Transfer Learning for Depression Detection on Social Networks*

Oliver Chi[1] and Xiaohui Tao[1]

School of Information System, University of Southern Queensland, Australia
{ochi,xtao}@usq.edu.au

Abstract. The abstract should briefly summarize the contents of the paper in 150–250 words.

Keywords: major depressive disorder · ensemble classification technique · supervised learning · psychological knowledge base.

* Supported by School of Information System, University of Southern Queensland, Australia

1 Introduction

Major depressive disorder also known simply as depression, is a global challenge for healthcare. In psychological domain, it is defined as a mental disorder consistent with at least two weeks of developing low mood across most situations [Zimmerman]. Major depressive disorder can be successfully diagnosed by interviewers normally psychologists, applying operational diagnostic criteria of depression. However, a wide range of depressive patients did not seek clinic advices or professional care at all [Huerta-Ramírez]. Health professionals hence often fail to approach a proper depressive patient at an early stage.

Early diagnosis of depression are among the priority actions for reducing the burden of depressive disorders [Huerta-Ramírez]. With the growing popularity of artificial intelligence, one of methods in early diagnosis is to apply machine learning technique in the processing of exploring depressive patients from wide range of the potential persons. Using algorithms to learn from individual health history and previous behaviours, machine learning enable computers to automatically distinguish person with depression from persons without depression. It is apparently quick comparing to traditional interview method. And it has a great potential to apply similar algorithms to crowdsource depression on online social networks which is able to approach millions of users without a significant cost of healthcare. However, there is a lack of research into an efficient machine learning classifier for detecting depression based on a large data.

Therefore, the objective of the present paper is to propose a suitable effective machine learning method in discriminating depression from collected health data for further interview diagnosis.

2 Related Work

2.1 Psychological Studies Of Clarifying Depression

In the domain of psychology, researchers created questionnaire-based scale instruments for roughly screening the mental status of patients. Several measures among them have been widely accepted as good and reliable tools, such as BDI [Huerta-Ramírez][Tsugawa], CES-D [Choudhury][Tsugawa], SCL-20 and PHQ-9 [PHQ-9].

Zimmerman et, al. [Zimmerman] compared a preferred scale measure "IDD" of depression to psychiatric interview and found that it had good reliability and was significantly associated with the result of clinician's diagnosis. The proposed method had a 97.2 % [Zimmerman] overall accuracy on the criterion of clinic diagnosis. And new method is suggested to be an inexpensive way to collect large homogeneous cases to screen depression. However, the research also appointed three apparent reasons why self-report depression scales generally performed poorer than psychiatric interview: first, scale measure covered less diagnostic criteria for major depressive disorder than interviews; second, non-criteria items in depression scale inventory decreased specificity of performance; and self-report scales did not assess symptom duration and exclusion criteria which may reverse the previous judgement [Zimmerman]. The classification via scale measure and clinical interview implied contrasting algorithms [Zimmerman] to screen depression but all based on the depression diagnostic criteria.

Sakado et, al. [Sakado] designed a 22-item inventory "IDL" to diagnose depression based clinical criteria of major depressive disorder. The proposed inventory had a sufficient discriminant validity in identifying a depressive disorder with or without previous medical record of depression. It had a couple of advantages comparing to traditional interview. For instance, it is inexpensive and speedy in examining a lifetime prevalence of major depression [Sakado]. And it avoided past history of depression distorting the current action of identifying depressive patient [Sakado]. In the experiment, 22-item inventory reached a higher sensitivity of 83 %.The method is thought to be a both inexpensive and quick instrument for preliminary screening of major depressive disorder.

Kroenke et, al [PHQ-9] examined the performance of Patient Health Questionnaire (PHQ-9) self-administrated diagnostic instrument by companion to 20-item Short-Form General Health Survey (SF-20) and an independent structured mental health professional interview. Using the psychological interview as the criterion standard, main score of PHQ-9 had a sensitivity of 88 % and a specificity of 88 % for major depression [PHQ-9]. The study assessed depression severity in PHQ-9 measure to improve the detection. PHQ-9 scores of 5, 10, 15, 20 represented valid thresholds of limits in mild, moderate, moderately severe and severe depression [PHQ-9]. For instance, persons with scores less than 10 were seldom diagnosed major depressive disorder in clinical interview. A very close association between PHQ-9 and SF-20 mental health inventory was observed. PHQ-9 scores were almost perfectly correlated with SF-20 scores in all five subjects [PHQ-9]. Moreover, the research highlighted one of PHQ-9 advantages that was ”its exclusive focus on the 9 diagnostic criteria for DSM-IV depressive disorder” [PHQ-9]. PHQ-9 method hence excluded the non-criteria symptoms for measure specifically for major depression. It was able to discriminate accurately major depression from other psychological disorders such as anxiety, hopelessness and distress.

In summary, there appears to be many comparable scale measures for major depressive disorder. Most of them are in the form of questionnaire such as IDD, IDL, SF-20 and PHQ-9. Their performance are strongly correlated with mental health interview. Some measures like PHQ-9 also demonstrated a high sensitivity in screening depression over time change [PHQ-9]. However, there are still inconvenience occurred during managing questionnaire for a large cases. Therefore, researchers recommended those measures being useful tools for monitoring outcomes of depression therapy [PHQ-9] and preliminary screening before formal mental health inventory [Sakado]. Those measures demonstrated successful instances that using mental health diagnostic criteria to establish a quick and inexpensive measurement for the detection of depression.

2.2 Complexity In Dataset From Social Media For Detecting Depression

The classification of depression-indicative actions on social media is an exciting direction for detecting major depression. Mining and analysis of social media activities in order to distinguish depressive people from a wide online community has a great momentum recently among researchers. One portion of those researches is to identify the diagnostic features which indicated depressive symptoms from online media.

Ophir et, al. [Ophir] collected 190 adolescents who received depression treatment and researched their Facebook status to look for depression indicators among their online social network activities. The research used a multiple regression analysis to reveal features that estimated depression scores [Ophir]. They extracted 13 status update features that discriminated mental status update in both depressive direction and non-depressive path. Four significant features among them enabled to predict the worse status in depression [Ophir]:

a) depressive symptoms according to diagnostic criteria,
b) cognitive distortions,
c) poetic-dramatic verbal function,
d) attitudes towards others.

By the judgement of ten psychological experts, three features had the higher depression correlation scores as depressive symptoms (0.839), cognitive distortions (0.748) and content valence (0.698) [Ophir]. Although there was a lack of predictive validity in the study, it still laid the ground for research aimed at detecting depression online [Ophir]. It made contribution to extract features from online activities for identifying depressive status.

Mowery et, al. [Mowery] developed an annotated corpus of 9300 tweets from Twitter APIs using depression-related keywords. The study also used nature language processing to detect both depression symptoms and psychological stressors in tweet corpus and investigated the correlations between them. The analyses of depressive tweets suggested that only searching keywords is insufficient to predict depressive tweets because context can change the meaning of keywords in the tweet [Mowery]. In fact, depression symptoms and psychosocial stressors were observed in the tweets without depression-related keyword [Mowery]. The association between several depression symptoms and psychosocial stressors were observed as well, such as disturbed sleep correlated with educational problem. One advantage of this methodology was that it could capture relevant depressive symptoms without a formal diagnosis [Mowery]. However, there was a limitation of the proposed dataset that no more than one depression symptom or psychosocial stressor were observed in each tweet [Mowery].

From these two studies on extracting depressive features, the task that aims at establishing a suitable dataset from online social media to predict depression based on extracted features is actually intricate. And there is a lack of machine learning methodology in these researches to automate the proceeding of classifying depression from non-depression.

2.3 Research Of Classifying Depression On Social Media

From various online contents of depressive symptoms and non-depression indicators, it is possible to use machine learning techniques to develop automatic detection systems for major depressive disorders. Unusual actions and uncommon patterns of interaction [Wongkoblap] in social networks can be classified into cases or non-cases of depression through existing learning algorithms, such as Support Vector Machine algorithm, Naive Bayes method and Random Forest technique.

Choudhury et, al. [Choudhury] developed a probabilistic model to train crowdsourcing Twitter posts to determine if depression-related by support vector machine algorithm. Using the model, a social media depression index was created to characterise the levels of depression in population. It confirmed that the depression index from the proposed model had a strong correlation with national depression statistics [Choudhury]. It also provided solid evidence that understanding peoples' social environment was useful for detecting depression severity. The study used a Support Vector Machine classifier with RBF kernel for identifying depressive instances. Five-fold cross validation was used to validate the performance of classifier. The results indicated that the best model yielded an average accuracy of 73 % and high precision of 82 % [Choudhury].

Tsugawa et, al. [Tsugawa] also build a SVM supervised learning model to use features from online tweet activities for predict users' current depression status. The study showed that an accuracy of 69 % can be reached through the prediction of depressive users by the proposed classifier [Tsugawa]. The trusted status (critical standard) of users were generated by CES-D and BDI screening scales of all participants. Features used for predicting depression were extracted from the activity history of users, not like other researches from depressive symptoms. It pointed that long observation periods for collecting data may decrease accuracy [Tsugawa].

Moreover, researchers compared several supervised learning techniques to achieve the best performance of predicting depression. For instance, Hassan et, al. [Hassan] used majority vote for classification and regression of depression on top of predictions of three single classifiers, Support Vector Machine (SVM) classifier, Naive Bayes (NB) classifier, and Maximum Entropy (ME) classifier. The study illustrated how to find individual depression scale by observing and extracting emotions from the text, using machine learning techniques and natural language processing techniques on different social media platforms [Hassan]. The performance was observed that the accuracy of SVM is 91 %, 83 % and 80 % respectively for NB and ME classifiers.

