

## RESEARCH ARTICLE

# Using a Classifier Ensemble for Preventing Burnout in University Students: A Study Case in Valparaíso

LARRAÍN GONZALO<sup>1</sup>, ROJAS-MORALES NICOLÁS<sup>1</sup> , (Member, IEEE),  
GONZALEZ NICOLÁS<sup>2</sup>, AND OLCAY DANIEL<sup>2</sup>

<sup>1</sup>Computer Science Department, Universidad Técnica Federico Santa María, Valparaíso 2390123, Chile

<sup>2</sup>Servicios y Soluciones de Calidad de Vida, Valparaíso 2381604, Chile

Corresponding author: Rojas-Morales Nicolás (nicolas.rojasm@usm.cl)

This work was supported in part by the Project INNOVA REGION under Grant 22IR-215189. The work of Rojas-Morales Nicolás was supported by Universidad Técnica Federico Santa María (UTFSM) Dirección General de Investigación, Innovación y Emprendimiento (DGIIE) Funding Project under Grant PI\_LII\_2022\_03.

**ABSTRACT** University students are constantly exposed to high-tension situations and peaks of stress caused by the difficulty from their many responsibilities. These situations can produce disorders in students, such as psychosomatic, behavioral, or emotional disorders. Burnout is a work-related syndrome that can be observed in students. This syndrome considers depersonalization, emotional exhaustion, and diminished feelings of personal accomplishment. We aim to detect anxiety, depression, and burnout symptoms in university students to prevent further negative consequences. For this, we design a questionnaire using well-known instruments to detect these signs. We propose to use an ensemble of classifiers, including random forest and artificial neural networks, to predict a set of four possible disturbances in persons. The proposal will be used in the *Human Place* project to suggest strategies to tackle four disturbance types. This study considers the participation of 93 persons from the Valparaíso region in Chile. Results show that the evaluated algorithms can predict the presence or absence of the disturbances with high accuracy levels and a low number of false negative cases. We also present a detailed analysis of which questions were relevant in the classification task of each algorithm.


**INDEX TERMS** Burnout syndrome, mental health, neural networks, random forest.

## I. INTRODUCTION

University students can be exposed to high-stress levels, considering the possibilities of success or failure, the educational expectations, the continual adaptation process between several amounts of work and wellness, and the effective management of their responsibilities. On the other hand, several other factors can also influence the student's mental health, such as competitiveness, economic instability, family pressure and their expectations, and employment prospects, among other aspects. These situations can lead to a syndrome named burnout. The concept of burnout was initially introduced by Freudenberger, who described it as a syndrome characterized by symptoms of anxiety, depression,

and a profound lack of energy [1]. This syndrome can also manifest through emotional exhaustion, depersonalization, and diminished feelings of personal accomplishment [2].

Student burnout syndrome emerges as a psychosocial phenomenon considering the long-term exposure to stressors inherent and complicated factors in the educational process, producing psychosomatic, behavioral, and emotional disorders. Consequently, students can have complex symptoms such as cardiac alterations, fatigue, migraines, sleeping disturbances, substance abuse, eating disorders, the desire to drop out of studies, irritability, difficulty focusing, low self-esteem, and depression [3]. This work is involved in a project named *Human Place* whose objective is the early detection of somatic symptoms of burnout syndrome in university students and to generate an early intervention to support students in their life and learning processes.

The associate editor coordinating the review of this manuscript and approving it for publication was Sunil Karamchandani .

We propose to use an ensemble of classifiers to predict four disturbances: physical, emotional, cognitive, and organizational. To detect signs of these disturbances, we used relevant well-known instruments from literature, such as the World Health Organization Quality of Life instrument and the Beck Anxiety Inventory, among others. The described scenario considers a multi-label classification problem because each person can simultaneously have more than one disturbance. To perform successful and accurate predictions, we propose transforming this problem into a set of binary classification problems. We evaluate the usage of random forests and neural networks algorithms, for detecting each disturbance type. This study considers 93 participants from the Valparaíso region of Chile. To evaluate our work, we generated a set of synthetic datasets based on the participant's responses, followed by a set of guidelines provided by the team's psychologist. The proposed ensemble is a *Human Place* mobile application component that suggests strategies and routines to tackle the mentioned disturbances. Given the financial and logistical barriers usually associated with accessing mental health treatment, we aim to provide strategies that allow university students to self-address potential burnout symptoms. However, it is important to emphasize that the substitution of the critical role of therapy and the work of specialists such as psychologists and psychiatrists is not our objective.

The contributions of our work are:

- The proposal of a classifier ensemble approach to suggest mental-health strategies,
- The assessment of the proposed ensemble considering synthetic and real data.

This work is an expanded version of [4], which includes further details of the study, an improved related work and proposal sections, implementing, tuning, and evaluating of a random forest algorithm for each class, the design and generation of new datasets, results of the algorithms in the new datasets, and an improved relevant feature analysis.

The article is organized in the following structure: Section II briefly presents some definitions and a revision of existing articles in literature, Section III presents the details of our study, Section IV presents the proposed classifier ensemble that contains Random Forest and Neural Networks, Section V details the experimental setup, the synthetic dataset generation procedure and the hyperparameter tuning process, Section VI presents the obtained results of each classifier and a detailed analysis of the most relevant features of each classifier. Section VIII presents the conclusions of our work and some possible paths for future work.

## II. PREVIOUS DEFINITIONS AND RELATED WORK

A classification task is the natural process of assign classes to cases or group things. For this process, we can formally consider a feature vector  $x_k$ , and a class  $c$  from a set of existing classes. In a supervised learning scenario, considering a set of training examples composed of pairs  $(x_k, c)$ , the idea is to

obtain a function  $f(x)$  that maps each feature vector  $x_k$  to a label  $c$ . When only two mutually exclusive classes are considered, the problem is considered as a binary-classification one. On the other hand, when classes are not mutually exclusive, and more than one label can be assigned to an instance, it is known as a *multi-label classification problem*. Classifying genres for movies is an example of a multi-label classification problem, where a movie can be labeled as science-fiction, drama, and comedy at the same time [5]. Several articles work with multi-label classification problems, and some examples are related to the analysis of social media content [6], classification of user customer reviews [7], [8], and topic recommendations for software repositories [9]. The existing methods to tackle multi-label classification can be separated in *problem transformation methods* or *algorithm adaptation methods* [10]. The first-mentioned methods transform a multi-label classification problem into one or more single-label classification problems. The second one extends the single-label approaches for solving multi-label classification problems. In this work, we use problem transformation methods to work with the detection of multiple disturbances in participants of the study.

## A. LITERATURE REVIEW

In recent years, Machine Learning (ML) techniques have been mainly used for decision support, detection, and diagnosis of mental health conditions. This section briefly reviews some existing works that apply ML in the mental health research area.

In [11], a ML framework was proposed for the detection of multiple levels of stress. Particularly, the framework contains Support Vector Machine (SVM), Logistic Regression (LogR), and Naive Bayes Classification (NBC) algorithms. In this study, twenty-two participants were stress-induced performing mental arithmetic tasks based on the Montreal Imaging Stress Task Paradigm. Four levels of stress were observed, considering different difficult mathematical tasks and observing time pressure, distraction, and evaluative pressure. In the stress condition, the arithmetic tasks were presented along with negative comments to the participants. Also, a control condition was considered without negative comments and extra time for the tasks. Electroencephalogram (EEG) signal was used to analyze the stress of participants, considering the following extracted features: absolute power, relative power, coherence, amplitude asymmetry, and phase lag. Data was standardized using a standard score, and a feature selection procedure was performed (ROC AUC, t-test, and Bhattacharya distance). Results show that the NBC algorithm has a higher accuracy level for levels 1, 3, and 4 of stress, and for level 2, the SVM approach.

A study of the Post-Traumatic Stress Symptoms (PTSS) in physicians on the frontlines of COVID-19 was presented in [12]. The objectives of the study were (1) to evaluate the symptoms of PTSS among frontline physicians compared to second-line ones, (2) to identify the higher-risk group among

them, (3) to predict the PTSS risk in the higher-risk group using ML algorithms, and (4) determine possible patterns between the predictors. The study participants were 1,017 physicians from the US states with the greatest COVID-19 caseloads. In the study, to evaluate the PTSS, they used the following questionnaires: the Patient Health Questionnaire (PHQ-9), the 5-point Burnout scale, the Post-Traumatic Stress Disorder Checklist (PCL-5), the Survey of Perceived Organizational Support scale, the Connor-Davison Resilience Scale, stressors in the work environment (work-load, non-routine work, perceived stigma from treating COVID-19 patients, among others), and demographics and workplace characteristics (age, sex, ethnicity, among others). The authors evaluate the following algorithms: LogR, Bagging, NBC, SVM, Gradient Boosting Method, Bayesian Additive Regression Trees, Random Forest (RF), and Neural Networks. Considering the accuracy, ROC AUC, recall, precision, F1-score, the interpretability of all the mentioned algorithms, the authors selected RF as the most suitable technique for their prediction problem. Also, the authors presented the set of key predictors where the PHQ-9 and the Burnout Score were the most important among the 20 used predictors.

A study that aims to discover factors causing anxiety and sleep disorders during the COVID-19 lockdown is presented in [13]. In this study, 704 persons from India responded to a survey that considered: the Pittsburgh Sleep Quality Index (PSQI) questionnaire, the General Anxiety Disorder (GAD-7) questionnaire, demographic information, and occupation-targeted questions. The participants were school and college students, working professionals, healthcare persons, and retired persons, among others. The study has two objectives (1) to detect the presence or absence of sleep disorders through the responses of college students using a Random Forest algorithm and (2) to cluster participants based on their GAD-7 and PSQI scores using a K-means clustering algorithm. Results show that 67% of the participants had poor sleep quality, and 20% had a high anxiety score. For the RF model, the features are anxiety score, worry about the inability to understand concepts taught online, involvement of parents, college hours, worrying about other workloads, and deadlines.

In [14], Italian university students' mental health was investigated, considering students from different areas (e.g., Arts, Law, Economics, Medicine, Psychology and Mental Health, Neuroscience, and Pharmacy). The study considered two stages, the first one with 1,388 participants, and six months later, 557 persons participated in the second stage. The participants give personal information about their demographics, health, lifestyle habits, and economic and financial situation. Also, the Obsessive Compulsive Inventory-Revised (OCI-R), Eating Habits Questionnaire (EHQ), BAI, and other instruments were used in the study. This study also used regression and machine learning techniques to determine associations between demographic variables and instruments (using Multiple Regression), to assess the contribution of

demographic variables and clinically relevant features (using binomial Regression Models), to detect variables that had a significant role in determine the severity of symptoms (using Random Forest), to determine which variables could be leveraged to predict possible variations in depressive symptoms at the first stage (using Random Forest), to reveal what factors were predictive of a change in depressive symptoms and suicidal ideation at the first stage (using Random Forest). The results show that one out of five students have severe depressive symptoms or suicidal ideation. Moreover, Random Forest has high accuracy levels in students who maintained well-being or in the absence of suicidal ideation. However, the accuracy of detecting worsening students was lower than 50%.

A recently published survey reviews 300 works published between 2004 and 2018 [15]. Authors consider only peer-reviewed publications that apply an ML approach to address mental health (conceptualized using the World Health Organization's definition). For each article, the authors analyzed the mental health application, the ML technique, the data type, and the study results. Mainly, four domains of mental health applications were identified:

- Detection and diagnosis of mental health conditions in individuals,
- Prognosis, treatment, and support of the progress of mental health conditions, or exploration of treatment or support opportunities,
- Public health applications to monitor mental health conditions, and
- Research and clinical administration to improve the administrative process in clinical work.

The results show that nearly 90% of the works considered supervised learning and classification approaches (support vector machines, naive Bayes, decision trees), and 8% used unsupervised learning and clustering approaches (k-nearest neighbors, k-means clustering). Also, regarding mental health conditions, 30% of the articles addressed depression, 15% addressed Alzheimer's disease and another cognitive decline, 10% schizophrenia, 10% stress, and 6% suicide. Authors highlight a set of possible paths for future work, considering: (1) the detection of other conditions (e.g., anxiety disorders, eating disorders, and neurodevelopmental disorders), (2) the usage of social media data can be further considered, (3) the importance of greater collaboration between researchers and clinicians to access to better data sets, and (4) the usage of less structured prospective data for real-time ML analysis.

Several other applications and psychological studies have been published in the literature that present the usage of machine learning techniques in mental health conditions, such as: a pre-clinical mental health dataset to classify anxiety problems with SVM, Multilayer Perceptron, and RF [16], the usage of 10-fold cross-validation, SVM, and gradient boosting tree models to classify major depression disorder patients based on immunometabolic and oxidative stress biomarkers and lifestyle habits [17], the usage of

Synthetic Minority Oversampling Technique (SMOTE) and RF to detect depression on women from Malasia [18], the application of RF to predict some mental disorders and drug abuse [19], an analysis of data from Twitter to detect depression using recurrent neural network and convolutional neural network [20], the proposal of a framework for mental health detection in educational scenarios named CASTLE [21] using multi-view social network embedding (MOON), SMOTE and DNN.

### III. MATERIALS AND METHODS

This section presents the main features of our study. Mainly, it details the study's design, a description of the participants, the instruments used in our questionnaire, and the defined classes.

#### A. DESIGN

The research design was experimental and observational. The participants of this study remotely answered the questionnaires considering the restrictions imposed by the COVID-19 pandemic and to reduce the contagion risks. The participation was anonymous and voluntary. This procedure was performed during May and July 2022. The Human Place team considered different recommendations to avoid some risks, bias, and difficulties that can happen in the remote application of the questionnaires related to [22]: (1) technology (e.g., ensure that each person has the required resources during the evaluation), (2) personal features (e.g., ensure the privacy, comfort during the process, and discretion about their responses), and (3) the selection and adaptation of the tests (adaptation of the questions to be remotely answered, respecting their viability, validity, and norms).

