Psychiatry Research 322 (2023) 115127

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

# Machine learning-decision tree classifiers in psychiatric assessment: An application to the diagnosis of major depressive disorder

Daiana Colledani *, Pasquale Anselmi, Egidio Robusto

*Department of Philosophy, Sociology, Education and Applied Psychology, University of Padova, Via Venezia 14, 35131, Padova, Italy*

ARTICLE INFO

*Keywords:*
Machine learning
Psychodiagnostic test
PHQ-9
Sensibility
Specificity

ABSTRACT

This work illustrates the advantages of using machine learning classifiers in psychiatric assessment. Machine learning-decision trees (ML-DTs) represent a new approach to scoring and interpreting psychodiagnostic test data that allows for increasing assessment accuracy and efficiency. The approach is outlined in an easy yet detailed way, and its application is illustrated on real psychodiagnostic test data. Specifically, cross-sectional data concerning nonclinical and clinical Japanese populations were taken from a panel registered with an internet survey company. Responses to the Patient Health Questionnaire-9 (PHQ-9) underwent receiver operating characteristic (ROC) curve, DSM algorithm, and ML-DT analyses. The results showed greater diagnostic accuracy for ML-DT (0.71–0.75) compared with the DSM algorithm (0.69) and ROC curves (0.70–0.71). Moreover, ML-DT enabled classifying participants as having or not having a diagnosis of depression using, on average, the information from 2.99 out of 9 items ($SD = 1.35$). The application showed that ML-DTs can provide information of high clinical value to integrate traditional psychometric methods. The resulting assessments are informative, accurate, and efficient.

## 1. Introduction

Formulating timely and accurate diagnoses is crucial to maximizing the efficacy of therapeutic actions and reducing costs. The diagnostic process often requires the administration of multiple questionnaires and the execution of complex, expensive, and sometimes invasive clinical examinations. Understanding which symptoms are the most informative for each patient might not be easy. Medicine and psychology have recently begun to take advantage of machine learning (ML; Battineni et al., 2020; Witten et al., 2016), a subfield of artificial intelligence that aims to learn new pieces of knowledge from a set of data to use them for effectively predicting new cases (Witten et al., 2016).

This work aims to illustrate how ML can contribute to the diagnostic process in psychiatry. Specifically, the paper outlines the usefulness of ML classifiers as a new approach to interpreting the data from psychodiagnostic tests. ML can improve the efficiency of the diagnostic process while guaranteeing its accuracy. Moreover, it can provide information of high clinical value that supplements that of traditional methods. In the next section, ML is described, and its strengths and weaknesses are compared with those of traditional methods. Then, the results of an application on real psychodiagnostic test data are presented and discussed.

### 1.1. Psychological assessment and diagnosis with classical methods and ML classifiers

Usually, the testing process starts with the administration of a set of items to an individual and ends with the computation of an aggregate score that informs the clinician about the level of the assessed characteristic in that individual. The diagnosis of a certain disease is formulated if the score is above a specific value. Clinical cut-off scores are usually identified using receiver operating characteristic (ROC) curves (Carter et al., 2016). Given a diagnosis obtained with a gold standard (e.g., clinical interview), the ROC curve method identifies the score, among all possible ones, that allows for correctly classifying the largest number of persons as having or not having the disease. In ROC curves, sensitivities (i.e., true positive rates) and specificities (i.e., true negative rates) of all possible scores are tabulated. The score that maximizes both sensitivity and specificity is selected as the clinical cut-off score for future classifications and diagnoses. Two elements are needed to compute ROC curves: the test scores of a group of respondents and an independent gold standard indicating the condition of each of them,

* Corresponding author.
*E-mail address:* daiana.colledani@unipd.it (D. Colledani).

https://doi.org/10.1016/j.psychres.2023.115127
Received 26 October 2022; Received in revised form 17 February 2023; Accepted 19 February 2023
Available online 22 February 2023

namely "diagnosed" versus "non-diagnosed" (Zhou et al., 2011).

This situation closely resembles that of classification in supervised ML algorithms. These models require a set of data containing a number of predictors (e.g., the items of a test), which are measured on all instances (e.g., the subjects), and one dependent variable identifying the condition of each instance (e.g., diagnosed vs. non-diagnosed; Baştanlar and Özuysal, 2014; Witten et al., 2016). Starting from these data, ML classifiers learn a classification function that aims to predict the value of the dependent variable using the information provided by the predictors (Gonzalez, 2021a). Typically, ML algorithms implement a cross-validation approach in which a model is built using one part of the dataset (called the "training dataset"), and its predictive performance is evaluated on a different part of the dataset (called the "test dataset"; Yarkoni and Westfall, 2017). The predictors and the outcome variable are observed in both the training and test datasets. The predictive value of the model is evaluated by considering the accuracy of the prediction obtained in the test dataset when applying the model instructed in the training dataset (Witten et al., 2016).

