Information Systems 127 (2025) 102465

Contents lists available at ScienceDirect

Information Systems

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

Using AI explainable models and handwriting/drawing tasks for psychological well-being

Francesco Prinzi [a,*], Pietro Barbiero [b], Claudia Greco [c], Terry Amorese [c], Gennaro Cordasco [c,d], Pietro Liò [e], Salvatore Vitabile [a], Anna Esposito [c,d]

[a] *Department of Biomedicine, Neuroscience and Advanced Diagnostics (BiND), University of Palermo, Palermo, Italy*
[b] *Università della Svizzera Italiana, Lugano, Switzerland*
[c] *Department of Psychology, Università degli Studi della Campania "Luigi Vanvitelli", Caserta, Italy*
[d] *International Institute for Advanced Scientific Studies (IIASS), Vietri sul Mare, Italy*
[e] *Department of Computer Science and Technology, University of Cambridge, Cambridge, UK*

ARTICLE INFO

Keywords:
Psychological well-being
Explainable AI
Machine learning
Handwriting features

ABSTRACT

This study addresses the increasing threat to Psychological Well-Being (PWB) posed by Depression, Anxiety, and Stress conditions. Machine learning methods have shown promising results for several psychological conditions. However, the lack of transparency in existing models impedes practical application. The study aims to develop explainable machine learning models for depression, anxiety and stress prediction, focusing on features extracted from tasks involving handwriting and drawing.

Two hundred patients completed the Depression, Anxiety, and Stress Scale (DASS-21) and performed seven tasks related to handwriting and drawing. Extracted features, encompassing pressure, stroke pattern, time, space, and pen inclination, were used to train the explainable-by-design Entropy-based Logic Explained Network (e-LEN) model, employing first-order logic rules for explanation. Performance comparison was performed with XGBoost, enhanced by the SHAP explanation method.

The trained models achieved notable accuracy in predicting depression (0.749 ±0.089), anxiety (0.721 ±0.088), and stress (0.761 ±0.086) through 10-fold cross-validation (repeated 20 times). The e-LEN model's logic rules facilitated clinical validation, uncovering correlations with existing clinical literature. While performance remained consistent for depression and anxiety on an independent test dataset, a slight degradation was observed for stress prediction in the test task.

1. Introduction

In recent years it has been observed that Psychological Well-Being (PWB) is threatened by three widely spread conditions, often closely connected, namely Depression, Anxiety, and Stress; in fact, it has been observed that low levels of PWB are highly associated with these mental disorders [1,2]. Moreover, Depression, Anxiety, and Stress appear to be connected for two main reasons: first, stressful events could result in negative affective states as feelings of anxiety and depression [3]; second, even though anxiety and depression are considered two separate classes of disorders, they often occur in comorbidity and share common symptoms [4].

Depression is also known as Major Depression, Major Depressive Disorder, or Clinical Depression. According to the World Health Organization (2021), approximately 280 million people live with depression, making it one of the most common illnesses worldwide; it affects approximately 5% of adults, and its incidence increases with increasing age, in fact around 5.7% of adults aged 60 years and older suffer from depression.

Depression is characterized by both physical and purely psychological/emotional symptoms, such as changes in body weight, sleep patterns and psychomotor changes, depressed mood, decreased interest in all activities, feelings of worthlessness, reduced ability to concentrate and recurring thoughts of death. Anxiety can be considered as a complex emotion characterized by fear, apprehension and worry. However, anxiety must be distinguished from anxiety disorders, characterized by excessive fear and anxiety, and related behavioral disorders [5]. In 2019, Around 301 million people were reported living with an anxiety

[*] Corresponding author.
E-mail addresses: francesco.prinzi@unipa.it (F. Prinzi), barbip@usi.ch (P. Barbiero), claudia.greco@unicampania.it (C. Greco), terry.amorese@unicampania.it (T. Amorese), gennaro.cordasco@unicampania.it (G. Cordasco), pl219@cam.ac.uk (P. Liò), salvatore.vitabile@unipa.it (S. Vitabile), anna.esposito@unicampania.it (A. Esposito).

https://doi.org/10.1016/j.is.2024.102465
Received 31 July 2024; Received in revised form 23 September 2024; Accepted 25 September 2024
Available online 28 September 2024

disorder, making it among the most widespread in the world [6]. Stress occurs when a person perceives that the surrounding environment demands more than his/her adaptive capacity [7]; the exposure to stressful events can lead to various clinical conditions described in the Diagnostic and Statistical Manual of Mental Disorders (2013) and included in the category of Disorders Related to Traumatic and Stressful Events.

Unfortunately, the necessity to provide people living with the above-mentioned mental health disorders adequate and effective care is often hindered by factors such as poor financial funding for mental health care services, the shortage of properly trained professionals, and the social stigma associated with mental disorders [8]. In addition to the difficulty in providing adequate treatments, a further issue concerns the diagnosis of these disorders, also made difficult by the fact that, as highlighted above, Depression, Anxiety and Stress are characterized by symptoms of various kinds, not only cognitive or emotional, but by behavioral symptoms as well. The lack of established and objective criteria for assessing psychiatric disorders is a major factor contributing to incorrect or delayed diagnoses.

Currently, anxiety and depression diagnosis rely heavily on subjective methods such as clinical interviews and self-reported questionnaires. However, the inclusion within the diagnostic process of objective behavioral signals, such as handwriting analysis, could offer unbiased information to support it [9]. Efforts have been made to leverage new technologies for their early detection and identification, aiming to reduce associated costs. A specific area of research is dedicated to examining the behavioral symptoms linked to these mental issues, particularly through the analysis of handwriting and drawing patterns. Statistical approaches have proven effective in discerning various individual characteristics, including indicators of negative moods, and depressive states [10,11].

Machine learning models have proven to be a valuable tool to support the diagnostic process, but a lack of transparency hinders their use in practice [12,13]. Some examples have shown that AI-based models are often used by clinicians to inform decision-making and can improve conventional diagnostic capabilities [14]. However, it is essential to properly trust an AI-supported medical decision because a misdiagnosis can significantly impact patients [15]. This makes systems explainability not merely a technological issue, but also an ethical, legal, and social issue [16]. In some clinical situations, global and local model explanations are mandatory prerequisites for validating and justifying decisions [17]. Several *post-hoc* explanation methods have been proposed to elucidate the well-defined black-box algorithms, including tree ensemble and neural networks. These methods aim to provide explanations after model training. Although some of these methods are widely used, they do not provide the relationship of these features to perform predictions. Specifically, traditional methods focus on assessing the significance of individual features for prediction on a global and local scale. However, they do not discern specific relationships between multiple features that may contribute to the prediction. To solve this issue, the entropy-based Logic Explained Network (e-LEN) [18] was proposed. The e-LEN model incorporates constraints in both the architecture and the learning process, allowing for the emergence of simple rule-based formula explanations. The e-LENs seek to exploit the advantages of rule-based expert systems, which employ a first-order formalism to achieve explainable decision-making, and the advantages of neural networks, which excel at discovering relationships within data.

In this work, machine learning methods are employed for the prediction of mental health conditions such as depression, anxiety, and stress. Features extracted from handwriting and drawing tasks are used for model training. Considering the intelligible nature of handwriting and drawing features and the implementation of explainable AI methods, the aim is to develop both accurate and explainable systems [19]. The XGBoost model was compared with the explainable-by-design e-LEN, in an attempt to overcome the trade-off between explainability and accuracy. Introducing a model explanation through logic rules represents a new method for validating data-driven systems, allowing the analysis of the interactions between the involved features.

Table 1
Class distributions for training and test sets, for depression, anxiety, and stress conditions.

	Total		Training		Test	
	Healty	Sub-clinical	Healty	Sub-clinical	Healty	Sub-clinical
Depressed	105	95	89	81	16	14
Anxiety	104	96	88	82	16	14
Stress	99	101	84	86	15	15

The remaining of the paper is organized as follows: Section 2 describes the conducted study, detailing the dataset employed, each step of the processing pipeline, the machine learning models used and the explainability methods; Section 3 illustrates the obtained experimental results, concerning each specific built-up models and the achieved explainability; Section 4 discuss the obtained results, highlighting the clinical viewpoint of the findings; finally, conclusions are provided in Section 5.

