EASR

Engineering and Applied Science Research

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

Comparison of capability of data classification models to predict consistent results for depression analysis based on user-behaviour tracking and facial expression recognition during PHQ-9 assessment

Natratanon Kanraweekultana[1)], Sajjaporn Waijanya*[1)], Nuttachot Promrit[1)], Undaman Nopnapaporn[1)], Apisada Korsanan[2)] and Sansanee Poolphol[2)]

[1)]Center of Excellence in AI and NLP, Department of Computing, Faculty of Science, Silpakorn University, Sanam Chandra Palace Campus, Nakhon Pathom 73000, Thailand
[2)]Psychiatry and Drug Addiction Subdivision, Somdetphraphutthaloetla Hospital, Samut Songkhram Province 75000, Thailand

Received 4 March 2023
Revised 25 October 2023
Accepted 27 October 2023

Abstract

This research compares the capability of data classification models to predict consistent results for a subject's depression potentiality, track the subject behaviour and recognise facial expressions during PHQ-9 assessments. This research is motivated by the necessity for depression screening and diagnosis, which traditionally relies on observations by experienced physicians or clinical psychologists of symptoms in conjunction with data from questionnaires. However, the field still requires a suitable technological approach that gives more accurate and consistent results. All data used in the present research were collected by combining technologies and compared by using classification models, the goal being to find the machine-learning model that most accurately predicts consistent results for the subjects' PHQ-9 assessment, behaviours and emotions. The subjects were screened by clinical psychologists and divided into three groups: (i) subjects suffering from depression but not receiving treatment (undertreated subjects), (ii) subjects undergoing depression treatment (subjects undergoing treatment) and (iii) subjects without depression disorder (normal subjects). Related studies have compared the accuracy of classification models to one another. The four most frequently applied classification models in depression-related studies are (i) decision tree (ii) support vector machine, (iii) naïve Bayes and (iv) neural network. All models were analysed, designed and developed before being tested experimentally. The accuracy of the experimental results was tested by using the data analysis tool RapidMiner Studio. The results show that the decision tree model is not only the most accurate for predicting depression potentiality, tracking behaviour and recognising facial expressions during PHQ-9 assessments but also the most suitable.

Keyword: Efficiency comparison, Classification models, PHQ-9, Depression

1. Introduction

Major depression is a psychiatric disorder whereby patients often exhibit severe emotional distress, possibly leading to suicidal behaviour. According to the World Health Organization, approximately 1 million people with depression die by suicide each year, and about 340 million people around the world suffer from depression. The characteristics of depression and some of its limitations are huge obstacles to accurate diagnosis. A precise diagnosis requires experienced physicians or clinical psychologists with well-designed questionnaires and behavioural observations [1, 2]. In addition, the accuracy of the diagnosis depends on the patient's collaboration [3]. Depression assessment can depend on the subject's age, educational level, gender or even heredity. These factors can prompt the subjects to reply to questionnaires with drastically distorted answers, causing the data obtained to be distorted and not accurately portray the subject's psychological condition.

The connectivity of nerves is bidirectional, meaning that the nerves stimulate muscle contraction when signalled by the brain and also communicate information back to the brain. Nearly all facial muscles are innervated by the facial nerve (also called cranial nerve VII), which produces normal facial expressions or a particular type of abnormality that can be transmitted through a person's facial features during assessment [4]. Such abnormalities are beyond control and are indicative of depression. In addition, facial expressions cannot be measured or evaluated by PHQ-9 assessment. Facial expressions and behaviour must be diagnosed by a physician or clinical psychologist.

Systems have been developed to determine depression potentiality, monitor behaviour and recognise facial expressions during PHQ-9 assessment to improve the assessment's accuracy and assist professionals in diagnosing their patients' clinical symptoms before offering them medical treatment [5]. Four clinical symptoms serve as criteria for depression classification: (i) emotional symptoms, (ii) neurovegetative symptoms, (iii) psychomotor symptoms and (iv) cognition symptoms [6]. Although diagnoses are conducted by clinical psychologists or experienced physicians based on patients' PHQ-9 scores, behaviour and emotions, the data are likely to be inconsistent because patients undergoing PHQ-9 assessment may avoid answering truthfully, which can cause the system to malfunction. This tendency limits screening for depression and produces data that may not reduce the workload of medical personnel.

*Corresponding author.
Email address: waijanya_s@silpakorn.edu
doi: 10.14456/easr.2024.2

The purpose of this research is to compare the predictive capability of data classification models by using machine learning to predict PHQ-9 scores, behaviour and emotions and analyse the consistency between these results. The data obtained from the subjects are used to train the models, which are then used to improve the prediction and classification of behavioural symptoms of the individuals being assessed. The most accurate model may be used by a person with no psychological knowledge and may also reduce the workload of mental health and medical professionals. Thus, the objective of this research is to compare the predictive capability of data classification models and the consistency between the predicted depression symptoms (e.g. behavioural tendencies and facial expressions during PHQ-9 assessment).

2. Materials and methods

2.1 Related work

Major depression disorder (MDD) is one of the most commonly encountered mental illnesses and the second leading cause of disability worldwide. Major symptoms of MDD are chronic depressive emotion, hypoxia, decreased cognitive functioning, and, especially in severe cases, suicidal tendencies. These symptoms all shift a considerable burden onto the economy and society. Traditionally, clinical psychiatrists and psychologists have diagnosed MDD based on patients' mental conditions, symptoms and personal experiences, possibly leading to excessively broad or varied diagnoses [7]. Because human brains contain complex neural networks with a multitude of interconnected structures and functions, the pathological underpinnings of neuropsychiatric diseases require effective screening, especially for depressive people with interactive indicators such as depressive behaviour or facial expressions. Consequently, more effective screening methods have been developed by applying artificial intelligence, machine learning and deep learning.