### 1.2. Decision tree classifiers

Decision trees (DTs) are one of the most interesting ML applications in the clinical field. Compared with other algorithms, they provide a clear indication of the role and importance of each variable in the prediction (Higa, 2018; Zhao and Zhang, 2008). DTs are generally built using a top-down approach and with a recursive divide-and-conquer process (Breiman et al., 2017; Witten et al., 2016). A standard tree consists of a root and a series of branches, nodes, and leaves (Fig. S1 in the supplementary materials). The root is the origin of the tree, namely, the node from which all subsequent branches develop. It includes the entire sample. Each node is represented by a specific attribute (e.g., item/variable), and a chain of nodes from the root to a leaf (end of a branch) constitutes a branch. In each node, the instances (e.g., subjects) are divided into a series of branches (i.e., subsamples of subjects) determined by the values assumed by the attribute (e.g., the item score) that constitutes the node. If an attribute-node is nominal, the number of possible branches is determined by the number of possible values of the attribute. Conversely, if the attribute-node is numeric, DT algorithms work to identify the value of the attribute-node that splits the instances into subsamples that are as similar as possible relative to the outcome variable (Witten et al., 2016). Imagine a situation in which a binary variable (diagnosed, non-diagnosed) represents the outcome to be predicted by a DT, and an item (scored from 1 to 7) from a psychodiagnostic test constitutes one of the nodes of the tree. The algorithm could suggest generating two branches from the item-node that split subjects into two subsamples defined by a score $\leq 2$ or $> 2$. In this case, a score of 2 for the item is identified as the score that classifies the individuals into the two levels of the outcome variable (diagnosed vs. non-diagnosed) as accurately as possible. In other words, the score of 2 is the value of the node (i.e., the item) that minimizes the misclassification rate of the outcome in each subsample (Witten et al., 2016).

The first step in developing a DT is the selection of the attribute to be placed at the root node and from which branches recursively develop. At each node, all the present instances are progressively "distributed" into the branches, which are created according to the criteria identified at each step. If, at any step, all instances have the same classification, branch development stops. Branch development also stops when the data cannot be split any further or when growing other branches does not improve the classification. The selection of the nodes to be included in the tree and the root node is driven by the need to obtain the smallest possible tree. Consequently, at each step, the attribute is selected that produces the "purest" node (i.e., that containing the largest number of instances with the same classification; Witten et al., 2016). Node selection is typically based on information gain (Criminisi et al., 2012; Gupta et al., 2017; Witten et al., 2016), which increases with the average purity of the subsets generated by the division of instances and can be evaluated by considering the entropy of the class distribution (Gray, 2011). Entropy is maximum when all classes are equally probable and minimum (i.e., 0) when one of the classes has probability of 1. Uncertainty is minimal when entropy is 0.

### 1.3. ROC curve and DT as a means for interpreting the responses to psychodiagnostic tests

Although DTs and ROC curves start from the same data and pursue the same goal, they achieve it differently. Each method has advantages and disadvantages. ROC curves are a well-known and consolidated method, easy to calculate and understand. A weakness is that they rely only on an aggregate score to formulate a diagnosis. The same aggregate score can be obtained via different patterns of item responses so that equal scores could correspond to qualitatively different profiles. For example, a response of 1 to an item investigating the frequency of suicidal ideations coupled with a response of 4 to an item asking about sleep disturbances produces a score of 5. This score can also be obtained with the opposite pattern of responses, which indicates a very different profile. In situations like this, aggregate scores hide qualitative differences among individuals. On the contrary, DT algorithms value the qualitative role and information provided by each item. DTs guide the diagnosis through a sort of flowchart in which the item (or symptom of the disease) to consider at a specific step is based on the responses to the items already considered. The order in which the items appear in the DT provides information about their role in the diagnostic process (Zhao and Zhang, 2008). The DT provides a representation of the diagnostic process in which items/symptoms are evaluated individually but, at the same time, in relation to each other. This makes the diagnostic process more complete and, sometimes, faster. A weakness of DTs is that they do not provide any estimate of the intensity with which a trait or symptom is present in the individual.

The following section illustrates the results of an application of an ML-DT classifier to real psychodiagnostic test data. It is shown how ML-DT classifiers can be profitably used to score and interpret psychodiagnostic test data through a highly informative and efficient process. The performance of this method is compared with those of classical methods, and their advantages and disadvantages are discussed.

## 2. Method

### 2.1. Participants

Data were taken from a large web-based survey investigating emotions and psychopathology among Japanese people (Ito et al., 2015). The dataset (available on the FigShare repository; Doi et al., 2018) contains the responses of 2830 individuals (mean age = 42.44; $SD$ = 10.39; males $N$ = 1283) to the Japanese version of the Patient Health Questionnaire-9 (PHQ-9; Muramatsu et al., 2007). For each respondent, the presence of current or past psychiatric diagnoses was also indicated. The sample included 1163 individuals who belonged to the nonclinical population because they declared that, at the time, they had not been diagnosed with nor were they being treated for any psychiatric disorders. Among them, 509 declared that they had been diagnosed with or treated for psychiatric disorders in the past (however, they were neither diagnosed nor being treated at the time; nonclinical population with clinical history), whereas 654 declared that they had never been diagnosed with nor treated for psychiatric disorders (nonclinical population without clinical history). The remaining 1667 respondents belonged to the clinical population because they declared being, at that time, diagnosed with or treated for one or more psychiatric disorders (i.e., panic disorder, social anxiety disorder, obsessive-compulsive disorder, or other psychiatric disorders). Further details about the data can be found in Doi et al. (2018).

