Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Research paper

Predicting depression and anxiety of Chinese population during COVID-19 in psychological evaluation data by XGBoost

Zhanxiao Tian a, Wei Qu a, Yanli Zhao a, Xiaolin Zhu a, Zhiren Wang a, Yunlong Tan a, Ronghuan Jiang b, Shuping Tan a, *

a Peking University HuiLongGuan Clinical Medical School, Beijing HuiLongGuan Hospital, Beijing 100096, China
b The First Medical Center of Chinese People's Liberation Army General Hospital, No.100 West Fourth Ring Road, Fengtai District, Beijing 100853, China

ARTICLE INFO
Keywords:
Machine learning
Depression
Anxiety
Resilience
Social support
COVID-19 pandemic

ABSTRACT

Background: Due to the onset of sudden stress, COVID-19 has greatly impacted the incidence of depression and anxiety. However, challenges still exist in identifying high-risk groups for depression and anxiety during COVID-19. Studies have identified how resilience and social support can be employed as effective predictors of depression and anxiety. This study aims to select the best combination of variables from measures of resilience, social support, and alexithymia for predicting depression and anxiety.

Methods: The eXtreme Gradient Boosting (XGBoost 1) model was applied to a dataset including data on 29,841 participants that was collected during the COVID-19 pandemic. Discriminant analyses on groups of participants with depression (DE 2), anxiety (AN 3), comorbid depression and anxiety (DA 4), and healthy controls (HC 5), were performed. All variables were selected according to their importance for classification. Further, analyses were performed with selected features to determine the best variable combination.

Results: The mean accuracies achieved by three classification tasks, DE vs HC, AN vs HC, and DA vs HC, were 0.78, 0.77, and 0.89. Further, the combination of 19 selected features almost exhibited the same performance as all 56 variables (accuracies = 0.75, 0.75, and 0.86).

Conclusions: Resilience, social support, and some demographic data can accurately distinguish DE, AN, and DA from HC. The results can be used to inform screening practices for depression and anxiety. Additionally, the model performance of a limited scale including only 19 features indicates that using a simplified scale is feasible.

1. Introduction

Since its outbreak, COVID-19 rapidly became a pandemic (Wang et al., 2020a). Several factors including demographic characteristics (e.g., gender, occupation, education level, health status) and those related to COVID-19 (e.g., physical symptoms, contact history, worry level, and preventive measures) significantly impacted people's mental health, which, in some cases, further developed into psychiatric disorders (Banerjee and Rai, 2020; Minihan et al., 2020; Wang et al., 2020c; de Figueiredo et al., 2021), such as depression, anxiety, insomnia, and post-traumatic stress symptoms (Bao et al., 2020; Huang and Zhao, 2020; Luo et al., 2020; Shader, 2020; Li et al., 2022). Typically, diagnoses for depression and anxiety depend on the clinical evaluation of symptoms, as well as scales, such as the Patient Health Questionnaire-9 (PHQ-9) and Generalized Anxiety Disorder-7 (GAD-7). However, medical resource shortages during the COVID-19 pandemic made it increasingly challenging to identify these psychiatric disorders and intervene (Emmanuel et al., 2020). This necessitated the development of psychiatric screening tools with minimal demand on the already limited resources of clinical staff. Although the aforementioned measures are readily accessible, they only offer short-term evaluations based on patients' subjective experiences, which may only detect the recent abnormal (last

* Corresponding author.
E-mail address: shupingtan@126.com (S. Tan).
1 The eXtreme Gradient Boosting.
2 Depression.
3 Anxiety.
4 Depression and anxiety comorbidity.
5 Health control.

https://doi.org/10.1016/j.jad.2022.11.044
Received 16 March 2022; Received in revised form 27 October 2022; Accepted 18 November 2022
Available online 30 November 2022
0165-0327/© 2022 Elsevier B.V. All rights reserved.
two weeks) psychological fluctuations of such patients (Garabiles et al., 2020). Therefore, it is difficult for PHQ-9 and GAD-7 to effectively describe the risk of depression or anxiety. We hope to use some indicators that can describe the risk of depression or anxiety to predict depression and anxiety, so as to quantify the probability of depression and anxiety. In addition, because there are many risk factors related to depression and anxiety, it is difficult for participants to complete if all the factors are included. It may ultimately affect the prediction results. Thus, we hope to find some stable key variables to simplify the whole process without affecting the prediction effect.

