Supervised Learning on Health Survey for Depression Detection

Oliver Chi, *Member, IEEE,* Xiaohui Tao, *Fellow, OSA,*

Abstract—Major depressive disorder is a serious worldwide issue for healthcare. This study was to transfer current psychological knowledge on depression diagnosis into an analysis technique of health survey data. It aimed to provide psychologists a proper algorithm for automate screening a large number of samples to identify depressed cases among. The research proposed an effective ensemble binary classifier for distinguishing depressive instances from a wide range of healthcare data. On the experimental evaluation on the NCHS health dataset, it has a significant measure 0.976 of F1 score in the prediction, without any incorrectly identified depression instance. Only about 4% instances had been mistakenly classified into depressed cases with a significant Accuracy 95.4% comparing to the result from PHQ-9 mental screen inventory. The presented ensemble binary classifier performed comparably better than each baseline algorithms in all measures and all experiment. It approved that the ensemble system is stable and robust to preliminary screening of depressive instances from a large number of health data. In this research, we introduced how to transfer the psychological knowledge into the classification methodology on health dataset. Meanwhile, we analysed the ensemble technique and several supervised learning classification algorithms for building a solid classification method. Finally, we trained the ensemble model on processed dataset, tested and evaluated with the result of mental screen inventory, discussed the comparable predictions, and pointed out the future research directions.

Index Terms—major depressive disorder, ensemble classification technique, supervised machine learning, mental screen inventory, psychological domain knowledge, python, scikit-learning.

1 Introduction

MENTAL disorder is one of the most serious and prevalent healthcare issues worldwide [1]. According to 2012 world health journal by the World Health Organisation (WHO), more than 350 million people have major depressive disorder which can leads to suicide in the worst circumstance. Major depressive disorder also known simply as depression, has been a global challenge for healthcare over years. In psychological domain, it is defined as a mental disorder consistent with at least two weeks of developing low mood across most situations [2]. 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 [3]. 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 [3]. 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.

- *M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. E-mail: see http://www.michaelshell.org/contact.html*
- *J. Doe and J. Doe are with Anonymous University.*

Manuscript received April 19, 2005; revised August 26, 2015.

2 Related Work

By analysis of various contents on depressive criteria and symptoms, it is possible to use machine learning techniques to develop automatic detection systems for major depressive disorders. People who having specific actions or particular patterns of interaction can be classified into cases or non-cases of depressed group through existing learning algorithms [1], such as Support Vector Machine algorithm [5] [4] [12], Naive Bayes method [10] and Random Forest technique [11].

Choudhury et, al. [5] 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 [5]. 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 % [5].

Tsugawa et, al. [4] 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 [4]. 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 [4] [13].

Moreover, researchers compared several supervised learning techniques to achieve the best performance of predicting depression. For instance, Hassan et, al. [10] 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 [10]. The performance was observed that the accuracy of SVM is 91 %, 83 % and 80 % respectively for NB and ME classifiers.

Fatima et, al. [11] 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 predefined 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 [11]. They recommended that RF is a very powerful classifier in algorithm for establishing an accurate model for multi-class classification [11]. In the experiment, RF performed better in comparison with SVM method [11]. The proposed model achieved about 90 % and 95 % accuracy in classifying the depressive posts and depressed communities, respectively.

Peng, Hu and Dang [12] 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 [12]. 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 [12]. 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 [12].

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 [13]. However, the results showed that their predictive method for prediagnosis was rather conservative and tended to detect no depression in all instances [13].

Despite an increasing number of impressive researches detecting depression on massive resources, particularly rich data from social media, some common problems still persist. Successfully distinguishing depressive users from some data source is problematic, "not only due to biases associated with the collection methods, but also with regard to managing consent and selecting appropriate analytics techniques" [1]. Choudhary and Gianey [26] stated that every learning algorithm differs according to area of application and no algorithm is more powerful than the other in all scenarios. 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 Research Problem

This research aims to design an effective classification method for automatically detecting depressed cases in healthcare dataset and also to provide a solid fundation for future preliminary screening of major depressive disorder on social networks. In order to clearly describe the research, 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 cases waiting for detection into two classes: depression instances and non-depressed examples. 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 non-depression.