In this work, a subsample of 2205 individuals (mean age = 42.65; $SD$ = 10.51; males $N$ = 1023) was extracted according to a single eligibility

criterion, namely, belonging to the nonclinical population or being currently diagnosed with major depressive disorder (MDD). The subsample included 1163 individuals from the nonclinical population (52.7%; mean age = 42.64; *SD* = 11.55; males *N* = 474) and 1042 individuals with MDD (47.3%; mean age = 42.66; *SD* = 9.22; males *N* = 549; MDD only *N* = 406; mixed diagnoses *N* = 636; for further details, see Table S1 in the supplementary materials). There were no missing data. The total sample comprised 625 people diagnosed with psychiatric disorders different from depression (mean age = 41.68; *SD* = 9.92; males *N* = 260). They were not used to train the DT in order to reduce noise (Uğuz, 2011).

### 2.2. Measures

The PHQ-9 (Kroencke et al., 2001) consists of nine items based on the nine DSM-IV criteria for major depression, and it evaluates the frequency with which people experienced the symptoms over the previous 2 weeks (4-point scale from 0 "not at all" to 3 "nearly every day"). The instrument, including the Japanese version considered in this study, showed good validity and reliability. In the current sample, the Cronbach's α reliability coefficient was 0.94 (0.92 for the respondents with MDD and 0.91 for those from the nonclinical population).

Two methods are commonly used for scoring the PHQ-9. One is an algorithm that scores the test items according to the DSM-IV criteria for MDD diagnosis. This algorithm, hereafter denoted as the DSM algorithm, requires that at least five items are scored at least 2 (i.e., "more than half the days"). Among these items, one should be either "loss of interest (or pleasure)" or "depressed mood". The item concerning suicidal ideation contributes to the count if it is scored at least 1 ("several days"; Kroencke et al., 2001). The sensitivity and specificity of this algorithm were 0.806 and 0.895, respectively (Muramatsu et al., 2018). The second scoring method consists of computing the sum score of the test and using a cut-off score to detect MDD. Many studies indicated 10 as the optimal cut-off score, with a sensitivity and specificity of 0.905 and 0.766, respectively (Manea et al., 2012; Spitzer et al., 1999).

### 2.3. Analysis

The J48 classifier was run through the open-source software WEKA 3.8.5 (Waikato Environment for Knowledge Analysis, University of Waikato, New Zealand). The J48 algorithm is a Java extension of the better-known Quinlan C4.5 algorithm (Quinlan, 1993). It aims to find a knowledge structure among a set of input data to produce a DT that allows for classifying new data into the groups of a specific class variable. The J48 uses a recursive divide-and-conquer approach based on a greedy algorithm. At each node, the algorithm selects the attribute that most effectively splits the data into a series of subsets that are as similar as possible relative to the outcome variable. The splitting criterion is the normalized information gain provided by each variable (calculated as the difference in entropy). The attribute with the largest normalized information gain is chosen to make the classification. In this work, the J48 was run to develop a DT that uses the nine items of the PHQ-9 as independent variables and points to classify individuals as having or not having a diagnosis of MDD.

In the building phase, the algorithm develops the tree and its branches (Lin, 2001; Prabhakar et al., 2002; Sugumaran et al., 2007). It uses entropy-based information gain to select the items that must be placed at each node (i.e., those that, at each step, most accurately categorize individuals) and to identify the splitting rules (i.e., the scores generating branches) to distribute individuals. In the pruning phase, the nodes whose removal does not affect classification accuracy are removed iteratively. Setting a minimum size for each leaf allows for developing an interpretable DT, preventing overfitting, and obtaining higher prediction accuracy and generalizability (Dekker et al., 2009; Song et al., 2011). In this work, the minimum size was set to 10.

A stratified 10-fold cross-validation procedure was used to validate the model created by applying the J48 to the entire dataset (*N* = 2205). It involves randomly splitting the entire dataset into 10 approximately equally sized subsets, where the instances of each group of the class variable are represented in approximately the same proportions as in the entire dataset (i.e., the subsets are as similar as possible to the entire dataset with respect to the class variable). Each subset is in turn used for testing the model, and the remaining nine are used together for training it (i.e., nine-tenths of the dataset are used for training the model, and one-tenth is used for testing it). The procedure is repeated 10 times so that, in the end, every instance (i.e., every respondent) has been used exactly once for testing (Witten et al., 2016). The 10-fold cross-validation procedure is highly recommended to assess the generalizability of a model (Bock and Gough, 2003; Bouckaert, 2003; Martin and Hirschberg, 1996) and is one of the best options to obtain accurate estimates (Witten et al., 2016).

To show the advantages associated with the use of an ML-based approach, its sensitivity (i.e., true positive/(true positive + false negative)), specificity (i.e., true negative/(true negative + false positive)), accuracy (i.e., (true positive + true negative)/total cases), positive predictive value (i.e., true positive/(true positive + false positive)), and negative predictive value (i.e., true negative/(true negative + false negative)) are compared with those of two traditional methods, namely the DSM algorithm and the ROC-based cut-off scores. Two ROC-based cut-off scores were considered: the cut-off score suggested in the literature (i.e., $\geq 10$) and a cut-off score calculated on this specific dataset (*N* = 2205). The latter cut-off score was identified using the package "pROC" (Robin et al., 2011) for the R environment for statistical computing (R Core Team, 2018). The analysis was run using the independent categorical diagnosis (i.e., being currently diagnosed with or treated for MDD) and the PHQ-9 sum score. The score maximizing sensitivity and specificity was selected as the cut-off score. The performance of the cut-off score calculated on this dataset is the best that can be obtained using a ROC-based cut-off score. This would be analogous to training and testing an ML algorithm on the same dataset (unlike cross-validation, which trains and tests the algorithm on different subsets of the dataset). In both cases, overfitting occurs. The results of the DT trained and tested on the same dataset (*N* = 2205; the DT is the same one validated with the 10-fold cross-validation procedure) are compared with those of the cut-off score calculated on the same dataset because they represent a fairer comparison between the two methods.

