Journal of Eating Disorders

RESEARCH

Open Access

Predicting anorexia nervosa treatment efficacy: an explainable machine learning approach

Giulia Brizzi[1,2*], Chiara Pupillo[2,3], Elena Sajno[2,3], Margherita Boltri[1], Federico Brusa[4], Federica Scarpina[5,6], Leonardo Mendolicchio[4] and Giuseppe Riva[1,2,7]

Abstract

Introduction Anorexia nervosa (AN) is a psychopathology with an alarmingly high mortality rate. The growing number of individuals seeking help, coupled with the limited resources of clinics, highlights the critical need to identify factors that can predict treatment efficacy. Machine learning (ML) techniques hold great promise in this regard. This data-driven approach offers an unbiased means to uncover predictors of specific outcomes, advancing the understanding and management of this challenging condition.

Objective Six supervised ML algorithms (e.g., Decision Tree and Random Forest) were applied to develop a binary classification model predicting short-term weight recovery/stabilization in AN inpatients and identify the most critical factors influencing this outcome.

Methods Change in Body Mass Index (BMI) from admission to discharge (ΔBMI) was used as the outcome, allowing to classify patients into "improved" (BMI stability or increase) and "aggravation" (BMI decrease). Predictors included clinically relevant psychological tests and physical parameters. Scikit-learn features importance, and SHAP (SHapley Additive exPlanations) analyses were used to investigate predictor importance.

Results The Random Forest model achieved an accuracy of 0.77, an AUC-ROC of 0.72, and a PR curve score of 0.88. Body Uneasiness, Personal Alienation, and Interpersonal Problems subscales emerged as best predictors. SHAP analysis confirmed these results at the individual prediction level.

Discussion Results encouraged interventions focused on body-self experience in addition to interpersonal relationships, including body-swapping experiences and metaverse activities, respectively. This could maximize treatment efficacy, effectively allocating limited resources to achieve clinically relevant outcomes.

Keywords Machine learning, Anorexia nervosa, Eating disorders, Body image, Social relationships

Plain Language Summary

Anorexia nervosa (AN) is a serious eating disorder with one of the highest mortality rates of any mental illness. As the number of people seeking help increases, clinics face challenges in providing effective treatment due to limited resources. This study explored the use of supervised machine learning (ML) to predict weight change in a sample of inpatients with AN. Six supervised ML models were trained to predict changes in body mass index

*Correspondence:
Giulia Brizzi
giulia.brizzi@unicatt.it
Full list of author information is available at the end of the article

(BMI) from admission to discharge, categorising patients into two groups: "improving" (BMI stable or increasing at the end of hospitalisation) and "worsening" (BMI decreasing at the end of hospitalisation). After finding the best performing model, predictor importance extraction was performed to identify which psychological and physical parameters influenced BMI change. The Random Forest model was found to be the best model, with an accuracy of 77%. It highlighted specific psychological factors - such as body image dissatisfaction, feelings of personal disconnection, and relationship difficulties - as the most influential predictors of treatment outcome. These findings suggest that focusing on patients' body image and interpersonal relationships during therapy, including innovative approaches such as virtual reality and metaverse-based activities, may improve treatment success. This research highlights the potential of ML to tailor interventions for people with AN, ensuring that limited resources are used effectively to achieve the best possible outcomes.

Introduction

Anorexia Nervosa (AN) is a severe Eating Disorder (ED) marked by extreme self-imposed food restriction, significantly low body weight, intense fear of gaining weight, and a distorted body image. This distortion reflects an altered perception of the body and a dysfunctional relationship between body and self, where physical appearance heavily influences self-worth and identity [1, 2]. The development of AN arises from a complex interplay of biological, psychological, and sociocultural factors. Genetic variations influence metabolic and appetite-regulating hormones like leptin and dopamine [3]. Dysfunctional thought patterns about body weight, perfectionism, and novelty-seeking personality traits further increase vulnerability [4, 5]. Sociocultural pressures, including thinness ideals and parental attitudes toward food, also play a significant role [6, 7].

The interplay of these factors contributes to AN's chronicity and severity, with studies showing that 19–26% of individuals meet diagnostic criteria even after 20–30 years [8]. This chronic nature makes AN a life-threatening mental health condition with severe physical and psychological implications. Recent research showed that AN affects up to 3% of young women and has the highest mortality rate of any psychiatric disorder, with approximately 5% of patients dying within four years of diagnosis [9, 10]. These epidemiological statistics portray AN as a public health emergency, particularly given that lifetime prevalence ranges from 0.5 to 3.5% in women and 0.1–2.0% in men [11]. The situation has worsened following the COVID-19 pandemic [12], which resulted in a 48% increase in hospital admission rates to specialized ED units compared to previous years [12].