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

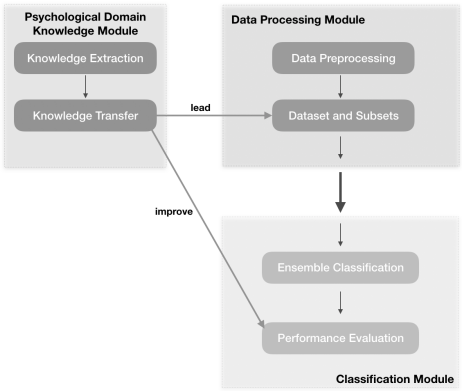

Fig. 1: Conceptual Framework

ensemble classifier and performs the comparative prediction of depressive risk for participants.

4.2 Psychological Knowledge

Kroenke et, al. [6] 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).

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 [6]

Associations of health related functionings with depression have been observed in many previous studies at psychological domain. For instance, Clark et, al. [14] 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. [15] 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-Saenz de Tejada et, al. [16] 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 $\mathbb{U} = \langle u_m, u_s, u_r, u_{pa}, u_{ph}, u_g \rangle$, every u $\in \mathbb{U}$ is an independent function of user, where

- u_m presents individual mental disorder symptoms;
- u_s presents diagnostic criteria in the social activities;
- u_r presents diagnostic criteria in the role functionality;
- u_{pa} presents diagnostic criteria in the pain domain ;
- u_{ph} presents diagnostic criteria in the physical category;
- u_g presents 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

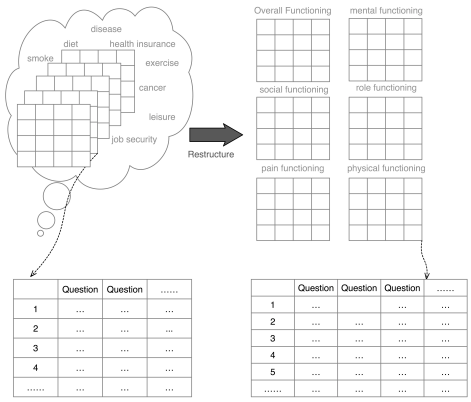

Fig. 3: Data Restructure based on Psychological Knowledge

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 $\mathbb{D}_o = \{(x_1, x_2, ..., x_n, y), x_i \in R^m, y \in \{0,1\}^m\}$, and sub-datasets of different 6 functional categories $\mathbb{D}_m$, $\mathbb{D}_s$, $\mathbb{D}_r$, $\mathbb{D}_{pa}$, $\mathbb{D}_{ph}$ and $\mathbb{D}_g$, where

- $|\mathbb{D}_o| = |\mathbb{D}_m| = |\mathbb{D}_s| = |\mathbb{D}_r| = |\mathbb{D}_{pa}| = |\mathbb{D}_{ph}| = |\mathbb{D}_g| = m$;

- Sum of Features in $\mathbb{D}_m$, $\mathbb{D}_s$, $\mathbb{D}_r$, $\mathbb{D}_{pa}$, $\mathbb{D}_{ph}$, $\mathbb{D}_g$ = Number of Features in $\mathbb{D}_o$ = 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 [17]. There are several reasons why ensemble methods having a better performance [18]:

(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 [18]. For example, Lu, Cheung and Tang [19] 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 [20] 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 [21] 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 [18] 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 pre-processed data.

4.4.1 *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" [22]. 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 [17]:

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

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.

4.4.2 Adapted Classification Methods

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 [22]:

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 [22]. 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.

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

4.4.3 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:

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.

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 on $\mathbb{S}$;
end
Calculate weightings $w_i = \frac{1-E_i}{\sum_{k=1}^{h}(1-E_k)}$;
Obtain the ensemble model $\mathbb{M}_e = \sum_{g=1}^{h} w_i M_i$;

Algorithm 1: Ensemble Modelling

5.1 Experiment Design

In 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 k-fold cross validation algorithm to value the complete predictive performance. The overall look of all experiment proceedings is illustrated in Fig. 4.

