Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 167 (2020) 1258–1267

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2019)

Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms

Anu Priya[a], Shruti Garg[a,*], Neha Prerna Tigga[a]

[a]*Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi – 835215, India*

Abstract

In the fast-paced modern world, psychological health issues like anxiety, depression and stress have become very common among the masses. In this paper, predictions of anxiety, depression and stress were made using machine learning algorithms. In order to apply these algorithms, data were collected from employed and unemployed individuals across different cultures and communities through the Depression, Anxiety and Stress Scale questionnaire (DASS 21). Anxiety, depression and stress were predicted as occurring on five levels of severity by five different machine learning algorithms – because these are highly accurate, they are particularly suited to predicting psychological problems. After applying the different methods, it was found that classes were imbalanced in the confusion matrix. Thus, the f1 score measure was added, which helped identify the best accuracy model among the five applied algorithms as the Random Forest classifier. Furthermore, the specificity parameter revealed that the algorithms were also especially sensitive to negative results.

Keywords: Decision Tree(DT); K-NN; Naïve Bayes(NB); Random Forest Tree(RFT); Support Vector Machine(SVM)

1. Introduction

Humans are, by nature, becoming ambitious nowadays and seek every possible opportunity to grow professionally. Anxiety, depression, stress frustration and dissatisfaction have become so commonplace that people

* *E-mail address:* gshruti@bitmesra.ac.in

10.1016/j.procs.2020.03.442

now believe them to be part and parcel of professional life. The World Health Organization (WHO) has observed that depression is the most prevalent mental disorder affecting more than 300 million people worldwide, and the severity of the issue has led many health researchers to focus their studies in this area. Differentiating anxiety, depression and stress from one another is problematic for machines; hence, an appropriate learning algorithm is required for an accurate diagnosis. According to WHO, a healthy person possesses a healthy brain along with physical wellness. [1]

The standard diagnosis criterion for depression is the Patient Health Questionnaire (PHQ), whilst the Depression, Anxiety and Stress Scale (DASS 21), which has 21 questions, is used for screening the symptoms related to these mental illnesses [5-6].

The main symptoms of depression [3] from a clinical point of view are loss of memory; lack of concentration; an inability to make decisions; loss of interest in recreational activities and hobbies including sex; overeating and weight gain; low appetite and weight loss; feelings of guilt, worthlessness, helplessness, restlessness and irritation; as well as suicidal thoughts. In case [2], these symptoms were found to have a significant effect on important areas of an individual's life – such as in education, employment and social activities, and this provides a vital clue for forming a clinical diagnosis.

The symptoms of GAD (Generalised Anxiety Disorder) [3] are irritability, nervousness, fatigue, insomnia, gastro intestinal problems, panic, and a sense of impending danger, increased heart rate, sweating, rapid breathing and difficulty concentrating.

The symptoms of stress [4] are feeling upset or agitated, an inability to relax, low energy levels, chronic headaches, frequent overreaction and persistent colds or infections. Thus, stress, anxiety and depression have many common symptoms including insomnia, chest pain, fatigue, increased heart rate and inability to concentrate, all of which makes classification challenging for machines.

This paper is structured as follows: Section 2 explores related studies on anxiety, depression and stress along with the methods and techniques that were adopted. Section 3 describes the materials and methodology used in the research herein, whilst Section 4 shows the results that were gained after applying the classification algorithms. Finally, section 5 is the conclusion, which summarises the study in its entirety.

2. Related Studies

Many researchers have worked on predicting anxiety and depression with machine learning algorithms, such as Random Forest Tree (RFT), the Support Vector Machine (SVM) and the Convolution Neural Network (CNN) for the collection and subsequent classification of data from blog posts. For encoding the text, various techniques have been used, that is topic modelling, Bag-of-Words (BOW) and Term Frequency–Inverse Document Frequency (TF–IDF). Moreover, Python programming has been used for modelling experiments, with the best results among all the classifiers[2] being produced by the CNN, whose accuracy and recall scores were found to be 78% and 0.72, respectively.

