Accuracy of Supervised Machine Learning in Predicting Depression, Anxiety and Stress Using Web-based Big Data: Preserving the Humanistic Intellect

Article *in* Malaysian Journal of Medicine and Health Sciences · December 2022

DOI: 10.47836/mjmhs.18.s19.14

2 authors:

Edre Mohammad Aidid
International Islamic University Malaysia
34 PUBLICATIONS **46** CITATIONS

SEE PROFILE

Ramli Musa
International Islamic University Malaysia
103 PUBLICATIONS **1,047** CITATIONS

SEE PROFILE

ORIGINAL ARTICLE

Accuracy of Supervised Machine Learning in Predicting Depression, Anxiety and Stress Using Web-based Big Data: Preserving the Humanistic Intellect

Edre Mohammad Aidid[1], Ramli Musa[2]

[1] Department of Community Medicine, Kulliyyah of Medicine, International Islamic University Malaysia, Jalan Sultan Ahmad Shah, 25200 Kuantan, Pahang.
[2] Department of Psychiatry, Kulliyyah of Medicine, International Islamic University Malaysia, Jalan Sultan Ahmad Shah, 25200 Kuantan, Pahang.

ABSTRACT

Introduction: One of the most useful tool to assess the extent of depression, anxiety and stress symptoms is the validated Depression, Anxiety and Stress Scale, 21 items (DASS-21). The availability of online mental health resource centre provides big data capable of machine learning analytics for early detection of mental health issues. However, prediction accuracy of these data using machine learning method remains elusive. **Methods:** A cross sectional study was conducted, using secondary data of respondents who answered an online DASS-21 questionnaire from an online resource center. Depression, anxiety and stress were measured using DASS21 as either the outcome or predictor, depending on the model. The model includes sociodemographic predictors such as gender, age, race, marital status, education level and occupational status. A feed-forward artificial neural network was constructed based on multilayer perceptron machine learning procedure using IBM SPSS version 23. **Results:** A total of 339,781 respondents data were obtained. The observed prevalence of depression, anxiety and stress was 39.9%, 48.5% and 13.4%, respectively. This resulted in 76.4% prediction accuracy for depression, 76.3% accuracy for anxiety and 87.4% prediction accuracy for stress. Stress and anxiety were the most important factors contributing to the disease model. **Conclusion:** The prediction models have high accuracy to predict the true observed depression, anxiety and stress prevalence. The clinical relevance of these prediction models still needs the human intellect judgment based on Maqasid al-Shariah principles. Machine learning therefore should not be abused but to help in decision-making towards early detection and prompt treatment.
Malaysian Journal of Medicine and Health Sciences (2022) 18(19) 87-92. doi:10.47836/mjmhs.18.s19.14

Keywords: Supervised machine learning, Depression, Anxiety, Stress, Big data

Corresponding Author:
Edre Mohammad Aidid, DrPH
Email: edreaidid@iium.edu.my
Tel: +609-5704605

INTRODUCTION

Depression, anxiety and stress has become one of the leading causes of not just psychospiritual morbidity, but also physical morbidity. More and more people succumb to this chronic but treatable disease or condition. The symptoms range from mild to severe, leading to potential adverse events such as suicide which is preventable if detected early. The condition is worse during the COVID-19 pandemic, where the prevalence is not just higher but the rate of change is worrisome, where an increment of 27.6% for depression, 25.6% for anxiety was observed globally. Southeast Asian countries including Malaysia showed lesser but significant increment from pre-pandemic times to during pandemic for depression (11.5%) and anxiety (13.1%) (1). One of the tools that has been widely used to assess depression, anxiety and stress symptoms is the Depression, Anxiety and Stress Scale - 21 items (DASS-21). Locally, the prevalence of DASS-21 depression, anxiety and stress is 15%, 34% and 17%, respectively, among adolescent (2). Another study showed higher prevalence with 46% depression, 59% anxiety and 38% stress using the same tool and population (3). The aim is to recognize these symptoms early in order to begin early management that can potentially prevent serious complications such as suicide. So far, clinicians rely on passive surveillance

where patient come to the clinic to seek help for these conditions. There is a need for a larger safety net and this can be made available by having an online information and survey (4).