The accuracy of each model was compared with the no-information rate, which is the proportion of the largest class (e.g., Hastie et al., 2009). A one-tailed binomial test was used to determine whether the accuracies were larger than the no-information rate (i.e., 0.527, which is the proportion of individuals belonging to the nonclinical population). Significant results ($p < .05$) indicate that model predictions are unlikely to be the result of chance.

The McNemar test was computed to compare the performance of the DT algorithm with those of the other considered methods (i.e., the two ROC-based cut-off scores and the DSM algorithm). The McNemar test statistic has a $\chi^2$ distribution with 1 degree of freedom.

## 3. Results

The mean of the PHQ-9 sum score was 6.96 (*SD* = 6.46) for the 1163 individuals from the nonclinical population and 14.52 (*SD* = 7.53) for the 1667 individuals with MDD. The study and its main results are illustrated in Fig. 1.

Fig. 2 depicts the DT developed by applying the J48 algorithm to the data. Item 3 ("Trouble falling asleep, staying asleep, or sleeping too much") was placed at the root node, meaning that it was the most effective in classifying individuals. Specifically, a response $> 1$ or $\leq 1$ to this item allowed for dividing all individuals into two large groups, containing 1007 and 1198 individuals, respectively. Among the individuals who provided a response $> 1$ to Item 3, Item 1 ("Little interest or pleasure in doing things"; splitting rule $> 1$ vs. $\leq 1$) was the item with

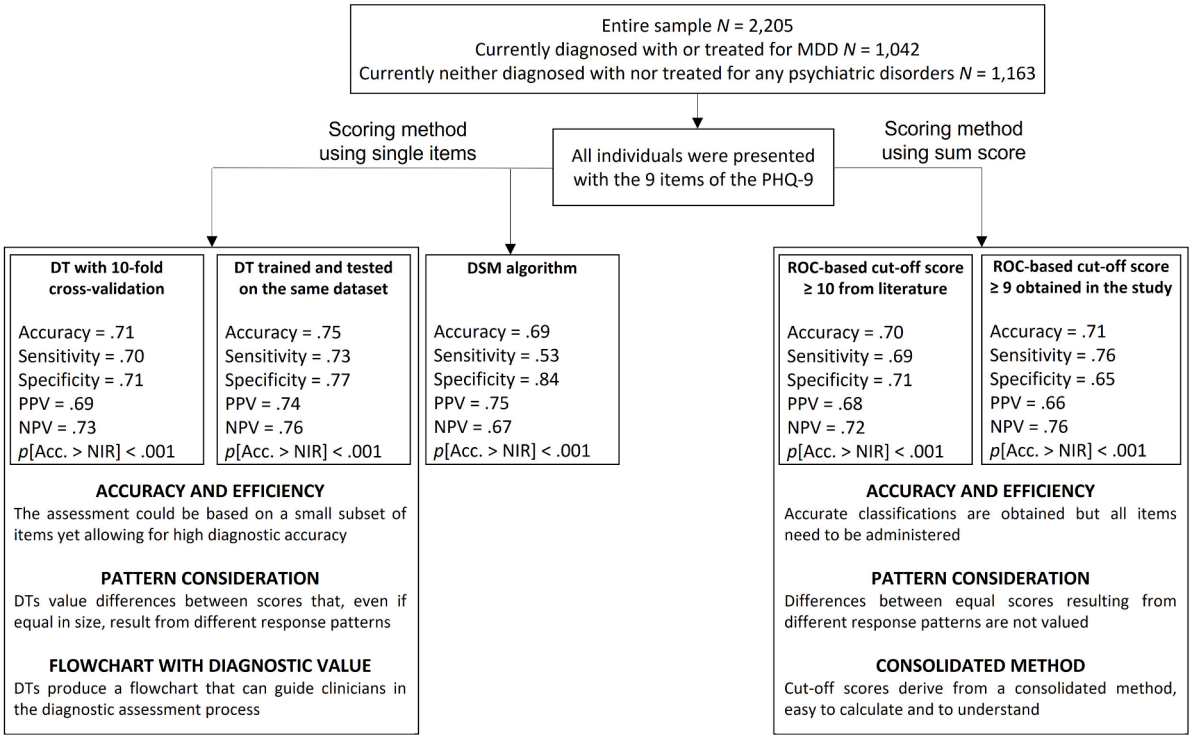


**Fig. 1.** Flowchart illustrating the study and its main results
*Note*. DT = decision tree; PPV = positive predictive value; NPV = negative predictive value; NIR = no-information rate (i.e., 0.527). The DT tested and trained on the same dataset was significantly more accurate than the cut-off score ≥ 10 ($\chi^2 = 28.82$, $p < .001$), the cut-off score ≥ 9 ($\chi^2 = 28.01$, $p < .001$), and the DSM algorithm ($\chi^2 = 36.33$, $p < .001$).

