Title

Personalized multi-task attention for multimodal mental health detection and explanation

Permalink

https://escholarship.org/uc/item/26b016s4

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors

Song, Donglei
Shen, Qiang
Feng, Haotian
et al.

Publication Date

2023

Peer reviewed

Personalized Multi-task Attention for Multi-modal Mental Health Detection and Explanation

Donglei Song[1], Qiang Shen[1], Haotian Feng[1], Rui Song[2], Fausto Giunchiglia[3], Hao Xu[1,*]
[1]College of Computer Science and Technology, Jilin University, Changchun, China
[2]School of Artificial Intelligence, Jilin University, Changchun, China
[3]DISI, University of Trento, Trento, Italy
[*]Corresponding Author: xuhao@jlu.edu.cn

Abstract

The unprecedented spread of smartphone usage and its various boarding sensors have been garnering increasing interest in automatic mental health detection. However, there are two major barriers to reliable mental health detection applications that can be adopted in real-life: (a)The outputs of the complex machine learning model are not explainable, which reduces the trust of users and thus hinders the application in real-life scenarios. (b)The sensor signal distribution discrepancy across individuals is a major barrier to accurate detection since each individual has their own characteristics. We propose an explainable mental health detection model. Spatial and temporal features of multiple sensory sequences are extracted and fused with different weights generated by the attention mechanism so that the discrepancy of contribution to classifiers across different modalities can be considered in the model. Through a series of experiments on real-life datasets, results show the effectiveness of our model compared to the existing approaches.

Keywords: Mental health; Mobile sensing; Attention; Deep learning; Explainable machine learning

Introduction

Mental health problems are on the rise globally and affect a large number of people of all ages worldwide (Woodward et al., 2020; R. Wang et al., 2018). A negative mental health state not only leads to a severe negative impact on daily life such as work and school performance but also contributes to a high proportion of illness-related burdens. Therefore, there is a need to understand what is leading to negative mental health and how to detect the mental health state automatically and dynamically so that worse affection can be avoided.

Traditionally, clinical physical and mental health assessment visits rely on periodic self-reports, which require a lot of time and effort. In addition, the diagnosis result usually represents a very specific time window into patients' lives, which means that the optimal timing of treatment has been missed. Therefore, if daily mental health state can be measured in time, it can provide early warnings and prevent severe disorders. As the smartphone is ubiquitous and adopted as a computing platform with richer functionality in recent years, it gives the ability to continuously inform clinical inferences and timely intervention on mental health states. Automatic detection of mental well-being (R. Wang et al., 2016, 2014a) can serve healthcare applications, digital personal assistants, aging systems, and many other domains.

Machine Learning(ML) is a widely-used method to detect the mental health state accurately. Conventional ML approaches have made tremendous progress on mental health detection by adopting algorithms such as decision trees, support vector machines, naive Bayes, and hidden Markov models. As the dimension and quantity of the data continuously increase, however, those methods may heavily rely on heuristic hand-crafted feature extraction, which is limited by human domain knowledge. In addition, high-level features of sensory readings can not be learned with conventional ML models. In recent years, due to the similar structure of text data and sensor data, deep learning techniques convolutional neural network (CNN) and recurrent neural network (RNN) (Sharif Razavian, Azizpour, Sullivan, & Carlsson, 2014; Sundermeyer, Schlüter, & Ney, 2012) are adopted on the task of sensory series classification such as emotion and health detection. The studies of detecting and predicting mood, stress, and mental health using sensory data collected from smartphones have increased interest (Tsai, Lai, Chiang, & Yang, 2013; Tsai, Tsai, Chiang, & Yang, 2018; Tsai, Lai, & Vasilakos, 2014). For instance, many studies focusing on detecting mental health using the deep learning models (Almaslukh, AlMuhtadi, & Artoli, 2017; Suhara, Xu, & Pentland, 2017; Jaques, Taylor, Sano, Picard, et al., 2017; Li & Sano, 2020; Li, Yu, & Sano, 2019)

However, the black-box deep learning models can hardly earn users' trust in healthcare because of their unexplainable nature and risk factors. Although existing deep learning approaches can automatically and accurately learn features to infer mental health states, most are suffering a significant drawback—the lack of explanation of the model. In addition, the multi-modality sensors vary from the data type, sampling rates, etc (Radu et al., 2018). Most of the existing works fuse the multiple modalities of sensor readings without considering the discrepancy of data and its importance, which leads to multi-modality streaming sensors fusion challenges.