Different machine learning algorithms such as Logistic Regression, Catboost, Naïve Bayes, RFT and SVM were applied for classification in [7]. In this study, 470 seafarers were interviewed and information on the occupations, socio-demographics and health of the participants was collected via 16 characteristics including age, academic qualifications, monthly income, employment status, BMI, duration of service, family type, marital status, presence (if any) of hypertension, diabetes or ischemic heart disease, job profile, rank within the organisation, types of vessels posted to and dummy variables for academic qualifications and marital status. As a result, the researchers found that Catboost produced the highest levels of accuracy and precision among all the classifiers – i.e. 82.6% and 84.1%, respectively.

Sau *et al.* (2017) manually collected data from the Medical College and Hospital of Kolkata, West Bengal on 630 elderly individuals, 520 of whom were in special care. After applying different classification methods Bayesian Network, logistic, multiple layer perceptron, Naïve Bayes, random forest, random tree, J48, sequential random optimization, random sub-space and K star they observed that random forest produced the best accuracy rate of 91% and 89% among the two data sets of 110 and 520 people, respectively. For feature selection and classification, WEKA tool were used in [1].

These days, social media is rapidly turning into a healthcare evaluation tool for predicting various types of illness. Saha *et al.* [8] selected topics and psycholinguistic attributes appearing in posts on the LiveJournal website. These were then inputted into a joint modelling framework, so as to categorise the mental problems occurring in online communities with an interest in depression. The proposed joint modelling framework outperformed the existing single task learning (STL) and multi task learning (MLT) baselines, and the study showed that discussions in online communities went beyond feelings of being depressed.

Reece *et al.* [9] focused on the predictors of depression and Post Traumatic Stress Disorder (PTSD) among Twitter users. The Hidden Markov Model (HMM) was used to recognise increases in the probability of PTSD. Of the entire dataset, 31.4% and 24% were observed to be affected by depression and PTSD. Braithwaite *et al.* [10] collected tweets from 135 participants recruited from Amazon Mechanical Turk (MTurk) and applied decision tree classification to measure suicide risk. The accuracy level for the prediction of suicide rate was observed to be 92%.

Du *et al.* [11] extracted streaming data from Twitter and used psychiatric stressors to annotate tweets that had been deemed suicidal. The Convolution Neural Network (CNN) outperformed the Support Vector Machine (SVM) and extra trees (ET) etc. with a precision of 78% in recognising tweets with suicidal tendencies.

The audio-text approach can also be used to model depression, where the researcher collects data from individuals with depression. The long short-term memory neural network model was used for detecting depression in [12], which observed that the context-free model produced the best results for audio (weighted, sequence and multi-model).

Depression was also predicted in [13] in the early stages through social media content. Data collection was carried out using CLEF eRisk. After evaluating five systems, it was discovered that a combination of machine learning and information retrieval gave the optimum result. In Hou *et al.*, a big data approach was used to predict depression based on a person's reading habits. The features of Chinese text were extracted in order to develop a book classifier and after applying five classifications, naïve Bayes was found to be the most appropriate [14].

Post-traumatic stress disorder has detected in [15] using supervised machine learning classifiers. Their study is on ex-serviceman UK militants, the parameters used in their study alcohol misuse, gender and deployment status. As results satisfactory sensitivity was obtained for multiple supervised Machine Learning classifiers, but the outcomes were not very sensitive to false negative diagnoses.

Anxiety and mood disorder were detected in [16] by scanning patient facial emotions and applying cross validation and better precise results were found that is verified by different statistical measures.

Imbalance classification was applied in [20] and ensemble machine learning methods were discussed in [21].

Different researchers have applied different machine leaning algorithms for the prediction of psychological disorders, and the performances of different algorithms have been found to vary, depending on the scenario; no fixed algorithm has been determined as most suitable in all cases. Thus, in the present study, all the machine learning algorithms were applied to identify the symptoms of anxiety, depression and stress.

3. Materials and Methodology

This research focused on detecting anxiety, depression and stress using the Depression, Anxiety and Stress Scale questionnaire (DASS 21). Data were collected from a total of 348 participants via Google forms and subsequently classified using five machine learning algorithms – namely Decision Tree, Random Forest Tree, Naïve Bayes, Support Vector Machine and KNN.

