stats

MDPI

Communication

Assessing Mental Health Without Relying on Psychiatric Data

Juan de Dios Sanjuán-Antúnez [1], Ernesto Curbelo [1], Inmaculada Peñuelas-Calvo [2], Enrique Baca-García [3,4], Elena Pérez-García [5], Hilario Blasco-Fontecilla [6,7] and David Delgado-Gómez [1,*]

[1] Departament of Statistics, School of Engineering, University Carlos III of Madrid, 28911 Leganés, Spain; jsanjuan@est-econ.uc3m.es (J.d.D.S.-A.); ecurbelo@est-econ.uc3m.es (E.C.)
[2] Foundation for Biomedical Research of the Hospital 12 de Octubre, 28041 Madrid, Spain; inmaculada.penuelas@salud.madrid.org
[3] Instituto de Investigación Sanitaria de la Fundación Jiménez Díaz, 28003 Madrid, Spain; ebaca@quironsalud.es
[4] Department of Psychology, Catholic University of the Maule, Talca 3460000, Chile
[5] Independent Researcher, 28521 Rivas Vaciamadrid, Spain; eperezg2@gmail.com
[6] Institute for Transfer and Research (ITEI), International University of La Rioja (UNIR), 26006 Logroño, Spain; hilariomanuel.blasco@unir.net
[7] Department of Psychiatry, Institute of Mental Health and Wellbeing Emooti, 28010 Madrid, Spain
* Correspondence: ddelgado@est-econ.uc3m.es

Abstract

Approximately one in seven people worldwide is affected by a mental disorder, making early identification crucial for preventing symptom progression and chronicity. However, diagnosis remains challenging, with discrepancies between referring physicians' initial assessments and specialist psychiatric diagnoses reported in approximately half of cases. In recent years, machine learning has shown promising results in mental health assessment, with supervised techniques such as decision trees, random forests, and support vector machines among the most widely used approaches. However, these methods require large and representative datasets to capture the substantial heterogeneity of mental disorders, which can be difficult to obtain due to stigma, disease severity, and limited access to mental healthcare. This work investigates the feasibility of identifying mental disorders using data exclusively from a non-psychiatric population. The underlying hypothesis is that this population exhibits greater homogeneity and can therefore be characterized using a variational autoencoder with a Gaussian latent space. The proposed approach uses reconstruction errors from the non-psychiatric training population as a reference, with substantially higher errors in new observations indicating a potential mental disorder. Results show that the proposed VAE achieves performance comparable to that of the three aforementioned supervised techniques in the two dataset used. For the first dataset, the accuracy, sensitivity, specificity and ROC-AUC of the VAE were 0.88, 0.90, 0.85 and 0.95. These performance measures were 0.83, 0.86, 0.80, and 0.88 for decision trees; 0.88, 0.91, 0.85 and 0.94 for random forest; and 0.88, 0.91, 0.85 and 0.95 for support vector machines. Similar results were obtained for the second dataset. Moreover, the progressive increase in reconstruction errors across populations may reflect greater deviations from the patterns observed in the non-psychiatric population in groups with greater clinical complexity, supporting its potential use as a screening tool.

Keywords: mental disorders; variational autoencoders; decision trees; random forest; support vector machines

Check for updates

Academic Editor: Wei Zhu

Received: 21 August 2026
Revised: 4 October 2026
Accepted: 9 October 2026
Published: 14 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Mental disorders are defined as syndromes that involve clinically significant disturbances in an individual's cognition, emotional regulation, or behavior, reflecting dysfunctions in the underlying biological, psychological, or developmental processes [1]. According to the World Health Organization, in 2021, approximately one in seven people worldwide, or around 1.1 billion individuals, were living with a mental disorder [2]. Among these, anxiety and depressive disorders were the most common. In addition, the COVID-19 pandemic has been noted to have a significant impact on global mental health, leading to an estimated 25% increase in the prevalence of anxiety and depressive disorders [3].

Mental health disorders account for approximately 16% of the global disease burden, as measured in disability-adjusted life years (DALYs), highlighting that one in every six years of healthy life lost worldwide is due to mental illness [4,5]. In economic terms, poor mental health was estimated to cost the global economy approximately $2.5 trillion per year in healthcare costs and reduced productivity in 2010, with this figure projected to rise to $6 trillion by 2030 [6].

The situation becomes even more alarming when these conditions are not identified and treated in time. Untreated mental illness can deteriorate progressively, leading to chronic disability, social and occupational dysfunction, and a significant decline in overall quality of life [7,8]. Moreover, in many cases, it can result in death by suicide, which remains a major global public health concern, accounting for nearly 800,000 deaths each year [9], and standing as the second leading cause of death among individuals aged 15 to 29 worldwide [10].

Fortunately, when mental health disorders are detected early and properly managed, the outcomes can improve significantly. Early intervention has been shown to reduce symptom severity, lower the risk of relapse, enhance quality of life, and in many cases, prevent the onset of more severe complications, including suicide. Prompt access to psychological support, medication when necessary, and community-based services can make a critical difference in recovery and long-term well-being [11,12].

However, achieving an early and accurate diagnosis in mental health currently remains one of the greatest challenges for professionals in the field [13]. A recent observational study underscored this issue, showing that the initial impressions of referring physicians differed from the final diagnoses made by psychiatric specialists in nearly 50% of cases [14]. In standard clinical practice, diagnosis relies primarily on a structured or semi-structured medical interview. After ruling out biological or neurological causes, the psychiatrist makes a diagnosis based on information provided by the patient and, in some cases, by family members or caregivers. Consequently, the accuracy of the diagnosis depends not only on the clinician's expertise and knowledge but also on various factors related to the informants [15]. Variables such as the informants' age, gender, emotional state, limited self-awareness, concern about stigma, and specific personal interests may hinder the diagnostic process [16–18].

To support clinical assessment, medical professionals often rely on standardized psychometric scales and questionnaires, such as the Beck Depression Inventory (BDI) [19], Social Readjustment Rating Scale (SRRS) [20], the International Personality Disorder Examination Screening Questionnaire (IPDE-SC) [21], and the Patient Health Questionnaire (PHQ) [22], among others. However, these instruments present several limitations, including variable effectiveness, cultural bias, and challenges related to respondents' comprehension and engagement [23,24]

In recent years, various machine learning (ML)-based approaches have been developed to improve the accuracy and early detection of mental health diagnoses [25]. This progress is reflected in several systematic reviews published in the literature, which examine the

application of these techniques either across a broad range of mental health disorders [26,27] or focusing on specific conditions such as suicide risk [28], depression [29], autism spectrum disorder [30], and attention deficit hyperactivity disorder (ADHD) [31]. These reviews reveal that the vast majority of the employed techniques are supervised, requiring each observation to be assigned a label indicating whether the individual is a case or a control. Consequently, model performance depends heavily on the representativeness of the labeled groups. This assumption might be particularly problematic in mental health research, where case populations are often highly heterogeneous and may fail to adequately characterize all clinically relevant subpopulations [32,33].

This study is based on the hypothesis that, while the psychiatric population is inherently heterogeneous due to the presence of multiple clinical subtypes and symptom profiles, the non-psychiatric population exhibits a comparatively less structured and more homogeneous distribution. Under this assumption, an unsupervised framework is proposed in which only the non-psychiatric population is modeled, while individuals with mental disorders are treated as outliers. To identify these anomalous patterns, variational autoencoders (VAEs) are employed to learn a low-dimensional latent representation of the non-psychiatric population, approximated as a Gaussian distribution. Individuals with mental disorders are subsequently detected through elevated reconstruction errors, enabling their differentiation from the reference population.

The remainder of this work is structured as follows. Section 2 presents the related work and positions the proposed approach within the existing literature, highlighting its novelty and contributions. Section 3 presents a brief introduction to the fundamentals of neural networks and VAEs to provide the necessary background, followed by the presentation of the proposed approach. Section 4 describes the materials and methods, including participant characteristics, the assessment instruments used (scales and questionnaires), the classifiers employed, the experimental design, and the evaluation metrics used for performance comparison. Section 5 presents the obtained results and findings. Section 6 provides a discussion and interpretation of the results in the context of existing literature. Finally, Section 7 presents the main conclusions and summarizes the key contributions.

2. Literature Review

Given the growing diversity and availability of data sources related to mental health, there has been a notable increase in the development and application of analytical techniques aimed at the detection and characterization of mental disorders. This growing interest is reflected in the emergence of several literature reviews published in recent years.

In this context, Shatte, Hutchinson, and Teague conducted a scoping review of 300 studies focused on the application of machine learning in mental health, identifying depression, schizophrenia, Alzheimer's disease, stress, and suicide as the most frequently addressed conditions [27]. Their review also highlighted that the most commonly employed analytical techniques were supervised learning approaches, particularly support vector machines (SVMs), random forests (RFs), and decision trees (DTs).

Iyortsuun et al. provided a systematic review of 33 published studies on the use of machine learning and deep learning methods for the diagnosis of mental health disorders [26]. The review covers several disorders, including schizophrenia, depression, anxiety, bipolar disorder, post-traumatic stress disorder (PTSD), anorexia nervosa, and ADHD. The findings indicate that traditional machine learning methods remain dominant, accounting for approximately 62.5% of the reviewed studies, while deep learning-based approaches represent around 25%, and hybrid machine learning and deep learning methods account for 12.5%. Among the most frequently used techniques, Support Vector Machine (SVM) and Random Forest (RF) stand out, appearing in more than half of the studies.

Pigoni et al. conducted a systematic review focused on the prediction of suicidal behavior, including a total of 81 studies. Consistent with previous research, they found that SVMs and RFs were the most frequently used techniques and generally achieved the best predictive performance [28]. In addition, the authors also highlighted the limited presence of unsupervised approaches, with only four studies applying clustering techniques.

These reviews are particularly relevant, as they reveal the predominance of supervised models in the literature and suggest that unsupervised methodologies capable of identifying complex latent structures remain underexplored.

In relation to the use of VAEs in the field of mental health, their application has mainly developed along two directions: data augmentation and latent representation learning, with the resulting latent space typically being used to train downstream classifiers.

With respect to studies applying VAEs for data augmentation, the work by Nagrath et al. is particularly relevant [34]. In this study, the VAE was not employed as a classification model, but rather as a synthetic data generation mechanism to increase dataset size and address class imbalance in clinical datasets. Their findings indicate that VAE-generated samples can improve the performance of traditional machine learning classifiers, including RFs and SVMs, particularly when an appropriate proportion of synthetic data is incorporated into the training process. However, the benefits of data augmentation are not uniform across models, with ensemble-based methods demonstrating greater robustness to synthetic data integration, whereas linear models appear more sensitive to distributional shifts.