The incidence of depression and anxiety, especially during COVID-19, were affected by many factors, such as knowledge and concerns related to COVID-19 (Tee et al., 2020; Wang et al., 2021a), more physical symptoms (Wang et al., 2021b), facemask use (Wang et al., 2020b), loss of confidence in doctors (Wang et al., 2021a), and number of children in the family (Li et al., 2020). In addition, the degree of enthusiasm about the government’s response was also very important. A meta analysis showed that reduction in the prevalence of depression was significantly related to a rapid and strict response from the government (Lee et al., 2021). We plan to build upon the collective and uncertain factors like government response by examining more individual and stable indicators. Psychological resilience is an individual’s ability to recover from negative experiences and flexibly adapt to the changing external environment (Werner, 1995) or withstand a high level of destructive changes without being significantly influenced (Lazarus, 1993). This is regarded as a dynamic mechanism for mitigating the impact of adverse events (Tuasie and Dyer, 2004). Many studies have demonstrated that resilience is an essential factor affecting depression and anxiety (Kanako et al., 2018; Morete et al., 2018). For example, resilience can mitigate the adverse effects of stress (Garmezy and Masten, 1986; Sheerin et al., 2018), regulate depressive symptoms caused by personality characteristics and family dysfunctions (Chang et al., 2019; Gong et al., 2019), and help reduce the risk of depression for individuals with negative childhood experiences (A et al., 2017). Moreover, adolescents with low resilience levels are at high risk of lifelong use of antidepressants and anxiolytics (Ayako et al., 2015). Therefore, evaluating an individual’s resilience could help predict mental health outcomes.

Another significant factor affecting the incidence of depression and anxiety is social support, which generally includes objective support, subjective experience, and the utilization of social support. Presently, many reports have confirmed the relationship between social support and psychiatric disorders (Rothon et al., 2012; Koelmel et al., 2016; Tomás et al., 2016; Cao et al., 2018). For example, patients with depression generally have an abnormal social support system (Nsr et al., 2020). The lack of social support exerts an adverse effect on depression by serving as a stressor (Li et al., 2017). Improvements in depressive symptoms positively correlate with improved utilization of social support (Gariepy et al., 2016). Therefore, these studies revealed that social support could be an important factor for depression.

Alexithymic patients are unable to properly describe their emotional experience, and lack fantasy and practical thinking (Hogev and Grafman, 2021). Alexithymia is positively related to the severity of mental symptoms (Mcgillivray et al., 2017). Specifically, one study found that there was an indirect relation between alexithymia and affective disorder symptoms with emotion regulation as the intermediate variable (Preece et al., 2022). This means that alexithymia, as a relatively stable risk factor, may make an individual more prone to affective disorder by influencing emotional regulation. Thus, alexithymia can also be employed as a good predictor of depression and anxiety.

The current study aimed to predict depression and anxiety using psychological resilience, social support, and alexithymia as predictors, and also select some key predictors for simplifying the whole process.

2. Methods

2.1. Participants

Participants were from different occupations in different provinces of China and were recruited online during the early stages of the COVID-19 pandemic (between February 2020 and May 2020). Participants were required to provide their personal information and complete the Connor-Davidson resilience scale (CD-RISC), Social Support Questionnaire (SSQ), Diagnostic Criteria for Psychosomatic Research (DCPR), followed by the PHQ-9 and GAD-7 evaluations. Participants were informed of the purpose and significance of the study and signed informed consent before undergoing any assessment. Data quality was controlled by employing the following rules: 1) each participant must have a unique IP address; 2) all the items were accomplished. A total of 31,017 participants were registered in our online evaluation system. Of these, 1,176 failed to pass the quality control, so 29,841 were included in the final analysis.