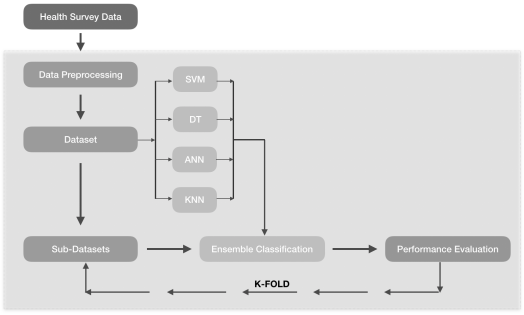

Fig. 4: The overall look of experiment proceedings

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

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

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

Obtain dataset $\mathbb{D}_o$ by data pre-processing on survey data $\mathbb{H}$;
Aggregate features manually referred on 6 psychological functionalities ;
$\mathbb{D}_o = \mathbb{D}_{ph} \cup \mathbb{D}_r \cup \mathbb{D}_m \cup \mathbb{D}_s \cup \mathbb{D}_{pa} \cup \mathbb{D}_g$, for each pair of ($\mathbb{D}_i$, $\mathbb{D}_j$), where both $\mathbb{D}_i$ and $\mathbb{D}_j \in \{\mathbb{D}_{ph}, \mathbb{D}_r, \mathbb{D}_m, \mathbb{D}_s, \mathbb{D}_{pa}, \mathbb{D}_g\}$, $\mathbb{D}_i \cap \mathbb{D}_j = \phi$;
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* $\{\mathbb{D}_o, \mathbb{D}_{ph}, \mathbb{D}_r, \mathbb{D}_m, \mathbb{D}_s, \mathbb{D}_{pa}, \mathbb{D}_g\}$ **do**
    /* *ensemble classification*/
    Do ensemble classification on $\mathbb{D}_i$;
    Validate its predictive performance;
end

Algorithm 2: Experiment Design

5.2 Dataset

5.2.1 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.

5.2.2 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 [6]. 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, Tedders et, al. [27] in 2011 researched the relationship between depression and low cholesterol among household population using

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

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 $\mathbb{R} = \{(x_1, x_2, ..., x_n)\}$ and the testing set $\mathbb{S} = \{y\}$, 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**
        $y_e = 1$ /* *non-depression* */
    else
        $y_e = 0$ /* *depression* */
    end
    Test y_e on 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

NHANES data; Merikangas et, al. [28] in 2012 proposed a association between major depressive disorder and obesity by assessing 2001-2004 NHANES collections; Ubani and Zhang [29] 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. [30] in 2016 explored depressed adults information in social support and health service use of NHANES data; Nguyen et, al. [31] 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. [6], 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".

5.2.3 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.

5.2.4 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. Grouping 98 features into separate functionalities by Definition. 4, we have six sub-datasets (see table. 1).

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

TABLE 1: Features and Sub-datasets

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.

5.3.1 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. 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.

5.4 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)**. F1 score is a balanced measure of both the precision (PPV) and the recall (TPR) of the validation:

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

6 Results and Discussions

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 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:

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

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

TABLE 2: Performance and weights for sub-models

Accuracy is the fraction of predictions our model got right. It indicates the number of correct predictions made in all occurrences of both labels. In our experiment, it presents all corrected prediction based on the result of PHQ-9 metal screen inventory. 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. Recall is the ability of a classifier to find all positive instances. It measures how many mental healthy people are correctly identified. As features and specificity of the overall dataset and each sub-datasets varied, the divided performances are expected (see table.3, 4, 5.).

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

TABLE 3: Features and performances of ensemble classifier

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 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	0.921	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 in Accuracy

Unsurprisingly, ensemble classifier performs better comparing to all baselines, as F1 score 0.976 VS 0.959 (SVM/DT best), Accuracy 0.954 VS 0.925 (DT best). Performances in functionality subsets is compromised in this experiment but is still comparable to other machine learning methodologies [11] [10] [12] [13]. The prediction performance in mental functionality subset is close to the whole dataset but having much less features involved. It may approve that features for metal functionality is more depression-related than other features in other categories because non-criteria items in depression scale decreased specificity of performance [2]. As Recall measures are the successful rate of non-depressive predictions and are almost equal to 1 in the experiment (see table. 3), ensemble classifier is absolutely successful in the prediction of non-depressive cases.