2. Materials and methods

2.1. Dataset description

The dataset is composed of two-hundred participants (M = 102; F = 98), aged between 15 and 75 years old (Mean = 26.24; S.D. = 0.6). Each row in the dataset corresponds to a single participant's data. To preserve participants' privacy, they were assigned an ID code. For each of them, the dataset reports age, gender, three separate scores obtained from the depression, anxiety, and stress sub-scales of the DASS-21 questionnaire [20], and the extracted quantitative values of the considered handwriting and drawing features. Scores of each DASS-21 sub-scale were used to define, according to predefined thresholds, two balanced groups (healthy and sub-clinical) depending on participants' psychological well-being level: healthy participants are those who scored below the recommended cut-off points in each sub-scale of the DASS-21. Sub-clinical participants are those who scored above the threshold of the tool. The term "sub-clinical" refers to the absence of a formal medical diagnosis, but indicates a severity of psychological symptoms (measured with the DASS-21) that does not characterize the healthy population. The sub-sample composition is the following (also, see Table 1 for total, training, and test sub-sample distribution):

- Stress: The healthy group is composed of 99 participants (M = 52, F = 47; mean age = 25; SD = 0.24), while the sub-clinical group consists of 101 participants (M = 46, F = 55; mean age = 27.45; SD = 1.07).
- Anxiety: Healthy participants were 104 (M = 54, F = 50; age mean = 25.01; SD = 0.35). The sub-clinical group comprised 96 participants (M = 44, F = 52; age mean = 27.56; SD = 1.09).
- Depression: The healthy group consisted of 105 participants (M = 50, F = 55; age mean = 25.12; SD = 0.39), whereas the sub-clinical group was composed of 95 participants (M = 48, F = 47; age mean = 27.47; SD = 1.09).

The Venn diagram displayed in Fig. 1 reports the distribution of each sub-sample by also considering the frequencies overlapping among the three conditions.

All participants performed the same experimental protocol. In detail, they were asked to sit down in front of a laptop and a digitalized tablet located in a quiet room. The experimenter explained the experimental procedure, the research purposes, and all the other information about data treatment and confidentiality. Participants were informed about the experimental instruction, consisting of completing seven

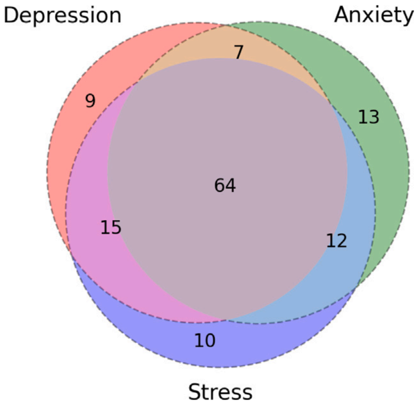
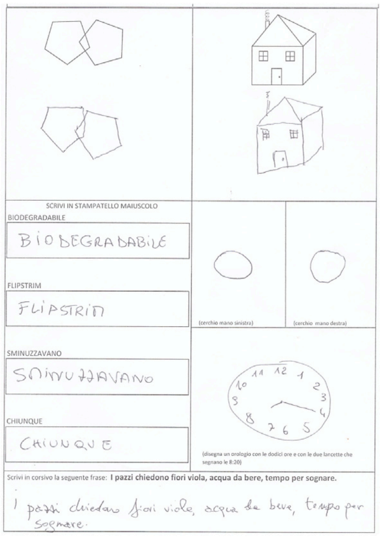


Fig. 1. Comorbidity distributions among the sub-clinical subjects. Healthy subjects were 70.



SCRIVI IN STAMPATELLO MAIUSCOLO

BIODEGRADABILE

BIODEGRADABILE

FLIPSTRIM

FLIPSTRIM

SMINUZZAVANO

SMINUZZAVANO

CHIUNQUE

CHIUNQUE

(cerchio mano sinistra)

(cerchio mano destra)

(disegna un orologio con le dodici ore e con le due lancette che segnano le 8:20)

Scrivi in corsivo la seguente frase: **I pazzi chiedono fiori viola, acqua da bere, tempo per sognare.**

I pazzi chiedono fiori viole, acqua da beve, tempo per sognare.

Fig. 2. Handwriting and drawing tasks performed.

different handwriting and drawing tasks (see Fig. 2) on a paper-sheet placed on a WACOM INTUOS PRO series 4 digital tablet. To preserve the naturalness of the handwriting process, participants used an Intuos Inkpen which works as a common pen, so that they could visualize the traits while they carried out the tasks. This technology allows to collect three different types of strokes depending on the pen status: *on-paper*; *in-air* (the pen is very close to the paper); *idle* (the stroke is not recorded but recognized by using time-stamps). Please note that the term “stroke” denotes the longest uninterrupted points series associated with the same pen status (either on-paper or in-air). Hardware specifics and further information about the data-saving process can be found in [9,11].

Concerning the handwriting and drawing features, seventeen quantitative measurements about five different categories were considered for the current study. Fig. 3 report their classification and description.

2.2. Problem definition

In this paper, we deal with three binary classification problems, one for each condition: Depression, Anxiety, and Stress. For each condition, participants were divided into two equally divided groups (healthy and sub-clinical), according to their DASS-21 score, associated with the considered condition. Then for each condition (or classification job), the goal is to find a model that maps the above-described handwriting and drawing features, to the identified groups. Fig. 4 shows the general workflow.

2.3. Feature preprocessing

2.3.1. Gender harmonization

For several disciplines, such as psychology, historical document analysis, and handwriting biometrics, the ability to identify a writer’s gender based on his or her handwriting is extremely important [21]. Recent applications have shown promising results for gender classification from handwriting [22], proving how certain features can be highly discriminated. For example, in [23] features of space and time have been introduced, particularly demonstrating that those of pressure and irregularity are discriminative. As a result, the distributions of features associated with the two genders can differ significantly and may represent a confounding factor for machine learning models.

To test this phenomenon in our case, the statistical significance of the extracted features concerning gender was evaluated. The Mann–Whitney test was used to compare the distributions of the male and female groups. Fig. 5 shows the number of features statistically significant for gender, in orange considering $p < 0.1$ and in gray $p < 0.5$ Specifically, a strong correlation was observed between time and pressure features with gender. Consequently, the dataset was harmonized as a proactive measure to reduce gender-related variability. The Combat method was applied for feature harmonization [24,25]. ComBat was developed to adjust for inter-site variability in the data while preserving variability related to the variables of interest [26]. It is a location-scale method that estimates the location-scale parameters (mean and variance) of each cohort and aligns the distributions using empirical Bayes shrinkage [27]. ComBat was originally developed to align distributions data for genetic studies [26], as well as widely used to reduce batch-effect in imaging [28] and many other scenarios [27].

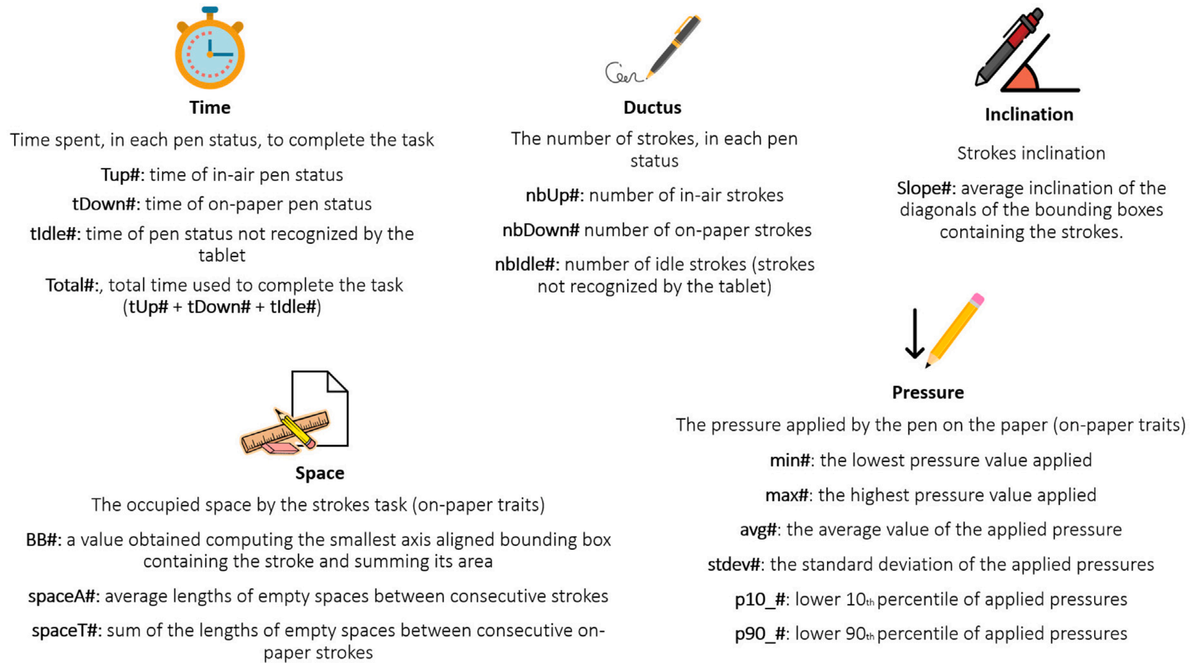
2.3.2. Feature selection

A total of 119 handwriting and drawing features (17 features × 7 tasks) were collected. Firstly, the correlated features using the Spearman Correlation coefficient were discarded, considering $|\rho| < 0.9$ as the threshold. The Spearman test is an unsupervised statistical test and therefore is not dependent on the condition analyzed. A total of 88 features passed the specified threshold. Then, the Sequential Forward Selector (SFS) [29] was employed for feature selection considering the 88 features selected. The XGBoost model was used as a classifier, a stratified 10-fold cross-validation repeated 20 times was used for SFS evaluation, and accuracy was the metric to maximize. The same procedure was repeated for depression, anxiety, and stress conditions.

