

RESEARCH

Open Access



# Detecting depression using an ensemble classifier based on Quality of Life scales

Xiaohui Tao<sup>1\*</sup> , Oliver Chi<sup>2</sup>, Patrick J. Delaney<sup>1</sup>, Lin Li<sup>3</sup> and Jiajin Huang<sup>4</sup>

## Abstract

Major depressive disorder (MDD) is an issue that affects 350 million people worldwide. Traditional approaches have been to identify depressive symptoms in datasets, but recently, research is beginning to explore the association between psychosocial factors such as those on the quality of life scale and mental well-being, which will lead to earlier diagnosis and prediction of MDD. In this research, an ensemble binary classifier is proposed to analyse health survey data against ground truth from the SF-20 Quality of Life scales. The classifier aims to improve the performance of machine learning techniques on large datasets and identify depressed cases based on associations between items on the QoL scale and mental illness by increasing predictive performance. On the experimental evaluation on the National Health and Nutrition Examination Survey (NHANES), the classifier demonstrated an F1 score of 0.976 in the prediction, without any incorrectly identified depression instances. Only about 4% of instances had been mistakenly classified into depressed cases, with a significant accuracy of 95.4% comparing to the result from PHQ-9 mental screen inventory. The presented ensemble binary classifier performed comparably better than each baseline algorithm in all measures and all experiments. We trained the ensemble model on the processed NHANES dataset, tested and evaluated the results of its performance against mental screen inventory and discussed the comparable predictions. Finally, we provided future research directions.

**Keywords:** Major depressive disorder, Ensemble classification, Supervised machine learning

## 1 Introduction

Major depressive disorder (MDD) is one of the most prevalent healthcare issues worldwide [1]. According to the 2012 world health journal by the World Health Organization (WHO), more than 350 million people experience this illness, with many people having their quality of life impacted [2]. MDD has a range of different symptoms and at its most critical can result in suicidal thoughts, making it a global challenge for healthcare professionals for the past few decades. Depression can be successfully diagnosed by health experts when applying an operational diagnostic criteria of depression, such as a mental screening tool. However, an issue remains that a wide range of people with depressive symptoms

do not seek clinic advice or professional care, resulting in healthcare professionals being unable to intervene in major cases of MDD [3].

In most situations, MDD is characterised by at least two weeks of developing low mood [4]. Therefore, early diagnosis of MDD is essential for early intervention [3]. A traditional approach to early diagnosis is the use of a self-assessment tool such as the Patient Health Questionnaire (PHQ-9), which is then followed up through a validation interview with a Mental Health Professional (MHP) [5]. The disadvantage of this process is that it may take time and incur costs, which is one of the barriers to patients with undiagnosed MDD seeking early assistance. Therefore, ways to automatically detect depressed cases in datasets may assist in early diagnosis of MDD and ensure depressed individuals engage with MHPs early.

Healthcare experts note that predicting depression can be based on lifestyle choices. The SF-20 Quality of Life Scale (QOLS) is an instrument used to measure

\*Correspondence: xiaohui.tao@usq.edu.au

<sup>1</sup> School of Sciences, University of Southern Queensland, Toowoomba, Australia

Full list of author information is available at the end of the article

association between chronic illness and lifestyle behaviours of people using survey data. These measures are based on the Mental Health Inventory (MHI-5), a short five-question instrument used to assess mental health [6]. Information from the MHI-5 can be checked against other contributing physiological and psychosocial factors to fully understand what contributes to low mood and mental disorders in an individual. To this end, QOLS contains conceptual categories that provide the basis for contributing factors, including material and physical well-being, relationships with other people, social, community and civic activities, personal development and fulfillment, and recreation [7]. While machine learning (ML) researchers have explored the link between these conceptual categories and other chronic diseases, ML-related research into the correlation between mental illness and other contributing factors is gaining attention. In a systematic review of ML techniques in mental health, Thieme et al. [8] found a growing application of multiple classification methods to structured datasets such as health questionnaires to detect depressive and mood symptoms using extracted features. The studies identified in their review apply techniques to data containing health history and previous behaviours, to more effectively distinguish between depressive and non-depressive persons. Mowery et al. [9] conducted a study into the association between depression and psychosocial stressors, which contained 12 categories ranging from demographic, socio-economic and environmental factors. These studies indicate that the need to further investigate the link between depressive symptoms and other factors that either cause or exacerbate them, which may provide a foundation for researchers and MHPs to increase their understanding of how we can predict MDD based on the physiological and psychosocial experiences of individuals.

Shatte et al. [10] identified two central themes in modern ML approaches, being the development of pre-screening diagnosis tools and the development of models to identify a person's predisposition or risk for progressing to a severe mental health condition. Some of the most frequently reported challenges include robustly measuring and labelling mental health data as complex and dynamic due to noise or ambiguous data, generating low-dimensional features for reducing data into quantifiable categories for appropriate modelling and selection of models and training algorithms [8]. In studies involving ground truth especially, the challenge of low-dimensional labels can impact the accuracy of results. To this end, there is a need to improve the identification of appropriate features that can be used to help predict an individual's disposition toward MDD [11]. One recommendation for improvement,

being that future studies incorporate patient histories to improve the predictive capabilities of the diagnosis [10].

The National Health and Nutrition Examination Survey (NHANES) incorporates healthcare validation tools for measuring health status, such as the Patient Health Questionnaire (PHQ-9). The PHQ-9 tool is a 9-item screening instrument for measuring the severity of depressive symptoms from no depression to major depressive disorder. The PHQ-9 is the only integrated measurement for depression in the NHANES because it is simple, reliable and widely used in clinical evaluations [5]. Several studies have used NHANES data and the integrated PHQ-9 tool to explore correlations between depression and health issues, such as an investigation into the relationship between depression and low cholesterol [12], the association between MDD and obesity [13], the relationship between serum leptin and depressive episodes [14], and associations between blood folate concentrations in reproductive aged women and MDD [15]. To date, very few studies have used the NHANES to evaluate MDD against multiple psychosocial functionalities using ML techniques. Very little work has been done to incorporate psychological domain knowledge into the development of classifiers and no previous ensemble classifier approaches have used the NHANES dataset, with most ML techniques using it to explore physical disorders such as diabetes.

The objective of the present paper is to propose a suitable ML method to discriminate depression from collected health data for further interview diagnosis. The work done in this research builds an ensemble classifier using psychological domain knowledge from the SF-20. This ensemble classifier is applied to 98 features extracted from the NHANES healthcare dataset relating to mental health, which are used to explore associations between mental illness and items of health-related functionality on the QOLS: social, general, role, pain and physical functionalities.

The contributions of this research include:

- Development of an ensemble classifier using psychological domain knowledge from the PHQ-9 and SF-20.
- Improving predictive performance for MDD by incorporating 98 features extracted from the NHANES dataset.
- Further enhancing the utility of a ground truth technique similar to [5] to distinguish between depressive and non-depressive persons based on self-identification in health questionnaires.
- Using this ground truth technique with the classifier to demonstrate a higher accuracy of detecting

depressed cases in a large healthcare datasets such as the NHANES.

After discussing the contributions of this work, the remainder of this paper will provide background to the study and introduce the new ensemble classifier. This paper will firstly discuss related work on ML techniques in diagnosing MDD and discuss the use of ensemble classifiers in this domain. It will then discuss the Research Objective, followed by the Approach. The Experiment is then detailed, followed by the Results and Discussion of the experiment. A conclusion will present suggested future work.