#### B. LOCATION AND INVITATION PROCESS OF PARTICIPANTS

Human Place team is based in the Valparaíso region of Chile. Moreover, they chose the Valparaiso region to perform this study because it contains an important population of university students from Chile. Considering a regional report from 2021 [23], this region has nearly 10% of the traditional universities in Chile (7 out of 68), nearly 9% of the private universities of Chile (9 out of 105), nearly 13% of the professional institutes of Chile (out of a total of 141), 14% of the technical formation centers of Chile (17 out of 120 in Chile).

The authors invited nearly 200 persons to answer a preliminary survey via Google Forms about mental health, their quality of life, and their usage of mobile applications for time organization. Authors sent the invitation through different platforms (e.g., email, and social media). We received the answer to this preliminary survey from 120 persons. Then, the authors sent back an invitation to participate in this study. Participants were invited to answer a questionnaire with a set of psychometric instruments voluntarily. Finally, this study considers a population of 93 participants from the Valparaiso region of Chile. Table 1 shows details

TABLE 1. Details of the participants.

Feature of participants	Value	Percentage
Gender	Female	68%
	Male	32%
Age	16-19	11%
	20-29	63%
	30-39	23%
	40-49	3%
Occupancy	Student	34%
	Employee	53%
	Student+Employee	13%
City	Valparaíso	41%
	Viña del Mar	31%
	Quilpué	14%
	Concón	7%
	Villa Alemana	5%
	Quillota	2%

of the study participants, considering their gender, age, occupancy, and the city where they live. The Human Place team invited an equal number of males and females to participate in this study. However, most of the people that accepted the invitation were females. Regarding the age and occupancy of participants, the Human Place project mainly focuses on young persons considering students, recently graduated, or persons that study and work simultaneously. Consequently, persons between 16 and 49 years and students and employees were invited to be part of the study. However, most participants are current employees or students between 20 and 29 years old. About the city where the participants live, around 80% of the participants live in seaside cities such as Valparaíso, Viña del Mar, and Concón. On the other hand, nearly 20% of the participants live in inland cities such as Quilpué, Villa Alemana, and Quillota.

#### C. INSTRUMENTS

For this study, we construct a questionnaire with 83 questions, considering four instruments: *Beck Depression Inventory* (BDI) [24], *Beck Anxiety Inventory* (BAI) [25], the *Student Burnout One-Dimensional Scale* (EUBE)<sup>1</sup> [26], and the *World Health Organization Quality of Life* (WHOQOL) [27].

##### 1) BAI

This test aims to measure anxiety levels that can provoke hyperactivity, attention deficit, or depressive syndromes [28]. This instrument has been severally used in literature to diagnose anxiety in teenagers and adult patients [29]. The BAI instrument presents a list of twenty-one anxiety symptoms, which participants are invited to answer considering how they fell during the last month and the day the instrument is answered. Some examples of these symptoms are:

- “Dizzy or lightheaded”
- “Fear of losing control”

<sup>1</sup>Originally entitled: *Escala Unidimensional de Burnout Estudiantil*.

The answers are Likert-type scale with four possible levels:

- 0 - “Not at all”
- 1 - “Mildly but it didn’t bother me much”
- 2 - “Moderately - it wasn’t pleasant at times”
- 3 - “Severely - it bothered me a lot”

Raw scores range from 0 to 63. The scores are classified as minimal anxiety (0 to 7), mild anxiety (8 to 15), moderate anxiety (16 to 25), and severe anxiety (30 to 63).

## 2) BDI

The BDI is a self-score instrument designed in 1961 for measuring depression, and that has been severally applied in the literature [30]. The test presents 21 questions related to symptoms that can appear in depression. The symptoms are: (1) mood, (2) pessimism, (3) sense of failure, (4) lack of satisfaction, (5) guilt feelings, (6) sense of punishment, (7) self-dislike, (8) self-accusation, (9) suicidal wishes, (10) crying, (11) irritability, (12) social withdrawal, (13) indecisiveness, (14) distortion of body image, (15) work inhibition, (16) sleep disturbance, (17) fatigability, (18) loss of appetite, (19) weight loss, (20) somatic preoccupation, and (21) loss of libido. The scale of the answers is Likert-type, with the lowest score representing the absence of the symptom (0) and (3) the symptom is severally present. The scores are classified in:

- 0 - 9 - “minimal depression”
- 10 - 18 - “mild depression”
- 19 - 29 - “moderate depression”
- 30 - 63 - “severe depression”

## 3) EUBE

This instrument was proposed to detect burnout syndrome in students [26]. It has been applied in students to determine the presence of symptoms of burnout and their relationship with emotional intelligence and academic achievement [31], [32]. The EUBE instrument consists of fifteen affirmations about how participants feel about their student responsibilities, their interest in attending classes, school grades, and paying attention in classes. Some examples of affirmations are:

- “I feel sleepy during classes”.
- “It is increasingly difficult for me to pay attention to the teacher”.

The answers are in a Likert-type scale considering: (0) Never, (1) Sometimes, (2) Commonly, and (3) Always. The score can be classified four classes:

- 0 - 19 - absence of the syndrome,
- 20 - 38 - slight presence of the syndrome,
- 39 - 56 - moderate presence of the syndrome,
- 57 - 75 - deep presence of the burnout syndrome.

## 4) WHOQOL

The WHOQOL instrument was designed to measure the self-perception of quality of life of the participants, considering their position in life, expectations, cultural standards and concerns [33]. Particularly, in this study, we use the

reduced version (WHOQOL-BREF) of the WHOQOL-100 instrument that consists of 26 questions. The answers are rated on a five-point Likert scale, with (1) the lowest score and (5) the highest one. The instrument is divided in four domains:

- Physical Health, about activities of daily living, dependence on medicinal substances, sleep and rest, among others.
- Psychological, about bodily image and appearance, positive or negative feelings, self-esteem, among others.
- Social Relationships, about personal relationships, social support, and sexual activity.
- Environment, about home environment, transport, freedom, among others.

Some examples of the questions are “How satisfied are you with your sleep?” and “How satisfied are you with your capacity for work?”.

## D. CLASSES AND STRATEGIES

In analyzing the obtained results, observing possible response patterns, and considering each instrument’s scores, the *Human Place* team defines four classes:

- *Cognitive Disturbance* (COG), can affect learning, memory, and cognitive processes.
- *Emotional Disturbance* (EMO), where the self-perception of their feelings and emotions can be affected.
- *Organizational Disturbance* (ORG), where persons can have difficulties in planning their daily work/study time.
- *Physical Disturbance* (PHY), where critical factors can be reduced libido, drowsiness, and medication dependence.

It is important to mention that the classes are not mutually exclusive (e.g. a person can have a COG and a ORG disturbances simultaneously).

The *Human Place* mobile application provides a set of *wellness strategies* for managing the symptoms of the four disturbances. Each wellness strategy defines a routine of tasks that include four key habits: meditation, acceptance, planning, and gratitude. The *Human Place* team particularly defined the intern configuration of each wellness strategy.

In the PHY class, the wellness strategy involves a four-week daily routine, one week per habit. The objective is to enrich organizational skills and evaluate the daily achievements associated with routine activities.

For the EMO class, the strategy considers two weeks of gratitude activities, followed by one week of meditation and one week of planning exercises. This wellness strategy promotes emotional understanding and self-organization, improving the self-evaluation of daily actions.

The COG class strategy considers five weeks. The first week includes acceptance exercises, and the second one is focused on exercising gratitude. Then, the third week exercises the planning methods, and the last two ones are

focused on meditation. The strategy for the COG disturbance tries to develop self-acceptance, gratitude, and the ability to plan daily activities autonomously.

The ORG disturbance strategy focuses on working on the habit of planning. This strategy considers a four-week plan: (1) one week of planning exercises, (2) one week of acceptance, and (3) two more weeks of planning exercises. This strategy allows students to develop effective planning skills for better organization of their tasks.

When no disturbance is detected in a particular student, the app suggests a four-week routine considering one week per habit. This routine helps individuals maintain and reinforce their overall abilities.

#### IV. ALGORITHM PROPOSAL

Considering the presented description of the problem, its features, and defined classes, this problem is a multi-label classification problem. Particularly, there are four not mutually exclusive classes (one per disturbance). To tackle this problem, we propose to use an ensemble of single-label classifiers, considering one specifically trained classifier per class. Here, we use a problem transformation method to consider a multi-label classification problem as a set of binary classification problems, each considering the absence or presence of a particular disturbance. The main reason for using this design is related to the care that should be considered in classifying persons, considering information about their mental health. The *Human Place* app will suggest a (set of) strategy(es) to tackle each disturbance. Along the same line, the *Human Place* team requires that the detection of the presence or absence of each class can be successfully made.

Considering a unique approach that manages the four classes simultaneously can be a more general approach and easier for the client to be embedded in the mobile application. However, considering the No Free Lunch theorem, more general ML approaches tend to perform less accurately than specifically designed and tuned techniques [34]. For this reason, we will work with a set of binary classification problems with an ensemble of classifiers.

In order to evaluate the possibility of detecting if a person is healthy without the need to previously executing four classifiers, we include a fifth class named Control condition (CTRL). The idea is to contrast the response of the classifiers and, in the future, use this classifier's opinion to improve the classification processes' performance.

Figure 1 presents the structure of our proposed ensemble. Considering the responses of the  $Participant_k$  to the EUBE, BDI, BAI, and WHOQOL instruments, the answers are considered input for the five classifiers (four disturbances and the control condition). Data comes in the form of test answers on a Likert-type scale, with 83 columns representing each question and the raw classification from a psychiatrist. Here, a previous normalization process could be needed to ease the workload of the model. Then, each classifier performs a prediction related to  $Participant_k$ : presence or absence of

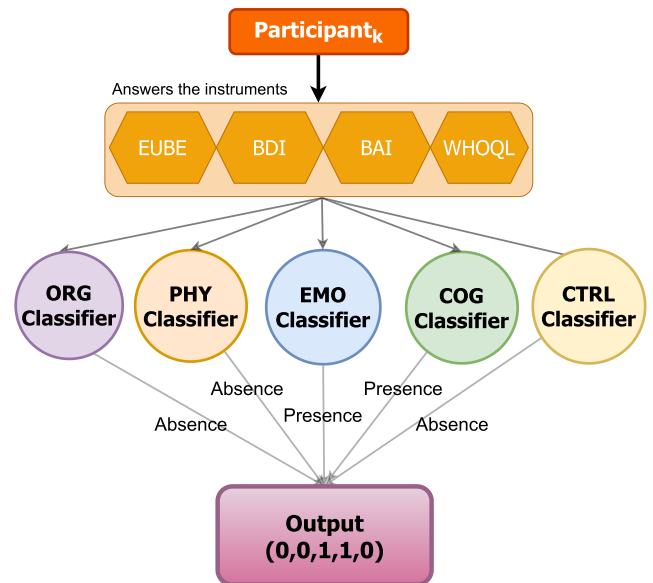


FIGURE 1. Ensemble Algorithm for predictions.

the disturbance. In the case of the control condition, the prediction is related to if the patient is healthy or if it is not. For example, suppose that this  $Participant_k$  has an emotional and a cognitive disturbance. The output shows a zero value in the absence of a disturbance and one in its presence. Then, the mobile application will assign the defined routine for each detected disturbance, guiding the users to accomplish all the assigned tasks.

The *Human Place* team defines a boundary condition as when a person has three or more disturbances. In this situation, the psychologist of *Human Place* suggests visiting a professional (psychologist or psychiatrist) to guide persons to manage the specific situation and prevent further consequences properly. As a consequence, the participants of this study can have zero, one, or at most, two disturbances.

#### A. CLASSIFIERS

The proposed ensemble will consider five classifiers. We here evaluate using two different algorithms for each class: a Random Forest (RF) or a Neural Network (NN) algorithm. The objective is to select the most suitable techniques for detecting each disturbance, considering the characteristics of the data and the number of participants of the study.

##### 1) RF

Random Forest is a Machine Learning technique combining multiple Decision Trees' outputs to reach a single result [35]. RF is an extension of the *Bagging Method*, considering a forest of Decision Trees [36]. The forest considers the output of each decision tree to perform a prediction and performs the case classification using majority voting criteria. In each decision tree, to perform a split in a particular node, the algorithm can use different functions such as *Gini*, *Entropy*, *Log Loss*, among others. RF considers a set of features

when looking for the best split (*max\_features*) and can control each decision tree's growing process with a defined maximum depth (*max\_depth*). We perform a hyper-parameter tuning process to define the number of trees in the forest (*n\_estimators*), the most suitable values for the minimum number of samples required to split an internal node (*min\_s*), the split function used (*criterion*), the *max\_depth*, and *max\_features*.

2) NN

A *multi-layer perceptron* is a NN that has input and output layers, with one or more hidden layers with many connected neurons. The neurons are combined and perform the classification process, considering a particularly defined set of weights to linearly combine the input information. The objective is to optimize the performance of the NN, minimizing a defined *loss function* that compares the target and predicted output values. The value of the output of each node/layer is determined by considering an activation function. In this study, we used a feed-forward multi-layer perceptron and the back-propagation algorithm to adjust the weights in the network. Considering the nature of the problem being tackled, we choose the Binary Cross-Entropy as the loss function, and each model layer has a *sigmoid* activation function. Moreover, each classifier considers a defined number of *n\_layer* layers, where each layer has *n\_neu* neurons. Also, each NN classifier considers an extra final layer with two neurons with the rectified linear unit activation function (*relu*) for the output. We perform a hyper-parameter tuning process to determine the values of the number of layers, the number of neurons per layer, the learning rate, and the batch and epoch size.