2.4. Models training

2.4.1. Baseline ML methods

Shallow learning methods are considered the baseline for small tabular dataset analysis. Decision trees are inherently interpretable due to their transparent structure, which involves a series of binary decisions based on features leading to a clear and intuitive representation. However, their interpretability can come at the cost of accuracy. Decision trees may oversimplify complex relationships in the data, leading to a lack of precision when capturing intricate patterns. Tree ensemble algorithms, such as XGBoost, have proven effective for classification in small datasets [30,31]. In addition, they are well established as a standard tool to process tabular data and, in this case, showed improved performance over deep architectures [32].



Time

Time spent, in each pen status, to complete the task

Tup#: time of in-air pen status

tDown#: time of on-paper pen status

tIdle#: time of pen status not recognized by the tablet

Total#:, total time used to complete the task (tUp# + tDown# + tIdle#)

Ductus

The number of strokes, in each pen status

nbUp#: number of in-air strokes

nbDown# number of on-paper strokes

nbIdle#: number of idle strokes (strokes not recognized by the tablet)

Inclination

Strokes inclination

Slope#: average inclination of the diagonals of the bounding boxes containing the strokes.

Space

The occupied space by the strokes task (on-paper traits)

BB#: a value obtained computing the smallest axis aligned bounding box containing the stroke and summing its area

spaceA#: average lengths of empty spaces between consecutive strokes

spaceT#: sum of the lengths of empty spaces between consecutive on-paper strokes

Pressure

The pressure applied by the pen on the paper (on-paper traits)

min#: the lowest pressure value applied

max#: the highest pressure value applied

avg#: the average value of the applied pressure

stdev#: the standard deviation of the applied pressures

p10_#: lower 10th percentile of applied pressures

p90_#: lower 90th percentile of applied pressures

Fig. 3. Features description and classification.

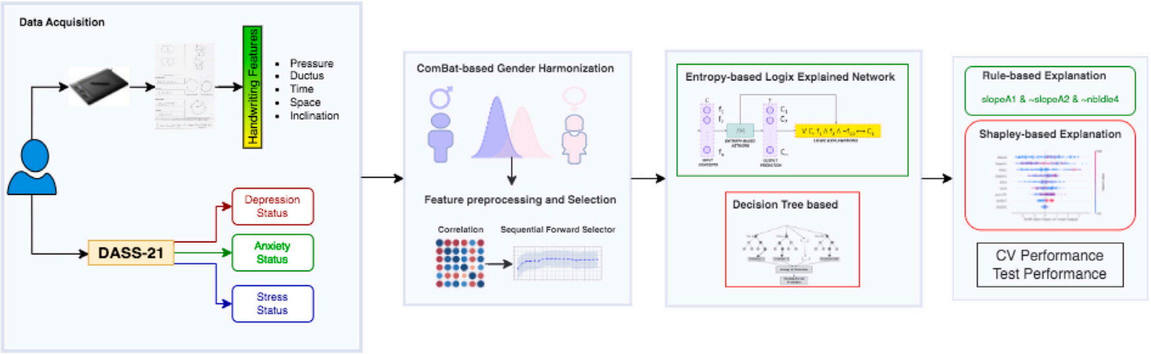


Fig. 4. General Workflow. Each patient underwent evaluation using the DASS-21 questionnaire to assess their depression, anxiety, and stress conditions. Additionally, the same patients completed seven handwriting and drawing tasks, from which handwriting and drawing features were extracted. The extracted features underwent preprocessing and feature selection steps and models were trained using both decision tree-based methods and e-LEN method. Finally, the two types of explanations provided by the conventional SHAP method and the rule-based explanations provided by e-LEN were compared.

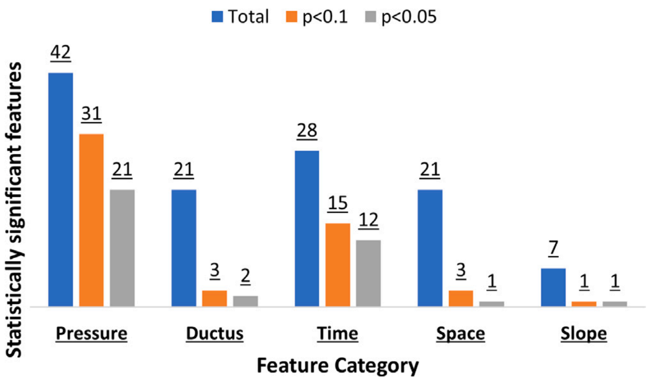


Fig. 5. Statistically significant number of features related to gender.

2.4.2. Entropy-based logic explained network

The primary objective of this study is to introduce an explainable model that facilitates the examination of key features and their interrelationships. While conventional neural networks boast state-of-the-art performance, they lack intrinsic explainability and necessitate post-hoc algorithms for explanation.

To address this limitation, the Entropy-based Logic Explained Network was proposed as an explainable-by-design algorithm, aiming to deliver both high performance and interpretability in neural networks. The e-LEN model adopts the formalism of First-Order Logic to provide the most important concepts for prediction. The Entropy layer, a pivotal component, is designed to compute: (i) the embeddings h_i (as to any linear layer), and (ii) a truth table T_i clarifying how the network leverages concepts to make predictions for the ith target class.

Consequently, the model's loss function takes into consideration the maximization of the concept of entropy alongside the minimization of a standard loss function for supervised learning. Specifically, the formalization of the entropy concept is expressed as:

$$\mathcal{H}(\alpha^i) = -\sum_{j=1}^{k} \alpha^i_j \log \alpha^i_j \tag{1}$$

where a_j^i represents the importance of each concept j and class i. Consequently, the complete loss function is:

$$\mathcal{L}(f, y, a_1, \dots, a_r) = L(f, y) + \lambda \sum_{i=1}^{r} \mathcal{H}(\alpha^i) \tag{2}$$

In this context, f denotes the predictions for class membership, y corresponds to the actual class, a_j signifies the significance of the j concepts for class i, and λ serves as the hyperparameter for balancing the importance of low-entropy solutions in the loss function. The e-LEN model ultimately furnishes both global and local explanations. The global explanation encompasses the most prevalent predicates associated with each class. Consequently, distinct sets of predicates are supplied for healthy and sub-clinical patients. Each predicate is linked to others using the ∧ (and) operator to formulate a rule, and all rules are interconnected by the ∨ (or) operator. Ultimately, the comprehensive explanation for each class is represented by the collection of all predicates.

2.5. Evaluation protocol and test

The dataset was divided into training and test subsets. The test dataset was used only for the final model evaluation and was not involved in training, preprocessing and feature selection steps. In the training set, a 10-fold cross-validation was repeated 20 times to achieve a fair performance estimation of the three employed classifiers. The best model in terms of accuracy found during cross-validation was selected for testing. Accuracy, Area Under the Receiver Operating Characteristic (AUROC), Sensitivity, Specificity, PPV, and NPV were computed to evaluate the models' performance.

2.6. Models explainability

The three classifiers implemented in this paper (e-LEN, XGB, and DT) are representative to discuss some considerations about the explainability and accuracy of the models. The highest model performance is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models [33]. This has created a huge dilemma in optimizing explainability or accuracy. In fact, in several papers, it is argued how accuracy should drive model development rather than accuracy [34–36]. Other works impose the use of transparent models [37].

