Transfer Learning for Depression Detection on Social Networks*

Oliver Chi[1] and Xiaohui Tao[1]

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

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

Keywords: psychological knowledge base · ensemble classification technique · supervised learning · depression.

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

1 Introduction

2 Related Work

2.1 Psychological Studies

2.2 Classification Research on Depression

3 Definitions/Research Problem

The research aims to design an effective classification method for automatically detecting depressive risk of users which has a potential to implement in the real environment of social network.

.... 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' ∈ 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* $\mathbb{V}$*:*

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

4 Framework

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

4.1 Conceptual Design

In this study, the framework consists of three modules:

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

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

4.2 Psychological Knowledge Base

Kroenke et, al. [PHQ-9] discovered that there was a strong association between increasing depression severity screen scores and worsening functionality on all 6 variables: mental, social, role, pain, physical and general functions. The research illustrated graphically the relationship between increasing PHQ-9 scores of depression and worsening functional variables (see Fig. 2).

Associations of health related functionings with depression have been observed in many previous studies at psychological domain. For instance, Clark et, al. [Clark, Bradley] explored the opposite association of depression and psychosocial functionings. The research examined the potential psychosocial benefits of wellness coaching in functionings which included the overall quality of life and the 5 domains of physical, social, emotional, cognitive, and spiritual functioning. It found that depression is associated with poor health status and negative health behaviours. It also addressed that participants significantly reduced their level

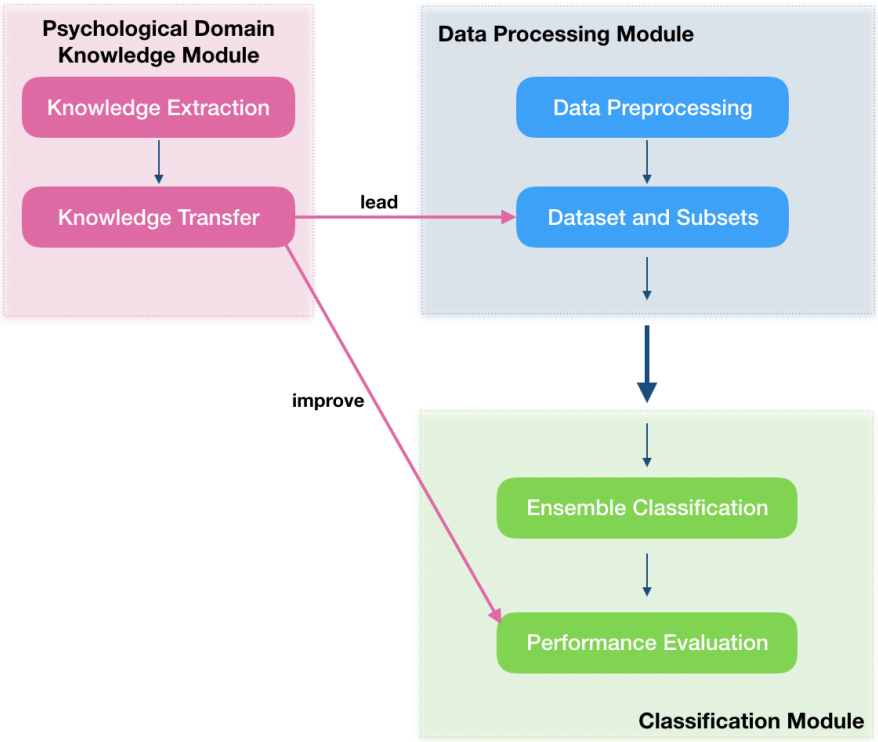

Fig. 1. Conceptual Framework

of depression after improving health functional status by wellness coaching. The researchers suggested that additional self-care on physical activity, health sleep, spirituality and social activities could help on long-term depression management. Ostir et, al. [Ostir] discovered that patients identified as not depressed showed greater improvement in functional status than other patient groups in stoke disease.The research varied previous reports on the association between depression and functional status. It suggested that early recognition and management of depression in person with stroke represents an important effort to improve health outcomes and facilitate functional independence. Moreover, Gonzalez-Saenz de Tejada et, al. [Gonzalez-Saenz de Tejada] explored the association of functional and psychological status of cancer patients. The study addressed that patients with depression showed lower gains in all health related functional domains than patients without depression. It confirmed again that patients with depression tended to show less improvement in all functional variables in health related quality of life (physical, role, emotional, cognitive, and social function, and global quality of life). And it also confirmed that depression were associated with changes in at least one pre-noted functional variable.