6.2 Discussions

Ensemble classifier is obviously superior than baseline models as performing advantageous at F1 measure and Accuracy.(see table. 2, 3.). It leads not only in the test of overall dataset but also all experiment in sub-datasets (see table.4, 5.). 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 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 [18].

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 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, 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 P_p and percentage of non-depressed instances as N_1, the overall prediction P_o of depressed instances is calculated as below:

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

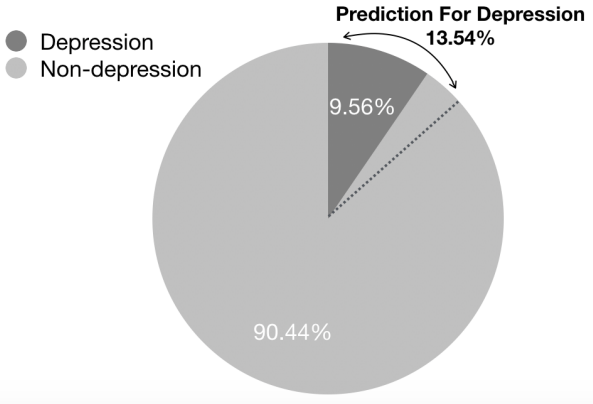

Fig. 5: Coverage in correct classification of depression

The coverage (see fig. 5) of depression cases is slightly larger than the real situation of mental health inventory. However, it is absolutely 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 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.

In comparison with the predictions in the different sub-datasets (see table. 3), ensemble classifier performs the best in the overall dataset and has a similar accuracy in mental subset. The importance of diagnostic criteria in mental symptoms has been manifested that mental criteria are certainly the major features for identifying depression. Meanwhile, both accuracy and F1 measures for predicting depression in physical, social and role are equal to the predictive performance in general subset. However, general subset (70 features) has much more features than physical (7 features), social (6 features) and role (9 features) three subsets. It means that there are a lot of features in general subset occurred without enough specificity for classifying depressed and non-depression labels. Partial of general functional features hence are useless in the detection of depression. Correspondingly, features in physical, social and role functional subsets present 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 metal 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 system which is able to preliminarily distinguish depressive cases from a wide range of health data without the missing identification of any potential depressed case. In the experimental evaluation on NCHS dataset, only 4% cases were mistakenly classified into depressed class and no depressed case has been detected incorrectly. The ensemble classifier on the whole dataset has a high F1 measure as 0.976 comparing to 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 of dataset. Moving forward, this research can help in preliminary screening of depressive cases from a large number of potential cases before formal clinical diagnosis. It is able to save huge cost in the psychological healthcare. It provides a much more efficient way to screen more people than traditional technologies and has a similar accuracy and coverage as current mental scale inventory. However, the reliability and sensitivity of this ensemble system need to be tested next on more datasets and text mining on social network platforms. In the conclusion, the presented work clearly illustrated the predictive capability of the proposed ensemble system to efficiently distinguish depressive instances from a board span of healthcare occasions.

As future work, there are several future research directions of detecting depression that need to be carefully analysed for applying our ensemble system. One trend is to utilise rich online social media sources to extract features on the classification. It certainly will help improve the reliability and sensitivity of ensemble system. It hence will be our next work on the schedule. Another interesting research direction is to examine images and brain EEG signals for mental clinic purpose. Deep learning has gained more momentum in such areas. Current ensemble system involves four supervised learning algorithms, however it will not suit for deep learning. In order to detect depression in intricate circumstance, deep learning technologies like DNN will integrate to increase the performance of ensemble system in future. As a major issue of automate depression detection, we will also consider how to simulate the proceeding of clinic diagnosis to improve the actions of effectively selecting mental diagnostic features.