Liu et al. [8] presents a depression-screening method that takes the form of a non-verbal self-association task and provides data relating to emotions and behaviour obtained from the subject's brain via non-verbal screening, thereby identifying the differences in physiology. This non-verbal technique supports depression screening at the individual level by using Lasso-type regularisation and the nonlinear gradient boosting model in conjunction with the Hamilton Depression Scale. This system is related to the one developed by Kanraweekultana et al. [5] to diagnose depression potentiality via behaviour tracking and facial expression recognition during PHQ-9 assessment. Souza Filho et al. [9] developed a screening tool to assist depression diagnosis by applying artificial intelligence in psychiatry and analysed the accuracy of their machine-learning algorithms for assessing MDD patients based on clinical, laboratory and social science data. Likewise, Zheng et al. [3] studied depression classification based on treatment determined by brain functional networks. Zheng pointed out that diagnosing MDD is challenging because it mainly depends both on the patient's collaboration and the psychiatrist's experience. Moreover, inexperienced psychiatrists correctly diagnose depression in only 50% of their patients. Thus, resting-state functional magnetic resonance imaging (rs-fMRI) and machine-learning algorithms were applied together for depression classification.

Related studies indicate that the different experiences and expertise of physicians and clinical psychologists in depression diagnosis are based on the cooperation or variability of the people they assessed for depression. Such differentiation led to other studies that aimed to classify differences or consistencies between specific conditions of interest. For example, Dong et al. [10] studied emotional abnormalities by applying the logistic regression model to categorise and analyse the data in the MATRICS Consensus Cognitive Battery to classify psychiatric disorders and emotional illnesses. Furthermore, machine-learning techniques were applied to classify stress, depression and anxiety by feature classification algorithms based on the same feature to compare the results and prove their accuracy.

Psychiatrists have applied the principal component analysis, gradient boosting algorithm, dimensionality reduction algorithm, K-nearest neighbour (KNN), decision tree, naïve Bayes and support vector machines (SVM) [11]. Likewise, Yang et al. [1] classified depression by using SVM, KNN and decision tree models for binary classification. Using a confusion matrix, the classification models were evaluated based on four cases: true-positive, false-positive, true-negative and false-negative. The evaluation metrics used to compare the classification performance of the models were accuracy, specificity, sensitivity, precision and F1 score. The study aimed to test the accuracy with which the models classified depression. The result showed that SVM was the most accurate (94.03%). The experimental groups consisted of individuals with depression, while the control group consisted of normal individuals. Similarly, Price et al. [12] collected data from three groups of subjects: (i) individuals with psychiatric disorders, (ii) individuals with depression and (iii) individuals in good health. The data collected were analysed using unsupervised machine learning via clustering methods [12].

In the medical field, machine-learning models were developed for diagnostic purposes. Various machine-learning algorithms, such as decision tree, multilayer perceptron, naïve Bayes, random forest and SVM, were used to build these models and evaluate their performance. The highest accuracy achieved was 88.90% [13]. In similar work, Joshi and Kanoongo [14] worked on depression detection by using artificial intelligence together with machine learning and compared the performance of several models for detecting and analysing emotions and depression, including naïve Bayes, SVM, long short-term memory, radial neural networks, logistic regression and linear support vector, which are all suitable to assess depression.

Four machine-learning model patterns were also used: probabilistic, nearest neighbour, neural network and tree-based patterns to assess anxiety, depression and stress [15]. In addition, Huang et al. [16] compared three models: random forest, SVM and decision tree, to predict suicidal ideation and depression among Chinese adolescents. Their research findings are relevant to social science and appropriate for medical prediction related to depression. Studies that compare the accuracy of depression-classification models include that of Aleem et al. [17], who summarised and compared machine-learning algorithms for deep-learning-based depression diagnosis. They used three supervised learning classifiers: (i) classification, (ii) regression and (iii) deep learning. The classification group included four models: naïve Bayes, KNN, neural network, SVM classifier and decision tree. For the deep-learning group, they concluded that their most suitable and frequently used neural network models for practical deep-learning-based depression diagnosis were SVM classifier and decision tree [17].

Thus, related studies have compared the capability of data classification models to predict data, suitability and classification accuracy of experimental results obtained from depression assessments, behaviour tracking and facial expression recognition during PHQ-9 assessment. These studies focused on the consistency between PHQ-9 assessment scores [18], symptoms and emotions by applying both machine- and deep-learning algorithms. In all the related studies, the four most used classification models in depression-related studies are (i) decision tree, (ii) SVM, (iii) naïve Bayes and (iv) neural network (see Table 1).

Table 1 Summary of machine-learning classification models used in related studies.

Research	Year	Algorithms					
		Decision tree	Naïve Bayes	Neural Network	SVM	KNN	Logistic Regression
Yang et al. [1]	2023	✓	✓		✓	✓	
Dong et al. [10]	2023						✓
Singh and Kumar [11]	2022	✓	✓		✓		✓
Mahoto et al. [13]	2023	✓	✓	✓	✓		
Joshi and Kanoongo [14]	2022	✓	✓	-	✓		✓
Kumar et al. [15]	2022	✓		✓		✓	
Huang et al. [16]	2022	✓			✓		
Aleem et al.[17]	2022	✓	✓	✓	✓	✓	
Uddin et al. [19]	2022			✓			
Rank		1	3	4	2	5	5

2.2 Methods

The present research used experimentation, data analysis and testing to compare the capability of classification models to predict accurate and consistent results for depression potentiality, behaviour tracking and facial expression recognition during PHQ-9 assessment. The research processes are detailed below.