3.1. Participants

This study was conducted on a total of 348 participants aged between 20 and 60 years, both males and females, employed and unemployed and with a wide range of responsibilities from household chores to professional duties who were asked to complete a questionnaire.

3.2. Questionnaires

The data for the study were collected through DASS-21, the Depression, Anxiety and Stress Scale questionnaire. DASS 21 comprises 21 questions, with 7 questions allocated to each of the scales of Stress, Anxiety and Depression. The possible answers for each question – which could be given in text or numeric form – are as follows:

0 did not applied to me
1 applied to me to some degree, or some of the time.
2 applied to me to a considerable degree or a good part of time.
3 applied to me very much or most of the time.

The questions asked from individuals are described in table 1.

Table 1. Questionnaires on anxiety, depression and stress.

	Anxiety	Depression	Stress
1.	Dryness of Mouth	Couldn't Experience the positive feeling	Found hard to wind down
2.	Difficulty in Breathing	Difficult to work up the initiative to do things	Overreact to situations
3.	Experience Trembling	Nothing to look forward	A lot of nervous energy
4.	Worried about panic and make a fool of themselves	Felt down-hearted and Blue	Getting Agitated
5.	Close to Panic	Unable to become enthusiastic	Difficult to Relax
6.	Aware of the action of the heart in the absence of physical exertion	Felt wasn't worth much as a person	Intolerant to getting what I was doing
7.	Felt scared without any good reason	Felt life was meaningless	Touchy

Following the data collection, the participants' responses were encoded using numeric values of 0 to 3, and the scores were then calculated by adding the values associated with each question and the below formula:

$$\text{score} = \text{Sum of rating points of each class} * 2 \tag{1}$$

Once the final scores had been calculated, these were labelled according to severity – i.e. Normal, Mild, Moderate, Severe and extremely severe (see Table 2).

Table 2. Severity levels.

	Anxiety	Depression	Stress
Normal	0-7	0-9	0-14
Mild	8-9	10-13	15-18
Moderate	10-14	14-20	19-25
Severe	15-19	21-27	26-33
Extremely severe	20+	28+	33+

3.4. Classification

The machine learning algorithms were applied in R programming language using Rstudio version 3.5. This predicts the percentage of people suffering from symptoms of stress, anxiety and depression, according to the level of severity. The dataset was divided into the ratio 70:30, representing the training and test sets, respectively. The working principles of each machine learning algorithm are described in the following subsections.

3.4.1. Decision Tree

The decision tree method of machine learning makes decisions at different levels using tree data structure – this is suitable for predictive problems because they are easy to interpret and the structure is stable. It covers both classification (tree models with a volatile target for a distinct set of values) as well as regression (a volatile target for an endless set of values) [17]. In figure 1 decision tree example is shown. In this a question is divided into yes or no (binary; 2 alternatives) into two branches (yes and no) driving out of the tree. One can get a greater number of choices than 2.

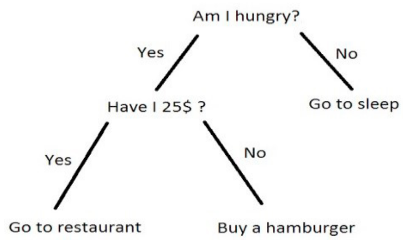

Fig. 1. Decision Tree example

3.4.2. Random Forest

The Random forest classifier creates multiple decision trees from randomly selected subset of training dataset as shown in figure 2. Then it aggregates the votes from different decision trees to decide the final class of test objects [19]. A random forest classification was proposed in [18] with reduced number of trees.

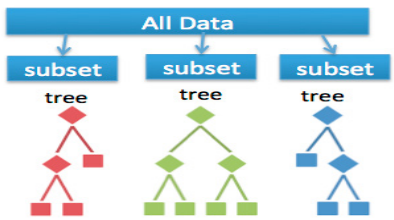

Fig. 2. Random forest

3.4.3. Support Vector Machine (SVM)

A Support Vector Machine [22] is a machine learning algorithm that works for both regression and classification tasks but is mainly used in classification. This classifier [23] has been utilised of late in many applications because of its exceptional classifying ability and presentation quality, dividing the data linearly into two separate classes (also known as hyperplanes), with the maximum distance between the two classes as represented in figure 3.

