Available online at www.HEFJournal.org

ISSN: 2785-2997

Journal of
Human, Earth, and Future

Vol. 5, No. 1, March, 2024

An Analysis of Machine Learning for Detecting Depression, Anxiety, and Stress of Recovered COVID-19 Patients

Tran Anh Tuan [1, 2*], Le Thanh Thao Trang [3], Tran Dai An [4], Nguyen Huu Nghia [5], Dao Thi Thanh Loan [6]

[1] *School of Informatics, Walailak University, Nakhon Si Thammarat, 80160, Thailand.*

[2] *Informatics Innovation Center of Excellence (IICE), Walailak University, Nakhon Si Thammarat, 80160, Thailand.*

[3] *Faculty of Basic Science and Public Health, Dong Thap Medical College, Dong Thap, 81000, Vietnam.*

[4] *Dong Thap Provincial Center for Disease Control, Dong Thap, 81000, Vietnam.*

[5] *Quang Binh Medial College, Quang Binh, 47100, Vietnam.*

[6] *Faculty of Foundational Sciences, Dak Lak College of Pedagogy, Dak Lak, 63000, Vietnam.*

Received 01 November 2023; Revised 16 January 2024; Accepted 06 February 2024; Published 01 March 2024

Abstract

Objectives: This study explores different machine learning models (KNN: k-nearest neighbor, MLP: Multilayer Perceptron, SVM: Support Vector Machine) to identify the optimal model for accurate and rapid mental health detection among the recovered COVID-19 patients. Other techniques are also investigated, such as feature selection (Recursive Feature Elimination (RFE) and Extra Trees (ET) methods) and hyper-parameter tuning, to achieve a system that could effectively and quickly indicate mental health. *Method/Analysis:* To achieve the objectives, the study employs a dataset collected from recovered COVID-19 patients, encompassing information related to depression, anxiety, and stress. Machine learning models are utilized in the analysis. Additionally, feature selection methods and hyper-parameter tuning techniques are explored to enhance the model's predictive capabilities. The performance of each model is assessed based on accuracy metrics. *Findings:* The experimental results show that SVM is the most suitable model for accurately predicting an individual's mental health among recovered COVID-19 patients (accuracy $\geq$ 0.984). Furthermore, the ET method is more effective than the RFE method for feature selection in the anxiety and stress datasets. *Novelty/Improvement:* The study lies in the understanding of predictive modeling for mental health and provides insights into the choice of models and techniques for accurate and early detection.

Keywords: Predictive Model; Machine Learning; Depression; Anxiety; Stress; DAS; Mental Health; COVID-19.

1. Introduction

The COVID-19 pandemic, caused by the coronavirus SARS-CoV-2, has left an indelible mark on the global population, affecting millions of lives physically, emotionally, and mentally. Numerous studies have been dedicated to understanding and mitigating the acute health risks of the virus. In recent times, there has been a growing body of research focusing on the long-term implications of COVID-19 recovery on mental and physical health, which has gained increasing significance, primarily due to the effects of COVID-19 on recovered patients [1-3].

* Corresponding author: tran.tu@wu.ac.th

http://dx.doi.org/10. 28991/HEF-2024-05-01-01

Recovered COVID-19 patients, often referred to as “long-haulers”, have reported a wide array of persistent symptoms extending well beyond the acute phase of the illness [4, 5]. These individuals experience a range of symptoms, both physical and mental, which can significantly impact their overall well-being. Among these symptoms, mental health concerns, such as depression, anxiety, and stress (DAS), have emerged as significant issues, affecting the quality of life and overall well-being of recovered patients [6].

Depression manifests as persistent feelings of sadness, disinterest, and energy depletion, with symptoms including indecisiveness, memory loss, and changes in appetite [7–9]. Anxiety, an emotional state marked by fear and uneasiness, is represented by symptoms such as fatigue, irritability, and difficulty concentrating [9]. Stress, characterized by emotional or physical tension, has symptoms such as low energy, agitation, and chronic headaches [10, 11]. Recognizing these conditions and their associated symptoms is crucial for accurate clinical diagnosis and the development of effective mental health support strategies. One of the popular diagnosis criteria for depression, anxiety, and stress is the 20-item Depression, Anxiety, and Stress Scale (DASS-21) [12, 13].

The common symptoms of depression, anxiety, and stress, such as chest pain, insomnia, and fatigue, pose challenges for machine classification. Consequently, numerous studies have delved into the contribution of machine learning (ML) to mental health datasets [14–35]. These datasets span diverse domains, encompassing the general population (including pregnant women and elders) [20, 21, 31, 32, 34], patients [23], college students [28, 30], and workers in technology and health fields [24–26, 29]. There is a lack of studies conducting mental health on recovered COVID-19 patients with ML. Moreover, recent studies have reported a high prevalence of depression, anxiety, and stress in this population [36–39]. Hence, our study aims to explore and identify an ML model tailored for mental health datasets from recovered COVID-19 patients.

This study has employed K-nearest neighbor (KNN), Multilayer Perceptron (MLP), and Support Vector Machine (SVM) to predict levels of depression, anxiety, and stress in recovered COVID-19 patients. Since the primary objective is to identify the best ML model among these models, achieving high accuracy. To achieve this, we explore the Recursive Feature Elimination (RFE) and Extra Trees (ET) methods for feature selection. Subsequently, we assess and determine the number of features that yield high accuracy using the Decision Tree (DT) algorithm and tune hyper-parameters for these ML models, seeking to identify the optimal model. It is crucial to note that k-fold cross-validation, with $k = 10$, is consistently applied in all assessments. This study primarily aims to compare these methods in finding hyper-parameter values for depression, anxiety, and stress level prediction in recovered COVID-19 patients and the optimal ML model identification, leading to a significant innovation in this study.

2. Research Methodology

2.1. Data Collection

This study has utilized survey data from [36], performing experiments under pertinent guidelines and regulations (refer to the “Materials and Methods” Section in [36]). The survey was undertaken after obtaining approval from the Human Research Ethics Committee, Walailak University (WU-EC-PU-0-214-65). The dataset comprises information from 549 participants in Dong Thap province, Vietnam, all of whom were previously infected with COVID-19 and had recovered, having been discharged from the hospital for more than six months.

2.2. The Proposed Method

Figure 1 illustrates the ML-based framework for depression, anxiety, and stress detection in recovered COVID-19 patients. The framework has four significant phases, including (1) *data pre-processing*, (2) *feature selection*, (3) *hyper-parameter tuning*, and (4) *optimal prediction model selection*. The input and output of the framework are the mental health dataset (including depression dataset (Data-D), anxiety dataset (Data-A), and stress dataset (Data-S)) and five classes, respectively.

2.2.1. Data Pre-processing

The dataset under examination in this study pertains to the responses to 21 questions (DASS-21) and sociodemographic details of recovered COVID-19 patients, containing 549 rows and 27 columns [36]. The dataset encompasses different types of variables, including categorical and ordinal. In this step, these variables are transformed into numerical values, utilizing encoding and normalizing techniques provided by the Scikit-learn library in Google Colab.

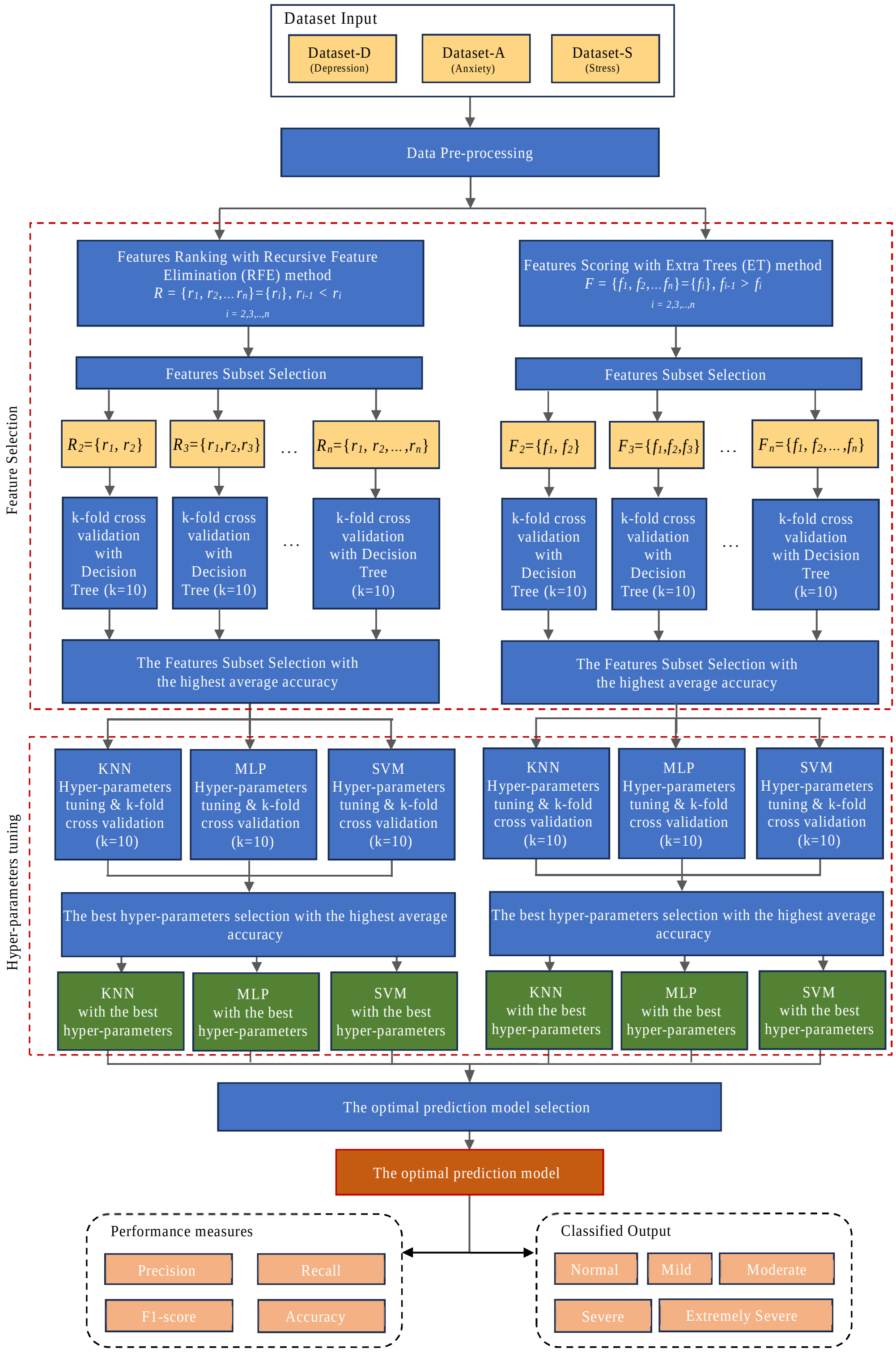


Figure 1. Workflow of proposed method for Depression, Anxiety, and Stress detection based on machine learning

2.2.2. Feature Selection

Feature selection contributes considerably to the model's performance improvement by removing unnecessary features [40]. After data pre-processing, the framework proceeds to the feature selection phase, employing RFE [41] and ET [42] methods. In this stage, the pre-processed data are inputted into each method to identify the optimal subset of features, aiming to enhance the accuracy of predictions.