2.2.1 Problem

Although numerous technologies have been applied for depression screening, participants who undergo depression assessments still deviate or conceal their behaviour, resulting in inaccurate PHQ-9 results that require further analysis from experienced psychologists and clinical psychiatrists. This problem obliges health professionals to study and compare the accuracy of data classification models to predict consistent results for behaviour tracking and facial expression recognition during PHQ-9 assessment. Success in this research would increase the efficiency of the depression-analysis system.

2.2.2 System analysis, design and development

The depression-analysis system used for this research was designed and developed to track subjects' behaviours and recognise facial expressions during PHQ-9 assessment. The system is also based on clinical and technological theories [20]. Figure 1 shows the four categories of the clinical symptoms referred to in this process.

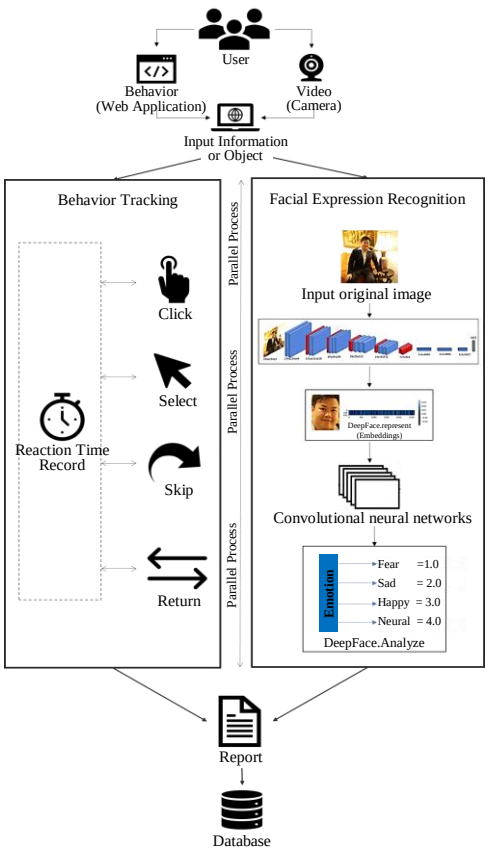


Figure 1 Development Framework

Emotional and cognitive symptoms were analysed by using DeepFace to recognise facial expressions. This approach forms the framework for facial recognition and facial-attribute analysis. Facial recognition has five steps: (i) detect, (i) align, (iii) normalise, (iv) represent and (v) verify. Next, the face-verification function was activated to examine whether a pair of input facial images represents the same individual and is verified by the encoded NumPy images. Subsequently, facial recognition was verified by using DeepFace with an out-of-the-box function. This function finds the input images and returns a list of pandas data frames as the output for the face recognition model. The face images were displayed as multidimensional vectors so that DeepFace need only display the output of the Represent function and return the embedded list. The result is a facial image in the form of a 2622-dimensional vector. To return the array via vector embedding, vectors were horizontally plotted in 2622 squares. Each square is consistent with the size of the embedded vectors. Facial recognition models, which are normal convolutional neural networks for facial-attribute analysis, were created by using DeepFace to predict four emotions: fear = 1.0, sad = 2.0, happy = 3.0 and neutral = 4.0. The most suitable tools obtained from the tool-matching process between the technological tools and the four clinical symptoms: emotion, neurovegetative, psychomotor and cognition [6]. The emotions betrayed by facial expressions were detected by OpenCV, SSD, Dlib, MTCNN, RetinaFace and MediaPipe and were analysed in real-time while a camera recorded the subjects' faces [6, 21-23]. Algorithm 1 details this procedure.

Algorithm 1 Procedure for capturing emotions from facial expressions.

```
Procedure EmotionPercentageCalculate (emotionPredict)

result = predict(string)

        emotion_result=
                {"1","2","3","4","5","6","7","8","9",: {"fear": 0, "happy": 0, "sad": 0, "neutral": 0}

           IF current_emotion = emotion_result THEN
                  emotion_result= current_emotion +1
           ELSE print("[INFO] Emotion not found...") THEN
           ENDIF

        result_percentage =
                {"1","2","3","4","5","6","7","8","9",: {"fear": 0, "happy": 0, "sad": 0, "neutral": 0}

        FOR questionList, emote_dict IN result.items DO
           sum_emote = sum(emote_dict)
        FOR emotion_result IN emote_dict.items DO
                IF result_percentage= (value * 100) / sum_emote
                ELSE result_percentage = 0.0
           ENDFOR
        ENDFOR

        RETURN result_percentage

        emotionPercent =
                {" fear ": 0, "sad": 0, " happy ": 0, " neutral": 0}

                        IF (emotion !== undefined)
                            IF result_percentage["Fear "] < emotion["fear"]))
                                result_percentage["Fear"] = emotion["fear"]
                            ENDIF
                            IF result_percentage["Sad"] < emotion["sad"]))
                                result_percentage["Sad"] = emotion["sad"]
                            ENDIF
                            IF result_percentage["Happy"] < emotion["happy"]))
                                result_percentage["Happy"] = emotion["happy"]
                            ENDIF
                            IF result_percentage["Neutral"] < emotion["neutral"]))
                                result_percentage["Neutral"] = emotion["neutral"]
                            ENDIF
                        ENDIF
                    RETURN emotionPercent

End Procedure
```

Psychomotor, neurovegetative and cognitive symptoms were analysed by user-behaviour tracking technology and PHQ-9 online assessment. For data collection via screen recording, the script used for the screen recording was embedded in advance in a web page used for recording data. Next, when the embedded script was called from the web page, the data obtained from the subjects' (or the users') behaviour were recorded in the database. At the same time, all behaviour was collected as data by jQuery, which is a JavaScript Library used to detect each click. If the input or the object is clicked, the data are saved in the database with the previous clicks. The time lapse between the previous and current clicks was calculated to reflect the users' decision tendencies for each question. All these processes ran continuously until the PHQ-9 assessment was completed [24, 25], as detailed in Algorithm 2.