Within the second category, a representative example is the work by Geenjaar et al. [35]. In this study, VAEs are used to learn a low-dimensional latent representation that fuses multimodal neuroimaging data. Specifically, the VAE compresses high-dimensional functional and structural brain signals into a compact latent space while preserving the most relevant information for downstream analysis. The resulting embeddings are then used as input to a SVM to distinguish patients with schizophrenia from controls. The model achieved strong discriminative performance, demonstrating that the learned latent space retains clinically meaningful information despite substantial dimensionality reduction.

The work most closely related to the present study lies outside the field of mental health. Shan et al. trained an adversarial autoencoder exclusively on normal electrocardiograms to detect abnormal signals as outliers [36]. The proposed approach combines reconstruction and discrimination scores to quantify deviations in new signals.

In contrast to the approach proposed by Geenjaar et al. [35], the present work does not use the learned latent representation as input to a downstream classifier. Instead, similar to Shan et al. [36], reconstruction errors are used to identify observations that deviate from the reference population. However, our approach uses a variational autoencoder to characterize the response patterns of the non-psychiatric reference population through a latent space constrained to follow a Gaussian distribution. To the best of our knowledge, this is the first study to apply this type of VAE-based approach to mental health assessment. To assess the effectiveness of the proposed approach, its performance will be compared against supervised machine learning methods. In particular, SVMs, RFs, and DTs will be considered, as these methods have been identified in previous reviews as the most frequently employed approaches in the mental health literature.

3. Proposed Methodological Framework

Before describing the proposed approach, a brief overview of the fundamental concepts underlying neural networks is presented to provide the necessary background. For a more comprehensive treatment of the topic, readers are referred to the textbook by Goodfellow, Bengio, and Courville [37].

3.1. Neural Networks

Neural networks are computational models inspired by the human brain that can be applied to supervised tasks such as classification and regression, as well as to unsupervised tasks such as clustering and representation learning. As illustrated in Figure 1, a neural network consists of interconnected units called neurons, organized into three types of layers: input, hidden, and output layers. In the following, each of these layers is explained.

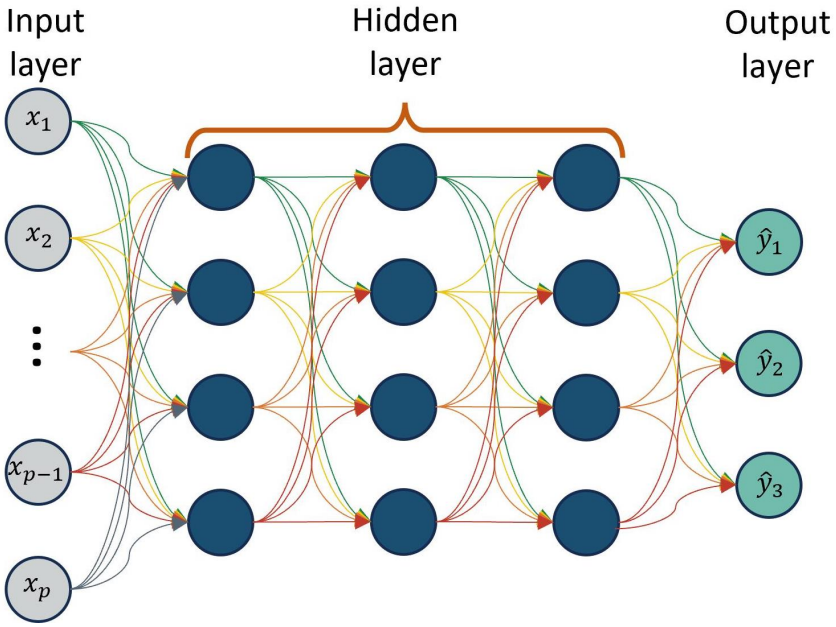

Figure 1. Representation of a neural network with p input variables and three output variables. Source: authors' own elaboration.

Input neurons. Given a dataset of n observations $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$, $i = 1, \ldots, n$, where $\boldsymbol{x}_i$ represents the predictor variables and $\boldsymbol{y}_i$ represents the dependent variable(s) associated with the ith observation, the input layer introduces the predictors $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{ip}]$ into the network.

Hidden neurons. Neurons in the intermediate, or hidden, layers progressively transform the network's input $\boldsymbol{x}_i$ to approximate the target value(s) $\boldsymbol{y}_i$ of the dependent variable(s). Each neuron j in layer i computes its output a_j^i, by taking a weighted sum of the signals received from the previous layer a_k^{i-1}, adding a bias term b_j^i, and then applying an activation function f, which can be either linear or non-linear, as follows:

$$a_j^i = f\left(\sum_{k=1}^{n_{i-1}} w_{kj}^{i-1} a_k^{i-1} + b_j^i\right),$$

where w_{kj}^{i-1} denotes the weight of the connection between neuron k in layer $i-1$ and neuron j in layer i. A comprehensive list of commonly used activation functions can be found in Apicella et al. [38].

Output neurons. Finally, the output layer neurons provide the estimates $\hat{\boldsymbol{y}}_i$ of the dependent variable(s) $\boldsymbol{y}_i$.

During network training, a loss function $\mathcal{L}(\mathbf{w})$ is optimized with respect to the weights. This function quantifies the discrepancy between the network output $\hat{\boldsymbol{y}}_i$ and the true value of the dependent variable $\boldsymbol{y}_i$ for each observation, thus guiding the adjustment of the weights to improve predictive accuracy. A comprehensive review of the various optimization methods can be found in the survey by Abdulkadirov, Lyakhov, and Nagornov [39], while the most commonly used loss functions are summarized by Li, Liu, and Liu [40].

The number of neurons in each layer and their pattern of interconnections determine the network's architecture and its potential applications. In the following, we focus on a particular type of architecture, called variational auto-encoders.

3.2. Variational Autoencoders

A variational autoencoder (VAE), illustrated in Figure 2, is a particular type of neural network designed to learn a probabilistic latent representation of the input data while reconstructing the original input from this representation. VAEs have been widely used for dimensionality reduction and latent representation learning.

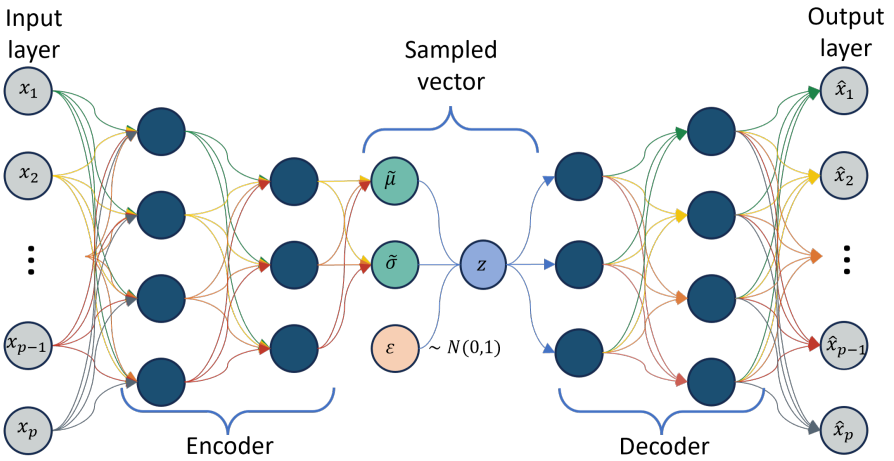

Figure 2. Representation of a Variational Autoencoder. Source: authors' own elaboration.

The architecture of a VAE consists of two main components: an encoder and a decoder, parametrized by $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, respectively. The encoder maps the input data to the parameters of a probability distribution in the latent space, typically represented by a mean vector $\boldsymbol{\mu}$ and a standard deviation vector $\boldsymbol{\sigma}$. A latent vector $\boldsymbol{z}$ is then sampled from this distribution. The decoder receives the latent representation $\boldsymbol{z}$ and attempts to reconstruct the original input.

During training, the network is optimized to minimize the reconstruction error while simultaneously encouraging the latent representations to follow a predefined probability distribution, typically a Gaussian distribution. This is achieved through the maximization of the Evidence Lower Bound (ELBO), as defined in Equation (1):

$$\mathrm{ELBO}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbb{E}_{z \sim q_{\boldsymbol{\phi}}(z|x)} \Big[\ln p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}) \Big] - D_{\mathrm{KL}} \Big(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p(\boldsymbol{z}) \Big), \tag{1}$$

where the first term corresponds to the expected log-likelihood of the data given the latent representation and the second term represents the Kullback–Leibler divergence between the approximate posterior distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})$ and the assumed prior over the latent space, $p(\boldsymbol{z})$.

To balance the trade-off between reconstruction quality and latent regularization, a scaling hyperparameter β can be introduced to weight the Kullback–Leibler divergence term [41]. This formulation follows the β-VAE framework proposed by Higgins et al. [41], in which the Kullback–Leibler divergence term is explicitly weighted to control the trade-off between reconstruction fidelity and latent-space regularization. The modified objective function is formulated as

$$\mathrm{ELBO}_{\beta}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbb{E}_{z \sim q_{\boldsymbol{\phi}}(z|x)} \Big[\ln p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}) \Big] - \beta\, D_{\mathrm{KL}} \Big(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) \parallel p(\boldsymbol{z}) \Big), \tag{2}$$

where $\beta > 0$ controls the degree of regularization applied to the latent space. Setting $\beta > 1$ enforces a stricter constraint on the latent distribution to adhere to the prior $p(\boldsymbol{z})$, whereas $\beta < 1$ relaxes this constraint to prioritize reconstruction fidelity.

The ability of VAEs to learn compact representations of a reference population makes them particularly suitable for anomaly detection tasks. Once the VAE has been trained, observations belonging to the reference population used to train the model are expected to be reconstructed accurately, resulting in low reconstruction errors. In contrast, observations that differ substantially from the patterns learned from the reference population are expected to produce higher reconstruction errors and can therefore be identified as potential anomalies.

3.3. Non-Psychiatric Modeling with Variational Autoencoders

The proposed methodology is based on the assumption that the non-psychiatric population exhibits relatively consistent response patterns, whereas the psychiatric population is more heterogeneous, reflecting the diversity of symptom profiles, severity levels, and behavioral manifestations observed across individuals.

Under this assumption, a VAE is trained exclusively on responses from the non-psychiatric population, with the expectation that it will produce low reconstruction errors for observations that follow the patterns present in the reference population. In contrast, observations that differ substantially from these learned patterns are expected to produce higher reconstruction errors. This difference allows the problem to be formulated as an anomaly detection task, in which deviations from the learned representation of the reference population are used to identify potential clinical cases.