## 2 Related work

### 2.1 Machine learning techniques in MDD

Several previous studies have applied learning algorithms to detect MDD symptoms on datasets containing patterns of behaviour among individuals, with the most common being [1]: Support Vector Machine algorithm [16–18], Naive Bayes method [19] and Random Forest technique [20]. In their systematic review of ML techniques in the mental health domain, Shatte et al. [10] identified regression and Decision Trees (DT) as other common approaches used. Along with KNN, these methods comprise typical representative techniques in machine learning experiments.

Predictive modelling for MDD symptoms on large datasets is a relatively new approach. Choudhury et al. [16] developed a probabilistic model to train crowdsourcing Twitter posts and develop a social media depression index to characterise the levels of depression in a sample population. The study used a SVM classifier with an RBF kernel for identifying depressive instances. Fivefold cross-validation was used to validate the performance of the classifier, with results yielding an average accuracy of 73% and high precision of 82% [16]. The depression index had a strong correlation with national depression statistics [16]. Importantly, this approach established the need to add social environment and external factors to MDD assessment.

Similarly, Tsugawa et al. [17] built a SVM supervised learning model to use features from online tweet activities to predict users' current depression status. Features used for predicting depression were extracted from the activity history of users. In this approach, an accuracy of 69% can be reached through the prediction of depressive users by the proposed classifier [17]. The trusted status (critical standard) of users were generated by CES-D and BDI screening scales of all participants. The limitation of this study found that long observation periods for collecting data may decrease accuracy [17, 21].

In combining a Random Forest (RF) algorithm and SVM technique, Fatima et al. [20] were able to discriminate depressive posts and communities from non-depressive posts and communities in the online social network LiveJournal. LiveJournal enables users to provide predefined “mood tags” on user posts, which were extracted features to measure depression levels among users who created posts and participated in communities. The study implemented Random Forest algorithm with SVM classifier for text classification to find the maximum margin between severe depressed, moderate depressive and non-depressed classes. In the experiment, RF performed better in comparison with a standard SVM method, as the proposed model achieved about 90% and 95% accuracy in classifying the depressive posts and depressed communities, respectively [20].

What many of these previous studies have in common is that they use single ML techniques and often use text-based data, such as datasets from social media, to predict MDD. Applying multiple techniques might improve the precision of detecting depressive symptoms in a wider range of data. The prior studies reviewed in this section adopt mainstream techniques such as DT, ANN, KNN and SVM as methods for machine learning experiments. As such, we can use these as typical representative techniques for baseline solutions as part of the comparative analysis against our ensemble classifier.

### 2.2 Ensemble classifiers

Instances of ensemble classifiers have proved promising in accurate diagnosis, particularly when using multiple features. Hassan et al. [19] used majority vote for classification and regression on top of predictions from three single classifiers: 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 as features from text on different social media platforms [19]. The performance accuracy of SVM is 91%, 83% and 80%, respectively, for NB and ME classifiers. In another study exploring data from Chinese social media network Weibo, Peng, Hu and Dang [18] used a multi-kernel SVM-based model on three categories of features, user microblog text, user profile and user behaviours [18]. The multi-kernel SVM method had a lowest error rate 16.5% for identifying depressed people [18]. This study demonstrated that an ensemble method can obtain better predictive performance using multiple learning algorithms than single traditional learning algorithms alone [18].

Such studies apply ensemble methods to rich textual data from social media, which is one of the most common applications of these approaches [8]. The use of survey data such as health questionnaires is gaining precedence

in studies, as the process of feature extraction can be useful in correlating external factors with possible MDD outcomes. Yang and Bath [22] built an ensemble classifier for predicting depression in the elderly, with different models (GBM, KNN, RGF and LR) used to explore different factors such as demographic, social engagement, physical health and disability, psychological and mental health, lifestyle and cognition factors. This improved predictive performance and provided greater insight into risk factors that could lead to severe depression in older adulthood, assisting early intervention. A study by Srividya et al. [23] developed an ensemble classifier to identify mentally distressed individuals in a target population of high school students and working individuals, with an accuracy score of 90%. Their study explored the association mental distress against education, socio-economic, life satisfaction and relationship quality factors, with the researchers indicating that the incorporation of additional parameters (e.g. physiological) could be added to extend predictive models to a range of other mental illnesses such as MDD. To date, studies have applied ensemble classifiers to mental illnesses such as stress [11] and internet addiction [24], with only a handful having applied it to depression [22].

For better predictive modelling, researchers need to incorporate additional parameters into an ML classifier. This leads to the challenge of both reducing the dimensionality of features and the selection of appropriate analytics techniques, which can each be affected by the quality of data collection [1]. These prior studies have employed multiple classification techniques, but there is still a need to test additional techniques for detecting depression, particularly in using large and varied samples. However, there is a lack of research into an efficient machine learning classifier for detecting depression in large data. The study by Hsieh et al. [24] utilised an ensemble classifier on internet addiction, but applications of this method to depression are in their infancy. These prior studies have incorporated parameters such as physiological, but there has been a small amount of work that uses psychological domain knowledge to build ensemble classifiers. Furthermore, few studies have used ground truth in combination with ensemble classification in this area. One study developed a ground truth dataset to test an ensemble classifier on mental health instances, but again, this study used social network data [25]. In [26], they used an ensemble classifier with ground truth to predict happiness based on a range of different data including physiological and behavioural, demonstrating the feasibility of the approach suggested in this research.

This work attempts to use psychological domain knowledge such as that of the SF-20 QOLS [5]. Only a handful of studies have used ensemble classifiers on

survey questionnaires, as indicated in a recent systematic review [8], with none using the NHANES data. This dataset is mostly used for physiological studies, however, the survey contains notable health-related variables that can influence mental health. To the best knowledge of the authors, this is the first study to attempt to use the NHANES data in such a work.

### 3 Research definitions

This research aims to design an effective ensemble classifier method for automatically detecting depressed cases in healthcare datasets. The objective is to develop the classifier based on psychological domain knowledge and use a process of ground truth to measure features in the NHANES survey data that are related to the functioning categories in the SF-20 QOLS.

To outline the objectives of the research, the first definition is:

**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  $s := \langle p_1, p_2, p_3, \dots, p_n \rangle$ , where

- $p$  is a visualisation or instance of an user property;
- $p$  is not a mental or depression closely 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.  $\square$

With clear definition of research objective, the research target is defined as:

**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 (1).  $\square$

From Definition 1, any given user property  $s \in \mathbb{S}$  is possibly overlapped with other user properties. While learning from related psychological researches, a set of user personal functionalities can present a reflection of a user's mental profile. This method can potentially detect depression by analysis of a set of user functionalities. Therefore, given the definition below:



**Definition 3** Let  $\mathbb{U} = \langle u_1, u_2, u_3, \dots, u_k \rangle$  be a sub-set of  $\mathbb{S}$ , any element  $u \in \mathbb{U}$  is a tuple  $u := \langle p'_1, p'_2, p'_3, \dots, p'_n \rangle$ , where

- $\mathbb{U}$  is a machine learning descriptive sub-set transferred from  $\mathbb{S}$  in psychological domain descriptive;
- $\forall p' \in u$  is assigned from an instance  $p \in a$  in Definition 1;
- $|\mathbb{D}^s|$  is limited due to the small functionalities defined in psychological domain.  $\square$

The research problem can be defined as an effective classification model  $\mathbb{M}$  that provides a reliable mapping function for a well-defined  $\mathbb{U}$  to map into  $\mathbb{V}$ :

$$\mathbb{U} \xrightarrow{\mathbb{M}} \mathbb{V} \text{ or } \mathbb{M}(\mathbb{U}) = \mathbb{V}.$$

Generally, we can label the cases waiting for detection into two classes: depression instances and non-depressed instances. The binary classification is seen as supervised learning because the objective is to use machine learning to automatically classify participants into two labelled categories of depression and non-depression.