Algorithm 2 Behaviour tracking algorithm.

```
Procedure behaviorCalculate (userAction)

    handleOnMouseLeave(value):
        index = value.index
        totalTime = TimeEnter - TimeExit
        currentHoverTime = hoverTime + totalTime

        IF currentHoverTime > 60 THEN
           SET currentHoverTime.behaver.over = true
        ENDIF

    behaviorAnswer (answerValue):
        value.behaver = answerValue. checkedValue
        answerValue. checkedValue = Number(value)

        IF value.behaver != -1 THEN
            answerValue.behaver.change = true
        ENDIF

            IF key < index AND answerValue == -1 THEN
                answerValue.behaver.skip = true
            ELSE IF key == index THEN
                RETURN
            ENDIF

    behavior = {Change: "N", Skip: "N", Overtime: "N", Submit: "N"}

                IF(value.behaver !== {})
                    IF(value.behaver.change === true)
                        behavior.Change = "Y"
                    ENDIF
                    IF(value.behaver.skip === true)
                        behavior.Skip = "Y"
                    ENDIF
                    IF (value.behaver.over === true)
                        behavior.Overtime = "Y"
                    ENDIF
                    IF (display_info.submit_count > 0)
                        behavior.Submit = "Y"
                    ENDIF
                ENDIF

End Procedure
```

The next step was to use the UX/UI design principles to design and develop a system that considers the users' moods and feelings when using the system. Colour was used according to psychological principles, and medical and technological concepts were implemented to develop the system: (i) facial emotion analysis represents emotion symptoms, (ii) user-behaviour tracking represents psychomotor and cognition symptoms and (iii) PHQ-9 assessment represents neurovegetative and psychomotor symptoms. The system's processes in the form of system algorithms are summarised in Algorithm 3.

Algorithm 3 System algorithm.

```
Procedure severityCalculate (questionList)

    totalScore = 0
    numQuestions = 9

    FOR i = 1 TO numQuestions DO
        totalScore = totalScore + questionList[i].score
    ENDFOR

    IF totalScore <= 4 THEN
            severity = "minimal"
    ELSE IF totalScore <= 9 THEN
            severity = "mild"
    ELSE IF totalScore <= 14 THEN
            severity = "moderate"
    ELSE IF totalScore <= 19 THEN
            severity = "moderately severe"
    ELSE
            severity = "severe"
    ENDIF

    RETURN severity

End Procedure
```

2.2.3 Experimentation

This research received approval from the Human Research Ethics Committee of Somdetphraphutthaloetla Hospital (COA No. 57) and Silpakorn University (COA No. 64-0913-129-5712). Subjects consenting to partake in the experiment were selected by a purposive sampling method. The subjects were divided into three groups: (i) subjects suffering from depression disorder without treatment (undertreated subjects), (ii) subjects undergoing depression treatment (subjects undergoing treatment) and (iii) subjects without depression disorder (normal subjects). All personal data and identity details used in the depression potentiality analysis system are confidential, including the recorded videos and users' interactions during the PHQ-9 assessment [18]. The experimentation procedures were implemented as detailed below.

- The experimental system was installed on a laptop with a high-quality webcam. Subject lighting ensured proper camera exposure and was controlled by psychiatrists. The experiments were conducted at the Psychiatry and Drug Addiction Subdivision of Somdetphraphutthaloetla Hospital. Figure 2 shows the experimental tool, which consisted of a system for analysing depression potentiality, behaviour tracking and facial recognition during PHQ-9 assessment [18].
- After the subjects submitted their answers to the PHQ-9 assessment, the system created the assessment report and saved five parts of the data: emotion, skip, change, click time, reaction time and group test. All experimental data contained a core PHQ-9 for each questionnaire with emotion and behaviour by map and tag: (i) emotion represents emotional symptoms, (ii) user-behaviour tracking: skip, change, click time, reaction time represent psychomotor and cognition symptoms and (iii) PHQ-9 assessment represent neurovegetative and psychomotor symptoms. into the database without specifying the subjects' personal information.

Figure 2 System for analysing depression potentiality, behaviour tracking and facial recognition during PHQ-9 assessment.

2.2.4 Data collection and analysis

The data collected from the experiment (depression potentiality, behaviour tracking and facial recognition during PHQ-9 assessment) were stored in the database. Ninety subjects were classified into three groups according to the experimental research principles [26]: (i) normal subjects, (ii) undertreated subjects and (iii) subjects undergoing treatment. The data collected were analysed to clarify the characteristics and the limitations of each data set. The data used in this research have 14 attributes and 90 records covering behavioural symptoms of depression. These data were used to develop MDD classification models and compare their predictive capabilities (see Table 2).

