Prediction Model for Common Mental Disorder and Depression in users of Psychoactive Drugs

Rhyan X. de Brito, Carlos A. R. Fernandes, Roberta M. M. Moreira and Eliany N. Oliveira

***Abstract*—Mental disorders are among the most prevalent diseases in the world. Many studies have observed the relationship between the use of psychoactive substances and mental diseases, such as Common Mental Disorder (CMD) or depression. The present paper aims to test the effectiveness of ML techniques as auxiliary tools in the pre-diagnosis of CMD and depression, through the classification of users of psychoactive substances. The main objective is to obtain a model for predicting the risk of depression and CMD, as well as to determine which factors contribute most to the risk of these mental diseases. The databases used in this work are composed of 605 samples from people from eight cities in the state of Ceará, Brazil, collected from January to July 2019. The results showed that the tested ML techniques reached an accuracy of 82.81% and 81.98% in the prediction of CMD and depression respectively, with the Support Vector Machine (SVM) and Sequential Backward Selection (SBS) methods. The results also showed that the use of tobacco derivatives, alcohol and cocaine/crack are the most significant factors for predicting these CMD and depression.**

***Index Terms*– common mental disorder, depression, psychoactive drugs, data mining, machine learning, prediction model.**

I. Introduction

Over the past 30 years, many epidemiological surveys around the world have shown that mental disorders have become very relevant from a public health care perspective due to their prevalence and persistence, accounting for approximately 12% of the global disease diagnostics [1]. In particular, the Common Mental Disorder (CMD), responsible for the reduction of the ability to concentrate and memory disorders, is considered the most prevalent mental suffering in the world population, being estimated to be among the biggest disabling causes in 2030 [2]. The CMD is characterized by depressive, anxious and somatic symptoms, such as irritability, fatigue, insomnia, excessive worry, among others.

Depression is also one of the most prevalent mental illnesses in the world. According to the World Health Organization, more than 350 million people worldwide suffer from depression, and it will likely be the main global disease by 2030 [3],[4]. Depression can be understood as a state of mind or a type of physiological problem that causes many symptoms, resulting in limitations of mental and physical functioning [5], [6]. Some biological issues may also contribute to depression, such as low levels of serotonin, dopamine and noradrenaline that are synthesized in the brain [7]. For authors [7], [8], the accumulation of homocysteine by genetic alteration of MTHFR C677T, as well as folate deficiency, decrease the synthesis of the neurotransmitters' dopamine, norepinephrine, epinephrine and serotonin, leading to depression due to the reduction of neurotransmitter synthesis. Individuals with depression suffer with melancholy, having difficulties in concentration and interaction with other people [5].

The use of psychoactive substances such as alcohol, tobacco, cocaine and crack, has a significant impact on the intensity and prevalence of CMD and depression, as they act on the central nervous system, causing effects on cognitive, behavioral and psychological functions, as well as causing changes in mood, behavior and consciousness [9]. Indeed, many studies have observed the relationship between the use of psychoactive drugs and several health problems, such as CMD and depression [9], [10], [11].

On the other hand, Data Mining (DM) and Machine Learning (ML) have become very popular in many areas of knowledge as auxiliary mechanisms for solving various problems. Currently, DM and ML have applications in a huge variety of knowledge fields and, in particular, they have become powerful tools in the fields of medicine, health and biology [12], [13]. The systems based on DM and/or ML developed within these areas aim to identify patterns in large amounts of data and to assist in clinical decisions, being a powerful tool for helping in pre-diagnosis and in predictive systems.

The present work aims to test the effectiveness of DM and ML techniques as auxiliary tools in the pre-diagnosis of CMD and depression, through the classification of users of psychoactive drugs. In particular, the main objective is to obtain a model for predicting the risk, based on data related to the use of psychoactive drugs and socio-economic data. Another objective is to determine which factors contribute most to the prediction of the risk of CMD and depression. These two objectives combined can assist health professionals and help the development of public health policies.

The prediction model is based on a classification system that follows the steps of the Knowledge Discovery in Databases (KDD) method for DM [13]. The KDD approach allows a better exploitation of the database, leading to an efficient use of the ML techniques. Several ML models are tested in the classification system, in order to find the one that has the best ability to model the considered database. In particular, the following nine classifiers were tested: k-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA),

Rhyan X. de Brito, Instituto Federal de Educação, Ciência e Tecnologia do Ceará, Tianguá, Ceará, Brasil, rhyan.brito@ifce.edu.br

Carlos Alexandre Rolim Fernandes, Universidade Federal do Ceará, Sobral, Ceará, Brasil, alexandrefernandes@ufc.br

Roberta M. M. Moreira, Faculdade IEducare, Tianguá, Ceará, Brasil, robertamoreiraenf@hotmail.com

Eliany N. Oliveira, Universidade Estadual Vale do Acaraú, Sobral, Ceará, Brasil, elianyy@gmail.com

Quadratic Discriminant Analysis (QDA), Naive LDA, Naive QDA, Extreme Learning Machine (ELM) and Random Forest (RF).

In addition, in order to increase the accuracy of the classifiers and to reduce the number of features, three techniques of feature selection/transformation were tested: Principal Component Analysis (PCA), *Sequential Feature Selection* (SFS) and *Sequential Backward Selection* (SBS). Regarding the objective of determining which factors contribute most to the prediction of CMD, an analysis based on the entropy (information gain) of the attributes is performed.

The database used in this work was built with data collected from 605 participants, from January to July 2019, in eight municipalities of the state of Ceara/Brazil that have mental health services and/or therapeutic communities that assist users of psychoactive substances. The data were collected in interviews supported by three instruments: a form for profile of sociodemographic, clinical and pattern of consumption, the Self-Reporting Questionare (SRQ-20) and the Patient Health Questionnaire-9 (PHQ-9) [14]. The results showed a good accuracy of the ML techniques in predicting the CMD and depression, reaching accuracy of 82.81% and 81.98%, respectively.