A decision threshold for identifying these potential cases is estimated from the distribution of reconstruction errors observed in the training data. A commonly used approach in anomaly detection problems is to define a threshold of the form $\mu + k \cdot \sigma$, where μ and σ represent the mean and standard deviation of the reconstruction errors, respectively. The value of k controls the trade-off between sensitivity and specificity. Recently, several studies addressing similar problems in other fields have successfully used $\mu + \sigma$ as the detection threshold [36,42–44]. However, a limitation of this approach is that the value of k has no direct interpretation in terms of the proportion of the reference population classified as atypical. Therefore, in this study, the threshold is defined using a percentile-based criterion, with the 90th percentile of the reconstruction errors in the training set used as the cutoff point. This threshold provides a direct and easily interpretable criterion while favoring sensitivity. It also avoids an overly restrictive cutoff that could exclude observations with potentially relevant deviations from the patterns observed in the reference population. This consideration is particularly relevant because the non-psychiatric reference population may include individuals with undiagnosed mental health conditions.

The behavior of the VAE is controlled by several hyperparameters, including the number of layers, the number of neurons per layer, the dimensionality of the latent space and the weight of the Kullback–Leibler divergence.

4. Materials and Methods

This section describes the participants included in the study, as well as the scales and questionnaires used to assess and characterize their mental health status. It also presents the analytical techniques employed to identify individuals who may be at risk of mental health problems, the experimental design adopted and the metrics used to evaluate the performance of the proposed method and the comparison methods.

4.1. Participants

Two samples obtained from different hospitals were used to evaluate the performance of the four classifiers.

4.1.1. First Dataset

A total of 774 individuals participated in the study. This sample size is comparable to those reported in other studies [45–48]. All participants were 18 years of age or older and provided written informed consent prior to participation. Individuals unable to provide informed consent were excluded from the study (e.g., due to the presence of a life-threatening medical condition or significant organic brain disease). The study was approved under research protocol PIC 46/2006 at Fundación Jiménez Díaz hospital.

Of the total sample, 413 participants were blood donors who reported never having been diagnosed with a mental disorder. In this group, 62.5% were male. The mean age was 34.09 ± 10.29 years for men and 33.87 ± 11.67 years for women. This group constituted the non-psychiatric population.

The second group consisted of 269 patients with clinically diagnosed mental disorders and no history of suicide attempts. In this group, 31.46% were male. The mean age was 40.13 ± 15.13 years for men and 35.31 ± 13.69 years for women. This group will be referred to as the mental disorder group.

The third group comprised 92 individuals with a first suicide attempt. The proportion of males was 31.46%. The mean age was 38.41 ± 12.64 years for men and 44.57 ± 15.16 years for women. This group will be referred to as suicide attempters.

4.1.2. Second Dataset

A total of 182 individuals participated in the study, all of them aged 18 or older and admitted to the emergency department of the Puerta de Hierro University Hospital (Madrid, Spain) between 1 June and 1 December 2013. All participants provided written informed consent prior to participation. The study was approved under research protocol PI 108-12 by the Ethics Committee of Puerta de Hierro University Hospital. This sample was previously described by Artieda-Urrutia et al. [49].

Of the total sample, 64 participants presented to the emergency department for medical reasons unrelated to mental health and reported no history of a mental disorder diagnosis. In this group, 40.63% were male. The mean age was 42.19 ± 14.70 years for men and 40.53 ± 15.71 years for women. This group constituted the non-psychiatric population.

The second group consisted of 57 patients who presented to the emergency department for psychiatric reasons other than suicidal behavior and had no history of suicide attempts. In this group, 36.84% were male. The mean age was 41.90 ± 11.58 years for men and 45.50 ± 14.48 years for women. This group will be referred to as the mental disorder group.

The third group comprised 61 individuals who presented to the emergency department after a suicide attempt. The proportion of males was 29.51%. The mean age was 38.67 ± 11.84 years for men and 36.30 ± 11.89 years for women. This group will be referred to as suicide attempters.

4.2. Scales and Questionnaires

A subset of items from the Social Readjustment Rating Scale (SRRS) and the International Personality Disorder Examination Screening Questionnaire (IPDE-SQ) was used [20,21]. These instruments have been previously identified as informative predictors of mental health status [50]. Then, these two instruments and the selected items are briefly described.

IPDE-SC. This scale, developed by Loranger et al. [21], is a screening instrument designed to identify personality traits and behavioural patterns associated with major categories of personality disorders according to international diagnostic classifications. The questionnaire consists of 77 true/false items. In this study, 14 items were selected.

The selected IPDE-SQ items assessed a range of personality-related dimensions. Emotional emptiness and affective instability were captured by item 40. Fear of abandonment and dependency-related concerns were represented by items 33, 34, 42 and 75. Difficulties in emotional regulation, including anger outbursts, were assessed by item 43. Interpersonal suspiciousness and perceived hostility or reputational threat were captured by item 66, while distrust in romantic relationships was reflected in item 36. Identity disturbance was assessed by item 4. Reduced sexual interest was measured by item 31. Perceived social eccentricity or oddness was captured by item 70. Affective instability was reflected in item 26. Reverse-keyed items capturing positive self-evaluation and well-being included items 15 and 1, respectively.

SRRS. The Social Readjustment Rating Scale is a life-events inventory composed of 43 major stressful life events designed to quantify cumulative life stress by assigning weighted scores to each event according to its estimated impact on required psychological readjustment. In this study, a subset of five items was selected to capture relevant stressors.

The selected items included changes in interpersonal conflict frequency, captured by item 19; modifications in personal routines and habits, represented by item 29; marital disruption, assessed by item 3; changes in family structure, reflected in item 14; and health-related stressors, measured by item 6.

4.3. Classifiers

In addition to the proposed VAE-based approach, the following classification techniques are considered for performance comparison. As discussed above, these techniques are selected because they are among the most widely used methods for mental health assessment.

Decision Trees (DTs). DTs are supervised learning models that make predictions by learning a sequence of simple decision rules from the data [51]. These rules are organized in a tree-like structure, where each internal node represents a test on a feature, each branch corresponds to an outcome of that test, and each terminal node (leaf) provides a final prediction.

This tree structure is built by recursively splitting the data at each node to create subsets that are increasingly homogeneous with respect to the target variable, based on selecting the feature and threshold that best separate the observations of each class according to an impurity measure. The impurity of a node, D, is typically measured with the following coefficients:

Gini index:

$$\text{Gini}(D) = 1 - \sum_{k=1}^{K} p_k^2, \tag{3}$$

Shannon entropy:

$$H(D) = -\sum_{k=1}^{K} p_k \log_2(p_k) \tag{4}$$

where p_k represents the proportion of class k present in the node, calculated as $p_k = \dfrac{n_k(D)}{\sum_{m=1}^{K} n_m(D)}$, with $n_m(D)$ being the number of observations of class m in the node D, and K is the total number of classes.

When working with imbalanced classes, the calculation of the proportion of each class in the node to be split can change to take into account this disparity. Each class is assigned a weight, δ_k, and the new proportions $\tilde{p}_k$ are calculated as

$$\tilde{p}_k = \frac{\delta_k n_k(D)}{\sum_{l=1}^{K} \delta_l n_l(D)}.$$

Under the imbalanced scheme, the impurity measures defined in Equations (3) and (4) are calculated using $\tilde{p}_k$ instead of p_k.

Because of their structure, DTs are easy to interpret and can naturally capture non-linear relationships between input features and the target variable.

The behavior of a decision tree is controlled by several hyperparameters that regulate the trade-off between model complexity and generalization performance, helping to prevent overfitting while maintaining interpretability and predictive accuracy. In this work, the hyperparameters considered include the impurity decision criterion (Gini index, entropy); the maximum depth of the tree that is, the longest path from the root node to any leaf; the minimum number of samples required to split an internal node; the minimum number of samples required on any leaf node; the maximum number of features considered when searching for the best split, and the class weighting scheme used to handle potential class imbalance.

Random Forests (RFs). RFs are supervised ensemble learning models that combine the predictions of multiple DTs to improve predictive performance and reduce overfitting [52]. Instead of relying on a single tree, the model constructs a collection of B trees, each of them trained on a different bootstrap sample $\boldsymbol{X}_b$, $b = 1, \ldots, B$, drawn with replacement from the training data. In addition, when evaluating candidate splits, each tree considers only a random subset of m features from the total of p available features. These sources of randomness promote diversity among the trees, reducing the variance of the model and improving its generalization capability.

For classification tasks, each tree produces an individual prediction, and the final class label is determined through majority voting across all trees in the forest, that is, the assigned class is the most frequent class among all trees in the forest. For probabilistic predictions, the forest returns the probability of the positive class as

$$\widehat{P}(y = 1 \mid \boldsymbol{x}) = \frac{1}{B} \sum_{b=1}^{B} \widehat{P}_b(y = 1 \mid \boldsymbol{x}),$$

where $\widehat{P}(y = 1 \mid \boldsymbol{x})$ represents the probability assigned to the positive class by the forest, while $\widehat{P}_b(y = 1 \mid \boldsymbol{x})$, $b = 1, \ldots, B$ is the probability assigned by individual trees. By aggregating the predictions of multiple trees, RFs can capture complex non-linear relationships while being more robust to noise and overfitting than a single DT. However, this ensemble approach sacrifices the interpretability of a single decision tree, resulting in a model harder to interpret.

When addressing imbalanced classes, the weighting scheme operates on each individual tree as explained in the Decision Trees section. Specifically, the weight w_k is assigned to each class when calculating the impurity metrics on each of the trees forming the forest.

The behavior of an RF is controlled by several hyperparameters that regulate both the complexity of the individual trees and the diversity of the ensemble. In this work, the hyperparameters considered include the number of trees in the forest, the maximum number of features considered when searching for the best split, the maximum depth of each tree, the minimum number of samples required to split an internal node, the minimum

number of samples required to form a leaf node, and the class weighting scheme (none or balanced) used to handle potential class imbalance.

Support Vector Machines (SVMs). SVMs are supervised learning models that aim to find the hyperplane that best separates the classes by maximizing the margin M, that is, the distance between the decision boundary and the closest training observation of each class. These observations, which lie closest to the boundary, are known as support vectors [53,54].