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

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

** SF-20 scores are adjusted for age, gender, race, education, study site, and number of physical disorders. Point estimates for the mean as well as 95% confidence intervals (±1.96 × standard error of the mean) are displayed.*
Most pairwise comparisons of mean SF-20 scores between each PHQ-9 level within each scale are significant at P < 0.05 using Bonferroni's correction for multiple comparisons. Only those pairwise comparisons that share a common superscript letter (a, b, or a,b) are not significant.

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

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 health-related functional conditions. 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 narrowed down to 6. The dataset of user mental profile need to be redefined:

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

- u_{mental} presents mental functional variables;
- u_{social} presents social functional variables;
- u_{role} presents role functional variables;
- u_{pain} presents pain functional variables;
- $u_{physical}$ presents physical functional variables;
- $u_{general}$ presents the overall functional variables.

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 functionality classes 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 classes of health functionings (see Fig. 3). They may involve different number of questions but they all have the same participants. Furthermore, those six tables can be rejoin into one big table due to same row index of them. By instant consideration of those tables, each table forms a new dataset where participants are cases and questions are features. We can therefore define the new datasets after data pre-processing as below:

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

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

4.4 Modelling

In this study, we use an ensemble classification approach to build the model for detecting depression. It implements the independent ensemble methodology which applies several classification techniques in parallel. They are Support

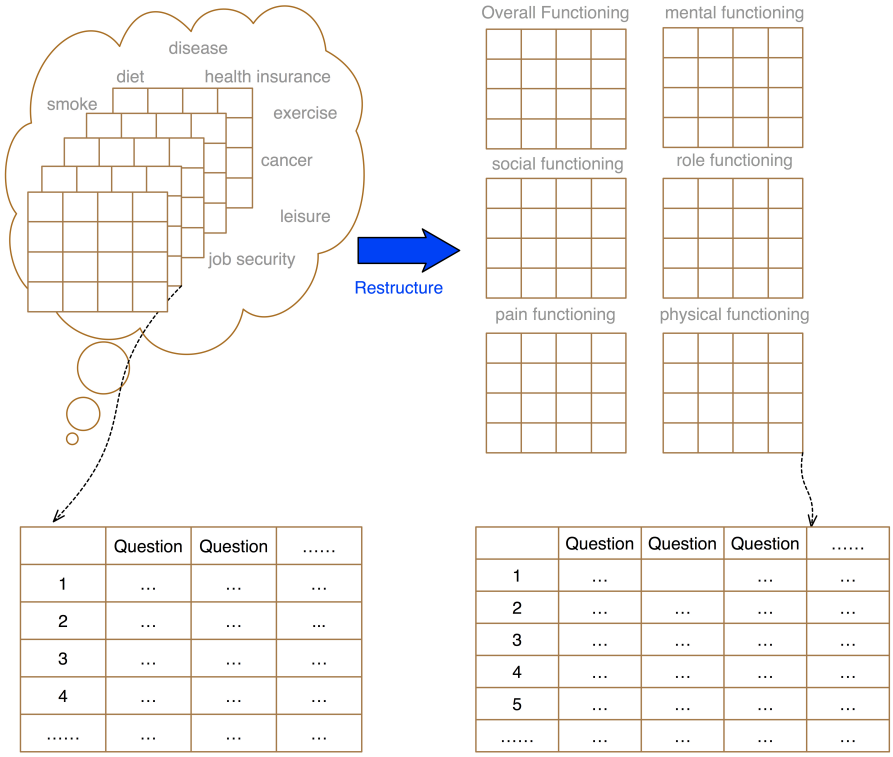

Fig. 3. Data Restructure based on Psychological Knowledge Base

Vector Machine (SVM) technique, Artificial Neutral Network(ANN) algorithm, K-Nearest Neighbour (KNN) method and Decision Tree (DT) method. Each composite classifier among them is trained on the same portion of the training set in one run. The performance of them are evaluated by k-fold cross validation algorithm. And amalgamating all outputs of composite classifiers into a single prediction, we consequently generates the ensemble classifier. The main idea of this ensemble classification approach is to collect various outputs of multiple independent classifiers and combines them to improve the predictive performance.

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

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

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

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

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

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

Baseline Classification Technique

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

a) Logic based algorithm: The algorithms use logic or rules to make a decision of selecting proper features during the learning. Decision tree method adopts this algorithm.