The rest of the work is divided as follows. In Section II, the works related to this article are presented. In Section III, the theoretical framework about CMD is addressed. In Section IV, the methods used in the work are presented. In Section V, the results are shown and discussed. In Section VI, the conclusion and future works are discussed.

II. Related works

This section presents the state of the art related to the use of DM and ML as supporting tools in problems related to CMD, depression and similar mental diseases.

In [15], a clinical study aiming to to differentiate healthy people from patients suffering from schizophrenia and depression based on the Electroencephalogram (EEG) rhythm is presented, using Artificial Neural Networks (ANNs) with three groups of patients: 10 normal, 10 schizophrenic and 10 depressive. The ANNs were able to correctly identify schizophrenia, depression and normal controls with an overall accuracy of 60% to 80%. However, the classification accuracy of the self-organized competitive network was 40% to 60%.

In [2], a study was carried out for classifying the different types of headache using several ML methods, with data collected from 2,177 patients diagnosed with headache at the Neurological Clinic, in the city of Joinville, Brazil, from January 2010 to November 2014. The reached accuracy was around 76%.

In [16], different classifiers, with a genetic algorithm being used for feature selection, were used to distinguish patients with depression from healthy individuals, based on Electroencephalography (EEG) records from 30 depressed and 30 healthy patients. The SVM classifier using a genetic algorithm for feature selection reached an accuracy of 88.6% in the classification of patients with depression.

In [17], a predictive model is presented to diagnose anxiety and depression in elderly patients from sociodemographic and health-related factors. Ten classifiers were evaluated with a data set of 510 geriatric patients. The highest precision was 89%, obtained with the RF classifier.

In [18], a computational tool was developed using sequencing data from the complete genome of people with mental disorders. The authors used a scoring system based on deep learning (*ncDeepBrain*) to analyze personal genome sequencing data. The authors observed that the use of deep learning and logistic regression to discriminate the disease variants from the neutral variants reached an accuracy of 82%.

In [19], the diagnosis of anxiety and depression in young children was addressed, through the use of 90-second fear induction, during which the participant's movements were monitored using a wearable sensor. Several sets of attributes, as well as several modeling approaches, were verified, the logistic regression providing the best performance with an accuracy of 80%.

In [20], ANNs were applied to a set of data to predict mental disorders based on lifestyle and psychometric data of people from all age groups. The data collection was based on an Android application through a questionnaire to analyze mental abilities, behavior, lifestyle and personality.

In [21], a neuro-fuzzy system was used to recognize mental disorders such as schizophrenia, phobia, depression, anxiety and obsessive compulsive disorder, using data mining. For the data collection, questionnaires about symptoms and types of disorders were used. For the fuzzy inference system, the Mamdani model was used, with 65 rules to determine the classification. The system obtained an accuracy of 81.94% for the test data.

In [22], ML was used to track bipolar disorder using the Mood Disorder Questionnaire. The data set was fed to a decision tree classifier to determine which resource is the most significant in the data set. The precision achieved in the experiment was 88.07%. In [23], the authors conducted a study for the early detection of bipolar disorder using data from 300 participants. The success rate in the experiment was around 99.2% and 99.6% according to the different parameters tested.

In [24], a study was made aiming to predict the severity of depression using acoustic and visual behavioral markers, based on three different sets of questions. The decision three classifier reached 96% of accuracy.

Although the above mentioned works have brought important contributions, some issues are not addressed in these studies. It can be concluded from the above bibliography review that, although there is a significant number of works using ML methods to study depression, there is a lack of works using ML along with CMD. Moreover, these works do not try to study the impact of psychoactive drugs in CMD and depression using ML, which is the main objective of the present paper. In addition, these works lack of testing a higher number of classifiers along with feature selection techniques, which may considerably increase the accuracy of the classification process.

Comparing the accuracy results obtained in the present work with the ones obtained in the works cited in this section , it can be concluded that the present paper has reached satisfactory

levels of accuracy (above 80%, as it will be presented in Section VI).

III. Common Mental Disorder (CMD)

The expression CMD was coined by Goldeber and Huxley, and the concept of CMD was developed in the 1970s, through researches on mental illness, in the context of primary public health care [9]. The CMD, also known as Minor Psychiatric Disorders (MPD), is characterized by intense psychological distress with important consequences in the individual's health and in many other aspects of life, such as work, studies and other daily activities. The CMD is manifested as a mixture of somatic, anxious and depressive symptoms, such as states of anxiety, irritability, fatigue, insomnia, memory and concentration problems and somatic complaints.

The patient with CMD has generally non-psychotic symptoms, such as complaints of anxiety, irritability, somatization, decreased vital energy and depressed mood [25]. The CMD is characterized by the presence of different symptoms for at least seven days. The evaluation of these symptoms allows an early diagnosis, as well as the monitoring of depressive disorders, anxiety, phobia, panic disorder and obsessive-compulsive disorder, which are characteristic of the types of CMD [11]. Indeed, an early and correct diagnosis of this disorder is essential to avoid physical and psychological damage to the individual and burden on the health system [26].

The prevalence of CMD varies worldwide, but it has generally a high frequency of occurrence in the population. In [11], approximately one third of the interviewees (31.47%) of the Central region of Brazil had CMD, with a higher prevalence in the Southeast (51.9 % to 53.3 %), Northeast (64,3%) and South (57.7%) regions.

The diagnosis for the CMD is performed through the application of the SRQ-20, developed by Harring and Mcmullin [27]. The questionnaire originally had 24 questions: twenty on non-psychotic disorders and four on psychotic disorders. The version currently applied in Brazil was validated by Mari and Williams, who observed a sensitivity of 83%, specificity of 80% and 19% of classification errors [27].

The SRQ-20 is an instrument for tracking non-psychotic mental disorders in which the answers are categorical yes/no. Each affirmative answer scores with a value of 1, the final score being calculated through the sum of these values. The scores obtained are related to the probability of the presence of non-psychotic disorder, ranging from 0 (no probability) to 20 (extreme probability) [28]. The answers enable the establishment of a score and, if this score is above 7, the individual is considered positive for CMD [29].