Given the set of observations $\{\boldsymbol{x}_i, y_i\}$, with labels $y_i \in \{-1, 1\}$, the SVM in its primal form is formulated as:

$$\begin{aligned} &\underset{\boldsymbol{\omega}, \omega_0, \boldsymbol{\xi}}{\text{minimize}} \quad \frac{\|\boldsymbol{\omega}\|^2}{2} + C \sum_{i=1}^{n} \delta_{y_i} \xi_i \\ &\text{subject to} \\ &y_i(\boldsymbol{\omega}^\top \boldsymbol{x}_i + \omega_0) \geq 1 - \xi_i, \quad i = 1, \ldots, n, , \\ &\xi_i \geq 0, \quad i = 1, \ldots, n, \end{aligned} \tag{5}$$

where $\boldsymbol{\omega}$ is the vector defining the orientation of the hyperplane, ω_0 is the bias term, ξ_i are the slack variables that allow for misclassifications, $C > 0$ is the regularization parameter controlling the trade-off between margin maximization and training error, and δ_{y_i} is the weight assigned to the class y_i to handle potential class imbalance. Once the optimal solutions for (5) are found, $\widehat{\boldsymbol{\omega}}$ and $\widehat{\omega}_0$, the decision function for a new observation is:

$$f(\boldsymbol{x}) = \text{sign}\left(\widehat{\boldsymbol{\omega}}^\top \boldsymbol{x} + \widehat{\omega}_0\right).$$

By solving the problem via Lagrange multipliers $\alpha_i \geq 0$, the parameter vector can be explicitly expressed as a linear combination of the support vectors, $\widehat{\boldsymbol{\omega}} = \sum_{i \in \mathcal{S}} \alpha_i y_i \boldsymbol{x}_i$, where $\mathcal{S}$ denote the set of support vectors. This allows rewriting the linear decision function in terms of inner products:

$$f(\boldsymbol{x}) = \text{sign}\left(\sum_{i \in \mathcal{S}} \alpha_i y_i (\boldsymbol{x}_i^\top \boldsymbol{x}) + \widehat{\omega}_0\right).$$

A larger margin generally leads to better generalization performance, making SVMs particularly effective in high-dimensional feature spaces.

When a linear boundary is not enough to get a proper separation of the classes, a kernel function, K, can be used to implicitly transform the data into a higher-dimensional representation where a separating boundary can be found. In this work, the Radial Basis Function (RBF) kernel is employed:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right), \tag{6}$$

allowing the model to capture complex non-linear relationships between the input features and the target variable. For a new observation $\boldsymbol{x}$, the final predicted class is determined by evaluating the kernel function against the support vectors ($\mathcal{S}$):

$$f(\boldsymbol{x}) = \text{sign}\left(\sum_{i \in \mathcal{S}} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + \widehat{\omega}_0\right) \tag{7}$$

where α_i are the optimal weights assigned to each support vector.

The behavior of an SVM is controlled by several hyperparameters that regulate the complexity of the decision boundary and the trade-off between fitting the training data and generalization performance. In this work, the hyperparameters considered include the regularization parameter, which controls the penalty associated with classification errors, the kernel coefficient, which determines the influence of individual training observations

when using the RBF kernel, and the class weighting scheme (none or balanced) used to handle potential class imbalance.

It is important to note that while basic SVM offers some interpretability through the hyperplane normal vector, $\boldsymbol{\omega}$, the RBF kernel implicitly transforms the data into an infinite-dimensional space, reducing the interpretability of the model in favor of capturing complex relationships between features and labels.

4.4. Hyperparameter Optimization and Tuning Space

To optimize model performance and ensure robust generalization, a comprehensive grid search was conducted across a diverse hyperparameter space. The configuration details for each candidate architecture are specified below:

Variational Autoencoder Model: The neural network architecture explored several hidden-layer configurations to evaluate different depth and compression trajectories. Given the sample size and the number of input variables, compact network architectures were considered to limit model complexity and reduce the risk of overfitting. Specifically, the evaluated hidden-layer configurations were $[16, 4, 2]$, $[8, 4, 2]$, $[16, 8, 4, 2]$, $[16, 8, 4]$, $[16, 12, 8, 4, 2]$, and $[16, 12, 8, 4]$, where the length of each list indicates the number of hidden layers and each element specifies the number of neurons in that layer.

The latent space dimensionality was initially evaluated using both one and two latent dimensions. When using two latent dimensions, the resulting latent representations showed an approximately one-dimensional structure, with the observations concentrated along a single direction and limited variation in the second dimension. This suggested that the second latent dimension provided limited additional information for the representation of the input data. Therefore, a one-dimensional latent space was selected for the final model, providing a more compact representation while avoiding the inclusion of an apparently redundant latent dimension.

Optimization was carried out for a maximum of 1000 epochs with a learning rate of 10^{-3}. Full-batch training was used, so that each optimization step was performed using the complete training set. The Adam optimizer [55] was used for model training, with a weight decay of 0.001. Adam was selected because its adaptive parameter updates facilitate stable and efficient optimization of neural network parameters. The KL-divergence regularization weight was varied across $\beta \in \{0.5, 1, 3, 4, 5, 10\}$. This range includes the standard VAE setting ($\beta = 1$) and values both below and above one, allowing different regularization strengths to be explored. No early-stopping criterion was applied. Instead, a callback monitored the validation loss throughout the 1000 training epochs and retained the model corresponding to the lowest validation loss among all epochs. For each candidate configuration, the mean reconstruction error was computed on the validation fold in each of the five folds and then averaged across folds. The configuration yielding the lowest mean reconstruction error across the five folds was selected as the final configuration.

In this work, the ReLU activation function was used in the hidden layers because of its simplicity and favorable gradient-propagation properties [56], which facilitate the optimization of neural networks. Sigmoid activation was used in the output layer because the input variables are binary, allowing the reconstructed values to be constrained to the $[0, 1]$ interval and interpreted as probabilities.

Decision Tree: Maximum depth was tested at $\{\text{None}, 2, 3, 4, 5, 15, 20, 30\}$. The minimum samples required to split an internal node spanned $\{2, 5, 10, 20, 30\}$, and the minimum leaf samples spanned $\{1, 2, 5, 10, 20\}$. Splitting criteria were evaluated using Gini impurity and Shannon entropy. Class weights were either uniform or balanced.

Random Forest: For the ensemble classifier, the number of trees was evaluated at 200, 400, 600, and 1000. The maximum tree depth ranged across $\{\text{None}, 2, 3, 4, 5, 15, 20, 30\}$.

Internal node splitting required a minimum sample size of 2, 5, 10, 20 or 30, while leaf nodes were constrained to a minimum sample size of 2, 5, 10, 20 or 30. The maximum number of features considered at each split was selected via square root (*sqrt*) or base-2 logarithm (*log2*). Both bootstrapped and non-bootstrapped trees were evaluated, alongside unweighted and class-balanced loss functions.

Support Vector Machine (SVM): The SVM classifier utilized a Radial Basis Function (RBF) kernel. The regularization parameter C was optimized over a log-spaced grid of 12 values ranging from 10^{-3} to 10^{3}. The kernel coefficient γ was systematically varied across library defaults (*scale*, *auto*) and explicit fractional values ($\{0.1, 0.025, 0.01, 0.001\}$). Class-imbalance corrections were evaluated by toggling between uniform and balanced class weights.

All analyses were performed in Python 3.13.12 [57]. The scikit-learn library [58] was used to implement the DT, RF, and SVM models, while PyTorch 2.11.0 [59] was used to implement the VAEs.

4.5. Experiment Design

To compare the performance of the classifiers, a repeated hold-out validation scheme was employed. The procedure was repeated 100 times, with a different random partition generated in each repetition. In each iteration, the available dataset was randomly divided into a training set (80% of the data) and a test set (20% of the data).

A distinction was made between the training data used by the proposed VAE and those used by the supervised classification methods. The VAE was trained exclusively on non-psychiatric participants, as its objective is to learn the underlying data distribution of the non-psychiatric population. In contrast, the supervised classifiers were trained using both psychiatric and non-psychiatric participants. Consequently, the supervised methods had access to a larger number of training observations than the VAE.

To control for the potential effect of training sample size on the comparison, the training data available to the supervised methods were randomly subsampled to match the total number of observations used to train the VAE. This subsampling was performed in a stratified manner, preserving the original proportion of the different participant groups within the supervised training data. Specifically, after the train/test partition, stratified sampling without replacement [60] was applied to the training sets only. The number of observations drawn from each group (non-psychiatric controls, psychiatric controls and individuals with suicidal behavior) was proportional to its size in the original training set. Within each group, each individual had the same probability of being selected, so the removal process was purely random and did not depend on any participant characteristic. This process ensures that the training set for the supervised models contains exactly the same number of observations as the VAE training set while keeping the class proportions intact. By matching the number of observations seen by each model during training, we obtain a fairer comparison, since classifier performance depends on the sample size [61].

The test set was kept identical for all classification methods and included participants from all relevant groups, including non-psychiatric controls, participants with a mental disorder without a history of suicide attempt, and suicide attempters. This common test set was used to evaluate the ability of each method to distinguish the target groups, regardless of the composition of its training data.

For each repetition, the training data were used for hyperparameter tuning through a 5-fold cross-validation procedure. This process was performed exclusively within the training set, without using the test set, to select the optimal hyperparameter configuration for each classifier. For the VAE, in each fold, the training portion was further divided into an inner training subset (80%) and an inner validation subset (20%). The VAE was trained

using the inner training subset, while the callback monitored the validation loss on the inner validation subset. The resulting model was then evaluated on the held-out fold.

Once the optimal hyperparameters had been identified, each model was retrained using the complete training data available to that method and subsequently evaluated on the test set. For the VAE, the same inner 80/20 split of the complete training data was used by the callback during this final training. As explained, all compared models were evaluated using the same test partition in each repetition, ensuring a consistent and fair comparison across methods.

No missing values were present in the first dataset. For the second dataset, participants with missing values had been excluded prior to the analysis described in the present study, as reported in the corresponding study [49]. No scaling or normalization was applied because all questionnaire items were binary. Reverse-coded items were retained in their original coding, as the models can learn the direction of each feature's contribution during training.

4.6. Metrics

To compare the performance of the proposed VAE with that of the three supervised classification techniques, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-score, the area under the ROC (ROC-AUC), and the area under the precision–recall curve (PR-AUC) were calculated on the test set in each repetition. The resulting values were summarized by their mean and standard deviation across repetitions, together with the 2.5th and 97.5th percentiles of their empirical distribution. Formal significance tests were not applied, since the overlap between the training and test sets of different repetitions violates the independence assumption of standard tests and leads to an underestimation of the variance [62,63].

The accuracy, sensitivity, and specificity are derived from the entries of the confusion matrix, namely true positives, true negatives, false positives, and false negatives. For the three supervised techniques, these quantities are directly derived from their predictions: the majority class of the leaf nodes in which the observations fall for DTs, the majority vote among the trees for RFs, and the predicted class labels provided by the SVM.

For the VAE, the predicted class labels are obtained by applying the previously defined decision threshold to the reconstruction error. Observations with a reconstruction error greater than or equal to this threshold are classified as positive, whereas those with a lower reconstruction error are classified as negative.

Based on the confusion matrix entries in Table 1, the performance metrics are defined as follows:

Table 1. Generic confusion matrix for a binary classification problem.