For the RFE method, features are ranked in descending order of importance ($R = \{r_1, r_2, \ldots, r_n\} = \{r_i\}$, where r_i is the feature i^{th}, n is the number of features, $n = 27$, and $r_{i-1} < r_i$ for $i = 2, 3, \ldots, n$). This process results in the selection of 26 subsets of features being selected. The rule for feature selection in each subset is as follows: "Each subset contains at least two features, and the features with higher rankings are selected first". For example, the first subset includes the first two features ($R_2 = \{r_1, r_2\}$), and the second subset contains the first three features ($R_3 = \{r_1, r_2, r_3\}$). Subsequently, each subset undergoes evaluation using the DT algorithm with k-fold cross-validation ($k = 10$). The subset achieving the highest mean accuracy is then chosen for the subsequent phase. For the ET method, the procedure that selects subsets of features is the same as the RFE method. In the ET method, features are scored in descending order of importance ($F = \{f_1, f_2, \ldots, f_n\} = \{f_i\}$, where f_i is the feature i^{th}, $f_{i-1} < f_i$). Hence, the selection procedure is based on scores of features.

2.2.3. Hyper-parameters Tuning

Three machine learning models (KNN, MLP, SVM) are deployed in this phase to assess and optimize hyper-parameters. The evaluation process utilizes k-fold cross-validation ($k = 10$). The hyper-parameters for the three machine learning models are presented in Table 1, with values assigned to each option derived from existing studies. Each machine learning model encompasses multiple options, each with several parameters automatically selected and generated in all possible combinations. For KNN, the '*metric*' parameter is employed to compute distance, '*n_neighbors*' determines the number of neighbors, and the '*algorithm*' parameter specifies the algorithm for computing the nearest neighbors. In the case of MLP, '*hidden_layer_sizes*' is the number of neurons in each hidden layer, '*activation*' represents the activation function used in the hidden layers, and '*solver*' determines the optimization algorithm for weight optimization during training. In SVM, 'C' serves as a regularization parameter influencing the trade-off between smooth decision boundaries and accurate classification of training points, the '*kernel*' parameter defines the type of kernel function, '*gamma*' determines how far the influence of a single training example reaches, and '*degree*' is relevant for the polynomial kernel function. For example, an option for KNN is *n_neighbors: [5, 10, 20, 50], metric: ['minkowski']*. Combinations of parameters for KNN are *{n_neighbors: 5, metric: 'minkowski'}; {n_neighbors: 10, metric: 'minkowski'}; {n_neighbors: 20, metric: 'minkowski'}; {n_neighbors: 50, metric: 'minkowski'}*. In each machine learning model, the hyper-parameters returned with the highest mean accuracy are selected.

Table 1. Hyper-parameters of the three machine learning models (KNN, MLP, SVM)

Machine learning model	Hyper-parameters
KNN	• Option 1: metric: ['minkowski'] • Option 2: n_neighbors: [5, 10, 20, 50], metric: ['minkowski'] • Option 3: n_neighbors: [5, 10, 20, 50], algorithm: ['auto', 'ball_tree', 'kd_tree', 'brute']
MLP	• Option 1: hidden_layer_sizes: [10, 20, 50, 100], activation: ['identity', 'logistic', 'tanh', 'relu'] • Option 2: hidden_layer_sizes: [10, 20, 50, 100], activation: ['identity', 'logistic', 'tanh', 'relu'], solver: ['lbfgs', 'sgd', 'adam']
SVM	• Option 1: C: [1, 10, 20], kernel: ['linear'], • Option 2: C: [1, 10, 20], gamma: [0.1, 0.01, 0.02], kernel: ['rbf'] • Option 3: degree: [1, 10, 20], gamma: [0.1, 0.01, 0.02], kernel: ['poly']

2.2.4. Optimal Prediction Model Selection

Six machine learning models from the previous stage, including three with the best hyper-parameters derived from the RFE method-based features and three with the best from the ET method-based features, are evaluated in this phase. The assessment is conducted as follows: "Models utilizing the same machine learning algorithm are compared, and the model with the higher mean accuracy is selected. Subsequently, the optimal prediction model is chosen based on the highest accuracy from the selected models". This process aims to identify the most effective prediction model.

2.2.5. Performance Measures

To evaluate and select the models with the best performance, we use the popular measurement metrics, including precision (P), recall (R), F1-scores, and accuracy in prediction issues [43-49].

$$\mathrm{P} = TP/(TP + FP) \tag{1}$$

$$\mathrm{R} = TP/(TP + FN) \tag{2}$$

$$\mathrm{F1-score} = (2 \times \mathrm{P} \times \mathrm{R})/(\mathrm{P} + \mathrm{R}) \tag{3}$$

$$\mathrm{Accuracy} = TP + TN/(TP + FP + TN + FN) \tag{4}$$

where *TP* (True positive): Observation is actually positive and is predicted positive. *FP* (False positive): Observation is actually negative and is predicted positive. *TN* (True negative): Observation is actually negative and is predicted negative. *FN* (False negative): Observation is actually positive and is predicted negative.

2.2.6. Classified Output of Depression, Anxiety, and Stress

The depression, anxiety, and stress levels serve as the classification output, with five distinct levels: normal, mild, moderate, severe, and extremely severe. These classifications are determined based on specific score ranges. In the case of depression, the five levels are categorized as normal (0-9), mild (10-13), moderate (14-20), severe (21-27), and extremely severe (≥28). Similarly, for anxiety, the levels are categorized as normal (0-7), mild (8-9), moderate (10-14), severe (15-19), and extremely severe (≥20). In the context of stress, the five levels are categorized as normal (0-14), mild (15-18), moderate (19-25), severe (26-33), and extremely severe (≥34) [50].

3. Results

3.1. Dataset Description

3.1.1. Characteristics of Dataset

The study utilizes a dataset related to the mental health of recovered COVID-19 patients [36], comprising 549 records of patients with 41 attributes. These attributes are divided into three datasets: Data-D (depression dataset), Data-A (anxiety dataset), and Data-S (stress dataset). The 20 attributes related to sociodemographic information, underlying diseases, and COVID-19 details include *gender*, *age*, *areas*, *education*, *marital status*, *monthly income*, *employment status*, *family infected*, *hypertension*, *diabetes*, *heart disease*, *cancer*, *respiratory disease*, *kidney disease*, *other diseases*, *non-communicable diseases* (No_NCDs), *COVID classification*, *COVID treatment*, *days in hospital*, and *body mass index* (BMI). The seven attributes related to depression are *lack of positive feelings* (Q3), *difficulty initiating tasks* (Q5), *lack of anticipation* (Q10), *feeling down-hearted* (Q13), *lack of enthusiasm* (Q16), *low self-worth* (Q17), and *sense of life being* (Q21). The seven attributes related to anxiety are: *mouth dryness awareness* (Q2), *breathing difficulty* (Q4), *trembling* (Q7), *worry about social panic* (Q9), *proximity to panic* (Q15), *heart awareness energy* (Q19), and *unexplained fear* (Q20). The seven attributes related to stress are *difficulty winding down* (Q1), *tendency to over-react* (Q6), *feeling of using nervous energy* (Q8), *agitation meaningless* (Q11), *difficulty relaxing* (Q12), *intolerance to interruptions* (Q14), and *sensitivity or touchiness* (Q18). Each of the three datasets has 27 attributes, which include the 20 attributes related to sociodemographic information, underlying diseases, COVID-19 details, and seven attributes representing depression/anxiety/stress.

3.1.2. Depression, Anxiety, and Stress among Recovered COVID-19 Participants

Out of 549 participants, 136 (24.77%) had depression, with severity ranging from mild (n=60) to extremely severe (n=11). Regarding anxiety, 228 (41.53%) participants were diagnosed with varying levels of severity, including mild (n=81), moderate (n=88), severe (n=33), and extremely severe (n=26). Stress was identified in 139 participants (25.32%), mainly at a mild level (n=67), followed by moderate (n=45), severe (n=23), and extremely severe (n=4). Among those who had recovered from COVID-19, 69 (12.57%) were diagnosed with all three mental health symptoms, including depression, anxiety, and stress. Table 2 illustrates the depression, anxiety, and stress levels among recovered COVID-19 participants.

Table 2. Levels of depression, anxiety, and stress among recovered COVID-19 participants

Levels	Number of participant (n = 549) (%)		
	Depression	Anxiety	Stress
Normal	413 (75.23)	321 (58.47)	410 (74.68)
Mild	60 (10.93)	81 (14.75)	67 (12.20)
Moderate	49 (8.93)	88 (16.03)	45 (8.20)
Severe	16 (2.91)	33 (6.01)	23 (4.19)
Extremely Severe	11 (2.00)	26 (4.74)	4 (0.73)

3.2. Depression Prediction for Recovered COVID-19 Patients

Figure 2 illustrates the importance of features in the depression dataset based on the RFE and ET methods. The RFE method identifies the top ten features, including *low self-worth* (Q17), *difficulty initiating tasks* (Q5), *lack of anticipation* (Q10), *sense of life being* (Q21), *feeling down-hearted* (Q13), *lack of positive feelings* (Q3), *lack of enthusiasm* (Q16), *Diabetes*, *Hypertension*, and *Cancer*, as the most important, while *Gender* and *Age* are ranked as the least important features. Meanwhile, the ET method returns the top ten features: *difficulty initiating tasks* (Q5), *low self-worth* (Q17), *lack of positive feelings* (Q3), *lack of enthusiasm* (Q16), *sense of life being* (Q21), *lack of anticipation* (Q10), *feeling down-hearted* (Q13), *non-communicable diseases* (No_NCDs), *diabetes*, and *hypertension*, as the most important, while *kidney disease* and *cancer* are identified as the least important features.