Fatima et, al. [Fatima] applied Random Forest (RF) algorithm and SVM technique to discriminate the depressive posts and communities from non-depressive ones based on online social contents. The research extracted features from an online communication platform "LiveJournal" as the input of the classification algorithm. LiveJournal provided pre-defined mood tags which enabled to indicate the level of depression in each community and post. Researchers implemented Random Forest algorithm with SVM classifier for text classification to find the maximum margin between severe depressed, moderate depressive and non-depressed classes [Fatima]. They recommended that RF is a very powerful classifier in algorithm for establishing an accurate model for multi-class classification [Fatima]. In the experiment, RF performed better in comparison with SVM method [Fatima]. The proposed model achieved about 90 % and 95 % accuracy in classifying the depressive posts and depressed communities, respectively.

Peng, Hu and Dang [Peng] proposed a multi-kernel SVM based model to recognise the depressed people based on Chinese social media Weibo. Three categories of features, user microblog text, user profile and user behaviours, are extracted from Weibo for classification [Peng]. Compared with Naive Bayes, Decision Trees, KNN and single-kernel SVM techniques, multi-kernel SVM method had a lowest error rate 16.5% for identifying the depressed people [Peng]. The research also compared with the latest ensemble method which can obtain better predictive performance using multiple learning algorithms than the traditional learning algorithms alone [Peng].

Expect these studies relied on text analysis, Reece and Danforth discovered a 100-tree Random Forests methodology for analysing photographic data from Instagram to predictively screen for depression. They employed a couple of machine learning algorithms but 100-tree Random Forests algorithm had the best performance of 70 % accuracy with a reasonably low number of mis-identities [Reece]. However, the results showed that their predictive method for pre-diagnosis was rather conservative and tended to detect no depression in all instances [Reece].

Despite an increasing number of studies investigating depression using social media data, some common problems persist [Wongkoblap]. Successfully distinguishing depressive users from rich data source like social media is problematic, "not only due to biases associated with the collection methods, but also with regard to managing consent and selecting appropriate analytics techniques" [Wongkoblap]. In order to improve the performance of preliminary screening major depressive disorders in a relatively large samples, we need to explore a more suitable and high-quality classification technique for detecting depression.

3 Definitions/Research Problem

This research aims to design an effective classification method for automatically detecting depressive risk of users in health data and is also able to potentially expend the model to be implemented in the real environment of social networks.

The research objective is defined:
Definition 1 *Let $\mathbb{S}$ be a set of user properties to present an effective user profile for depression, a user property $s \in \mathbb{S}$ is a tuple $s := \langle p_1, p_2, p_3, \cdots p_n \rangle$, where*

- *p is a visualisation or instance of an user property;*
- *p is not a mental or depression close-related symptom;*
- *n could be an infinite integer so the number of p elements could be unlimited;*
- *all p elements in the same user profile are generally independent.*

With clear definition of research objective, the research target is defined:
Definition 2 *Let $\mathbb{V}$ be a set of labeled user depression, a label of user depression $v \in \mathbb{V}$ is a screening result of personal depression, where*

- *when v is binary, it presents depression (1) or healthy (0);*
- *when v is scale, it presents the severity of depression from healthy (0) to most severe depression(1).*

From Definition 1, any given user property s $\in \mathbb{S}$ is possibly overlapped with other user properties. The overlapped information in user profile apparently doesn't suit for classification. While learning from related psychological researches, a set of user personal functionings can present a perfect reflection of user mental profile. It innovates a creative method that detecting user depression by analysis of a set of user functionings. Therefore, the research problem is defined:
Definition 3 *Let $\mathbb{U} = \langle u_1, u_2, u_3, \cdots u_k \rangle$ be a subset of $\mathbb{S}$, any element $u \in \mathbb{U}$ is a tuple $u := \langle p'_1, p'_2, p'_3, \cdots p'_{n'} \rangle$,where*

- *$\mathbb{U}$ is a machine-learning descriptive subset transferred from $\mathbb{S}$ in psychological domain descriptive;*
- *every p' $\in u$ is assigned from a instance $p \in s$ in Definition 1;*
- *$|\mathbb{D}^s|$ is limited due to the limited functionings defined in psychological domain.*

This research aims to discover an effective classification model $\mathbb{M}$ which provides a reliable mapping of a well-defined $\mathbb{U}$ into $\mathbb{V}$:

$$\mathbb{U} \overset{\mathbb{M}}{\Rightarrow} \mathbb{V} \text{ or } \mathbb{M}(\mathbb{U}) = \mathbb{V}$$

Generally, we can label the users into two classes: depression users and group away of depression. We naturally employ machine learning technique for classification to seek an effective solution. And the binary classification is seen as a supervised learning because the objective is to use machine learning to automatically classify participants into two labelled categories of depression and depression-less.

4 Framework

The Framework is the theoretical structure of research study to describe and explain all level models and classification methods. It comprises several modules that establish a completely detailed structure of research study. Implementation of the Framework is the procedure of research experiment. And it explains how the research problem is analysed and in which content the research problem is solved.

4.1 Conceptual Design

In this study, the framework consists of three modules:

1. Psychological domain knowledge transfer;
2. Data processing;
3. Classification Modelling.

The conceptual design of the framework is illustrated in Fig. 1. Psychological knowledge module learns the knowledge how to group health informatics in psychological domain. It is a guideline to direct the actions how to transform the dataset in data processing module. It also assists designing ensemble classification technique in classification modelling module. Data processing module contains all proceedings of data preprocessing, feature extraction and dataset establishment. The module converts the data from rare health statistics dataset into several normalised dataset being ready for classification. The last modelling module implements the classification of dataset. It builds an effective ensemble classifier and performs the comparative prediction of depressive risk for participants.

4.2 Psychological Knowledge Base

Kroenke et, al. [PHQ-9] discovered that there was a strong association between increasing depression severity screen scores and worsening functionality on all 6 categories: mental, social, role, pain, physical and general functions. These 6 categories were directly interpreted 5 items of mental health diagnostic criteria in Mental Health Inventory (MHI-5) and additionally mental disorder symptoms as mental category. The research illustrated graphically the relationship between increasing PHQ-9 scores of depression and worsening functional categories (see Fig. 2).

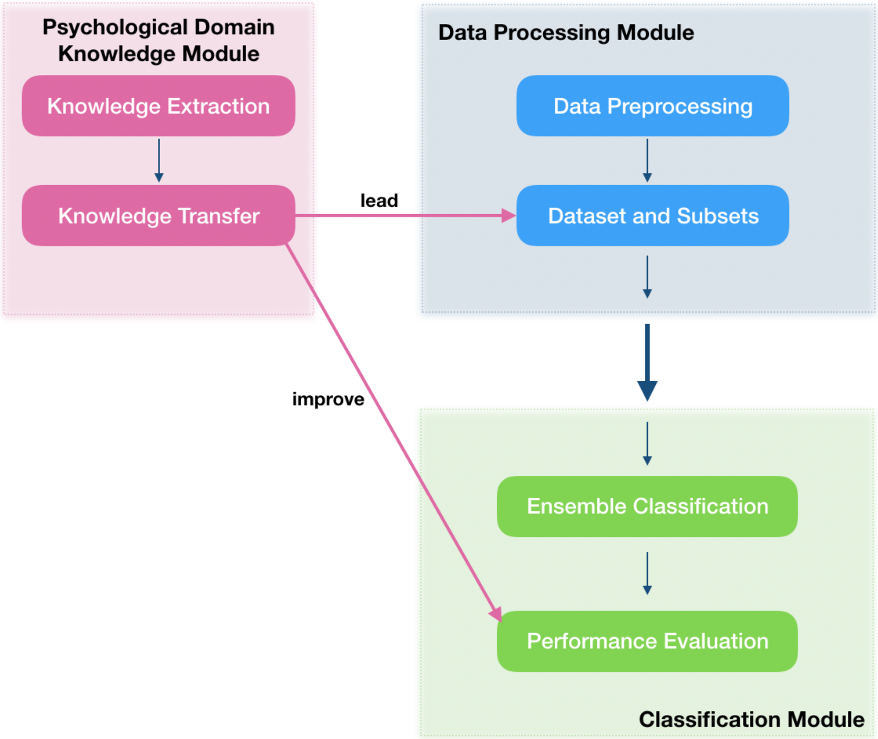


Fig. 1. Conceptual Framework

Associations of health related functionings with depression have been observed in many previous studies at psychological domain. For instance, Clark et, al. [Clark, Bradley] explored the opposite association of depression and psychosocial functionings. The research examined the potential psychosocial benefits of wellness coaching in functionings which included the overall quality of life and the 5 domains of physical, social, emotional, cognitive, and spiritual functioning. It found that depression is associated with poor health status and negative health behaviours. It also addressed that participants significantly reduced their level of depression after improving health functional status by wellness coaching. The researchers suggested that additional self-care on physical activity, health sleep, spirituality and social activities could help on long-term depression management. Ostir et, al. [Ostir] discovered that patients identified as not depressed showed greater improvement in functional status than other patient groups in stoke disease.The research varied previous reports on the association between depression and functional status. It suggested that early recognition and management of depression in person with stroke represents an important effort to improve health outcomes and facilitate functional independence. Moreover, Gonzalez-