A decision tree is recognized in the small set of existing interpretable models [38]. The paths from the root to the leaves of the decision tree represent the classification rules. A decision tree can be linearized into a set of decision rules with the if-then form [39]. In fact, rules-based models are recognized as interpretable models as well [12]. Their inherent interpretability makes it unnecessary to use Explainable AI methods to explain these trained models. Conversely, tree ensembles and neural networks are recognized as black-box and therefore require explanation. In the case of XGBoost, we employed the SHAP *post-hoc* explanation method for global explanation [40].

The e-LEN model tries to overcome this trade-off between accuracy and explainability. It is an explainable-by-design approach as it embeds additional constraints both in the architecture and in the learning process. This point of view is in contrast with *post-hoc* methods, which generally do not impose any constraint on classifiers [18]. For this reason, e-LEN does not require the use of some *post-hoc* explanation methods but learns during training the most important rules involved in classification. The logical rules are represented through the formalism of first-order logic, used in the implementation of rule-based systems widely recognized as transparent [12].

In addition to an explanation through the logical rules produced by the e-LEN model and the feature importance computed *via* SHAP for XGBoost, the complexity of the models was computed. For e-LEN, the complexity is computed by counting the number of predicates for the two classes. For the Decision Tree model, the complexity was the number of generated nodes during the training. Similar to the Decision Tree model, in XGBoost the complexity was computed considering the nodes generated by the trees added to the ensemble. Model complexity was calculated during cross-validation.

3. Experimental results

For each condition (depression, anxiety, and stress), a random hyperparameters search was performed. This phase was performed entirely within the cross-validation procedure. The test set was not involved in hyperparameter tuning or model selection. This ensured that the test set remained unseen until the final model evaluation. Specifically: for the e-LEN model, the number of linear layers was searched in the range of 1 to 5 layers, and the number of neurons per layer varied between 32 and 512. For decision-tree-based algorithms, the hyperparameter search focused on the number of estimators, with the range being explored between 20 and 500 estimators. Finally, for the depression and anxiety conditions, XGBoost and e-LEN were trained with the same hyperparameters, i.e., 100 estimators for XGBoost, while e-LEN was implemented considering one Entropy Layer with 400 neurons, followed by three linear layers with 400, 200, and 100 neurons, with the last layer composed of one neuron for classification. For stress detection, 50 estimators were used for XGBoost, while e-LEN was implemented considering one Entropy Layer with 500 neurons, followed by three linear layers with 500, 250, and 100 neurons, with the last layer composed of one neuron for classification. The ReLU was used as an activation function after each hidden layer and a binary cross entropy with logits as loss function. Adam was employed as optimizer. Features were normalized before model training.

3.1. Feature selected

The same preprocessing and feature selection protocol was applied for depression, anxiety, and stress conditions. In general, less than 19 features were selected for each condition, to allow a correct proportion between the number of samples and features used for model training [41]. Nineteen features were used for the depression condition, 15 for the anxiety condition and 18 for the stress condition. Fig. 6 shows the features selected for the three conditions. There are only two features selected simultaneously for the three conditions (nbUp5 and nbIdle5), both belonging to task 5. Furthermore, it appears that in terms of selected features, the depression condition overlaps significantly with both the anxiety and stress conditions. Conversely, the anxiety and stress conditions are quite different. Furthermore, it is quite clear that for the anxiety condition, the pressure features would be the most significant ones. The number of strokes in each pen status was the most important category for the depression condition. In general, for all conditions, features belonging to the 7 tasks and the 5 feature categories were equally informative.

3.2. Training and test performance

Tables 2–4 show the performances for the three models employed, i.e. e-LEN, XGB and DT, for depression, anxiety, and stress conditions, respectively. A 10-fold cross-validation was repeated 20 times to have a precise model performance estimation. Therefore, the performances calculated during the cross-validation procedure can be considered reliable. Furthermore, to exclude the overfitting problem, a small test set was used for the final evaluation. XGB in general provides higher AUROC compared with e-LEN and DT. However, accuracy seems to be significantly higher for e-LEN. XGB has a better balance between sensitivity and specificity for depression and anxiety conditions. Considering the test performance for depression and anxiety conditions, e-LEN and XGB have opposite behaviors in terms of sensitivity and specificity. We can conclude the XGB model is the best for the depression condition with an AUROC of 0.795 in the test set, and the e-LEN model for the

Table 2
Depression prediction performance computed using e-LEN, XGB and DT. The table above shows the mean and standard deviation values calculated for each metric calculated during the 10-fold cross-validation repeated 20 times. The table below shows the performance on the test set.

Model	20-Repeated 10-Fold CV metrics					
	Accuracy	AUROC	Sensitivity	Specificity	PPV	NPV
e-LEN	0.749 ± 0.089	0.681 ± 0.129	0.637 ± 0.194	0.851 ± 0.128	0.821 ± 0.132	0.738 ± 0.108
XGB	0.709 ± 0.112	0.773 ± 0.121	0.676 ± 0.167	0.738 ± 0.155	0.715 ± 0.141	0.725 ± 0.121
DT	0.612 ± 0.131	0.601 ± 0.131	0.589 ± 0.182	0.631 ± 0.171	0.600 ± 0.153	0.635 ± 0.143
	Test metrics					
e-LEN	0.667	0.688	0.500	0.8125	0.700	0.650
XGB	0.733	0.795	0.786	0.688	0.688	0.786
DT	0.533	0.527	0.429	0.625	0.500	0.555

Table 3
Anxiety prediction performance computed using e-LEN, XGB and DT. The table above shows the mean and standard deviation values calculated for each metric calculated during the 10-fold cross-validation repeated 20 times. The table below shows the performance on the test set.

Model	20-Repeated 10-Fold CV metrics					
	Accuracy	AUROC	Sensitivity	Specificity	PPV	NPV
e-LEN	0.721 ± 0.088	0.658 ± 0.139	0.542 ± 0.211	0.886 ± 0.125	0.851 ± 0.153	0.691 ± 0.097
XGB	0.727 ± 0.1	0.775 ± 0.110	0.686 ± 0.160	0.765 ± 0.134	0.743 ± 0.126	0.735 ± 0.116
DT	0.648 ± 0.121	0.647 ± 0.121	0.631 ± 0.177	0.664 ± 0.158	0.643 ± 0.141	0.667 ± 0.123
	Test metrics					
e-LEN	0.767	0.830	0.714	0.813	0.770	0.765
XGB	0.700	0.723	0.786	0.625	0.647	0.770
DT	0.667	0.674	0.786	0.563	0.611	0.750

Table 4
Stress prediction performance computed using e-LEN, XGB, and DT. The table above shows the mean and standard deviation values calculated for each metric calculated during the 10-fold cross-validation repeated 20 times. The table below shows the performance on the test set.

Model	20-Repeated 10-Fold CV metrics					
	Accuracy	AUROC	Sensitivity	Specificity	PPV	NPV
e-LEN	0.761 ± 0.086	0.699 ± 0.122	0.701 ± 0.163	0.816 ± 0.133	0.812 ± 0.114	0.749 ± 0.113
XGB	0.680 ± 0.114	0.740 ± 0.121	0.663 ± 0.175	0.699 ± 0.150	0.699 ± 0.132	0.684 ± 0.136
DT	0.649 ± 0.114	0.649 ± 0.114	0.637 ± 0.176	0.662 ± 0.165	0.668 ± 0.136	0.651 ± 0.133
	Test metrics					
e-LEN	0.600	0.618	0.600	0.600	0.600	0.600
XGB	0.600	0.618	0.667	0.533	0.588	0.616
DT	0.500	0.500	0.533	0.467	0.500	0.500

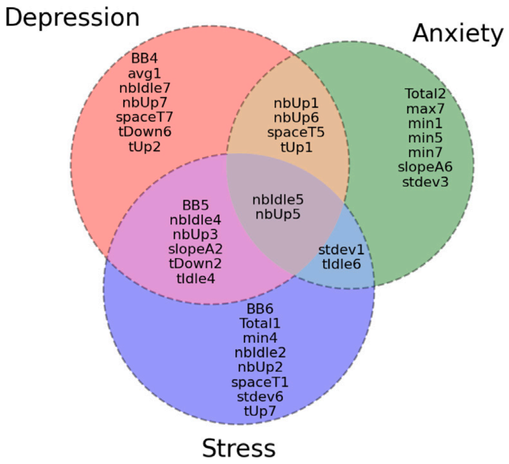


Fig. 6. Venn diagram of selected features for the three conditions.

anxiety condition with an AUROC of 0.830 in the test set. For the stress condition, the cross-validation performance resulted in balanced sensitivity and specificity. However, this condition was the most difficult compared with depression and anxiety. Lower test performance were computed compared with the depression and anxiety conditions.