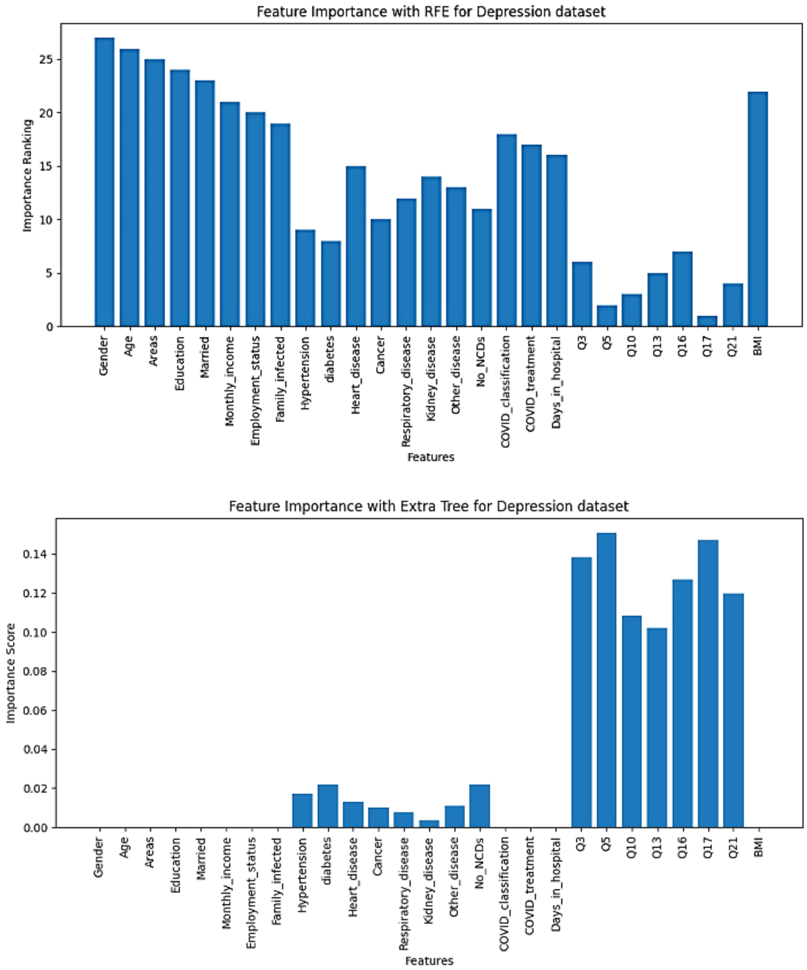
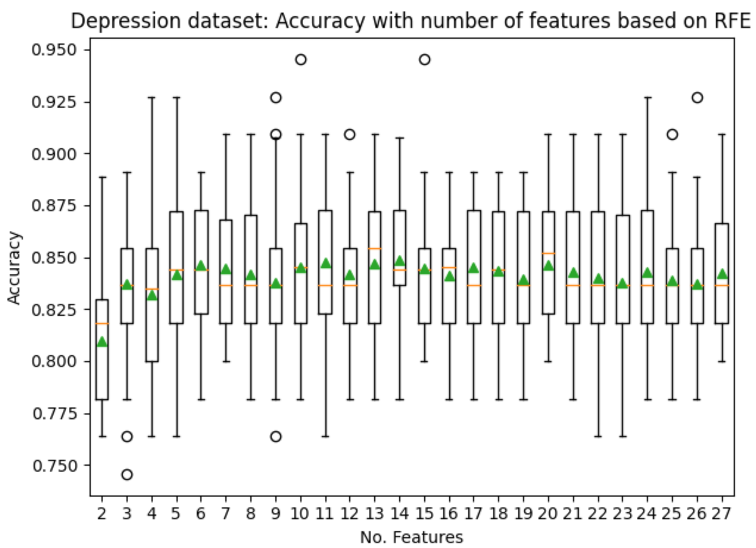
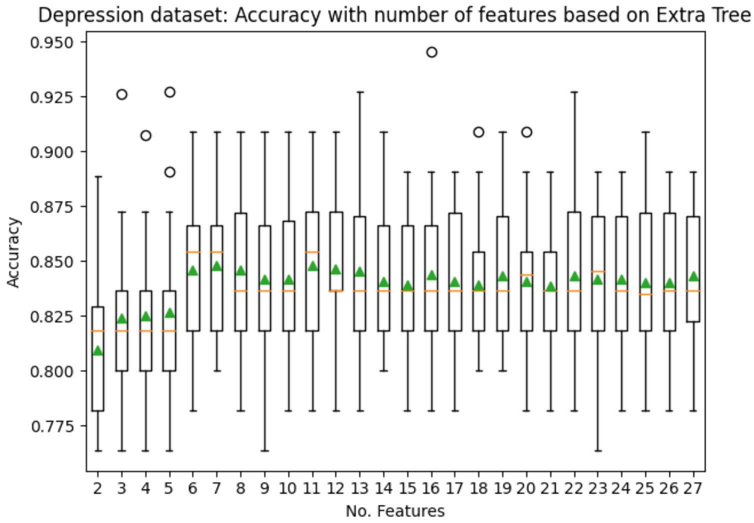


Figure 2. Feature importance based on RFE and ET methods for the depression dataset

Among the features ranked by the RFE method, the first 14, including *low self-worth* (Q17), *difficulty initiating tasks* (Q5), *lack of anticipation* (Q10), *sense of life being* (Q21), *feeling down-hearted* (Q13), *lack of positive feelings* (Q3), *lack of enthusiasm* (Q16), *diabetes*, *hypertension*, *cancer*, *non-communicable diseases* (No_NCDs), *respiratory disease*, *other disease*, and *kidney disease*, returned the highest mean accuracy (0.848). Meanwhile, the best performance (0.851) of the features scored by the ET method was observed in the first 11 features, comprising *difficulty initiating tasks* (Q5), *low self-worth* (Q17), *lack of positive feelings* (Q3), *lack of enthusiasm* (Q16), *sense of life being* (Q21), *lack of anticipation* (Q10), *feeling down-hearted* (Q13), *non-communicable diseases* (No_NCDs), *diabetes*, *hypertension*, and *heart disease* (see Figure 3).



(a) Accuracy with number of features based on RFE



(b) Accuracy with number of features based on ET

Figure 3. Accuracy with number of features for the depression dataset

In the depression dataset, 14 features based on the RFE method, and 11 features based on the ET method were selected to tune the hyper-parameters of three machine learning models (KNN, MLP, and SVM). The features derived from the ET method exhibited the best hyper-parameters, resulting in the highest mean accuracy compared to the RFE method-based features. The respective mean accuracy for each model was 0.880, 0.980, and 0.984. The best hyper-parameters for the three machine learning models were: KNN with the *algorithm: 'brute', n_neighbors: 5*; MLP with the *activation: 'identity', hidden_layer_sizes: 100, solver: 'lbfgs'*; and SVM with the *C: 1, kernel: 'linear'*. The results of hyper-parameter tuning and mean accuracy for these machine learning models are summarized in Table 3. In terms of accuracy, SVM with ET method-based feature selection and the best hyper-parameters emerged as the optimal model for depression prediction in recovered COVID-19 patients (accuracy = 0.984). Meanwhile, MLP, whose feature was selected by the ET method, performed well (accuracy = 0.980 and F1-score = 0.915) in predicting each level of depression in the recovered COVID-19 patients (see Figures 4 and 5).

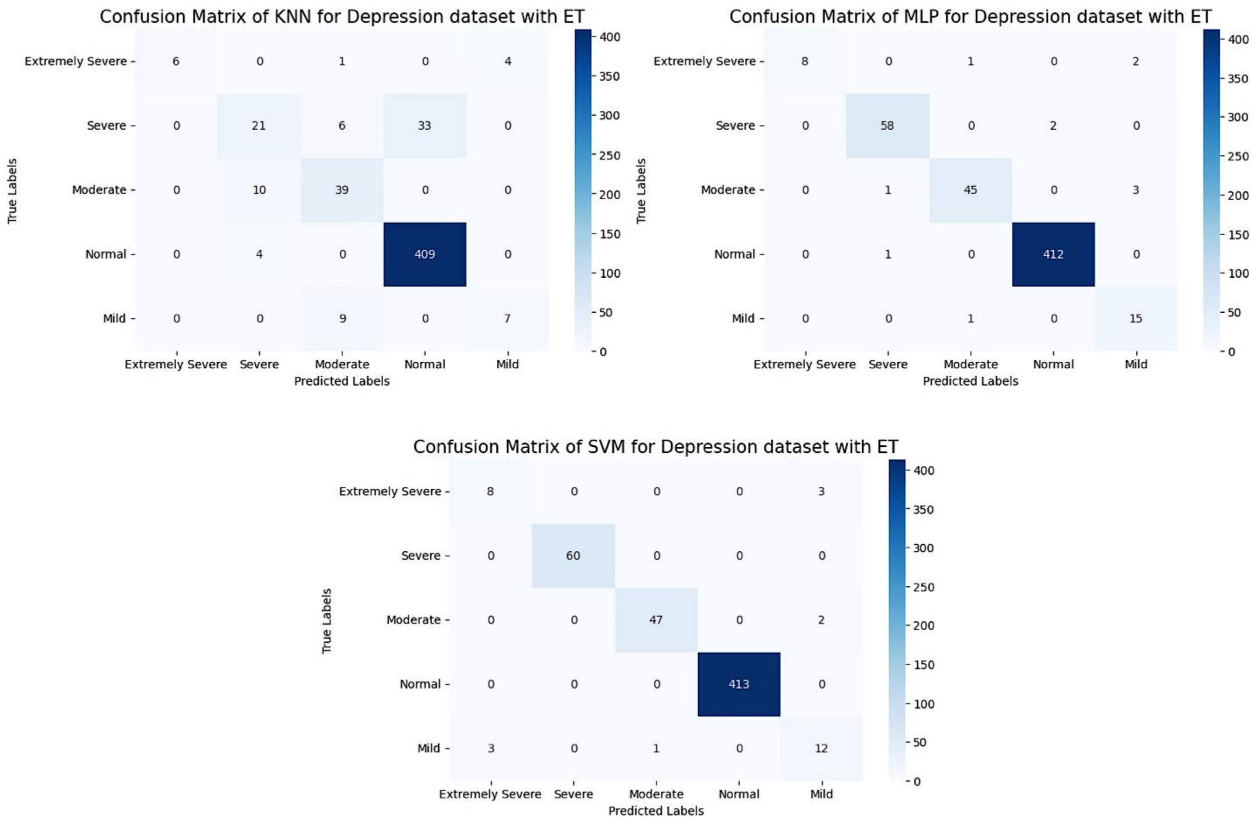


Figure 4. Confusion matrices of three machine learning models for the depression dataset