To address these issues above, we propose an attention-based model for mental health detection by generating explanations based on visualizing the attention mechanism. Attention mechanism (Vaswani et al., 2017) has been proven to perform well in natural language processing areas because of the ability of weighted averaging of a series of input vectors. In our proposed model, CNNs are firstly used to extract features of different modalities, whose outputs are then fed into an attention-based fusion layer considering the diverse importance of multiple modalities. Then an attention-based bidirectional LSTM (Bi-LSTM) extracts and fuses the relatively im-

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

portant features from multiple time windows. To promote the reliability and explainability of the model, the explanation of when and why mental health is detected is generated according to the visualization of attention weights. In addition, the contextual information is also taken into consideration in the model by being fused with features of multi-modal sensory features. To evaluate the proposed model, we conduct experiments on the real-life mobile sensing dataset, which is collected in the university to record students' daily-life mental health and smartphone sensory data. The results demonstrate that the proposed method outperforms other methods.

Related Work

Mobile Sensing for Mental Health Detection

Mental health refers to the wellbeings including emotional, psychological, and social state (Taylor & Brown, 1988; Gravatt, Lindzey, & Aronson, 2013), which affects how we think, feel, and perform. The relationships between depression and generic behavioral features from passive sensing have been studied in the previous work (Canzian & Musolesi, 2015; Min et al., 2016; Mohr, Zhang, & Schueller, 2017).

In today's life, human behaviors such as activity, mobility, social interaction, and smartphone usage can be sensed by multiple sensors embedded in smartphones and other wearable devices (Tsai, Lai, Chiang, & Yang, 2014). The study in (Wahle et al., 2016) uses behavioral features such as location, activity, smartphone usage, wifi, and log of calls and achieved an accuracy of 61.5%. (Farhan et al., 2016) detects negative mental health states using a machine learning model on the dataset with 79 college students over eight months. The work in (Thakur, 2020) extracts daily-life behavioral features from smartphone usage and sensory data to predict mental health state. To study depression in the university environment, (R. Wang et al., 2014b) conducts experiments to collect data from 48 students for ten weeks. They analyze further the correlations between depression scores and various behavioral features such as sleep duration and contextual information (X. Xu et al., 2019; Matthews, Abdullah, Gay, & Choudhury, 2014). The study in (Ahmed, Jeon, & Picciali, 2021) proposes an IoT-based non-invasive automated patient discomfort monitoring system, using a deep learning-based algorithm. (Tuli et al., 2020) proposed a novel framework for integrating ensemble deep learning in edge computing devices and deployed it for healthcare monitoring.

Multi-modal Sensor Fusion for Emotion Detection

To achieve accurate performance, various sensor readings are fed into the machine learning model for the detection of emotion (Costa, Rincon, Carrascosa, Julian, & Novais, 2019; G. Xu, Li, & Liu, 2020), activity (Wu, Liu, Zhu, Wang, & Zha, 2020; Yao, Hu, Zhao, Zhang, & Abdelzaher, 2017) and mental health (Lu et al., 2018) in the recent work, especially in the scenario of an open world (Saeed, Ozcelebi, Trajanovski, & Lukkien, 2018; Vaizman, Weibel, & Lanckriet, 2018). Existing sensor fusion work can be divided into two commonly adopted multi-modal fusion approaches: shallow classifiers and deep neural networks.

Shallow classifiers such as decision trees, random forest, and support vector machine (SVM) rely on the manually-extracted features (Bishop, 2006). Feature concatenation approaches combine multiple features from each modality into a single feature vector. The work in (Bulling, Ward, & Gellersen, 2012) fuses multiple features extracted from the eye movement data to recognize reading activity using SVM. (Hemminki, Nurmi, & Tarkoma, 2013) used AdaBoost to extract and concatenate accelerometer features capturing characteristics of transportation movement patterns so that generalization and robustness of the transportation mode detection model improved. (Kapoor & Picard, 2005) proposes a multi-modality sensor fusion model for an affect recognition system based on a mixture of Gaussian Processes.

Artificial neural networks (ANN) show a promising ability to tackle the challenge of multi-modality sensors fusion (Goodfellow, Bengio, Courville, & Bengio, 2016; Ramachandram & Taylor, 2017). The work in (Liu, Zheng, & Lu, 2016) adopts a multi-modal deep learning approach on multiple physiological sequential signals for the emotion recognition task. The results indicate that the fusion of multi-modality signals improves the performance of affective computing models. A multi-modality time series classification model combining CNN and RNN structure is proposed in (Yao et al., 2017), showing promising results in multiple sensors fusion. The study in (G. Xu et al., 2020) proposes an emotional classification model for multi-modal social media to capture users' emotions in social networks. However, the features of multiple modalities do not have equal contributions to the task, and thus the features should be treated differently when being fused.

Explainable AI in Mental Health Detection

Despite deep learning approaches achieving great performance, a major drawback is that the model results are not explainable (Riccardo et al., 2018). Explainability of AI refers to the model that can generate an explanation to humans for decision making, meanwhile, both an accurate proxy of the AI model and comprehensible to humans (Arrieta et al., 2020). Explainability is one of the most significant principles for AI models to be used in practical scenarios such as healthcare (Adadi & Berrada, 2020; Khedkar, Subramanian, Shinde, & Gandhi, 2019; Pawar, O'Shea, Rea, & O'Reilly, 2020; Holzinger, Biemann, Pattichis, & Kell, 2017).