2.2. Measures

This was a cross-sectional study. The online evaluation consisted of two parts: general information and psychological evaluation. The general information part consisted of demographic data (e.g., age, gender, education level, marital status, occupation) and information relating to COVID-19 included variables, such as contact with COVID-19 patients, worry about COVID-19, and general health status. Psychological assessments included CD-RISC, SSQ, DCPR, PHQ-9, and GAD-7.

2.3. Assessment of psychological resilience

The Chinese version of the CD-RISC (Kathryn et al., 2003) is a self-report measure employed to measure personal psychological resilience within the past 30 days. The CD-RISC consisted of 25 items with the options for each item rated from 0 to 4 (not at all (0), rarely (1), sometimes (2), often (3), and almost always (4)). The scale contains items measuring three factors: 1) tenacity (11–23), 2) strength (1, 5, 7, 8, 9, 10, 24, and 25), and 3) optimism (2, 3, 4, and 6). A high score represents high level of psychological resilience. The CD-RISC has demonstrated significant reliability and validity within different populations (Windle et al., 2011; Ye et al., 2017).

2.4. Assessment of social support

The SSQ (Sarason et al., 1983) is a self-report measure for evaluating the level of individual social support with a high score corresponding with a high social support level. The consistency of the total score, when retested using Chinese college students, was 0.92 ($p < 0.01$); each item was between 0.89 and 0.94, which corresponded to good reliability and validity.

2.5. Assessment of alexithymia

The DCPR (Porcelli and Sonino, 2007) is a simple, effective, and reliable regular interview tool, which was developed by an international psychosomatic research group and can be employed to screen and diagnose psychosomatic and psychophysiological disorders. In the revised DCPR, a minimum of three items were considered alexithymia from the following six items: 1) inability to utilize appropriate emotions, 2) tendency to describe details rather than feelings, 3) lack of an interesting life, 4) exhibiting thought patterns that are more related to external events than fantasies or emotions, (5) being unaware of the relationship between common physical reactions and the various emotional experiences, and (6) displaying occasionally violent and often inappropriate emotional behaviors.
2.6. Assessment of anxiety and depression

The PHQ-9 (Wang et al., 2014) is a self-report measure for identifying whether individuals are suffering from depression. Scores correspond with normal (0–4), mild (5–9), moderate (10–14), moderate to severe (15–19), and severe (20–27) depression. In the current study, a score of 4 was used as the boundary between healthy control and depression. The PHQ-9 exhibited strong reliability and validity in Chinese individuals (the internal consistency was 0.86). A recent online evaluation via smartphones, in addition to a paper evaluation, obtained similar results (Zhen et al., 2020). The GAD-7 (He et al., 2010) is a self-report measure for identifying whether subjects suffer from anxiety. Scores correspond with normal (0–4), mild (5–9), moderate (10–14), and severe (≥15) levels. The current study designated 4 as the boundary between healthy control and anxiety. The retest reliability for the Chinese version of the GAD-7 was 0.85.

2.7. Descriptive and data analysis

Based on results of the PHQ-9 and GAD-7, participants were labeled DE (depression, PHQ-9 ≥ 5 & GAD-7 ≤ 4), AN (anxiety, GAD-7 ≥ 5 & PHQ-9 ≤ 4), DA (depression and anxiety comorbidity, PHQ-9 ≥ 5 & GAD-7 ≥ 5), and HC (health control, PHQ-9 ≤ 4 & GAD-7 ≤ 4). The HC group was combined separately with DE, AN and DA group to form three new datasets.

We use univariate analysis to analyse the relations between features and labels. To be specific, the Mann–Whitney U test was employed to compare the continuous data of the non-normal distribution. Pearson chi-square (χ²) test was applied for categorical and dichotomous variables. Two-tailed test of significance used: *p < 0.01.

2.8. Machine learning model

After comparing several models based on performance, including the support vector machine, random forest, logistic regression, and Xgboost; Xgboost was selected as our classifier. Xgboost was first proposed by Tianqi Chen (Chen and Guestrin, 2016). It is a widely recognized and efficient machine learning technique, which assembles weak prediction models through continuous feature splitting, as well as the addition of new trees, to generate a more accurate model. It is also an open-source package.