In addition, the SRQ-20 is recommended by the World Health Organization (WHO) for community studies and primary health care, especially in developing countries, as it is easy to use and has a low cost, being used in several countries of different cultures for tracking non-psychotic disorders [28].

IV. Depression

According to [30], the depression characterized by the presence of sad, melancholic, empty or irritable mood, accompanied by somatic and cognitive changes that significantly affect the individual's ability to function, differing in the aspects of duration, moment or presumed etiology. [30] states that the abuse of substances such as psychoative drugs and some medications can be associated with depression.

The depression or melancholia has been recognized as a clinical syndrome for over 2,000 years and a fully satisfactory explanation for its intriguing and paradoxical features has not yet been found, with important unresolved questions about its nature, classification and etiology [31]. Among these questions, the authors of [31] cite the following points: (i) Is depression an exaggeration of a mood state experienced by normal individuals, or is it qualitatively and quantitatively different from a normal mood state? (ii) What are the causes, defining characteristics, outcomes and effective treatments of depression? (iii) Is depression a type of reaction or a disease? (iv) Is depression primarily caused by psychological stress and conflict, or is it primarily related to a biological disorder?

There are no definitive answers to these questions and there is a clear disagreement between clinicians and researchers who have studied depression, with considerable controversy over the classification of depression. The nature and etiology of depression are still issues that have not been clearly defined. Some authorities claim that depression is primarily a psychogenic disorder, others claim that the cause is related to organic factors. A third group advocates the concept of two different types of depression: one psychogenic and one organic [31].

Within this context, the ability to correctly identify individuals at risk of developing depression is essential in epidemiological studies, as it allows the estimation of the prevalence of the disease [32]. According to [32], among the instruments used to identify individuals at risk of depression, one can find the PHQ-9, derived from the Primary Care Evaluation of Mental Disorders (PRIME-MD), which was originally developed to identify five common mental disorders in primary health care: depression, anxiety, alcohol abuse, somatoform disorders and eating disorders. The PHQ-9 is an instrument of relatively quick application, containing nine questions, which is an advantage in epidemiological studies, compared to others validated instruments, such as the Beck Depression Inventory (BDI) often used by specialists for screening depression [32], [6].

V. Methods

This section describes the methods used in this work, being divided in two parts: collection and construction of the database, and steps of the classification system.

A. Database Collection and Construction

The database used in this work was built in the context of a research project called "Mental health and the risk of suicide in drug users", with a favorable opinion from the Research Ethics Committee in 2018 and No. 2,739,560. The participants' consent was legitimized through the Free and Informed Consent Form (FICF).

The data were collected from 605 participants in eight municipalities of the state of Ceará that have mental health

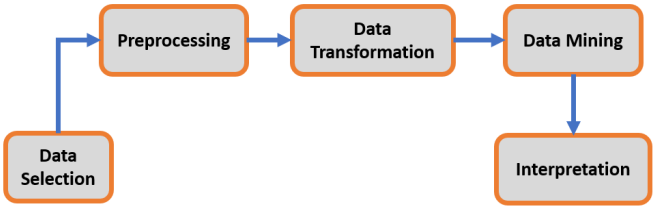


Fig. 1. Steps of the KDD method.

services to assist users of psychoactive drugs, such as the General Psychosocial Care Centers (CAPS Geral, from portuguese *Centros de Atenção Psicossociais Geral*), Psychosocial Care Centers for Alcohol and Others Drugs (CAPS AD, from portuguese *Centros de Atenção Psicossociais Álcool e outras drogas*) and therapeutic communities.

The survey took place from January to July 2019, through interviews using three instruments: one form for sociodemographic, clinical and consumption pattern profile, and the SRQ-20 and PHQ-9 forms, which can be found in [9], [14]. The inputs of the prediction model presented in this work are the data obtained from the sociodemographic, clinical and consumption pattern form, where as the outputs. i.e. the prediction of CMD and depression, are obtained from the SRQ-20 and PHQ-9 questionnaires, respectively.

The sociodemographic, clinical and consumption pattern form seeks to characterize the participants using variables such as gender, age, self-reported skin color/race, religion, education, occupation, marital status, number of children, family income, number of residents in the household and housing situation. As for the clinical aspects, the presence of clinical or psychiatric comorbidities are also investigated in the form, as well as its relationship with the use of psychoactive drugs. The SRQ-20 and PHQ-9, on the other hand, are well-established forms containing questions for tracking non-psychotic disorders [14].

A part of this database was analysed in [9] using inferential statistics and simple tests of association and correlation. This study found a high rate of people with depressed mood, anxiety and somatic symptoms. Moreover, the main social risk factors observed were the female gender and being young, while being catholic or evangelical, having a steady partner and children were found as protective factors. In [9], it was also found that the use of psychoactive drugs has a major impact in the risk of CMD and depression.

B. Steps of the Prediction Model

The prediction model presented in this work is based on a classification system that follows the stages of the KDD method for DM, with the following steps: (1) data selection; (2) preprocessing; (3) data transformation; (4) data mining and (5) interpretation [33]. A simplified scheme of the steps of the prediction model is shown in Fig. 1

In Step 1, some samples with missing attributes were discarded, as well as some questions that were not related to the classification. Moreover, some redundant information has also been discarded. Examples of redundant information are birth date and age, family income in *reais* (Brazilian currency) and family income in terms of the minimum wage. Besides, examples of irrelevant data, in view of the classification, are municipality of birth, municipality of residence and type of public service where the data were collected.

In Step 2, the questionnaire responses were coded and inconsistencies were corrected by discarding samples with missing information and eliminating some attributes that contain redundant information. Attributes that we considered unrelated to the classification problem were also removed. Some questions as, for instance, self-reported skin color/race, religion and type of habitation, were coded in several binary variables, one for each skin color/race, each religion and type of habitation. After the preprocessing, the database contained 84 attributes, 605 samples and 2 binary outputs: one output representing the prediction of CMD (yes or no), and one output output representing the prediction of depression (yes or not).