References

[1] A. Wongkoblap, M. A. Vadillo, and V. Curcin, "Researching mental health disorders in the era of social media: Systematic review," *J Med Internet Res*, vol. 19, no. 6, p. e228, 2017, wongkoblap, Akkapon Vadillo, Miguel A Curcin, Vasa eng Review Canada J Med Internet Res. 2017 Jun 29;19(6):e228. doi: 10.2196/jmir.7215. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28663166

[2] M. Zimmerman and W. Coryell, "The inventory to diagnose depression (idd): A self-report scale to diagnose major depressive disorder," *Journal of Consulting and Clinical Psychology*, vol. 55, no. 1, pp. 55–59, 1987, u Iowa Coll of Medicine, Iowa City. [Online]. Available: http://ezproxy.usq.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true\&db=pdh\&AN=1987-14494-001\&site=ehost-live

[3] R. Huerta-Ramírez, J. Bertsch, M. Cabello, M. Roca, J. M. Haro, and J. L. Ayuso-Mateos, "Diagnosis delay in first episodes of major depression: A study of primary care patients in spain," *Journal of Affective Disorders*, vol. 150, no. 3, pp. 1247–1250, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165032713004862

[4] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, "Recognizing depression from twitter activity," pp. 3187–3196, 2015. [Online]. Available: https://dl.acm.org/citation.cfm?doid=2702123.2702280

[5] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," pp. 47–56, 2013. [Online]. Available: https://dl.acm.org/citation.cfm?doid=2464464.2464480

[6] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The phq-9: validity of a brief depression severity measure," *J Gen Intern Med*, vol. 16, no. 9, pp. 606–13, 2001, kroenke, K Spitzer, R L Williams, J B eng Research Support, Non-U.S. Gov't Validation Studies 2001/09/15 10:00 J Gen Intern Med. 2001 Sep;16(9):606-13. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/11556941https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1495268/pdf/jgi_01114.pdf

[7] K. Sakado, T. Sato, T. Uehara, S. Sato, and K. Kameda, "Discriminant validity of the inventory to diagnose depression, lifetime version," *Acta Psychiatrica Scandinavica*, vol. 93, no. 4, pp. 257–260, 1996.

[8] Y. Ophir, C. S. C. Asterhan, and B. B. Schwarz, "Unfolding the notes from the walls: Adolescents' depression manifestations on facebook," *Computers in Human Behavior*, vol. 72, pp. 96–107, 2017.

[9] D. Mowery, H. Smith, T. Cheney, G. Stoddard, G. Coppersmith, C. Bryan, and M. Conway, "Understanding depressive symptoms and psychosocial stressors on twitter: A corpus-based study," *Journal of Medical Internet Research*, vol. 19, no. 2, 2017.

[10] A. U. Hassan, J. Hussain, M. Hussain, M. Sadiq, and S. Lee, "Sentiment analysis of social networking sites (sns) data using machine learning approach for the measurement of depression," pp. 138–140, 2017.

[11] I. Fatima, H. Mukhtar, H. F. Ahmad, and K. Rajpoot, "Analysis of user-generated content from online social communities to characterise and predict depression degree," *Journal of Information Science*, vol. 44, no. 5, pp. 683–695, 2018. [Online]. Available: http://journals.sagepub.com/doi/abs/10.1177/0165551517740835

[12] Z. Peng, Q. Hu, and J. Dang, "Multi-kernel svm based depression recognition using social media data," *International Journal of Machine Learning and Cybernetics*, 2017.

[13] A. G. Reece and C. M. Danforth, "Instagram photos reveal predictive markers of depression," *EPJ Data Science*, vol. 6, no. 1, p. 15, 2017. [Online]. Available: https://doi.org/10.1140/epjds/s13688-017-0110-zhttps://epjdatascience.springeropen.com/track/pdf/10.1140/epjds/s13688-017-0110-z

[14] M. M. Clark, K. L. Bradley, S. M. Jenkins, E. A. Mettler, B. G. Larson, H. R. Preston, J. T. Liesinger, B. L. Werneburg, P. T. Hagen, A. M. Harris, B. A. Riley, K. D. Olsen, and K. S. Vickers Douglas, "The effectiveness of wellness coaching for improving quality of life," *Mayo Clinic Proceedings*, vol. 89, no. 11, pp. 1537–1544, 2014.