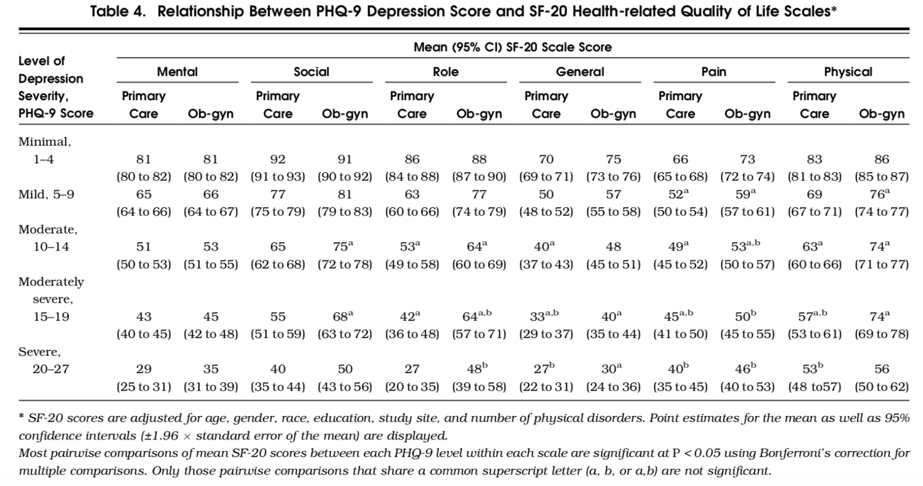


Table 4. Relationship Between PHQ-9 Depression Score and SF-20 Health-related Quality of Life Scales*

Level of Depression Severity, PHQ-9 Score	Mean (95% CI) SF-20 Scale Score											
	Mental		Social		Role		General		Pain		Physical	
	Primary Care	Ob-gyn	Primary Care	Ob-gyn	Primary Care	Ob-gyn	Primary Care	Ob-gyn	Primary Care	Ob-gyn	Primary Care	Ob-gyn
Minimal, 1–4	81 (80 to 82)	81 (80 to 82)	92 (91 to 93)	91 (90 to 92)	86 (84 to 88)	88 (87 to 90)	70 (69 to 71)	75 (73 to 76)	66 (65 to 68)	73 (72 to 74)	83 (81 to 83)	86 (85 to 87)
Mild, 5–9	65 (64 to 66)	66 (64 to 67)	77 (75 to 79)	81 (79 to 83)	63 (60 to 66)	77 (74 to 79)	50 (48 to 52)	57 (55 to 58)	52^{a} (50 to 54)	59^{a} (57 to 61)	69 (67 to 71)	76^{a} (74 to 77)
Moderate, 10–14	51 (50 to 53)	53 (51 to 55)	65 (62 to 68)	75^{a} (72 to 78)	53^{a} (49 to 58)	64^{a} (60 to 69)	40^{a} (37 to 43)	48 (45 to 51)	49^{a} (45 to 52)	53^{a,b} (50 to 57)	63^{a} (60 to 66)	74^{a} (71 to 77)
Moderately severe, 15–19	43 (40 to 45)	45 (42 to 48)	55 (51 to 59)	68^{a} (63 to 72)	42^{a} (36 to 48)	64^{a,b} (57 to 71)	33^{a,b} (29 to 37)	40^{a} (35 to 44)	45^{a,b} (41 to 50)	50^{b} (45 to 55)	57^{a,b} (53 to 61)	74^{a} (69 to 78)
Severe, 20–27	29 (25 to 31)	35 (31 to 39)	40 (35 to 44)	50 (43 to 56)	27 (20 to 35)	48^{b} (39 to 58)	27^{b} (22 to 31)	30^{a} (24 to 36)	40^{b} (35 to 45)	46^{b} (40 to 53)	53^{b} (48 to57)	56 (50 to 62)

** SF-20 scores are adjusted for age, gender, race, education, study site, and number of physical disorders. Point estimates for the mean as well as 95% confidence intervals (±1.96 × standard error of the mean) are displayed.*

Most pairwise comparisons of mean SF-20 scores between each PHQ-9 level within each scale are significant at P < 0.05 using Bonferroni's correction for multiple comparisons. Only those pairwise comparisons that share a common superscript letter (a, b, or a,b) are not significant.

Fig. 2. The relationship between depression severity and personal health-related functionalities[PHQ-9]

Saenz de Tejada et, al. [Gonzalez-Saenz de Tejada] explored the association of functional and psychological status of cancer patients. The study addressed that patients with depression showed lower gains in all health related functional domains than patients without depression. It confirmed again that patients with depression tended to show less improvement in all functional variables in health related quality of life (physical, role, emotional, cognitive, and social function, and global quality of life). And it also confirmed that depression were associated with changes in at least one pre-noted functional variable.

By analysis of the relationship between depression and variables of functional status, the scales of health-related functional variables have the similar trend as the severity of depression in statistics. Previous studies in related-work were more focused on detecting depressive symptoms and depression-related contents. Likewise, this relationship innovates a new potential method of predicting users' depression by sampling various diagnostic criteria of functionality. The classification technique and binary ground truth technique will enhance the strength of new type prediction as well. New method apparently has a couple benefits comparing to previous techniques:

a) There are more features available for classification due to enlarged inputs in various functional areas;
b) It is more easier to acquire functional data than sensitive data of depressive symptoms especially on social network;
c) It is more easier to cover sufficient specificities of one functional status than to cover all available types of depressive symptoms;
d) It is more accuracy and more comparable in the classification of six functional status group than in only one collection of depressive symptoms;

e) It can provide a real opportunity to apply the similar method on automatically detecting depression on social network.

Therefore, we can transfer psychological domain knowledge to information domain. $\mathbb{D}^s$ can be leveraged and divide into 6 sub-datasets. The dataset of user mental profile need to be redefined:

Definition 4 Let new redesigned $\mathbb{U} = \langle u_{mental}, u_{social}, u_{role}, u_{pain}, u_{physical}, u_{general} \rangle$, every u $\in \mathbb{U}$ is an independent function of user, where

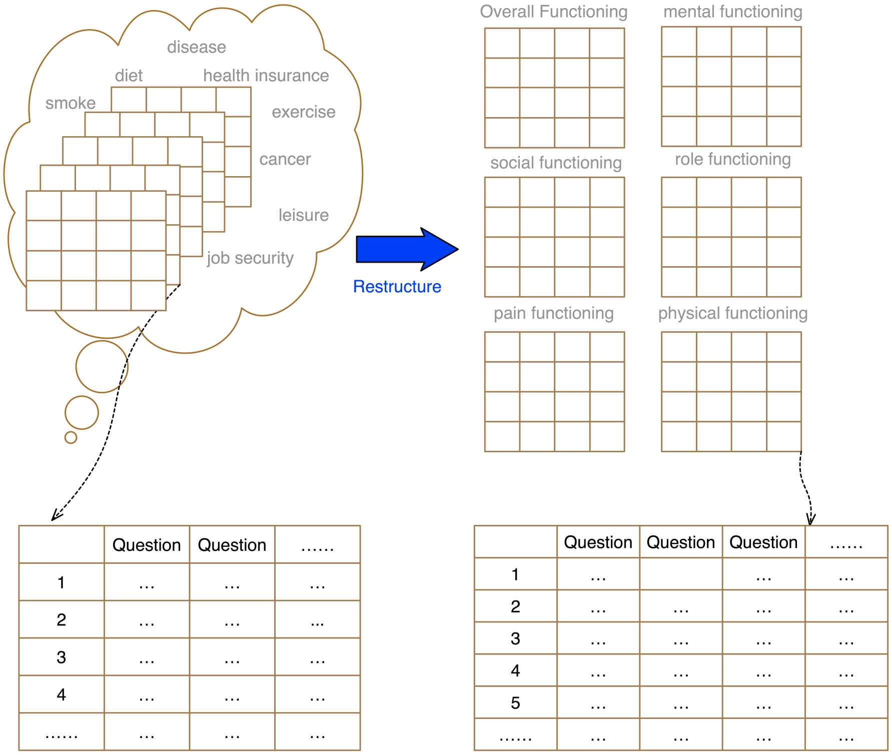
- u_{mental} presents individual mental disorder symptoms;
- u_{social} presents mental diagnostic criteria in the social activities;
- u_{role} presents mental diagnostic criteria in the role funcitonlity;
- u_{pain} presents mental diagnostic criteria in the pain domain ;
- $u_{physical}$ presents mental diagnostic criteria in the physical category;
- $u_{general}$ presents mental diagnostic criteria in the general actions.

4.3 Data Processing

In this research, we use the dataset that was directly collected from national health examination survey. It is generally used for health statistics, but unfortunately not for data mining. And we only use the survey question part which was one third of whole dataset. It was organised by variety of health survey questions which divided questions into columns and participants into rows amongst different tables of health domain. Since those tables were not organised in the same format and structure, the pre-processing of them is hence prominent for later classification.