Subsequently, in Step 3, some feature transformation/selection techniques were applied. In particular, the following techniques were tested: PCA, SFS and SBS. The PCA is a feature transformation technique whose main idea is to reduce the dimensionality and correlation of the dataset. The PCA transforms the original features into a new set of variables that are uncorrelated and organized so that the first components contain most of the variance of the original data set [34], which allows a reduction in the number of used features. As a consequence, the PCA may also decrease considerably the processing time. In this work, the number of PCA components was chosen based on the accuracy, after testing some values for this parameter.

On the other hand, in the feature selection method SFS, the features are added sequentially to the set of used features until the addition of another feature does not increase the accuracy of the classifier [35]. In contrast, in the feature selection algorithm SBS, the features are removed sequentially from a complete set of features, until the removal of another feature does not decrease the accuracy [35].

The SFS and SBS are search algorithms with relatively low computational cost that improve the efficiency of the classifier by decreasing the number of used features. Moreover, they may improve the predictive ability of the classifier [36].

After that, in Step 4, the data are sent to a binary classifying algorithm, for preforming the prediction of the classes of the participants, related to the presence of CMD or depression. Nine classifiers were tested in this work: KNN, MLP, SVM, LDA, QDA, Naive LDA, Naive QDA, ELM and RF. The Naive LDA and Naive QDA classifiers are versions of the standard LDA and QDA that assume that the features are correlated. We decided to test a high number of classifiers in order to verify which ones perform better in the classification of the considered database. These classifiers were chosen because they are well-established and popular in the literature, and they were used in similar applications [15].

The hyperparameters of each classifier were obtained through testing, based on the accuracy obtained during the test phase, using the *k-fold* cross-validation technique with $k = 10$ folds. The classifiers were tested with or without the

z-score normalization. This method normalizes the standard deviation and removes the mean of each attribute. The best results were obtained with the z-score normalization, hence, the results presented in the next section correspond to the case with z-score normalization.

As it will be viewed in results section, the MLP and SVM techniques provided the best accuracy. Due to this reason, the most part of the results were obtained using these classifiers. The MLP is a feedforward ANN with at least three layers of nodes (input layer, hidden layer and output layer) and neurons that use nonlinear activation functions. The SVM classifier performs the separation of the classes through hyperplanes that are optimized for generating the greatest possible distance between the classes, usually using kernel functions.

In the Step 5 of the KDD method, the results are visualized and analyzed, which is done in the next section. The figures of merit used in this work are the accuracy and confusion matrix, obtained in the $k = 10$ folds with binary classification of CMD (or depression). Regrading the analysis of the importance of the features, the information gain, based on the entropy, [37], [38], [39], is used to determine the best attributes.

VI. Results and Discussion

In this session, the results of the tests performed are presented and discussed. As earlier mentioned, the results were obtained using the database presented in the Subsection V-A, k-fold cross validation with $k = 10$ and z-score normalization. The accuracy showed below represent the average correct classification rate (number of correctly classified samples divided by the total number of samples) in the 10 folds during the binary classification of the CMD and depression. Firstly, the results related to CMD are shown and, then, the results associated with depression are presented.

TABLE I
HYPERPARAMETERS USED BY THE CLASSIFIERS (CMD)

Classifier	Hyperparameter
MLP	2 hidden layers with 10 neurons each, activation function: linear saturated, scaled conjugate gradient backpropagation, learning rate $= 0.2$, batch size $= 1$, number of epochs $= 30$
SVM	Polinomial *kernel* (non-homogeneous) with degree 1, $C = 1$, KernelScale: 1/sqrt(2*0.01), one-vs-one
ELM	1 hidden layer with 25 neurons
KNN	$K = 40$
RF	number of seeds $= 1$, number of trees$= 400$

TABLE II
ACCURACY IN THE PREDICTION OF CMD - WITH PCA

Classifier	N. of PCA comp	Accuracy
MLP	81	77,70%
SVM	80	77,54%
ELM	35	75,90%
KNN	55	75,90%
QDA	60	73,61%
RF	55	73,05%
Naive LDA	75	72,46%
LDA	75	71,64%
Naive QDA	25	71,64%

A. Common Mental Disorder (CMD)

1) Classification of CMD: The following results have the objective of evaluating the performance of the classification algorithms using the feature transformation method PCA and the feature selection techniques SFS and SBS. As earlier mentioned, many simulations were carried out to adjust the hyperparameters of these classifiers. Table I shows the classifiers' hyperparameters that provided the best results. In the rest of the simulations related to the classification of CMD, the results were obtained using the hyperparameters of Table I.

Table II shows the accuracy of the nine tested classifiers using the PCA, with the third column showing the number of used components. For each classifier, several values for the number of PCA components were tested, with the value that provides the highest accuracy being chosen.

It can be viewed from Table II that there is a significant difference in the accuracy obtained by the tested classifiers. This is due to the different abilities of each classifier for discriminating the data. It can also be viewed in this table that the MLP and SVM techniques obtained the highest accuracy, with 77.70% and 77.54% respectively. This result is expected, as the MLP and SVM are two of the most popular and efficient classifiers. On the other hand, the classifiers that obtained the worst results were the LDA and Naive QDA, with an accuracy of 71.64% for both the techniques. This is due to the fact that these classifiers assume that the features follow Gaussian distributions, which is not a valid hypothesis for the considered database.

Some simulations were carried out using the feature selection techniques SFS and SBS, and the MLP and SVM classifiers, which provided the best accuracy in Table II. In these simulations, the SVM provided the best results. Due to this, only the results obtained with the SVM are shown in the sequel. Tables III and IV show the confusion matrices with the true and predicted classes, obtained by the SVM along with the SFS and SBS techniques, respectively.

It can be viewed from Tables III and IV that both the feature selection techniques were able to increase the accuracy of the SVM with respect to the PCA case. In particular, the SVM with the SBS reached an accuracy of 82.81 %, which is 4,67 % higher than the rate obtained by the SVM with PCA. This performance gain is due to the fact that the SFS and SBS techniques select subsets of features that are best suited to the problem.