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

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

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

Artificial Neutral Network (ANN)
Artificial neutral network is a biologically inspired algorithm to simulate the manner of nerve cells in the brain. According to the book of Kumar [Kumar], ANN is made up of elements named as artificial neurons; the neurons are organised in network to simulate the anatomy of brain by a standard processing whose output is calculated by multiplying its input by a weight vector; they are aggregated into layers and layers are aggregated into the network to form highly interconnected processing structures; whilst the input layer doesn't process information, it simply sends the inputs, modified by a weight, to each of

the neurons in the next layer; and the next layer does the processing which can be a hidden layer or the output layer in a single layer design. ANNs are usually more able to easily provide incremental learning than decision trees as having a good multiple layers architecture. Fei and Li [Fei, Li] discovered that ANN are widely used in medical data mining methodology, and the combination of ANNs and some other algorithms will be able to achieve a better results in medical diagnosis and prediction. Likewise, ANN contains some weaknesses, including "poor general application of the architecture, inaccurate analysis for various indicators of the network and uncontrollable time of machine learning" [Fei, Li]. Kotsiantis et, al. [Kotsiantis] concluded that the most striking disadvantage of ANN is lack of ability to answer how the output in a specific way being effectively communicated. Generally, it is a problem to properly determining the size of the hidden layer. The underestimated neutrons in hidden layer can lead to poor approximation, while "excessive nodes can result in overfitting and eventually make the search for the global optimum more difficult" [Kotsiantis]. In spite of its several disadvantages, ANN is still a good competitor for other learning algorithms, which has been used on a variety of intricate problems including computer vision, speech recognition, recommendation filtering even medical diagnosis [Somevanshi][Kumar].

K-Nearest Neighbour (KNN)

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

Support Vector Machine (SVM)

Support Vector Machine algorithm is the newest classification technique among the proposed methods. In classification, SVM constructs a hyperplane or set of hyperplanes in the dimensional space; the hyperplane separates the training data into diverse two classes; and a good classification is achieved by the hyperplane's capability to make a larger margin between two classes of training data. As the application environment is dimensional space, input data of SVM are paired into vectors and vectors are defined in terms of a kernel function. Selection of proper hyperplane and proper parameters for kernel function gives more accurate results as compared to neural networks [Somevanshi]. As the model complexity of an SVM is unaffected by the number of features encountered in the training data,

SVM is well suited to deal with learning dataset with large number of features and training instances. Also, choice of an appropriate kernel leads to different SVM applications in linear, nonlinear and multiclass classification. The potential drawbacks of SVM are addressed including [Kotsiantis]: a) requiring full label of input data; b) being difficult to interpret parameters of the solved model; c) being unsuitable for non-binary multiple classification problems.

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

Ensemble Model

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

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

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

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

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

Definition 6 Let the ensemble model

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

where

- M_i presents a single base model;
- w_i presents the weighting metric of predictive performance at specific base model M_i;

- k is the order of base models;
- n is the total number of base models, and in our case $n = 4$;
- i is the order number of specific base model.

Algorithm

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

input : Dataset $\mathbb{D} = \left\{(x_1, x_2, ..., x_n, y), x_i \in R^m, y \in \{0,1\}^m\right\}$
output: Ensemble Model $\mathbb{M}_e$