3.3. Explainability and complexity

Models explainability was implemented in two levels. The first exploits the SHAP analysis, to provide a *post-hoc* explanation for XGB. Furthermore, the explainable-by-design e-LEN model was used to provide the explanations following the first-order logic formalism. As outlined in [42], SHAP and similar feature scoring methods are effective in identifying important features but do not provide insight into how the model combines these features during the decision-making process. In contrast, e-LEN not only identifies important features but also reveals how these features are combined to predict labels, offering a more comprehensive understanding of the model's reasoning.

Fig. 7 shows the global feature importance for the three conditions computed using SHAP. It is not possible to establish in general a feature category more important than others for the three conditions. What is apparent is that for the three conditions, the most predictive features and their value (high or low) leading toward positivity or negativity differ in terms of category and handwriting/drawing task. For example, it can be seen that especially for stress and anxiety conditions, time-related features suggest positivity when the time to perform the task is shortened. The same considerations can be made for the explanations provided by the e-LEN model.

Another e-LEN advantage lies in the fact that explanations are class-level, that is, predictive rules are provided for each specific class. Tables A.6–A.8 provide examples of logic explanations for each class of the three conditions. In the case of depression, all the provided

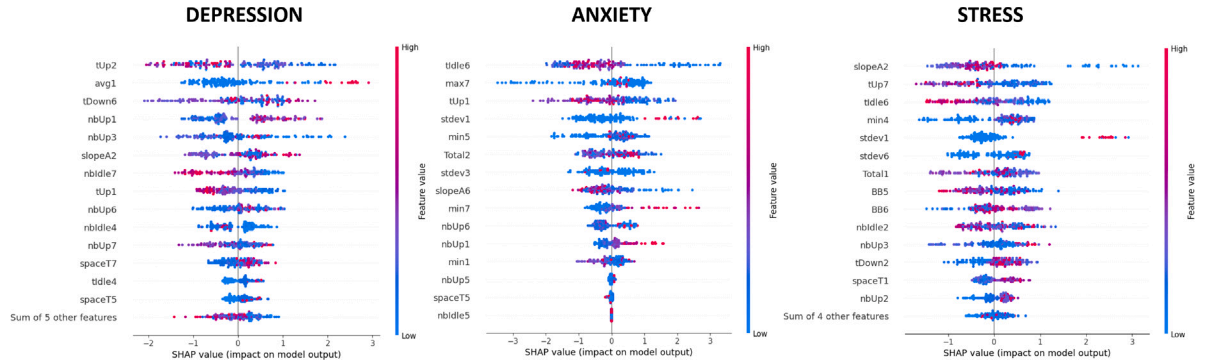


Fig. 7. SHAP beeswarm plots for depression, anxiety, and stress conditions.

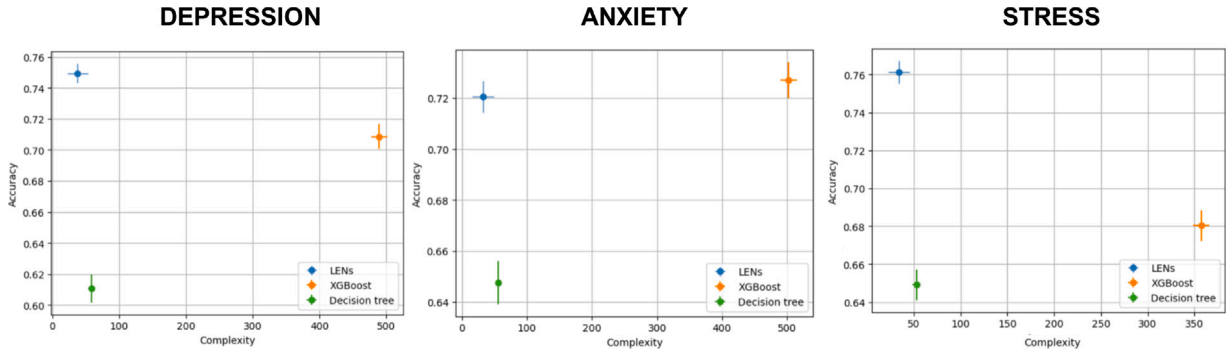


Fig. 8. Accuracy versus complexity plots for depression, anxiety, and stress conditions.

rules involve mainly the number of strokes in each pen status feature. As an example, the formula $(tUp1 \wedge nbIdle4 \wedge nbUp7 \wedge \neg nbIdle7)$ is predictive for depression because (i) the time the pen stands upper than average over the paper for task 1, (ii) there are more strokes than average unrecognized for task 4, (iii) there are more strokes than average in-air for task 4 and (iv) there are fewer strokes than average unrecognized for task 7. All predictive rules for depression positivity have a strong influence coming from the number of strokes in each pen status features. In the case of anxiety positivity, all the predictive rules involve the combination of time, ductus, and pressure features. For stress positivity, the rules involve all categories, and it is not possible to establish one category more impactful than another.

Another important aspect lies in the developed models' complexity. Fig. 8 show the ratio accuracy/complexity obtained for the three trained models. For e-LEN, the complexity is computed by counting the number of predicates for the two classes, while for decision tree-based models was the number of generated nodes during the training. XGB achieves more complex models compared with DT and e-LEN. In general e-LEN results in better accuracy and simpler models compared to DT and XGB. One of the most interesting aspects lies in the complexity of the rules produced by e-LEN for the two classes. The number of rules most used for predicting the positivity of the three conditions is greater than those predictive of negativity. This result is quite intuitive considering that the most difficult task is the prediction of positivity.

3.4. The impact of gender harmonization

The work emphasizes the importance of harmonizing the dataset with respect to gender, as it introduces statistically significant differences between male and female participants in handwriting analysis. This finding supported both statistically by Fig. 5 and by relevant literature [21–23], underscores the need for harmonization. To evaluate the impact of data harmonization on model performance, models were trained using the same pipeline but without the harmonization step for anxiety, depression, and stress tasks. Table 5 presents the accuracy and AUROC scores obtained for each model and task, showing significantly higher results using the harmonization step. Only in the case of e-LEN trained for the anxiety task is this effect marginally noticeable. Moreover, during the testing phase, models trained without harmonization demonstrated very poor generalization, with accuracy scores not exceeding 0.6 across all tasks.

4. Discussion

Trained models suggest that handwriting analysis may be predictive for detecting mental conditions, such as depression, anxiety, and stress. Regarding model accuracy, the prediction of positivity of the three mental disorders is more difficult than the prediction of negativity; in fact, all models have higher specificity than sensitivity. However, the main results concern the introspection of the models through SHAP and rule-based formulas provided by the e-LEN models. Overall, no single model was consistently better than the others; their performance varied depending on the specific conditions considered. The explainable-by-design method e-LEN has presented several advantages, enabling clinical validation through the extracted rules. When choosing between methods such as SHAP and e-LEN, it is critical to consider the importance of studying the relationships between variables in clinical models. Complex clinical phenomena often result from interactions between multiple factors. Understanding not only the importance of these factors but also how they combine to affect patient outcomes can provide deeper clinical insights. By elucidating these combinations through logical rules, e-LEN could guide clinicians in making more informed decisions and potentially lead to the adoption of new practices or interventions tailored to specific patient subgroups.

In our scenario, it was shown that for depression detection, each rule involves the number of strokes in each pen status. In the case of anxiety positivity, all the predictive rules involve the combination of time, ductus, and pressure features. In addition, the models require

Table 5
10-fold cross-validation (repeated 20 times) results for Depression, Anxiety, and Stress Prediction models, with (yes) and without (no) the harmonization step.