Table 2 Attribute Details

No	Attribute	Type of Data	Descriptive
1	TYPE_ID	TEXT	Types of the subjects
2	SCORE_PHQ9	Number	PHQ-9 assessment scores
3	RESULT_PHQ9	TEXT	PHQ-9 assessment results
4	CHANGE_BEHV	TEXT	The subjects show reluctant behaviors by changing their answers during the assessment.
5	SKIP_BEHV	TEXT	The subjects skip a question to answer another.
6	OVERTIME_BEHV	TEXT	The subjects spend overtime to answer a question.
7	SUBMIT_BEHV	TEXT	The subjects submit their answers after completing the assessment.
8	FEAR_EMO	Number	Fearful / Anxious
9	SAD_EMO	Number	Sad / Sorrowful
10	HAPPY_EMO	Number	Happy
11	NATURAL_EMO	Number	Natural
12	CONSISTENCY_LABLE	TEXT	The experimental result is consistent or inconsistent. (2 Classes)

2.2.5 Data preparation

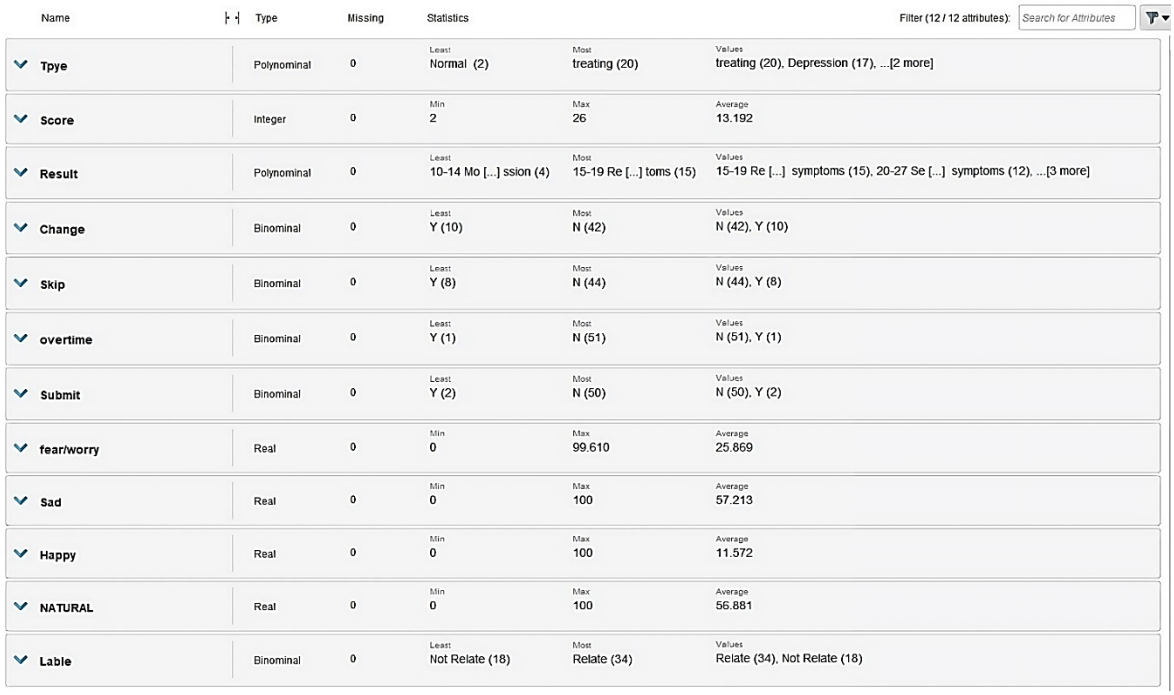
- Data Processing: the data were processed from the database containing the subjects' behaviours, emotions and PHQ-9 assessment scores.
- Data Cleaning: the error data from the database (i.e. incomplete or empty data) were deleted for the entire set. Next, the consistency of the subjects' PHQ-9 scores, behaviours and emotions was analysed. The results were submitted to consistency

analysis by experts (experienced clinical psychologists and psychiatrists) to determine classes for depression classification. The classes were categorised as either consistent or inconsistent. Table 3 lists the criteria and conditions defined for all the data.

- Data Transformation: the data were appropriately transformed and classified to suit all the tools and techniques for accuracy before being tested by the classification model. Figure 3 compares the accuracy of the data. The researcher transformed the data in three parts: (i) TYPE_ID transformed to Depression = 0, Normal = 1 and treating = 0 (ii) CONSISTENCY_LABLE transformed to Y (consistent) = 1 and N (inconsistent) = 0 and finally transformed the data. The functions CHANGE_BEHV, SKIP_BEHV, OVERTIME_BEHV and SUBMIT_BEHV transform to Y = 1 and N (inconsistent) = 0 via RapidMiner.

Table 3 Criteria and conditions defined by the experienced experts for the data.

PHQ-9	**Behavior**		**Emotion**		**Label**	
(Assessment Scores)	**Present**	**Absent**	**Negative**	**Positive**	**Consistent**	**Inconsistent**
0-4	✓		✓			✓
No or slightly	✓		✓	✓		✓
depressive behaviors.		✓	✓			✓
	✓			✓	✓	
		✓	✓	✓	✓	
		✓		✓	✓	
5-9	✓		✓			✓
Low-level depressive	✓		✓	✓		✓
behaviors		✓	✓			✓
	✓			✓	✓	
		✓	✓	✓	✓	
		✓		✓	✓	
10-14		✓	✓	✓	✓	
Medium-level	✓		✓		✓	
depressive behaviors	✓		✓	✓	✓	
		✓		✓		✓
		✓	✓	✓		✓
	✓			✓		✓
15-19	✓		✓		✓	
High-level depressive	✓		✓	✓	✓	
behaviors		✓	✓		✓	
	✓			✓		✓
		✓	✓	✓		✓
		✓		✓		✓
20-27	✓		✓		✓	
Severely depressive	✓		✓	✓	✓	
behaviors		✓	✓		✓	
	✓			✓		✓
		✓	✓	✓		✓
		✓		✓		✓