Data cleaning and transformation is prior in the whole procedure of data preparation because all data should be computer readable and not redundant. And data types in the dataset are justified in order to make each other compatible and comparative. The normalisation is also necessary to uniform the scale condition in various questions. Whilst data preprocessing is implemented, psychological domain knowledge in functional diagnostic criteria is applied in the reconstruction of data structure. According to Definition 4, we can lower the dimension of data set by reducing the number of tables. All tables need to be reconstructed into only six tables referred by six categories of depression diagnostic criteria in functionality (see Fig. 3). They may involve different number of questions but they all have the same participants. Furthermore, those six tables can be rejoin into one big table due to same row index of them. By instant consideration of those tables, each table forms a new dataset where participants are cases and questions are features. We can therefore define the new datasets after data pre-processing as below:

Definition 5 Let new overall dataset of m cases and n features $\mathbb{D}_{overall} = \{(x_1, x_2, ..., x_n, y), x_i \in R^m, y \in \{0,1\}^m\}$, and sub-datasets of different 6 functional categories $\mathbb{D}_{mental}$, $\mathbb{D}_{social}$, $\mathbb{D}_{role}$, $\mathbb{D}_{pain}$, $\mathbb{D}_{physical}$ and $\mathbb{D}_{general}$, where



	Question	Question	......
1	...	...	...
2	...	...	...
3	...	...	...
4	...	...	...
......	...	...	...

	Question	Question	Question	......
1	...		...	...
2	...	...	...	...
3	...	...	...	...
4	...	...	...	...
5	...	...	...	...
......	...	...	...	...

Fig. 3. Data Restructure based on Psychological Knowledge Base

- $|\mathbb{D}_{overall}| = |\mathbb{D}_{mental}| = |\mathbb{D}_{social}| = |\mathbb{D}_{role}| = |\mathbb{D}_{pain}| = |\mathbb{D}_{physical}| = |\mathbb{D}_{general}| = m$;
- Feature($\mathbb{D}_{mental}$) + Feature($\mathbb{D}_{social}$) + Feature($\mathbb{D}_{role}$) + Feature($\mathbb{D}_{pain}$) + Feature($\mathbb{D}_{physical}$) + Feature($\mathbb{D}_{general}$) = Feature($\mathbb{D}_{overall}$) = n .

4.4 Modelling

In this study, we use an ensemble classification approach to build the model for detecting depression. It implements the independent ensemble methodology which applies several classification techniques in parallel. They are Support Vector Machine (SVM) technique, Artificial Neutral Network(ANN) algorithm, K-Nearest Neighbour (KNN) method and Decision Tree (DT) method. Each composite classifier among them is trained on the same portion of the training set in one run. The performance of them are evaluated by k-fold cross validation algorithm. And amalgamating all outputs of composite classifiers into a single prediction, we consequently generates the ensemble classifier. The main idea of this ensemble classification approach is to collect various outputs of multiple independent classifiers and combines them to improve the predictive performance.

In general, the ensemble method provides higher accuracies and better predictive performance than a single algorithm [Rokach]. There are several reasons why ensemble methods having a better performance [Sagi]:

(i) Overfitting avoidance: ensemble methods improve the overall predictive performance by averaging different hypothesis to reduce the risk of choosing an incorrect hypothesis.

(ii) Computational advantage: ensemble methods decrease the risk of obtaining a local minimum by combining several learners, ensemble methods.

(iii) Strong representation: ensemble methods achieve a better fit to the data space due to combining different models and extending the search space.

Moreover, ensemble methods are considered the potential solution for several machine learning challenges like class imbalance, concept drift and curse of dimensionality [Sagi]. For example, Lu, Cheung and Tang [Lu, Cheung] proposed a new ensemble algorithm to utilise both undersampling and oversampling base sampling methods in data training; the proposed method specifically selected various sampling rate for each data set; they also illustrated that the proposed ensemble method significantly outperformed other traditional algorithms for class imbalanced problem. And about concept drift problem, Limsetto and Waiyamai [Limsetto, Waiyamai] considered that it can be solved in multiple ways such as robust classifier, data sampling, semi-supervised learning and cost-based learning. They proposed an ensemble method from many well-known models instead of one that resulting in less bias than previous baseline models; and the experimental results demonstrated that the ensemble model yielded better performance when class distribution of data set was not set uniformly. Furthermore, Serpen and Pathical [Serpen, Pathical] researched how the ensemble method solved curse of dimensionality problem in machine learning; they divided high-dimensional feature space into subspaces and assigned each subspace alongst a base learner within an ensemble machine learning context; their simulation of over 20,000 features indicated that the ensemble classifier had better performance in prediction accuracy and cpu time than other benchmark machine learners. Therefore, ensemble methods are obtained widely to avoid above problems and further improve the overall performance in classification.

The ensemble method also imitates human nature by seeking various solutions before making a final decision [Sagi] and therefore, it becomes a nature option for modelling. The ensemble method hence is considered as a optimised technology comparing to other baseline models in the classification of our preprocessed data.

Baseline Classification Technique

Our ensemble classification method involves several baseline supervised classification models. Supervised classification is one of most frequently applications in predictive data mining. We have concentrated on selecting intricate supervised learning algorithms within diverse advantages. The goal of each classification method is to build a concise model to achieve the best possible prediction accuracy. However, each classification method has diverse computing algorithm. There are several most important supervised machine learning techniques [Kotsiantis]:

a) Logic based algorithm: The algorithms use logic or rules to make a decision of selecting proper features during the learning. Decision tree method adopts this algorithm.
b) Perceptron-based techniques: The algorithms are based on the notion of perceptron to construct a pattern like layers of neutrons to learn different paths in the classification. Neutral network is its well-known representer.
c) Statistical learning algorithms: The algorithm uses statistical approaches to provide a probability that an instance belongs in each class. Under this category of classification algorithms, one can find Naive Bayesian network and k-Nearest Neighbour technique.
d) Support vector machines: Support Vector Machine is the newest supervised machine learning technique [Kotsiantis]. In classic, it uses a hyperplane to separate two data classes and the margin created by the separating hyperplane indicates how the success of classification is.

We propose to involve one method of each type in order to present sufficient algorithms in the limited number of sub-models. We thereby select four techniques for baseline models: Decision Tree method, Artificial Neutral Network technique, k-Nearest Neighbour method and Support Vector Machine algorithm.

Decision Tree (DT)

Decision trees are logic trees that classify instances by evaluating them based on attributes. Each internal node in a decision tree represents evaluating an attribute in an instance, each branch represents the outcome of evaluation and each leaf node represents a class label. Instances are classified starting at the root node and stop at one leaf node after computing all attributes on the path. Decision tree algorithm is the easiest algorithm and capable of classifying huge datasets [Somevanshi]. It is simple to understand and interpret. Kotsiantis et, al. [Kotsiantis] addressed that one of the most useful characteristics of decision trees is their comprehensibility, which makes users can easily understand why a decision tree classifies an instance as a specific class label. And decision tree method can make a decision even with little hard data. Somvanshi et, al. [Somvanshi] believed that decision tree algorithm can process the data which contains the missing values and errors. Their research showed that decision tree is able to work very good in the presence of redundant attributes. Likewise, a disadvantage of decision tree method is well-known. The algorithm is unstable that a small change in the data may change the overall look of decision tree. However, decision tree algorithm is still one of the most useful and powerful algorithm in supervised learning.

Artificial Neutral Network (ANN)
Artificial neutral network is a biologically inspired algorithm to simulate the manner of nerve cells in the brain. According to the book of Kumar [Kumar], ANN is made up of elements named as artificial neurons; the neurons are organised in network to simulate the anatomy of brain by a standard processing whose output is calculated by multiplying its input by a weight vector; they are aggregated into layers and layers are aggregated into the network to form highly interconnected processing structures; whilst the input layer doesn't process information, it simply sends the inputs, modified by a weight, to each of the neurons in the next layer; and the next layer does the processing which can be a hidden layer or the output layer in a single layer design. ANNs are usually more able to easily provide incremental learning than decision trees as having a good multiple layers architecture. Fei and Li [Fei, Li] discovered that ANN are widely used in medical data mining methodology, and the combination of ANNs and some other algorithms will be able to achieve a better results in medical diagnosis and prediction. Likewise, ANN contains some weaknesses, including "poor general application of the architecture, inaccurate analysis for various indicators of the network and uncontrollable time of machine learning" [Fei, Li]. Kotsiantis et, al. [Kotsiantis] concluded that the most striking disadvantage of ANN is lack of ability to answer how the output in a specific way being effectively communicated. Generally, it is a problem to properly determining the size of the hidden layer. The underestimated neutrons in hidden layer can lead to poor approximation, while "excessive nodes can result in overfitting and eventually make the search for the global optimum more difficult" [Kotsiantis]. In spite of its several disadvantages, ANN is still a good competitor for other learning algorithms, which has been used on a variety of intricate problems including computer vision, speech recognition, recommendation filtering even medical diagnosis [Somevanshi][Kumar].

K-Nearest Neighbour (KNN)
Conversely to intricate neutral networks, the K-Nearest Neighbor algorithm is a typical lazy learning algorithms. KNN is based on the principle that classifying instances is to find other similar instances that have proximate properties. "If the instances are tagged with a classification label, then the value of the label of an unclassified instance can be determined by observing the class of its nearest neighbours" [Kotsiantis]. In KNN assigning weight by the contributions of the neighbours, the nearer neighbours thereby contribute more to the average than other distant ones. This algorithm can be used for both classification and regression. And it is among the simplest machine learning algorithms, even no explicit training step is required. KNN has somewhat weaknesses in computational time and classification accuracy. Though KNN is very sensitive to the choice of the similarity function that is used to compare the contribution of neighbours [Kotsiantis], it is still a popular classification technique.