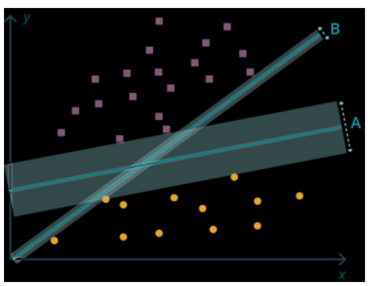

Fig. 3. Support vector machine representation

3.4.4. Naïve Bayes

This classifier superintends machine learning algorithms using Bayes theorem and works on the premise that features are analytically independent. This theorem depends on naïve assumption, in which input factors are independent of each other [24-25]. The formula for naive bayes is given as follows:

$$p(H \mid D) = \frac{p(H)p(D \mid H)}{p(D)} \tag{2}$$

Where,

$p(H|D)$ = this is posterior
$p(H)$ = this is the prior i.e. what you believed before you saw the evidence
$p(D|H)$ = this the likelihood of seeing that evidence if your hypothesis is correct
$p(D)$ = this is the normalizing of that evidence under any circumstances

3.4.5. K- Nearest Neighbour (K-NN)

K-NN is one of the most straightforward algorithms adopted in machine learning for classification and regression problems. Based on closest measures, KNN takes information and classifies recent information points. The information is then allotted to the class with the foremost closest neighbour. The illustration is shown in figure 4. KNN [26] is often used to classify future information due to its ease of execution and adequacy.

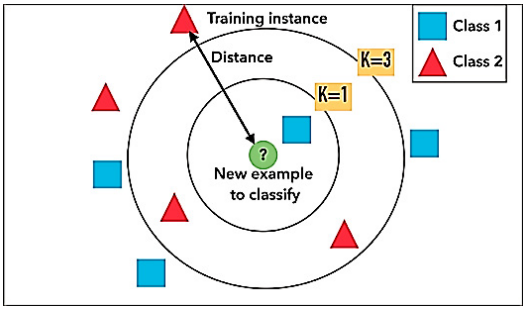

Fig. 4. K- nearest neighbour representation

4. Results and discussions

The application of all five methods – i.e. Decision Tree (DT), Random Forest Tree (RFT), Naïve Bayes (NB), Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) – to all three classes of Stress, Anxiety, and Depression, resulted in the confusion matrices depicted in Table 3. The rows in the matrices show the actual classes whilst the columns show the predicted classes. The numbers 1, 2, 3, 4 and 5 in the rows and columns represent normal, mild, moderate, severe and extremely severe cases, respectively.

Equations 3-8 below were used to calculate the accuracy and error rates, precision, recall and specificity in each confusion matrix.

$$\text{Accuracy Rate} = \frac{\text{Sum of diagonals (TP)}}{\text{Total number of instances}} \tag{3}$$

$$\text{Error Rate} = 1 - \text{Accuracy Rate} \tag{4}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{7}$$

$$\text{F1 Score} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \tag{8}$$

Whereas,

TP (True positive) = Diagonals of matrix
FN (False Negative) = Sum of the consistent row for class (excluding TP of that class)
FP (False Positive) = Sum of the corresponding column for class (excluding TP of that class)
TN (True Negative) = Sum of the all row and column (excluding row and column of that class)