[15] G. V. Ostir, I. M. Berges, A. Ottenbacher, and K. J. Ottenbacher, "Patterns of change in depression after stroke," *J Am Geriatr Soc*, vol. 59, no. 2, pp. 314–20, 2011, Comparative Study Multicenter Study Research Support, N.I.H., Extramural J Am Geriatr Soc. 2011 Feb;59(2):314-20. doi: 10.1111/j.1532-5415.2010.03266.x. Epub 2011 Jan 28. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21275930

[16] M. Gonzalez-Saenz de Tejada, A. Bilbao, M. Baré, E. Briones, C. Sarasqueta, J. M. Quintana, A. Escobar, and M. Baré, "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*, vol. 25, no. 8, pp. 891–897, 2016, psycho-Oncology Source Information: Aug2016, Vol. 25 Issue 8, p891; [Online]. Available: http://ezproxy.usq.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true\&db=pbh\&AN=117000366\&site=ehost-live

[17] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1, pp. 1–39, 2010. [Online]. Available: https://doi.org/10.1007/s10462-009-9124-7

[18] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1249

[19] Y. Lu, Y. M. Cheung, and Y. Y. Tang, "Hybrid sampling with bagging for class imbalance learning," R. Wang, J. Bailey, T. Washio, J. Z. Huang, L. Khan, and G. Dobbie, Eds. Springer Verlag, 2016, vol. 9651, pp. 14–26,[Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84963994580\&doi=10.1007%2f978-3-319-31753-3_2\&partnerID=40\&md5=c4e52d673e51d1db80beff81cf56e44c

[20] N. Limsetto and K. Waiyamai, "Handling concept drift via ensemble and class distribution estimation technique," ser. Advanced Data Mining and Applications. Springer Berlin Heidelberg, Conference Proceedings, pp. 13–26.

[21] G. Serpen and S. Pathical, "Classification in high-dimensional feature spaces: Random subsample ensemble," in *2009 International Conference on Machine Learning and Applications*, Conference Proceedings, pp. 740–745.

[22] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: a review of classification and combining techniques," *Artificial Intelligence Review*, vol. 26, no. 3, pp. 159–190, 2006. [Online]. Available: https://doi.org/10.1007/s10462-007-9052-3

[23] M. Somvanshi and P. Chavan, "A review of machine learning techniques using decision tree and support vector machine," in *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, Conference Proceedings, pp. 1–7.

[24] R. Kumar, *Fundamental of artificial neural network and fuzzy logic*, ser. Fundamentals of artificial neural network and fuzzy logic. New Delhi, India: University Science Press, An Imprint of Laxmi Publications Pvt. Ltd., 2010.

[25] Y. Fei and W.-q. Li, "Improve artificial neural network for medical analysis, diagnosis and prediction," *Journal of Critical Care*, vol. 40, p. 293, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0883944117308468

[26] R. Choudhary and H. K. Gianey, "Comprehensive review on supervised machine learning algorithms," in *2017 International Conference on Machine Learning and Data Science (MLDS)*, Conference Proceedings, pp. 37–43.

[27] S. H. Tedders, K. D. Fokong, L. E. McKenzie, C. Wesley, L. Yu, and J. Zhang, "Low cholesterol is associated with depression among us household population," *Journal of Affective Disorders*, vol. 135, no. 1-3, pp. 115–121, 2011.

[28] A. Merikangas, P. Mendola, P. Pastor, C. Reuben, and S. Cleary, "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*, vol. 35, no. 2, pp. 149–154, 2012.

[29] C. C. Ubani and J. Zhang, "The role of adiposity in the relationship between serum leptin and severe major depressive episode," *Psychiatry Research*, vol. 228, no. 3, pp. 866–870, 2015.

[30] S. B. Andrea, S. A. R. Siegel, and A. R. Teo, "Social support and health service use in depressed adults: Findings from the national health and nutrition examination survey," *General Hospital Psychiatry*, vol. 39, pp. 73–79, 2016.

[31] B. Nguyen, P. Weiss, H. Beydoun, and V. Kancherla, "Association between blood folate concentrations and depression in reproductive aged u.s. women, nhanes (2011–2012)," *Journal of Affective Disorders*, vol. 223, pp. 209–217, 2017.

PLACE PHOTO HERE

Xiaohui Tao Biography text here.