It can also be noted that the SBS provided a better result

TABLE III
CONFUSION MATRIX WITH THE TRUE AND PREDICTED CLASSES - SVM WITH SFS (CMD)

		True class		
		CMD	No CMD	Accuracy
Predicted	CMD	**387**	67	
class	No CMD	58	**93**	
Rate (%)		86.97	58.13	79.34

TABLE IV
CONFUSION MATRIX WITH THE TRUE AND PREDICTED CLASSES - SVM WITH SBS (CDM)

		True class		
		CMD	No CMD	Accuracy
Predicted	CMD	**395**	54	
class	No CMD	50	**106**	
Rate (%)		88.76	66.25	82.81

than the SFS. However, the SBS used much more features than the SFS. Indeed, the SFS selected only four features, while the SBS excluded only two features for the classification process. It is also noteworthy that the SFS reached good accuracy with only four features.

Another observation that can be made from these confusion matrices is that the class that represents the positive diagnosis for CMD obtained the correct classification rates than the class representing the absence of CMD, i.e. the sensitivity (true positive rate) is higher than the specificity (true negative rate) in Tables III and IV. From the point of view of public health policies, this can be viewed as a desired characteristic for a prediction system, as it has a smaller probability to miss the detection of CMD in people that are in a group of risk. In the best case (Table IV), the sensitivity is 88.76%.

2) Importance of Attributes: In this subsection, some results that analyse the importance of the features in the classification of the considered database are presented, in order to determine which factors are the most relevant to the prediction of CMD. The most common parameter used to measure the relevance of the features is the information gain, based on the entropy [37], [38], [39]. The information gain is commonly used in the design of decision tree classifiers, like the RF, but they also indicate how relevant the attributes are for a classification task.

Table V shows the 10 highest information gains of the features, with the respective feature descriptions. In this tables, the attributes "CID10-F19" and "CID10-F17" represent diagnostics for mental disorder due to multiple drug use and to smoking, respectively, and the attributes "psychoactive drug problem: tobacco" and "psychoactive drug problem: cocaine/crack" inform if the participant had psychoactive problems due tobacco derivatives and cocaine/rack, respectively.

Table V shows a high influence of the use of psychoactive drugs on the discrimination of CMD, with the use of cocaine/crack being the most relevant factor. The use of alcoholic beverages and tobacco derivatives is also an important factor for predicting the risk of CMD. Indeed, the consumption of multiple psychoactive substances can directly interfere in the users' mental health by increasing the probability of breaking

TABLE V
THE 10 HIGHEST INFORMATION GAINS OF THE FEATURES (CMD)

Information gain	Feature Description
0.0165	depression
0.0175	CID10-F19
0.0187	gastrointestinal disorders
0.0191	psychoactive drug problem: tobacco
0.0193	age
0.0209	most used drug: alcohol
0.0222	psychoactive drug problem: cocaine/crack
0.0223	no occupation
0.0241	CID10-F17
0.0324	most used drug: cocaine/crack

or weakening social relationships, causing a reduction in the self-esteem and, consequently, feelings of loneliness [40]. In addition, the use of psychoactive drugs can impair the clinical treatment of the individual, causing a greater risk for the worsening of the CMD [40].

Moreover, Table V shows that having no occupation (no job) is an important risk factory, as well the age of the participants, where being younger constitutes a risk factor. These results are in accordance with the conclusion of [9]. This table also indicate that depression is a risk factor for CMD. In [41], it was observed that there is a direct relationship between depression and CMD in users of psychoactive drugs. Indeed, depression and the use of psychoactive substances directly interfere with the quality of life of these individuals, especially the physical, social and mental health, contributing to the presence of CMD [41].

Besides, the disorder related to the use of multiple substances and gastrointestinal disorders are also important predictive factors for CMD. The association between gastrointestinal disorders and psychological symptoms has been evidenced in the works [42], [43], [44], [45]. Therefore, it should be highlighted the high number of relevant factors that are important for the CMD, which requires early intervention for a better prognosis. Finally, it is worth mentioning that, while the entropy of the database is equal to 0.833, the sum of the information gains of all the 84 features is equal to 0.440, which can be considered is a significant amount of information gain.

B. Depression

1) Classification of Depression: In this subsection, the performance of several classification algorithms in the prediction of the risk of depression is studied, using the PCA, SFS and SBS methods. Table VII shows the accuracy obtained by the nine tested classifiers using the PCA, as well as the number of used PCA components. As in Subsection VI-A, several simulations were carried out in order to adjust the hyperparameters of the classifiers. Table VI shows the hyperparameters that provided the best accuracy. In the rest of the simulations of the paper, the results were obtained using the hyperparameters of Table VI.

Comparing the accuracy of the different classifiers in Table VII, one may note that there is a difference of 4.4% in the accuracy provided by the best and worst methods. It can also be viewed in this table that the SVM and MLP techniques

TABLE VI
HYPERPARAMETERS USED BY THE CLASSIFIERS (DEPRESSION)

Classifier	Hyperparameter
MLP	2 hidden layers with 10 neurons each, activation function: tangent hiperbolic, resilient backpropagation, learning rate $= 0.1$, batch size $= 1$, number of epochs $= 20$
SVM	(Gaussian *kernel*), $C = 0.89$, KernelScale: 1/sqrt(2*0.02), one-vs-one
ELM	1 hidden layer with 40 neurons
KNN	$K = 70$
RF	number of seeds $= 1$, number of trees$= 100$

TABLE VII
ACCURACY IN THE PREDICTION OF DEPRESSION - WITH PCA

Classifier	N. of PCA comp	Accuracy
MLP	80	71.48%
SVM	79	71,31%
ELM	40	70.82%
RF	55	70.74%
KNN	30	70.49%
QDA	40	69.02%
Naive QDA	40	68.52%
Naive LDA	55	67.70%
LDA	35	67.05%

obtained the best results, as observed in the experiments carried out with the CMD, achieving accuracy of 71.48% and 71.31%, respectively. On the other hand, the Naive LDA and LDA provided the worst accuracy, with 67.70% and 67.05% respectively, as they assume that the attributes are Gaussian distributed, which is not a valid hypothesis.