Initially, the efforts such as feature visualization are made to investigate the black box of CNNs in the task of image recognition (Yosinski, Clune, Nguyen, Fuchs, & Lipson, 2015). The work in (K. Xu et al., 2015) uses attention mechanisms to visualize how the model can automatically learn the features. The study in (Ribeiro, Singh, & Guestrin, 2016) explains the outputs of machine learning by learning an explainable and reliable model for the classifier. (D. Wang, Yang, Abdul, & Lim, 2019) explores the application-specific explanation for the machine learning model and applies it to

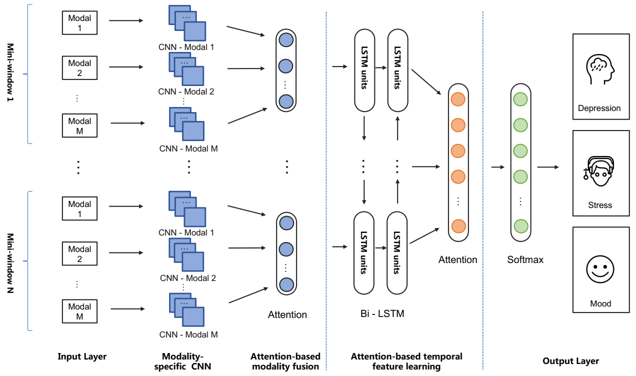

Figure 1: Structure of the model

healthcare applications. However, the process of time series classification tasks such as sensor-based activity and mental health recognition are less human-understandable since the model inputs are the sensor signals.

Methodology

The multi-modal sensor readings of smartphones (e.g., accelerometer, GPS, wifi) and contextual information of individuals are collected as the inputs of our model. The trained model can provide users with two functions: automatic mental health detection and an explanation of the result. To achieve these, an attention-based deep learning model for mental health detection is designed as shown in Fig. 1.

Problem Definition

Mental Health State. The notion of Mental Health State is formally defined as a three-tuple: $MH = \{DE, ST, MO\}$. DE denotes the level DEpression. ST refers to the answer to "how much STress are you suffering?". MO indicates the current MOod of the individual.

Multi-modal Sensory Data. The set of multi-modal sensor readings is denoted as: $\mathcal{S} = \{S_k\}, k \in \{1, \ldots, K\}$. Each sequence of single-modality sensor data from the smartphone is defined as: $S_k^t = \{s_k^1, s_K^2, s_t^3, \ldots, S_k^t, \}, t \in T$, where T refers to the length of the time window of sensory readings.

Task of Mental Health Detection. The detection of mental health can be formalized as follows. Given the multi-modal sensory series $\mathcal{S}$ and its corresponding mental health labels MH_n. The mental health detection task can be defined as finding a function $f_n : \mathcal{S} \rightarrow MH_n$ for each aspect. The goal is to minimize the loss $\mathcal{L}(f_n)$ for each aspect of the mental health state.

Single-modal Feature Extraction

In the proposed model, CNN is used to learn the feature representation of each modality of sensory readings. First, each instance's sensor sequence is split into multiple mini-windows. Each modality-specific CNN is used for learning the feature representations for each mini-window. We apply Fast Fourier Transform (FFT) to each sensor series S_k to extract more frequency domain information to obtain better local frequency patterns. We then stack both the time domain and frequency domain data into a tensor, and the set of resulting tensors for each modality is the input of the CNNs.

For each modality of sensor readings, an individual CNN is designed to extract the features in a mini-window. To extract both time domain and frequency domain features, we first apply 2d filters to learn the interaction among sensor measurement dimensions and local patterns in the frequency domain with the output. Then we apply 1d filters hierarchically to learn high-level relationships. Then we flatten the matrix into vectors and concatenate all the vectors into a K-row matrix, which is the input of the sensor fusion layer.

Attention for Multi-modality Features Fusion

In this work, attention is used to capture the varying levels of contribution from sensors at different modalities for classification. For instance, the depression state can be more related to the time of sleeping or battery rather than the Bluetooth sensor; while the stress level should have strong correlations with the activity that an individual is doing. In addition, visualizing the different contributions of features can explain the result of the black-box deep learning model. Specifically, the input can be represented as: $[v_{t1}, \ldots, v_{tk}, \ldots, v_{tK}]$, where v_{tk} represents the feature vector of modality k in the mini-window t. Then those feature vectors of all sensors are fused by using their attention scores as weights to form a uniform feature representation vector c_t. The self-attention structure can be formalized as follows:

$$\mu_{tk} = tanh(W_1 v_{tk} + b_1)) \tag{1}$$

$$\alpha_{tk} = \frac{\exp(\mu_{tk})^T w_1}{\sum_k \exp(\exp(\mu_{tk})^T w_1)} \tag{2}$$

$$c_t = \sum_k \alpha_{tk} v_{tk} \tag{3}$$

In order the learn μ_{tk}, here we compute the representation of v_{tk} through an MLP architecture $\{W1, b1, w1\}$ are parameters of the attention, which are randomly initialized and jointly learned during the training process. In this way, important modalities can be prioritized by the weights.