**B. MODEL GENERATION PROCESS**

To generate the RF or NN models, we followed the procedure described in algorithm 1. As we are working in a supervised learning scenario, this process considers the usage of a real data set labeled *rds* and uses a synthetic dataset *sds* (that will be explained in the following sections). Also, we required values for the hyperparameters *H* of the algorithms, and the *Type* of approach will be generated (RF or NN). First, the used data is prepared, managing missing data or normalization process if required (line 1). Then, for each *Class* in the *rds* (ORG, PHY, EMO, COG, and CTRL), a *Model<sub>i</sub>* will be created (line 3). The training and testing processes are performed with *sds* and *rds*, respectively (lines 4-5). At the end of this procedure, a set of models *M* will be created, one per *Class* in *rds*.

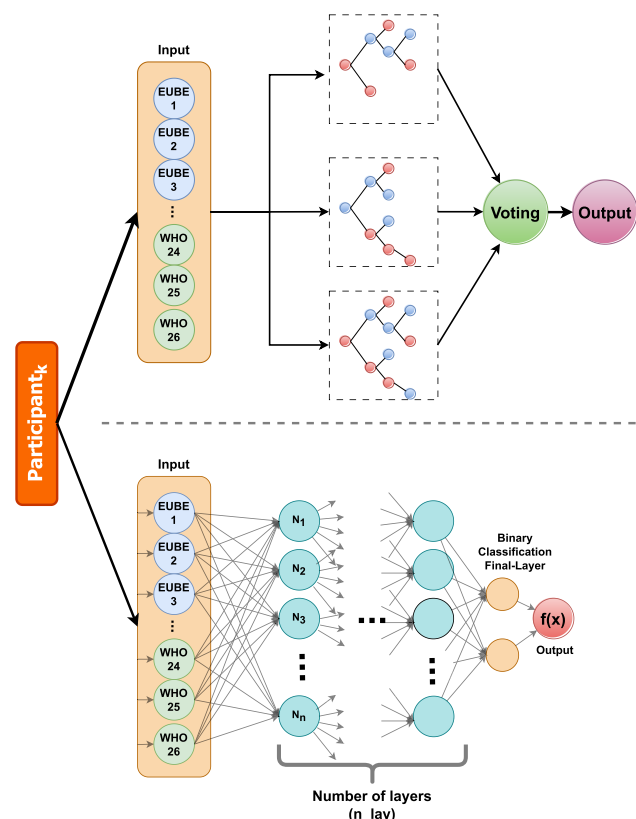
**C. MODEL PREDICTION PROCESS**

We evaluated the performance of one NN and one RF model for each classification task. Figure 2 shows the prediction process. Considering the answers *Participant<sub>k</sub>* to the questionnaire as input, one model will predict the presence or absence of each disturbance. For the final design

**Algorithm 1** Model generation process

```

Input: rds (real dataset), sds (synthetic dataset), H (hyper-parameter values), Type (RF or NN)
Output: M (set of models)
Prepare(rds, sds)
for each Class ∈ rds do
    Modeli ← CreateModel(Type)
    Train(Modeli, sds)
    TestAndEvaluate(Modeli, rds)
    M = M ∪ Modeli
end for
    
```



**FIGURE 2.** Evaluation of NN (up) and RF (down) models for each classification task.

of our ensemble, we will choose the models that optimize a set of metrics detailed in the following section.

Algorithm 2 shows the structure of the prediction procedure. The input dataset (*ds*) contains a set of cases that will be classified by the set of models *M*. To manage missing data and normalization tasks, the *ds* is prepared to be used by the models in line 2. Then, for each *case<sub>k</sub>*, each *Model<sub>i</sub>* will predict the presence or absence of its particular class. This model can be RF or NN, and its prediction is stored in *p<sub>k</sub>* (line 5). At the end, a set of predictions *P* is generated, with the response of all models for all the cases.

We generate synthetic datasets based on real data. This procedure will be explained in the following section. Then,

**Algorithm 2** Model prediction procedure

---

**Input:**  $ds$  (input dataset),  $M$  (set of models)  
**Output:**  $\mathcal{P}$  (prediction for all cases from all classifiers)

```

 $\mathcal{P} \leftarrow \emptyset$ 
Prepare( $ds$ )
for each  $case_k \in ds$  do
  for  $Model_i \in M$  do
     $p_k \leftarrow Predict(case_k, Model_i)$ 
     $\mathcal{P}_k = \mathcal{P}_k \cup p_k$ 
  end for
end for

```

---

we present the hyper-parameter tuning procedure for each RF and NN algorithm.

**V. EXPERIMENTAL SETUP**

The participants of this study were 93 persons. Considering the questionnaire presented in Section III, the psychologist in Human Place classified this data. From the total of 93 participants, thirty-three persons were classified with a physical (PHY) disturbance, twenty-eight persons with a cognitive (COG) disturbance, nine were classified with an emotional (EMO) disturbance, ten with an organizational (ORG), two persons required the attention of a specialist (three or more disturbances), nine persons have two simultaneous disturbances, and seventeen persons do not present symptoms of any disturbances (CTRL). We present a demographic analysis of the participants, considering the four features described in section III (gender, age, occupancy, and the city of the participants). Figure 3 shows histograms of the distribution of the classes in the real data, considering the mentioned features. In the four plots, the five classes and the multi-label scenarios are particularly colored. In general, it can be observed that the real data contains fewer multi-label cases compared to the single-label ones. Analyzing the gender, as we detailed in section III, only one-third of the participants are male. The classes with the highest number of cases among the female gender are the PHY (35%), COG (26%), and CTRL (15%) classes. These classes are also the most important ones for the male gender. However, these three classes have almost the same distribution among them. About the occupancy, the distribution of the classes is similar for the employees as the female participants. The class with the highest presence in the employee and student participants is the PHY class (5 out of 12 cases). On the other hand, the COG class has the highest presence among the student participants (37%), followed by the CTRL class (27%). About the age of the participants, the PHY, CTRL, and COG classes are important in the segments of [20, 29] and [30, 39] years. It can be observed that only a few participants are less than 20 and more than 40 years old. However, the PHY and CTRL classes are the most relevant through these segments. Analyzing the city where the participants live, it can be observed that all cities have participants classified with the PHY class, with a higher number of cases in Valparaíso and

**Algorithm 3** Synthetic dataset generation

---

**Input:**  $Q'$ ,  $Q$ ,  $rds$  (real dataset),  $N$  (Number of entries to be created per category),  $K_X$  (Number of questions to be modified per entry)  
**Output:**  $sds$  (synthetic dataset)

```

1:  $Q^* = Q - Q'$ ;
2:  $sds \leftarrow \emptyset$ 
3: for each  $Class \in rds$  do
4:   for  $i = 1$  to  $N$  entries do
5:      $entry \leftarrow SelectRandomEntryFrom(rds)$ 
6:      $SynthEntry \leftarrow ModifyEntry(K_X, entry, Q^*)$ 
7:      $sds = sds \cup SynthEntry$ 
8:   end for
9: end for

```

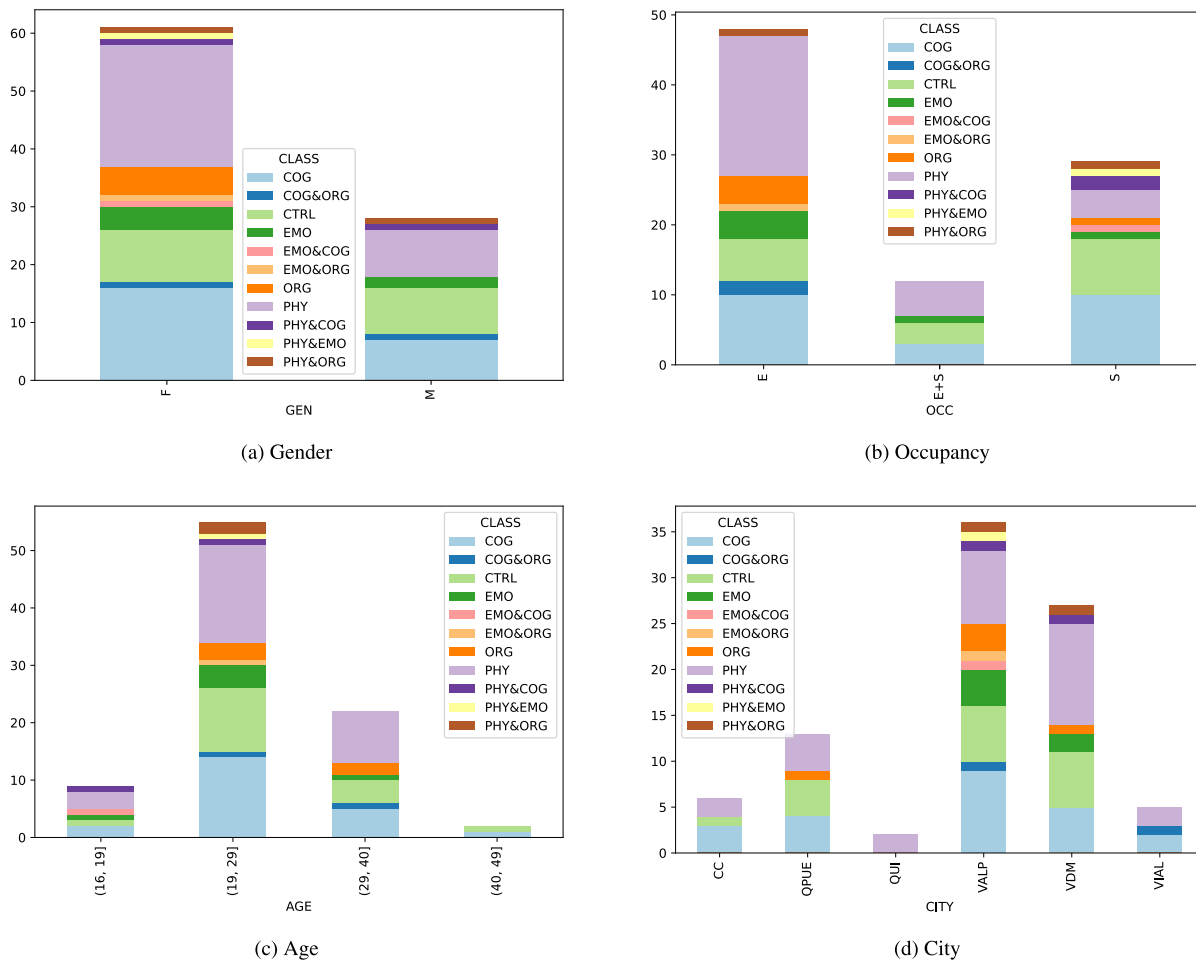
---

Viña del Mar. Moreover, Quillota city only has participants classified as PHY. Regarding the EMO disturbance and the multi-label cases of PHY&COG and PHY&ORG, all the existing cases are placed in Valparaíso and Viña del Mar cities. CTRL individuals are placed in Concón, Quilpué, Viña del Mar, and Valparaíso. Most of the participants with the ORG disturbance live in Valparaíso, Viña del Mar, and Quilpué. About the multi-labeled cases of EMO&ORG, EMO&COG, and EMO&PHY, all these participants live in Valparaíso.

**A. SYNTHETIC DATASET GENERATION**

We generate synthetic datasets for this study. The main reason is related to the number of participants of the study and the number of classes of the multilabel classification problem. We perform a data augmentation and a data balancing procedure, considering some guidelines detailed by the psychologist of Human Place. First, from the set  $Q$  that contains the 83 questions, the psychologist defines a subset of questions  $Q'$  that their answers are critical in the classification process of each class. Moreover, to avoid the possibility of modifying the classification provided by the psychologist, the participant's answers to the questions in  $Q'$  should not be modified. Consequently, all questions in  $Q^* = Q - Q'$  can be modified (reducing the value of each answer). Here, we produce changes in the answers to questions in  $Q^*$ , ensuring that every class has at least 20% of the total cases. We generate two sets of datasets: one set named  $\Pi_{Tun}$  for tuning the RF and NN algorithms' hyper-parameters and one set named  $\Pi_{Test}$  for testing the algorithms. For  $\Pi_{Tun}$ , we generate ten datasets considering 300 cases per datasets. Here, we randomly change  $K_{Tun} \in [15, 60]$  questions from  $Q^*$ , with an increment step of  $K_{Tun}$  of five. Algorithm 3 represents the procedure to generate one synthetic dataset ( $sds$ ). The generated  $sds$  includes  $N$  entries per class (line 5). For each randomly selected entry, it modifies  $K_X$  ( $X$  can be  $Tun$  or  $Test$ ) randomly selected questions from  $Q^*$  set (line 6). For  $\Pi_{Test}$ , we generate 26 datasets, each one with 1000 cases. Here, we modified  $K_{Test}$  randomly selected questions





**FIGURE 3.** Distribution of the classes through the gender (F: female, M: male), the occupancy (S: student, E: employee, and S+E: student and employee), the age, and the city (CC: Concón, QPUE: Quilpué, QUI: Quillota, VALP: Valparaíso, VDM: Viña del Mar, VIAL: Villa Alemana) of the participants.

from  $Q^*$ , with  $K_{Test} \in [10, 60]$  with an increment step of  $K_{Test}$  of two.