Furthermore, comparing the results of Tables II and VII, it can be viewed that the accuracy obtained with the depression database are significantly lower than those obtained with the CMD database, indicating that depression is more difficult to be modeled by the tested ML techniques along with PCA. This is probably due to the fact the depression presents symptoms with subjective aspects, as well as intriguing and paradoxical characteristics, with unresolved questions about its nature and classification [31].

Some simulations were carried out using the feature selection techniques SFS and SBS, and the MLP and SVM classifiers, which provided the best accuracy in Table VII. As well as in the CMD case, in these simulations, the SVM provided the best results. Due to this, only the results obtained with the SVM are shown in the sequel. Tables VIII and IX show the confusion matrices obtained by the SVM along with the SFS and SBS techniques, respectively. It can be

TABLE VIII
CONFUSION MATRIX WITH THE TRUE AND PREDICTED CLASSES - SVM WITH SFS (DEPRESSION)

		True class		
		Depres.	No Depres.	Accuracy
Predicted class	Depres.	**332**	41	
	No Depres.	78	**154**	
Rate (%)		80.98	78.97	80.33

TABLE IX
CONFUSION MATRIX WITH THE TRUE AND PREDICTED CLASSES - SVM WITH SBS (DEPRESSION)

		True class		
		Depres.	No Depres.	Accuracy
Predicted class	Depres.	**338**	37	
	No Depres.	72	**158**	
Rate (%)		82.44	81.03	81.98

TABLE X
THE 10 HIGHEST INFORMATION GAINS OF THE FEATURES (DEPRESSION)

Information gain	Feature Description
0.0191	age
0.0163	catolic religion
0.0217	no occupation
0.0158	number of household residents
0.0250	CID10-F17
0.0272	CID10-F19
0.0185	first drug used: tobacco
0.0352	must used drug: cacaine/crack
0.0177	psychoative drug problem: tobacco
0.0315	psychoative drug problem: cocaine/crack

viewed from these tables that the feature selection methods SFS and SBS provided significant gains in accuracy when compared to the PCA case. The best result was obtained with the SVM along with the SBS, reaching an accuracy of 81.98 %, representing a gain of 10,67 % with respect to the SVM with PCA.

It is worth noting that, as well as in the CMD database, the SBS technique presented better results than SFS. However, the SBS used more attributes, excluding only two features, while the FSF selected only four attributes. Another similarity between the results obtained with the two databases is that the sensitivity is higher than the specificity, which can be viewed as a positive aspect of the prediction system, as earlier mentioned. In the best case (Table IX), the sensitivity related to the detection of depression is 82.44 %.

2) Importance of Attributes: Table X shows the 10 highest information gains of the features, with the respective feature descriptions, for the depression database. As for the CMD database, the use of cocaine/crack stands out as the main factor for the detection of depression. The use of tobacco also represent a very relevant factor for the the detection of depression. Furthermore, the entropy of the depression database is equal to 0.906, while the sum of the information gains of the 84 attributes is equal to 0.470, which can be considered an expressive information gain.

These results corroborates with the discussion presented in Subsection VI-A2. Moreover, in [46], the authors reached

a conclusion that corroborates with these results, showing that the use of psychoactive drugs directly interferes in the prevalence of depression, being risk factors associated with the occurrence of negative effects on mental health, well-being and social performance.

VII. Conclusion and Future Work

This work presented a classification system based on DM and ML, in order to to classify people according to the risk of CMD and depression, based on the use of psychoactive substances and socioeconomic data, with the objective of assisting in the development of public policies for health. The database used is composed of 605 people from eight municipalities in the state of Ceará, Brazil, with data collected from January to July 2019.

The results showed the effectiveness of the prediction system as an auxiliary tool in the pre-diagnosis of CMD and depression. The SVM classifier, together with the SBS technique, achieved an accuracy of 82.81% for CMD and 81.98% for depression, which can be considered good accuracy, given the complexity of proposed problems. The study also looked at the most discriminant variables in the classification process, in order to identify the most significant factors in predicting the risk of CMD and depression. The results showed that the use of cocaine/crack is the most relevant factor for both CMD and depression. In addition, the use of alcohol and tobacco, and being unemployed are also important factors in predicting CMD. On the other hand, the results showed that the use of tobacco derivatives is a relevant factor for predicting depression.

Finally, when comparing the results related with the two diseases, it was observed that the depression provided better accuracy than the CMD. As future work, the application of deep learning architectures is considered, as well as the study the Autism Spectrum Disorders (ASD) in children. Moreover, a deeper analysis of the false negatives of CMD and depression, with their common factors, is also a perspective of work.

Acknowledgments

The authors would like to thank the Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (FUNCAP/Brazil) for the financial support to the research project entitled "Mental Health and the risk of suicide in drug users" through the BPI Program (Bolsa de Produtividade em Pesquisa, Estímulo à Interiorização e à Inovação Tecnológica), process N. BP3-0139-00310.01.00/18.

References

[1] P. Skapinakis, S. Bellos, S. Koupidis, I. Grammatikopoulos, P. N. Theodorakis, and V. Mavreas, "Prevalence and sociodemographic associations of common mental disorders in a nationally representative sample of the general population of greece," *BMC psychiatry*, vol. 13, no. 1, p. 163, 2013.

[2] A. T. Fenerich, M. T. A. Steiner, J. C. Nievola, K. B. Mendes, D. P. Tsutsumi, and B. S. dos Santos, "Diagnosis of headaches types using artificial neural networks and bayesian networks," *IEEE Latin America Transactions*, vol. 18, no. 01, pp. 59–66, 2020.

[3] S. Liu, J. Shu, and Y. Liao, "Depression tendency detection for microblog users based on svm," in *2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2021, pp. 802–806.

[4] C. Jiang, Y. Li, Y. Tang, and C. Guan, "Enhancing eeg-based classification of depression patients using spatial information," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 566–575, 2021.