		True class	
		1	0
Predicted class	1	TP	FP
	0	FN	TN

Accuracy corresponds to the proportion of correct predictions with respect to the total number of observations. It is calculated as the sum of true positives and true negatives divided by the total number of observations:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Sensitivity (recall or true positive rate) measures the model's ability to correctly identify positive cases. It is a particularly relevant metric in screening processes. It is defined as the proportion of true positives with respect to the total number of actual positives:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Specificity (true negative rate) measures the model's ability to correctly identify negative cases. It is defined as the proportion of true negatives with respect to the total number of actual negatives. This metric complements sensitivity, as both evaluate the model's performance from opposite perspectives: sensitivity for positive cases and specificity for negative cases:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Positive predictive value (PPV or precision) measures the proportion of correct positive predictions relative to all positive predictions made by the model. It evaluates the reliability of a positive result, making it a particularly relevant metric in scenarios where false positives carry a high cost. It is defined as the proportion of true positives with respect to the total number of predicted positives:

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Negative predictive value (NPV) measures the proportion of correct negative predictions relative to all negative predictions made by the model. It evaluates the reliability of a negative result, making it an essential metric in clinical and screening contexts where ruling out a condition with confidence is required. It is defined as the proportion of true negatives with respect to the total number of predicted negatives:

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

F1-score conveys the balance between precision and recall by computing their harmonic mean. It is a particularly useful metric when dealing with imbalanced class distributions, where evaluating accuracy alone can be misleading. It is defined as the harmonic mean of sensitivity and positive predictive value:

$$\text{F1-score} = 2 \cdot \frac{\text{PPV} \cdot \text{Sensitivity}}{\text{PPV} + \text{Sensitivity}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}$$

Finally, the ROC curve is a measure of the model's ability to discriminate between the two classes. Unlike the previous metrics, it does not depend on a single decision threshold but evaluates the performance of the model across all possible cutoff points. To construct the ROC curve, a continuous classification score is required for each observation. For the VAE, the reconstruction error is used as the classification score, with higher values indicating a greater likelihood of belonging to the positive class. For the DT, the proportion of positive-class observations in the terminal leaf node is used as the classification score. For the RF, the proportion of trees voting for the positive class is used as the classification score. For the SVM, the decision score generated by the fitted model is used as the classification score.

For each method, the ROC curve is obtained by varying the corresponding decision threshold across the range of classification scores and computing the resulting true positive rate (sensitivity) and false positive rate (1-specificity). The ROC-AUC corresponds to the area under this curve and can be interpreted as the probability that the model assigns a higher score to a randomly selected positive observation than to a randomly selected

negative observation. A value close to 1 indicates excellent discriminative ability, whereas a value close to 0.5 indicates performance close to random classification.

Similarly, the precision–recall curve is obtained by varying the decision threshold across the range of classification scores and computing the resulting PPV (precision) and sensitivity (recall). The PR-AUC corresponds to the area under this curve. Unlike the ROC-AUC, whose baseline value for a random classifier is 0.5, the baseline value of the PR-AUC equals the proportion of positive observations in the data, which makes it particularly informative in the presence of class imbalance.

5. Results

The results obtained using the experimental set-up described above, considering 100 repetitions, are presented below.

5.1. First Dataset Results

Table 2 summarizes the performance of the three supervised techniques (DT, RF, and SVM) and the proposed unsupervised VAE-based approach across 100 repetitions. In terms of accuracy, DT achieved the lowest performance (0.83 ± 0.04), whereas RF, SVM, and VAE obtained similar values (0.88 ± 0.02, 0.88 ± 0.03, and 0.88 ± 0.02, respectively). A similar pattern was observed for ROC-AUC, with DT showing the lowest value (0.88 ± 0.04) and the remaining models achieving values between 0.94 and 0.95.

Table 2. Performance comparison across models for the IPDE-SRRS questionnaires in the first dataset (Fundación Jiménez Díaz). Values are reported as mean ± standard deviation across the 100 repetitions, with the 2.5th–97.5th percentile interval shown below.

Model	DT	RF	SVM	VAE
Accuracy	0.83 ± 0.04 [0.76, 0.90]	0.88 ± 0.02 [0.84, 0.93]	0.88 ± 0.03 [0.83, 0.93]	0.88 ± 0.02 [0.83, 0.92]
Specificity	0.86 ± 0.06 [0.76, 0.95]	0.91 ± 0.04 [0.82, 0.96]	0.91 ± 0.04 [0.81, 0.98]	0.90 ± 0.04 [0.82, 0.98]
Sensitivity	0.80 ± 0.07 [0.66, 0.91]	0.85 ± 0.05 [0.75, 0.94]	0.85 ± 0.05 [0.75, 0.93]	0.85 ± 0.04 [0.76, 0.91]
ROC-AUC	0.88 ± 0.04 [0.78, 0.94]	0.94 ± 0.02 [0.91, 0.97]	0.95 ± 0.02 [0.92, 0.98]	0.95 ± 0.02 [0.92, 0.97]
PR-AUC	0.88 ± 0.04 [0.81, 0.94]	0.94 ± 0.02 [0.92, 0.97]	0.95 ± 0.02 [0.89, 0.97]	0.95 ± 0.01 [0.92, 0.97]
PPV	0.84 ± 0.05 [0.74, 0.94]	0.89 ± 0.04 [0.81, 0.96]	0.89 ± 0.04 [0.80, 0.97]	0.89 ± 0.04 [0.80, 0.97]
NPV	0.83 ± 0.05 [0.76, 0.92]	0.88 ± 0.04 [0.81, 0.94]	0.87 ± 0.03 [0.81, 0.93]	0.87 ± 0.03 [0.82, 0.92]
F1	0.82 ± 0.04 [0.74, 0.89]	0.87 ± 0.03 [0.82, 0.93]	0.87 ± 0.03 [0.82, 0.92]	0.87 ± 0.03 [0.82, 0.91]

Regarding specificity, RF and SVM achieved the highest values (0.91 ± 0.04), followed by the VAE (0.90 ± 0.04), while DT obtained 0.86 ± 0.06. The sensitivity was 0.85 ± 0.05 for RF, SVM, and VAE, and 0.80 ± 0.07 for DT. The remaining metrics (PR-AUC, PPV, NPV, and F1-score) followed the same pattern, with the VAE, RF, and SVM achieving practically identical values and DT obtaining slightly lower values. The 2.5th–97.5th percentile intervals of the VAE, RF, and SVM overlapped widely for all metrics. Overall, these results indicate that the proposed VAE-based approach achieved a performance

comparable to the obtained by the supervised techniques, despite being trained exclusively on non-psychiatric participants.

Figure 3 shows the distribution of VAE reconstruction errors for non-psychiatric participants in the training and test sets, and for participants with mental disorders and suicide attempts in the test set, for one representative repetition.

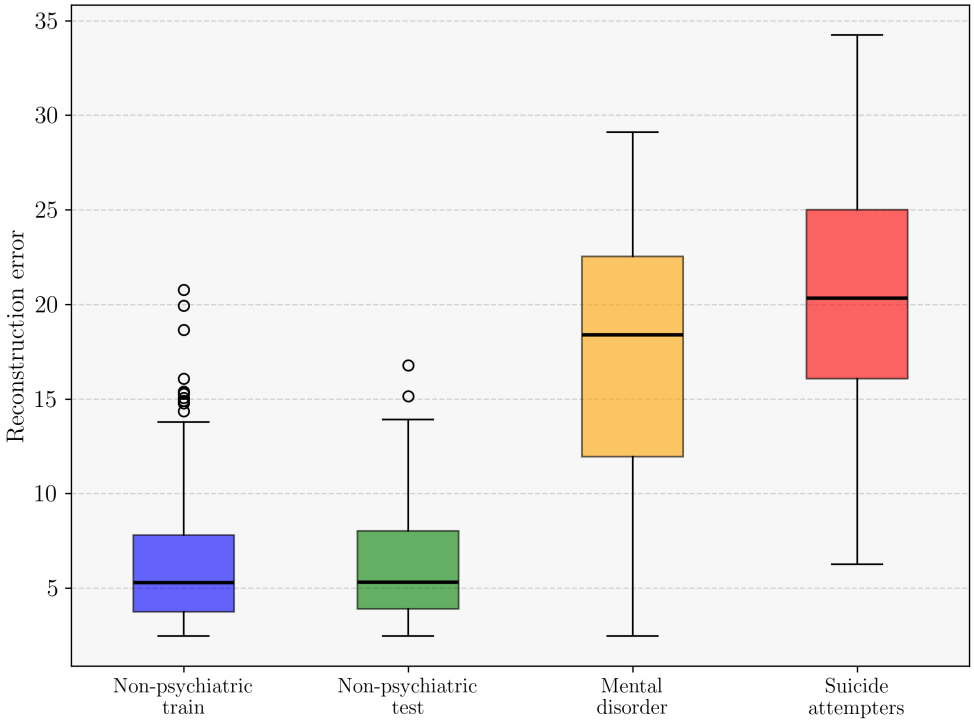

Figure 3. First dataset. Distribution of VAE reconstruction errors for non-psychiatric controls in the training and test sets, and for participants with mental disorders and suicide attempters in the test set, in one representative repetition. Source: authors' own elaboration based on the study data.

It is observed that both the training and test samples from the non-psychiatric population exhibited substantially lower reconstruction errors than the psychiatric groups. In addition, several outliers were identified within the non-psychiatric groups. It should be noted that this cohort consisted of blood donors who reported not having received a diagnosis of mental disorder. Therefore, although these individuals were considered part of the non-psychiatric population, the absence of underlying mental health conditions cannot be guaranteed with complete certainty. The elevated reconstruction errors observed in these cases may reflect individuals whose characteristics were not fully captured by the dominant patterns learned by the model or, alternatively, participants with unrecognized or undiagnosed mental health conditions. However, distinguishing between these possibilities would require further clinical evaluation.

Figure 4 shows the corresponding reconstruction-error distributions across all 100 repetitions. To improve the visualization, observations identified as outliers were excluded from the figure. This was done solely for visualization purposes and did not involve removing these observations from the analysis. A large number of points would otherwise correspond to the same observations being identified as outliers repeatedly across the 100 repetitions, resulting in substantial overlap and obscuring the overall distributions.

The reconstruction-error distributions of the non-psychiatric training and test sets were nearly identical, suggesting that the VAE generalized well to unseen individuals from the same population. Moreover, psychiatric inpatients without a recent suicide attempt showed markedly higher reconstruction errors, while the highest errors were observed among

patients who had attempted suicide within the previous 48 h. This progressive increase in reconstruction error across groups suggests that the latent representations learned from the non-psychiatric population capture deviations from the patterns observed in the reference population, with larger deviations observed in groups with greater clinical complexity.