Table 3. Confusion Matrix

	Anxiety						Depression						Stress					
Decision Tree		1	2	3	4	5		1	2	3	4	5		1	2	3	4	5
	1	21	0	4	0	0	1	31	4	2	0	0	1	39	3	1	0	0
	2	2	0	5	0	0	2	4	7	1	0	0	2	3	8	5	3	0
	3	5	0	30	0	0	3	0	0	24	1	0	3	6	8	8	1	0
	4	0	0	9	0	2	4	0	0	5	5	1	4	1	3	3	8	1
	5	0	0	1	0	26	5	0	0	0	5	14	5	0	0	0	1	3
Random Forest		1	2	3	4	5		1	2	3	4	5		1	2	3	4	5
	1	19	0	6	0	0	1	35	0	2	0	0	1	38	4	1	0	0
	2	6	0	1	0	0	2	9	1	2	0	0	2	3	13	3	0	0
	3	4	0	29	0	2	3	1	0	24	0	0	3	1	3	12	7	0
	4	0	0	8	0	3	4	0	0	4	5	2	4	0	0	6	10	0
	5	0	0	0	0	27	5	0	0	1	0	18	5	0	0	0	1	3
Naive Bayes		1	2	3	4	5		1	2	3	4	5		1	2	3	4	5
	1	20	0	5	0	0	1	30	7	0	0	0	1	41	2	0	0	0
	2	7	0	0	0	0	2	0	10	2	0	0	2	2	12	5	0	0
	3	4	1	30	0	0	3	0	0	25	0	0	3	0	5	16	2	0
	4	0	0	10	1	0	4	0	0	1	9	1	4	0	2	5	9	0
	5	0	0	1	0	26	5	0	0	0	4	15	5	0	0	0	4	0
Support Vector Machine		1	2	3	4	5		1	2	3	4	5		1	2	3	4	5
	1	19	0	6	0	0	1	33	2	2	0	0	1	38	4	1	0	0
	2	6	0	1	0	0	2	6	4	2	0	0	2	3	10	6	0	0
	3	4	0	31	0	0	3	1	0	22	2	0	3	1	5	10	7	0
	4	0	0	8	0	3	4	0	0	4	5	2	4	0	2	5	9	0
	5	0	0	0	0	27	5	0	0	1	0	18	5	0	0	0	1	3
K-NN		1	2	3	4	5		1	2	3	4	5		1	2	3	4	5
	1	20	0	5	0	0	1	35	0	2	0	0	1	37	5	1	0	0
	2	6	0	1	0	0	2	7	3	2	0	0	2	1	11	7	0	0
	3	4	0	31	0	0	3	5	0	16	4	0	3	1	4	11	7	0
	4	0	0	7	1	3	4	0	0	3	6	2	4	0	1	5	10	0
	5	0	0	1	1	25	5	0	0	0	4	15	5	0	0	0	1	3

Table4. Values of different measures for different classification methods.

Classifier	Mental illness	Accuracy	Error Rate	Precision	Recall	Specificity	F1 Score
Decision	Anxiety	0.733	0.267	0.458	0.532	0.923	0.492
Tree	Depression	0.778	0.222	0.731	0.714	0.909	0.723
	Stress	0.628	0.372	0.599	0.585	0.900	0.592
Random	Anxiety	0.714	0.286	0.431	0.51	0.919	0.470
Forest	Depression	0.798	0.202	0.881	0.678	0.910	0.766
	Stress	0.723	0.277	0.731	0.692	0.928	0.711
Naive	Anxiety	0.733	0.267	0.459	0.542	0.924	0.497
Bayes	Depression	0.855	0.145	0.822	0.850	0.917	0.836
	Stress	0.742	0.258	0.548	0.568	0.934	0.558
Support	Anxiety	0.678	0.322	0.403	0.504	0.914	0.448
Vector	Depression	0.803	0.197	0.820	0.716	0.908	0.765
Machine	Stress	0.667	0.333	0.672	0.631	0.921	0.651
K Nearest	Anxiety	0.698	0.302	0.449	0.530	0.913	0.527
Neighbour	Depression	0.721	0.279	0.750	0.634	0.892	0.687
	Stress	0.714	0.286	0.719	0.682	0.921	0.700