Table 3. Hyper-parameters tuning of three machine learning models for the depression, anxiety, and stress datasets

Model, hyper-parameters, accuracy		Data-D		Data-A		Data-S	
		RFE-based	**ET-based**	**RFE-based**	**ET-based**	**RFE-based**	**ET-based**
KNN	Hyper-parameters	• metric: ['minkowski'] • n_neighbors: [5, 10, 20,50], metric: ['minkowski'] • n_neighbors: [5, 10, 20,50], algorithm: ['auto', 'ball_tree', 'kd_tree', 'brute']					
	The best hyper-parameters	metric: 'minkowski', n_neighbors: 5	algorithm: 'brute', n_neighbors: 5	algorithm: 'brute', n_neighbors: 10	algorithm: 'ball_tree', n_neighbors: 50	algorithm: 'brute', 'n_neighbors': 5	algorithm: 'brute', n_neighbors: 10
	Accuracy	0.878	0.880	0.778	0.672	0.893	0.794
MLP	Hyper-Parameters	• hidden_layer_sizes: [10, 20,50,100], activation: ['identity', 'logistic', 'tanh', 'relu'] • hidden_layer_sizes: [10, 20,50,100], activation: ['identity', 'logistic', 'tanh', 'relu'], solver: ['lbfgs', 'sgd', 'adam']					
	The best hyper-parameters	hidden_layer_sizes: 10, activation: 'relu', solver: 'lbfgs'	hidden_layer_sizes: 100, activation: 'identity', solver: 'lbfgs'	hidden_layer_sizes: 20, activation: 'identity', solver: 'lbfgs'	hidden_layer_sizes: 10, activation: 'relu'	hidden_layer_sizes: 50, activation: 'identity', solver: 'lbfgs'	hidden_layer_sizes: 50, activation: 'tanh'
	Accuracy	0.971	0.980	1.00	0.681	0.989	0.801
SVM	Hyper-parameters	• C: [1, 10, 20], kernel: ['linear'] • C: [1, 10, 20], gamma: [0.1, 0.01, 0.02], kernel: ['rbf'] • degree: [1, 10, 20], gamma: [0.1, 0.01, 0.02], kernel: ['poly']					
	The best hyper-parameters	C: 1, kernel: 'linear'	C: 1, kernel: 'linear'	C: 1, kernel: 'linear'	C: 10, kernel: 'linear'	C: 1, kernel: 'linear'	C: 20, kernel: 'linear'
	Accuracy	0.983	0.984	1.00	0.685	0.991	0.803

Note: Accuracy is based on the best hyper-parameters

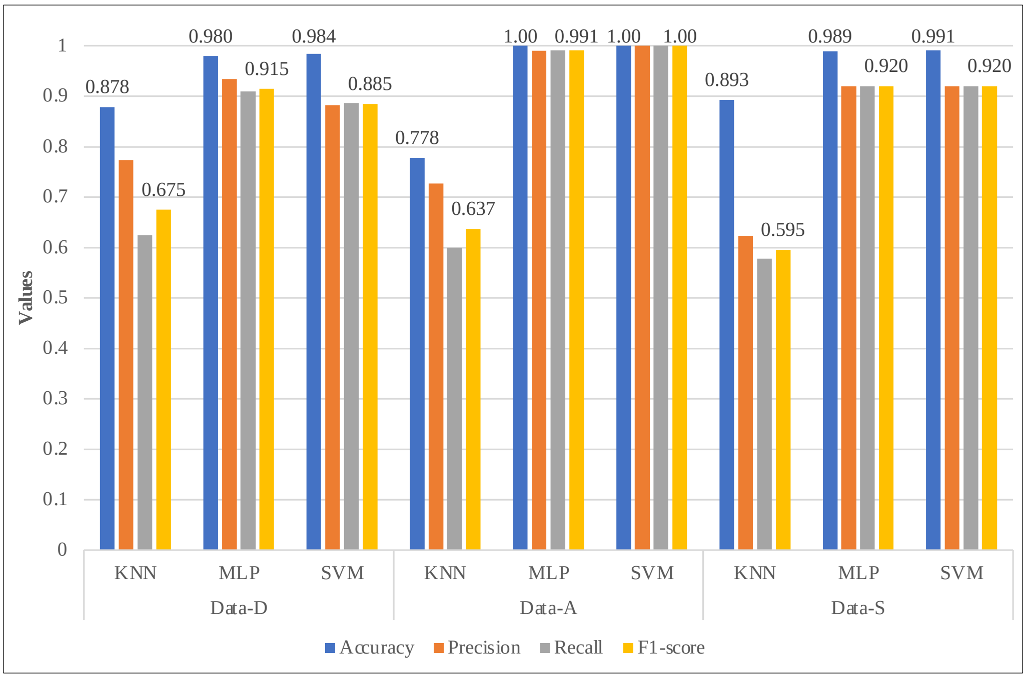


Figure 5. Comparison of machine learning models with the best hyper-parameters in three datasets

3.3. Anxiety Prediction for Recovered COVID-19 Patients

Figure 6 illustrates the importance of features in the anxiety dataset based on the RFE and ET methods. The RFE method identifies the top ten features, comprising *proximity to panic* (Q15), *heart awareness energy* (Q19), *trembling* (Q7), *unexplained fear* (Q20), *mouth dryness awareness* (Q2), *breathing difficulty* (Q4), *worry about social panic* (Q9), *respiratory disease*, *diabetes*, and *cancer* as the most important ranking, while *gender* and *BMI* are ranked as the least important features. The ET method returns the top ten features with the highest scores, including *breathing difficulty* (Q4), *trembling* (Q7), *mouth dryness awareness (Q2)*, *heart awareness energy* (Q19), *worry about social panic* (Q9), *unexplained fear* (Q20), *proximity to panic* (Q15), *non-communicable diseases* (No_NCDs), *hypertension*, and *heart disease*, as the most important, while *kidney disease* and *cancer* are identified as the least important features.

Among the features ranked by the RFE method, the first eight ranked features, comprising *proximity to panic* (Q15), *heart awareness energy* (Q19), *trembling* (Q7), *unexplained fear* (Q20), *mouth dryness awareness* (Q2), *breathing difficulty* (Q4), *worry about social panic* (Q9), and *respiratory disease*, returned the highest mean accuracy (0.805). Meanwhile, the best performance (0.751) of the features scored by the ET method was observed in the first two: *breathing difficulty* (Q4) and *trembling* (Q7) (see Figure 7).