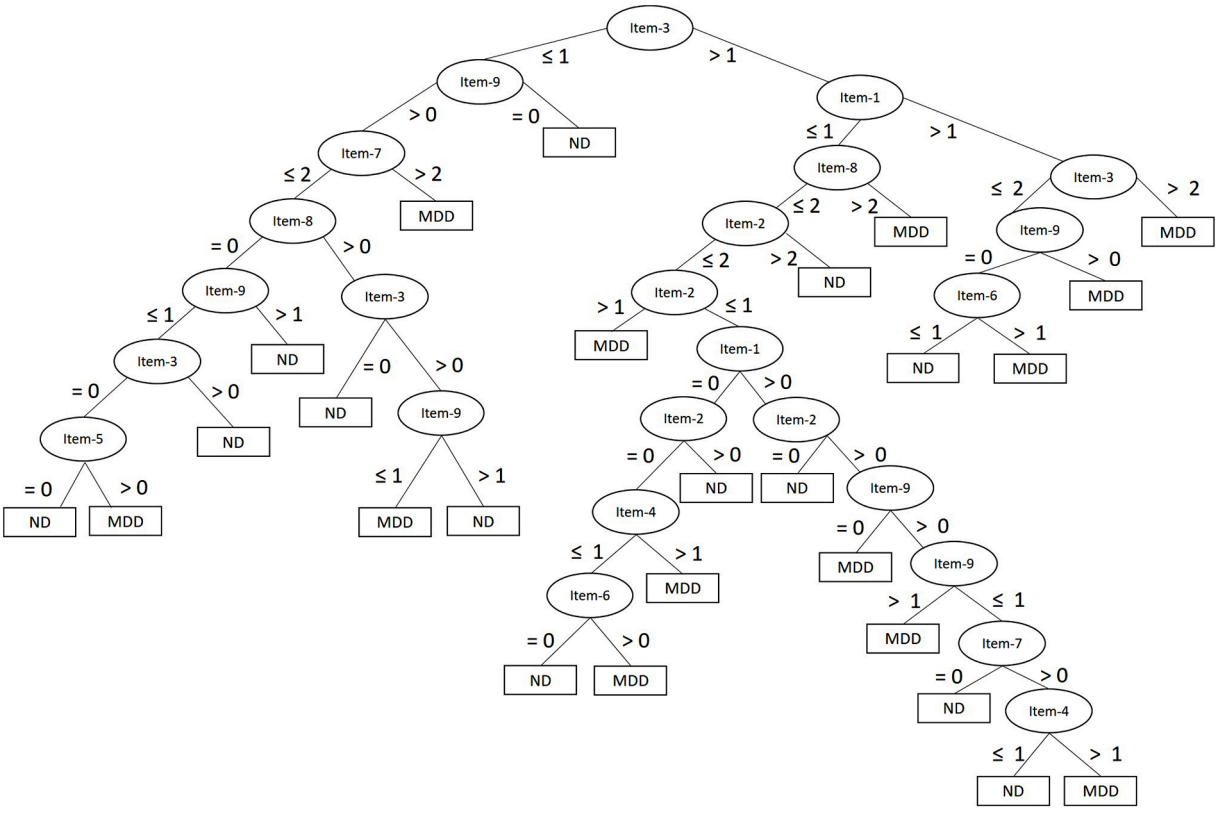


**Fig. 2.** Decision tree obtained running the J48 algorithm on the entire dataset ($N = 2205$)
*Note*. MDD = The algorithm classified the individuals falling into the branch as having major depressive disorder; ND = The algorithm classified the individuals falling into the branch as not having the diagnosis.

the greatest capability to differentiate individuals. Conversely, among the individuals who provided a response $\leq 1$ to Item 3, the item with the greatest discriminating power was Item 9 ("Thoughts that you would be better off dead or of hurting yourself in some way"; splitting rule $> 0$ vs. 0).

Interestingly, if the PHQ-9 items had been administered in a personalized way according to the process outlined by the DT structure, the respondents would have been classified as having or not having MDD using fewer than the nine available items. For instance, following the structure of the DT represented in Fig. 2, a respondent obtaining a score $\leq 1$ on Item 3 and a score of 0 on Item 9 could have been classified as not having MDD through the administration of these two items only. However, if the respondent had obtained a score $\leq 1$ on Item 3 and a score $> 0$ on Item 9, the administration of at least one more item would have been required to classify that respondent (e.g., it would have been necessary to observe a score $> 2$ on Item 7 to classify the respondent as having MDD). Following the procedure outlined by the DT, all respondents can be classified using a minimum of 2 and a maximum of 7 items (on average, 2.99 items out of 9, $SD = 1.35$; further details can be found Fig. S2 in the supplementary materials).

Table 1 shows the confusion matrices for all the considered methods, and Fig. 1 shows the accuracy, sensitivity, specificity, positive predictive value, and negative predictive value computed for each method (the ROC curves are shown in Fig. 3, and additional results can be found in Fig. S3 in the supplementary materials). For all the models, accuracies were significantly larger than the no-information rate, suggesting that the model predictions were unlikely to be the result of chance. The DT resulted in noticeable values for the five indicators, both when it was trained and tested on the same dataset and when the 10-fold cross-validation procedure was used. The results of the DT algorithm are compared with those of the DSM algorithm and two ROC-based cut-off scores, one recommended in the literature ($\geq 10$) and the other obtained from the sample under consideration ($\geq 9$). In general, the performance of the DT evaluated with cross-validation was analogous to that of the cut-off score $\geq 10$ (the DT showed slightly higher accuracy, sensitivity, and positive predictive value). Compared with the other methods, the cut-off score $\geq 9$ resulted in the largest sensitivity but the smallest specificity and positive predictive value. It should be noted that this cut-off score was specifically calculated on this sample and, thus, it represents the best result that can be obtained using a ROC-based cut-off score. This would be analogous to training and testing an ML algorithm on the same dataset (i.e., without cross-validation). Looking at the performance of the DT trained and tested on the same dataset, it can be noted that it exceeded that of the cut-off score $\geq 9$ (except for sensitivity, which was slightly higher for the cut-off score $\geq 9$). The DSM algorithm outperformed all the other methods in specificity, but it showed the lowest sensitivity and negative predictive value.