Support Vector Machine (SVM)

Support Vector Machine algorithm is the newest classification technique among the proposed methods. In classification, SVM constructs a hyperplane or set of hyperplanes in the dimensional space; the hyperplane separates the training data into diverse two classes; and a good classification is achieved by the hyperplane's capability to make a larger margin between two classes of training data. As the application environment is dimensional space, input data of SVM are paired into vectors and vectors are defined in terms of a kernel function. Selection of proper hyperplane and proper parameters for kernel function gives more accurate results as compared to neural networks [Somevanshi]. As the model complexity of an SVM is unaffected by the number of features encountered in the training data, SVM is well suited to deal with learning dataset with large number of features and training instances. Also, choice of an appropriate kernel leads to different SVM applications in linear, nonlinear and multiclass classification. The potential drawbacks of SVM are addressed including [Kotsiantis]: a) requiring full label of input data; b) being difficult to interpret parameters of the solved model; c) being unsuitable for non-binary multiple classification problems.

Meanwhile, Choudhary and Gianey [Choudhary, Gianey] stated that every learning algorithm differs according to area of application and no algorithm is more powerful than the other in all scenarios. They concluded that the choice of a suitable algorithm depends on the type of problem and the given data, and the accuracy can be improved by using two or more algorithms together. We therefore comprise above four algorithms into ensemble model for this study.

Ensemble Model

After a better understand of the strengths and limitations of each model, the ensemble of integrating four algorithms together is possible to maximum the predictive performance. ”The objective is to utilise the strengths of one method to complement the weaknesses of another” [Kotsiantis]. While more specifically each independent sub-model is trained, more targeted concepts are covered by the ensemble classifier and more accuracy it becomes.

In order to combine all baseline classifiers' outputs, our modelling procedure adopts weighting ensemble method. Weighting ensemble method is very genetic when all base classifiers have uniform comparable outputs. The weight of each classifier can be set proportional to its accuracy performance on a validation set [Rokach]:

$$w_i = \frac{1 - E_i}{\sum_{k=1}^{n}(1 - E_k)} \tag{1}$$

where E_i is a normalisation factor which is based on the predictive performance of classifier i on the validation set.

In view of the fact that the ensemble classifier combines weighted outputs of all base classifiers, we can define the ensemble classifier as below:

Definition 6 Let the ensemble model

$$\mathbb{M}_e = \sum_{k=1}^{n} w_i M_i \tag{2}$$

where

- M_i presents a single base model;
- w_i presents the weighting metric of predictive performance at specific base model M_i;
- k is the order of base models;
- n is the total number of base models, and in our case $n = 4$;
- i is the order number of specific base model.

In this ensemble method, the driving principle is to build a couple of estimators independently and then to average their predictions. The combined estimator is usually better than any of the single base estimator because instances' variance is moderated.

Algorithm

Given a well-preprocessed dataset of m examples and n features $\mathbb{D} = \{(x_1, x_2, ..., x_n, y), x_i \in R^m, y \in \{0, 1\}^m\}$, we can generate a suitable ensemble model $\mathbb{M}_e$ to present a mapping of $\{x_1, x_2, ..., x_n\}$ to $\{y\}$ by applying h various types of baseline model M_i:

```
input  : Dataset $\mathbb{D} = \{(x_1, x_2, ..., x_n, y), x_i \in R^m, y \in \{0, 1\}^m\}$
output: Ensemble Model $\mathbb{M}_e$
Set the training set as $\mathbb{R} = \{(x_1, x_2, ..., x_n), x_i \in R^m\}$, and the testing
  set as $\mathbb{S} = \{y, y \in \{0, 1\}^m\}$;
for $i \leftarrow 1$ to $h$ do
    /* validate baseline model */
    Do training $M_i$ on the training set $\mathbb{R}$ ;
    Get the performance $E_i$ while validating the training result on $\mathbb{S}$ ;
end
Calculate $w_i = \frac{1-E_i}{\sum_{k=1}^{h}(1-E_k)}$; /* calculate performance weightings */
Obtain the ensemble model $\mathbb{M}_e = \sum_{g=1}^{h} w_i M_i$;
```

Algorithm 1: Ensemble Modelling

5 Experiment

We employ an ensemble supervised learning experiment to classify depressive users from a rare health survey dataset $\mathbb{H}$. We follow psychological knowledge to reduce the dimension of dataset by split dataset into sub-sets. It will not only benefit the processing of classification but also provide a great opportunity to compare the performance of overall dataset and subsets for support of further solution on the real condition with less features.

5.1 Experiment Design

In experiment, we first obtain dataset $\mathbb{D}_{overall}$ by data preprocessing on survey data $\mathbb{H}$; next, we aggregate all features of $\mathbb{D}_{overall}$ into 6 health-related functional classes and follow the same procedure to divide $\mathbb{D}_{overall}$ into 6 sub-sets $\mathbb{D}_{physical}$, $\mathbb{D}_{role}$, $\mathbb{D}_{mental}$, $\mathbb{D}_{social}$, $\mathbb{D}_{pain}$ and $\mathbb{D}_{general}$; and we train dataset $\mathbb{D}_{overall}$ by four baseline models (DT, ANN, KNN, SVM) to obtain the relevant performances; then we build the ensemble model $\mathbb{M}_e$ by calculating the performance weight w_i of each baseline model M_i; furthermore, we train all 6 sub-datasets by the ensemble classifier $\mathbb{M}_e$; and the final step is to use k-fold cross validation algorithm to value the complete predictive performance. The overall look of all experiment proceedings is illustrated in Fig. 4.

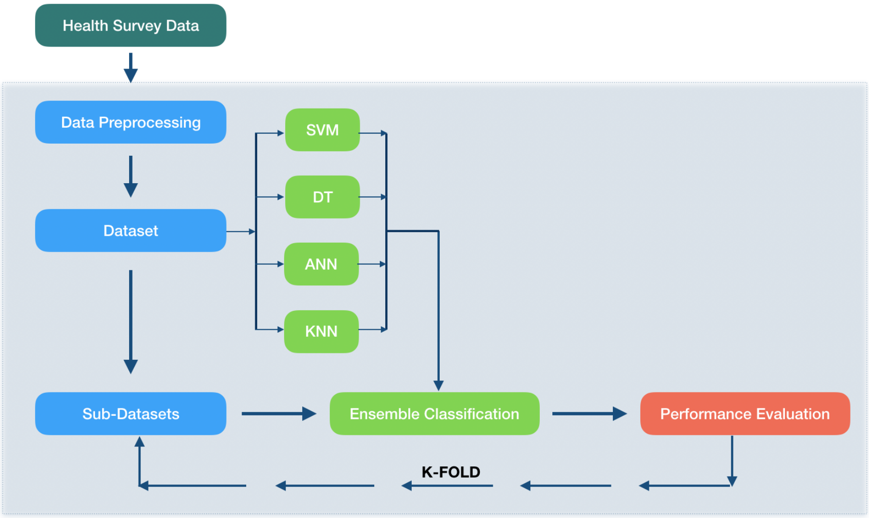


Fig. 4. The overall look of experiment proceedings

From the proceeding details of classification, we can define the algorithm of whole experiment as below:

```
input  : a rare health survey dataset $\mathbb{H}$
output: Ensemble Classifier $\mathbb{F}_e$ and the complete prediction
Obtain dataset $\mathbb{D}_{overall}$ by data pre-processing on survey data $\mathbb{H}$;
Aggregate features manually referred on 6 psychological functionalities ;
Divide $\mathbb{D}_{overall}$ into $\{\mathbb{D}_{physical}, \mathbb{D}_{role}, \mathbb{D}_{mental}, \mathbb{D}_{social}, \mathbb{D}_{pain}, \mathbb{D}_{general}\}$;
Supervised learning on $\mathbb{D}_{overall}$ for ensemble model $\mathbb{M}_e = \sum_{g=1}^{h} w_i M_i$;
foreach sub-dataset $\mathbb{D}_i$ in $\{\mathbb{D}_{overall}, \mathbb{D}_{physical}, \mathbb{D}_{role}, \mathbb{D}_{mental}, \mathbb{D}_{social}, \mathbb{D}_{pain}, \mathbb{D}_{general}\}$ do
    /* ensemble classification*/
    Do ensemble classification on $\mathbb{D}_i$ ;
    Validate its predictive performance;
end
```

Algorithm 2: Experiment Design

The ensemble classification can be expressed in algorithm as well: Given a well-preprocessed dataset of m examples and n features $\mathbb{D} = \left\{(x_1, x_2, ..., x_n, y), x_i \in R^m, y \in \{0,1\}^m\right\}$, we can obtain the ensemble classifier $\mathbb{F}_e = w_{svm} \cdot f_{svm} +$