## 4 Approach

### 4.1 NHANES survey data

In this study, we use the dataset from the National Health and Nutrition Examination Survey (NHANES). NHANES is designed to collect health-related information about the U.S. household population. It is a rich data source for health professionals and researchers for 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 is gathered and protected with the requirement of Federal Law of U.S. and for health research purposes only. Collections of NHANES in the last decade are free for researchers and published on the website of the NCHS.

We employ the questionnaire data in NHANES 2013–2014 collection as input data  $\mathbb{H}$  for the experiment. The age of participants is set to 18+, because data for teenagers and children are only partially published. As our objective is to classify general individuals into non-depressive and depressive groups, the features only involved with a single gender are excluded.

### 4.2 Build ground truth

Using the integrated PHQ-9 screen measurement, we can establish ground truth label information on whether or not a participant has depression) for the whole dataset. The PHQ-9 measurement scales contain five levels of

depression severity, from minimal-level to severe-level. In the research of Kroenke et al. [5], patients who were identified at the moderate-level (score  $\geq 10$ ) of depression in the PHQ-9 measurement had a sensitivity of 88% and a specificity of 88% for MDD. We therefore choose the separation at PHQ-9 score 10. Participants with a PHQ-9 score less than 10 are considered non-depressed and vice versa. We label these non-depressed people as the logical truth or “1”; conversely, those depressive people are labeled as the logical false or “0”.

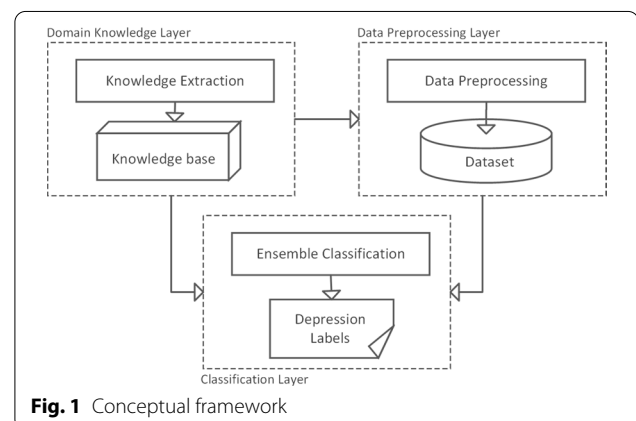
We rationalise our classification of study respondents into depressive and non-depressive groups because our ground truth is based on their self-reporting in the PHQ-9 component of the NHANES dataset. Given that our profile of study respondents is based only on their self-identification in the survey and not on clinical diagnosis of these individuals by medical professionals, we cannot separate them into different levels of depression (i.e. severe to minimal). Without the clinical diagnosis and their medical reports, we are dependent on their self-reporting of their state of depression using the tool to be reliable, hence we use the standard of the two-classification approach employed by [5]. Additionally, to predict mental status, we do not include their feelings and expressions in the features.

### 4.3 Conceptual framework

The conceptual framework is the theoretical structure encompassing all level models and classification methods. In this study, the framework consists of three layers:

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

The conceptual design of the framework is illustrated in Fig. 1.



The framework is implemented in the research experiment. The psychological knowledge layer learns how to group health informatics in the psychological domain. It directs the actions for the data processing layer to transform the dataset. It also assists designing the ensemble classification technique in the classification modelling layer. The data processing layer contains all proceedings of data preprocessing, feature extraction and establishment of the dataset. The layer converts the health statistics dataset into several normalised datasets ready for classification. The classification layer implements the classification of dataset. It builds an effective ensemble classifier and performs the comparative prediction of depressive risk for participants.

#### 4.3.1 Psychological knowledge

This layer is informed by the PHQ-9 instrument and the SF-20 QOLS. The PHQ-9 measures the level of depression severity, while the SF-20 contains questions around health-related functionalities related to quality of life. Kroenke et al. [5] discovered a strong association between increasing depression severity screen scores on the PHQ-9 and worsening functionality on all 6 categories of the SF-20 Quality of Life scale. The 6 categories are mental, social, role, pain, physical and general functions. Five items are derived from health diagnostic criteria in the Mental Health Inventory (MHI-5) and the sixth classifies mental disorder symptoms as mental category.

Associations of health functionalities with MDD have been observed in many previous studies. In research

by Clark et al. [27], they examined the opposite association of depression and psychosocial functionalities. Using 5 domains of physical, social, emotional, cognitive, and spiritual functioning. It found that depression is associated with poor health status and negative health behaviours. This layer intends to implement a similar approach with the health-related functionalities of the SF-20. Figure 2 from [5] illustrates the relationship between increasing PHQ-9 scores of depression and worsening functional categories (see Fig. 2).

The relationship between MDD and the scales of health-related functionalities have the similar trend as the severity of depression in statistics. Previous studies in related-work focused on detecting depressive symptoms and depression-related contents. Likewise, the relationship used in the implementation here presents a new potential method of predicting users' depression by sampling various diagnostic criteria of functionality. The combination of the classification technique and the binary ground truth technique could potentially enhance the strength of new predictive approaches. This new method has the following advantages compared to previous techniques:

- There are more features available for classification due to enlarged inputs in various functional areas;
- It is easier to acquire functional data than sensitive data of depressive symptoms, especially on social networks;

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 <sup>a</sup> (50 to 54)	59 <sup>a</sup> (57 to 61)	69 (67 to 71)	76 <sup>a</sup> (74 to 77)
Moderate, 10–14	51 (50 to 53)	53 (51 to 55)	65 (62 to 68)	75 <sup>a</sup> (72 to 78)	53 <sup>a</sup> (49 to 58)	64 <sup>a</sup> (60 to 69)	40 <sup>a</sup> (37 to 43)	48 (45 to 51)	49 <sup>a</sup> (45 to 52)	53 <sup>a,b</sup> (50 to 57)	63 <sup>a</sup> (60 to 66)	74 <sup>a</sup> (71 to 77)
Moderately severe, 15–19	43 (40 to 45)	45 (42 to 48)	55 (51 to 59)	68 <sup>a</sup> (63 to 72)	42 <sup>a</sup> (36 to 48)	64 <sup>a,b</sup> (57 to 71)	33 <sup>a,b</sup> (29 to 37)	40 <sup>a</sup> (35 to 44)	45 <sup>a,b</sup> (41 to 50)	50 <sup>b</sup> (45 to 55)	57 <sup>a,b</sup> (53 to 61)	74 <sup>a</sup> (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 <sup>b</sup> (39 to 58)	27 <sup>b</sup> (22 to 31)	30 <sup>a</sup> (24 to 36)	40 <sup>b</sup> (35 to 45)	46 <sup>b</sup> (40 to 53)	53 <sup>b</sup> (48 to 57)	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 ( $\pm 1.96 \times$  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 [5]

- c) It is easier to cover sufficient specificities of one functional status than to cover all available types of depressive symptoms;
- d) Using six functional status groups rather than only one collection of depressive symptoms will produce more accurate and comparable classification.