2.9. Data preprocessing

Each dataset was split into the training and testing sets in an 8:2 ratio. Further, a 10-fold cross-validation was conducted within the training set to optimize the algorithm. The holdout testing set was only employed to measure the performance of the model.

2.10. Predictions and evaluation

The area under the curve (AUC) was employed as a primary indicator to evaluate the model. An AUC of 0.8–0.9 is generally considered to be good, while an AUC of >0.9 is considered excellent (Hosmer and Lemeshow, 2000). Other performance indicators, including the overall accuracy, precision, recall, and F1 score, were also employed.

2.11. Feature selection

“Gain” is a built-in method of the Xgboost model, employed to determine the significance of selected features during prediction. The F-score represents the degree of feature significance (the higher the F-score, the more significant the feature). Afterward, feature selection was performed based on the significance of the feature, as well as its predicted performance. Feature combination was also performed to fit three groups of people according to the feature selection result.

3. Results

3.1. Descriptive and data analysis of demographic characteristics, COVID-19 related factors, current health status, and psychological factors

The results of the descriptive analysis of the 29,841 participants (male: 10,592, female: 19,249) is presented in Table 1.

The demographic data reveals that older, higher education level and divorced women are more likely to suffer from depression and anxiety (p < 0.01). As for the factors related to the COVID-19 epidemic, having patients with infection (including family members, friends or colleagues) around them, having COVID-19 contact history or being infected are tend to depression and anxiety (p < 0.01). In terms of general health status, the worse the general health status, the more prone to depression and anxiety (p < 0.01). In psychological assessment, social support and resilience are protective factors of depression and anxiety (p < 0.01). The higher the score of SSQ or CD-RISC, the less likely to suffer from depression and anxiety.

The results of the descriptive analysis showed all factors are related to depression and anxiety. However, it is hard to explain the impact of these variables on depression and anxiety, so it is necessary to further quantify the predictive effect of these variables on depression and anxiety with machine learning models.

3.2. Predictive performance

Among the four groups in the discriminant analysis, the prediction of DA was the most accurate, followed by DE and AN. The AUC of the three prediction tasks were ≥0.85, indicating that the model exhibited high stability and reliability for the three tasks. The predictive performance of the model for the three tasks is summarized in Table 2, and their receiver operating characteristics (ROC) are shown in Fig. 1.

3.3. Feature selection and feature importance

3.3.1. Feature selection

The feature screening results for DE, AN and DA is summarized in Fig. 2. The red label indicates that the accuracy of the model was relatively high when enough features were selected. (DE: feature numbers = 22, accuracy = 0.76; AN: feature numbers = 19, accuracy = 0.77; DA: feature numbers = 28, accuracy = 0.88.)

3.3.2. Feature importance

The top features of DE, AN and DA are plotted in Fig. 3. Fig. 3(a) shows the top 22 features of DE. The inclusion of five demographic characteristics accounted for 22.7 % of all the 22 features. Therefore, “AGE,” “WorkingPlace,” “EducationLevel,” and “WorryofnewCvirus” accounted for the top four features among the 22; eleven CD-RISC items (50.0 %) were included, and Item 3 (“Sometimes fate or god can help”) scored the highest; six SSQ items (27.3 %) were included, and Item 6 (“Whether there is someone to talk when encountering troubles in the past year”) was the most significant; and no DCPR items were included in the top 22 features.

Fig. 3(b) shows the top 19 features of AN. The inclusion of six demographic characteristics accounted for 31.6 % of all the 19 features. “AGE,” “WorkingPlace,” and “EducationLevel” accounted for the top three features among the 19; eight CD-RISC items (42.1 %) were included, and “Sometimes fate or god can help” still scored the highest; five SSQ items (26.3 %) were included, and “Whether there is someone to talk when encountering troubles in the past year” was the most significant. DCPR did not account for any item in the 19 features.