**B. HYPER-PARAMETER TUNING**

We use the well-known tuner *irace* for choosing the hyper-parameter values for the RF and the NN algorithms [37]. As each classifier is focused on a particular disturbance, we tuned the hyper-parameters of each classifier independently, considering 5,000 evaluations per algorithm. We consider the datasets in  $\Pi_{Tun}$  for the training process and the real data for the validation process of each algorithm.

1) TUNING NN

During the tuning process of the NN algorithms, we optimize the accuracy (ACC) of the classification:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

considering the number of True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN) in a classical confusion matrix. The hyper-parameters

**TABLE 2.** Tuned hyper-parameters for the NN algorithm, their types, and domains.

Parameter	Type	Domain
Number of layers (n_lay)	Integer	[1, 6]
Number of neurons (n_neu)	Integer	[2, 30]
Epochs (n_ep)	Integer	[1, 30]
Batch size (batch_s)	Integer	[5, 10]
Learning rate (l_rate)	Categorical	$(10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 3 \times 10^{-3}, 5 \times 10^{-3}, 10^{-2})$

tuned, their domains, and data types are shown in table 2. We follow the guidelines presented in [38] and [39] to determine the domains of the number of nodes and layers. Considering that the real data contains only 93 cases, we tried to avoid possible overfitting with small values for the batch, learning rate values, and epoch values. However, considering the capabilities of *irace*, we expand the size of each domain to allow the tuner to evaluate and further explore different parameter configurations. The obtained values by *irace* for each NN classifier are detailed in table 3. The differences

**TABLE 3. Hyper-parameter configurations obtained by irace for the NN.**

Classifier	$n_{lay}$	$n_{neu}$	$n_{ep}$	$batch_s$	$L_{rate}$
COG	2	8	10	7	$5 \cdot 10^{-3}$
EMO	1	15	21	7	$10^{-2}$
PHY	3	13	26	9	$3 \cdot 10^{-3}$
ORG	6	21	22	6	$10^{-3}$
CTRL	4	24	17	6	$10^{-4}$

**TABLE 4. Tuned hyper-parameters for the RF algorithm, their types, and domains.**

Parameter	Type	Domain
Number of estimators ( $n_{estimators}$ )	Integer	[80, 120]
Maximum features ( $max\_features$ )	Integer	[0, 120]
Minimum number of samples split ( $min\_s$ )	Integer	[2, 10]
Maximum depth ( $max\_depth$ )	Integer	[5, 13]
Criterion ( $L_{rate}$ )	Categorical	( $gini, entropy, log\_loss$ )

**TABLE 5. Hyper-parameter configurations for RF obtained by irace.**

Classifier	$n_{estimators}$	$max\_features$	$min\_n$	$max\_depth$	$criterion$
COG	116	28	6	10	$log\_loss$
EMO	100	36	2	11	$gini$
PHY	100	74	2	11	$gini$
ORG	99	71	3	12	$log\_loss$
CTRL	116	5	3	10	$log\_loss$

in the obtained hyper-parameter values for each classifier show the importance of performing these tuning processes to optimize the algorithms' performance. The differences can mainly be observed in the number of layers, number of neurons per layer, number of epochs, and the learning rate.

2) TUNING RF

We optimize the F1-score during the tuning process of the RF algorithms:

$$F1 = 2 \cdot \frac{PREC \cdot REC}{PREC + REC} \tag{2}$$

where  $PREC$  is the precision metric ( $PREC = TP/(TP+FP)$ ) and the  $REC$  is the recall metric ( $REC = TP/(TP + FN)$ ). Table 4 shows the hyper-parameters tuned for each RF classifier, their data type, and domain. The values obtained for each classifier are presented in table 5. Some similarities can be observed in the values obtained from the tuning process through the classifiers. For example, for COG and CTRL classifiers, their values in  $n_{estimators}$ ,  $criterion$ , and  $max\_depth$  are the same. A similar situation happens between the EMO and PHY classifiers, where different values were only obtained for  $max\_features$  hyper-parameter. For the five classifiers, the main differences among the obtained configurations can be observed in the  $max\_features$  values.

**VI. RESULTS AND ANALYSIS**

This section presents the evaluation of our proposal. First, we select each class's model, considering the datasets in  $\Pi_{Tun}$  and the suggested hyper-parameter values reported in section V-B. We select the best NN and RF model for each

class, considering the accuracy, the number of false negatives, and the area under the receiver-operation Curve (AUC). The AUC considers the True Positive Rate ( $TPR = TP/(TP + FN)$ ) and the False Positive Rate ( $FPR = FP/(FP + TN)$ ). Here, it is essential to analyze the number of false negative cases, considering that this project aims to prevent student burnout symptoms. As a false negative case represents a person with symptoms and is classified as healthy, we expect to minimize this metric.

*Accuracy:* Table 6 shows the accuracy obtained by each classifier for each dataset in  $\Pi_{Test}$ , and it shows the average (AVG) and the standard deviation (SD) over all the datasets. Results in bold highlight when an algorithm outperforms the other one in a particular dataset. About the COG class, both algorithms can detect the presence of this class with accuracy values higher than 90%. For this class, it can be observed that the RF algorithm obtains a higher accuracy in all the datasets compared to the NN model. However, on average, the difference is near 3%. Considering the EMO class, in the datasets with K ranging from 10 to 30, the NN model has a higher accuracy in predicting the presence of the EMO class. Then, when K is between 34 and 60, the RF algorithm has a higher accuracy value. On average, both algorithms have almost the same accuracy value (near 93%), with a slight difference that favors the NN model (with a lower standard deviation value). Analyzing the PHY class datasets, the NN obtained an average accuracy near 95%, showing the highest NN accuracy performance (compared to all the other NN models). The NN outperforms the RF algorithm in this class on seventeen datasets (with K ranging from 10 to 42). Moreover, the average accuracy of the RF is near 91%, which is the class where the RF had the lowest accuracy performance compared to all the other RF models. In the ORG and CTRL classes, the performance of RF is better than the NN in terms of accuracy, outperforming the NN in almost all datasets. About the ORG, the average accuracy is near 99%, which is the best performance of an RF algorithm in this study. Analyzing the variation of the  $K_{Test}$  value and the performance of all the classifiers, the algorithms change their performance as the number of modified questions increases. For example, the RF models improve their accuracy in the EMO and the PHY scenarios as the  $K_{Test}$  value increases. On the other hand, when the number of cases increases, the performance of the NN models slightly decreases for the EMO and ORG classes. This situation is related to the method we used to generate, train and select the models. As we used the best models for each class, these models were generated with datasets with a particular number of modified questions that can be similar to some  $K_{Test}$  values. However, only slight performance differences exist through the  $K_{Test}$  values, and all the selected models perform well.

*False Negative Rate:* This section presents an analysis of the False Negative Rate, computed as:

$$FNR = \frac{FN}{FN + TN + FP + TP} \tag{3}$$

**TABLE 6.** Accuracy obtained in each dataset by each RF and NN algorithm for each class. The name of each dataset is the number of modified questions from the real case scenario. Also, the average (AVG) and the standard deviation (SD) of the metric, considering all the datasets, is presented. The classes are COG (cognitive disturbance), EMO (emotional disturbance), ORG (organizational disturbance), PHY (physical disturbance), and CTRL (control scenario).

Dataset - $K_{Test}$ value	COG		EMO		PHY		ORG		CTRL	
	RF	NN	RF	NN	RF	NN	RF	NN	RF	NN
10	<b>0.953</b>	0.945	0.855	<b>0.982</b>	0.826	<b>0.953</b>	<b>0.966</b>	0.951	0.931	<b>0.966</b>
12	<b>0.961</b>	0.938	0.866	<b>0.978</b>	0.839	<b>0.946</b>	<b>0.974</b>	0.947	0.936	<b>0.957</b>
14	<b>0.959</b>	0.940	0.860	<b>0.975</b>	0.848	<b>0.946</b>	<b>0.971</b>	0.946	0.948	<b>0.961</b>
16	<b>0.959</b>	0.938	0.873	<b>0.969</b>	0.855	<b>0.948</b>	<b>0.982</b>	0.951	<b>0.963</b>	0.950
18	<b>0.966</b>	0.937	0.878	<b>0.968</b>	0.847	<b>0.947</b>	<b>0.986</b>	0.945	<b>0.961</b>	0.950
20	<b>0.964</b>	0.941	0.886	<b>0.952</b>	0.865	<b>0.947</b>	<b>0.983</b>	0.950	<b>0.971</b>	0.950
22	<b>0.968</b>	0.942	0.887	<b>0.958</b>	0.881	<b>0.941</b>	<b>0.986</b>	0.952	<b>0.966</b>	0.952
24	<b>0.967</b>	0.947	0.894	<b>0.960</b>	0.883	<b>0.944</b>	<b>0.986</b>	0.944	<b>0.972</b>	0.946
26	<b>0.965</b>	0.938	0.910	<b>0.957</b>	0.891	<b>0.930</b>	<b>0.984</b>	0.942	<b>0.974</b>	0.943
28	<b>0.971</b>	0.938	0.924	<b>0.950</b>	0.883	<b>0.942</b>	<b>0.987</b>	0.937	<b>0.981</b>	0.948
30	<b>0.969</b>	0.945	0.915	<b>0.946</b>	0.895	<b>0.943</b>	<b>0.990</b>	0.942	<b>0.981</b>	0.950
32	<b>0.973</b>	0.942	0.945	0.945	0.915	<b>0.943</b>	<b>0.990</b>	0.945	<b>0.982</b>	0.936
34	<b>0.977</b>	0.946	<b>0.946</b>	0.942	0.920	<b>0.948</b>	<b>0.991</b>	0.934	<b>0.986</b>	0.958
36	<b>0.972</b>	0.943	<b>0.946</b>	0.925	0.928	<b>0.942</b>	<b>0.990</b>	0.934	<b>0.980</b>	0.945
38	<b>0.972</b>	0.936	<b>0.958</b>	0.942	0.927	<b>0.942</b>	<b>0.996</b>	0.928	<b>0.983</b>	0.940
40	<b>0.974</b>	0.939	<b>0.953</b>	0.936	0.930	<b>0.948</b>	<b>0.996</b>	0.934	<b>0.990</b>	0.936
42	<b>0.975</b>	0.945	<b>0.962</b>	0.922	0.931	<b>0.946</b>	<b>0.994</b>	0.934	<b>0.987</b>	0.943
44	<b>0.969</b>	0.934	<b>0.971</b>	0.923	0.947	0.947	<b>0.992</b>	0.930	<b>0.990</b>	0.933
46	<b>0.973</b>	0.941	<b>0.977</b>	0.917	<b>0.948</b>	0.947	<b>0.992</b>	0.934	<b>0.990</b>	0.940
48	<b>0.974</b>	0.940	<b>0.980</b>	0.913	<b>0.959</b>	0.946	<b>0.990</b>	0.927	<b>0.991</b>	0.935
50	<b>0.966</b>	0.949	<b>0.985</b>	0.915	<b>0.969</b>	0.945	<b>0.991</b>	0.922	<b>0.987</b>	0.938
52	<b>0.971</b>	0.942	<b>0.989</b>	0.912	<b>0.973</b>	0.949	<b>0.999</b>	0.928	<b>0.990</b>	0.936
54	<b>0.968</b>	0.945	<b>0.988</b>	0.907	<b>0.971</b>	0.951	<b>0.994</b>	0.921	<b>0.991</b>	0.942
56	<b>0.973</b>	0.938	<b>0.991</b>	0.904	<b>0.975</b>	0.950	<b>0.990</b>	0.924	<b>0.994</b>	0.930
58	<b>0.973</b>	0.943	<b>0.991</b>	0.898	<b>0.972</b>	0.950	<b>0.994</b>	0.922	<b>0.996</b>	0.928
60	<b>0.969</b>	0.939	<b>0.991</b>	0.901	<b>0.972</b>	0.951	<b>0.991</b>	0.911	<b>0.992</b>	0.931
AVG	0.969	0.941	0.935	0.938	0.913	0.946	0.988	0.936	0.977	0.944
SD	0.006	0.004	0.047	0.026	0.048	0.005	0.008	0.011	0.017	0.010

As we mentioned, it is crucial to observe the behavior of all the models in detecting the presence of a class and reducing the number of people with symptoms classified as healthy. Observing the COG class, the RF model has the lowest FNR in all datasets but one ( $K = 10$ ), reaching an average of 0.012. However, the average FNR of the NN model is very similar to the one of RF (with the same standard deviation). A similar situation happens for the EMO class, where the RF model outperforms the NN model in 18 out of 26 datasets. Regarding the PHY class, the NN model outperforms the RF model, reaching the lowest FNR in almost all datasets and with an average FNR of 2%. As in the accuracy analysis, this class is where the RF had the worst performance with a 5% of average FNR. In the ORG class, the RF outperforms the NN model in all datasets, having its best performance with only 1% as the average FNR. As in the accuracy analysis, the worst NN performance happens in the CTRL class, with an average FNR of 5%. Observing the FNR through the different  $K_{Test}$  values, there are some cases when a relationship can be observed. The FNR for RF in the EMO class tends to increase with the  $K_{Test}$  value. An opposite situation happens in the PHY class with the RF model, where the FNR values tend to decrease with the increment of the  $K_{Test}$  value. However, in most cases, there is no observable relationship between the increment of  $K_{Test}$  and the FNR values (e.g., RF and NN for COG, NN for EMO, RF and NN for PHY, NN for CTRL).

*AUC*: Table 8 shows the AUC obtained by each classifier for the five classes in all the datasets. Generally, it can be clearly observed that the RF models outperform the NN models in most datasets in all the classes. Considering the average AUC values, the lowest one is 0.97 for the RF models and 0.88 for the NN models. The results show that the AUC values of the RF models increase and get closer to one as the  $K_{Test}$  increases for all the classes. However, the opposite behavior can be observed for the NN algorithms, decreasing the AUC values when the  $K_{Test}$  values increase. As we mentioned in the accuracy analysis, this behavior is related to the method we use to generate, train, and select the best models.