		Depression		Anxiety		Stress	
		Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC
e-LEN	Yes	0.749 ± 0.089	0.681 ± 0.129	0.721 ± 0.088	0.658 ± 0.139	0.761 ± 0.086	0.699 ± 0.122
	No	0.724 ± 0.083	0.651 ± 0.133	0.723 ± 0.086	0.676 ± 0.112	0.691 ± 0.084	0.606 ± 0.133
XGB	Yes	0.709 ± 0.112	0.773 ± 0.121	0.727 ± 0.100	0.775 ± 0.110	0.680 ± 0.114	0.740 ± 0.121
	No	0.688 ± 0.114	0.735 ± 0.125	0.704 ± 0.102	0.748 ± 0.121	0.636 ± 0.113	0.677 ± 0.132
DT	Yes	0.612 ± 0.131	0.601 ± 0.131	0.648 ± 0.121	0.647 ± 0.121	0.649 ± 0.114	0.649 ± 0.114
	No	0.597 ± 0.117	0.596 ± 0.118	0.612 ± 0.125	0.612 ± 0.125	0.560 ± 0.116	0.559 ± 0.116

more rules to predict the positive classes, proving the major difficulty of positivity prediction. By examining the obtained result, some considerations can be drawn. First of all, it has been found that the depression condition overlaps in terms of selected features with both anxiety and stress conditions. The overlapping of the three conditions is also common in the clinical field; indeed, a stress response prolonged in time may lead to both physiological and behavioral changes, by altering the typical homeostatic functioning and the ability to properly process emotional responses [43,44]. These impairments may generate anxious and depressive symptoms, by establishing a vicious circle where the ability to cope with the original stressors is increasingly diminished by the onset of these symptoms [45]. Literature widely supports the high comorbidity level among depressive symptomatology and anxiety and stress-related disorders [46]. To this regard, a survey conducted on worldwide scale reported that 45.7% individuals suffering from depression, have also been diagnosed with anxiety disorders [47]. Moreover, literature reports that these conditions may share part of the symptomatology related to psychomotor functioning, such as slowed thought processes, motor hypoactivity, hyperactivity, or alternations between the two, restlessness, and perceived fatigue, which may be detected through handwriting and drawing activities [48,49].

Concerning the differences in the observed predictive power of the three conditions due to the specific features, for the anxiety condition the pressure features seem to be the most significant ones. The hypothesis that such features are effective in identifying anxious states has been supported by other research [50,51]. Likely, this occurred since anxiety is strictly related to muscle tension and feelings of unrelaxation usually reported by individuals suffering from this type of disorder, which could lead them to experience tightness in hand muscles [52]. Differently, in the depression condition temporal and ductus features are the most predictive. For instance, results report that compared to the average, depressed participants spent more time with the pen in-air in task 1, made more in-air and unrecognized traits in task 4, and fewer unrecognized traits in task 7. The longer time that depressed participants need to perform a handwriting or drawing task has been also reported by [9–11]. A possible explanation of this result could be found in the reduced processing velocity that is observed in depressed individuals with psychomotor retardation, which lead to slower reactions during activities entailing action planning and attention processes [53]. For what concerns the number of in-air and unrecognized strokes, results could reflect a less stable trait and greater indecisiveness characterizing depressive states. Such impaired progression of the motor activity could be ascribed to alterations of serotoninergic and dopaminergic neurotransmission in motivational and control neural networks [54]. As regards the stress-related data, these were the most difficult to model. Indeed, stress conditions reported lower test performance compared to depression and anxiety ones. Probably, the prediction accuracy of the handwriting and drawing may have depended on the nature of the stress concept. Indeed, stress is not necessarily a pathological mental health condition, such as anxiety or depression. Rather, it is a natural adaptive strategy of the organism to alert and overcome what is perceived as an emergency situation. The stress response becomes pathological when it is prolonged in time and it does not wear off when the triggering threat has disappeared [55].

Another consideration concerns the absence of homogeneous effects of the different categories' features in each condition. However, this lack of consistency does not jeopardize the efficacy of the models. Indeed, it is conceivable that the effects of different features depend on the conditions. Although stress, anxiety, and depression share part of the symptoms, they remain three different mental health conditions with their specificity in severity and phenomenology. As well, the type of tasks plays a role in the predictive accuracy of the features. Such relation may be ascribed to the different contributions of psychomotor skills in carrying out the different tasks. To support this, developmental and neuroimaging studies suggest that, for instance, drawing and writing activities are regulated by different processing systems [56–59].

Finally, to support the gender harmonization step performed in the pre-processing stage of the current study, we show that the literature investigating gender differences in handwriting and drawing processes. In fact, it reports discrepancies both in neural correlates underlying such activities and psychological factors (e.g., self-efficacy; self-awareness), which result in a different production of handwriting movements [60,61]. Due to these differences, studies carrying out handwriting signal processing and machine learning techniques suggest normalizing the weight of gender to decrease biases likelihood in predictive models for healthcare and medical applications [62–64].

Overall, results obtained in the current study support the adoption of this explainable AI model in the diagnostic process, with significant implications for clinical practice. By offering transparent explanations for its predictions, the model can help clinicians better identify putative biomarkers that may be indicative of specific medical conditions such as depression, anxiety, and stress [65]. This enables a more informed and confident diagnostic process, where AI-generated insights complement clinical expertise. These techniques contribute to the detection of quantitative, noninvasive indicators of psychiatric and psychological disorders, which are not always easily recognizable due to potential overlap with other conditions and the heterogeneous nature of their symptomatology. Considering that such complexity could hinder the diagnostic process, the use of these methods can provide more objective support, helping clinicians differentiate between disorders and tailor interventions more effectively. In addition, this would potentially reduce diagnostic times, enabling earlier intervention, preventing the worsening of symptoms, and ultimately improving patient outcomes. Finally, it could lower the overall burden of the disease, both in terms of mental health and healthcare costs [66]. In summary, integrating explainable AI into clinical practice not only enhances the diagnostic process but also fosters more personalized treatment plans, offering significant benefits in improving patient care and reducing long-term impacts.

5. Conclusion

In conclusion, the contemporary advancement of high-performance models necessitates a concurrent emphasis on their explainability. This study exploits the traditional and extensively acknowledged SHAP explanation and the explanations offered by the explainable-by-design e-LEN model. Leveraging the first-order logic rules inherent in e-LEN, this research substantiates certain findings previously established by the

Table A.6
First-order rules involved in the e-LEN training for depression condition.

Class	Formulas
Negative	$(nbUp7 \land \neg nbIdle7 \land \neg avg1) \lor (nbUp7 \land \neg spaceT7 \land \neg avg1) \lor (nbIdle7 \land \neg nbUp6 \land \neg avg1) \lor (\neg nbIdle7 \land \neg spaceT7 \land \neg avg1) \lor (nbUp6 \land spaceT7 \land avg1 \land \neg nbUp7 \land \neg nbIdle7)$
Positive	$(tUp1 \land nbIdle4 \land nbUp7 \land \neg nbIdle7) \lor (tUp1 \land nbIdle4 \land nbIdle7 \land \neg slopeA2) \lor (slopeA2 \land \neg tUp1 \land \neg nbIdle4 \land \neg nbUp7) \lor (nbIdle4 \land \neg tUp1 \land \neg nbUp7 \land \neg nbIdle7) \lor (tUp1 \land \neg slopeA2 \land \neg nbIdle4 \land \neg nbUp7 \land \neg nbIdle7)$

Table A.7
First-order rules involved in the e-LEN training for anxiety condition.

Class	Formulas
Negative	$(min1 \land \neg min7 \land \neg max7) \lor (\neg nbUp1 \land \neg min7 \land \neg max7) \lor (nbUp1 \land nbUp6 \land \neg min1 \land \neg max7) \lor (nbUp1 \land min1 \land \neg nbUp6 \land \neg max7)$
Positive	$(tUp1 \land \neg nbUp6 \land \neg min1) \lor (nbUp1 \land tUp1 \land nbUp6 \land min1 \land min7) \lor (nbUp1 \land nbUp6 \land \neg tUp1 \land \neg min1) \lor (nbUp1 \land \neg tUp1 \land \neg min1 \land \neg min7) \lor (min1 \land min7 \land \neg nbUp1 \land \neg tUp1 \land \neg nbUp6)$

Table A.8
First-order rules involved in the e-LEN training for stress condition.

Class	Formulas
Negative	$(spaceT1 \land \neg nbUp5 \land \neg stdev1) \lor (slopeA2 \land \neg nbUp5 \land \neg stdev1) \lor (spaceT1 \land slopeA2 \land \neg nbIdle4 \land \neg stdev1) \lor (nbIdle4 \land nbUp5 \land \neg spaceT1 \land \neg slopeA2 \land \neg stdev1)$
Positive	$(tDown2 \land \neg slopeA2 \land \neg BB6) \lor (stdev6 \land \neg slopeA2 \land \neg tIdle6) \lor (stdev6 \land \neg tIdle6 \land \neg BB6) \lor (\neg tDown2 \land \neg tIdle6 \land \neg BB6) \lor (slopeA2 \land stdev6 \land \neg tDown2 \land \neg BB6) \lor (BB6 \land \neg tDown2 \land \neg slopeA2 \land \neg stdev6)$

clinical literature of the analyzed domain. The explanations provided by e-LEN introduce valuable benefits, particularly in terms of clinical validation, fostering confidence in these systems, and facilitating their integration into clinical practice.