In the advent of Industrial Revolution (IR) 4.0, machine learning has become one of the most sought-after data analytic method due to its ability to handle big amounts of data (5). It is part of artificial intelligence (AI), an area where the decisions are made based on rigorous computation. Its main advantage over statistical methods is its ability to handle big data which is usually non-linear and complex (5). Statistical methods often needs various assumptions in order to make correct statistical decisions, in which can be daunting when dealing with big data. If assumptions are neglected, it can lead to false clinical decisions based on the results. The robustness of supervised machine learning can overcome this. Moreover, supervised machine learning is used as the data is labeled, making the analysis more structured to the clinical question. Supervised machine learning helped in feature selection of prognostic variables important in cognitive behavioral therapy impact on depression (6). However, a key component here is the human touch and interpretability in healthcare that makes a difference especially when handling health conditions such as depression, anxiety and stress (7).

Traditionally, DASS-21 was analysed as part of early identification of depression, anxiety and stress symptoms using prevalence as the measure of morbidity. This is important in stratification of patients in need for referral for further psychiatric and psychological assessment. The analysis tends to have modest number of samples and have limited generalizability, as it tends to focus on specific groups such as single mothers, students and other vulnerable groups. However, there are paucity of studies utilizing big online community data and machine learning methods which can capture abnormal trends almost real-time from an open source perspective. Furthermore, the supervised machine learning method is needed to study about human behaviours and its relation with their background characteristics in people who have the access to the online survey which might be people who are hesitant initially to seek psychiatric evaluation. The key component to this measurement of association is the accuracy of the machine learning model.

The objective of this study is to measure the accuracy of supervised machine learning in predicting the observed prevalence of depression, anxiety and stress, as well as to determine the most important predictor of depression, anxiety and stress.

MATERIALS AND METHODS

A cross sectional study was done using data from an online Mental Health Information and Research Centre (4). It is an online health promotional website, aimed to empower the public and their families on the knowledge of their psychiatric and social issues as well as the treatments available. This online centre was made publicly available and catered thousands of hits and visitors per day. The data is validated by a consultant psychiatrist.

Sample size calculation was based on artificial neural network (ANN) model for a dichotomous outcome; minimum sample size of 50 times the number of weights in the ANN (8). Weight is the parameter within a neural network that transforms input data within the network's hidden layers. Taking into account the total number of categories in the independent variable that contributes to this weight of 30, the minimum sample size required was 1500. Thus, since we already have the complete data of 339,781 respondents, we feel that all of the data should be analyzed to get a more representative picture. Sampling was done via universal sampling method; all data that were inputted in the website were taken as of 17 January 2020.

Outcomes were depression, anxiety and stress measured using DASS-21, which was answered by the general community online. The questions such as "I find it hard to wind down" was asked in written form via online platform. The raw scores were converted into two categories; having depression (10 or more), anxiety (8 or more), stress (15 or more), or not having the condition (9). Each outcome was modelled with the rest of the outcome, plus gender, age, race, marital status, education level and occupational status. A feed-forward artificial neural network was modelled using multilayer perceptron machine learning procedure using IBM SPSS version 23. The variable importance, or strength of association, of a specific independent variable for the outcome variables was determined by identifying all weighted connections between the neural network nodes of interest. On the other hand, normalized variable importance is useful when the data has variable scales and artificial neural networks does not make assumptions about the distribution of the data (10). The dataset was divided into training (70%) and testing (30%) dataset for validation.

Research and ethical approval were obtained from Kulliyyah of Medicine Research Committee, International Islamic University Malaysia (IIUM) (ID: 825) and IIUM Research Ethics Committee (IREC 2022-071), respectively. The secondary data from the online database was kept anonymous since the start as there were no personal identifiers of the users collected. Data is pulled from the website after online consent. Variables such as depressive symptom score, anxiety symptom score, stress symptom score, gender, age, race, marital status, education level and occupational status were used for the analysis. Data cleaning was done by checking for missing data. Little's missing completely at

random (MCAR) test was done on the variables using expectation maximization (EM) estimation method, which was statistically not significant. Thus, the missing data was assumed as MCAR and not included in the analysis. Cross entropy error was used as a loss function, where the lower the value the better the prediction model is by setting difference between the estimated probability with the desired outcome (10).

RESULT

A total of 339,781 respondents were included in the analysis. From the database, majority of those with depressed symptoms were female, over 75 years old, other race, divorced, no formal and unemployed. The prevalence of anxiety symptoms are higher in female, over 75 years old, other race, single, no formal and students. On the other hand, stress is more prevalent in female, over 75 years old, Indians, divorced, no formal education and unemployed. Results are shown in Table I.