According to the McNemar test, in the current sample ($N = 2205$), the DT trained and tested on the same dataset was significantly more accurate than the cut-off score $\geq 10$ ($\chi^2 = 28.82$, $p < .001$), the cut-off score $\geq 9$ ($\chi^2 = 28.01$, $p < .001$), and the DSM algorithm ($\chi^2 = 36.33$, $p < .001$).

The performance of all the considered methods was also evaluated on the 625 individuals reporting psychiatric disorders different from MDD who, although present in the total sample ($N = 2830$), were not included in the analysed dataset ($N = 2205$) to reduce noise. All these 625 individuals had to be classified as "non-diagnosis" because, even if they reported some psychiatric disease, this was not MDD. Interestingly, the DT algorithm correctly classified 56.6% of them, whereas the cut-off score methods correctly classified lower percentages (52.8% and 47.8%, for the cut-off scores $\geq 9$ and $\geq 10$, respectively).

## 4. Discussion

This work aimed to illustrate how ML-DT can be effectively used to score and interpret psychodiagnostic test data. The algorithm generates a flowchart that can guide clinicians in the diagnostic assessment process. Moreover, it allows for obtaining accurate classifications. In this study, it showed greater accuracy compared with all the other considered methods. It also exhibited larger sensitivity and negative predictive values than both the DSM algorithm and the cut-off score $\geq 10$, as well as larger specificity and positive predictive values than the cut-off score $\geq 9$.

DTs allow for clarifying and explaining differences between scores that, even if equal in magnitude, result from different patterns of item responses. In the analysed dataset, individuals with different psychiatric conditions (with or without MDD) obtained the same PHQ-9 sum score from different response patterns. Inevitably, the cut-off score methods do not allow for appreciating and valuing the differences between these individuals and always lead to the same (sometimes correct and sometimes incorrect) conclusion. For example, two individuals in the sample obtained the same sum score of 12. The first person was diagnosed with MDD and responded with 2 on Items 2, 3, 4, and 9; 1 on Item 1; and 3 on Item 6. The second person belonged to the nonclinical population and responded with 3 on Items 3, 4, and 5; 1 on Item 2; and 2 on Item 7. Even though these two individuals obtained the same sum score, the DT algorithm correctly classified them in a different way. Conversely, the two cut-off scores and the DSM algorithm failed to correctly classify the second person as not having the diagnosis.

Finally, DTs contribute to the efficiency of the assessment. Individuals can be classified by presenting them with a limited number of items, selected in a personalized way. In the case of the PHQ-9, the relevance of this aspect is not marked because it is a short test. However, psychodiagnostic tests are often far longer, and obtaining accurate diagnoses using a limited number of items would be of great importance. In this regard, it is expected that, in the future, ML and some of its specific applications will become a crucial resource for the improvement and development of personalized and abbreviated assessment tools (Gonzalez, 2021b).

This work illustrated the application of ML-DT on data from a single psychodiagnostic test. Although there is no reason to believe that the proposed approach should not be a viable option for other psychodiagnostic tests as well, future studies are needed to generalize its usefulness and effectiveness.

## Funding statement

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

## Contributorship statement

Daiana Colledani planned the study, conducted the literature search, analysed and interpreted the data, and wrote the article. Pasquale

**Table 1**
Confusion matrices for cut-off scores $\geq 10$ and $\geq 9$, the DSM algorithm, and the decision tree (DT) algorithm.

| | Cut-off score $\geq 10$ | | Cut-off score $\geq 9$ | | DSM algorithm | | DT with 10-fold cross-validation | | DT trained and tested on the same dataset | |
|---|---|---|---|---|---|---|---|---|---|---|
| Diagnosis | Positive | Negative | Positive | Negative | Positive | Negative | Positive | Negative | Positive | Negative |
| Positive | 722 | 320 | 797 | 245 | 551 | 491 | 727 | 315 | 759 | 283 |
| Negative | 333 | 830 | 402 | 761 | 186 | 977 | 332 | 831 | 267 | 896 |

*Note*. Diagnosis = Variable indicating whether the respondent had major depressive disorder (Positive) or not (Negative).