**A. RELEVANT FEATURES ANALYSIS**

In order to understand and explain the decisions performed by the classifiers, we perform an analysis to observe the most relevant features. For this, we use the Shapley Additive Explanations (SHAP) method to explain the predictions made by the NN and the RF classifiers [40]. Particularly, the SHAP method computes the contribution of each feature to the predictions made. In this section, the analysis is performed considering the  $K_{Test} = 60$  dataset. We perform this analysis considering all the other datasets, but we do not observe important differences in the set of relevant features per classifier (at least 80% were the same).

**TABLE 7.** False Negative Rate obtained by each RF and NN algorithm for each class in each dataset from the test set. The name of each dataset is the number of modified questions from the real case scenario. Also, the average (AVG) and the standard deviation (SD) of the metric, considering all the datasets, are presented. The classes are COG (cognitive disturbance), EMO (emotional disturbance), ORG (organizational disturbance), PHY (physical disturbance), and CTRL (control scenario).

Dataset - $K_{Test}$ value	COG		EMO		PHY		ORG		CTRL	
	RF	NN	RF	NN	RF	NN	RF	NN	RF	NN
10	0.030	<b>0.027</b>	<b>0.000</b>	0.009	0.122	<b>0.019</b>	<b>0.023</b>	0.047	0.069	<b>0.014</b>
12	<b>0.021</b>	0.028	<b>0.000</b>	0.010	0.105	<b>0.022</b>	<b>0.023</b>	0.049	0.064	<b>0.027</b>
14	<b>0.022</b>	0.027	<b>0.000</b>	0.011	0.100	<b>0.022</b>	<b>0.018</b>	0.044	0.052	<b>0.024</b>
16	<b>0.020</b>	0.028	<b>0.000</b>	0.006	0.094	<b>0.022</b>	<b>0.016</b>	0.041	0.037	<b>0.036</b>
18	<b>0.018</b>	0.032	<b>0.000</b>	0.007	0.105	<b>0.023</b>	<b>0.011</b>	0.043	0.039	<b>0.035</b>
20	<b>0.016</b>	0.031	<b>0.000</b>	0.009	0.080	<b>0.019</b>	<b>0.015</b>	0.038	<b>0.029</b>	0.035
22	<b>0.014</b>	0.025	<b>0.000</b>	0.006	0.074	<b>0.024</b>	<b>0.012</b>	0.031	0.034	<b>0.033</b>
24	<b>0.016</b>	0.020	<b>0.000</b>	0.004	0.074	<b>0.023</b>	<b>0.012</b>	0.035	<b>0.028</b>	0.034
26	<b>0.012</b>	0.023	<b>0.001</b>	0.007	0.072	<b>0.030</b>	<b>0.011</b>	0.039	<b>0.026</b>	0.038
28	<b>0.008</b>	0.023	<b>0.002</b>	0.007	0.066	<b>0.023</b>	<b>0.008</b>	0.039	<b>0.019</b>	0.039
30	<b>0.008</b>	0.019	<b>0.000</b>	0.008	0.071	<b>0.021</b>	<b>0.006</b>	0.032	<b>0.019</b>	0.039
32	<b>0.012</b>	0.020	<b>0.002</b>	0.006	0.053	<b>0.019</b>	<b>0.008</b>	0.027	<b>0.018</b>	0.051
34	<b>0.006</b>	0.018	<b>0.002</b>	0.004	0.048	<b>0.019</b>	<b>0.008</b>	0.033	<b>0.014</b>	0.035
36	<b>0.006</b>	0.018	<b>0.001</b>	0.007	0.037	<b>0.023</b>	<b>0.006</b>	0.032	<b>0.020</b>	0.042
38	<b>0.006</b>	0.023	<b>0.004</b>	0.004	0.036	<b>0.018</b>	<b>0.004</b>	0.029	<b>0.017</b>	0.047
40	<b>0.010</b>	0.019	<b>0.001</b>	0.004	0.042	<b>0.017</b>	<b>0.003</b>	0.030	<b>0.010</b>	0.054
42	<b>0.007</b>	0.020	<b>0.004</b>	0.008	0.039	<b>0.019</b>	<b>0.006</b>	0.026	<b>0.013</b>	0.192
44	<b>0.008</b>	0.019	0.005	0.005	0.028	<b>0.016</b>	<b>0.008</b>	0.025	<b>0.009</b>	0.051
46	<b>0.007</b>	0.016	<b>0.003</b>	0.005	0.023	<b>0.015</b>	<b>0.007</b>	0.020	<b>0.010</b>	0.060
48	<b>0.005</b>	0.016	0.007	<b>0.004</b>	0.021	<b>0.015</b>	<b>0.009</b>	0.024	<b>0.009</b>	0.057
50	<b>0.010</b>	0.013	0.008	<b>0.003</b>	<b>0.013</b>	0.016	<b>0.009</b>	0.024	<b>0.012</b>	0.059
52	<b>0.005</b>	0.012	0.007	<b>0.003</b>	0.017	<b>0.013</b>	<b>0.001</b>	0.015	<b>0.010</b>	0.060
54	<b>0.013</b>	0.015	0.010	<b>0.003</b>	<b>0.010</b>	0.012	<b>0.006</b>	0.018	<b>0.009</b>	0.062
56	<b>0.006</b>	0.017	0.009	<b>0.004</b>	<b>0.011</b>	0.014	<b>0.010</b>	0.015	<b>0.006</b>	0.058
58	<b>0.007</b>	0.011	0.008	<b>0.006</b>	<b>0.012</b>	0.014	<b>0.006</b>	0.015	<b>0.004</b>	0.072
60	<b>0.009</b>	0.014	0.009	<b>0.002</b>	<b>0.011</b>	0.013	<b>0.009</b>	0.015	<b>0.008</b>	0.068
AVG	0.012	0.021	0.003	0.006	0.052	0.019	0.010	0.030	0.023	0.046
SD	0.006	0.006	0.003	0.002	0.035	0.004	0.005	0.010	0.017	0.015

1) SUMMARY PLOTS FOR RF

We here present summary plots for each class, where in the y-axis, the ten most relevant features are shown (from all the questions in  $Q$ ). The higher the position of the questions, the higher the importance of the feature in the classification process. The x-axis details the impact on the model’s output (SHAP value). Also, the color of each dot reflects the value of the feature (a high value is red, a lower value is blue) and, when a SHAP value is positive, is related to the presence of a disturbance (and vice versa). Figure 4 presents summary plots for the five classes.

It can be observed that some questions are repeated through the different classifiers. We will analyze some of the cases where a question is repeated in three or more classes. For example, the question *WHO 16* (“How satisfied are you with your sleep?”) is relevant for the COG, PHY, and CTRL classes. However, the importance, their values, and their relationship with the classification process are different through the classes. For COG, the *WHO 16* is in the first place, being the most crucial question in the classification process of this algorithm. This question has a third place for the PHY and the CTRL classifiers. Analyzing the SHAP values, it can be observed that this question, for the COG class, the highest values and their distribution are related to the presence of cognitive disturbance. In the case of the PHY

class, this question is relevant for detecting the absence of this disturbance. For the CTRL class, the distribution of the values for this question favors the presence of any disturbance. However, the highest values are focused on classifying a person as healthy.

The *BAI 6* question is also relevant for ORG, PHY, and CTRL classifiers (if the participant feels “*Dizzy or lightheaded*”). For the ORG class, this question is the most relevant in the classification process, with their highest values focused on predicting the absence of the organization disturbance. In the case of the PHY class, this question has the tenth place, having a distribution of the values close to the zero value but with the highest ones favoring the detection of the presence of the this disturbance. About the CTRL class, this question has the fourth place of importance, having its highest values focused on classifying a person with the presence of any disturbance.

Another essential question is *EUBE 3* (“*I feel sleepy during classes*”), which is relevant for the PHY, the COG, and the EMO algorithms. In the case of the PHY model, this question has the fourth place, showing their highest values in detecting the absence of physical disturbance. However, most of the SHAP values are close to zero. For the COG class, most values are close to zero, but the highest ones favor the absence of cognitive disturbance classification. About the

**TABLE 8.** AUC obtained by each RF and NN algorithm for each class in each dataset from the test set. The name of each dataset is the number of modified questions from the real case scenario. Also, the average (AVG) and the standard deviation (SD) of the metric, considering all the datasets, are presented. The classes are COG (cognitive disturbance), EMO (emotional disturbance), ORG (organizational disturbance), PHY (physical disturbance), and CTRL (control scenario).

Dataset - $K_{Test}$ value	COG		EMO		PHY		ORG		CTRL	
	RF	NN	RF	NN	RF	NN	RF	NN	RF	NN
10	<b>0.991</b>	0.931	0.948	<b>0.975</b>	0.903	<b>0.936</b>	<b>0.995</b>	0.965	<b>0.997</b>	0.953
12	<b>0.992</b>	0.919	0.964	<b>0.968</b>	0.910	<b>0.927</b>	<b>0.996</b>	0.959	<b>0.998</b>	0.923
14	<b>0.994</b>	0.921	0.963	0.963	0.923	<b>0.927</b>	<b>0.996</b>	0.951	<b>0.998</b>	0.931
16	<b>0.993</b>	0.919	<b>0.969</b>	0.943	0.927	<b>0.931</b>	<b>0.997</b>	0.957	<b>0.999</b>	0.901
18	<b>0.994</b>	0.922	<b>0.984</b>	0.942	<b>0.931</b>	0.930	<b>0.998</b>	0.947	<b>1.000</b>	0.903
20	<b>0.996</b>	0.928	<b>0.989</b>	0.911	<b>0.934</b>	0.925	<b>0.998</b>	0.951	<b>0.999</b>	0.903
22	<b>0.995</b>	0.923	<b>0.984</b>	0.919	<b>0.945</b>	0.920	<b>0.999</b>	0.945	<b>0.999</b>	0.908
24	<b>0.995</b>	0.926	<b>0.991</b>	0.921	<b>0.950</b>	0.924	<b>0.998</b>	0.935	<b>1.000</b>	0.903
26	<b>0.995</b>	0.913	<b>0.994</b>	0.919	<b>0.964</b>	0.907	<b>0.999</b>	0.936	<b>1.000</b>	0.893
28	<b>0.995</b>	0.913	<b>0.994</b>	0.904	<b>0.953</b>	0.921	<b>0.999</b>	0.926	<b>1.000</b>	0.894
30	<b>0.996</b>	0.921	<b>0.996</b>	0.897	<b>0.967</b>	0.920	<b>0.999</b>	0.927	<b>1.000</b>	0.896
32	<b>0.996</b>	0.917	<b>0.997</b>	0.891	<b>0.972</b>	0.918	<b>0.999</b>	0.926	<b>1.000</b>	0.864
34	<b>0.998</b>	0.922	<b>0.998</b>	0.882	<b>0.976</b>	0.927	<b>0.999</b>	0.912	<b>1.000</b>	0.908
36	<b>0.997</b>	0.917	<b>0.997</b>	0.850	<b>0.977</b>	0.921	<b>0.999</b>	0.910	<b>1.000</b>	0.887
38	<b>0.996</b>	0.910	<b>0.997</b>	0.882	<b>0.977</b>	0.915	<b>1.000</b>	0.894	<b>1.000</b>	0.874
40	<b>0.996</b>	0.911	<b>0.998</b>	0.869	<b>0.985</b>	0.925	<b>1.000</b>	0.908	<b>1.000</b>	0.859
42	<b>0.997</b>	0.923	<b>0.996</b>	0.845	<b>0.983</b>	0.924	<b>1.000</b>	0.902	<b>1.000</b>	0.869
44	<b>0.996</b>	0.901	<b>0.998</b>	0.843	<b>0.985</b>	0.922	<b>1.000</b>	0.893	<b>1.000</b>	0.846
46	<b>0.997</b>	0.911	<b>0.999</b>	0.830	<b>0.991</b>	0.921	<b>1.000</b>	0.894	<b>1.000</b>	0.856
48	<b>0.997</b>	0.909	<b>0.998</b>	0.820	<b>0.990</b>	0.919	<b>1.000</b>	0.886	<b>1.000</b>	0.849
50	<b>0.997</b>	0.922	<b>0.997</b>	0.823	<b>0.993</b>	0.918	<b>1.000</b>	0.876	<b>1.000</b>	0.849
52	<b>0.997</b>	0.908	<b>0.999</b>	0.816	<b>0.995</b>	0.922	<b>1.000</b>	0.876	<b>1.000</b>	0.844
54	<b>0.997</b>	0.917	<b>0.998</b>	0.806	<b>0.992</b>	0.925	<b>1.000</b>	0.866	<b>1.000</b>	0.855
56	<b>0.997</b>	0.907	<b>0.997</b>	0.801	<b>0.994</b>	0.925	<b>1.000</b>	0.868	<b>1.000</b>	0.827
58	<b>0.997</b>	0.909	<b>0.998</b>	0.791	<b>0.995</b>	0.925	<b>1.000</b>	0.864	<b>1.000</b>	0.820
60	<b>0.997</b>	0.905	<b>0.997</b>	0.791	<b>0.996</b>	0.926	<b>1.000</b>	0.841	<b>1.000</b>	0.829
AVG	0.996	0.916	0.990	0.877	0.966	0.923	0.999	0.912	1.000	0.879
SD	0.002	0.008	0.014	0.057	0.029	0.006	0.001	0.035	0.001	0.034

EMO model, the *EUBE 3* question has the tenth place, with all the values close to zero but with the highest ones favoring the presence of this disturbance.