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

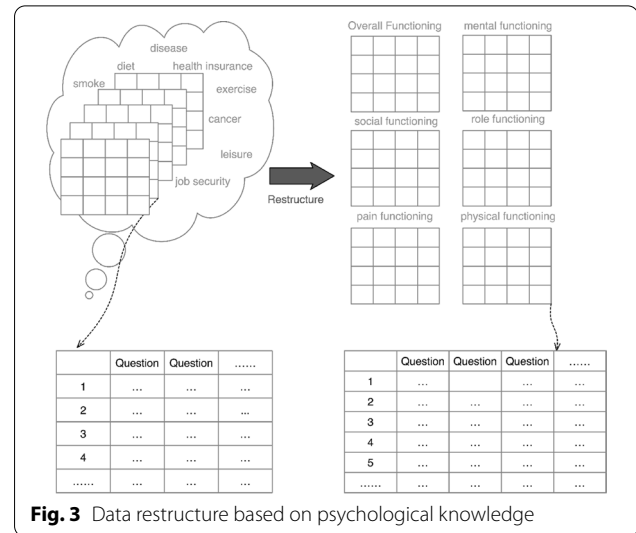
**Definition 4** Let new redesigned  $\mathbb{U} = \langle u_m, u_s, u_r, u_{pa}, u_{ph}, u_g \rangle$ , in which every  $u \in \mathbb{U}$  is an independent function of a user where

- $u_m$  is an individual mental disorder symptom;
- $u_s$  is a diagnostic criteria in the social activities;
- $u_r$  is a diagnostic criteria in the role functionality;
- $u_{pa}$  is a diagnostic criteria in the pain domain;
- $u_{ph}$  is a diagnostic criteria in the physical category;
- $u_g$  is a diagnostic criteria in the general actions.  $\square$

#### 4.3.2 Data processing

The survey questions in the NHANES dataset are spread across columns and participants are placed into rows, separated into different tables of health domains. Since the tables are not organised in the same format and structure, preprocessing is required for classification in the experiment. We will only use the survey questions component, which is one third of the dataset.

Data cleaning and transformation was used to eliminate redundancies and ensure the data is computer-readable. The data types are justified to make the dataset compatible and comparative. Normalisation was also necessary to uniform the scale condition in various questions. Whilst data preprocessing is implemented, the 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 rejoined into one table due to same row index. 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 preprocessing as below:



**Fig. 3** Data restructure based on psychological knowledge

**Definition 5** Let new overall dataset of  $m$  cases and  $n$  features  $\mathbb{D}_o = \{(x_1, x_2, \dots, x_n, y), x_i \in R^m, y \in \{0, 1\}^m\}$ , and sub-datasets of different six functional categories  $\mathbb{D}_m, \mathbb{D}_s, \mathbb{D}_r, \mathbb{D}_{pa}, \mathbb{D}_{ph}$  and  $\mathbb{D}_g$ , where

- $|\mathbb{D}_o| = |\mathbb{D}_m| = |\mathbb{D}_s| = |\mathbb{D}_r| = |\mathbb{D}_{pa}| = |\mathbb{D}_{ph}| = |\mathbb{D}_g| = m$ ;
- The space of features in  $\mathbb{D}_m, \mathbb{D}_s, \mathbb{D}_r, \mathbb{D}_{pa}, \mathbb{D}_{ph}, \mathbb{D}_g$  = the space of features in  $\mathbb{D}_o = n$ .  $\square$

#### 4.3.3 Classification modelling

We use an ensemble classification approach to build the model. It implements the independent ensemble methodology which applies several classification techniques in parallel. The model implements Support Vector Machine (SVM) technique, Artificial Neural 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 the approach is evaluated by a k-fold cross validation-algorithm. By amalgamating all outputs of composite classifiers into a single prediction, we generate the ensemble classifier. This ensemble classification approach collects various outputs of multiple independent classifiers and combines them to improve the predictive performance.

In general, an ensemble method provides higher accuracy and better predictive performance than a single algorithm [28]. Several advantages in performance include [29]:

- (i) *Overfitting avoidance* by averaging different hypothesis to reduce the risk of choosing an incorrect hypothesis.
- (ii) *Computational advantage* in decreasing the risk of obtaining a local minimum by combining several learning ensemble methods.
- (iii) *Strong representation* in achieving a better fit to the data space by 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 [29]. The ensemble method also imitates human nature by seeking various solutions before making a final decision. The ensemble method for this experiment is considered an optimised technology comparing to other baseline models in the classification of our preprocessed data.

#### 4.4 Ensemble model

The integration of four methods is expected to optimise predictive performance. As each independent sub-model is trained, the ensemble classifier covers more target concepts. To combine all baseline classifiers outputs, our modelling procedure adopts the weighting ensemble method. Weighting ensemble method is very generic 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 [28]:

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

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

Because the ensemble classifier combines the weighted outputs of all base classifiers, we can define the ensemble classifier as below:

**Definition 6** Let the ensemble model

$$\mathbb{M}_e = \sum_{k=1}^n w_i M_i, \quad (2)$$

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.  $\square$

The principle of this ensemble approach is to build estimators independently and then find the average of their predictions. The combined estimator is usually better than any single base estimator because instances variance is moderated.

#### 4.5 Adapted classification methods

Our ensemble classification method involves several baseline supervised classification models, consistent with predictive data mining. We selected supervised learning algorithms with diverse advantages. Each classification method has a diverse computing algorithm, with the goal being to build a concise model to achieve the best possible prediction accuracy. The supervised machine learning techniques used include [30]:

- a) Logic-based algorithm: the algorithms use logic or rules to make a decision of selecting proper features during the learning. DT method adopts this algorithm.
- b) Perceptron-based techniques: the algorithms are based on the notion of perceptron to construct pattern-like layers of neurons 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 are Naive Bayesian network and k-Nearest Neighbour technique.
- d) Support vector machines: SVM uses a hyperplane to separate two data classes and the margin created by the separating hyperplane indicates the success of the classification [30].

#### 4.6 Baseline models

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 [31]. In Section 2.1, a number of prior studies were discussed who implemented several typical representative ML techniques to analyse data. The techniques are taken as mainstream standards with the studies cited being contemporary and exploring similar work to what is demonstrated here. Hence, the ML techniques identified in Section 2.1 are the baseline models for our experiment. We thus propose one method of each type to present sufficient algorithms in the limited number of sub-models, selecting four techniques as our baseline models: Decision Tree method (DT), Artificial Neutral Network technique (ANN), k-Nearest Neighbour



(KNN) method and Support Vector Machine (SVM) algorithm.