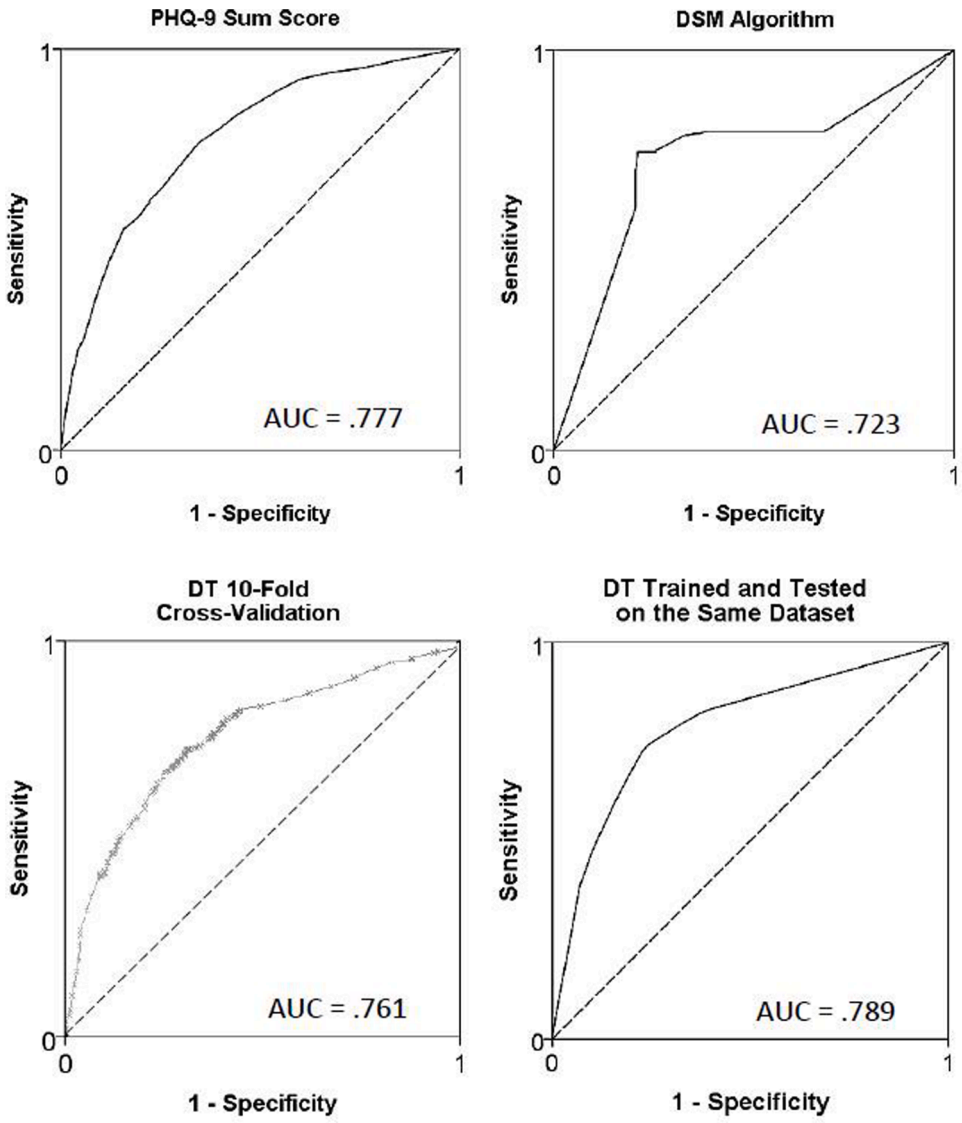


**Fig. 3.** Receiver operating characteristic (ROC) curves for the Patient Health Questionnaire-9 (PHQ-9) sum score, DSM algorithm, decision tree (DT) with 10-fold cross-validation, and DT trained and tested on the same dataset ($N = 2205$)
*Note*. AUC = area under the curve.

Anselmi analysed and interpreted the data, and wrote the article. Egidio Robusto interpreted the data and wrote the article.

Daiana Colledani is responsible for the overall content as guarantor.

## Declaration of Competing Interest

The authors declare no competing interests.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2023.115127.

## References

Baştanlar, Y., Özuysal, M., 2014. Introduction to machine learning. miRNomics: MicroRNA Biology and Computational Analysis. Springer Science Business Media, New York, pp. 105–128.

Battineni, G., Sagaro, G.G., Chinatalapudi, N., Amenta, F., 2020. Applications of machine learning predictive models in the chronic disease diagnosis. J. Pers. Med. 10 (2), 21. https://doi.org/10.3390/jpm10020021.

Bock, J.R., Gough, D.A., 2003. Whole-proteome interaction mining. Bioinformatics 19 (1), 125–135. https://doi.org/10.1093/bioinformatics/19.1.125.

Bouckaert, R.R., 2003. Choosing between two learning algorithms based on calibrated tests. In: Proceedings, Twentieth International Conference on Machine Learning.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (2017). Classification and regression trees. In *Classification and Regression Trees*. 10.1201/9781315139470.

Carter, J.V., Pan, J., Rai, S.N., Galandiuk, S., 2016. ROC-ing along: evaluation and interpretation of receiver operating characteristic curves. Surgery 59 (6), 1638–1645. https://doi.org/10.1016/j.surg.2015.12.029.

Criminisi, A., Shotton, J., Konukoglu, E., 2012. Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Found. Trends® Comput. Graphic. Vis. 7 (2–3), 81–227.

Dekker, G.W., Pechenizkiy, M., Vleeshouwers, J.M., 2009. Predicting students drop out: a case study. In: EDM'09 - Educational Data Mining 2009: 2nd International Conference on Educational Data Mining.

Doi, S., Ito, M., Takebayashi, Y., Muramatsu, K., Horikoshi, M., 2018. Factorial validity and invariance of the Patient Health Questionnaire (PHQ)-9 among clinical and non-clinical populations. PLoS ONE 13 (7), e0199235. https://doi.org/10.1371/journal.pone.0199235.

