study to other populations for developing a more generalized wellbeing states prediction model, such as patients with mental disorders, and caregivers. Meanwhile, we will develop a just-in-time adaptive intervention system based on early identified risks to help users reduce stress.

Declaration of competing interest

Acknowledgments

References

Cao, W., Fang, Z., Hou, G., Han, M., Xu, X., Dong, J., et al. (2020). The psychological impact of the COVID-19 epidemic on college students in China. *Psychiatry Research*, *287*, Article 112934.

Cohen, S., Kamarck, T., Mermelstein, R., et al. (1994). Perceived stress scale. *10*, In *Measuring stress: A guide for health and social scientists* (pp. 1–2).

Drew, J. H., Glen, A. G., & Leemis, L. M. (2000). Computing the cumulative distribution function of the Kolmogorov–Smirnov statistic. *Computational Statistics & Data Analysis*, *34*(1), 1–15.

Feldman, G., Hayes, A., Kumar, S., Greeson, J., & Laurenceau, J.-P. (2007). Mindfulness and emotion regulation: The development and initial validation of the cognitive and affective mindfulness scale-revised (CAMS-R). *Journal of Psychopathology and Behavioral Assessment*, *29*(3), 177–190.

Ferdous, R., Osmani, V., & Mayora, O. (2015). Smartphone app usage as a predictor of perceived stress levels at workplace. In *2015 9th international conference on pervasive computing technologies for healthcare* (pp. 225–228). IEEE.

Fonti, V., & Belitser, E. (2017). Feature selection using lasso. VU Amsterdam research paper in business analytics. vol. 30. (pp. 1–25).

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 1189–1232.

Hwang, T.-J., Rabheru, K., Peisah, C., Reichman, W., & Ikeda, M. (2020). Loneliness and social isolation during the COVID-19 pandemic. *International Psychogeriatrics*, *32*(10), 1217–1220.

Jaques, N., Taylor, S., Sano, A., Picard, R., et al. (2017). Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation. In *IJCAI 2017 workshop on artificial intelligence in affective computing* (pp. 17–33). PMLR.

Jiang, J.-Y., Chao, Z., Bertozzi, A. L., Wang, W., Young, S. D., & Needell, D. (2019). Learning to Predict Human Stress Level with Incomplete Sensor Data from Wearable Devices. In *Proceedings of the 28th ACM International conference on information and knowledge management*. (pp. 2773–2781).

King, Z. D., Moskowitz, J., Egilmez, B., Zhang, S., Zhang, L., Bass, M., et al. (2019). Micro-stress EMA: A passive sensing framework for detecting in-the-wild stress in pregnant mothers. *3*, In *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* (3), (pp. 1–22). ACM New York, NY, USA.

Mastin, D. F., Bryson, J., & Corwyn, R. (2006). Assessment of sleep hygiene using the sleep hygiene index. *Journal of behavioral medicine*, *29*(3), 223–227.

Mossman, S. A., Luft, M. J., Schroeder, H. K., Varney, S. T., Fleck, D. E., Barzman, D. H., et al. (2017). The generalized anxiety disorder 7-item (GAD-7) scale in adolescents with generalized anxiety disorder: Signal detection and validation. *Annals of Clinical Psychiatry: Official Journal of the American Academy of Clinical Psychiatrists*, *29*(4), 227.

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, *3*(Mar), 1371–1382.

Roberts, R. E., Lewinsohn, P. M., & Seeley, J. R. (1993). A brief measure of loneliness suitable for use with adolescents. *Psychological Reports*, *72*(3_suppl), 1379–1391.

Tennant, K. (2015). Assessment of fatigue in older adults: the FACIT fatigue scale (version 4). *Supportive Care in Cancer*, *23*(5), 1355–1364.

Umematsu, T., Sano, A., Taylor, S., & Picard, R. W. (2019). Improving students' daily life stress forecasting using lstm neural networks. In *2019 IEEE EMBS international conference on biomedical & health informatics* (pp. 1–4). IEEE.

Weakliem, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research*, *27*(3), 359–397.

Yu, H., Klerman, E. B., Picard, R. W., & Sano, A. (2019). Personalized wellbeing prediction using behavioral, physiological and weather data. In *2019 IEEE EMBS international conference on biomedical & health informatics* (pp. 1–4). IEEE.

Zhang, Z., Lai, Z., Xu, Y., Shao, L., Wu, J., & Xie, G.-S. (2017). Discriminative elastic-net regularized linear regression. *IEEE Transactions on Image Processing*, *26*(3), 1466–1481.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *67*(2), 301–320.