Name	Type	Missing	Statistics		Filter (12 / 12 attributes): Search for Attributes
Tpye	Polynominal	0	Least: Normal (2)	Most: treating (20)	Values: treating (20), Depression (17), ...[2 more]
Score	Integer	0	Min: 2	Max: 26	Average: 13.192
Result	Polynominal	0	Least: 10-14 Mo [...] ssion (4)	Most: 15-19 Re [...] toms (15)	Values: 15-19 Re [...] symptoms (15), 20-27 Se [...] symptoms (12), ...[3 more]
Change	Binominal	0	Least: Y (10)	Most: N (42)	Values: N (42), Y (10)
Skip	Binominal	0	Least: Y (8)	Most: N (44)	Values: N (44), Y (8)
overtime	Binominal	0	Least: Y (1)	Most: N (51)	Values: N (51), Y (1)
Submit	Binominal	0	Least: Y (2)	Most: N (50)	Values: N (50), Y (2)
fear/worry	Real	0	Min: 0	Max: 99.610	Average: 25.869
Sad	Real	0	Min: 0	Max: 100	Average: 57.213
Happy	Real	0	Min: 0	Max: 100	Average: 11.572
NATURAL	Real	0	Min: 0	Max: 100	Average: 56.881
Lable	Binominal	0	Least: Not Relate (18)	Most: Relate (34)	Values: Relate (34), Not Relate (18)

Figure 3 Data processing.

2.2.6 Modelling

The models developed to classify data by data mining were (i) decision tree, (ii) SVM, (3) naïve Bayes and (iv) neural networks. RapidMiner Studio was used to analyse and determine the testing required to compare the models' classification capabilities.

2.2.7 Testing model accuracy

K-fold cross-validation is a way to validate errors in model predictions. In this step, 10-Fold cross-validation was implemented by dividing the data into ten equal sets. One set was selected as the test set; the other nine sets were the training sets. The criteria tested each classification model to improve the prediction accuracy. The model parameters were optimised by using Optimize Parameters (Grid) Operator is a nested Operator and the most effective results were obtained using cross-validation. Because some hyperparameters work well on training data but not on test data, hyperparameter optimisation was vital because it allowed us to avoid using some hyperparameters. Therefore, we compared the number of folds in the RapidMiner program (five or ten folds). RapidMiner executes the subprocess for all combinations of selected values of the parameters and then delivers the optimal parameter values through the parameter set port. The performance vector for optimised parameters was delivered through the performance port, and the associated model consisted of the following values:

Accuracy: the accuracy of each model is calculated as follows based on all relevant classes:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + \ True\ Negative + \ False\ Negative}$$

Precision: the precision of each model is calculated as follows based on each class:

$$Precision\ = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: the accuracy of each model is calculated as follows based on each class:

$$Recall\ = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

F1-Score: the precision and recall of each model are calculated as follows based on each class:

$$F - Score = \frac{2 \times Precision\ \times Recall}{Precision + Recall}$$

3. Results

The accuracy of the four classification models (algorithms) (i.e. decision tree, SVM, naïve Bayes and neural network) is compared via the number of folds in RapidMiner (five or ten). To avoid using all the hyperparameters, they were optimised. Some hyperparameters work well on training data. Comparing the performance of five folds with that of ten folds shows that the decision tree model is more accurate than the other three models (see Table 4).

Table 4 Model comparison based on number of folds.

Efficiencies	Number of folds	Algorithms			
		Decision Tree	Naïve Bayes	Neural Network	Support Vector Machines (SVM)
Accuracy (%)	5 folds	76.91	68.18	69.27	70.00
	10 folds	84.62	83.01	83.01	83.01
Precision (%)	5 folds	76.25	64.06	65.63	60.00
	10 folds	75.00	72.53	70.43	71.48
Recall (%)	5 folds	70.59	63.34	64.71	67.25
	10 folds	73.33	70.84	65.00	67.92
F1-Score (%)	5 folds	73.31	63.70	65.16	63.42
	10 folds	80.00	88.47	70.00	79.24

According to the research results, the best method based on accuracy, precision, recall and F1-Score is the 10-fold cross-validation model. Decision tree achieves an accuracy of 84.62%, a precision of 83.01%, a recall of 83.01% and a F1-Score of 83.01%. Naïve Bayes achieves an accuracy of 75.00%, a precision of 72.53%, a recall of 70.43% and a F1-Score of 71.48%. Neural Network achieves an accuracy of 75.00%, a precision of 72.53%, a recall of 70.43% and a F1-Score of 67.92% 4) Finally, SVM achieves an accuracy of 80.00%,a precision of 88.47%, a recall of 70.00% and a F1-Score of 79.24%. These results are summarised in Figure 4 and Table 5.