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

Algorithm 1: Ensemble Modelling

5 Experiment

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

5.1 Experiment Design

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

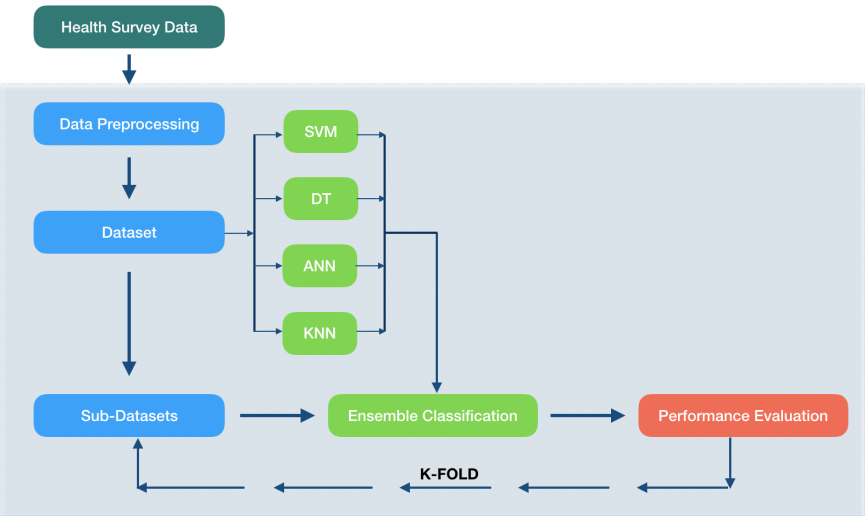

Fig. 4. The overall look of experiment proceedings

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

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

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

Algorithm 2: Experiment Design

The ensemble classification can be expressed in algorithm as well: Given a well-preprocessed dataset of m examples and n features $\mathbb{D} = \{(x_1, x_2, ..., x_n, y), x_i \in R^m, y \in \{0,1\}^m\}$, 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 : 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**
    $y_e = 1$;
  else
    $y_e = 0$;
  end
  Test y_e on the testing set $\mathbb{S}$ and report predictive performance p_k ;
end
/ 5-fold cross validation */*
Validate the predictive performance by calculating $p_e = \frac{\sum_{k=1}^{5} p_k}{5}$;
Generate the optimised ensemble classifier $\mathbb{F}_e = \text{Median}(\mathbb{F}_1, \mathbb{F}_2, \mathbb{F}_3, \mathbb{F}_4, \mathbb{F}_5)$

Algorithm 3: Ensemble Classification Procedure

5.2 Dataset

NHANES Survey Data

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

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

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

Build Ground Truth

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

Overall Dataset

5.3 Baseline Models

Decision Tree

Artificial Neutral Network

K-Nearest Neighbour

Support Vector Machine

5.4 Performance Measure

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

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

And four following terminologies are hence defined as below:

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

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

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

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

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

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

d) **Accuracy (ACC)**:

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

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

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

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

6 Results and Discussions

$\mathbb{D} = \mathbb{D}_{mental} + \mathbb{D}_{social} + \mathbb{D}_{role} + \mathbb{D}_{pain} + \mathbb{D}_{physical} + \mathbb{D}_{general}$

6.1 Experimental Results

Sub-Model Performance

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

Ensemble Classifier

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

Models	Overall (mean)	Physical (mean)	Role (mean)	Social (mean)	Mental (mean)	Pain (mean)	General (mean)
SVM	0	0.9498	0	0	0	0	0
ANN	0	0.9437	0	0	0	0	0
KNN	0	0.9468	0	0	0	0	0
DT	0	0.9496	0	0	0	0	0
Ensemble	0	0.9658	0	0	0	0	0

Table 1. F1 Measure

Predictive Performance

6.2 Discussions

Models	Overall (mean)	Physical (mean)	Role (mean)	Social (mean)	Mental (mean)	Pain (mean)	General (mean)
SVM	0	0.9044	0	0	0	0	0
ANN	0	0.8954	0	0	0	0	0
KNN	0	0.9002	0	0	0	0	0
DT	0	0.9047	0	0	0	0	0
Ensemble	0	0.9343	0	0	0	0	0

Table 2. Accuracy

Models	Overall (mean)	Physical (mean)	Role (mean)	Social (mean)	Mental (mean)	Pain (mean)	General (mean)
SVM	0	0.9044	0	0	0	0	0
ANN	0	0.9187	0	0	0	0	0
KNN	0	0.9147	0	0	0	0	0
DT	0	0.9100	0	0	0	0	0
Ensemble	0	0.9348	0	0	0	0	0

Table 3. Precision

7 Analysis

7.1 Sensitivity Study

Models	Overall (mean)	Physical (mean)	Role (mean)	Social (mean)	Mental (mean)	Pain (mean)	General (mean)
SVM	0	1.0000	0	0	0	0	0
ANN	0	0.9703	0	0	0	0	0
KNN	0	0.9812	0	0	0	0	0
DT	0	0.9928	0	0	0	0	0
Ensemble	0	1.0000	0	0	0	0	0

Table 4. Recall

8 Conclusion and Future Work

References

1. R. L. Spitzer, K. Kroenke, and J. B. Williams, "Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire," JAMA, vol. 282, no. 18, pp. 1737-44, Nov 10 1999.
2. 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, Sep 2001.
3. M. Résibois, P. Kuppens, I. Van Mechelen, P. Fossati, and P. Verduyn, "Depression severity moderates the relation between self-distancing and features of emotion unfolding," Personality and Individual Differences, vol. 123, pp. 119-124, 2018.
4. B. T. Stegenga et al., "Depression, anxiety and physical function: exploring the strength of causality," Journal of Epidemiology and Community Health, vol. 66, no. 7, pp. e25-e25, 2012.
5. H. Xiao, J. Y. Yoon, and B. Bowers, "Living Arrangements and Quality of Life," Western Journal of Nursing Research, vol. 38, no. 6, pp. 738-752, 2015.
6. Rokach, L.: 'Ensemble-based classifiers', Artificial Intelligence Review, 2010, 33, (1), pp. 1-39.