Table 4 shows the accuracy, error rate, precision, recall, specificity and f1 score of each class obtained by the different algorithms. From the observation made in Table 4, the highest accuracy for all three scales ofanxiety, depression and stress were achieved by Naïve Bayes. Nevertheless, the results in Table 3 show that the classes were imbalanced, because the confusion matrices of anxiety, depression and stress produced 25, 37 and 43 instances of normal but 7, 12 and 19 instances of mild, respectively. Similarly, there were 35, 25 and 23 instances of moderate; 11, 11 and 16 instances of severe; and 27,19 and 4 instances of extremely severe for the scales of Anxiety, Depression and Stress respectively; thus, the classes here were imbalanced. In such cases, accuracy alone is not a sufficient measure, and the f1 score becomes an important measure for determining the best model. The f1 score is a harmonic mean of precision and recall, whose value is higher when both precision and recall are higher. Consequently, the best model in cases of classification imbalance is the one whose f1 score is higher even if its accuracy is lower. The f1 score of Random Forest was the highest for Stress and Naïve Bayes for depression whereas it is low for all the algorithms for anxiety. The specificity of all the algorithms was found to be around 90% or above in all three cases. Moreover, this is an important parameter in healthcare because it shows that the negative cases(patients without diseases) are also classified appropriately. All of the algorithms applied in this study produced highly accurate results for negative cases as well.

(a)

```
Variable importance
                          Scared_without_any_good_reason
                                                      26
                                     Breathing_difficulty
                                                      20
                                           Close_to_panic
                                                      19
Aware_of_the_action_of_heart_in_the_absence_of_physical_exertion
                                                      15
                                         Dryness_of_mouth
                                                      15
                                     Experienced_trembling
                                                       5
```

(b)

```
Variable importance
                         Life_was_meaningless
                                           19
                      Nothing_to_look_forward_to
                                           18
                           Down_hearted_and_blue
                                           18
                    Wasnt_worth_much_as_a_person
                                           16
         Unable_to_become_enthusiastic_about_anything
                                           14
             could_nt_experience_any_positive._feeling
                                           11
Difficult_to_work_up_the_initiative_to_do_things
                                            5
```

(c)

```
Variable importance
            Difficult_to_relax                 Getting_agitated
                            23                               20
   I_felt_that_I_was_rather_touchy                Hard_to_wind_down
                            20                               13
                nervous_energy Intolerant_to_getting_what_I_was_doing
                            12                               12
```

Fig. 5. Variable importance (a) Anxiety (b) Depression (c) Stress.

Figure 5 shows the variable importance ratings for Anxiety, Depression and Stress, respectively; the higher the number, the more important the variable. That is, the variable 'Scared_without_any_good_reason' proved to be most important on the Anxiety scale; the variable 'Life_was_meaningless' was found to be important on the Depression scale whilst'Difficult_to_relax' was found to be important on the Stress scale.

4. Conclusion

In this paper, machine learning algorithms were applied to determine five different severity levels of anxiety, depression and stress. Data were collected using a standard questionnaire measuring the common symptoms of anxiety, depression and stress (DASS-21). Subsequently, five different classification techniques were applied – Decision Tree (DT), Random Forest Tree (RFT), Naïve Bayes, Support Vector Machine (SVM) and K- Nearest Neighbour (KNN). The accuracy of naïve Bayes was found to be the highest, although Random Forest was identified as the best model. Due to the fact that this problem produced imbalanced classes, the best-model selection was made on the basis of the f1 score, which is used for cases of imbalanced partitioning. The important variables were found to be 'scared_without_any_good_reason', 'Life_was_meaningless' and 'Difficult_to_relax' for the scales of Anxiety, Depression and Stress, respectively. As such, these variables were considered to be most important in detecting psychological disorder.

References

[1] Sau, A., Bhakta, I. (2017)"Predicting anxiety and depression in elderly patients using machine learning technology."*Healthcare Technology Letters* **4 (6)**: 238-43.

[2] Tyshchenko, Y. (2018)"Depression and anxiety detection from blog posts data."*Nature Precis. Sci., Inst. Comput. Sci., Univ. Tartu, Tartu, Estonia.*

[3] https://adaa.org/understanding-anxiety/depression/symptoms

[4] https://www.webmd.com/balance/stress-management/stress-symptoms-effects_of-stress-on-the-body#1

[5] Oei, T. P., Sawang, S., Goh, Y. W., Mukhtar, F. (2013) "Using the depression anxiety stress scale 21 (DASS-21) across cultures." *International Journal of Psychology* **48 (6):** 1018-1029.

[6] Kroenke, K., Spitzer, R. L.,Williams, J. B. (2001) "The PHQ-9: validity of a brief depression severity measure."*Journal of general internalmedicine* **16 (9)**: 606-613.