Funding

This work was supported by the European Union - Next Generation EU, PNRR Mission 4 Component 2 Investment 1.1 – D.D n.1409 del 14-09-2022 PRIN 2022 – Code “P20222MYKE” - CUP: B53D2302598000 - Project Title: “A Pilot analysIs of behavioRal and opeRational data for detEcting Socio-emotional PrEcursors of mild CogniTIVE impairments (MCI) and demEntia (IRRESPECTIVE)”; and by the projects TRUST-ME (205121L_214991).

CRediT authorship contribution statement

Francesco Prinzi: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Pietro Barbiero:** Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Claudia Greco:** Writing – original draft, Visualization, Validation, Data curation. **Terry Amorese:** Writing – original draft, Visualization, Validation, Data curation. **Gennaro Cordasco:** Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Pietro Liò:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation. **Salvatore Vitabile:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition. **Anna Esposito:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix. Logic formulas

See Tables A.6–A.8.

References

[1] A.J. Romain, J. Marleau, A. Baillot, Impact of obesity and mood disorders on physical comorbidities, psychological well-being, health behaviours and use of health services, J. Affect. Disord. 225 (2018) 381–388, http://dx.doi.org/10.1016/j.jad.2017.08.065.

[2] L. Iani, R.M. Quinto, M. Lauriola, M.L. Crosta, G. Pozzi, Psychological well-being and distress in patients with generalized anxiety disorder: The roles of positive and negative functioning, PLoS One 14 (11) (2019) http://dx.doi.org/10.1371/journal.pone.0225646.

[3] S. Cohen, R. Kessler, U.L. Gordon, Strategies for measuring stress in studies of psychiatric and physical disorder, in: Measuring Stress: A Guide for Health and Social Scientists, Oxford University Press, New York, NY, USA, 1995, pp. 3–26, http://dx.doi.org/10.1093/oso/9780195086416.003.0001P.

[4] M.H. Pollack, Comorbid anxiety and depression, J. Clin. Psychiatry 66 (22) (2005).

[5] American Psychiatric Association, Anxiety disorders, in: Diagnostic and Statistical Manual of Mental Disorders: DSM-5, Vol. 5, 2013, (5).

[6] Institute of Health Metrics and Evaluation, Global health data exchange (GHDx), 2023, https://vizhub.healthdata.org/gbd-results/, (Accessed 12 December 2023).

[7] S. Cohen, D. Janicki-Deverts, G.E. Miller, Psychological stress and disease, Jama 298 (14) (2007) 1685–1687, http://dx.doi.org/10.1001/jama.298.14.1685.

[8] M. Large, Study on suicide risk assessment in mental illness underestimates inpatient suicide risk, BMJ 352 (2016) http://dx.doi.org/10.1136/bmj.i267.

[9] C. Greco, G. Raimo, T. Amorese, M. Cuciniello, G. Mcconvey, G. Cordasco, M. Faundez-Zanuy, A. Vinciarelli, Z. Callejas-Carrion, A. Esposito, Discriminative power of handwriting and drawing features in depression, Int. J. Neural Syst. (2023) http://dx.doi.org/10.1142/S0129065724500060.

[10] G. Cordasco, F. Scibelli, M. Faundez-Zanuy, L. Likforman-Sulem, A. Esposito, Handwriting and drawing features for detecting negative moods, in: Quantifying and Processing Biomedical and Behavioral Signals, Vol. 27, Springer, 2019, pp. 73–86, http://dx.doi.org/10.1007/978-3-319-95095-2_7.

[11] G. Raimo, M. Buonanno, M. Conson, G. Cordasco, M. Faundez-Zanuy, G. McConvey, S. Marrone, F. Marulli, A. Vinciarelli, A. Esposito, Handwriting and drawing for depression detection: A preliminary study, in: Applied Intelligence and Informatics: Second International Conference, AII 2022, Reggio Calabria, Italy, September 1–3, 2022, Proceedings, Springer, 2023, pp. 320–332, http://dx.doi.org/10.1007/978-3-031-24801-6_23.

[12] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. (CSUR) 51 (5) (2018) 1–42, http://dx.doi.org/10.1145/3236009.

[13] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, Benchmarking and survey of explanation methods for black box models, 2021, arXiv preprint arXiv:2102.13076.

[14] P. Tschandl, N. Codella, B.N. Akay, G. Argenziano, R.P. Braun, H. Cabo, D. Gutman, A. Halpern, B. Helba, R. Hofmann-Wellenhof, et al., Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study, Lancet Oncol. 20 (7) (2019) 938–947, http://dx.doi.org/10.1016/S1470-2045(19)30333-X.

[15] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, et al., Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions, 2023, arXiv preprint arXiv:2310.19775.

[16] T. Han, S. Srinivas, H. Lakkaraju, Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations, Adv. Neural Inf. Process. Syst. 35 (2022) 5256–5268, http://dx.doi.org/10.48550/arXiv.2206.01254.

[17] Z.C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery, Queue 16 (3) (2018) 31–57, http://dx.doi.org/10.1145/3236386.3241340.

[18] P. Barbiero, G. Ciravegna, F. Giannini, P. Lió, M. Gori, S. Melacci, Entropy-based logic explanations of neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 6046–6054, http://dx.doi.org/10.1609/aaai.v36i6.20551, (6).

[19] F. Prinzi, C. Militello, N. Scichilone, S. Gaglio, S. Vitabile, Explainable machine-learning models for COVID-19 prognosis prediction using clinical, laboratory and radiomic features, IEEE Access 11 (2023) 121492–121510, http://dx.doi.org/10.1109/ACCESS.2023.3327808.

[20] P.F. Lovibond, S.H. Lovibond, The structure of negative emotional states: Comparison of the depression anxiety stress scales (DASS) with the beck depression and anxiety inventories, Behav. Res. Ther. 33 (3) (1995) 335–343, http://dx.doi.org/10.1016/0005-7967(94)00075-u.

[21] I. Rabaev, M. Litvak, Automated gender classification from handwriting: a systematic survey, Appl. Intell. 53 (13) (2023) 17154–17177, http://dx.doi.org/10.1007/s10489-022-04347-w.

[22] S. Dargan, M. Kumar, A. Mittal, K. Kumar, Handwriting-based gender classification using machine learning techniques, Multimedia Tools Appl. (2023) 1–25, http://dx.doi.org/10.1007/s11042-023-16354-1.

[23] A.-Q. Najla, M. Khayyat, C.Y. Suen, Novel features to detect gender from handwritten documents, Pattern Recognit. Lett. 171 (2023) 201–208, http://dx.doi.org/10.1016/j.patrec.2022.08.016.

[24] J.-P. Fortin, N. Cullen, Y.I. Sheline, et al., Harmonization of cortical thickness measurements across scanners and sites, NeuroImage 167 (2018) 104–120, http://dx.doi.org/10.1016/j.neuroimage.2017.11.024.

[25] W.E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, Biostatistics 8 (1) (2006) 118–127, http://dx.doi.org/10.1093/biostatistics/kxj037.

[26] D. Dima, A. Modabbernia, E. Papachristou, G.E. Doucet, I. Agartz, M. Aghajani, T.N. Akudjedu, A. Albajes-Eizagirre, D. Alnæs, K.I. Alpert, et al., Subcortical volumes across the lifespan: Data from 18,605 healthy individuals aged 3–90 years, Hum. Brain Mapp. 43 (1) (2022) 452–469, http://dx.doi.org/10.1002/hbm.25320.

[27] Y. Nan, J. Del Ser, S. Walsh, C. Schönlieb, M. Roberts, I. Selby, K. Howard, J. Owen, J. Neville, J. Guiot, et al., Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions, Inf. Fusion 82 (2022) 99–122, http://dx.doi.org/10.1016/j.inffus.2022.01.001.

[28] J. Xu, Y. Su, J. Fu, X. Wang, B.A. Nguchu, B. Qiu, Q. Dong, X. Cheng, Glymphatic dysfunction correlates with severity of small vessel disease and cognitive impairment in cerebral amyloid angiopathy, Eur. J. Neurol. 29 (10) (2022) 2895–2904, http://dx.doi.org/10.1111/ene.15450.

[29] S. Raschka, Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack, J. Open Source Softw. 3 (24) (2018) http://dx.doi.org/10.21105/joss.00638.

[30] M.M. Ghiasi, S. Zendehboudi, Application of decision tree-based ensemble learning in the classification of breast cancer, Comput. Biol. Med. 128 (2021) 104089, http://dx.doi.org/10.1016/j.compbiomed.2020.104089.