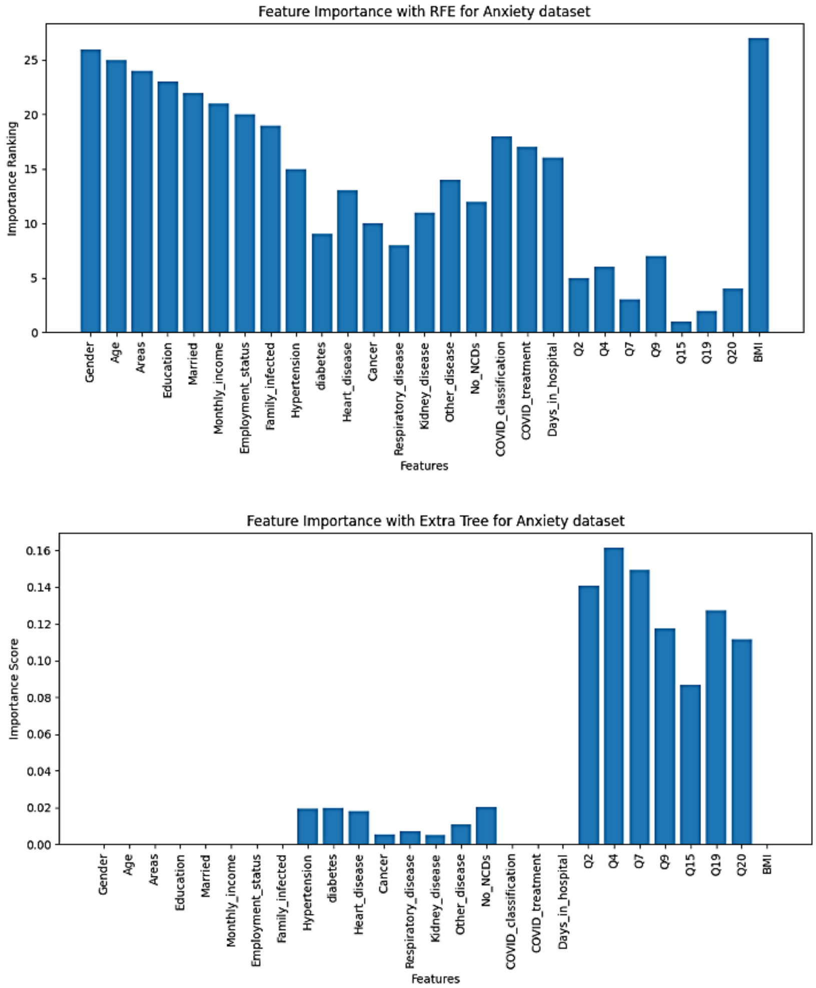
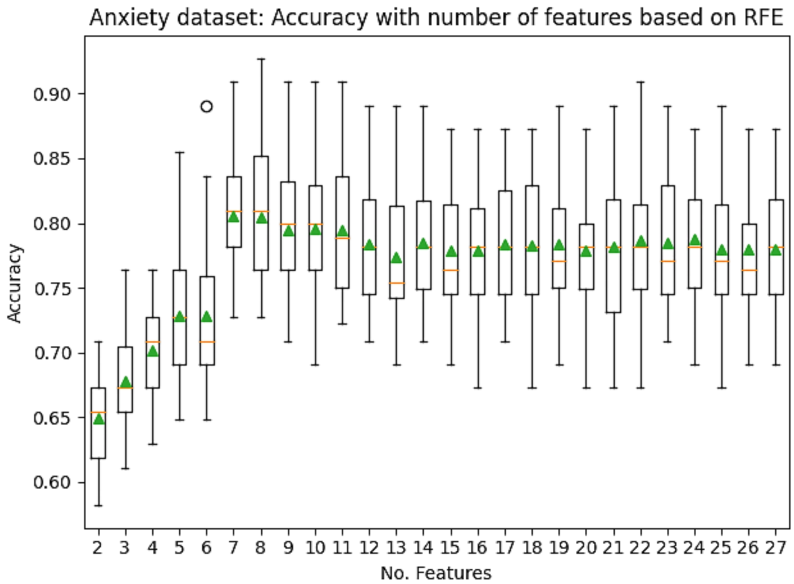
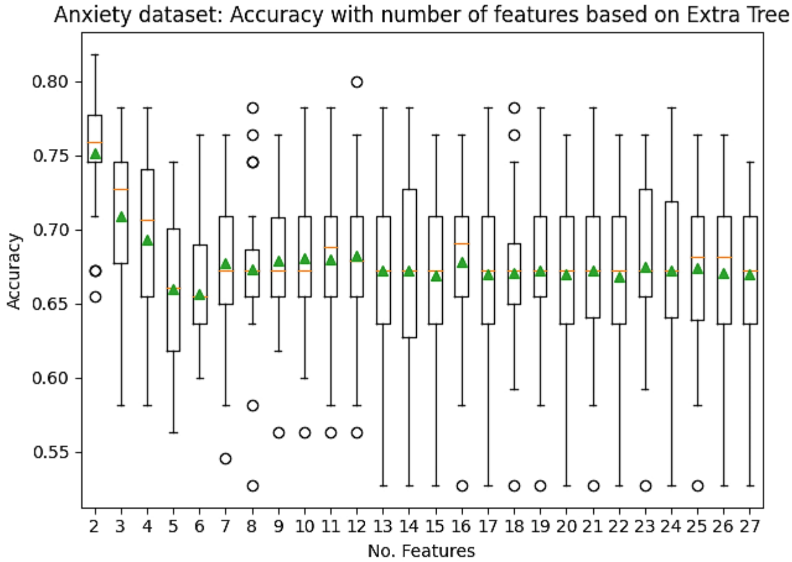


Figure 6. Feature importance based on RFE and ET methods for the anxiety dataset



(a) Accuracy with number of features based on RFE



(b) Accuracy with number of features based on ET

Figure 7. Accuracy with number of features for the anxiety dataset

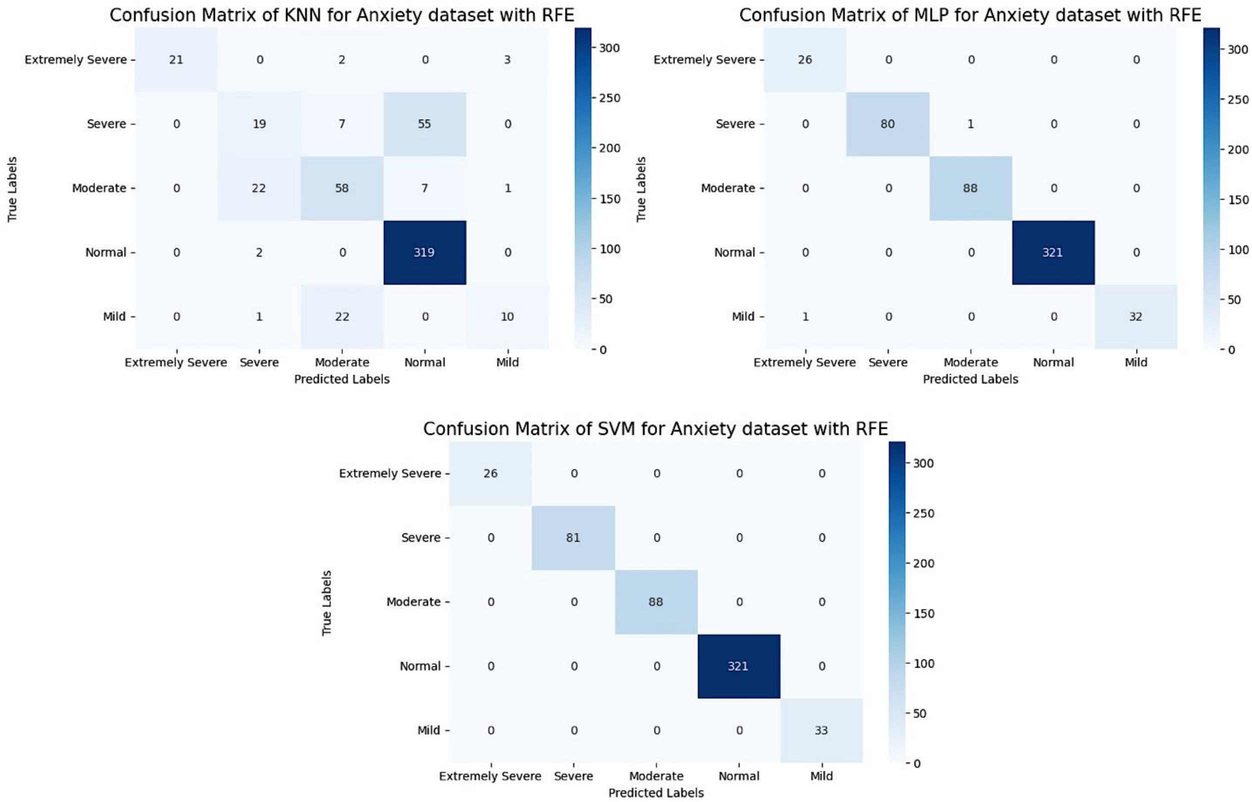


Figure 8. Confusion matrices of three machine learning models for the anxiety dataset

In the anxiety dataset, eight features based on the RFE method, and two features based on the ET method were selected to tune the hyper-parameters of KNN, MLP, and SVM. The features derived from the RFE method showed the best hyper-parameters, resulting in the highest mean accuracy compared to ET method-based features in the three machine learning methods (KNN with 0.778, MLP with 1.00, and SVM with 1.00). The best hyper-parameters for the three machine learning models were KNN with the *algorithm 'brute', n_neighbors: 10*, MLP with the *activation 'identity', hidden_layer_sizes: 20, solver: 'lbfgs'*, and SVM with the *C: 1, kernel: 'linear'*. Table 3 presents the details of tuned hyper-parameters with accuracy for these models, which are depicted in Table 3. Both models (SVM and MLP with RFE method-based feature selection and the best hyper-parameters) revealed the best results in terms of accuracy (accuracy = 1.00 for both) and F1-score (F1-score >0.99 for both) in predicting anxiety levels from the recovered COVID-19 patients (see Figures 5 and 8).

3.4. Stress Prediction for Recovered COVID-19 Patients

Figure 9 illustrates the importance of features in the stress dataset based on the RFE and ET methods. The RFE method identifies the top ten features, including *agitation meaningless* (Q11), *intolerance to interruptions* (Q14), *feeling of using nervous energy* (Q8), *difficulty relaxing* (Q12), *difficulty winding down* (Q1), *tendency to over-react* (Q6), *sensitivity or touchiness* (Q18), *other disease*, *respiratory disease,* and *cancer*, as the most important, while *Gender* and *Age* are ranked as the least important features. The ET method returns the top ten features, comprising the *feeling of using nervous energy* (Q8), *intolerance to interruptions* (Q14), *tendency to over-react* (Q6), *difficulty relaxing* (Q12), *sensitivity or touchiness* (Q18), *difficulty winding down* (Q1), *agitation meaningless* (Q11), *non-communicable diseases* (No_NCDs), *hypertension*, and *diabetes*, as the most important, while *cancer* and *heart disease* are identified as the least important features.

Among the features ranked by the RFE method, the first nine ranked features, including *agitation meaningless* (Q11)*, intolerance to interruptions* (Q14)*, feeling of using nervous energy* (Q8)*, difficulty relaxing* (Q12)*, difficulty winding down* (Q1)*, tendency to over-react* (Q6)*, sensitivity or touchiness* (Q18)*, other disease,* and *respiratory disease,* returned the highest mean accuracy (0.874). Meanwhile, the best performance (0.746) of the features scored by the ET method was observed in the first three features, comprising the *feeling of using nervous energy* (Q8)*, intolerance to interruptions* (Q14)*,* and *tendency to over-react* (Q6) (see Figure 10).