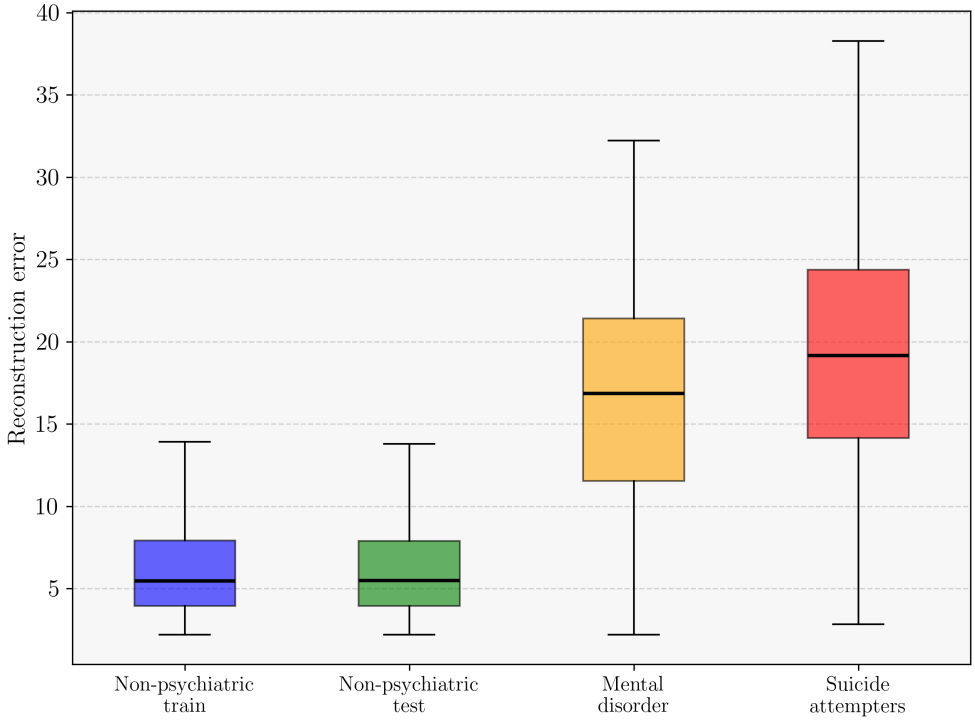

Figure 4. First dataset. Distribution of VAE reconstruction errors for non-psychiatric controls in the training and test sets, and for participants with mental disorders and suicide attempters in the test set, aggregated across the 100 repeated train/test splits. Source: authors' own elaboration based on the study data.

5.2. Second Dataset Results

Table 3 summarizes the results obtained in the second dataset. RF achieved the highest values for most metrics, with an accuracy of 0.90 ± 0.05, a sensitivity of 0.93 ± 0.06, and a ROC-AUC of 0.96 ± 0.03. The VAE obtained an accuracy of 0.86 ± 0.06 and a ROC-AUC of 0.94 ± 0.04, similar to those of SVM (0.87 ± 0.06 and 0.95 ± 0.03) and higher than those of DT (0.83 ± 0.07 and 0.87 ± 0.07). The sensitivity of the VAE (0.89 ± 0.07) was between those of SVM (0.87 ± 0.09) and RF, while its specificity (0.81 ± 0.11) was similar to that of DT (0.80 ± 0.12) and slightly lower than those of RF (0.84 ± 0.10) and SVM (0.86 ± 0.11).

Because this dataset is smaller, the variability across repetitions was larger than in the first dataset, and the percentile intervals of all methods overlapped widely. In addition, the positive class (participants with mental disorders and suicide attempters) represented approximately 65% of this sample, which explains the high PR-AUC values obtained by all methods.

As shown in Figure 5, the reconstruction errors followed the same pattern observed in the first dataset: the non-psychiatric training and test sets showed similar and low reconstruction errors, whereas participants with mental disorders and suicide attempters showed markedly higher errors, with the highest median among suicide attempters.

Table 3. Performance comparison across models for the IPDE-SRRS questionnaires in the second dataset (Puerta de Hierro University Hospital). Values are reported as mean ± standard deviation across the 100 repetitions, with the 2.5th–97.5th percentile interval shown below.

Model	DT	RF	SVM	VAE
Accuracy	0.83 ± 0.07 [0.70, 0.95]	0.90 ± 0.05 [0.77, 0.97]	0.87 ± 0.06 [0.73, 0.96]	0.86 ± 0.06 [0.74, 0.97]
Specificity	0.80 ± 0.12 [0.54, 1.00]	0.84 ± 0.10 [0.61, 1.00]	0.86 ± 0.11 [0.62, 1.00]	0.81 ± 0.11 [0.58, 1.00]
Sensitivity	0.85 ± 0.09 [0.64, 0.98]	0.93 ± 0.06 [0.81, 1.00]	0.87 ± 0.09 [0.64, 1.00]	0.89 ± 0.07 [0.73, 1.00]
ROC-AUC	0.87 ± 0.07 [0.73, 0.97]	0.96 ± 0.03 [0.89, 1.00]	0.95 ± 0.03 [0.88, 1.00]	0.94 ± 0.04 [0.85, 0.99]
PR-AUC	0.93 ± 0.04 [0.84, 0.99]	0.98 ± 0.02 [0.94, 1.00]	0.97 ± 0.02 [0.93, 1.00]	0.97 ± 0.02 [0.92, 0.99]
PPV	0.89 ± 0.06 [0.76, 1.00]	0.91 ± 0.05 [0.80, 1.00]	0.92 ± 0.05 [0.81, 1.00]	0.90 ± 0.05 [0.79, 1.00]
NPV	0.76 ± 0.11 [0.55, 0.96]	0.87 ± 0.09 [0.69, 1.00]	0.80 ± 0.11 [0.59, 1.00]	0.80 ± 0.10 [0.61, 1.00]
F1	0.87 ± 0.06 [0.74, 0.96]	0.92 ± 0.04 [0.83, 0.98]	0.89 ± 0.05 [0.77, 0.96]	0.89 ± 0.05 [0.79, 0.98]

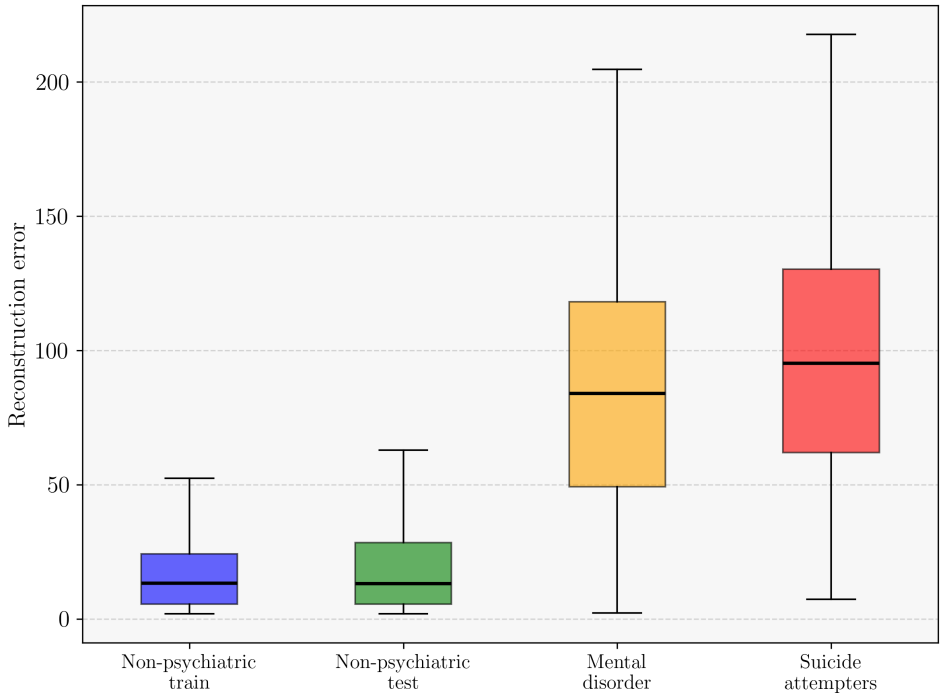

Figure 5. Second dataset. Distribution of VAE reconstruction errors for non-psychiatric controls in the training and test sets, and for participants with mental disorders and suicide attempters in the test set, aggregated across the 100 repeated train/test splits. Source: authors' own elaboration based on the study data.

5.3. Sensitivity to the Reconstruction-Error Threshold

To assess the influence of the decision threshold on the performance of the VAE, the 90th-percentile threshold used in the main analysis was compared with alternative percentile-based thresholds corresponding to the 80th, 85th, and 95th percentiles of the

reconstruction errors of the non-psychiatric training population, and with a $\mu + \sigma$ criterion, where μ and σ denote the mean and standard deviation of the reconstruction errors of the same population (Table 4).

Table 4. Performance of the VAE using alternative reconstruction-error thresholds. Values represent the mean across the 100 repeated train/test partitions. The threshold used in the main analysis (90th percentile) is shown in bold.

Threshold	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	ROC-AUC
First dataset							
$\mu + \sigma$	0.87	0.88	0.86	0.85	0.90	0.87	0.95
P80	0.86	0.92	0.80	0.81	0.92	0.86	0.95
P85	0.87	0.89	0.85	0.84	0.90	0.87	0.95
P90	**0.88**	**0.85**	**0.90**	**0.89**	**0.87**	**0.87**	**0.95**
P95	0.87	0.77	0.95	0.93	0.83	0.84	0.95
Second dataset							
$\mu + \sigma$	0.86	0.91	0.77	0.88	0.84	0.90	0.94
P80	0.86	0.94	0.72	0.86	0.87	0.90	0.94
P85	0.86	0.92	0.76	0.88	0.84	0.90	0.94
P90	**0.86**	**0.89**	**0.81**	**0.90**	**0.80**	**0.89**	**0.94**
P95	0.85	0.84	0.87	0.93	0.76	0.88	0.94

In both datasets, the results showed the expected trade-off between sensitivity and specificity. In the first dataset, the 80th percentile yielded the highest sensitivity (0.92) and the lowest specificity (0.80), whereas the 95th percentile yielded the lowest sensitivity (0.77) and the highest specificity (0.95). A similar pattern was observed in the second dataset, where sensitivity ranged from 0.94 to 0.84 and specificity from 0.72 to 0.87. Since the ROC-AUC does not depend on the decision threshold, the discriminative capacity of the VAE, comparable to that of RF and SVM in both datasets, is not affected by this choice. The 90th percentile, used in the main analysis, provides an intermediate operating point between sensitivity and specificity in both datasets.

6. Discussion

In this work, we proposed the identification of individuals with potential mental health disorders using a VAE framework, where the non-psychiatric population is modeled as a latent normal distribution and individuals with mental health disorders are represented as deviations from this distribution, characterized by higher reconstruction errors.