Table I: Prevalence of depression, anxiety and stress according to sociodemographic characteristics

Background*		Depression		Anxiety		Stress	
		Frequency (n)	Prevalence (%)	Frequency (n)[a]	Prevalence (%)	Frequency (n)	Prevalence (%)
Gender	female	121712	42.0	147601	50.9	42290	14.6
	male	24981	32.2	30736	39.6	7077	9.1
Age	18-24	79102	43.0	96280	52.4	26383	14.4
	25-34	42483	36.9	50758	44.1	13764	12.0
	35-44	7929	24.9	11022	34.6	2377	7.5
	45-54	1027	14.2	1718	23.7	237	3.3
	55-64	175	12.0	287	19.6	51	3.5
	65-74	30	25.0	37	30.8	12	10.0
	over 75	27	57.4	31	66.0	21	44.7
	under 12	102	31.0	129	39.2	61	18.5
Rrace	Chinese	1465	36.3	1680	41.6	516	12.8
	Indian	1044	38.9	1265	47.1	454	16.9
	Malay	133164	39.9	161209	48.3	44105	13.2
	other	11001	40.9	14162	52.6	4287	15.9
Marital status	divorced/ widowed	2538	44.8	2840	50.1	967	17.1
	married	25962	27.8	34859	37.3	8875	9.5
	single	118205	44.0	140651	52.4	39530	14.7
Education level	college/ university	110048	37.8	135750	46.6	35188	12.1
	none	2487	51.8	2701	56.3	900	18.8
	primary	788	41.8	902	47.9	351	18.6
	secondary	33382	47.9	38997	56.0	12933	18.6
occupation	professional	32422	30.7	42487	40.2	10736	10.2
	semi-skilled	14122	42.4	16382	49.2	4492	13.5
	skilled	8639	34.2	10996	43.5	2938	11.6
	student	69907	43.3	86562	53.6	24048	14.9
	unemployed	21615	51.4	21923	52.1	7158	17.0

*Result shown excluded any missing data

Based on Table II, the observed prevalence of depression, anxiety and stress was 39.9%, 48.5% and 13.4%, respectively. The prediction accuracy for depression, anxiety and stress were 76.4%, 76.3% and 87.4%, respectively. The lowest cross entropy error was for the stress model, and the highest was the anxiety.

Table II: Model summary for the three prediction models

Aspect	Depression	Anxiety	Stress
Prevalence (%)	39.9	48.5	13.4
Accuracy (%)	76.4	76.3	87.4
Cross entropy error	48895.9	53049.7	26510.4
AUC	0.8	0.8	0.9
Most important predictor	Stress	Stress	Anxiety

On the other hand, stress and anxiety was the most important predictor contributing to the three disease model (normalized importance= 100%).

Fig. 1 until 3 shows the breakdown of predictors based on depression, anxiety and stress machine learning model, respectively, according to its importance. Stress and anxiety was the most important factors contributing to the disease model.

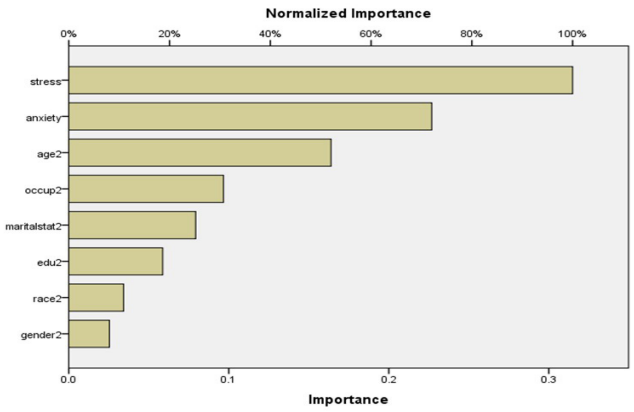

Fig. 1: Predictors of depression which are (in descending order of importance) stress, anxiety, age, occupation, marital status, education, race and gender.

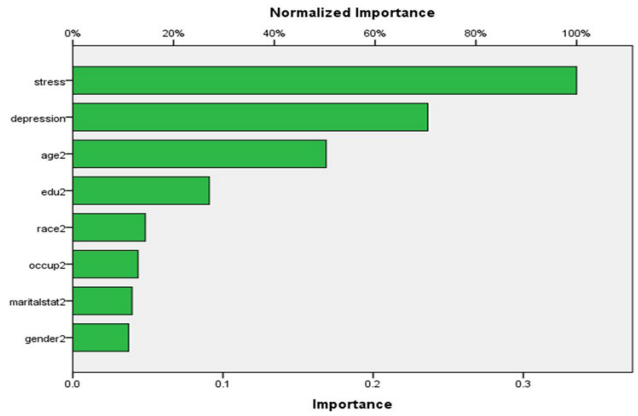

Fig. 2: Predictors of anxiety which are (in descending order of importance) stress, depression, age, education, race, occupation, marital status and gender