Fig. 3(c) shows the top 28 features of DA. The inclusion of five demographic characteristics accounted for 17.9 % of all the 28 features; “AGE” was still the most significant feature among the 28; 14 CD-RISC items (50.0 %) were included, and Item 2 (“I have close and safe relationship”) scored the highest; seven SSQ items (25.0 %) were included,
Table 1

| Sample characteristics based on DE, AN, DA, and HC (n = 29,841). |
|-----------------------------------------------|
| DE (n = 4632) | AN (n = 1219) | DA (n = 8055) | HC (n = 15,993) |
|-----------------------------------------------|
| **Sex:** N (%) | | | |
| Female | 2714 (58.6) | 765 (62.8) | 5259 (65.3) | 6268 (65.3) |
| Male | 1918 (41.4) | 454 (37.2) | 2896 (34.7) | 2627 (34.7) |
| **Age:** Mean (SD) | | | |
| 30.9 ± 7.6 | 32.9 ± 7.6 | 31.0 ± 7.6 | 30.0 ± 7.6 |
| **Education level:** Mean (SD) | | | |
| 15.7 ± 2.5 | 16.4 ± 2.5 | 15.9 ± 2.5 | 15.5 ± 2.5 |
| **Marital status:** N (%) | | | |
| Married | 221 (4.8) | 756 (62.0) | 4201 (52.2) | 7693 (52.2) |
| Divorce | 107 (2.3) | 27 (2.2) | 238 (3.0) | 204 (3.0) |
| Cohabitation | 38 (0.8) | 12 (1.0) | 85 (1.0) | 49 (0.3) |
| Single | 2231 (48.2) | 409 (33.6) | 3408 (42.3) | 7855 (49.3) |
| Others | 44 (0.9) | 15 (1.2) | 123 (1.5) | 134 (0.8) |
| **COVID-19 exposure:** N (%) | | | |
| FFI | 871 (18.8) | 234 (19.2) | 1526 (19.0) | 3656 (22.9) |
| CNI | 1785 (38.5) | 426 (34.9) | 3259 (40.5) | 6522 (40.9) |
| CH | 33 (0.7) | 11 (0.9) | 36 (0.4) | 53 (0.3) |
| SC | 112 (2.4) | 31 (2.5) | 220 (2.7) | 145 (0.9) |
| NCH | 1814 (39.2) | 513 (42.1) | 2993 (37.2) | 5541 (34.8) |
| Patient | 17 (0.4) | 4 (0.3) | 21 (0.2) | 18 (0.1) |
| **Worried status:** N (%) | | | |
| Not at all | 649 (14.0) | 114 (9.4) | 729 (9.1) | 3214 (20.2) |
| A little | 2165 (46.7) | 489 (40.1) | 3281 (40.7) | 7209 (45.2) |
| Some | 456 (9.8) | 148 (12.1) | 971 (12.1) | 1224 (7.7) |
| Worry | 843 (18.2) | 268 (22.0) | 1602 (19.9) | 2688 (16.9) |
| Very much | 519 (11.2) | 200 (16.4) | 1472 (18.3) | 1600 (9.9) |
| **Current health status:** N (%) | | | |
| Good | 3886 (83.4) | 1036 (85.0) | 5532 (68.7) | 15,327 (96.2) |
| Ok | 693 (15.0) | 172 (14.1) | 2200 (27.3) | 579 (3.6) |
| Not very well | 45 (1.0) | 11 (0.9) | 277 (3.4) | 23 (0.1) |
| Bad | 8 (0.2) | 0 (0.0) | 46 (0.6) | 6 (0.0) |
| **Social support:** Mean (SD) | | | |
| Objective support | 8.2 (2.9) | 9.1 (3.0) | 7.4 (2.8) | 10.4 (3.3) |
| Subjective support | 21.8 | 23.5 | 20.1 | 26.1 (4.7) |
| Used of support | 7.8 (1.9) | 8.2 (1.9) | 7.1 (1.9) | 9.4 (2.0) |
| Total support | 37.8 | 40.8 | 34.7 | 45.9 (8.0) |
| Psychological resilience: Mean (SD) | | | |
| Toughness | 30.4 | 32.1 | 26.4 | 37.7 (8.1) |
| Strength | 22.1 | 22.9 | 19.2 | 26.2 (4.5) |
| Optimism | 9.5 (2.5) | 9.8 (2.5) | 8.3 (2.6) | 10.7 (2.5) |
| Total resilience | 62.0 | 64.8 | 53.9 | 74.6 (13.6) |
| Alexithymia: Mean (SD) | | | |
| 2.2 (1.1) | 2.2 (1.1) | 2.2 (1.1) | 2.3 (0.9) |