## 2) SUMMARY PLOTS FOR NN

Figure 5 shows the summary plot for the NN algorithms. We here analyze some cases where questions are repeated at least three times through all the models. The question *BDI 5* was the most repeated one, considered in the relevant features for the PHY, ORG, EMO, and CTRL classes.<sup>2</sup> In the case of the physical disturbance, this question had the eighth place, with their highest values favoring the presence's detection. This question has the same place of importance for the ORG class, but its higher values favor the absence of organization disturbance as opposed to the PHY class. About the EMO class, this question is the second most important, and their highest values favor the absence of emotional disturbance, and one crucial portion is near the zero value. In the CTRL class, this question has a similar behavior than for the EMO class, but it has the tenth place of importance.

Another relevant question is *BAI 14* (if the participants feel "Fear of losing control"), which is essential for the EMO,

<sup>2</sup>In this question, participants should select one of the following statements: "I don't feel particularly guilty", "I feel bad or unworthy a good part of the time", "I feel quite guilty", "I feel bad or unworthy practically all the time now", or "I feel as though I am very bad or worthless".

PHY, and CTRL classes. In the three scenarios, the highest SHAP value was related to the absence of each particular class. However, the ranking of importance was different in each model (EMO fifth place, PHY third place, CTRL second place).

The question *BAI 16* was also relevant for the NN models (if participants feel "Fear of dying"). Particularly, for the EMO, PHY and COG models. For EMO and PHY models, this question was relevant to detect the absence of their particular classes. However, it was the third more relevant for the EMO class and the sixth one for the PHY class. About the COG model, the *BAI 16* question was the sixth more relevant one, with its highest values for detecting the presence of the cognitive disturbance.

## B. DEMOGRAPHIC ANALYSIS

This section presents an analysis of the performance of the algorithms in sub-datasets, particularly generated considering the gender, age and occupancy of the participants. For this analysis, we use the same dataset considered in the relevant features analysis section ( $K_{Test} = 60$ ). The objective is to disaggregate the performance analysis and to understand possible scenarios where a classifier can be better for specific tasks.

Table 9 presents the accuracy of the RF and NN classifiers for the sub-datasets of gender, age, and occupancy. The results highlighted in bold represent when a classifier

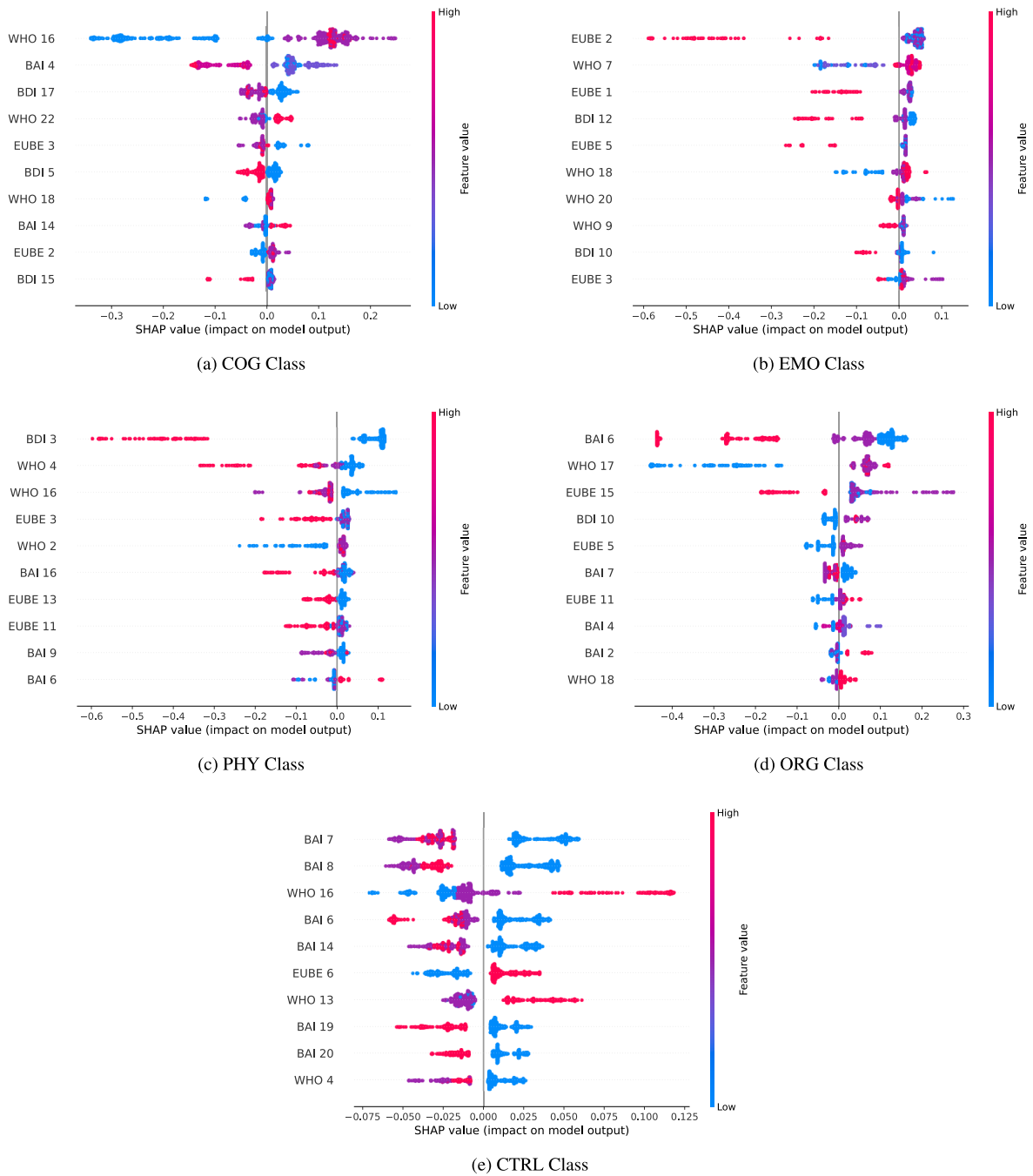


FIGURE 4. SHAP summary plots for the RF models.

outperforms the other in a particular sub-dataset. It can be observed that the RF classifiers outperform the NN ones in most cases. However, for the COG class, the NN classifier obtained a better accuracy in gender (both female and male) and employee participants. Moreover, the performance is almost the same in some scenarios between the classifiers. For example, some ties can be observed in the occupancy (S+E) and in the 16-19 (ORG) and 40-49 (EMO, PHY, ORG) age ranges.

Considering that this project will be working particularly with university students, we can observe that, in the occupancy-S sub-dataset, the RF outperforms the NN classifiers for all the classes. Also, a similar situation can be observed for the 16-19 and 20-29 age ranges.

VII. DISCUSSION

This section presents a discussion of the results, an analysis of the participants' obtained scores in the used instruments,

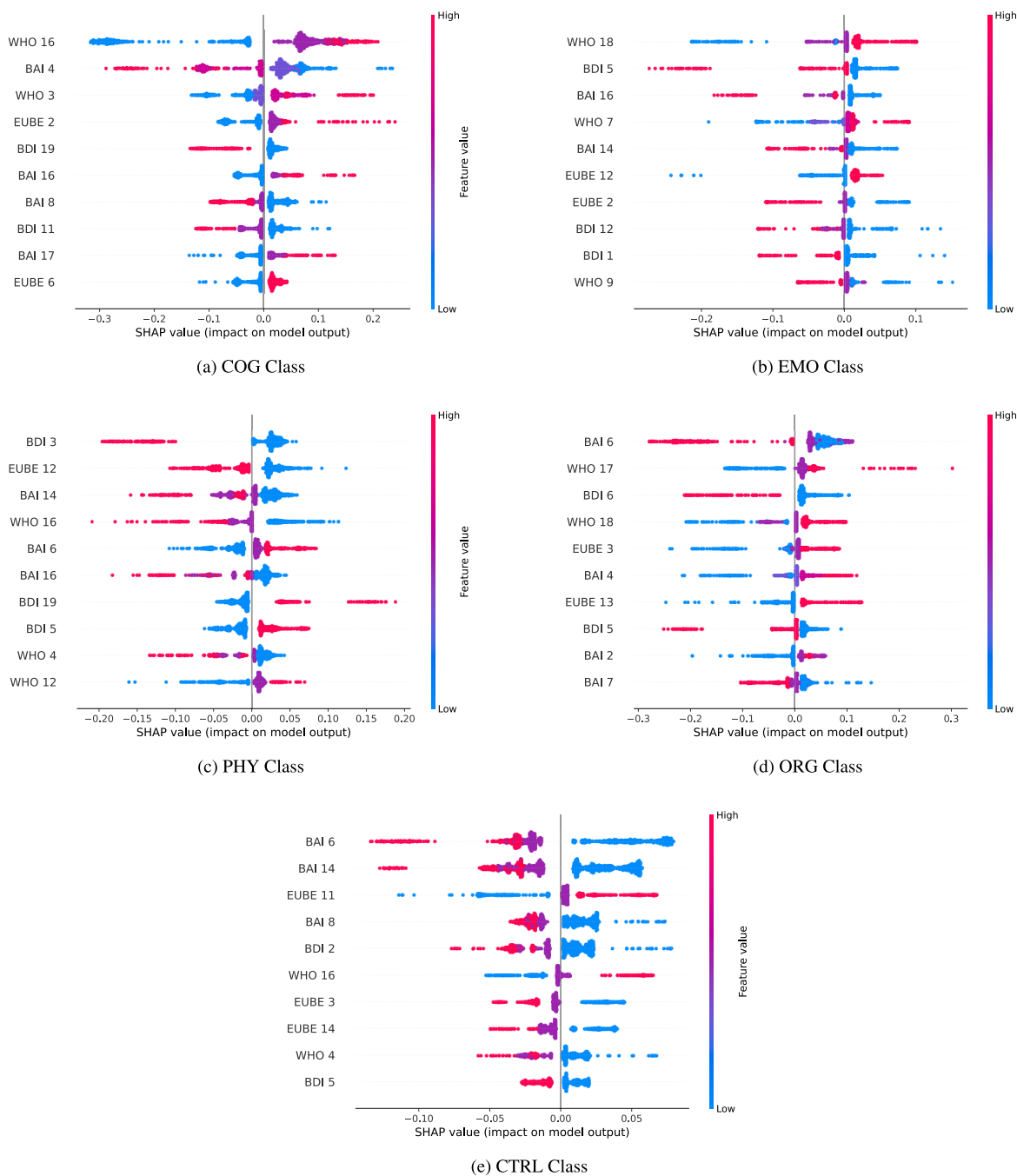


FIGURE 5. SHAP summary plots for the NN models.

TABLE 9. Accuracy of each classifier in different groups of participants considering their demographic features: gender (F: female, M: male), age, and occupancy (S: student, E: employee, S+E: student and employee).

Classifier	Gender				Age								Occupancy					
	F		M		16-19		20-29		30-39		40-49		S		E		S+E	
	RF	NN	RF	NN	RF	NN	RF	NN	RF	NN	RF	NN	RF	NN	RF	NN	RF	NN
COG	0.970	<b>0.942</b>	0.930	<b>0.953</b>	<b>0.906</b>	0.854	<b>0.970</b>	0.958	0.942	0.942	<b>1.000</b>	0.875	<b>0.929</b>	0.894	0.967	<b>0.969</b>	1.000	1.000
EMO	<b>0.991</b>	0.878	0.990	<b>1.000</b>	<b>1.000</b>	0.991	<b>0.995</b>	0.889	<b>0.971</b>	0.959	1.000	1.000	<b>0.991</b>	0.917	<b>0.989</b>	0.905	1.000	1.000
PHY	<b>0.972</b>	0.957	<b>0.963</b>	0.893	0.991	<b>1.000</b>	<b>0.973</b>	0.959	<b>0.942</b>	0.879	1.000	1.000	<b>0.970</b>	0.932	<b>0.967</b>	0.933	0.981	<b>1.000</b>
ORG	<b>0.997</b>	0.952	<b>0.976</b>	0.940	1.000	1.000	<b>0.986</b>	0.877	<b>1.000</b>	0.951	1.000	1.000	<b>0.979</b>	0.923	<b>0.996</b>	0.899	1.000	1.000
CTRL	<b>1.000</b>	0.921	<b>0.970</b>	0.913	<b>1.000</b>	0.974	<b>0.986</b>	0.933	<b>1.000</b>	0.947	<b>1.000</b>	0.945	<b>0.973</b>	0.873	<b>1.000</b>	0.916	<b>1.000</b>	0.918

**TABLE 10.** Most relevant features in common between RF and NN models.

Class	Number of coincidences	Questions
COG	4	BAI 4, BAI 17, EUBE 2, WHO 16
EMO	5	BDI 12, EUBE 2, EUBE 9, WHO 7, WHO 9, WHO 18
PHY	5	BAI 6, BAI 16, BDI 3, WHO 4, WHO 16
ORG	6	BAI 2, BAI 4, BAI 6, BAI 7, WHO 17, WHO 18
CTRL	5	BAI 6, BAI 8, BAI 14, WHO 4, WHO 16

and a comparison with the literature. Analyzing the results, the Random Forest and Neural Network models were available to correctly classify participants in the presence or absence of cognitive, emotional, physical, and organizational disturbances and the control case participants. Moreover, both algorithms have successful accuracy, false negative rate, and AUC metric values. On average, considering all the classes, the RF has 96% accuracy, 2% of false negative rate, and 0.99 of AUC. Regarding the NN, the average accuracy is 94%, 2.4% of the average false negative rate, and an average AUC of 0.90. Considering the design of the ensemble and the obtained results, we chose to use the Random Forest algorithm for the cognitive disturbance, the organizational disturbance, and the control condition. For the physical disturbance, we chose to use the Neural Networks model. Regarding emotional disturbance, any of the two algorithms can be suitable, considering that the results of both algorithms are similar.