[7] Sau, A., Bhakta, I. (2018) "Screening of anxiety and depression among the seafarers using machine learning technology."*Informatics in Medicine Unlocked* :100149.

[8] Saha, B., Nguyen, T., Phung, D., Venkatesh, S. (2016) "A framework for classifying online mental health-related communities with an interest in depression."*IEEE journal of biomedical and health informatics* **20 (4)**: 1008-1015.

[9] Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., Langer, E. J.(2016) "Forecasting the Onset and Course of Mental Illness with Twitter Data."*Scientific reports* **7 (1)**: 13006.

[10]Braithwaite, S. R., Giraud-Carrier, C., West, J., Barnes, M. D., Hanson, C.L. (2016) "Validating machine learning algorithms for Twitter data against established measures of suicidality."*JMIR mental health* **3 (2)**: e21.

[11]Du, J., Zhang, Y., Luo, J., Jia, Y., Wei, Q., Tao, C., Xu, H. (2018) "Extracting psychiatric stressors for suicide from social media using deep learning."*BMC medical informatics and decision making* **18 (2)**: 43.

[12]Al Hanai, T., Ghassemi, M. M., Glass, J.R. (2018) "Detecting Depression with Audio/Text Sequence Modeling of Interviews." *InInterspeech* : 1716-1720.

[13]Ramiandrisoa, F., Mothe, J., Benamara, F., Moriceau, V. (2018) "IRIT at e-Risk 2018."*E-Risk workshop*: 367-377.

[14] Hou, Y., Xu, J., Huang, Y., Ma, X. (2016) "A big data application to predict depression in the university based on the reading habits."*3rd IEEE International Conference on Systems and Informatics (ICSAI)*: 1085-1089.

[15] Leightley, D., Williamson, V., Darby, J., Fear, N.T. (2019) " Identifying probable post-traumatic stress disorder: applying supervised machine learning to data from a UK military cohort." *Journal of Mental Health*. **28 (1)**: 34-41.

[16]Young, C., Harati, S., Ball, T., Williams, L. (2019) "Using Machine Learning to Characterize Circuit-Based Subtypes in Mood and Anxiety Disorders."*Biological Psychiatry* **85 (10)**: S310.

[17] Li, L., Zhang, X. (2010) "Study of data mining algorithm based on decision tree." *In2010 International Conference On Computer Design and Applications IEEE* **1**: V1-155.

[18] Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintha, A. R., Kundu, S. (2018)"Improved random forest for classification."*IEEE Transactions on Image Processing* **27 (8)**: 4012-4024.

[19] Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J. P. (2012) "An assessment of the effectiveness of a random forest classifier for land-cover classification."*ISPRS Journal of Photogrammetry and Remote Sensing* **67**: 93-104.

[20] Liu, X. Y., Wu, J., Zhou, Z. H. (2009)"Exploratory undersampling for class-imbalance learning."*IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **39 (2)**: 539-50.

[21]Dietterich, T. G. (2000) "Ensemble methods in machine learning."*InInternational workshop on multiple classifier systems*Springer, Berlin, Heidelberg: 1-15.

[22] Saitta, L., (2000) "Support-Vector Networks." *Kluwer Acad. Publ. Bost.* : 273–297.

[23] Hamed, T., Dara, R., Kremer, S. C. (2014) "An accurate, fast embedded feature selection for SVMs." *In2014 13th International Conference on Machine Learning and Applications*IEEE :135-140.

[24] Martinez-Arroyo, M., Sucar, L. E. (2006) "Learning an optimal naive bayes classifier."*In18th International Conference on Pattern Recognition (ICPR'06)* IEEE **3**: 1236-1239.

[25] Cheng, J., Greiner, R. (1999) "Comparing Bayesian network classifiers."*InProceedings of the Fifteenth conference on Uncertainty in artificial intelligence*Morgan Kaufmann Publishers Inc : 101-108.

[26] Taneja, S., Gupta, C., Goyal, K., Gureja, D. (2014) "An enhanced k-nearest neighbor algorithm using information gain and clustering." *Fourth International Conference on Advanced Computing & Communication Technologies* IEEE : 325-329.