DE: depression; AN: anxiety; DA: depression and anxiety comorbidity; HC: health control; FFI: family or friends infected; CNI: colleagues or neighbors infected; CH: has contact history; SC: suspected case; NCH: does not have contact history.

* Significant difference between DE and HC.

** Significant difference between AN and HC.

*** Significant difference between DA and HC.

p < 0.01.

Table 2

| Performance of the three groups based on the Xgboost model. |
|-----------------------------------------------|
| Groups | Accuracy | Precision | Recall | F1-score | AUC |
|-----------------------------------------------|
| DE* vs HC* | 0.78 | 0.77 | 0.80 | 0.78 | 0.86 |
| AN* vs HC | 0.77 | 0.78 | 0.80 | 0.79 | 0.85 |
| DA* vs HC | 0.89 | 0.89 | 0.89 | 0.89 | 0.95 |

DE*: depression group (n = 4632); AN*: anxiety group (n = 1219); DA*: depression and anxiety comorbidity group (n = 8055).

and Item 7 ("Whether there is someone for help when encountering trouble in the past year") ranked 2nd among the 28 features. Two DCPR items: Item 6 ("Have you ever had an occasional but violent outburst of anger, crying or joy, which is incompatible with the relationship between the events at that time or your normal behavior") and Item3 ("Do you often fantasize") (7.1 %) were included, and they ranked as the last two among the 28 features.

3.4. Feature combination

After the first feature-selection step, twenty overlapping features were observed among the DE, AN, and DA groups. We built a combination of 19 selected features, including only the overlapping variables, but excluding the features with a correlation coefficient of >0.8 (SSQ7).

The Xgboost model based on 19 selected features achieved a good predictive power (accuracy: DE = 0.75; AN = 0.75; DA = 0.86). A description of the definition for the 19 selected features can be found in Online supplementary Table S1.

4. Discussion

Based on the Xgboost model, this study explored the use of the CD-RISC, SSQ, DCPR, and demographic characteristics for predicting DE, AN and DA and achieved good model performances. Among the three groups of classification tasks, the DA group performed significantly better than the other two. On the one hand, there was a certain gap between the data volumes of the three groups. Thus, we down-sampled the data of the normal control group according to the sample size of the patient group to avoid the bias caused by unbalanced data. However, since downsampling is actually just reducing the overall sample size, the small sample size might result in underfitting, followed by reduced accuracy. On the other hand, DA combines the depression and anxiety characteristics, so this makes their recognition easier compared to normal controls.

Following feature selection, it was easy for the participants to accept the first 19 features, and the performance of such a combination was only slightly lower than those of 28 and 36 features, indicating that these features were key features for predicting DE, AN, and DA. As a result, our study provides a simplified instrument to screen depression and anxiety disorders, which may improve the efficiency of clinical evaluation.

"AGE" was the most significant feature during feature selection. The demographic data show that the age distribution of the patient groups were higher than those of the HC, indicating that older people might have higher risks of depression, anxiety, or comorbidity during the pandemic than younger people, possibly due to the cognitive control deficits (Dotson et al., 2020). A recent study demonstrated that young age was related to reduced depressive symptoms (Alono Debreceni and Bailey, 2021), which aligns with the findings reported here. Additionally, "WorkingPlace," "EducationLevel," "WorryofnewCvirus," and "Contactwithvirus" accounted for the overlapping features, and they all ranked high among the screened features of the three groups. The demographic data revealed that there were significant differences between these four variables among the groups of DE, AN, DA and HC (p < 0.01).