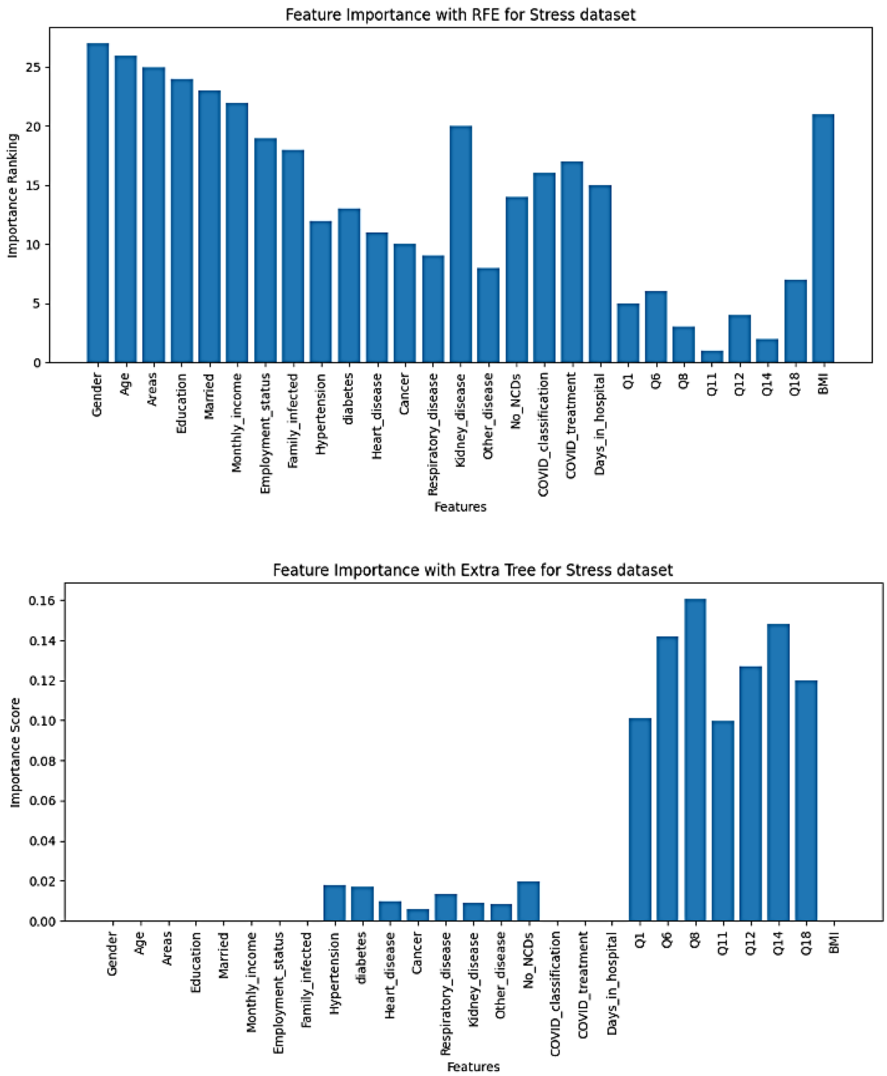
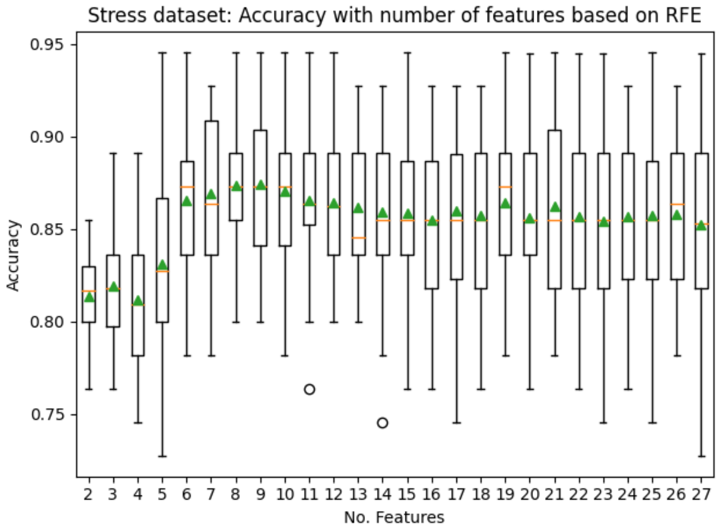
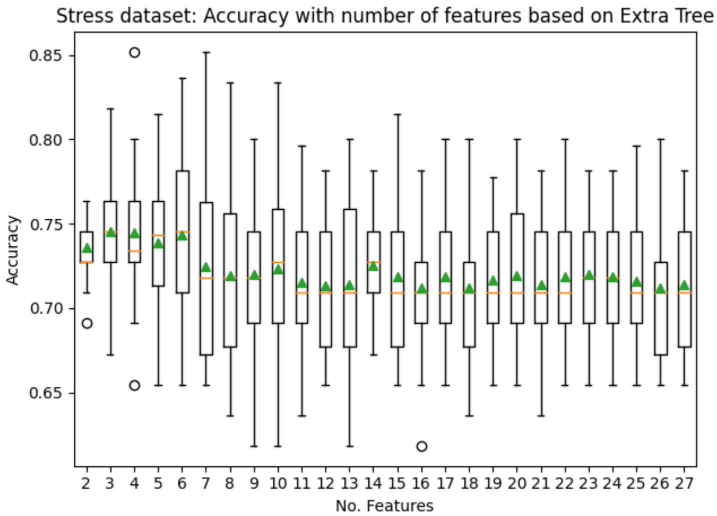


Figure 9. Feature importance based on RFE and ET methods for the stress dataset



(a) Accuracy with number of features based on RFE



(b) Accuracy with number of features based on ET

Figure 10. Accuracy with number of features for the stress dataset

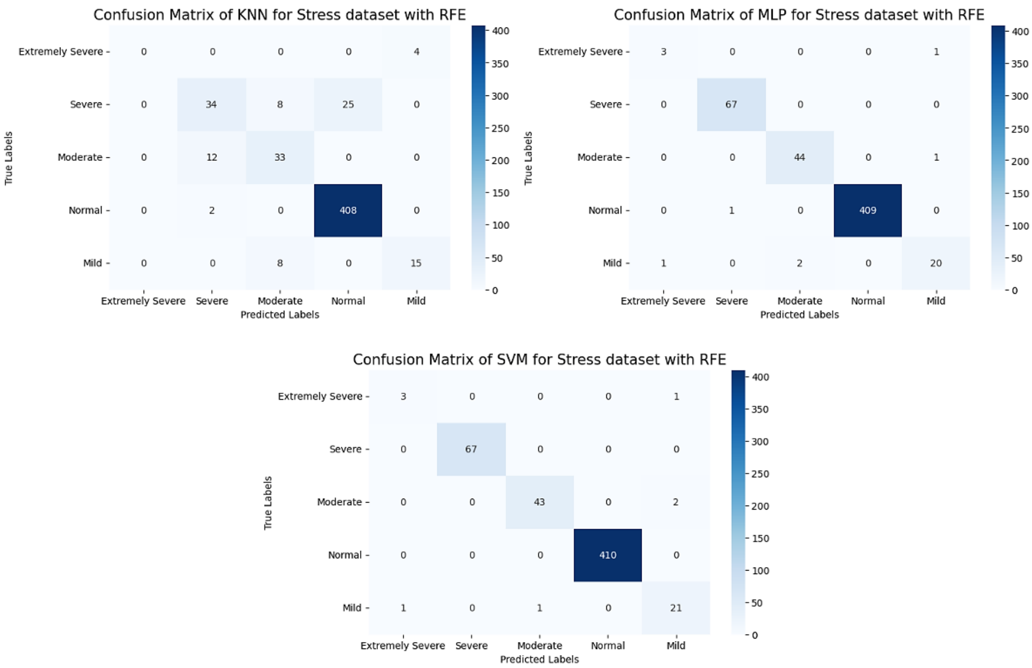


Figure 11. Confusion matrices of three machine learning models for the stress dataset

In the stress dataset, nine features based on the RFE method and three features based on the ET method were selected to tune the hyper-parameters of three machine learning models (KNN, MLP, and SVM). The features derived from the RFE method revealed the best hyper-parameters, resulting in the highest mean accuracy compared to the ET method-based features. The respective mean accuracy for each model was 0.893, 0.989, and 0.991. The best hyper-parameters for the three machine learning models were KNN with the *algorithm: 'brute', n_neighbors: 5*, MLP with the *activation: 'identity', hidden_layer_sizes: 50, solver: 'lbfgs'*, and SVM with the *C: 1, kernel: 'linear'*. Table 3 depicts the details of tuned hyper-parameters with accuracy for these models. SVM, which had RFE method-based feature selection and the best hyper-parameters, revealed the best results in terms of accuracy and F1-score (accuracy = 0.991 and F1-score = 0.920) in predicting stress levels of recovered COVID-19 patients (see Figures 5 and 11).

3.5. Optimal Machine Learning Models for Depression, Anxiety, and Stress of Recovered COVID-19 Patients

In terms of precision, recall, and F1-score, MLP achieved the highest F1-score (0.915) in the depression dataset compared to SVM and KNN. On the other hand, SVM achieved the highest F1-score (1.00) in the anxiety dataset, while both SVM and MLP shared the top F1-score (0.992) in the stress dataset (see Figure 12). Across all three datasets, SVM emerged with the highest accuracy scores (0.984, 1.00, and 0.991, respectively) (see Figure 5). The optimal models for the depression, anxiety, and stress datasets in recovered COVID-19 patients were SVM with hyperparameters (C: 1 and kernel: 'linear'). The depression, anxiety, and stress datasets featured 11, eight, and nine selected features, respectively, with the ET and RFE methods (see Table 4).

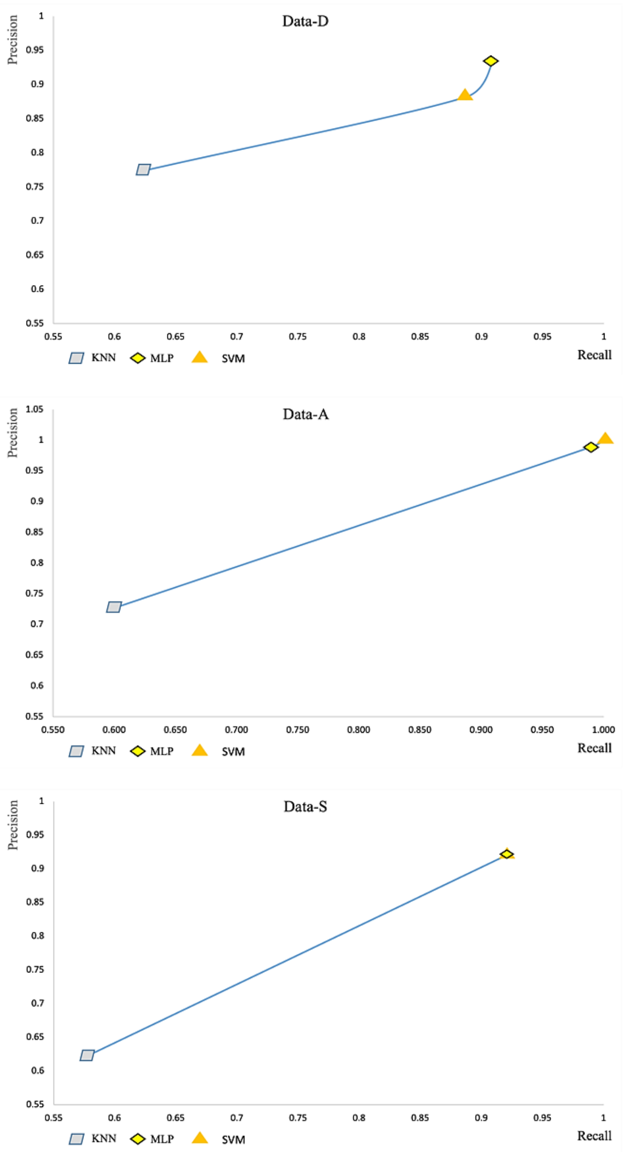


Figure 12. Comparison of machine learning models with the best hyper-parameters in three datasets