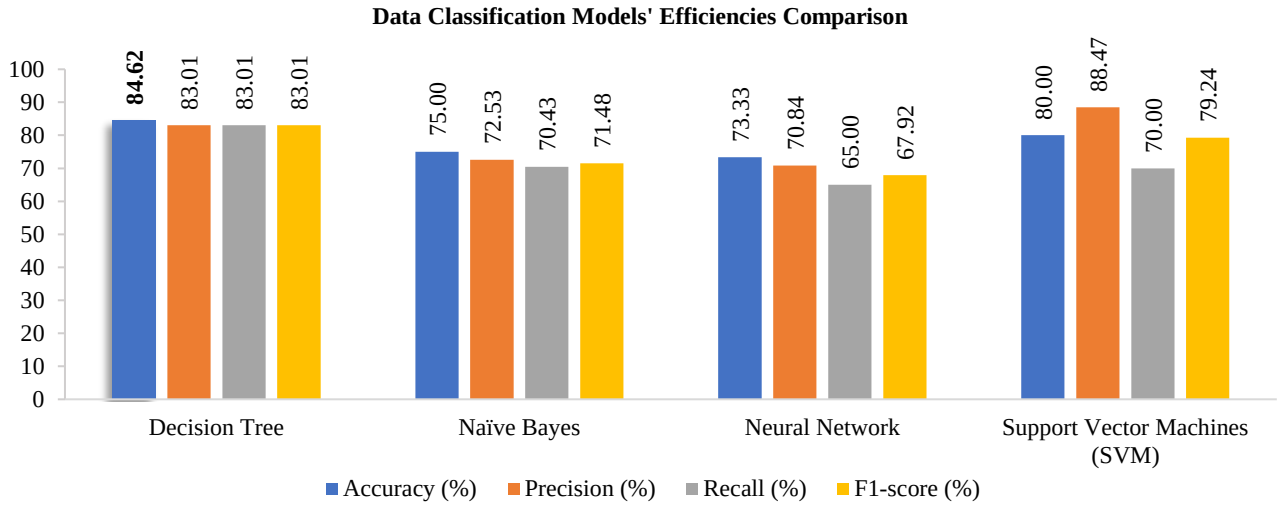


Figure 4 Comparison of results of data classification models.

Table 5 Algorithm results (10-fold cross-validation).

Algorithms	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	84.62	83.01	83.01	83.01
Naïve Bayes	75.00	72.53	70.43	71.48
Neural Network	73.33	70.84	65.00	67.92
Support Vector Machines (SVM)	80.00	88.47	70.00	79.24

The data obtained from the subjects reflect the characteristics of MDD that still require analysis by experienced clinical psychiatrists or psychologists. On the one hand, some data from the subjects' PHQ-9 assessment scores, behaviours and emotions are consistent with MDD symptoms and are beneficial for diagnosis. On the other hand, some data are inconsistent, affecting the subjects' future treatment plans and signalling that they should immediately receive treatment. Thus, highly effective classification models should be implemented to ensure the accuracy of the MDD analysis system. In this case, decision tree is the most accurate and appropriate model due to its accuracy of 84.62%, precision of 83.01%, recall of 83.01% and F1-Score of 83.01%. Decision Tree can increase the depression potentiality, behaviour tracking and facial recognition analysis system during PHQ-9 assessment and may thus be deployed as a tool to assist medical personnel in forming a preliminary diagnosis.

4. Discussion

The purpose of this research is to analyse the classification models to determine whether their predictions of depression potentiality, behaviour tracking and facial recognition during PHQ-9 assessment are mutually consistent. To this end, we compare all the classification models by using the data analysing tool RapidMiner Studio to find the most suitable model for MDD classification. Four data mining techniques are applied in this research: decision tree, naïve Bayes, neural network and SVM. These are evaluated based on the strongest predictive statistics: accuracy, precision, recall and F1-Score.

Testing these models by using 10-fold cross-validation shows that decision tree is the most accurate model (84.62% compared with 75.00% for naïve Bayes, 73.33% for neural network and 80.00% for SVM). This result is consistent with those of [1, 11], which experimented on depression classification by using SVM, KNN and decision tree.

Mahoto et al. [13] created and compared machine-learning classification models (decision tree, multilayer perceptron, naïve Bayes, random forest and SVM). They reported that their decision tree method was the most accurate model. Likewise, Joshi and Kanoongo [14] detected depressive behaviour by deploying artificial intelligence and machine learning to compare the detection accuracy of tree-based models [15]. Furthermore, based on tests and comparisons with other models, Huang et al. [16] reported that decision tree is the most accurate model. They thus encouraged the use of decision tree models to predict MDD potentiality. Aleem et al. [17] presented the use of classification machine-learning algorithms by decision tree and asserted its suitability for the in-depth analysis of MDD and applied decision tree to actual medical cases. Aleem et al. [17] and Uddin et al. [19] reported that decision tree is the most frequently used model for MDD classification.

We thus compared the predictions of four classification models to determine whether they predicted mutually consistent results for MDD potentiality, behaviour tracking and facial recognition during PHQ-9 assessment. The focus on the consistency of the subjects' PHQ-9 scores, behaviours and emotions results in the decision tree model being the most accurate, which means it can significantly advance the accuracy of MDD diagnoses [1, 11, 13-17, 19]. Finally, the findings and related research indicate that, for diagnosing depression, the decision tree model is appropriate because it contains fewer conditional data and less classification complexity. Moreover, the use of the decision tree model indicates the classification conditions and the number of conditions that led to data classification (see Table 6).

Table 6 Comparison of research results with related works.

Discussion and Conclusion	Accuracy (%)			
	Decision Tree	**Naïve Bayes**	**Neural Network**	**Support Vector Machines**
The Comparison Results of this Research's Classification Models' Efficiencies	84.62	75.00	73.33	80.00
Related Works' Results (Average Accuracy)	83.05	87.15	90.75	77.59
Yang et al. [1]	85.27	84.72	-	89.74
Singh and Kumar [11]	83.70	79.17	-	76.67
Mahoto et al. [13]	92.00	94.70	97.30	55.55
Joshi and Kanoongo [14]	67.00	90.00	-	82.00
Huang et al. [16]	87.30	-		84.00
Uddin et al. [19]	-	-	84.20	-

5. Conclusion

This study compares the capability of classification models to predict mutually consistent results for depression potentiality, behaviour tracking and facial recognition during PHQ-9 assessment. The study is implemented according to psychological principles, and the results should advance related academic fields and can be applied to other related medical contexts. However, a limitation of the study is that it uses only four classification models. In addition, the parameters of the experiments were not reviewed by scholars, experts, clinical psychiatrists, or psychologists. In other words, these professionals did not participate in the experimental elements of this research.

In the future, the decision tree classification model should be applied to help analyse MDD potentiality, behaviour tracking and facial recognition and produce an accurate preliminary diagnosis. Other types of models related to the medical context should also be explored and experimented with to develop new opportunities and for the sake of academic and technological advancement in the hopes of obtaining more accurate and efficient diagnoses.