Global Attention for Temporal Feature Learning

We apply a Bi-LSTM network to learn the temporal dependency of sensor readings. In our model, the outputs of feature fusion layer $[x_1, x_2, \ldots, x_N]$ are fed into the Bi-LSTM network. Specifically, this layer transforms the input into the hidden layer output by multiple gate units worked as follows:

$$f_t = sigm(W_{xf} x_t + W_{hf} h_{t-1} + b_f) \tag{4}$$

$$i_t = sigm(W_{xi} x_t + W_{hi} h_{t-1} + b_i) \tag{5}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c) \tag{6}$$

$$o_t = sigm((W_{xo} x_t + W_{ho} h_{t-1} + b_o) \tag{7}$$

$$h_t = o_t \circ tanh(C_t) \tag{8}$$

where C_t and h_t are the outputs of the temp unit, f_t is the probability of how much information from the previous cell should be forgotten, i_t is the probability of how much information in temp unit should be updated, o_t is the output gate that determines how to calculate the output. In the equation above, W is a weight metric, representing the weights of the input, Operator $\circ$ represents for element-wise multiplication, $\{W_{xf}, W_{hf}, W_{xi}, W_{hi}, W_{xc}, W_{hc}, W_{xo}, W_{ho}, b_f, b_i, b_c, b_o\}$ are trainable variables, which will be updated in each training step. We use the Bi-LSTM network to extract the forward and backward features. The output of the bi-directional network can be described as follows:

$$RNNout_t = \{F_{h_t}, F_{c_t}, B_{h_t}, B_{c_t}\} \tag{9}$$

where F and B refer to forward and backward directions. Then the attention mechanism is used again to compute the sum of all hidden states weighted by their attention weights so that the temporal features can be fused considering the different contributions of different time windows.

Classification Layer

After the spatial and temporal feature extraction layers, the output of the attention-based Bi-LSTM network is fed into a classification layer. A fully connected layer and a softmax function are used to transform the outputs of the Bi-LSTM network to the probability of each mental health state, and then infer the label by finding the mental health level with maximum probability.

Data Collection

To evaluate the model's effectiveness in real life, a large-scale dataset involving multiple individuals is conducted with the support of the SmartUNI project, which aims to study university students' lifestyles, mobility, and mental health. The data collection process lasted two weeks at the university, from November 25 to December 8, 2019. Specifically, the students enrolled in the academic year 2019-2020 and 2018-2019 who were interested in the pilot were invited to an introductory presentation where they got the basic information about the project and the pilot's aims. Note that informed consent was signed to inform the students of privacy and ethics. Overall, 60 students accepted to participate and were allowed to quit at any time during the pilot. Finally, 54 students (23 males and 31 females) contributed their data to the pilot.

The pilot relied on the i-Log app (Zeni, Zaihrayeu, & Giunchiglia, 2014; Giunchiglia, Bignotti, & Zeni, 2017), which provided sensor data collection and time diaries. All the participants were required to install the app on their smartphones. The app recorded multiple sensors, both hardware (e.g., GPS, accelerometer) and software (e.g., running applications). The app also generated time diaries every 30 minutes as silent notifications to track the participants' personal context. The time diary was composed of six questions on activities(What are you doing?), locations(Where are you?), social relations(With whom are you?), mood, stress, and depression. The mental-related questions use a 5-point Likert scale. Every time diary could be answered within 150 minutes. It was possible to accumulate up to 5 notifications in the phone, after which the oldest would expire and be set to null.

Table 1: Dataset description.

Label	Depression	Stress	Mood
1 (very positive)	10284	8853	1040
2 (positive)	5383	5369	1534
3 (moderate)	2314	4667	7904
4 (negative)	1455	1404	5538
5 (very negative)	1052	1195	4472

Evaluation

This section introduces the empirical experiments for evaluating the proposed model. We will first describe the experimental setup. Secondly, the numerical results of the comparison between the state-of-the-art approaches and our proposed model are shown and discussed. Then, the impact of parameters (e.g., time window size) on the performance is analyzed.

Experiment Settings

To construct the dataset of machine learning, we cleanse extract three 5-class fine-grained level annotations about mental health (i.e., depression, stress, and mood) as truth labels of the dataset and all of the rest data, including multiple sensors and contextual information as the input of the model. Specifically, the input data is composed of multiple 10-minute sequential data, containing the 5-minute before and after answering the question. The overall description is shown in Table 1. The dataset is then divided into the training set, valid set, and test set with a ratio of 8:1:1. We use the Pytorch framework and train the model on a GPU Titan RTX to implement the proposed model. Considering the dataset size is slightly limited, the batch size of the training process is set to 32, and the network is optimized using a learning rate of 0.0001, where cross-entropy is adopted to compute the loss. The parameters are initialized by the default setting, and the model has trained 1000 iterations for each aspect of mental health. The numeric performance of the model is evaluated using accuracy.