The results showed that the proposed approach achieved performance comparable to that of three widely used supervised machine learning techniques (DT, RF, and SVM). Importantly, this was accomplished without requiring data from individuals with mental health disorders during training. This represents a significant advantage, as the collection of large, representative, and accurately labeled psychiatric datasets is often challenging due to practical, ethical, and privacy-related constraints. In contrast, the proposed approach only requires data from the non-psychiatric population, making it potentially easier to deploy and adapt across different settings and populations.

An additional relevant finding was the relatively high sensitivity achieved by the VAE-based model. From a mental health screening perspective, sensitivity is a particularly important metric, as it reflects the ability to correctly identify individuals who may require further clinical assessment. In both datasets, the VAE achieved a sensitivity comparable to that of the supervised models (0.85 versus 0.80–0.85 in the first dataset, and 0.89 versus 0.85–0.93 in the second). This is particularly relevant in suicide prevention. Previous studies have shown that a substantial proportion of individuals who attempt or die by suicide

have sought help from primary care services in the months preceding the suicidal act, highlighting a critical window for early detection and intervention [64]. In this context, the proposed approach could be integrated into primary care or community-based screening settings to support the early identification of individuals exhibiting atypical response patterns indicative of increased vulnerability, thereby facilitating earlier preventive and clinical actions.

The ROC-AUC and PR-AUC, which evaluate performance across all possible cutoffs, were practically identical for the VAE, RF, and SVM in both datasets, indicating that the discriminative ability of the VAE does not depend on the specific threshold selected. The threshold should therefore be understood as an operational component of the screening strategy: lower thresholds increase sensitivity at the expense of specificity, which may be preferable in screening settings, whereas higher thresholds favor specificity.

Although the results were consistent across two hospital samples, further external validation involving multiple hospitals, populations, and countries is required before the proposed approach can be considered for broader clinical or community screening applications.

A particularly relevant finding was the progressive increase in reconstruction error across groups. Specifically, reconstruction errors were substantially lower in the non-psychiatric population, intermediate among participants with mental disorders without a history of suicide attempts, and highest among suicide attempters. This pattern suggests an association between higher reconstruction errors and greater clinical complexity across the study groups.

From a clinical perspective, the proposed VAE-based framework could be used as a supportive tool for early screening and risk stratification rather than as a diagnostic system. By assigning higher reconstruction errors to individuals who deviate more strongly from the non-psychiatric population, the model may help prioritize cases that require further assessment or closer monitoring. This is particularly relevant in resource-constrained settings, where rapid identification of individuals at higher risk could contribute to more efficient allocation of clinical attention. In addition, the proposed approach could potentially be adapted to large-scale screening contexts, such as educational or community settings, where early detection of individuals who may benefit from professional evaluation is of interest. Importantly, the model is intended to complement, rather than replace, standard clinical evaluation and decision-making processes.

An additional finding was the presence of outliers within the non-psychiatric reference population. Although the absence of a previous psychiatric diagnosis does not guarantee the absence of undiagnosed mental health conditions, this limitation is inherent to the definition of reference populations in mental health research. Despite this possibility, the proposed approach showed consistent separation between the predefined groups across both datasets and achieved performance comparable to that of the supervised techniques. The observed outliers may reflect individuals whose characteristics were not adequately captured by the patterns learned by the VAE or, alternatively, individuals with unrecognized or undiagnosed mental health conditions. Distinguishing between these possibilities would require further clinical assessment and represents an interesting direction for future research.

A limitation of the proposed approach is that its performance depends on the extent to which the input features consistently capture a coherent underlying construct. Since the model learns the structure of the non-psychiatric population from questionnaire responses, the presence of heterogeneous or mixed constructs within the selected items may reduce the effectiveness of the learned latent representation. In such cases, reconstruction errors may reflect variability associated with different psychological dimensions, some of which

may be less directly related to mental health conditions, potentially reducing the specificity of the reconstruction error as an indicator of deviations associated with mental health. Consequently, the method may perform less effectively when applied to feature sets that combine multiple, partially unrelated constructs, highlighting the importance of careful feature selection and construct validity in the design of the input data.

Future work will focus on improving the robustness of the proposed framework by incorporating feature selection and representation learning mechanisms that can better identify the underlying latent constructs captured by the data. In particular, approaches aimed at disentangling different psychological dimensions could help isolate the most relevant trait for the detection task, while reducing the influence of items that introduce noise or reflect unrelated constructs.

7. Conclusions

This study has shown the potential of a VAE-based framework for the identification of individuals with potential mental health disorders using data exclusively from a non-psychiatric population. By learning the characteristics of the reference population and using reconstruction errors to identify deviations from this learned pattern, the proposed approach achieved performance comparable to the three supervised machine learning methods considered (DTs, RFs, and SVMs), without requiring data from individuals with mental health disorders during training. The progressive increase in reconstruction errors across populations may reflect increasing deviations from the patterns observed in the non-psychiatric population across groups with greater clinical complexity.

Overall, these findings support the use of unsupervised VAE-based approaches as a complementary strategy for mental health screening, particularly in settings where the collection of large, representative, and accurately labeled psychiatric datasets is challenging. Although the results were consistent across two hospital samples, further external validation involving multiple hospitals, populations, and countries is required before the proposed approach can be considered for broader clinical or community screening applications.

Author Contributions: Conceptualization, J.d.D.S.-A., E.C., I.P.-C., E.B.-G., E.P.-G., and D.D.-G.; methodology, J.d.D.S.-A., E.C. and D.D.-G.; software, J.d.D.S.-A. and E.C.; formal analysis, E.C.; investigation, J.d.D.S.-A., E.B.-G., I.P.-C., and E.P.-G.; resources, E.B.-G. and H.B.-F.; data curation, J.d.D.S.-A.; writing—original draft preparation, E.C. and D.D.-G.; writing—review and editing, J.d.D.S.-A., E.C. and D.D.-G.; supervision, D.D.-G. All authors have read and agreed to the published version of the manuscript.

Funding: Project PID2024-158631OB-I00 funded by MICIU/AEI/10.13039/501100011033/FEDER, UE. This research was supported by CIBER-Consorcio Centro de Investigación Biomédica en Red-(CB/07/09/0025), the Instituto de Salud Carlos III with the support of the European Regional Development Fund (ISCIII PI23/00614), Project PMP24/00026, funded by the Carlos III Health Institute (ISCIII) and the "European Union NextGenerationEU/Recovery and Resilience Facility (RRF)/PRTR." by Fundació La Marató de TV3 (202226-31) and by CaixaResearch Health 2023 LCF/PR/HR23/52430033.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki. The first dataset was approved by the Ethics Committee of Fundación Jiménez Díaz (protocol code PIC 46/2006, date of approval 22 November 2005). The second dataset was approved by the Ethics Committee of Puerta de Hierro University Hospital (protocol code PI 108-12, meeting number 285, date of approval 25 February 2013).

Informed Consent Statement: Informed consent was obtained from all subjects involved in both studies. Written informed consent has been obtained from the patients to publish this paper.

Data Availability Statement: The datasets analyzed in this study are not publicly available due to restrictions imposed by the data provider. The data were obtained from a hospital collaborator and cannot be redistributed or made publicly available without the provider's authorization. Researchers interested in accessing the data may contact EBG for further information.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*; American Psychiatric Association: Washington, DC, USA, 2013.
2. World Health Organization. Mental Disorders. 2022. Available online: https://www.who.int/news-room/fact-sheets/detail/mental-disorders (accessed on 10 August 2026).
3. World Health Organization. COVID-19 Pandemic Triggers 25% Increase in Prevalence of Anxiety and Depression Worldwide. 2022. Available online: https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide (accessed on 10 August 2026).
4. Arias, D.; Saxena, S.; Verguet, S. Quantifying the global burden of mental disorders and their economic value. *eClinicalMedicine* **2022**, *54*, 101675. [CrossRef] [PubMed]
5. World Health Organization. Mental Health. 2024. Available online: https://www.who.int/health-topics/mental-health (accessed on 10 August 2026).
6. The Lancet Global Health. Mental health matters. *Lancet Glob. Health* **2020**, *8*, e1352. [CrossRef] [PubMed]
7. Voros, V.; Fekete, S.; Tenyi, T.; Rihmer, Z.; Szili, I.; Osvath, P. Untreated depressive symptoms significantly worsen quality of life in old age and may lead to the misdiagnosis of dementia: A cross-sectional study. *Ann. Gen. Psychiatry* **2020**, *19*, 52. [CrossRef] [PubMed]
8. Dell'Osso, B.; Glick, I.D.; Baldwin, D.S.; Altamura, A.C. Can long-term outcomes be improved by shortening the duration of untreated illness in psychiatric disorders? A conceptual framework. *Psychopathology* **2012**, *46*, 14–21. [CrossRef] [PubMed]
9. Weaver, N.D.; Bertolacci, G.J.; Rosenblad, E.; Ghoba, S.; Cunningham, M.; Ikuta, K.S.; Moberg, M.E.; Mougin, V.; Han, C.; Wool, E.E.; et al. Global, regional, and national burden of suicide, 1990–2021: A systematic analysis for the Global Burden of Disease Study 2021. *Lancet Public Health* **2025**, *10*, e189–e202. [CrossRef] [PubMed]
10. Arensman, E.; Scott, V.; De Leo, D.; Pirkis, J. Suicide and suicide prevention from a global perspective. *Crisis* **2020**, *41*, S3–S7. [CrossRef] [PubMed]
11. Tahmazov, E.; Bosse, J.; Glemain, B.; Nabbe, P.; Guillou, M.; Blachier, A.; Walter, M.; Lemey, C. Impact of Early Intervention for Early Psychosis on Suicidal Behavior—A Meta-Analysis. *Acta Psychiatr. Scand.* **2025**, *151*, 127–141. [CrossRef] [PubMed]
12. Sediqzadah, S.; Portnoy, A.; Kim, J.J.; Keshavan, M.; Pandya, A. Cost-effectiveness of early intervention in psychosis: A modeling study. *Psychiatr. Serv.* **2022**, *73*, 970–977. [CrossRef] [PubMed]
13. Bradford, A.; Meyer, A.N.; Khan, S.; Giardina, T.D.; Singh, H. Diagnostic error in mental health: A review. *BMJ Qual. Saf.* **2024**, *33*, 663–672. [CrossRef] [PubMed]
14. AlSalem, M.; AlHarbi, M.A.; Badeghiesh, A.; Tourian, L. Accuracy of initial psychiatric diagnoses given by nonpsychiatric physicians: A retrospective chart review. *Medicine* **2020**, *99*, e23708. [PubMed]
15. Potts, M.K.; Burnam, M.A.; Wells, K.B. Gender differences in depression detection: A comparison of clinician diagnosis and standardized assessment. *Psychol. Assess. J. Consult. Clin. Psychol.* **1991**, *3*, 609–615. [CrossRef]
16. MacKinnon, R.A.; Michels, R.; Buckley, P.J. *The Psychiatric Interview in Clinical Practice*; American Psychiatric Pub: Washington, DC, USA, 2015.
17. Grant, J.B.; Bruce, C.; Batterham, P.J. Predictors of personal, perceived and self-stigma towards anxiety and depression. *Epidemiol. Psychiatr. Sci.* **2016**, *25*, 247–254. [CrossRef] [PubMed]
18. Graves, L.V.; Hamill, S.; Larry, M.; Williams, D. Informant characteristics influence reports of participant functioning and their associations with neuropsychological performance in non-Hispanic Black adults. *Arch. Clin. Neuropsychol.* **2023**, *38*, 1054–1067. [CrossRef] [PubMed]
19. Beck, A.T.; Steer, R.A.; Brown, G.K. Beck depression inventory II. *Psychol. Assess.* **1996**. [CrossRef]
20. Holmes, T.H.; Rahe, R.H. The social readjustment rating scale. *J. Psychosom. Res.* **1967**, *11*, 213–218. [CrossRef] [PubMed]
21. Loranger, A.W.; Sartorius, N.; Andreoli, A.; Berger, P.; Buchheim, P.; Channabasavanna, S.; Coid, B.; Dahl, A.; Diekstra, R.F.; Ferguson, B.; et al. The international personality disorder examination: The World Health Organization/Alcohol, Drug Abuse, and Mental Health Administration international pilot study of personality disorders. *Arch. Gen. Psychiatry* **1994**, *51*, 215–224. [PubMed]
22. Spitzer, R.L.; Kroenke, K.; Williams, J.B.; Patient Health Questionnaire Primary Care Study Group. Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study. *JAMA* **1999**, *282*, 1737–1744. [PubMed]