Table 4. Characteristics of optimal machine learning models of recovered COVID-19 patients

Characteristics	Data-D	Data-A	Data-S
Feature selection method	ET method	RFE method	RFE method
Number of selected features	11	8	9
Most importance feature	*difficulty initiating tasks* (Q5) *low self-worth* (Q17) *lack of positive feelings* (Q3) *lack of enthusiasm* (Q16) *sense of life being* (Q21) *lack of anticipation* (Q10) *feeling down-hearted* (Q13) *non-communicable diseases* (No_NCDs) *diabetes* *hypertension* *heart disease*	*proximity to panic* (Q15) *heart awareness energy* (Q19) *trembling* (Q7) *unexplained fear* (Q20) *mouth dryness awareness* (Q2) *breathing difficulty* (Q4) *worry about social panic* (Q9) *respiratory disease*	*agitation meaningless* (Q11) *intolerance to interruptions* (Q14) *feeling of using nervous energy* (Q8) *difficulty relaxing* (Q12) *difficulty winding down* (Q1) *tendency to over-react* (Q6) *sensitivity or touchiness* (Q18) *other disease* *respiratory disease*
ML model	SVM	SVM	SVM
Hyper-parameters	C: 1, kernel: 'linear'	C: 1, kernel: 'linear'	C: 1, kernel: 'linear'
Measurement Metric			
Accuracy	0.984	1.00	0.991
Precision	0.882	1.00	0.920
Recall	0.887	1.00	0.920
F1-score	0.885	1.00	0.920

4. Discussion

This study undertook a comprehensive exploration of mental health issues among recovered COVID-19 patients, with a specific focus on depression, anxiety, and stress. Leveraging a dataset that encompassed sociodemographic factors, underlying diseases, and mental health attributes, our analysis utilized machine learning models, including KNN, MLP, and SVM, and revealed promising results in accurately predicting the mental health conditions of recovered COVID-19 patients. SVM emerged as the most effective model across the three datasets. Our findings agree with prior studies on depression, anxiety, and stress, corroborating the importance of understanding mental health among individuals recovered from COVID-19 [25, 27, 32].

In terms of disease conditions, our results highlighted associations between depression, anxiety, and stress with underlying diseases, such as non-communicable diseases, hypertension, diabetes, heart disease, and respiratory disease. This aligns with existing research emphasizing that underlying diseases are significant risk factors contributing to the severity of symptoms related to depression, anxiety, and stress [36–38], underscoring the need for heightened mental health awareness, particularly among those with underlying health issues.

While sociodemographic and COVID-related details did not prove essential for the optimal machine learning models detecting depression, anxiety, and stress in recovered COVID-19 patients, their inclusion exhibited high accuracies (all > 0.700). This observation, portrayed through the number of features and accuracy in the depression, anxiety, and stress datasets, suggests that sociodemographic information and COVID-19-related details may indeed influence the mental well-being of recovered COVID-19 patients, aligning with findings from other studies [28, 36–38].

However, our study has notable limitations. Relying on a single-phase data collection approach may overlook the temporal dynamics of mental health conditions, potentially missing nuances in symptom progression. Additionally, the exclusive focus on depression, anxiety, and stress neglects other crucial dimensions of mental health, potentially limiting the model’s comprehensive applicability. These limitations should be considered when interpreting the existing findings and planning for future research.

5. Conclusion

This study proposed the ML-based framework for depression, anxiety, and stress (DAS) detection from a dataset of recovered COVID-19 patients (e.g., sociodemographic factors, underlying diseases, and mental health attributes) with machine learning models (e.g., KNN, MLP, and SVM), which demonstrated accuracy in predicting mental health conditions. The comprehensive exploration of feature selection methods, particularly RFE and ET, underscored their pivotal role in refining the models for accurate mental health predictions. In the experiment, SVM emerged as the optimal model, surpassing 0.984 accuracy, highlighting its robustness in predicting mental health disorders among recovered COVID-19 patients. The ET method is the most effective feature selection method for the anxiety and stress datasets, while the RFE method performs better in the depression dataset. There are intriguing opportunities with markers, such as physiological and biochemical indicators, to provide a more comprehensive understanding of mental health conditions. In the future, we plan to integrate these markers into survey data to enhance mental health support. This integration holds the potential for personalized intervention strategies tailored to individuals based on machine learning predictions.

6. Declarations

6.1. Author Contributions

Conceptualization, T.T.A., L.D.T.T., and T.L.T.T.; methodology, T.T.A., L.D.T.T., and N.N.H.; software, N.N.H. and A.T.D.; validation, T.T.A., L.D.T.T., and N.N.H.; formal analysis, T.L.T.T. and A.T.D.; resources, T.L.T.T. and A.T.D.; data curation, T.T.A. and T.L.T.T.; writing—original draft preparation, T.T.A., L.D.T.T., T.L.T.T., A.T.D., and N.N.H.; writing—review and editing, T.T.A. and L.D.T.T.; visualization, A.T.D. and N.N.H.; supervision, T.T.A. and L.D.T.T. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support to conduct the research.

6.4. Acknowledgements

The authors would thank School of Informatics and Informatics Innovation Center of Excellence (IICE), Walailak University, Thailand and Dong Thap Medical College, Vietnam for support this research.

6.5. Institutional Review Board Statement

The study was conducted using the reused data without individual information of participants (secondary data) and approved by the Dong Thap Medical College Research Ethics Committee (Approval Number: 01/180/QĐ-CĐYT).

6.6. Informed Consent Statement

Not applicable.

6.7. Declaration of Competing Interest

The authors declare that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7. References

[1] Lopez-Leon, S., Wegman-Ostrosky, T., Perelman, C., Sepulveda, R., Rebolledo, P. A., Cuapio, A., & Villapol, S. (2021). More than 50 long-term effects of COVID-19: a systematic review and meta-analysis. Scientific Reports, 11(1), 16144. doi:10.1038/s41598-021-95565-8.

[2] Rass, V., Beer, R., Schiefecker, A. J., Kofler, M., Lindner, A., Mahlknecht, P., Heim, B., Limmert, V., Sahanic, S., Pizzini, A., Sonnweber, T., Tancevski, I., Scherfler, C., Zamarian, L., Bellmann-Weiler, R., Weiss, G., Djamshidian, A., Kiechl, S., Seppi, K., … Helbok, R. (2021). Neurological outcome and quality of life 3 months after COVID-19: A prospective observational cohort study. European Journal of Neurology, 28(10), 3348–3359. doi:10.1111/ene.14803.

[3] de Oliveira Almeida, K., Nogueira Alves, I. G., de Queiroz, R. S., de Castro, M. R., Gomes, V. A., Santos Fontoura, F. C., Brites, C., & Neto, M. G. (2023). A systematic review on physical function, activities of daily living and health-related quality of life in COVID-19 survivors. Chronic Illness, 19(2), 279–303. doi:10.1177/17423953221089309.

[4] Pei, H., Wu, Q., Xie, Y., Deng, J., Jiang, L., & Gan, X. (2021). A Qualitative Investigation of the Psychological Experiences of COVID-19 Patients Receiving Inpatient Care in Isolation. Clinical Nursing Research, 30(7), 1113–1120. doi:10.1177/10547738211024807.

[5] Engel, F. D., Da Fonseca, G. G. P., Cechinel-Peiter, C., Backman, C., Da Costa, D. G., & De Mello, A. L. S. F. (2023). Impact of the COVID-19 Pandemic on the Experiences of Hospitalized Patients: A Scoping Review. Journal of Patient Safety, 19(2), E46–E52. doi:10.1097/PTS.0000000000001084.

[6] Sun, N., Wei, L., Shi, S., Jiao, D., Song, R., Ma, L., ... & Wang, H. (2020). A qualitative study on the psychological experience of caregivers of COVID-19 patients. American journal of infection control, 48(6), 592-598. doi:10.1016/j.ajic.2020.03.018.

[7] ADAA. (2023). Depression symptoms. Anxiety and Depression Association of America, Maryland, United States. Available online: https://adaa.org/understanding-anxiety/depression/symptoms. (accessed on May 2023).

[8] NIMH Resources. (2023). Depression symptoms. National Institutes of Mental Health, Maryland, United States. Available online: https://www.nimh.nih.gov/health/topics/depression. (accessed on May 2023).

[9] Tyshchenko, Y. (2018). Depression and anxiety detection from blog posts data. Institute of Computer Science Computer Science Curriculum, University of Tartu, Tartu, Estonia.

[10] Kralj, M. M. (1989). Life-change stress and stress symptoms. Journal of College Student Development, 30, 333.

[11] Marks, H. (2023). Stress Symptoms. Health & Balance. Stress Management, WedMD website, 1–4. Available online: https://www.webmd.com/balance/stress-management/stress-symptoms-effects_of-stress-on-the-body#1

[12] Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. Journal of General Internal Medicine, 16(9), 606–613. doi:10.1046/j.1525-1497.2001.016009606.x.

[13] Oei, T. P. S., Sawang, S., Goh, Y. W., & Mukhtar, F. (2013). Using the Depression Anxiety Stress Scale 21 (DASS-21) across cultures. International Journal of Psychology, 48(6), 1018–1029. doi:10.1080/00207594.2012.755535.

[14] Bakkeli, N. Z. (2023). Predicting psychological distress during the COVID-19 pandemic: do socioeconomic factors matter?. Social Science Computer Review, 41(4), 1227-1251. doi:10.1177/08944393211069622.