$w_{nb} \cdot f_{nb} + w_{knn} \cdot f_{knn} + w_{dt} \cdot f_{dt}$ by applying supervised learning on dataset $\mathbb{D}$:

```
input  : Dataset $\mathbb{D} = \big\{(x_1, x_2, ..., x_n, y), x_i \in R^m, y \in \{0,1\}^m\big\}$
output: the optimised ensemble classifier $\mathbb{F}_e$ and its predictive
        performance $p_e$
Divide dataset $\mathbb{D}$ into k portions, each portion has $\frac{m}{k}$ examples;
for $k \leftarrow 1$ to $5$ do
    Select all portions except $k^{th}$ portion to form new dataset $\mathbb{D}^{'}$ ;
    Use $\mathbb{D}^{'}$ to generate the training set $\mathbb{R} = \big\{(x_1, x_2, ..., x_n)\big\}$ and the
      testing set $\mathbb{S} = \big\{y\big\}$, where $|\mathbb{D}^{'}| = |\mathbb{R}| = |\mathbb{S}| = \frac{4}{5}|\mathbb{D}| = \frac{4m}{5}$;
    /* baseline model */
    foreach one classification method of SVM, ANN, KNN, DT do
        Training on the training set $\mathbb{R}$ and obtain classifier $f$;
        Obtain predictive value $y^p = f(\sum_{i=1}^{n}(x_i))$ ;
    end
    /* ensemble */
    Calculate the ensemble classifier $\mathbb{F}_k = w_{svm} \cdot f_{svm} + w_{nb} \cdot f_{nb} +$
      $w_{knn} \cdot f_{knn} + w_{dt} \cdot f_{dt}$ ;
    Calculate a float predictive value $y_e = w_{svm} \cdot y^p_{svm} + w_{nb} \cdot y^p_{nb} +$
      $w_{knn} \cdot y^p_{knn} + w_{dt} \cdot y^p_{dt}$ ;
    /* sensitivity */
    if $y_e > 0.5$ then
        $y_e = 1$;
    else
        $y_e = 0$;
    end
    Test $y_e$ on the testing set $\mathbb{S}$ and report predictive performance $p_k$ ;
end
/* 5-fold cross validation */
Validate the predictive performance by calculating $p_e = \frac{\sum_{k=1}^{5} p_k}{5}$ ;
Generate the optimised ensemble classifier $\mathbb{F}_e =$
  Median$(\mathbb{F}_1, \mathbb{F}_2, \mathbb{F}_3, \mathbb{F}_4, \mathbb{F}_5)$
```

Algorithm 3: Ensemble Classification Procedure

5.2 Dataset

NHANES Survey Data

In this study, we use dataset of National Health and Nutrition Examination Survey (NHANES). NHANES is a population-based survey designed to collect health-related information of the U.S. household population. It is a very rich resource for health professionals and researchers to expand our knowledges of various modern health problems. It is conducted by the National Center for Health Statistics (NCHS), which is part of the Centers for Disease Control and

Prevention (CDC). All information in NHANES are gathered and protected with the requirement of Federal Law of U.S. and for health research purposes only. Collections of NHANES in last decade are all free for researchers and published on the website of NCHS.

We employ the questionnaire data in NHANES 2013 - 2014 collection as input data $\mathbb{H}$ of experiment. We also limit the age of participants to 18+ because data of teenage and children are only partially published. As our objective is to classify general person into healthy and depressive groups, the features only involved with single gender are excluded.

Build Ground Truth

NHANES integrates health tools for measuring health status like Patient Health Questionnaire (PHQ-9) depression screen tool. PHQ-9 tool is a 9-item screening instrument to measure depressive severity from no depression to major depressive disorder. In NHANES, PHQ-9 measure is the only integrated measurement for depression because it is a simple, reliable and valid measure of depression severity [PHQ-9]. And it has been a useful clinical and research tool in years. There are plenty of health researches assigning with NHANES data and integrated PHQ-9 tool to study depression related health issues. For instance, Stuart et, al. [Stuart] in 2011 researched the relationship between depression and low cholesterol among household population using NHANES data; Alison et, al. [Alison] in 2012 proposed a association between major depressive disorder and obesity by assessing 2001-2004 NHANES collections; Ubani and Zhang [Ubani] in 2015 published a research of NHANES data to study the role of adiposity in the relationship between serum leptin and severe major depressive episode; and Andrea et, al. [Andrea] in 2016 explored depressed adults information in social support and health service use of NHANES data; Nguyen et, al. [Nguyen] in 2017 research the association between blood folate concentrations and depression in reproductive aged U.S. women in NHANES 2011 - 2012 collection.

Based on the integrated PHQ-9 screen measurement, we can establish ground-truth label information (on whether or not participant has depression) for whole dataset. In scales of PHQ-9 measurement, there are five level of depression severity from minimal level to severe level. In the research of Kroenke et, al. [PHQ-9], they found that patients who were identified at least on the moderate level (score ≥ 10) of depression in PHQ-9 measurement had a sensitivity of 88% and a specificity of 88% for major depression. We thereby choose the separation at PHQ-9 score 10. Participant who has a PHQ-9 score less than 10 is considered as a healthy person of depression or vice versa. We label these depression-less people as the logical truth or "1"; reversely those depressive people as the logical false or "0".

Principles Of Data Preprocessing

In spite of the fact that NHANES questionnaire collection was very organised and carefully preserved, it still existed some errors and missing values. And naturally part of participants did not complete all questions in the questionnaire. Furthermore, the questionnaire involves "Refuse" option and "Don't Know" option for nearly every question, because the design of it toke a very cautious consideration of personal privacy and individual interests. It is hence essential to fill, correct and normalise those meaningless inputs. In order to uniform all actions taken in data cleaning, we design a couple of presumption and principles to manage the proceeding:

a) we assume that missing inputs belong to the persons who have on depressive risk;
b) the choice of "Refuse" option or "Don't Know" option is presumed normal which can be corrected by the statistical mean of inputs;
c) all inputs of survey questions should be converted into binary, range and numbers due to the design of answer options;
d) the final value of each input should be normalised and have a limited byte size.

The Overall Dataset

After data preprocessing, we has an overall dataset that involves 5398 participants with 516 (9.56%) depressive persons and 4882 (90.44%) depression-less people among. The features are directly converted from the original major questions in the health survey, which means a major question of NHANES simply presents one feature in our dataset. After rejecting several irrelevant questions that limited by the age or gender, we get a total of 98 features. Among them, inputs in 49 feature are binary, 36 features are range data and the rest 14 features are float numbers.

5.3 Baseline Models

Many machine learning packages and tools are accessible to implement common classification algorithms. Scikit-learn library from Python is one of the most well-designed machine learning package. It provides simple and efficient tools for data mining and data analysis. And it nearly contains all supervised learning methods for both binary and multi-class classification. We thereby choose scikit-learn Python package to implement four baseline models.

Kernel and Parameters How to select suitable kernel and parameters is common task for classification but it is also complex for specific examples. We only balance the settings of baseline models instead searching a perfect for the parameter because it is uncertain if the settings could maximum the performance in utter instances. And the predictive performance is expected being improved by ensemble classification. Therefore, we employ common values for kernel and parameters in four sub-models:

I) Decision Tree:

1) scikit-learn uses an optimised version of the CART decision tree algorithm;
2) the depth of the tree is limited at maximum of 3 in order to prevent overfitting;
3) m̈in_samples_leaf=1ïs often the best choice due to few classes in classification;
4) the criteria of the quality of split is default as the Gini impurity;
5) sample weight is preferred by normalising the sum of the sample weights to prevent the tree from being biased toward the dominant class.

II) Artificial Neutral Network:

1) ANN model implements a multi-layer perception algorithm within back-propagation training method;
2) there are three hidden layers in the proposed network and each of them exists 30 neutrons;
3) Activation function for the hidden layer is the rectified linear unit function;
4) Adam stochastic gradient-based optimiser is preferred due to relatively large amount of training samples;
5) the learning is initialised from a small step-size 0.001 for a reasonably big learning rate;
6) maximum number of iterations is set as 10000 to prevent run out of memory while doing convergence;
7) the seed is randomly generated by random state computer generator to achieve the maximum of variability.

III) K-Nearest Neighbour:

1) we use the basic nearest neighbours classifier with k is an integer 5;
2) the contribution of neighbours is treated equal so the weights between neighbours is set as ”uniform”;
3) the algorithm used to compute the nearest neighbours is set as äutob̈ecause the number of features is varied in overall dataset and sub-datasets.
4) the size of leaf is default as 30 which may affect the speed of construction and query;
5) the default metric is Euclidean due to standard metric configuration.

IV) Support Vector Machine:

1) a popular Radial Basis Function (RBF) kernel is used in SVM sets;
2) the parameter C̈ẗhat trades off misclassification of training examples against simplicity of the decision surface is set as 1;
3) the parameter g̈ammad̈efines how much influence a single training example has is set as ”0.1”.

Moreover, all four base models are configured for binary classification and their predictive performances are weighted in both labelled classes.

5.4 Performance Measure

The predictive performance of each base classifier in our model is evaluated by F1 score which is generated on confusion matrix of validation. In confusion matrix, we simply let the number of real mental healthy cases in the training set as **condition positive (P)** and let the number of real depressive cases in the training set as **condition negative (N)**. And four derivations are defined to make it clear if the classifier is confusing binary classes (P and N):

(i) let the number of cases where the classifier correctly predicts P as **true positive (TP)**
(ii) let the number of cases where the classifier correctly predicts N as **true negative (TN)**
(iii) let the number of cases where the classifier incorrectly predicts P as **false positive (FP)**
(iv) let the number of cases where the classifier incorrectly predicts N as **false negative (FN)**

And four following terminologies are hence defined as below:

a) **Recall or True Positive Rate (TPR)**:

$$TPR = \frac{TP}{TP + FN}$$

b) **Specificity or True Negative Rate (TNR)**:

$$TNR = \frac{TN}{TN + FP}$$

c) **Precision or Positive Predictive Value (PPV)**:

$$PPV = \frac{TP}{TP + FP}$$

d) **Accuracy (ACC)**:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Based on these terminologies, F1 score is a balanced measure of both the precision (PPV) and the recall (TPR) of the validation:

$$F1 = \frac{2}{\frac{1}{TPR} + \frac{1}{PPV}} = \frac{2TP}{2TP + FP + FN} \tag{3}$$

Referred to normalisation factor E in equation (1), F1 score need to be normalised first and then is applied to equation (1) to calculate the weight fo each base classifier.