In response to these concerning trends and the multifaceted nature of AN, the current situation presents significant challenges for healthcare systems worldwide. AN remains a challenging disorder to comprehend and treat, given its complex etiology and diverse manifestations, and the rising demand for healthcare providers and therapeutic interventions has placed strain on existing services. This pressing situation has steered the ED research field, particularly in anorexia, towards more efficient identification of outcome predictors and treatment efficacy factors. The goal is to strategically allocate limited resources to maximize clinical impact and help the greatest number of individuals effectively [13, 14].

Longitudinal studies investigated predictors of treatment outcomes in AN. For instance, Fichter et al. [15] identified Body Mass Index (BMI) at admission and Eating Disorder Inventory Maturity Fears subscale at admission as critical factors in determining the successful recovery—intended as the absence of AN diagnosis—in 112 patients. However, these types of studies are difficult to conduct due to high costs, low power, high attrition rates, and testing of specific or limited hypotheses (e.g. a limited set of predictors), leading to difficulties in replication and limited study comparability [10].

In recent years, technological advances have allowed for more sophisticated and effective techniques than clinical longitudinal research, thanks to Machine Learning (ML). ML facilitates the development of predictive models that can be used to test hypotheses, make inferences from data, and employ flexible, data-driven approaches to maximize predictive power [16, 17]. This study specifically focuses on supervised ML methods, where models are trained on labeled datasets containing input–output pairs with known outcomes. This approach involves selecting predictors based on existing literature and clinical expertise, defining an outcome label such as treatment effectiveness, and developing a model able to predict this label based on selected features [16]. In other words, supervised ML models detect patterns and relationships that traditional analysis might not detect. Supervised ML has been applied in emerging use cases in EDs domain, including risk factor identification, monitoring of patients, and predicting treatment response and prognosis in clinical populations (for a review, see [18]).

In the mental health research field, one line of research is applying ML models to predict treatment outcomes and identify critical factors linked to clinically positive

outcomes [19]. For example, ML has been used to predict treatment outcomes for depressive disorder (for a review, see [20]). Chekroud et al. [21] employed supervised ML to predict patients' responses to antidepressant medication by analyzing clinical variables, including baseline symptom severity, treatment history, and sociodemographic information. In the field of cognitive rehabilitation for Mild Cognitive Impairment and dementia, ML techniques have been applied to predict disease progression, conversion from MCI to dementia, and rehabilitation outcomes starting from neuropsychological tests and sociodemographic data [22].

In the EDs research field, current research has primarily concentrated on two areas: (i) identifying biomarkers through neuroimaging (e.g., [23]) and (ii) detecting at-risk individuals (e.g., [24]). For instance, [25] reviewed the use of ML and Natural Language Processing methods to predict AN symptom from social media posts and comments, reporting relatively good performance levels (F1-score ranging from 0.67 to 0.93). Frank et al. [10] developed a ML model to predict BMI at a 6-month follow-up in a sample of individuals with AN treated in a 6-day per week partial hospital program, finding that BMI change from admission to discharge was the most important predictor, strongly correlating with BMI at follow-up ($r=0.55$), while clinical questionnaire scores (e.g., State-Trait Anxiety Inventory, Beck Depression Inventory-II) were less predictive. Sandoval-Araujo et al. [26] applied ML to distinguish between typical and atypical AN using BMI, reinforcing its role as a key diagnostic parameter. Another line of research instead used ML methods to classify individuals with a history of AN at two stages of recovery from healthy controls using brain magnetic resonance images, observing differences in cortical thickness and gray matter volume across several regions (i.e., insula, lateral orbitofrontal, and temporal pole,[24]). Similarly, Lavignino et al. [23] identified six brain regions (i.e., cerebellum white matter, choroid plexus, putamen, nucleus accumbens, the diencephalon, and the third ventricle) to be relevant in distinguishing individuals with AN from healthy controls.

While ML applications in EDs, particularly AN, are emerging, their integration into clinical practice remains limited. Moreover, a major gap persists in studies investigating the key variables that influence treatment outcomes and long-term recovery trajectories. This underscores the need for broader ML applications to support clinical decision-making, treatment development, and personalized intervention strategies in AN care.

The present study aims to address this gap by expanding ML applications focusing on understanding factors that influence treatment outcomes in AN. Specifically, this research employs supervised ML techniques with two primary objectives:

- Create a model to predict the effectiveness of ED treatment in a cohort of hospitalized patients affected by AN,
- Identify the most influential physical and psychological variables for predicting positive treatment outcomes, specifically in terms of weight recovery.