Studies have demonstrated that the incidence of depression and anxiety increased following the COVID-19 pandemic (Choi et al., 2020; Luo et al., 2020). In our study, "WorkingPlace", "WorryofnewCvirus" and...
“Contact with virus” were associated with the COVID-19 pandemic; which corresponds with findings from previous studies. Conversely, one study showed that COVID-19 patients did not show significant depression and anxiety tendencies compared with psychiatric patients and healthy controls (Hao et al., 2020). A meta-analysis on the association between patients with mood disorders and COVID-19 outcomes, found that mood disorders were a high-risk factor for COVID-19 infection and death (Ceban et al., 2021a). Although our study shows that there is a significant difference in the proportion of COVID-19 patients among the four groups ($p < 0.01$), after feature selection, the proportion of COVID-19 patients does not show a good predictive effect (i.e., it is not among the top variables), which is also consistent with previous studies to some extent. Since the proportion of COVID-19 patients in this study is <1%, it may not be enough to conclude that the proportion of COVID-19 patients will affect the prediction results. In addition, when examining the association between patients with COVID-19 symptoms that subsided within 12 weeks and the high incidence of depression, a systematic review found that the severity of acute COVID-19 does not lead to an increase in the incidence of depression (Renaud-Charest et al., 2021). However, fatigue and cognitive impairment are more likely to occur after 12 weeks of COVID-19 infection (Ceban et al., 2021b). While patients with depression have extensive cognitive impairment (Wang et al., 2022), perhaps cognitive impairment is the mediator of depression 12 weeks or more after a COVID-19 infection. Finally, the studies involving the China family panel that explored the relationship between depression and educational achievement confirmed that high educational achievement reduced the risk of depression (Shen, 2020), similar to what was observed in our research. In our study, the patient groups generally had high education levels. Additional data would be required to further explore the reason.

Regarding the CD-RISC scale, Items 2 (“I have close and safe relationship”), 3 (“Sometimes fate or god can help”) and 20 (“I had to act on my hunch”) ranked as the top three, respectively. According to the three-dimensional score method of CD-RISC (Yu et al., 2007), Items 2 and 3 represented optimism, indicating that these patient groups might be less-optimistic. A strong negative correlation has been noted between optimism and the COVID-19 pandemic (Ran et al., 2020). Item 20 belongs in the tenacity category. A recent study indicated that tenacity correlated with depression and anxiety (Ran et al., 2020). The more tenacious the participants were, the less likely they were to be affected by the pandemic.

For SSQ, Items 3 (“How is your relationship with your neighbors”), 6 (“Whether there is someone to talk when encountering troubles in the past year”), and 7 (“Whether there is someone for help when encountering troubles in the past year”) ranked as the top three. Item 3 represented the participant’s relationship with their neighbors. A previous study indicated that the neighborhood was key to reducing depressive symptoms by building a kind of social cohesion among the members of a

![ROC Curve](image-url)

**Fig. 1.** ROC curves and AUC values for the three prediction tasks.
community (Miao et al., 2019). Items 6 and 7 represent how the participants confided and asked for help when they encountered troubles, respectively. They exhibited very strong correlation ($r = 0.8$). The patient group tended to talk to and ask for help from fewer people compared with the other groups ($p < 0.01$). A systematic review revealed that the presence of confidants was identified as a factor of social relations that was significantly associated with depression (Schwarzbach et al., 2014).

The results of this study demonstrated that alexithymia was one of the etiological factors that determined generalized anxiety disorder and depression (Lenzo et al., 2020). It is generally believed that alexithymia abnormally regulates emotional processes, and this is a critical risk factor regarding the occurrence and development of mood disorders and psychosomatic diseases (Panayiotou et al., 2021). Item3 (“Do you often fantasize”) and Item 6 (“Have you ever had an occasional but violent outburst of anger, crying or joy, which is incompatible with the relationship between the events at that time or your normal behavior”) are included in the DA group, similar to previous studies (Lisha et al., 2018; Palser et al., 2018). Considering the role of emotion regulation in alexithymia and affective disorder, cognitive behavior therapy (CBT), as a kind of intervention aimed at emotion regulation can be used as a part of comprehensive treatment of alexithymia (Ho et al., 2020). Especially during the COVID-19 pandemic, online CBT has great potential (Zhang and Ho, 2017).