Convergence Processes

Fig. 2 shows the loss and accuracy of different numbers of iterations using our model. It shows that the model's accuracy has been improved, and the losses are decreasing along with the increase in the number of iterations. From Fig. 2, the accuracy of depression detection is the highest among the three aspects, which reaches 86.48%. The accuracy rates of stress and mood can reach 77.42% and 81.20%, respectively.

Attention Weights for Explanation

To explain the output, we provide a visualization of the attention weights, which can be used to evaluate the different

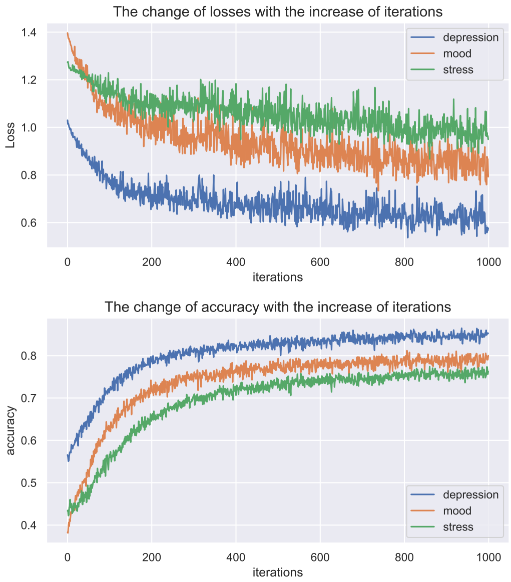

Figure 2: Classification accuracy of different numbers of iterations

impacts of various sensor modalities and different duration. Fig. 3 shows the distributions of spatial and temporal attention weights. The lighter the color, the greater the contribution of this sensor to the prediction of this mental state. Furthermore, we compute and visualize the attention weights from the perspective of multiple sensors. Fig. 4 shows the different weights of multiple sensor inputs for detecting the three aspects of mental health. For example, WiFi and context are more important than other features in predicting moods.

Comparison with Other Methods

We compare our proposed model with the following algorithms on our dataset collected in the wild:

1. **Deepsense (Yao et al., 2017)**: This is the state-of-the-art model on several types of time series classification tasks, which uses a CNN network to extract features of each sensor and another convolutional layer for sensor fusion, then it used an LSTM network to learn temporal features.

2. **CNN (LeCun, Bengio, & Hinton, 2015)**: A single CNN model with three convolutional layers, and a classification layer.

3. **Random forest (Liaw, Wiener, et al., 2002)**: The random forests are an ensemble classification model constructing multiple decision trees.

4. **SVM (Hearst, Dumais, Osuna, Platt, & Scholkopf, 1998)**: A simple support vector machine (SVM) with radial basis function (RBF) kernel.

Notice that we extract all time-domain features for the shallow models (i.e., RF and SVM) following the method in (Figo, Diniz, Ferreira, & Cardoso, 2010), including mean, std, etc. In addition, semantic features of locational information are extracted manually by point-of-interest techniques. The results of the model comparison are shown in Table 2. Compared with the existing shallow and deep models, our proposed model improves the performance of detecting all three aspects of mental health, which is mainly attributed to its capability to extract important temporal and sensory features. From the table, besides, the accuracy obtained using conventional ML classifiers is limited. Overall, the performance obtained from the deep model is significantly higher than the SVM and Random forest model.

Table 2: Comparisons of accuracy between different models.

	Depression	Stress	Mood
SVM	58.25%	43.23%	46.54%
Random Forest	58.29%	51.25%	53.92%
CNN	60.33%	53.83%	56.15%
Deepsense	68.25%	74.90%	78.02%
Proposed model	**86.48%**	**78.02%**	**81.20%**

Parameter Sensitivity

To study the influence of the parameter in our model, we evaluate the impact of the size of the sliding time window in the proposed model. We compare four lengths of sliding time windows, whose performances are shown in Fig. 5. It can be found that the size $1s$ for depression detection is the best parameter. As for mood detection, the performance reaches the peak at $1s$. And $2s$ is the best size for detecting mood.

Discussion and Conclusion

The practical adoption of AI-based healthcare applications in real-life scenarios suffers two main barriers: How to fuse multi-model data to reduce the uncertainty in the complex environment and how to explain the outputs of the model. The proposed model attempts to solve these challenges in the task of mental health detection. First, it can improve the performance of the mental health detection task compared to the deep learning and conventional machine learning approaches. Second, the explanation can be generated according to the adaptive fusion of temporal and spatial features.