Gonzalez, O., 2021a. Psychometric and machine learning approaches for diagnostic assessment and tests of individual classification. Psychol. Methods 26 (2), 236. https://doi.org/10.1037/met0000317.

Gonzalez, O., 2021b. Psychometric and machine learning approaches to reduce the length of scales. Multivariate Behav. Res. 56 (6), 903–919. https://doi.org/10.1080/00273171.2020.1781585.

Gray, R.M. (2011). Entropy and information theory. In *Entropy and Information Theory*. 10.1007/978-1-4419-7970-4.

Gupta, B., Rawat, A., Jain, A., Arora, A., Dhami, N., 2017. Analysis of various decision tree algorithms for classification in data mining. Int. J. Comput. Appl. 163 (8), 15–19. https://doi.org/10.5120/ijca2017913660.

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Vol. 2. Springer.

Higa, A., 2018. Diagnosis of breast cancer using decision tree and artificial neural network algorithms. Int. J. Comput. Appl. Technol. Res. 1 (7), 23–27. https://doi.org/10.7753/ijcatr0701.1004.

Ito, M., Bentley, K.H., Oe, Y., Nakajima, S., Fujisato, H., Kato, N., Miyamae, M., Kanie, A., Horikoshi, M., Barlow, D.H., 2015. Assessing depression related severity and functional impairment(warning) the overall depression severity and Impairment Scale (ODSIS). PLoS ONE 10 (4), e0122969. https://doi.org/10.1371/journal.pone.0122969.

Kroencke, K., Spitzer, R.L., Williams, J.B., Kroenke, K., Spitzer, R.L., Williams, J.B., 2001. The PHQ-9: validity of a brief depression severity measure [Electronic version]. J. Gen. Intern. Med. 16 (9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x.

Lin, J., 2001. Feature extraction of machine sound using wavelet and its application in fault diagnosis. NDT E Int. 34 (1), 25–30. https://doi.org/10.1016/S0963-8695(00)00025-6.

Manea, L., Gilbody, S., McMillan, D., 2012. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. CMAJ 184 (3), E191–E196. https://doi.org/10.1503/cmaj.110829.

Martin, J.K. & Hirschberg, D.S. (1996). Small sample statistics for classification error rates I: error rate measurements. *Technical Report*.

Muramatsu, K., Miyaoka, H., Kamijima, K., Muramatsu, Y., Tanaka, Y., Hosaka, M., … & Shimizu, E. (2018). Performance of the Japanese version of the Patient Health Questionnaire-9 (J-PHQ-9) for depression in primary care. *Gen. Hosp. Psychiatry*, 52, 64–69. 10.1016/j.genhosppsych.2018.03.007.

Muramatsu, K., Kamijima, K., Yoshida, M., Otsubo, T., Miyaoka, H., Muramatsu, Y., Gejyo, F., 2007. The patient health questionnaire, Japanese version: validity according to the mini-international neuropsychiatry interview-plus. Psychol. Rep. 101 (3), 952–960. https://doi.org/10.2466/PR0.101.3.952-960.

Prabhakar, S., Mohanty, A.R., Sekhar, A.S., 2002. Application of discrete wavelet transform for detection of ball bearing race faults. Tribol. Int. 35 (12), 793–800. https://doi.org/10.1016/S0301-679X(02)00063-4.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann.

R Core Team. (2018). R: a language and environment for statistical computing [Computer software]. http://www.Rproject.org/.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., Müller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform. 12 (1), 1–8. http://www.biomedcentral.com/1471-2105/12/77.

Song, E., Huang, D., Ma, G., Hung, C.C., 2011. Semi-supervised multi-class Adaboost by exploiting unlabeled data. Expert Syst Appl 38 (6), 6720–6726. https://doi.org/10.1016/j.eswa.2010.11.062.

Spitzer, R.L., Kroenke, K., Williams, J.B., 1999. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary care evaluation of mental disorders. Patient health questionnaire. JAMA 282 (18), 1737–1744. https://doi.org/10.1001/jama.282.18.1737.

Sugumaran, V., Muralidharan, V., Ramachandran, K.I., 2007. Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing. Mech. Syst. Signal Process 21 (2), 930–942. https://doi.org/10.1016/j.ymssp.2006.05.004.

Uğuz, H., 2011. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowl. Based Syst. 24 (7), 1024–1032. https://doi.org/10.1016/j.knosys.2011.04.014.

Witten, I.H., Frank, E., Hall, M.A. & Pal, C.J. (2016). Data mining: practical machine learning tools and techniques. In *Data Mining: Practical Machine Learning Tools and Techniques*. 10.1016/c2009-0-19715-5.

Yarkoni, T., Westfall, J., 2017. Choosing prediction over explanation in psychology: lessons from machine learning. Perspect. Psychol. Sci. 12 (6), 1100–1122. https://doi.org/10.1177/1745691617693393.

Zhao, Y., Zhang, Y., 2008. Comparison of decision tree methods for finding active objects. Adv. Space Res. 41 (12), 1955–1959. https://doi.org/10.1016/j.asr.2007.07.020.

Zhou, X.H., Obuchowski, N.A. & McClish, D.K. (2011). Statistical methods in diagnostic medicine. In *Statistical Methods in Diagnostic Medicine*.