In conclusion, based on the CD-RISC, SSQ, and DCPR measures, as well as demographic data, an excellent performance regarding the predictions of DE, AN, and DA was obtained. The performance supported the possibility for screening potential patients of depression and anxiety online. After the feature selection, the final 19 variables almost achieved the same performance as the total 56 variables, indicating that the simplified scale was theoretically feasible and is expected to improve the efficiency of screening patients online. Similar to our study, Ren et al. (2021a) collected data on the demographic characteristics of Chinese college students, such as gender, major, and grade, in addition to variables on personal views associated with COVID-19, and predicted the impact of COVID-19 on college students' mental health using a machine learning model (logistic regression). The results showed that the accuracy of 12 variables in predicting anxiety and depression were 81 % and 74 %, respectively. The AUC value for each model was >0.8, indicating that those models are as stable and reliable as ours. There were also

![Feature selections of DE (a), AN (b) and DA (c) prediction](image-url)

Fig. 2. Feature selections of DE (a), AN (b) and DA (c) prediction (the X-axis represents the feature numbers, as screened by the “Gain” method, and the Y-axis represents accuracy, which is the relative number of features that were employed for the prediction). The red circle and square represent the ideal accuracy and related feature numbers, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
some differences between two studies. First, the sample size of Ren’s study was quite small compared to ours (478 vs 29,841), as mentioned in the study’s limitations. Second, Ren’s study included more variables only relevant to college students, such as internship status, examination scores, and school-related situations. In contrast, social support, resilience, and alexithymia can be applicable to a wider range of people, not just school students.

The current study was cross-sectional, which is insufficient for providing evidence on the causal relationship between the selected features and the incidences of depression and anxiety. In a four-week longitudinal study, the authors found that the scores of the DASS-21 subscale in patients with depression and anxiety had no statistically significant change. Factors associated with higher DASS-21 subscale scores included physical symptoms, general health status, chronic medical history (Cw et al., 2020), but did not include social support, psychological resilience, or alexithymia. Therefore, additional longitudinal studies are needed to verify the role of these features in predicting the risks of depression and anxiety. Further, extensively retaining participant’s data and synthesizing it for the machine learning model may introduce an unbalanced data distribution, which might be a potential factor affecting performance. Finally, the variables in the current study were limited. Studies also have shown that suicidal ideation and suicide attempts increased during the COVID-19 pandemic (Ec et al., 2020; Berardelli et al., 2021). Although in some countries, national prevention and control led to a decline in the suicide rate (McIntyre et al., 2021), it began to rise significantly after the lock-down (Montalbani et al., 2020). It is still unclear how suicide interacts with depression and COVID-19. Another study found that the willingness to vaccinate is related to the severity of depression and anxiety (Hao et al., 2021). In addition, some groups, such as adolescents (Ren et al., 2021b) and pregnant women (Nguyen et al., 2022), have different effects for depression and anxiety. These factors, in addition to other factors, such as post-traumatic stress symptoms, social state, and the government’s response to the mental health crisis during COVID-19 should also be measured and included in future studies examining the impact on mental health.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jad.2022.11.044.

Funding

This work was supported by the Beijing Natural Science Foundation Grant [grant number 7202072]; the Beijing Municipal Science and Technology Commission Grant [grant number Z191100006619104]; and Beijing Hospitals Authority Ascent Plan [grant number DFL20192001].

CRediT authorship contribution statement

Zhanxiao Tian: Investigation, Validation, Formal analysis, Data curation, Writing - original draft. Wei Qu: Software, Validation, Formal analysis, Data curation. Yanli Zhao: Formal analysis, Data curation. Xiaolin Zhu: Formal analysis, Data curation. Zhiren Wang: Conceptualization, Methodology, Formal analysis. Yunlong Tan: Conceptualization, Resources, Project administration, Writing - review. Ronghuan Jiang: Resources, Project administration. Shuping Tan: Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are grateful to all of the project participants and all of the individuals involved in the data collection.