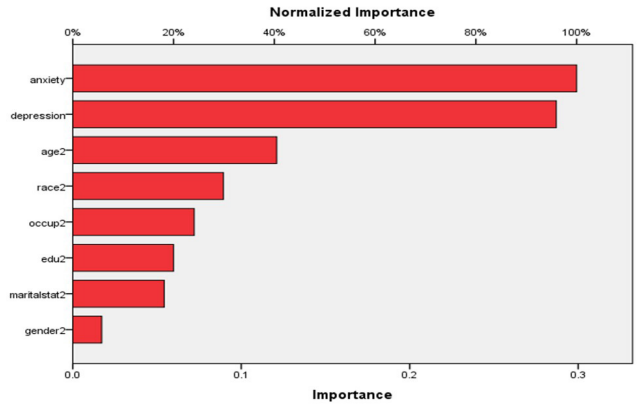

Fig. 3: Predictors of stress which are (in descending order of importance) anxiety, depression, age, race, occupation, education, marital status and gender

DISCUSSION

The study utilized machine learning methods of analysis as compared to the traditional statistical regression models. The advantage here is the big data. The data provides a platform to train the model as much as possible to improve the accuracy.

The background prevalence based on the sociodemographic characteristics correlates well with the national data especially among divorcee whom reported more depressive and stress symptoms compared to those who were married as well as single individuals (11,12). Some of them stay as single mothers and have many children to take care, contributing to potentially other illnesses. The prevalence of depression was 39.9%, higher than previous study among Malaysian adults (20.5%) (13). In terms of anxiety, it was 48.5% which was also higher according to previous study (44.5%) while stress was about 3% higher than the same previous study (13). The higher prevalence was surprising, considering the comparison study utilized adults attending primary care clinic for medical reasons. We expected that the perceived healthy community to have lower prevalence than that of psychiatric patients under follow-up, however our data suggest otherwise. This can be a red flag, or warning sign, telling that community mental health need to be looked upon, as some of them are possibly afraid to express their mental health concerns and need help.

The prediction used feedforward neural network and classification to determine model accuracy. The machine learning model correctly predicted 76% of the depressive cases. The high prediction level here correlates with good area under the curve (AUC) of 0.8 and modest cross entropy error, signifying the predictive power is good. Results here were far more realistic than a study in India which have high accuracy (97%) and receiver operating characteristics (ROC) (0.99) but small sample size and different factors being modelled (14). The strength of this model is in the number of data obtained from the online open source survey. We foresee the potential in using this data and make interval analysis of trends. From there, the drop or increase in prevalence can be seen.

Interestingly, the top five predictors of depression in descending order of was stress, anxiety, age, occupation and marital status. Divorcee also contributes to this importance in the machine learrning model, where they are most vulnerable to depression. However, for age it is unclear whether higher or younger are are more prone to get depressive symptoms in the model. When compared to a statistical model, the machine learning model identified the same, most important predictor which is stress (15). In terms of anxiety and stress models, the neural network uncovered stress and anxiety as the most important predictor, respectively. It is comparable with a study showing stress was a directly related to anxiety (16). Occupation played more role in stress as compared to the anxiety model. In other words, stress due to workplace demands are more frequently encountered as compared to anxiety triggered due to certain working conditions. This is also in line with a previous study which showed job demand was directly related with stress in a structural equation model analysis (16). Thus, it is very important here that stress and anxiety should always be monitored from time to time, and a quick online survey can help in detecting this early for early intervention. Workplace policies might not have the platform for daily assessment of employees' mental health status, hence making the online screening tools a useful alternative.

The limitation of this is the data veracity, or truthfulness of the data supplied online. The respondents are unanimous, and how they understand and answer the data honestly will determine the true accuracy of the data.

CONCLUSION

The supervised machine learning method utilized in this study was able to accurately classify more than 75% of the true depression, anxiety and stress cases via online methods. This is invaluable in pattern recognition for early signs of mental health disorders. The careful use of machine learning methods should be complemented with humanistic intellect, which is part of the Muslim Maqasid al-Shariah and thus, not to be abused.

We recommend that the online, open source information centre to be the complementary surveillance hub for not just depression, anxiety and stress condition but also for other mental health issues as well. Future studies should look at stress and anxiety as potential mediating factor to depression. Early interventions should be planned and have collaboration with health professionals and data scientists which in turn will help reduce mental health pandemic that we are currently facing.

ACKNOWLEDGEMENTS

Parts of this study were presented to the 3rd World Congress on Integration and Islamicisation 2021: Mental Health and Well Being in the 4th Industrial Revolution (3rd WCII), June 4–6, 2021, and were published in abstract form (17).