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

```

input : Dataset  $\mathbb{D} = \{(x_1, x_2, \dots, x_n, y), x_i \in R^m, y \in \{0, 1\}^m\}$ 
output: Ensemble Model  $\mathbb{M}_e$ 
1 Set the training set as  $\mathbb{R} = \{(x_1, x_2, \dots, x_n), x_i \in R^m\}$ ,
  and the testing set as  $\mathbb{S} = \{y, y \in \{0, 1\}^m\}$ ;
2 for  $i \leftarrow 1$  to  $h$  do
3   /* validate baseline model */
4   Do training  $M_i$  on the training set  $\mathbb{R}$ ;
5   Get the performance  $E_i$  while validating on  $\mathbb{S}$ ;
6 end
7 Calculate weightings  $w_i = \frac{1-E_i}{\sum_{k=1}^h (1-E_k)}$ ;
8 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 in the health survey dataset  $\mathbb{H}$ . Using psychological knowledge, we reduce the dimension of the dataset by splitting it into sub-sets. This will benefit the processing of classification, while comparing the performance of the overall dataset and sub-sets for support of further solution on the real condition with less features.

### 5.1 Data preprocessing

The NHANES dataset contained errors and missing values. Not all questions were completed by participants. Further to this, the questionnaire involves “Refuse” and “Don’t Know” options for nearly every question due to considerations of personal privacy and the right not to disclose. Data preprocessing involved filling, correcting and normalising these inputs to prepare it for the experiment. In order to uniform all actions taken in data cleaning, we design a couple of presumption and principles to manage the proceeding:

- We assume that missing inputs belong to the persons who have on depressive risk;

- The choice of “Refuse” option or “Don’t Know” option is presumed normal which can be corrected by the statistical mean of inputs;
- All inputs of survey questions should be converted into binary, range and numbers due to the design of answer options;
- The final value of each input should be normalised and have a limited byte size.

After preprocessing, an overall dataset was produced with a total of 5398 participants. Among them, 516 ( 9.56% ) are depressive persons and 4882 ( 90.44% ) are non-depressive. The features of the dataset are variables representing a major question in the NHANES survey. After rejecting several irrelevant major questions, a total of 98 features were produced. Among them, inputs in 49 features are binary, 36 features consists of range data and the remaining 13 features are float numbers. Grouping 98 features into separate functionalities by Definition. 4, we generated six sub-datasets (see Table 1).

### 5.2 Experiment design

In the experiment, we first obtain dataset  $\mathbb{D}_o$  by data preprocessing on survey data  $\mathbb{H}$ ; next, we aggregate all features of  $\mathbb{D}_o$  into 6 health-related functional classes and follow the same procedure to divide  $\mathbb{D}_o$  into 6 sub-sets  $\mathbb{D}_{ph}$ ,  $\mathbb{D}_r$ ,  $\mathbb{D}_m$ ,  $\mathbb{D}_s$ ,  $\mathbb{D}_{pa}$  and  $\mathbb{D}_g$ ; and we train dataset  $\mathbb{D}_o$  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 a k-fold cross-validation algorithm to determine the value of the complete predictive performance. The experimental procedure is depicted in Fig. 4.

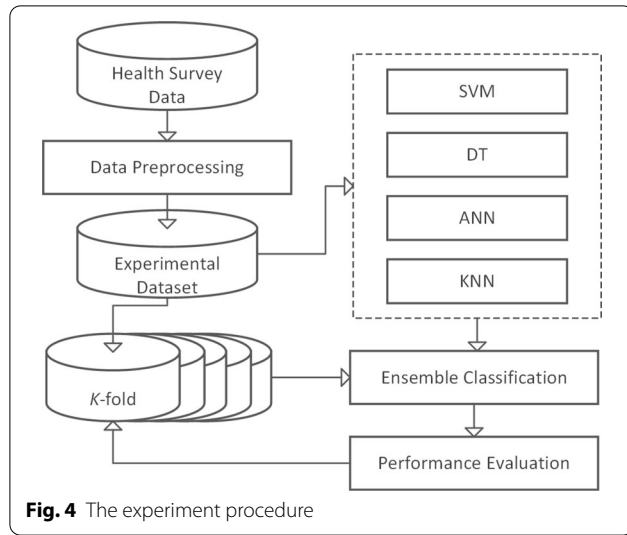
From the proceeding details of classification, we can define the algorithm of whole experiment as below (see Algorithm 2):

The ensemble classification can be expressed in algorithm as well: Given a well-preprocessed dataset of  $m$  examples and  $n$  features  $\mathbb{D} = \{(x_1, x_2, \dots, x_n, y), x_i \in R^m, y \in \{0, 1\}^m\}$ , 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  $\mathbb{D}$ :

Many machine learning packages and tools are accessible to implement common classification algorithms. The scikit-learn library from Python provides simple and efficient

**Table 1** Features and sub-datasets

Dataset	$\mathbb{D}_o$	$\mathbb{D}_{ph}$	$\mathbb{D}_r$	$\mathbb{D}_s$	$\mathbb{D}_m$	$\mathbb{D}_{pa}$	$\mathbb{D}_g$
Features	98	7	9	6	4	2	70



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

**input** : a rare health survey dataset  $\mathbb{H}$   
**output**: Ensemble Classifier  $\mathbb{F}_e$  and the complete prediction

- 1 Obtain dataset  $\mathbb{D}_o$  by data pre-processing on survey data  $\mathbb{H}$ ;
- 2 Aggregate features manually referred on 6 psychological functionalities ;
- 3  $\mathbb{D}_o = \mathbb{D}_{ph} \cup \mathbb{D}_r \cup \mathbb{D}_m \cup \mathbb{D}_s \cup \mathbb{D}_{pa} \cup \mathbb{D}_g$ , for each pair of  $(\mathbb{D}_i, \mathbb{D}_j)$ , where both  $\mathbb{D}_i$  and  $\mathbb{D}_j \in \{\mathbb{D}_{ph}, \mathbb{D}_r, \mathbb{D}_m, \mathbb{D}_s, \mathbb{D}_{pa}, \mathbb{D}_g\}$ ,  $\mathbb{D}_i \cap \mathbb{D}_j = \phi$ ;
- 4 Supervised learning on  $\mathbb{D}_{overall}$  for ensemble model  
 $\mathbb{M}_e = \sum_{g=1}^h w_i M_i$ ;
- 5 **foreach** sub-dataset  $\mathbb{D}_i$  in  $\{\mathbb{D}_o, \mathbb{D}_{ph}, \mathbb{D}_r, \mathbb{D}_m, \mathbb{D}_s, \mathbb{D}_{pa}, \mathbb{D}_g\}$  **do**
- 6     /\* ensemble classification\*/
- 7     Do ensemble classification on  $\mathbb{D}_i$  ;
- 8     Validate its predictive performance;
- 9 **end**

**Algorithm 2:** Experiment Design

Python package to implement four baseline models.

### 5.2.1 Kernel and parameters

Selecting suitable kernel and parameters is a common task for classification but it is also complex for specific examples. We only balance the settings of baseline

**Table 2** Performance and weights for sub-models

Models	Accuracy	F1 score	1 - F1	Weight
SVM	0.921	0.958	0.042	0.228
ANN	0.905	0.948	0.052	0.283
KNN	0.908	0.951	0.049	0.266
DT	0.925	0.959	0.041	0.223

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 to be improved by ensemble classification. We thereby employ common values for kernel and parameters. All four sub-models are configured for binary classification and their predictive performances are weighted in both labelled classes.

### 5.3 Performance measure

The predictive performance of each base classifier in our model is evaluated by F1 score, which is generated on a confusion matrix of validation. In the confusion matrix, we simply set the number of real non-depressive cases in the training set as *condition positive* ( $P$ ) and let the number of real depressive cases in the training set as *condition negative* ( $N$ ). F1 score is a balanced measure of both the precision (PPV) and the recall (TPR) of the validation:

$$F1 = \frac{2}{\frac{1}{TPR} + \frac{1}{PPV}} = \frac{2TP}{2TP + FP + FN}. \quad (3)$$

## 6 Results and discussion

### 6.1 Experimental results

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

$$\mathbb{F}_e = 0.228 \cdot f_{svm} + 0.283 \cdot f_{nb} + 0.266 \cdot f_{knn} + 0.223 \cdot f_{dt}. \quad (4)$$