A deeper understanding of the critical factors contributing to effective intervention programs can help target the core aspects influencing treatment outcomes. This allows for a more strategic allocation of limited resources (e.g., financial, personnel, and physical space) to enhance overall treatment outcomes and reach a broader population. Moreover, this approach offers valuable insights into AN mechanisms, stressing key variables that play a pivotal role through a data-driven perspective.

Table 1 Sample descriptives

	Mean (sd)	Minimum	Maximum
Age	24.11 (12.41)	13	66
Duration of Illness (years)	10.14 (12.53)	0	58
Age first diagnosis (years)	16.23 (6.41)	10	50
BMI (admission)	14.13 (1.58)	9.73	16.9
BMI (discharge)	14.49 (1.45)	10.22	17.6
Hospitalization length (days)	35.83 (9.07)	21	59

BMI = Body Mass Index

Methods

Data source

Data for this study was retrieved from the study by Brusa et al. [27]. It contains clinically relevant physical and psychological parameters assessed before starting the rehabilitation program at the IRCCS Istituto Auxologico Piancavallo (Italy).

The dataset contains information related to severe 72 patients (68 females, 4 males) with a diagnosis of AN (54 restrictive, 18 binge-purge subtypes) admitted to a multidisciplinary hospitalization program for EDs between August 2021 and July 2022. Inclusion criteria were: (a) a primary diagnosis of AN, (b) a BMI at admission equal or lower to 17 kg/m^2, indicating moderate to extreme severity, and (c) compliance with the rehabilitation program. Table 1 presents sample descriptives.

Patients were exposed to a multidisciplinary treatment, proposing both individual and group activities (for a more detailed explanation of the treatment see

Table 2 Paired sample t—test

		Student's t	df	*p*	Mean Diff (SE)	Cohen's d
BMI (admission)	BMI (discharge)	− 3.92	71.0	< 0.01	− 0.35 (0.09)	− 0.46

The table presents the paired sample t test results to investigate differences between BMI at admission and discharge. SE = standard error, Diff = difference, df = degree of freedom

Supplementary Materials—Table 8 -and the original study by [27]).

Clinically relevant parameters were assessed at the beginning of hospitalization (T0) and the end of the rehabilitation program (T1).

Because of the severity of patients (BMI mean = 14.13, sd = 1.58) and the short treatment length (Days mean = 35.83, sd = 9.07) we preliminary performed a paired sample t-test to investigate significant differences in BMI before and after the treatment. The analysis showed a significant difference, with the BMI ad admission being significantly higher than the BMI at the discharge (Table 2).

Procedure

Procedure and results reporting were conducted based on Flanagin et al. [28] and Serino et al. [108].

Study design and outcome measure

We developed and evaluated ML models to predict treatment success for AN through binary classification of ΔBMI (i.e., weight improvement). Models were developed and trained to predict two different classes reflecting patients' changes in BMI from admission to discharge (ΔBMI), where class 1 indicated weight recovery or stability (ΔBMI ≥ 0; 53 subjects) and class 0 indicated weight loss (ΔBMI < 0; 19 subjects). The choice of using BMI as an index of treatment effectiveness based on prior research [26, 30, 31]. Notably, our decision to consider BMI maintenance (no change) as a positive outcome is supported by research showing that severe anorexia (BMI < 15) and the restrictive subtype may have no change pre-post treatment in the short term [32]. Indeed, In severe cases of AN, maintaining weight can be a critical first step toward recovery: studies indicate that patients with extreme weight loss often experience significant physiological and psychological challenges that make immediate weight gain difficult. Then, stabilizing weight can prevent further health deterioration and provide a foundation for gradual recovery [33].

Features

The database contains clinical physical and psychological parameters commonly assessed in clinical practice. Below are presented main measures, whereas detailed information is provided in Supplementary Materials (Tables 5, 6).

Psychological variables

- The *Body Uneasiness Test* (BUT; [34]) is a 71-item questionnaire to assess body-related discomfort. It comprises two main subscales: subscale A refers to weight phobia, body image concerns, avoidance, compulsive self-monitoring, detachment, and depersonalization from the body (34 items), whereas subscale B measures worry related to different body parts (37 body areas). Examples of items of scale A include: "I spend a lot of time in front of the mirror" and "I feel I am fatter than others tell me", whereas scale B asks participants to rate how much they hate different body areas such as height, arms, and stomach. This questionnaire has been validated in the Italian language, with a Cronbach's alpha between 0.69 and 0.90 [34].
- The *Eating Disorder Inventory-3* (EDI-3; [35]) is a 91-item questionnaire that assesses three ED-specific pathological symptom categories (drive for thinness, bulimia, body dissatisfaction) and nine general symptom categories (low self-esteem, personal alienation, interpersonal insecurity, interpersonal alienation, interoceptive deficits, emotional dysregulation, perfectionism, asceticism, maturity fears),additionally, 6 composite scales can also be calculated to characterize better the condition (eating concerns composite, ineffectiveness composite, interpersonal problems composite, affective problems composite, overcontrol composite, global psychological maladjustment). Examples of items include "I eat sweets and carbohydrates without feeling nervous" and "I think my stomach is too big". The EDI-III has been validated in Italian language with Cronbach's alpha values for the subscales in ranging from 0.70 to 0.94.