REFERENCES

1. Santomauro DF, Herrera AM, Shadid J, Zheng P, Ashbaugh C, Pigott DM, Abbafati C, Adolph C, Amlag JO, Aravkin AY, Bang-Jensen BL. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. The Lancet. 2021 Oct 8. doi: 10.1016/S0140-6736(21)02143-7
2. Yong XY, Sui CF, Liew MY, San Chong TW, Liew JY. Psychological distress screening for depression, anxiety and stress among medical ward patients in hospital Tapah, Malaysia: a cross-sectional study using the Depression, Anxiety and Stress Scale (DASS-21). Journal of Health Science and Medical Research. 2022 Feb 23;40(3):317-33. doi: 10.31584/jhsmr.2021841
3. Latiff LA, Tajik E, Ibrahim N, Bakar AS, Ali SS. Psychosocial problem and its associated factors among adolescents in the secondary schools in Pasir Gudang, Johor. Malaysian Journal of Medicine and Health Sciences. 2017 Jan 1;13(1):35-44.
4. Ramli M. MaHIR Centre IIUM: Mental Health Information and Research Centre [internet]; 2008 (cited 2021 June 11). Available from: www.ramlimusa.com
5. Koppe G, Meyer-Lindenberg A, Durstewitz D. Deep learning for small and big data in psychiatry. Neuropsychopharmacology. 2021 Jan;46(1):176-90. doi: 10.1038/s41386-020-0767-z
6. Delgadillo J, Gonzalez Salas Duhne P. Targeted prescription of cognitive–behavioral therapy versus person-centered counseling for depression using a machine learning approach. Journal of Consulting and Clinical Psychology. 2020 Jan;88(1):14. doi: 10.1037/ccp0000476
7. Ahmad MA, Eckert C, Teredesai A. Interpretable machine learning in healthcare. InProceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics 2018 Aug 15 (pp. 559-560). doi: 10.1145/3233547.3233667
8. Alwosheel A, van Cranenburgh S, Chorus CG. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. Journal of choice modelling. 2018 Sep 1;28:167-82. doi: 10.1016/j.jocm.2018.07.002
9. Lovibond SH, Lovibond PF. (1995). Manual for the Depression Anxiety & Stress Scales. Sydney: Psychology Foundation. 1995;(2)
10. Garson GD. Interpreting neural network connection weights. Artificial Intelligence Expert. 1991;6(4):46–51.
11. Imran A, Azidah AK, Asrenee AR, Rosediani M. Prevalence of depression and its associated factors among elderly patients in outpatient clinic of Universiti Sains Malaysia Hospital. The Medical Journal of Malaysia. 2009 Jun 1;64(2):134-9.
12. Fatimah S, Maideen K, Sidik SM, Rampal L, Mukhtar F. Prevalence, associated factors and predictors of depression among adults in the community of Selangor. PloS One. 2014;9(4):1-4. doi: 10.1371/journal.pone.0095395
13. Abd Rahman LR, Idris IB, Ibrahim H. Risk factors of depression, anxiety and stress among adults attending primary health clinics in an urban area in Klang Valley, Malaysia. Malaysian Journal of Medicine and Health Sciences. 2020;16(1):240-6.
14. Sau A, Bhakta I. Artificial neural network (ANN) model to predict depression among geriatric population at a slum in Kolkata, India. Journal of clinical and diagnostic research: JCDR. 2017 May;11(5):VC01. doi: 10.7860/JCDR/2017/23656.9762
15. Yeoh SH, Tam CL, Wong CP, Bonn G. Examining depressive symptoms and their predictors in Malaysia: Stress, locus of control, and occupation. Frontiers in psychology. 2017 Aug 22;8:1411. doi: 10.3389/fpsyg.2017.01411
16. Rusli BN, Edimansyah BA, Naing L. Working conditions, self-perceived stress, anxiety, depression and quality of life: a structural equation modelling approach. BMC public health. 2008 Dec;8(1):1-2. doi: 10.1186/1471-2458-8-48
17. Edre MA, Ramli M. Supervised machine learning in predicting depression, anxiety and stress using web-based big data: Preserving the humanistic intellect. In: Mat Zin N, Rahmat S, Mohammad Aidid E, Mohd Ali AS, Haris @ Harith MS, Zaini

S, editors. Proceedings of the 3rd World Congress on Integration and Islamicisation 2021. 2021 June 4-6; Kuantan, Pahang, Malaysia. Malaysian Journal of Medicine and Health Sciences. 2021; Vol.17 Supp 5, August 2021, p. 21.