6 Results and Discussions

6.1 Experimental Results

Sub-Model Performance By comparing the performances of four base models (illustrated in table. 1), Decision tree algorithm has the highest accuracy and F1 measure. SVM technique has a similar performance as decision tree algorithm but only a bit lagging in precision. ANN method performs more advantageous in precision score and KNN method operates more fitting in recall measure. Almost F1 scores and recall measures are significant in correlation ($\geq 95\%$). However, baseline models perform less satisfied in accuracy and precision.

Models	Accuracy (mean)	Precision (mean)	Recall (mean)	F1 score (mean)
SVM	0.921	0.927	0.990	**0.958**
ANN	0.905	**0.941**	0.955	0.948
KNN	0.908	0.913	**0.993**	0.951
DT	**0.925**	0.938	0.981	**0.959**

Table 1. Performances of sub-models in the overall dataset

Ensemble Classifier We use F1 measure for the main indicator of model's performance. According to equation (1) and (2), we can calculate the weight for each base model (see at table 2.) and further generate the complete form of ensemble classifier:

$$\mathbb{F}_k = 0.228 \cdot f_{svm} + 0.283 \cdot f_{nb} + 0.266 \cdot f_{knn} + 0.223 \cdot f_{dt} \tag{4}$$

Models	F1 score (mean)	1 - F1	Weight
SVM	0.958	0.042	0.228
ANN	0.948	0.052	**0.283**
KNN	0.951	0.049	0.266
DT	0.959	0.041	0.223

Table 2. Calculation of weights for sub-models

We use F1 score and accuracy measure as the main predication indicators of the ensemble classification. As features and specificity of the overall dataset and each sub-datasets varied, the divided performances are expected (see table. 3). Unsurprisingly, ensemble classifier performs better in the overall dataset and mental sub-set. Performances in other sub-sets is compromised in this experiment but is still comparable to other machine learning methodologies [Fatima][Hassan][Peng][Reece]. As DT performs best (0.925 in Accuracy) in the sub-models and ensemble model has a overall Accuracy of 0.954, the improvement I for detecting depression in ensemble classifier is calculated as below (let percentage of depressed instances as N_0):

$$I = \frac{Accuracy_{new} - Accuracy_{old}}{N_0} = \frac{0.954 - 0.925}{9.56\%} \cdot 100\% = 30.3\% \tag{5}$$

Therefore, the proposed ensemble approach increases the predictive performance by 30.3% for distinguishing depression from non-depressed participants.

Dataset	Features	F1 score	Accuracy
$\mathbb{D}_{overall}$	98	**0.976**	**0.954**
$\mathbb{D}_{physical}$	7	0.964	0.931
$\mathbb{D}_{role}$	9	0.963	0.929
$\mathbb{D}_{social}$	6	0.964	0.931
$\mathbb{D}_{mental}$	4	**0.975**	**0.953**
$\mathbb{D}_{pain}$	2	0.961	0.925
$\mathbb{D}_{general}$	70	0.964	0.931

Table 3. Features and performances of ensemble classifier

Predictive Performance

F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. F1 measure equally considers both precision and recall in the performance measurement. By the comparison of F1 measures in both different algorithms and datasets (see table. 4), ensemble method has a better F1 score than other baseline algorithms. The overall dataset and mental subset enable to be distinguished from other datasets due to a higher level of F1 measures.

Models	Overall (mean)	Physical (mean)	Role (mean)	Social (mean)	Mental (mean)	Pain (mean)	General (mean)
SVM	0.9577	0.950	0.950	0.950	0.957	0.950	0.951
ANN	0.948	0.944	0.935	0.942	0.961	0.950	0.930
KNN	0.951	0.947	0.945	0.944	0.958	0.938	0.949
DT	0.959	0.950	0.949	0.950	0.960	0.950	0.950
Ensemble	**0.976**	0.964	0.963	0.964	**0.975**	0.961	0.964

Table 4. Performance in F1 score

Accuracy is the fraction of predictions our model got right. It indicates the number of correct predictions made in all occurrences of both labels.

Models	**Overall (mean)**	**Physical (mean)**	**Role (mean)**	**Social (mean)**	**Mental (mean)**	**Pain (mean)**	**General (mean)**
SVM	0.921	0.904	0.904	0.904	0.919	0.904	0.907
ANN	0.905	0.895	0.879	0.892	0.928	0.904	0.873
KNN	0.908	0.900	0.896	0.895	0.923	0.886	0.904
DT	0.924	0.905	0.904	0.905	0.926	0.904	0.906
Ensemble	**0.954**	0.931	0.929	0.931	0.953	0.925	0.931

Table 5. Performances of Accuracy

Precision is the ability of a classifier not to label an instance positive that is actually negative. It measures how effective to diagnose person's psychological health. For the performances of precision, ensemble classifier has the best precision in each dataset and it surprisedly performs better in metal sub-dataset than in the overall dataset. It may approve that features for metal disorder symptoms is more depression-related than other features in other categories because non-criteria items in depression scale decreased specificity of performance [Zimmerman].

Models	**Overall (mean)**	**Physical (mean)**	**Role (mean)**	**Social (mean)**	**Mental (mean)**	**Pain (mean)**	**General (mean)**
SVM	0.927	0.904	0.904	0.904	0.921	0.904	0.907
ANN	0.941	0.919	0.911	0.917	0.939	0.904	0.924
KNN	0.913	0.915	0.907	0.913	0.938	0.913	0.908
DT	0.938	0.910	0.906	0.910	0.939	0.904	0.912
Ensemble	0.956	0.934	0.929	0.931	**0.960**	0.925	0.930

Table 6. Performances of Precision

Recall is the ability of a classifier to find all positive instances. It measures how many healthy people are correctly identified. For ensemble classifier, nearly no person with depression is diagnosed into the health group. The recall per-

Models	**Overall (mean)**	**Physical (mean)**	**Role (mean)**	**Social (mean)**	**Mental (mean)**	**Pain (mean)**	**General (mean)**
SVM	0.990	1.000	1.000	1.000	0.996	1.000	0.999
ANN	0.955	0.970	0.961	0.969	0.985	1.000	0.937
KNN	0.993	0.981	0.986	0.977	0.980	0.968	0.994
DT	0.9814	0.992	0.997	0.993	0.982	1.000	0.992
Ensemble	**1.000**	1.000	1.000	1.000	0.999	1.000	1.000

Table 7. Performances of Recall

formances of ensemble classifier is about 1 in the overall dataset and all sub-datasets. According to the definition of recall measure $Recall = \frac{TP}{TP+FN}$, it means that only when false negative measurement (FN) is 0, recall measure is equal to 1. In our experiment, FN presents the number of depressed users who were incorrectly identified as non-depressed. As FN is zero, it indicates that no depressed instances in the experiment has been mistakenly classified. The coverage in correct classification of depressed participants is perfect, only slightly larger than the results of psychological screening (illustrated in Fig. 5). Let the predicted precision as $Precision_p$ and percentage of non-depressed instances as N_1, the overall prediction $p_{overall}$ of depressed instances is calculated as below:

$$p_{overall} = 1 - (Precision_p \cdot N_1) = (1 - 0.956 \cdot 90.44\%) \cdot 100\% = 13.54\% \quad (6)$$

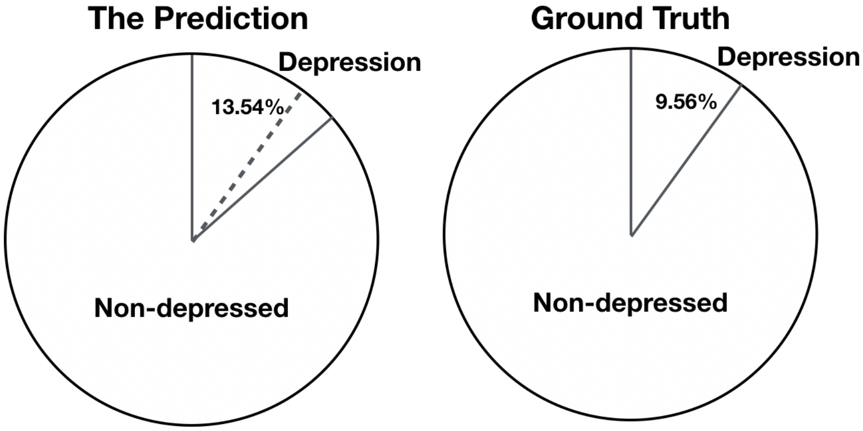


Fig. 5. Coverage in correct classification of depression

6.2 Discussions

performance of ensemble is great Ensemble classifier is obviously superior than baseline models as performing advantageous at F1 measure, Accuracy, Precision and Recall (see table. 4. 5. 6. 7.). It leads not only in the test of overall dataset but also all experiment in sub-datasets. It triumphantly gathers different predictions of baseline models and combine them into a better prediction. It is more stable and robust than any involved baseline algorithm. And this experiment uses random under-sampling technique with ensemble method to leverage the class imbalance problem where non-depression instances is about 10 times large as depressed instances. The propose ensemble method has significantly improve predictive performance by 30.3% with class imbalance. It enables to promote diversity among baseline models and convert that specificity into the performance. The ensemble method is very simple, closing to bagging and major voting ensemble methods. Other boost ensemble methods are also suggested to improve the prediction performance further like the EUSBoost method [Sagi].

the coverage of depression is greate By analysis of the performance in Recall measure (see table. 7), the preferred ensemble method covers all depressed cases in PHQ-9 screening measurement where no depressed instance has been mistakenly labelled as non-depression. The coverage (see fig. 5) of depression cases is slightly larger than the real situation of mental health inventory. However, it is absolutely acceptable for large sampling that there is no missing of depression case and only about 4% of total cases have been incorrectly labelled as depression in the prediction. The proposed ensemble method is perfect for preliminary screening of major depressive disorder in order to provide limited cases for further clinical diagnosis while without missing any potential depression case.

enlarge the area for collecting data, more features

7 Sensitivity Study

SVM: A low C makes the decision surface smooth, while a high C aims at classifying all training examples correctly. The larger gamma is, the closer other examples must be to be affected. Proper choice of C and gamma is critical to the SVM's performance. One is advised to use C and gamma spaced exponentially far apart to choose good values.