By visualizing the attention weights, we can clearly see which sensors contribute more to the detection of different emotions. The correlation between the contribution of these sensors and emotions will have a more important impact on the detection of emotions in the real world. In order to detect human behavior and emotions more accurately, considering the weights of different sensors for different detection objects will be a feasible way to further improve detection accuracy and efficiency. For example, the results of attention weights remind us that in detecting mood, appropriately increasing the weight of wifi and content may improve prediction performance and accuracy. Our model explains how emotions are detected, which can not only help to adjust the weights of model inputs in the next step but also show users so that users

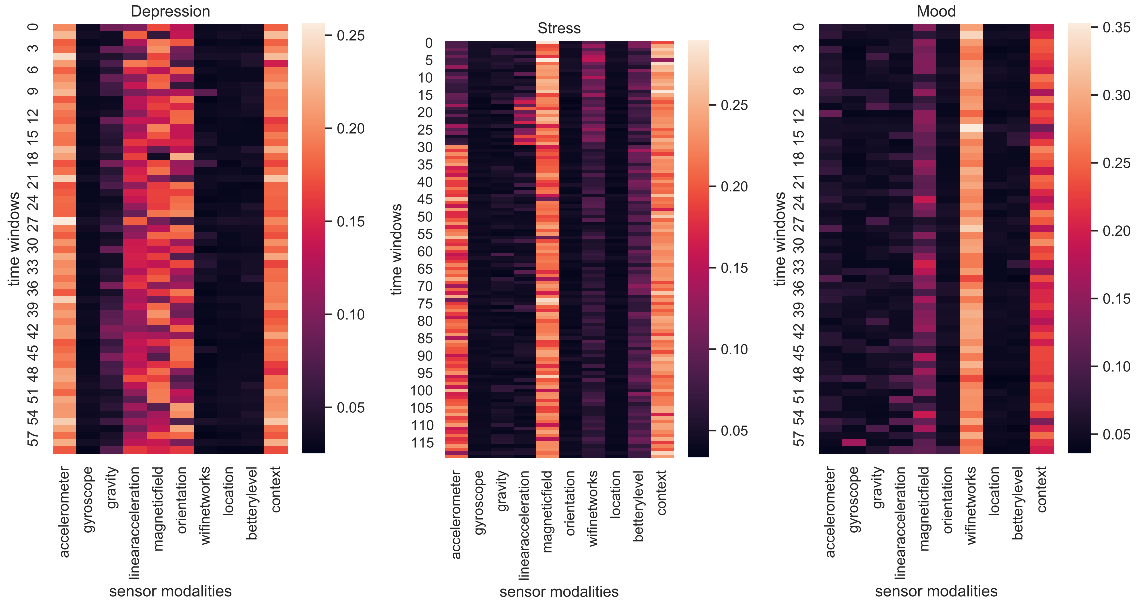

Figure 3: Visualization of attention weights for depression, stress, and mood

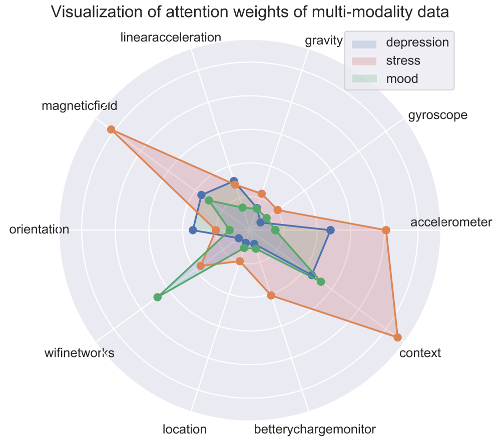

Figure 4: Visualization of multi-modality weights for depression, stress, and mood

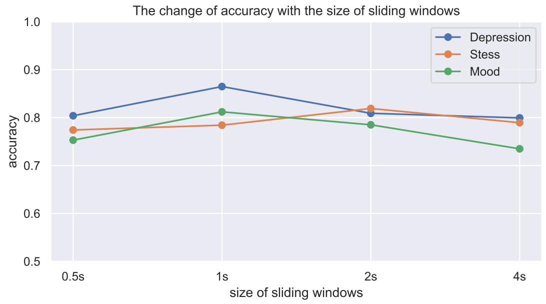

Figure 5: Performances of different parameters for depression, stress, and mood detection

Artificial intelligence that can be understood and explained put more trust and better promotes the application of artificial intelligence in the field of health.

However, one limitation in the model is the heterogeneity of the multiple modality sensory data in real life. Specifically, individuals have their behavioral patterns and personalities, and different individuals can have various behavior features sensed by smartphones when they are in different mental health states. Therefore, an adaptive machine learning model is significantly needed to accurately detect the cross-individual discrepancy. We intuitively argue that strategies such as transfer learning can improve the model, which is our next-step work. In addition, the use of unstructured data to predict mental health indicators has certain limitations. Specifically, it is necessary to take further measures in combination with people's overall conditions under the premise of considering privacy and ethics.