[5] A. Ashraf, T. S. Gunawan, F. D. A. Rahman, M. Kartiwi, N. Ismail, and Ulfiah, "A summarization of the visual depression databases for depression detection," in *2020 6th International Conference on Wireless and Telematics (ICWT)*, 2020, pp. 1–6.

[6] S. Hashempour, R. Boostani, M. Mohammadi, and S. Sanei, "Continuous scoring of depression from eeg signals via a hybrid of convolutional neural networks," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 176–183, 2022.

[7] E. Bedson, D. Bell, D. Carr, B. Carter, D. Hughes, A. Jorgensen, H. Lewis, K. Lloyd, A. McCaddon, S. Moat *et al.*, "Folate augmentation of treatment–evaluation for depression (folated): randomised trial and economic evaluation." *Health Technology Assessment (Winchester, England)*, vol. 18, no. 48, p. vii, 2014.

[8] P. Bhatia and N. Singh, "Homocysteine excess: delineating the possible mechanism of neurotoxicity and depression," *Fundamental & clinical pharmacology*, vol. 29, no. 6, pp. 522–528, 2015.

[9] R. M. M. Moreira, E. N. Oliveira, R. E. Lopes, M. V. de Oliveira Lopes, P. C. de Almeida, and H. L. Aragão, "Common mental disorder in users of psychoactive substances (in Portuguese)," *Enfermagem em Foco*, vol. 11, no. 1, 2020.

[10] A. I. O. Lima, M. Dimenstein, R. Figueiró, J. Leite, and C. Dantas, "Prevalence of common mental disorders and use of alcohol and drugs among prison agents (in Portuguese)," *Psicologia: Teoria e Pesquisa*, vol. 35, 2019.

[11] R. Lucchese, P. C. D. Silva, T. C. Denardi, R. L. de Felipe, I. Vera, P. A. de Castro, A. de Assis Bueno, and I. L. Fernandes, "Common mental disorder among individuals who abuse alcohol and drugs: cross-sectional study (in Portuguese)," *Texto & Contexto Enfermagem*, vol. 26, no. 1, pp. 1–7, 2017.

[12] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G. Yang, "Deep learning for health informatics," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 4–21, 2017.

[13] R. A. Pazmiño-Maji, F. J. García-Peñalvo, and M. Conde-González, "Statistical implicative analysis approximation to KDD and data mining: A systematic and mapping review in knowledge discovery database framework," 2017.

[14] R. M. M. Moreira, "Mental disorder and the risk of suicide in users of psychoactive substances (in Portuguese)," MSc Dissertation, Universidade Federal do Ceará, 2020.

[15] Y.-j. Li and F.-y. Fan, "Classification of schizophrenia and depression by EEG with ANNs," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE, 2006, pp. 2679–2682.

[16] B. Hosseinifard, M. H. Moradi, and R. Rostami, "Classifying depression patients and normal subjects using machine learning techniques," in *2011 19th Iranian Conference on Electrical Engineering*. IEEE, 2011, pp. 1–4.

[17] A. Sau and I. Bhakta, "Predicting anxiety and depression in elderly patients using machine learning technology," *Healthcare Technology Letters*, vol. 4, no. 6, pp. 238–243, 2017.

[18] A. Khan and K. Wang, "A deep learning based scoring system for prioritizing susceptibility variants for mental disorders," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2017, pp. 1698–1705.

[19] R. S. McGinnis, E. W. McGinnis, J. Hruschak, N. L. Lopez-Duran, K. Fitzgerald, K. L. Rosenblum, and M. Muzik, "Rapid anxiety and depression diagnosis in young children enabled by wearable sensors and machine learning," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 3983–3986.

[20] D. Sapkal, C. Mehta, M. Nimgaonkar, R. Devasthale, and S. Phansalkar, "Prediction of mental disorder using artificial neural network and psychometric analysis," in *Data Management, Analytics and Innovation*. Springer, 2021, pp. 369–377.

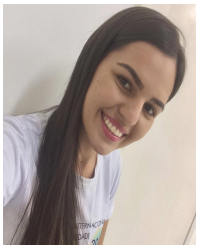
[21] M. Silvana, R. Akbar, M. Audina *et al.*, "Development of classification features of mental disorder characteristics using the fuzzy logic mamdani method," in *2018 International Conference on Information Technology Systems and Innovation (ICITSI)*. IEEE, 2018, pp. 410–414.

[22] R. Jadhav, V. Chellwani, S. Deshmukh, and H. Sachdev, "Mental disorder detection: Bipolar disorder scrutinization using machine learning," in