The behaviour of the model is very sensitive to the gamma parameter. If gamma is too large, the radius of the area of influence of the support vectors only includes the support vector itself and no amount of regularisation with C will be able to prevent overfitting.

When gamma is very small, the model is too constrained and cannot capture the complexity or "shape" of the data. The region of influence of any selected support vector would include the whole training set. The resulting model will behave similarly to a linear model with a set of hyperplanes that separate the centres of high density of any pair of two classes.

KNN: The optimal choice of the value is highly data-dependent: in general a larger suppresses the effects of noise, but makes the classification boundaries less distinct. The basic nearest neighbours classification uses uniform weights: that is, the value assigned to a query point is computed from a simple majority vote of the nearest neighbours. Under some circumstances, it is better to weight the neighbours such that nearer neighbours contribute more to the fit. This can be accomplished through the weights keyword. The default value, weights = 'uniform', assigns uniform weights to each neighbour. weights = 'distance' assigns weights proportional to the inverse of the distance from the query point. Alternatively, a user-defined function of the distance can be supplied to compute the weights.

8 Conclusion and Future Work

References

1 Andrea, S.B., Siegel, S.A.R., and Teo, A.R.: ‘Social Support and Health Service Use in Depressed Adults: Findings From the National Health and Nutrition Examination Survey’, General Hospital Psychiatry, 2016, 39, pp. 73-79

2 Choudhary, R., and Gianey, H.K.: ‘Comprehensive Review On Supervised Machine Learning Algorithms’, in Editor (Ed.)ˆ(Eds.): ‘Book Comprehensive Review On Supervised Machine Learning Algorithms’ (2017, edn.), pp. 37-43

3 Clark, M.M., Bradley, K.L., Jenkins, S.M., Mettler, E.A., Larson, B.G., Preston, H.R., Liesinger, J.T., Werneburg, B.L., Hagen, P.T., Harris, A.M., Riley, B.A., Olsen, K.D., and Vickers Douglas, K.S.: ‘The Effectiveness of Wellness Coaching for Improving Quality of Life’, Mayo Clinic Proceedings, 2014, 89, (11), pp. 1537-1544

4 De Choudhury, M., Counts, S., and Horvitz, E.: ‘Social media as a measurement tool of depression in populations’, 2013, pp. 47-56

5 Fatima, I., Mukhtar, H., Ahmad, H.F., and Rajpoot, K.: ‘Analysis of user-generated content from online social communities to characterise and predict depression degree’, Journal of Information Science, 2018, 44, (5), pp. 683-695

6 Fei, Y., and Li, W.-q.: ‘Improve artificial neural network for medical analysis, diagnosis and prediction’, Journal of Critical Care, 2017, 40, pp. 293

7 Gonzalez-Saenz de Tejada, M., Bilbao, A., Baré, M., Briones, E., Sarasqueta, C., Quintana, J.M., Escobar, A., and Baré, M.: ‘Association of social support, functional status, and psychological variables with changes in health-related quality of life outcomes in patients with colorectal cancer’, Psycho-Oncology, 2016, 25, (8), pp. 891-897

8 Hassan, A.U., Hussain, J., Hussain, M., Sadiq, M., and Lee, S.: ‘Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression’, in Editor (Ed.)ˆ(Eds.): ‘Book Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression’ (IEEE, 2017, edn.), pp. 138-140

9 Huerta-Ramírez, R., Bertsch, J., Cabello, M., Roca, M., Haro, J.M., and Ayuso-Mateos, J.L.: ‘Diagnosis delay in first episodes of major depression: A study of primary care patients in Spain’, Journal of Affective Disorders, 2013, 150, (3), pp. 1247-1250

10 Kotsiantis, S.B., Zaharakis, I.D., and Pintelas, P.E.: ‘Machine learning: a review of classification and combining techniques’, Artificial Intelligence Review, 2006, 26, (3), pp. 159-190

11 Kroenke, K., Spitzer, R.L., and Williams, J.B.: ‘The PHQ-9: validity of a brief depression severity measure’, J Gen Intern Med, 2001, 16, (9), pp. 606-613

12 Kumar, R.: ‘Fundamental of artificial neural network and fuzzy logic’ (University Science Press, An Imprint of Laxmi Publications Pvt. Ltd., 2010. 2010)

13 Limsetto, N., and Waiyamai, K.: ‘Handling Concept Drift via Ensemble and Class Distribution Estimation Technique’, in Editor (Ed.)^(Eds.): ‘Book Handling Concept Drift via Ensemble and Class Distribution Estimation Technique’ (Springer Berlin Heidelberg, 2011, edn.), pp. 13-26

14 Lu, Y., Cheung, Y.M., and Tang, Y.Y.: ‘Hybrid sampling with bagging for class imbalance learning’, in Editor (Ed.)^(Eds.): ‘Book Hybrid sampling with bagging for class imbalance learning’ (Springer Verlag, 2016, edn.), pp. 14-26

15 Merikangas, A., Mendola, P., Pastor, P., Reuben, C., and Cleary, S.: ‘The association between major depressive disorder and obesity in US adolescents: results from the 2001–2004 National Health and Nutrition Examination Survey’, Journal of Behavioral Medicine, 2012, 35, (2), pp. 149-154

16 Mowery, D., Smith, H., Cheney, T., Stoddard, G., Coppersmith, G., Bryan, C., and Conway, M.: ‘Understanding Depressive Symptoms and Psychosocial Stressors on Twitter: A Corpus-Based Study’, Journal of Medical Internet Research, 2017, 19, (2)

17 Nguyen, B., Weiss, P., Beydoun, H., and Kancherla, V.: ‘Association between blood folate concentrations and depression in reproductive aged U.S. women, NHANES (2011–2012)’, Journal of Affective Disorders, 2017, 223, pp. 209-217

18 Ophir, Y., Asterhan, C.S.C., and Schwarz, B.B.: ‘Unfolding the notes from the walls: Adolescents’ depression manifestations on Facebook’, Computers in Human Behavior, 2017, 72, pp. 96-107

19 Ostir, G.V., Berges, I.M., Ottenbacher, A., and Ottenbacher, K.J.: ‘Patterns of change in depression after stroke’, J Am Geriatr Soc, 2011, 59, (2), pp. 314-320

20 Peng, Z., Hu, Q., and Dang, J.: ‘Multi-kernel SVM based depression recognition using social media data’, International Journal of Machine Learning and Cybernetics, 2017

21 Reece, A.G., and Danforth, C.M.: ‘Instagram photos reveal predictive markers of depression’, EPJ Data Science, 2017, 6, (1), pp. 15

22 Rokach, L.: ‘Ensemble-based classifiers’, Artificial Intelligence Review, 2010, 33, (1), pp. 1-39

23 Sagi, O., and Rokach, L.: ‘Ensemble learning: A survey’, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8, (4), pp. e1249

24 Sakado, K., Sato, T., Uehara, T., Sato, S., and Kameda, K.: ‘Discriminant validity of the inventory to diagnose depression, lifetime version’, Acta Psychiatrica Scandinavica, 1996, 93, (4), pp. 257-260

25 Serpen, G., and Pathical, S.: ‘Classification in High-Dimensional Feature Spaces: Random Subsample Ensemble’, in Editor (Ed.)(̂Eds.): ‘Book Classification in High-Dimensional Feature Spaces: Random Subsample Ensemble’ (2009, edn.), pp. 740-745

26 Somvanshi, M., and Chavan, P.: ‘A review of machine learning techniques using decision tree and support vector machine’, in Editor (Ed.)(̂Eds.): ‘Book A review of machine learning techniques using decision tree and support vector machine’ (2016, edn.), pp. 1-7

27 Tedders, S.H., Fokong, K.D., McKenzie, L.E., Wesley, C., Yu, L., and Zhang, J.: ‘Low cholesterol is associated with depression among US household population’, Journal of Affective Disorders, 2011, 135, (1-3), pp. 115-121

28 Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., and Ohsaki, H.: ‘Recognizing Depression from Twitter Activity’, 2015, pp. 3187-3196

29 Ubani, C.C., and Zhang, J.: ‘The role of adiposity in the relationship between serum leptin and severe major depressive episode’, Psychiatry Research, 2015, 228, (3), pp. 866-870

30 Wongkoblap, A., Vadillo, M.A., and Curcin, V.: ‘Researching Mental Health Disorders in the Era of Social Media: Systematic Review’, J Med Internet Res, 2017, 19, (6), pp. e228

31 Zimmerman, M., and Coryell, W.: ‘The Inventory to Diagnose Depression (IDD): A self-report scale to diagnose major depressive disorder’, Journal of Consulting and Clinical Psychology, 1987, 55, (1), pp. 55-59