This work is an initial study on the explainable mental health detection model and its application. In the future, a large-scale experiment on data collection will be conducted to support further studies as follows:(a) Model Personalization. The behavioral and biological differences across individuals introduce the heterogeneous data for the machine learning model, which is ill-suited to predicting outcomes. The inability to account for individual discrepancies is necessary for accurate models and personal applications. (b) Continue to explore the explanation of the machine learning model deeply and conduct more empirical experiments on the explanation, such as correlation analysis. (c) More application functions will be designed and implemented to improve the user experience in mental healthcare.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No. 62077027), the Ministry of Science and Technology of the People's Republic of China(No. 2018YFC2002500), the Jilin Province Development and Reform Commission, China (No. 2019C053-1), the Education Department of Jilin Province, China (No. JJKH20200993K), the Department of Science and Technology of Jilin Province, China (No. 20200801002GH), and the European Union's Horizon 2020 FET Proactive project "WeNet-The Internet of us"(No. 823783).

References

Adadi, A., & Berrada, M. (2020). Explainable ai for healthcare: from black box to interpretable models. In *Embedded systems and artificial intelligence: Proceedings of esai 2019, fez, morocco* (pp. 327–337).

Ahmed, I., Jeon, G., & Piccialli, F. (2021). A deep learning-based smart healthcare system for patient's discomfort detection at the edge of internet of things. *IEEE Internet of Things Journal*, *PP*(99).

Almaslukh, B., AlMuhtadi, J., & Artoli, A. (2017). An effective deep autoencoder approach for online smartphone-based human activity recognition. *Int. J. Comput. Sci. Netw. Secur*, *17*(4), 160–165.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... others (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, *58*, 82–115.

Bishop, C. M. (2006). Pattern recognition. *Machine learning*, *128*(9).

Bulling, A., Ward, J. A., & Gellersen, H. (2012, March). Multimodal recognition of reading activity in transit using body-worn sensors. *ACM Trans. Appl. Percept.*, *9*(1).

Canzian, L., & Musolesi, M. (2015). Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *the 2015 acm international joint conference.*

Costa, A., Rincon, J. A., Carrascosa, C., Julian, V., & Novais, P. (2019). Emotions detection on an ambient intelligent system using wearable devices. *Future Generation Computer Systems*, *92*, 479–489.

Farhan, A. A., Yue, C., Morillo, R., Ware, S., Lu, J., Bi, J., ... Wang, B. (2016). Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. In *2016 ieee wireless health (wh)* (p. 1-8). doi: 10.1109/WH.2016.7764553

Figo, D., Diniz, P. C., Ferreira, D. R., & Cardoso, J. M. (2010). Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing*, *14*(7), 645–662.

Giunchiglia, F., Bignotti, E., & Zeni, M. (2017). Personal context modelling and annotation. In *2017 ieee international conference on pervasive computing and communications workshops (percom workshops)* (pp. 117–122).

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1) (No. 2). MIT press Cambridge.

Gravatt, A. E., Lindzey, G., & Aronson, F. (2013). The handbook of social psychology. *Mental Health*, *6*(2), 86-86.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), 18–28.

Hemminki, S., Nurmi, P., & Tarkoma, S. (2013). Accelerometer-based transportation mode detection on smartphones. In *Proceedings of the 11th acm conference on embedded networked sensor systems* (pp. 1–14).

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923.*

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).

Kapoor, A., & Picard, R. W. (2005). Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual acm international conference on multimedia* (pp. 677–682).

Khedkar, S., Subramanian, V., Shinde, G., & Gandhi, P. (2019). Explainable ai in healthcare. In *Healthcare (april 8, 2019). 2nd international conference on advances in science & technology (icast).*

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Li, B., & Sano, A. (2020). Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(2), 1–26.

Li, B., Yu, H., & Sano, A. (2019). Toward end-to-end prediction of future wellbeing using deep sensor representation learning. In *2019 8th international conference on affective computing and intelligent interaction workshops and demos (aciiw)* (pp. 253–257).

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, *2*(3), 18–22.

Liu, W., Zheng, W.-L., & Lu, B.-L. (2016). Multimodal emotion recognition using multimodal deep learning. *arXiv preprint arXiv:1602.08225.*

Lu, J., Shang, C., Yue, C., Morillo, R., Ware, S., Kamath, J., ... Bi, J. (2018). Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), 1–21.

Matthews, M., Abdullah, S., Gay, G., & Choudhury, T. (2014). Tracking mental well-being: Balancing rich sensing and patient needs. *Computer*, *47*(4), 36-43.