23. Eack, S.M.; Greeno, C.G.; Lee, B.J. Limitations of the Patient Health Questionnaire in identifying anxiety and depression in community mental health: Many cases are undetected. *Res. Soc. Work. Pract.* **2006**, *16*, 625–631. [CrossRef] [PubMed]
24. Kaiser, B.N.; Ticao, C.; Anoje, C.; Boglosa, J.; Gafaar, T.; Minto, J.; Kohrt, B.A. Challenges in simultaneous validation of mental health screening tools in multiple languages: Adolescent assessments in Hausa and Pidgin in Nigeria. *SSM-Ment. Health* **2022**, *2*, 100168. [CrossRef] [PubMed]
25. Sharma, S.K.; Alutaibi, A.I.; Khan, A.R.; Tejani, G.G.; Ahmad, F.; Mousavirad, S.J. Early detection of mental health disorders using machine learning models using behavioral and voice data analysis. *Sci. Rep.* **2025**, *15*, 16518. [CrossRef] [PubMed]
26. Iyortsuun, N.K.; Kim, S.H.; Jhon, M.; Yang, H.J.; Pant, S. A review of machine learning and deep learning approaches on mental health diagnosis. *Healthcare* **2023**, *11*, 285. [CrossRef] [PubMed]
27. Shatte, A.B.; Hutchinson, D.M.; Teague, S.J. Machine learning in mental health: A scoping review of methods and applications. *Psychol. Med.* **2019**, *49*, 1426–1448. [CrossRef] [PubMed]
28. Pigoni, A.; Delvecchio, G.; Turtulici, N.; Madonna, D.; Pietrini, P.; Cecchetti, L.; Brambilla, P. Machine learning and the prediction of suicide in psychiatric populations: A systematic review. *Transl. Psychiatry* **2024**, *14*, 140. [CrossRef] [PubMed]
29. Sajjadian, M.; Lam, R.W.; Milev, R.; Rotzinger, S.; Frey, B.N.; Soares, C.N.; Parikh, S.V.; Foster, J.A.; Turecki, G.; Müller, D.J.; et al. Machine learning in the prediction of depression treatment outcomes: A systematic review and meta-analysis. *Psychol. Med.* **2021**, *51*, 2742–2751. [CrossRef] [PubMed]
30. Cavus, N.; Lawan, A.A.; Ibrahim, Z.; Dahiru, A.; Tahir, S.; Abdulrazak, U.I.; Hussaini, A. A systematic literature review on the application of machine-learning models in behavioral assessment of autism spectrum disorder. *J. Pers. Med.* **2021**, *11*, 299. [CrossRef] [PubMed]
31. Caselles-Pina, L.; Quesada-López, A.; Sújar, A.; Hernandez, E.M.G.; Delgado-Gómez, D. A systematic review on the application of machine learning models in psychometric questionnaires for the diagnosis of attention deficit hyperactivity disorder. *Eur. J. Neurosci.* **2024**, *60*, 4115–4127. [CrossRef] [PubMed]
32. Coppersmith, D.D.; Kleiman, E.M.; Millner, A.J.; Wang, S.B.; Arizmendi, C.; Bentley, K.H.; DeMarco, D.; Fortgang, R.G.; Zuromski, K.L.; Maimone, J.S.; et al. Heterogeneity in suicide risk: Evidence from personalized dynamic models. *Behav. Res. Ther.* **2024**, *180*, 104574. [CrossRef] [PubMed]
33. Arnett, A.B.; Flaherty, B.P. A framework for characterizing heterogeneity in neurodevelopmental data using latent profile analysis in a sample of children with ADHD. *J. Neurodev. Disord.* **2022**, *14*, 45. [CrossRef] [PubMed]
34. Nagrath, P.; Saini, I.; Zeeshan, M.; Komal.; Kalla, D. Predicting Mental Health Disorders with Variational Autoencoders. In *Proceedings of the International Conference on Data Analytics & Management*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 38–51.
35. Geenjaar, E.; Lewis, N.; Fu, Z.; Venkatdas, R.; Plis, S.; Calhoun, V. Fusing multimodal neuroimaging data with a variational autoencoder. In *Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*; IEEE: New York, NY, USA, 2021; pp. 3630–3633.
36. Shan, L.; Li, Y.; Jiang, H.; Zhou, P.; Niu, J.; Liu, R.; Wei, Y.; Peng, J.; Yu, H.; Sha, X.; et al. Abnormal ECG detection based on an adversarial autoencoder. *Front. Physiol.* **2022**, *13*, 961724. [CrossRef] [PubMed]
37. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; The MIT Press: Cambridge, MA, USA, 2016.
38. Apicella, A.; Donnarumma, F.; Isgrò, F.; Prevete, R. A survey on modern trainable activation functions. *Neural Netw.* **2021**, *138*, 14–32. [CrossRef] [PubMed]
39. Abdulkadirov, R.; Lyakhov, P.; Nagornov, N. Survey of optimization algorithms in modern neural networks. *Mathematics* **2023**, *11*, 2466. [CrossRef]
40. Li, C.; Liu, K.; Liu, S. A Survey of Loss Functions in Deep Learning. *Mathematics* **2025**, *13*, 2417. [CrossRef]
41. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
42. Abeku, T.A.; Hay, S.I.; Ochola, S.; Langi, P.; Beard, B.; de Vlas, S.J.; Cox, J. Malaria epidemic early warning and detection in African highlands. *Trends Parasitol.* **2004**, *20*, 400–405. [CrossRef] [PubMed]
43. Alzamel, M. Deep recurrent models for forecasting infectious diseases. *Front. Public Health* **2025**, *13*, 1726819. [CrossRef] [PubMed]
44. Heidari, R.; Samavati, F.F. A new dissimilarity metric for anomaly detection in management zones delineation constructed from time-varying satellite images. *Agriculture* **2024**, *14*, 688. [CrossRef]
45. Oh, J.; Yun, K.; Hwang, J.H.; Chae, J.H. Classification of Suicide Attempts through a Machine Learning Algorithm Based on Multiple Systemic Psychiatric Scales. *Front. Psychiatry* **2017**, *8*, 192. [CrossRef] [PubMed]
46. Barros, J.; Morales, S.; Echávarri, O.; García, A.; Ortega, J.; Asahi, T.; Moya, C.; Fischman, R.; Maino, M.P.; Núñez, C. Suicide detection in Chile: Proposing a predictive model for suicide risk in a clinical sample of patients with mood disorders. *Braz. J. Psychiatry* **2017**, *39*, 1–11. [CrossRef] [PubMed]

47. Mikolas, P.; Vahid, A.; Bernardoni, F.; Süß, M.; Martini, J.; Beste, C.; Bluschke, A. Training a machine learning classifier to identify ADHD based on real-world clinical data from medical records. *Sci. Rep.* **2022**, *12*, 12934. [CrossRef] [PubMed]
48. Hettige, N.C.; Nguyen, T.B.; Yuan, C.; Rajakulendran, T.; Baddour, J.; Bhagwat, N.; Bani-Fatemi, A.; Voineskos, A.N.; Mallar Chakravarty, M.; De Luca, V. Classification of suicide attempters in schizophrenia using sociocultural and clinical features: A machine learning approach. *Gen. Hosp. Psychiatry* **2017**, *47*, 20–28. [CrossRef] [PubMed]
49. Artieda-Urrutia, P.; Delgado-Gómez, D.; Ruiz-Hernández, D.; García-Vega, J.M.; Berenguer, N.; Oquendo, M.A.; Blasco-Fontecilla, H. Short Personality and Life Event scale for detection of suicide attempters. *Rev. Psiquiatr. Salud Ment. (Engl. Ed.)* **2015**, *8*, 199–206. [CrossRef]
50. Blasco-Fontecilla, H.; Delgado-Gomez, D.; Ruiz-Hernandez, D.; Aguado, D.; Baca-Garcia, E.; Lopez-Castroman, J. Combining scales to assess suicide risk. *J. Psychiatr. Res.* **2012**, *46*, 1272–1277. [CrossRef] [PubMed]
51. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; CRC Press: Boca Raton, FL, USA, 1984; Volume 40, p. 358.
52. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
53. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
54. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000.
55. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
56. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
57. Van Rossum, G.; Drake, F.L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, USA, 2009.
58. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
59. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8024–8035.
60. Cochran, W.G. *Sampling Techniques*, 3rd ed.; Wiley: New York, NY, USA, 1977.
61. Perlich, C.; Provost, F.; Simonoff, J.S. Tree induction vs. logistic regression: A learning-curve analysis. *J. Mach. Learn. Res.* **2003**, *4*, 211–255.
62. Dietterich, T.G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* **1998**, *10*, 1895–1923. [CrossRef] [PubMed]
63. Nadeau, C.; Bengio, Y. Inference for the generalization error. *Mach. Learn.* **2003**, *52*, 239–281. [CrossRef]
64. Luoma, J.B.; Martin, C.E.; Pearson, J.L. Contact with mental health and primary care providers before suicide: A review of the evidence. *Am. J. Psychiatry* **2002**, *159*, 909–916. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.