6. References

[1] Yang J, Zhang Z, Fu Z, Li B, Xiong P, Liu X. Cross-subject classification of depression by using multiparadigm EEG feature fusion. Comput Methods Programs Biomed. 2023;233:107360.

[2] World Health Organization. Depression [Internet]. 2021 [cited 2021 Aug 5]. Available from: https://www.who.int/news-room/fact-sheets/detail/depression.

[3] Zheng Y, Chen X, Li D, Liu Y, Tan X, Liang Y, et al. Treatment-naïve first episode depression classification based on high-order brain functional network. J Affect Disord. 2019;256:33-41.

[4] Hossain S, Umer S, Rout RK, Tanveer M. Fine-grained image analysis for facial expression recognition using deep convolutional neural networks with bilinear pooling. Appl Soft Comput. 2023;134:109997.

[5] Kanraweekultana N, Waijanya S, Promrit N, Korsanan A, Poolphol S. Depression analysis using behavior tracking and facial expression recognition during PHQ-9 assessment. ICIC Express Lett. 2023;17(3):303-14.

[6] Faculty of Medicine Ramathibodi Hospital. Ramathibodi Essential Psychiatry. Bangkok: Mahidol University Press; 2015. (In Thai)

[7] Li Y, Chu T, Liu Y, Zhang H, Dong F, Gai Q, et al. Classification of major depression disorder via using minimum spanning tree of individual high-order morphological brain network. J Affect Disord. 2023;323:10-20.

[8] Liu YS, Song Y, Lee NA, Bennett DM, Button KS, Greenshaw A, et al. Depression screening using a non-verbal self-association task: a machine-learning based pilot study. J Affect Disord. 2023;310:87-95.

[9] Souza Filho EM, Veiga Rey HC, Frajtag RM, Arrowsmith Cook DM, Dalbonio de Carvalho LN, Pinho Ribeiro AL, et al. Can machine learning be useful as a screening tool for depression in primary care?. J Psychiatr Res. 2021;132:1-6.

[10] Dong W, He Y, Wang J, Shi C, Niu Q, Yu H, et al. Differential diagnosis of schizophrenia using decision tree analysis based on cognitive testing. Eur J Psychiatry. 2022;35(4):246-51.

[11] Singh A, Kumar D. Detection of stress, anxiety and depression (SAD) in video surveillance using ResNet-101. Microprocess Microsyst. 2022;95:104681.

[12] Price GD, Heinz MV, Zhao D, Nemesure M, Ruan F, Jacobson NC. An unsupervised machine learning approach using passive movement data to understand depression and schizophrenia. J Affect Disord. 2022;316:132-9.

[13] Mahoto NA, Shaikh A, Sulaiman A, Reshan MSA, Rajab A, Rajab K. A machine learning based data modeling for medical diagnosis. Biomed Signal Process Control. 2023;81:104481.

[14] Joshi ML, Kanoongo N. Depression detection using emotional artificial intelligence and machine learning: a closer review. Mater Today: Proc. 2022;58(1):217-26.

[15] Kumar P, Garg S, Garg A. Assessment of anxiety, depression and stress using machine learning models. Procedia Comput Sci. 2020;171:1989-98.

[16] Huang Y, Zhu C, Feng Y, Ji Y, Song J, Wang K, et al. Comparison of three machine learning models to predict suicidal ideation and depression among Chinese adolescents: a cross-sectional study. J Affect Disord. 2022;319:221-8.

[17] Aleem S, Huda N, Amin R, Khalid S, Alshamrani SS, Alshehri A. Machine learning algorithms for depression: diagnosis, insights, and research directions. Electronics. 2022;11(7):1111.

[18] Faculty of Medicine Ramathibodi Hospital. PHQ-9 Depression Assessment [Internet]. 2022 [cited 2021 Aug 5]. Available from: https://www.rama.mahidol.ac.th/th/depression_risk. (In Thai)

[19] Uddin MZ, Dysthe KK, Følstad A, Brandtzaeg PB. Deep learning for prediction of depressive symptoms in a large textual dataset. Neural Comput Appl. 2022;34:721-44.

[20] Sun Y, Fu Z, Bo Q, Mao Z, Ma X, Wang C, et al. The reliability and validity of PHQ-9 in patients with major depressive disorder in psychiatric hospital. BMC Psychiatry. 2020;20:474.

[21] Dede B, Delk L, White BA. Relationships between facial emotion recognition, internalizing symptoms, and social problems in young children. Pers Individ Differ. 2021;171:110448.
[22] Panichkriangkrai C, Silapasuphakornwong P, Saenphon T. Emotion recognition of students during e-learning through online conference meeting. Sci Eng Health Stud. 2021;15:1-10.
[23] Scoppetta O, Cassiani-Miranda CA, Arocha-Díaz KN, Cabanzo-Arenas DF, Campo-Arias A. Validity of the patient health questionnaire-2 for the detection of depression in primary care in Colombia. J Affect Disord. 2021;278:576-82.
[24] Faculty of Medicine Ramathibodi Hospital. Depression in detail [Internet]. 2022 [cited 2021 Aug 5]. Available from: https://med.mahidol.ac.th/ramamental/generalknowledge/ general/09042014-1017. (In Thai)
[25] Costantini L, Pasquarella C, Odone A, Colucci ME, Costanza A, Serafini G, et al. Screening for depression in primary care with Patient Health Questionnaire-9 (PHQ-9): a systematic review. J Affect Disord. 2021;279:473-83.
[26] Visser LNC, et al. Methodological choices in experimental research on medical communication using vignettes: the impact of gender congruence and vignette modality. Patient Educ Couns. 2022;105(6):1634-41.