```

input : Dataset  $\mathbb{D} = \{(x_1, x_2, \dots, x_n, y), x_i \in R^m, y \in \{0, 1\}^m\}$ 
output: the optimised ensemble classifier  $\mathbb{F}_e$  and its predictive performance  $p_e$ 

1 Divide dataset  $\mathbb{D}$  into  $k$  portions, each portion has  $\frac{m}{k}$  examples;
2 for  $k \leftarrow 1$  to 5 do
3   Select all portions except  $k^{th}$  portion to form new dataset  $\mathbb{D}'$ ;
4   Use  $\mathbb{D}'$  to generate the training set  $\mathbb{R} = \{(x_1, x_2, \dots, x_n)\}$  and the testing set  $\mathbb{S} = \{y\}$ , where  $|\mathbb{D}'| = |\mathbb{R}| = |\mathbb{S}| = \frac{4}{5}|\mathbb{D}| = \frac{4m}{5}$ ;
5   /* baseline model */
6   foreach one classification method of SVM, ANN, KNN, DT do
7     Training on the training set  $\mathbb{R}$  and obtain classifier  $f$ ;
8     Obtain predictive value  $y^p = f(\sum_{i=1}^n (x_i))$ ;
9   end
10  /* ensemble */
11  Calculate the ensemble classifier  $\mathbb{F}_k = w_{svm} \cdot f_{svm} + w_{nb} \cdot f_{nb} + w_{knn} \cdot f_{knn} + w_{dt} \cdot f_{dt}$ ;
12  Calculate a float predictive value  $y_e = w_{svm} \cdot y_{svm}^p + w_{nb} \cdot y_{nb}^p + w_{knn} \cdot y_{knn}^p + w_{dt} \cdot y_{dt}^p$ ;
13  /* sensitivity */
14  if  $y_e > 0.5$  then
15    |  $y_e = 1$  /* non-depression */
16  else
17    |  $y_e = 0$  /* depression */
18  end
19  Test  $y_e$  on testing set  $\mathbb{S}$  and report predictive performance  $p_k$ ;
20 end
21 /* 5-fold cross validation */
22 Validate the predictive performance by calculating  $p_e = \frac{\sum_{k=1}^5 p_k}{5}$ ;
23 Generate the optimised ensemble classifier  $\mathbb{F}_e = \text{Median}(\mathbb{F}_1, \mathbb{F}_2, \mathbb{F}_3, \mathbb{F}_4, \mathbb{F}_5)$ 

```

**Algorithm 3:** Ensemble Classification Procedure

Accuracy indicates the number of correct predictions made in all occurrences of both labels. It presents all corrected predictions based on the results of the PHQ-9 mental screen inventory. Precision is the ability of a classifier not to label an instance positive that is actually negative. Here, we measure how effectively the model can diagnose a person's psychological health. Recall is the ability of a classifier to find all positive instances. It measures how many non-depressed people are correctly identified. As the features and specificity of the overall dataset and each sub-datasets varied, the divided performances are expected (see Tables 3, 4 and 5).

The ensemble classifier performed better compared to the baseline models, including an F1 score of 0.976 vs 0.959 achieved by DT, and an accuracy of 0.954 vs 0.924, which was the again achieved by DT. Performances in functionality sub-sets is compromised in this experiment, but is still comparable to other machine learning methodologies [18–21].

**Table 3** Features and performances of the ensemble classifier

Dataset	F1 score	Accuracy	Precision	Recall
$\mathbb{D}_o$	0.976	0.954	0.956	1.000
$\mathbb{D}_{ph}$	0.964	0.931	0.934	1.000
$\mathbb{D}_r$	0.963	0.929	0.929	1.000
$\mathbb{D}_s$	0.964	0.931	0.931	1.000
$\mathbb{D}_m$	0.975	0.953	0.960	0.999
$\mathbb{D}_{pa}$	0.961	0.925	0.925	1.000
$\mathbb{D}_g$	0.964	0.931	0.930	1.000

The results in Table 3 indicate that for the mental functionalities sub-set, the F1 score (0.975) and accuracy (0.953) is closest to the scores for the overall dataset. The accuracy and F1 scores in physical, social and role functionalities in isolation are below the performance of the overall dataset. This indicates that mental functionalities are most relevant to the classifier.  $\mathbb{D}_m$  had 4 features, which suggests that these are the most relevant features to the psychological knowledge used to underpin the method. This is consistent with domain knowledge.

The prediction performance in the mental functionality sub-set shown in Table 5 is close to the whole dataset even though it has less features involved. This may indicate that features for mental functionality are more depression-related than features in other categories, because non-criteria items in the depression scale decreased in specificity of performance [4].

Recall measures encompass the successful rate of non-depressive predictions and are almost equal to 1 in the experiment as displayed in Table 6, the ensemble classifier was successful in the prediction of non-depressive cases.

## 6.2 Discussion

The ensemble classifier is superior to the baseline models in both F1 measure and Accuracy. It led the test results of both the overall dataset and all experiment results in sub-datasets as shown in Tables 4 and 5, respectively. It gathered different predictions from the baseline models and combined them into a better prediction. The ensemble proved more stable and robust than any involved baseline algorithm during the experiment. We utilised a random under-sampling technique with ensemble method to leverage the class imbalance problem where non-depression instances is about 10 times larger than depressed instances.

The proposed ensemble method significantly improved predictive performance with class imbalance. By analysis of the performance in recall measure (see Table 6), the

**Table 4 Performances in F1 score**

Models	$\mathbb{D}_o$	$\mathbb{D}_{ph}$	$\mathbb{D}_r$	$\mathbb{D}_s$	$\mathbb{D}_m$	$\mathbb{D}_{pa}$	$\mathbb{D}_g$
SVM	0.958	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 5 Performances in accuracy**

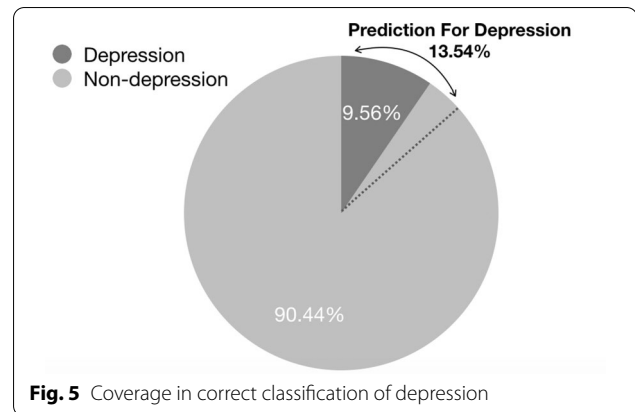
Models	$\mathbb{D}_o$	$\mathbb{D}_{ph}$	$\mathbb{D}_r$	$\mathbb{D}_s$	$\mathbb{D}_m$	$\mathbb{D}_{pa}$	$\mathbb{D}_g$
SVM	0.921	0.904	0.904	0.904	0.919	0.904	0.907
ANN	0.905	0.895	0.879	0.892	0.928	0.904	0.873
KNN	0.908	0.900	0.896	0.895	0.923	0.886	0.904
DT	0.924	0.905	0.904	0.905	0.926	0.904	0.906
Ensemble	0.954	0.931	0.929	0.931	0.953	0.925	0.931

**Table 6 Performances in recall**

Models	$\mathbb{D}_o$	$\mathbb{D}_{ph}$	$\mathbb{D}_r$	$\mathbb{D}_s$	$\mathbb{D}_m$	$\mathbb{D}_{pa}$	$\mathbb{D}_g$
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	0.981	0.992	0.997	0.993	0.982	1.000	0.992
Ensemble	1.000	1.000	1.000	1.000	1.000	1.000	1.000

preferred ensemble method covers all depressed cases in PHQ-9 screening measurement where no depressed instance has been mistakenly labelled as non-depression. The recall performances 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, the 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 were mistakenly classified. The coverage in correct classification of depressed participants is perfect, only slightly larger than the results of the psychological screening (illustrated in Fig. 5). If we let the predicted precision be  $P_p$  and percentage of non-depressed instances as  $N_1$ , the overall prediction  $P_o$  of depressed instances is calculated as below:

$$P_o = 1 - (P_p \cdot N_1) = 1 - 0.956 \cdot 90.44\% = 13.54\%. \quad (5)$$

**Fig. 5** Coverage in correct classification of depression

The coverage (see Fig. 5) of depression cases is slightly larger than the real situation of the MHI-5. However, it is acceptable for large sampling that there is no missing of any depression case and only about 4% of total cases have been incorrectly labelled as depressed in the prediction. The proposed ensemble method is perfect for preliminary screening of MDD cases for further clinical



diagnosis without missing any potential depression case.