2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2019, pp. 304–308.
[23] D. Fitriati, F. Maspiyanti, and F. A. Devianty, "Early detection application of bipolar disorders using backpropagation algorithm," in *2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. IEEE, 2019, pp. 40–44.
[24] M. Muszynski, J. Zelazny, J. M. Girard, and L.-P. Morency, "Depression severity assessment for adolescents at high risk of mental disorders," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 70–78.
[25] C. B. Falco, J. M. G. Fabri, E. B. Oliveira, A. V. Silva, M. G. de Araújo Faria, and C. C. F. Kestenberg, "Common mental disorder among nursing residents: an analysis based on the self-reporting questionnaire (in Portuguese)," *Revista Enfermagem UERJ*, vol. 27, p. 39165, 2019.
[26] B. D. M. Parreira, B. F. Goulart, V. J. Haas, S. R. da Silva, J. C. dos Santos Monteiro, and F. A. Gomes-Sponholz, "Common mental disorder and associated factors: study with women from a rural area (in Portuguese)," *Revista da Escola de Enfermagem da USP*, vol. 51, p. e03225, 2017.
[27] M. C. d. S. Minayo, E. R. d. Souza, and P. Constantino, *Prevent and protect mission: living, working and health conditions for military police in Rio de Janeiro (in Portuguese)*. Editora Fiocruz, 2008.
[28] D. M. Gonçalves, A. T. Stein, and F. Kapczinski, "Performance evaluation of the self-reporting questionnaire as a psychiatric screening tool: a comparative study with the structured clinical interview for DSM-IV-TR (in Portuguese)," *Cadernos de Saúde Pública*, vol. 24, pp. 380–390, 2008.
[29] M. C. P. Lima, M. de S. Domingues, and A. T. de A. R. Cerqueira, "Prevalence and risk factors for common mental disorders among medical students (in Portuguese)," *Revista de Saúde Pública*, vol. 40, no. 6, pp. 1035–1041, 2006.
[30] A. P. Association *et al.*, *DSM-5: Manual diagnóstico e estatístico de transtornos mentais*. Artmed Editora, 2014.
[31] A. T. Beck and B. A. Alford, *Depressão: causas e tratamento*. Artmed Editora, 2016.
[32] I. S. Santos, B. F. Tavares, T. N. Munhoz, L. S. P. d. Almeida, N. T. B. d. Silva, B. D. Tams, A. M. Patella, and A. Matijasevich, "Sensibilidade e especificidade do patient health questionnaire-9 (PHQ-9) entre adultos da população geral," *Cadernos de Saúde Pública*, vol. 29, pp. 1533–1543, 2013.
[33] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth *et al.*, "Knowledge discovery and data mining: Towards a unifying framework," in *KDD*, vol. 96, 1996, pp. 82–88.
[34] I. JOLLIFFE, "Principal component analysis," *Wiley Online Library*, p. 63, 2002.
[35] S. Visalakshi and V. Radha, "A literature review of feature selection techniques and applications: Review of feature selection in data mining," in *2014 IEEE International Conference on Computational Intelligence and Computing Research*, 2014, pp. 1–6.
[36] A. U. Haq, J. Li, M. H. Memon, M. H. Memon, J. Khan, and S. M. Marium, "Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection," in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*. IEEE, 2019, pp. 1–4.
[37] M. Kubat, *An introduction to machine learning*. Springer, 2017.
[38] C. Aggarwal Charu, "Data mining: The textbook," *Switzerland: Springer*, 2015.
[39] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[40] P. Dalgalarrondo, *Psychopathology and semiology of mental disorders. (in Portuguese)*. Artmed Editora, 2018.
[41] A. Adan, J. M.-A. E, and G. Gilchrist, "Comparison of health-related quality of life among men with different co-existing severe mental disorders in treatment for substance use," *Health and quality of life outcomes*, vol. 15, no. 1, pp. 1–12, 2017.
[42] S.-Y. Lee, M.-C. Park, S.-C. Choi, Y.-H. Nah, S. E. Abbey, and G. Rodin, "Stress, coping, and depression in non-ulcer dyspepsia patients," *Journal of Psychosomatic Research*, vol. 49, no. 1, pp. 93–99, 2000.
[43] H. Strid, M. Norström, J. Sjöberg, M. Simren, J. Svedlund, H. Abrahamsson, and E. Björnsson, "Impact of sex and psychological factors on the water loading test in functional dyspepsia," *Scandinavian journal of gastroenterology*, vol. 36, no. 7, pp. 725–730, 2001.
[44] L.-T. Chou, C.-Y. Wu, H.-P. Chen, C.-S. Chang, P.-G. Wong, C.-W. Ko, and G.-H. Chen, "The correlation of depression and gastric dysrhythmia in functional dyspepsia," *Journal of clinical gastroenterology*, vol. 33, no. 2, pp. 127–131, 2001.
[45] K. Mine, F. Kanazawa, M. Hosoi, N. Kinukawa, and C. Kubo, "Treating nonulcer dyspepsia considering both functional disorders of the digestive system and psychiatric conditions," *Digestive diseases and sciences*, vol. 43, no. 6, pp. 1241–1247, 1998.
[46] M. Schenker and M. C. d. S. Minayo, "Fatores de risco e de proteção para o uso de drogas na adolescência," *Ciência & Saúde Coletiva*, vol. 10, pp. 707–717, 2005.

Rhyan X. de Brito received the BSc degree in computer science from the Universidade Estadual Vale do Acaraú and the MSc in Electrical and Computer Engineering at Universidade Federal do Ceará, Brazil, in 2021. He is effective professor at the Instituto Federal de Educação, Ciência e Tecnologia do Ceará, Brazil. His research interest lies in the area of machine learning and data mining.

C. Alexandre R. Fernandes received the BSc degree in electrical engineering from the Universidade Federal do Ceará (UFC), Brazil, in 2003, MSc degrees from the UFC and University of Nice Sophia-Antipolis, France, in 2005, and the double PhD degree from the UFC and UNSA, in 2009, in the field of signal processing. In 2008 and 2009, he was a Teaching Assistant with the UNSA/FR and, from July 2009 to February 2010, he was a Postdoctoral Fellow at UFC. In 2010, he joined the UFC, where he works as a full professor with the Department of Computer Engineering, in Sobral. He is the founder and former head of the Graduate Program in Electrical and Computer Engineering at UFC. He is the head of the Group of Assistive and Educational Technologies, a research group with several projects in the area of assistive technologies for people with disabilities and in the area of helth and educational technologies. His research interest lies in the area of machine learning, assistive technologies, data mining and tensor algebra.



Roberta M. M. Moreira received the BSc degree in Nursing from Universidade Estadual Vale do Acaraú, Brazil, in 2017 and the MSc degree in Family Health from the Universidade Federal do Ceará, Brazil, in 2020. She is currently professor at Faculdade Ieducare, Brazil. Her research interest lies in public health, mental health, violence and risk of suicide in drug users.

Eliany N. Oliveira received the BSc degree in Nursing from the Universidade Federal do Ceará (UFC), Brazil, in 1992, the MSc degree in Nursing from the UFC in 1999 and PhD degree in Nursing from the UFC in 2004. She was a Postdoctoral Fellow at University of Porto, Portugal, in 2016 and 2017. In 2011, she joined the Universidade Estadual Vale do Acaraú as full professor and she is currently head of the Interdisciplinary League on Mental Health-LISAM. Her research interest lies in public health, mental health, family Health and quality of life.