Min, S., Alquaddoomi, F., Hsieh, C. K., Rabbi, M., Yang, L., Pollak, J. P., … Choudhury, T. (2016). Leveraging multimodal sensing for mobile health: A case review in chronic pain. *IEEE Journal of Selected Topics in Signal Processing*, *10*(5), 962-974.

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology*, *13*, 23–47.

Pawar, U., O'Shea, D., Rea, S., & O'Reilly, R. (2020). Explainable ai in healthcare. In *2020 international conference on cyber situational awareness, data analytics and assessment (cybersa)* (pp. 1–2).

Radu, V., Tong, C., Bhattacharya, S., Lane, N. D., Mascolo, C., Marina, M. K., & Kawsar, F. (2018). Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(4), 1–27.

Ramachandram, D., & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, *34*(6), 96–108.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).

Riccardo, G., Anna, M., Salvatore, R., Franco, T., Fosca, G., & Dino, P. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1-42.

Saeed, A., Ozcelebi, T., Trajanovski, S., & Lukkien, J. (2018). Learning behavioral context recognition with multi-stream temporal convolutional networks. *arXiv preprint arXiv:1808.08766*.

Sharif Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 806–813).

Suhara, Y., Xu, Y., & Pentland, A. (2017). Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *Proceedings of the 26th international conference on world wide web* (pp. 715–724).

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association.*

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193-210.

Thakur, S. S. (2020). Predicting mental health using smartphone usage and sensor data. *Journal of Ambient Intelligence and Humanized Computing*.

Tsai, C. W., Lai, C., Chiang, M. C., & Yang, L. T. (2014). Data mining for internet of things: A survey. *Communications Surveys Tutorials IEEE*, *16*(1 Partsupplement), 77-97.

Tsai, C.-W., Lai, C.-F., Chiang, M.-C., & Yang, L. T. (2013). Data mining for internet of things: A survey. *IEEE Communications Surveys & Tutorials*, *16*(1), 77–97.

Tsai, C.-W., Lai, C.-F., & Vasilakos, A. V. (2014). Future internet of things: open issues and challenges. *Wireless Networks*, *20*(8), 2201–2217.

Tsai, C.-W., Tsai, P.-W., Chiang, M.-C., & Yang, C.-S. (2018). Data analytics for internet of things: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(5), e1261.

Tuli, S., Basumatary, N., Gill, S. S., Kahani, M., Arya, R. C., Wander, G. S., & Buyya, R. (2020). Healthfog: An ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated iot and fog computing environments. *Future Generation Computer Systems*, *104*, 187–200.

Vaizman, Y., Weibel, N., & Lanckriet, G. (2018). Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(4), 1–22.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wahle, F., Kowatsch, T., Fleisch, E., Rufer, M., Weidt, S., et al. (2016). Mobile sensing and support for people with depression: a pilot trial in the wild. *JMIR mHealth and uHealth*, *4*(3), e5960.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–15).

Wang, R., Aung, M. S., Abdullah, S., Brian, R., Campbell, A. T., Choudhury, T., … others (2016). Crosscheck: toward passive sensing and detection of mental health changes in people with schizophrenia. In *Proceedings of the 2016 acm international joint conference on pervasive and ubiquitous computing* (pp. 886–897).

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., … Campbell, A. T. (2014a). Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 acm international joint conference on pervasive and ubiquitous computing* (pp. 3–14).

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., … Campbell, A. T. (2014b). Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 acm international joint conference on pervasive and ubiquitous computing* (p. 3–14). New York, NY, USA: Association for Computing Machinery.

Wang, R., Wang, W., DaSilva, A., Huckins, J. F., Kelley, W. M., Heatherton, T. F., & Campbell, A. T. (2018). Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), 1–26.

Woodward, K., Kanjo, E., Brown, D. J., McGinnity, T. M., Inkster, B., Macintyre, D. J., & Tsanas, A. (2020). Beyond mobile apps: a survey of technologies for mental wellbeing. *IEEE Transactions on Affective Computing*, *13*(3), 1216–1235.

Wu, H., Liu, J., Zhu, X., Wang, M., & Zha, Z.-J. (2020). Multi-scale spatial-temporal integration convolutional tube for human action recognition.

Xu, G., Li, W., & Liu, J. (2020). A social emotion classification approach using multi-model fusion. *Future Generation Computer Systems*, *102*, 347–356.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., … Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048–2057).

Xu, X., Chikersal, P., Doryab, A., Villalba, D. K., Dutcher, J. M., Tumminia, M. J., … Dey, A. K. (2019, September). Leveraging routine behavior and contextually-filtered features for depression detection among college students. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, *3*(3).

Yao, S., Hu, S., Zhao, Y., Zhang, A., & Abdelzaher, T. (2017). Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th international conference on world wide web* (pp. 351–360).

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

Zeni, M., Zaihrayeu, I., & Giunchiglia, F. (2014). Multi-device activity logging. In *Proceedings of the 2014 acm international joint conference on pervasive and ubiquitous computing: Adjunct publication* (pp. 299–302).