In comparison with the predictions in the different sub-datasets (see Table 3), the ensemble classifier performs the best in the overall dataset and has a similar accuracy in the mental sub-set. The importance of diagnostic criteria in mental symptoms demonstrates that mental criteria are the major features for identifying depression. Meanwhile, both accuracy and F1 measures for predicting depression in physical, social and role functionalities are equal to the predictive performance in general sub-set. The features in the mental sub-set have the most relevance to the conceptual framework of the classifier, which was expected given the domain knowledge used to build it. This demonstrated that even without clinical process and using the two-classification approach of depressive and non-depressive groups, the study still demonstrated that mental functionalities remain the most significant predictor for depression. This is most evident in the precision score, which was higher than the overall dataset precision shown in Table 3. However, while mental functionalities give reasonable accuracy and F1 score, adding other features from the QOLS increases the overall predictive performance of the model and shows that adding other life scales can assist in detecting depression. Mental features appear to be the most contributing factors which is consistent with psychological domain knowledge. In the overall dataset, further experiments might involve removing sub-sets from the overall dataset such as general features, which might impact the overall accuracy. Further study is needed on the features from the QOLS that have the strongest relationship with mental attributes.

Additional experiments adding other significant mental features to the sub-set and comparing it against the overall dataset might provide further insight into this correlation. The general sub-set (70 features) has more features than the physical (7 features), social (6 features) and role (9 features) sub-sets. This indicates that many features in the general sub-set occurred without enough specificity for classifying depressed and non-depressed labels. Partial general functional features hence are useless in the detection of depression. Correspondingly, features in physical, social and role functioning sub-sets are more correlational in the classification. Weak depression indicator is not only helpless in the classification, but also incline the overall predictive accuracy. Therefore, it is extremely critical for depression diagnostic approaches to select a limited number of suitable features to distinguish depressed cases from a wide range. From the result of this research, we suggest an algorithm for feature selection which first involves as more mental symptoms as possible according to depression diagnostic criteria

and pluses no more than 50% features in health criteria in physical, role and social functionality. This algorithm ensures the majority of features consisted by mental diagnostic criteria and mixes partial health criteria to avoid the scenario that temporary mental status change occurs by sudden events like losing close relatives. It simulates the proceedings that psychologist did in the standard clinical interview.

## 7 Conclusion and future work

This work presented a binary ensemble classifier which is able to distinguish depressive cases from non-depressive cases in a wide ranging health survey dataset. Importantly, the ensemble method does not miss any identification of potentially depressed case. In the experimental evaluation using the NHANES dataset, only 4% cases were mistakenly classified into depressed class and no depressed case were incorrectly detected. The ensemble classifier on the whole dataset has a high F1 measure of 0.976 compared to the PHQ-9 depression screen inventory, 95.4% and 95.6%, for Accuracy and Precision, respectively. It also demonstrated that the ensemble system is stable and robust for detecting depression on a partial dataset. The approach and the experiment also demonstrated that the combination of a classification technique with binary ground truth can provide stronger predictive performance compared to baseline standards. The ensemble method is very simple, close to the bagging and major voting ensemble methods. Other boost ensemble methods are also suggested to improve the prediction performance further, like the EUSBoost method [29].

Moving forward, this research presents a method that can assist in the preliminary screening of depressive cases in a large number of potential cases before formal clinical diagnosis. The significance is we demonstrate that an ensemble classifier outperforms baseline models in both distinguishing depressed and non-depressed cases, and predicting potential MDD diagnosis based on mental health severity scales in the PHQ-9 and health-related functionalities in the SF-20. With these two assessment instruments being widely used in healthcare, the system provides an efficient way to screen more people than traditional technologies and has a similar accuracy and coverage as the current PHQ-9. However, the reliability and sensitivity of this ensemble system need to be tested on additional datasets. In particular, adding additional features to the mental sub-set would provide further evidence about the importance of mental functionalities. Mental features appear to be the most significant contributing factors to predicting depression, however, further tests of these sub-sets in isolation and using relevant QOLS data will further improve the classification

performance and understanding about the relationship between features and depression. Several possible future research directions for applying our ensemble method include using rich online social media sources to extract features for classification, as is the current trend in ML approaches [8]. Using this classification method on textual data will assist in improving the reliability and sensitivity of the ensemble system. Furthermore, deep learning techniques like DNN would increase the range of the ensemble classification. Hence, this will be our next work in improving this method.

#### Authors' contributions

XT contributions include conceptualization, methodology, formal analysis, investigation and supervision. OC contributed with methodology, software, validation, investigation and writing—original draft; PJD with investigation, data curation, and writing—review and editing; LL and JH with conceptualization and methodology. All authors read and approved the final manuscript.

#### Funding

This work is partially supported by the National Basic Research Program of China (No. 2014CB744600), National Natural Science Foundation of China (No. 61420106005).

#### Availability of data and materials

The authors would like to acknowledge the use of the National Health and Nutrition Examination Survey (NHANES) in the study and thank the Centers for Disease Control and Prevention of the Department of Health and Human Services, the United States for making the data set public for research purpose.

#### Ethics approval and consent to participate

The work is conducted under the approval from the Human Research Ethics Committee of the University of Southern Queensland, Australia (Approval ID: H18REA049).

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