We analyzed the importance of the features in the classification process of the RF and the NN models. We considered the set of ten most relevant questions per classifier. The five most relevant questions, considering all the used models are: BAI 6 and WHO 16 with six occurrences, and EUBE 3, BDI 5, and BAI 14 with five occurrences. Table 10 shows the number of coincidences between the most relevant features of the RF and the NN models through the different classes. The idea is to analyze and compare how the RF and the NN models performed the decision classification process. The RF and NN models with the highest number of questions in common occur for the ORG class, considering the same six questions. For the CTRL, PHY, and EMO classes, five-question matches occur per scenario, and in the COG class, four matches occur. Regarding the ten models and their ten most relevant features, 31 of the 100 features are from the BAI instrument, 26 from the WHOQOL, 23 from the EUBE, and 18 from the BDI.

Section V-A mentions that the psychologist defines a set of  $Q'$  questions that answers should be fixed to maintain the assigned class to each case. In order to analyze the ability of the evaluated algorithms to observe new and different patterns present in the data, we compare the most relevant features with the set  $Q'$  of questions. Table 11 shows the number of questions that match between the set  $Q'$  and the ten most relevant features per algorithm per class. It can be

**TABLE 11.** Number of questions matches between the questions remaining fixed provided by the psychologist and the ten more relevant features per classifier.

Class	Number of Questions Matches	
	RF	NN
COG	1	1
EMO	1	4
PHY	1	2
ORG	3	4

observed that, in most cases, the RF models have the lowest number of matches. Considering the current scenario, with a lack of enough real data and the described procedure to generate synthetic datasets, these results show the ability of RF to detect new and different patterns, further than detecting the ones that remained fixed. On the other hand, the NN algorithms consider relevant four out of the six questions from  $Q'$  for the EMO class. This behavior shows that the NN should be considered for future work, as it is a powerful tool for detecting this kind of relationships.

**A. ANALYSIS OF THE INSTRUMENTS SCORE**

This section presents an analysis of the scores obtained by the participants of this study in the four applied instruments: BAI, BDI, EUBE, and WHOQOL-Bref. Table 12 presents the mean, standard deviation, median, and interquartile range (IQR) considering all the participants and separated by gender (female and male). The WHOQOL-Bref scores were transformed according to the guidelines in [41] to be compared to WHOQOL scores (0-100 points). Considering the general mean scores, the participants can be classified with the presence of *moderate anxiety* by the BAI instrument and with the presence of *mild depression* by the BDI instrument. On the other hand, if the standard deviation values are considered, some participants are classified with *severe anxiety* (BAI) or *moderate depression* (BDI). Analyzing the statistics separated by gender, female participants can be classified with *moderate anxiety* and male participants with *mild anxiety*. Depression levels do not change considering the gender of participants. On the other hand, considering the results of the EUBE instrument, participants do not have burnout syndrome, and the statistics of the scores of female and male participants are almost the same. However, a slight presence of burnout can be observed in the cases with the highest EUBE score values. Regarding the WHOQOL scores, the higher the value, the better the quality of the participant’s life (with a maximum score of 100). In the case of our study, the mean values of the four domains are in the middle of the possible range (near 50 points). Here, the physical domain has the highest score compared to the other three domains. Also, considering the standard deviation, participants reached the lowest score values, nearly 35 points in the social domain. Considering the scores separated by gender, statistic values are similar to the general scenario. However, the main differences can be observed in IQR values. For example, in the psychological domain, the first quartile



**TABLE 12.** Statistics of the obtained scores by participants: mean, standard deviation (SD), median, and interquartile range (IQR).

Instrument	General				Female				Male				
	Mean	SD	Median	IQR	Mean	SD	Median	IQR	Mean	SD	Median	IQR	
BAI	16.48	10.31	15.0	[7,23]	17.13	10.75	16.0	[7,23]	15.10	9.31	14.0	[7,23]	
BDI	11.90	8.16	9.0	[5,17]	12.02	8.73	9.0	[5,19]	11.65	6.94	9.0	[5,16]	
EUBE	13.77	5.12	14.0	[10,18]	13.75	5.17	13.5	[10,18]	13.79	5.10	15.0	[11,18]	
WHOQOL	Physical	58.75	9.47	60.7	[53.57,64.29]	58.98	9.34	60.7	[57.14,64.28]	58.25	9.87	60.7	[50.00,64.28]
	Psychological	50.86	12.16	54.1	[39.58,62.50]	50.40	12.15	54.2	[38.54,62.50]	51.86	12.32	54.2	[50.00,62.50]
	Social	53.75	9.84	50.0	[50.00,58.33]	53.09	9.45	50.0	[50.00,58.33]	55.17	10.06	58.3	[50.00,58.33]
	Environmental	50.10	5.82	50.0	[46.87,53.13]	50.40	5.85	53.1	[46.87,53.12]	49.35	5.81	50.0	[46.87,53.13]

of female participants scored around 38 points and male participants nearly 50 points. Also, in the physical domain, the first quartile of male participants is near 50 points, and female participants' first quartile is near 57.

### B. COMPARISON WITH THE LITERATURE

This section compares our study with existing works in the literature regarding the algorithmic design and the instruments used. Also, we present a comparison of the participant's scores of our work with existing related investigations from other regions of Chile and other countries where the same instruments were used (BAI, BDI, EUBE, or WHOQOL). Considering the literature described in Section II, the proposed ensemble can be considered as a supervised learning and classification approach, as the 90% of the articles reviewed in [15]. In the literature review, RF was suitable in contexts similar to what we tackle in this article [11], [12]. However, different questionnaires were considered in these articles, such as Patient Health Questionnaire (PHQ-9) and General Anxiety Disorder (GAD-7). In our case, RF was the most suitable technique considering that it performed better in more scenarios than the NN. Regarding the questionnaires, the PHQ-9 and the GAD-7 had the advantage of containing fewer questions than all of the questionnaires we used.

In order to analyze possible tendencies among university students of different regions of Chile and other countries, we present a comparison of the scores obtained by the participants of our work with other existing investigations in the literature. It is important to mention that the considered investigations are not strictly comparable, considering the existing differences in the design of the studies (e.g., the ages or the careers of the participants, the health context of the COVID 19 pandemic, among others). However, we present a descriptive comparison considering that the instruments used are the same.

In [42], the authors investigate the prevalence of anxious and depressive symptomatology in 277 medical students of the University of Chile (Santiago City, Metropolitan Region of Chile). The results on the BDI and BAI instruments show that more than 50% of the participants presented some depressive symptomatology, and 65% presented some anxiety symptomatology. Here, the mean score for the BDI instrument was 17.01, with a standard deviation of

11.16 points. The BAI instrument's mean score was 12.2, with a standard deviation 8.59. In a study performed at the University of Concepción (Concepción City, Bio-Bio Region of Chile), a similar investigation was performed considering 632 students from different faculties [43]. The BDI and BAI instruments results show that 16.4% of the students showed anxiety syndrome, and 23.4% showed depression. In this study, the mean BDI score was 11.95 with a standard deviation value of 8.60, and for the BAI instrument, the mean score was 11.31 and 9.39 as the standard deviation. It can be observed that the participants of our work have similar depression levels in terms of the mean score for the BDI instrument compared to the investigation performed at the University of Concepción. However, the mean score of the BAI instrument of the participants of our study shows higher anxiety levels than the two mentioned investigations.

An investigation of depressive symptoms in medical students residing in high southern latitudes of Chile (Magallanes Region) is presented in [44], considering 102 students from the University of Magallanes in Punta Arenas. The students answered the WHOQOL-Bref instrument and were classified into four groups, considering the number of months that they have lived in the region: less than 18 months (G1), between 19 and 36 months (G2), more than 37 months (G3), and born in the region (G4). About the G1 group, the mean scores reached by the Magallanes students are higher than our study's scores in the four domains (more than 64 points). In the G2 group, their physical domain scores are close to the values of our study ( $59.00 \pm 13.75$ ). However, in all the other domains, they have higher mean values than the ones in this study. Considering the G3 group, this study has higher score values for the physical domain ( $53.17 \pm 19.07$ ) but has lower values on all the other domains. The participants born in the Magallanes region have higher quality of life scores in all the domains compared to the values of this study.

Considering the two-stage study made with Italian university students reviewed in Section II, we will compare the reported anxiety levels with our study. First, the BAI scores of Italian university students show that the anxiety levels between males and females differed. Most female participants had mild anxiety levels, and most male participants did

not have anxiety symptoms. About the BAI scores, female participants of the first stage had a median of 13 points (IQR of [7.00, 21.00]) and 9 points (IQR of [4.00, 16.00]) for male participants. During the second stage, female participants had a median BAI score of 11 points (IQR of [6.00, 19.00]) and males 10 (IQR of [5.00, 17.00]). In our study, female participants also have slightly higher anxiety values than male participants. Moreover, in our study, female participants are classified as having *moderate anxiety* and *mild anxiety* for male participants. Considering both genders, the medians of the BAI scores of this study are higher compared to the medians of both stages in [14]. Also, regarding the distribution of BAI scores, this study's IQR values of both genders are similar to those of female participants of the first stage.

In Section VI, we observe the importance of some questions from the EUBE instrument, considering the most relevant ones for each evaluated model. In these analyses, the following questions were the most repeated among all the models: EUBE 3 (repeated five times), EUBE 2 (repeated four times), EUBE 11 (repeated three times), EUBE 5, and EUBE 13 (both repeated two times). We observe that some of these questions are also considered relevant or have a high percentage of presence in studies performed in other countries, such as Mexico, Spain, and Cuba, among others. For example, a study was performed in the University Autonoma medical faculty (in Sinaloa, Mexico) to detect burnout in medicine students [45]. The study considered 843 students from the first to the fifth year of medicine. The results show that 85.9% of the students had a slight burnout syndrome. Moreover, questions EUBE 3, EUBE 7, and EUBE 4 were the three most influential in the study, followed by EUBE 1, 2, 5, and 14 (with the same average score). A similar subset of relevant questions was found in [46], where Ph.D. students from the Anglo-Español Institute and the Pedagogic University (Durango, Mexico). The results show a slight presence of burnout syndrome in the participants of these Universities. Also, the percentage of presence of the most relevant questions of the EUBE instrument were EUBE 7 (44.7%), EUBE 3 (43.5%), and EUBE 2, 5, and 9 had a 40.7% of presence. An investigation to detect burnout syndrome considering 1146 students of the University of Granada is presented in [47]. Here, the five most relevant questions, in decreasing order are: EUBE 7, EUBE 1, EUBE 4, EUBE 3 and EUBE 2.

## VIII. CONCLUSION

This work presents a study for predicting university students' cognitive, physical, emotional, and organizational disturbances. We use four well-known instruments to detect the presence or absence of stress, anxiety, and depression symptoms: *EUBE*, *BAI*, *BDI*, and *WHOQOL*. The objective is to use Machine Learning techniques to automatize the classification process of students to assign *wellness strategies* that can prevent further consequences in their mental states. We propose the usage of a classifier ensemble to solve a

multi-label classification problem that considers these four not mutually exclusive classes.

We evaluate the usage of a Neural Network or a Random Forest algorithm for each binary classification task. Also, we consider a control condition where participants do not have any disturbances. We perform a procedure to generate synthetic datasets from a set of cases provided by the Human Place team. Results show that both the Neural Network and the Random Forest models properly classify the participants, with good accuracy and ROC AUC levels, and reduce the number of false negative cases. We present a feature analysis that reveals how the evaluated models decide on the classification of participants. Here, we observe the ability of the evaluated classifiers to obtain patterns and relationships considering questions that were not from the set of key questions. In conclusion, the Random Forest models were the most suitable for detecting cognitive and organizational disturbances and the control condition. Both evaluated models can be used regarding the emotional disturbance, considering that both obtained similar performance results. Regarding the physical disturbance, the Neural Network model obtained the best performance.

The main limitation of this work is the volume of real data. As detailed in Section III, the pandemic restrictions (and all the consequences of this scenario) made fewer persons participate in this study than we expected. This situation makes difficult the possibility of generalizing the obtained results and conclusions. In the same line, the predictions can only be extrapolated carefully. To tackle this complex situation, we generate synthetic datasets carefully following the guidelines defined by the team's psychologist, reaching datasets with 300 and 1000 total cases.

We are implementing this work at a university from the Valparaíso region in Chile to prevent burnout symptoms in university students and help them manage anxiety and depression. In this new scenario, we will work with a higher volume of real data, where we are not interested in replacing mental health professionals, expecting to replace patients' attention. Our work is focused on preventing the extension of symptoms and (hopefully) their early detection.

For our project's future work, we are interested in:

- 1) Considering the inclusion (or replacement) of questionnaires,
- 2) Considering the evaluation of some of the algorithms mentioned in Section II (SVM, NBC, among others),
- 3) Implementing a voting ensemble for more complex scenarios (for example, perform a contrast between the answer of the CTRL classifier with the other models), and
- 4) Compare the behavior and conclusions of this article with a new scenario with a higher number of university students.
- 5) We are interested in working with universities from different regions of Chile and other countries, adapting the design of our proposal to the specific scenarios of each place.