[31] V. Di Stefano, F. Prinzi, M. Luigetti, M. Russo, S. Tozza, P. Alonge, A. Romano, M.A. Sciarrone, F. Vitali, A. Mazzeo, et al., Machine learning for early diagnosis of ATTRv amyloidosis in non-endemic areas: A multicenter study from Italy, Brain Sci. 13 (5) (2023) 805, http://dx.doi.org/10.3390/brainsci13050805.

[32] R. Shwartz-Ziv, A. Armon, Tabular data: Deep learning is not all you need, Inf. Fusion 81 (2022) 84–90, http://dx.doi.org/10.1016/j.inffus.2021.11.011.

[33] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017) http://dx.doi.org/10.5555/3295222.3295230.

[34] A. Bell, I. Solano-Kamaiko, O. Nov, J. Stoyanovich, It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 248–266, http://dx.doi.org/10.1145/3531146.3533090.

[35] L.G. McCoy, C.T. Brenna, S.S. Chen, K. Vold, S. Das, Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based, J. Clin. Epidemiol. 142 (2022) 252–257, http://dx.doi.org/10.1016/j.jclinepi.2021.11.001.

[36] A.J. London, Artificial intelligence and black-box medical decisions: accuracy versus explainability, Hastings Cent. Rep. 49 (1) (2019) 15–21, http://dx.doi.org/10.1002/hast.973.

[37] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (5) (2019) 206–215, http://dx.doi.org/10.1038/s42256-019-0048-x.

[38] A.A. Freitas, Comprehensible classification models: A position paper, SIGKDD Explor. Newsl. 15 (1) (2014) http://dx.doi.org/10.1145/2594473.2594475.

[39] J.R. Quinlan, Generating production rules from decision trees, in: Proceedings of the 10th International Joint Conference on Artificial Intelligence - Volume 1, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987, pp. 304–307, http://dx.doi.org/10.5555/1625015.1625078.

[40] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (1) (2020) 56–67, http://dx.doi.org/10.1038/s42256-019-0138-9.

[41] N. Papanikolaou, C. Matos, D.M. Koh, How to develop a meaningful radiomic signature for clinical use in oncologic patients, Cancer Imaging 20 (2020) 1–10, http://dx.doi.org/10.1186/s40644-020-00311-4.

[42] G. Ciravegna, P. Barbiero, F. Giannini, M. Gori, P. Lió, M. Maggini, S. Melacci, Logic explained networks, Artificial Intelligence 314 (2023) 103822, Logic explained networks.

[43] E. de Kloet, M. Joëls, F. Holsboer, Stress and the brain: from adaptation to disease, Nat. Rev. Neurosci. 6 (2005) 463–475, http://dx.doi.org/10.1038/nrn1683.

[44] J. Koolhaas, Stress revisited: a critical evaluation of the stress concept, Neurosci. Biobehav. Rev. 35 (2011) 1291–1301, http://dx.doi.org/10.1016/j.neubiorev.2011.02.003.

[45] D. Wheatley, Stress, anxiety and depression, Stress Med. 13 (3) (1997) 173–177.

[46] N.H. Kalin, The critical relationship between anxiety and depression, Am. J. Psychiatry 177 (5) (2020) 365–367, http://dx.doi.org/10.1176/appi.ajp.2020.20030305.

[47] R.C. Kessler, et al., Anxious and non-anxious major depressive disorder in the world health organization world mental health surveys, Epidemiol. Psychiatr. Sci. 24 (2015) 210–226, http://dx.doi.org/10.1017/S2045796015000189.

[48] P. Singh, H. Yadav, Influence of neurodegenerative diseases on handwriting, Forens. Res. Criminol. Int. J. 9 (3) (2021) 110–114.

[49] E. Elkjaer, M.B. Mikkelsen, J. Michalak, D.S. Mennin, M.S. O'Toole, Motor alterations in depression and anxiety disorders: A systematic review and meta-analysis, J. Affect. Disord. 317 (2022) 373–387, http://dx.doi.org/10.1016/j.jad.2022.08.060.

[50] S.D. LaRoque, J.E. Obrzut, Pencil pressure and anxiety in drawings, J. Psychoeduc. Assess. 24 (4) (2006) 381–393, http://dx.doi.org/10.1177/0734282906288520.

[51] N. Vyawahare, A. Ashtaputre-Sisode, Relation between stress, anxiety and handwriting, J. Maharaja Sayajirao Univ. Baroda (2022).

[52] M. Pluess, A. Conrad, F.H. Wilhelm, Muscle tension in generalized anxiety disorder: A critical review of the literature, J. Anxiety Disord. 23 (1) (2009) 1–11, http://dx.doi.org/10.1016/j.janxdis.2008.03.016.

[53] D.K. Ahorsu, H.W. Tsang, Do people with depression always have decreased cognitive processing speed? Neuropsychiatry 8 (4) (2018) 1227–1231.

[54] I. Grahek, A. Shenhav, S. Musslick, R.M. Krebs, E.H.W. Koster, Motivation and cognitive control in depression, Neurosci. Biobehav. Rev. 102 (2019) 371–381, http://dx.doi.org/10.1016/j.neubiorev.2019.04.011.

[55] H. Yaribeygi, Y. Panahi, H. Sahraei, T.P. Johnston, A. Sahebkar, The impact of stress on body function: A review, EXCLI J. 16 (2017) 1057–1072, http://dx.doi.org/10.17179/excli2017-480.

[56] A.R. Potgieser, A. van der Hoorn, B.M. de Jong, Cerebral activations related to writing and drawing with each hand, PLoS One 10 (5) (2015) http://dx.doi.org/10.1371/journal.pone.0126723.

[57] L. Taverna, M. Tremolada, F. Sabattini, Drawing and writing. Learning of graphical representational systems in early childhood, in: Proceedings of the 2nd International and Interdisciplinary Conference on Image and Imagination: IMG 2019, 2020, pp. 216–229, http://dx.doi.org/10.1007/978-3-030-41018-6_20.

[58] G. Pinto, O. Incognito, The relationship between emergent drawing, emergent writing, and visual-motor integration in preschool children, Infant Child Dev. 31 (2) (2022) e2284, http://dx.doi.org/10.1002/icd.2284.

[59] A. Baumann, I. Tödt, A. Knutzen, C.A. Gless, O. Granert, S. Wolff, K.E. Zeuner, Neural correlates of executed compared to imagined writing and drawing movements: a functional magnetic resonance imaging study, Front. Hum. Neurosci. 16 (2022) 829576, http://dx.doi.org/10.3389/fnhum.2022.829576.

[60] C. Cordeiro, S.L. Castro, T. Limpo, Examining potential sources of gender differences in writing: The role of handwriting fluency and self-efficacy beliefs, Writ. Commun. 35 (4) (2018) 448–473, http://dx.doi.org/10.1177/0741088318788843.

[61] Y. Yang, F. Tam, S.J. Graham, G. Sun, J. Li, C. Gu, R. Tao, N. Wang, H.Y. Bi, Z. Zuo, Men and women differ in the neural basis of handwriting, Hum. Brain Mapp. 41 (10) (2020) 2642–2655, http://dx.doi.org/10.1002/hbm.24968.

[62] H. Wang, Z. Wu, E.P. Xing, Removing confounding factors associated weights in deep neural networks improves the prediction accuracy for healthcare applications, Pac. Symp. Biocomput. 24 (2019) 51–65, http://dx.doi.org/10.1142/9789813279827_0006.

[63] D. Cirillo, S. Catuara-Solarz, C. Morey, E. Guney, L. Subirats, S. Mellino, A. Gigante, A. Valencia, M.J. Rementeria, A.S. Chadha, N. Mavridis, Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare, NPJ Digit. Med. 3 (81) (2020) http://dx.doi.org/10.1038/s41746-020-0288-5.

[64] M. Faundez-Zanuy, J. Mekyska, Analysis of gender differences in online handwriting signals for enhancing e-health and e-security applications, Cogn. Comput. 15 (2023) 208–2019, http://dx.doi.org/10.1007/s12559-023-10116-9.

[65] J. Sun, Q.-X. Dong, S.-W. Wang, Y.-B. Zheng, X.-X. Liu, T.-S. Lu, K. Yuan, J. Shi, B. Hu, L. Lu, et al., Artificial intelligence in psychiatry research, diagnosis, and therapy, Asian J. Psychiatry (2023) 103705, http://dx.doi.org/10.1016/j.ajp.2023.103705.

[66] A. Ray, A. Bhardwaj, Y.K. Malik, S. Singh, R. Gupta, Artificial intelligence and psychiatry: An overview, Asian J. Psychiatry 70 (2022) 103021, http://dx.doi.org/10.1016/j.ajp.2022.103021.