#### Author details

<sup>1</sup> School of Sciences, University of Southern Queensland, Toowoomba, Australia. <sup>2</sup> Advanced Analytics Institute, University of Technology, Sydney, Australia. <sup>3</sup> School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, China. <sup>4</sup> International WIC Institute, Beijing University of Technology, Beijing 100124, China.

Received: 30 September 2020 Accepted: 23 October 2020

Published online: 15 February 2021

#### References

- Wongkoblap A, Vadillo MA, Curcin V (2017) Researching mental health disorders in the era of social media: systematic review. *J Med Internet Res* 19(6):228. <https://doi.org/10.2196/jmir.7215>
- Karyotaki E, Cuijpers P, Albor Y, Alonso J, Auerbach RP, Bantjes J, Bruffaerts R, Ebert DD, Hasking P, Kiekens G et al (2020) Sources of stress and their associations with mental disorders among college students: results of the world health organization world mental health surveys international college student initiative. *Front Psychol* 11:1759
- Huerta-Ramírez R, Bertsch J, Cabello M, Roca M, Haro JM, Ayuso-Mateos JL (2013) Diagnosis delay in first episodes of major depression: a study of primary care patients in Spain. *J Affective Disord* 150(3):1247–1250. <https://doi.org/10.1016/j.jad.2013.06.009>
- Zimmerman M, Coryell W (1987) The inventory to diagnose depression (idd): a self-report scale to diagnose major depressive disorder. *J Consulting Clin Psychol* 55(1):55–59. <https://doi.org/10.1037/0022-006X.55.1.55>
- Kroenke K, Spitzer RL, Williams JB (2001) The phq-9: validity of a brief depression severity measure. *J Gen Intern Med* 16(9):606–13. <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>
- Rumpf H-J, Meyer C, Hapke U, John U (2001) Screening for mental health: validity of the mhi-5 using dsm-iv axis i psychiatric disorders as gold standard. *Psychiatry Res* 105(3):243–253
- Burckhardt CS, Anderson KL (2003) The quality of life scale (qols): reliability, validity, and utilization. *Health Quality Life Outcomes* 1(1):60
- Thieme A, Belgrave D, Doherty G (2020) Machine learning in mental health: a systematic review of the hci literature to support the development of effective and implementable ml systems. *ACM Trans Comp Hum Interact (TOCHI)* 27(5):1–53
- Mowery D, Smith H, Cheney T, Stoddard G, Coppersmith G, Bryan C, Conway M (2017) Understanding depressive symptoms and psychosocial stressors on twitter: a corpus-based study. *J Med Internet Res*. <https://doi.org/10.2196/jmir.6895>
- Shatte AB, Hutchinson DM, Teague SJ (2019) Machine learning in mental health: a scoping review of methods and applications. *Psychol Med* 49(9):1426–1448
- Alonso S, Torre-Díez I, Hamrioui S, López-Coronado M, Barreno D, Nozaleda L, Franco M (2018) Data mining algorithms and techniques in mental health: a systematic review. *J Med Syst* 42(9):1–15
- Tedders SH, Fokong KD, McKenzie LE, Wesley C, Yu L, Zhang J (2011) Low cholesterol is associated with depression among us household population. *J Affect Disorders* 135(1–3):115–121. <https://doi.org/10.1016/j.jad.2011.06.045>
- Merikangas A, Mendola P, Pastor P, Reuben C, Cleary S (2012) The association between major depressive disorder and obesity in us adolescents: results from the 2001–2004 national health and nutrition examination survey. *J Behav Med* 35(2):149–154. <https://doi.org/10.1007/s10865-011-9340-x>
- Ubani CC, Zhang J (2015) The role of adiposity in the relationship between serum leptin and severe major depressive episode. *Psychiatry Res* 228(3):866–870. <https://doi.org/10.1016/j.psychres.2015.05.009>
- Andrea SB, Siegel SAR, Teo AR (2016) Social support and health service use in depressed adults: findings from the national health and nutrition examination survey. *General Hospital Psychiatry* 39:73–79. <https://doi.org/10.1016/j.genhosppsych.2015.11.009>
- De Choudhury M, Counts S, Horvitz E (2013) Social media as a measurement tool of depression in populations. <https://doi.org/10.1145/2464464.2464480>
- Tsugawa S, Kikuchi Y, Kishino F, Nakajima K, Itoh Y, Ohsaki H (2015) Recognizing depression from twitter activity. pp. 3187–3196. <https://doi.org/10.1145/2702123.2702280>
- Peng Z, Hu Q, Dang J (2017) Multi-kernel svm based depression recognition using social media data. *Int J Mach Learn Cybern*. <https://doi.org/10.1007/s13042-017-0697-1>
- Hassan AU, Hussain J, Hussain M, Sadiq M, Lee S (2017) Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. *IEEE, New York*. <https://doi.org/10.1109/ICTC.2017.8190959>
- Fatima I, Mukhtar H, Ahmad HF, Rajpoot K (2018) Analysis of user-generated content from online social communities to characterise and predict depression degree. *J Inform Sci* 44(5):683–695. <https://doi.org/10.1177/0165551517740835>
- Reece AG, Danforth CM (2017) Instagram photos reveal predictive markers of depression. *EPJ Data Sci* 6(1):15. <https://doi.org/10.1140/epjds/s13688-017-0110-z>
- Yang H, Bath PA (2019) Automatic prediction of depression in older age. In: *Proceedings of the Third International Conference on Medical and Health Informatics 2019. ICMHI 2019. Association for Computing Machinery, New York*. pp. 36–44. <https://doi.org/10.1145/3340037.3340042>
- Srividya M, Mohanavalli S, Bhalaji N (2018) Behavioral modeling for mental health using machine learning algorithms. *J Med Syst* 42(5):88

24. Hsieh W-H, Shih D-H, Shih P-Y, Lin S-B (2019) An ensemble classifier with case-based reasoning system for identifying internet addiction. *Int J Environ Res Public Health* 16(7):1233
25. Islam MR, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A (2018) Depression detection from social network data using machine learning techniques. *Health Inform Sci Syst* 6(1):8
26. Jaques N, Taylor S, Azaria A, Ghandeharioun A, Sano A, Picard R (2015) Predicting students' happiness from physiology, phone, mobility, and behavioral data. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, New York. pp. 222–228
27. Clark MM, Bradley KL, Jenkins SM, Mettler EA, Larson BG, Preston HR, Liesinger JT, Werneburg BL, Hagen PT, Harris AM, Riley BA, Olsen KD, Vickers Douglas KS (2014) The effectiveness of wellness coaching for improving quality of life. *Mayo Clinic Proceedings* 89(11):1537–1544. <https://doi.org/10.1016/j.mayocp.2014.04.028>
28. Rokach L (2010) Ensemble-based classifiers. *Artif Intell Rev* 33(1):1–39. <https://doi.org/10.1007/s10462-009-9124-7>
29. Sagi O, Rokach L (2018) Ensemble learning: a survey. *Wiley Interdisciplinary Rev Data Mining Knowl Discov* 8(4):1249. <https://doi.org/10.1002/widm.1249>
30. Kotsiantis SB, Zaharakis ID, Pintelas PE (2006) Machine learning: a review of classification and combining techniques. *Artif Intell Rev* 26(3):159–190. <https://doi.org/10.1007/s10462-007-9052-3>
31. Choudhary R, Gianey HK Comprehensive review on supervised machine learning algorithms. In: 2017 International Conference on Machine Learning and Data Science (MLDS). pp. 37–43. <https://doi.org/10.1109/MLDS.2017.11>

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

---

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)

© The Author(s) 2021. This work is published under  
<http://creativecommons.org/licenses/by/4.0/>(the “License”). Notwithstanding  
the ProQuest Terms and Conditions, you may use this content in accordance  
with the terms of the License.