[15] Garriga, R., Mas, J., Abraha, S., Nolan, J., Harrison, O., Tadros, G., & Matic, A. (2022). Machine learning model to predict mental health crises from electronic health records. Nature medicine, 28(6), 1240-1248. doi:10.1038/s41591-022-01811-5.

[16] Pabreja, K., Singh, A., Singh, R., Agnihotri, R., Kaushik, S., & Malhotra, T. (2021). Stress prediction model using machine learning. Proceedings of International Conference on Artificial Intelligence and Applications, 57-68. doi:10.1007/978-981-15-4992-2_6.

[17] Vilca, L. W., Chávez, B. V., Fernández, Y. S., Caycho-Rodríguez, T., & White, M. (2023). Impact of the fear of catching COVID-19 on mental health in undergraduate students: A Predictive Model for anxiety, depression, and insomnia. Current Psychology, 42(16), 13231-13238. doi:10.1007/s12144-021-02542-5.

[18] Dinga, R., Marquand, A. F., Veltman, D. J., Beekman, A. T., Schoevers, R. A., van Hemert, A. M., ... & Schmaal, L. (2018). Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: a machine learning approach. Translational Psychiatry, 8(1), 241. doi:10.1038/s41398-018-0289-1.

[19] Perzow, S. E., Hennessey, E. M. P., Hoffman, M. C., Grote, N. K., Davis, E. P., & Hankin, B. L. (2021). Mental health of pregnant and postpartum women in response to the COVID-19 pandemic. Journal of affective disorders reports, 4, 100123. doi:10.1016/j.jadr.2021.100123.

[20] Qasrawi, R., Amro, M., VicunaPolo, S., Abu Al-Halawa, D., Agha, H., Abu Seir, R., Hoteit, M., Hoteit, R., Allehdan, S., Behzad, N., Bookari, K., AlKhalaf, M., Al-Sabbah, H., Badran, E., & Tayyem, R. (2022). Machine learning techniques for predicting depression and anxiety in pregnant and postpartum women during the COVID-19 pandemic: A cross-sectional regional study. F1000 Research, 11. doi:10.12688/f1000research.110090.1.

[21] Nguyen, H. V., & Byeon, H. (2022). Explainable Deep-Learning-Based Depression Modeling of Elderly Community after COVID-19 Pandemic. Mathematics, 10(23), 4408. doi:10.3390/math10234408.

[22] Hawes, M. T., Schwartz, H. A., Son, Y., & Klein, D. N. (2023). Predicting adolescent depression and anxiety from multi-wave longitudinal data using machine learning. Psychological Medicine, 53(13), 6205-6211. doi:10.1017/S0033291722003452.

[23] Trivedi, N. K., Tiwari, R. G., Witarsyah, D., Gautam, V., Misra, A., & Nugraha, R. A. (2022). Machine Learning Based Evaluations of Stress, Depression, and Anxiety. Proceedings - International Conference Advancement in Data Science, E-Learning and Information Systems, ICADEIS 2022, 1–5. doi:10.1109/ICADEIS56544.2022.10037336.

[24] Chung, J., & Teo, J. (2023). Single classifier vs. ensemble machine learning approaches for mental health prediction. Brain Informatics, 10(1), 1–10. doi:10.1186/s40708-022-00180-6.

[25] Ajith Sankar, R., & Juliet, S. (2023). Investigations on Machine Learning Models for Mental Health Analysis and Prediction. 2023 2nd International Conference on Electrical, Electronics, Information and Communication Technologies, ICEEICT 2023, 1–7. doi:10.1109/ICEEICT56924.2023.10157385.

[26] Deplancke, C., Somerville, M. P., Harrison, A., & Vuillier, L. (2023). It’s all about beliefs: Believing emotions are uncontrollable is linked to symptoms of anxiety and depression through cognitive reappraisal and expressive suppression. Current Psychology, 42(25), 22004-22012. doi:10.1007/s12144-022-03252-2.

[27] Nemesure, M. D., Heinz, M. V., Huang, R., & Jacobson, N. C. (2021). Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. Scientific reports, 11(1), 1980. doi:10.1038/s41598-021-81368-4.

[28] Malik, S. S., & Khan, A. (2023). Anxiety, Depression and Stress prediction among College Students using Machine Learning Algorithms. 2023 2nd International Conference on Electrical, Electronics, Information and Communication Technologies, ICEEICT 2023, 1–5. doi:10.1109/ICEEICT56924.2023.10157693.

[29] Prout, T. A., Zilcha-Mano, S., Aafjes-van Doorn, K., Békés, V., Christman-Cohen, I., Whistler, K., Kui, T., & Di Giuseppe, M. (2020). Identifying Predictors of Psychological Distress during COVID-19: A Machine Learning Approach. Frontiers in Psychology, 11, 586202. doi:10.3389/fpsyg.2020.586202.

[30] Nison, P., Vuttipittayamongkol, P., Boonyapuk, P., & Kemavuthanon, K. (2023). A Machine Learning Approach for Depression Screening in College Students Based on Non-Clinical Information. 2023 International Conference on Cyber Management and Engineering, CyMaEn 2023, 413–417. doi:10.1109/CyMaEn57228.2023.10051001.

[31] Zulfiker, M. S., Kabir, N., Biswas, A. A., Nazneen, T., & Uddin, M. S. (2021). An in-depth analysis of machine learning approaches to predict depression. Current Research in Behavioral Sciences, 2, 100044. doi:10.1016/j.crbeha.2021.100044.

[32] Knolle, F., Ronan, L., & Murray, G. K. (2021). The impact of the COVID-19 pandemic on mental health in the general population: a comparison between Germany and the UK. BMC psychology, 9, 1-17. doi:10.1186/s40359-021-00565-y.

[33] Rezapour, M., & Hansen, L. (2022). A machine learning analysis of COVID-19 mental health data. Scientific Reports, 12(1), 14965. doi:10.1038/s41598-022-19314-1.

[34] Cho, S. E., Geem, Z. W., & Na, K. S. (2021). Predicting depression in community dwellers using a machine learning algorithm. Diagnostics, 11(8), 1429. doi:10.3390/diagnostics11081429.

[35] Jain, T., Jain, A., Hada, P. S., Kumar, H., Verma, V. K., & Patni, A. (2021). Machine Learning Techniques for Prediction of Mental Health. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 1606-1613. doi:10.1109/ICIRCA51532.2021.9545061.

[36] Trang, L. T. T., Le, C. N., Chutipatana, N., Shohaimi, S., & Suwanbamrung, C. (2023). Prevalence and Predictors of Depression, Anxiety, and Stress among Recovered Covid-19 Patients in Vietnam. Annals of the National Institute of Hygiene, 74(2), 217–230. doi:10.32394/rpzh.2023.0261.

[37] Phu, D. H., Maneerattanasak, S., Shohaimi, S., Trang, L. T. T., Nam, T. T., Kuning, M., Like, A., Torpor, H., & Suwanbamrung, C. (2023). Prevalence and factors associated with long COVID and mental health status among recovered COVID-19 patients in southern Thailand. PLoS ONE, 18(7 July), 289382. doi:10.1371/journal.pone.0289382.

[38] Suwanbamrung, C., Pongtalung, P., Trang, L. T. T., Phu, D. H., & Nam, T. T. (2023). Levels and risk factors associated with depression, anxiety, and stress among COVID-19 infected adults after hospital discharge in a Southern Province of Thailand. Journal of Public Health and Development, 21(1), 72–89. doi:10.55131/jphd/2023/210106.

[39] Huynh, G., Nguyen, H. V., Vo, L. Y., Le, N. T., & Nguyen, H. T. N. (2022). Assessment of insomnia and associated factors among patients who have recovered from COVID-19 in Vietnam. Patient preference and adherence, 1637-1647. doi:10.2147/PPA.S371563.

[40] Iguyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3(March), 1157–1182.

[41] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1–3), 389–422. doi:10.1023/A:1012487302797.

[42] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63(1), 3–42. doi:10.1007/s10994-006-6226-1.

[43] Vaishnavi, K., Kamath, U. N., Rao, B. A., & Reddy, N. S. (2022). Predicting mental health illness using machine learning algorithms. Journal of Physics: Conference Series, IOP Publishing, 2161, 012021. doi:10.1088/1742-6596/2161/1/012021.

[44] Singh, S., Gupta, H., Singh, P., & Agrawal, A. P. (2022). Comparative Analysis of Machine Learning Models to Predict Depression, Anxiety and Stress. Proceedings of the 2022 11th International Conference on System Modeling and Advancement in Research Trends, SMART 2022, 1199–1203. doi:10.1109/SMART55829.2022.10047752.

[45] Jha, A., Abirami, M. S., & Kumar, V. (2023). Predictive Model for Depression and Anxiety Using Machine Learning Algorithms. Communications in Computer and Information Science, 1719 CCIS, 133–147. doi:10.1007/978-3-031-27622-4_11.

[46] Priya, A., Garg, S., & Tigga, N. P. (2020). Predicting anxiety, depression and stress in modern life using machine learning algorithms. Procedia Computer Science, 167, 1258-1267. doi:10.1016/j.procs.2020.03.442.

[47] Shobhika, Kumar, P., & Chandra, S. (2022). Prediction and comparison of psychological health during COVID-19 among Indian population and Rajyoga meditators using machine learning algorithms. Procedia Computer Science, 218, 697–705. doi:10.1016/j.procs.2023.01.050.

[48] Nayan, M. I. H., Uddin, M. S. G., Hossain, M. I., Alam, M. M., Zinnia, M. A., Haq, I., ... & Methun, M. I. H. (2022). Comparison of the performance of machine learning-based algorithms for predicting depression and anxiety among University Students in Bangladesh: A result of the first wave of the COVID-19 pandemic. Asian Journal of Social Health and Behavior, 5(2), 75-84. doi:10.4103/shb.shb_38_22.

[49] Kim, S. W., & Chang, M. C. (2023). The usefulness of machine learning analysis for predicting the presence of depression with the results of the Korea National Health and Nutrition Examination Survey. Annals of Palliative Medicine, 12(4), 748–756. doi:10.21037/apm-23-78.

[50] Lovibond, S. H., & Lovibond, P. F. (1995). Manual for the Depression Anxiety Stress Scales. Psychology Foundation of Australia, 56, 42.