## REFERENCES

- [1] H. J. Freudenberger, "Staff burn-out," *J. Social Issues*, vol. 30, no. 1, pp. 159–165, Jan. 1974.
- [2] C. Maslach and S. E. Jackson, "The measurement of experienced burnout," *J. Organizational Behav.*, vol. 2, no. 2, pp. 99–113, Apr. 1981.
- [3] R. Y. Rosales, "Revista de la asociación española de neuropsiquiatría," *Revista Asociación Española de Neuropsiquiatría*, vol. 32, no. 116, pp. 795–803, 2012.
- [4] G. B. Larrain, N. Rojas-Morales, Psy. D. O. Jeneral, and N. G. Rogel, "Towards a classifier ensemble to prevent burnout syndrome on university students," in *Proc. 41st Int. Conf. Chilean Comput. Sci. Soc. (SCCC)*, Santiago, Chile, Nov. 2022, pp. 1–8, doi: [10.1109/SCCC57464.2022.10000313](https://doi.org/10.1109/SCCC57464.2022.10000313).
- [5] J. Wehrmann and R. C. Barros, "Movie genre classification: A multi-label approach based on convolutions through time," *Appl. Soft Comput.*, vol. 61, pp. 973–982, Dec. 2017.
- [6] S. Huang, W. Peng, J. Li, and D. Lee, "Sentiment and topic analysis on social media: A multi-task multi-label classification approach," in *Proc. 5th Annu. ACM Web Sci. Conf.*, H. Davis, H. Halpin, A. Pentland, M. Bernstein, and L. Adamic, Eds., 2013, pp. 172–181.
- [7] M. Messaoud, I. Jenhani, N. Jemaa, and M. Mkaouer, "A multi-label active learning approach for mobile app user review classification," in *Proc. Int. Conf. Knowl. Sci., Eng. Manag.*, in Lecture Notes in Computer Science, C. Douligieris, D. Karagiannis, and D. Apostolou, Eds. Cham, Switzerland: Springer, 2019, pp. 805–816.
- [8] E. Deniz, H. Erbay, and M. Coşar, "Multi-label classification of e-commerce customer reviews via machine learning," *Axioms*, vol. 11, no. 9, p. 436, Aug. 2022.
- [9] M. Izadi, A. Heydarnoori, and G. Gousios, "Topic recommendation for software repositories using multi-label classification algorithms," *Empirical Softw. Eng.*, vol. 26, no. 5, pp. 1–33, Sep. 2021.
- [10] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehousing Mining*, vol. 3, no. 3, pp. 1–13, Jul. 2007, doi: [10.4018/jdwm.2007070101](https://doi.org/10.4018/jdwm.2007070101).
- [11] A. R. Subhani, W. Mumtaz, M. N. B. M. Saad, N. Kamel, and A. S. Malik, "Machine learning framework for the detection of mental stress at multiple levels," *IEEE Access*, vol. 5, pp. 13545–13556, 2017, doi: [10.1109/ACCESS.2017.2723622](https://doi.org/10.1109/ACCESS.2017.2723622).
- [12] S. Mukherjee, L. Rintamaki, J. L. Shucard, Z. Wei, L. E. Carlasare, and C. A. Sinsky, "A statistical learning approach to evaluate factors associated with post-traumatic stress symptoms in physicians: Insights from the COVID-19 pandemic," *IEEE Access*, vol. 10, pp. 114434–114454, 2022, doi: [10.1109/ACCESS.2022.3217770](https://doi.org/10.1109/ACCESS.2022.3217770).
- [13] L. J. Anbarasi, M. Jawahar, V. Ravi, S. M. Cherian, S. Shreenidhi, and H. Sharen, "Machine learning approach for anxiety and sleep disorders analysis during COVID-19 lockdown," *Health Technol.*, vol. 12, no. 4, pp. 825–838, Jul. 2022, doi: [10.1007/s12553-022-00674-7](https://doi.org/10.1007/s12553-022-00674-7).
- [14] N. Meda, S. Pardini, P. Rigobello, F. Visioli, and C. Novara, "Frequency and machine learning predictors of severe depressive symptoms and suicidal ideation among university students," *Epidemiol. Psychiatric Sci.*, vol. 32, p. e42, Jan. 2023.
- [15] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," *Psychol. Med.*, vol. 49, no. 9, pp. 1426–1448, Jul. 2019.
- [16] E. S. Mohamed, T. A. Naqishbandi, S. A. C. Bukhari, I. Rauf, V. Sawrikar, and A. Hussain, "A hybrid mental health prediction model using support vector machine, multilayer perceptron, and random forest algorithms," *Healthcare Analytics*, vol. 3, Nov. 2023, Art. no. 100185.
- [17] Y. Sánchez-Carro, A. de la Torre-Luque, I. Leal-Leturia, N. Salvat-Pujol, C. Massaneda, A. de Arriba-Arnau, M. Urretavizcaya, V. Pérez-Solà, A. Toll, A. Martínez-Ruiz, R. Ferreirós-Martínez, S. Pérez, J. Sastre, P. Álvarez, V. Soria, and P. López-García, "Importance of immunometabolic markers for the classification of patients with major depressive disorder using machine learning," *Prog. Neuro-Psychopharmacol. Biol. Psychiatry*, vol. 121, Mar. 2023, Art. no. 110674.
- [18] L. K. Xin and N. b. A. Rashid, "Prediction of depression among women using random oversampling and random forest," in *Proc. Int. Conf. Women Data Sci. Taif Univ.*, Mar. 2021, pp. 1–5.
- [19] H. Abou-Warda, N. Belal, Y. El-Sonbaty, and S. Darwish, "A random forest model for mental disorders diagnostic systems," in *Proc. Int. Conf. Adv. Intell. Syst. Inform.*, in Advances in Intelligent Systems and Computing, vol. 533, A. Hassanien, K. Shaalan, T. Gaber, A. T. Azar, and M. Tolba, Eds., 2016, pp. 670–680.
- [20] Md. H. A. Banna, T. Ghosh, Md. J. A. Nahian, M. S. Kaiser, M. Mahmud, K. A. Taher, M. S. Hossain, and K. Andersson, "A hybrid deep learning model to predict the impact of COVID-19 on mental health from social media big data," *IEEE Access*, vol. 11, pp. 77009–77022, 2023.
- [21] T. Guo, W. Zhao, M. Alrashoud, A. Tolba, S. Firmin, and F. Xia, "Multimodal educational data fusion for students' mental health detection," *IEEE Access*, vol. 10, pp. 70370–70382, 2022, doi: [10.1109/ACCESS.2022.3187502](https://doi.org/10.1109/ACCESS.2022.3187502).
- [22] P. Elosua, "Remote tests administration: Risks and recommendations," *Papeles Psicólogo*, vol. 42, no. 1, pp. 33–37, 2021.
- [23] B. Biblioteca and C. N. de Chile. (2021). *Reportes Estadísticos 2021 Región de Valparaíso*. [Online]. Available: <http://bit.ly/bcnvalp>
- [24] A. T. Beck, R. A. Steer, and M. G. Carbin, "Psychometric properties of the beck depression inventory: Twenty-five years of evaluation," *Clin. Psychol. Rev.*, vol. 8, no. 1, pp. 77–100, Jan. 1988.
- [25] P. L. Hewitt and G. R. Norton, "The beck anxiety inventory: A psychometric analysis," *Psychol. Assessment*, vol. 5, no. 4, pp. 408–412, Dec. 1993, doi: [10.1037/1040-3590.5.4.408](https://doi.org/10.1037/1040-3590.5.4.408).
- [26] A. Barraza, "Validación psicométrica de la escala unidimensional del burnout estudiantil," *Revista Intercontinental Psicología Educación*, vol. 13, no. 2, pp. 51–74, 2011.
- [27] T. Whoqol Group, "Development of the world health organization WHOQOL-BREF quality of life assessment," *Psychol. Med.*, vol. 28, no. 3, pp. 551–558, May 1998.
- [28] A. Osman, J. Hoffman, F. X. Barrios, B. A. Kopper, J. L. Breitenstein, and S. K. Hahn, "Factor structure, reliability, and validity of the beck anxiety inventory in adolescent psychiatric inpatients," *J. Clin. Psychol.*, vol. 58, no. 4, pp. 443–456, 2002.
- [29] O. T. Leyfer, J. L. Ruberg, and J. Woodruff-Borden, "Examination of the utility of the beck anxiety inventory and its factors as a screener for anxiety disorders," *J. Anxiety Disorders*, vol. 20, no. 4, pp. 444–458, Jan. 2006.
- [30] P. Richter, J. Werner, A. Heerlein, A. Kraus, and H. Sauer, "On the validity of the beck depression inventory," *Psychopathology*, vol. 31, no. 3, pp. 160–168, 1998.
- [31] J. Casanova Diez, "Burnout, inteligencia emocional y rendimiento académico: Un estudio en alumnado de medicina," *ReiDoCrea, Revista Electrónica Investigación Docencia Creativa*, vol. 5, pp. 1–6, Jan. 2016.
- [32] A. D. Domínguez-González, M. T. Velasco-Jiménez, D. M. Meneses-Ruiz, G. Guzmán Valdivia-Gómez, and M. G. Castro-Martínez, "Síndrome de burnout en aspirantes a la Carrera de medicina," *Investigación Educación Médica*, vol. 6, no. 24, pp. 242–247, Oct. 2017.
- [33] *WHOQOL-BREF: Introduction, Administration, Scoring and Generic Version of the Assessment: Field Trial Version*, World Health Organization, Geneva, Switzerland, Dec. 1996.
- [34] D. Gómez and A. Rojas, "An empirical overview of the no free lunch theorem and its effect on real-world machine learning classification," *Neural Comput.*, vol. 28, no. 1, pp. 216–228, Jan. 2016.
- [35] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [36] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
- [37] M. López-Ibáñez, J. Dubois-Lacoste, L. Pérez Cáceres, M. Birattari, and T. Stützle, "The irace package: Iterated racing for automatic algorithm configuration," *Oper. Res. Perspect.*, vol. 3, pp. 43–58, Jan. 2016.
- [38] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review," *J. Biomed. Informat.*, vol. 35, nos. 5–6, pp. 352–359, Oct. 2002.
- [39] Z. Han, S. Yu, S.-B. Lin, and D.-X. Zhou, "Depth selection for deep ReLU nets in feature extraction and generalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1853–1868, Apr. 2022.
- [40] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Long Beach, CA, USA, 2017, pp. 4765–4774.
- [41] *Programme on Mental Health: Whoqol User Manual*, World Health Organization, Geneva, Switzerland, 1998.
- [42] L. Villacura, N. Irrarrazabal, and I. Lopez, "Evaluation of depressive and anxiety symptomatology in medical students at the University of Chile," *Mental Health Prevention*, vol. 7, pp. 45–49, Jan. 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2212657016300101>

- [43] F. Cova, W. Alvial, M. Arod, A. Bonitette, M. Hernández, and C. Rodríguez, "Mental health problems in students from the university of concepción," *Terapia Psicológica*, vol. 25, pp. 105–112, Oct. 2007.
- [44] C. Alvarado-Aravena, C. Estrada-Goic, and C. Nunez-Espinosa, "Depression and seasonal sensitivity among medical students residing in high southern latitudes," *Revista Medica Chile*, vol. 149, no. 3, pp. 357–365, 2021.
- [45] J. Martínez-García, A. Canizalez-Román, and N. León-Sicairos, "Prevalence of burnout syndrome in students of a medical school," *Revista Médica Universidad Autónoma Sinaloa Revmeduas*, vol. 11, no. 1, pp. 37–47, 2022.
- [46] A. Macías, F. Muñoz, and M. Muñoz, "Síndrome de burnout en alumnos de los doctorados en educación de Durango (México)," *Enseñanza Investigación Psicología*, vol. 17, no. 2, pp. 377–386, 2012.
- [47] A. Fernandez-Castillo, M. Fernandez-Prados, and N. Roldan-Molina, "Factorial structure and internal consistency of the student burnout scale in Spanish youth population," *Revista Iberoamericana Diagnostico Evaluacion-E Avaliacao Psicologica*, vol. 4, no. 65, pp. 35–45, 2022.



**LARRAÍN GONZALO** born in Santiago de Chile in 1997. He is a Computer Science Engineer from Universidad Técnica Santa María. His research is focused on applying AI-related techniques to solve current world problems such as mental health.



**ROJAS-MORALES NICOLÁS** (Member, IEEE) received the Ph.D. degree from Universidad Técnica Federico Santa María, Chile, in 2018. He is a Young Researcher with the Computer Science Department, Universidad Técnica Federico Santa María. His research efforts have resulted in funded research projects, journal publications, and high-level heuristic search conferences, such as GECCO, ANTS, and CEC. His research interests include opposition-inspired learning strategies, parameter setting problems, local optima networks, and machine learning techniques for detecting mental health conditions. He has served as a reviewer for different conferences in evolutionary computation and journals.



**GONZALEZ NICOLÁS** received the Civil Engineering degree in computer science from Universidad Técnica Federico Santa María, Valparaíso, Chile, in 2017. Since 2020, he has been leading the multidisciplinary team as a CEO of Servicios y Soluciones de Calidad de Vida SpA. His research interests include finding patterns in college students' mental health through machine learning algorithms and providing gamified technological solutions.



**OLCAY DANIEL** received the B.S. degree in psychology from the University of Tarapacá, Chile, in 2014. He is currently pursuing the M.S. degree in gender and social intervention with Bernardo O'Higgins University, Santiago, Chile. Since 2021, he has been the Psychology and Quality of Life Director with Servicios y Soluciones de Calidad de Vida SpA. His research interests include mental health, cognitive behavioral therapies, finding patterns in college students' mental health through machine learning algorithms, and providing gamified technological solutions.

...