- The *Psychological General Well-Being Index questionnaire* (PGWBI; [105]) is a 22-item questionnaire to assess Quality of Life (HRQoL). It comprehends six subscales: anxiety, depression, positivity and well-being, self-control, general health, and vitality. Examples of items are "How often were you bothered by any illness, bodily disorder, aches or pains during the past month?" and "Did you feel depressed during the past month?". It has been validated in the Italian language with Cronbach's alpha coefficients ranging from 0.92 to 0.94.
- The *Symptom Checklist-90* (SCL-90; [36]) is a 90-item questionnaire to assess nine symptomatologic dimensions: somatization; obsessive–compulsive behaviours and thoughts, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation, and psychoticism. Items are rated on a 5-point Likert scale from 0 (not at all) to 4 (extremely). Examples of items include the rating of how much participants were bothered by "headaches" and "feeling others are to blame for most of your troubles". The questionnaire has been validated in the Italian language, showing Cronbach's alpha coefficients between 0.70 and 0.96.
- *The Frost Multidimensional Perfectionism Scale* (FMPS; [37]) is a 35-item questionnaire assessing five dimensions of perfectionism, namely personal standards, concern over mistakes, parental expectations, doubting of actions, and organization. Examples of items include "I feel that I have made a lot of mistakes in my life" and "If I fail at work/school, I will be a failure as a person". The questionnaire has been validated in the Italian language, showing Cronbach's alpha ranging from 0.86 to 0.89.

Physical variables

- The *Bioelectrical impedance analysis* (BIA; [107]) is a method used to estimate body composition focused on the amounts of body fat and muscle mass. This technique relies on measuring the body's impedance, namely the body's opposition to the flow of an electric current (i.e., resistance and reactance). From this parameter, it is possible to calculate extracellular water, body cell mass, and phase angle.

The final features set consisted of 52 features, including questionnaire subscale scores and physical parameters assessing patients' state at the beginning and the end of the hospitalization. For this study, only parameters at the beginning of the treatment program were considered.

A more in-depth explanation of predictors (features) is available as Supplementary Material.

The model did not include demographic variables based on Frank et al. [10].

Data preparation and feature selection

The dataset was initially examined for quality and completeness. Missing values were imputed using the K-Nearest Neighbors Imputer, which has shown robustness in handling missing data in clinical datasets [38]. All features were scaled to the same values using StandardScaler, which is considered a necessary practice to improve the accuracy of ML model predictions [39]. The Variance Inflation Factor (VIF) was calculated for each predictor variable to assess multicollinearity. No variables exhibited a VIF greater than 10, indicating that multicollinearity was not a significant concern in our dataset [109].

Machine learning pipeline

ML pipeline (see Fig. 1) started with data preprocessing and cleaning with subsequent delta BMI label calculation. A Dummy Classifier (DC) was initially generated as a baseline model to establish a performance benchmark for randomness. In addition, Linear Discriminant Analysis (LDA) was applied to compare classification performance with the other models [40].

Subsequently, six supervised ML algorithms were trained based on ones commonly used in the EDs research field: specifically, Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), Logistic Regression (LR), and k-Nearest Neighbors (kNN) were tested based on Gosh et al. [18].

The dataset was randomly split into a 70/30 ratio, with 70% allocated for training to allow the models to learn underlying patterns and 30% reserved for testing to evaluate their generalization ability on unseen data. This split follows standard ML practices, balancing the need for sufficient training data while ensuring a robust test set for performance assessment [41]. fivefold and tenfold cross-validation with 5 repeats were performed for each algorithm to measure the models' performance. This process allowed the assessment of model stability and generalizability across different data subsets. Following the cross-validation phase, hyperparameter optimization for all six algorithms was performed by using three main techniques: Randomized Search, Bayesian Search, and Grid Search [42]. This phase aimed to refine the parameters of each model to achieve optimal performance. After this, a new cross-validation on the optimized models was performed to assess the performance improvement.

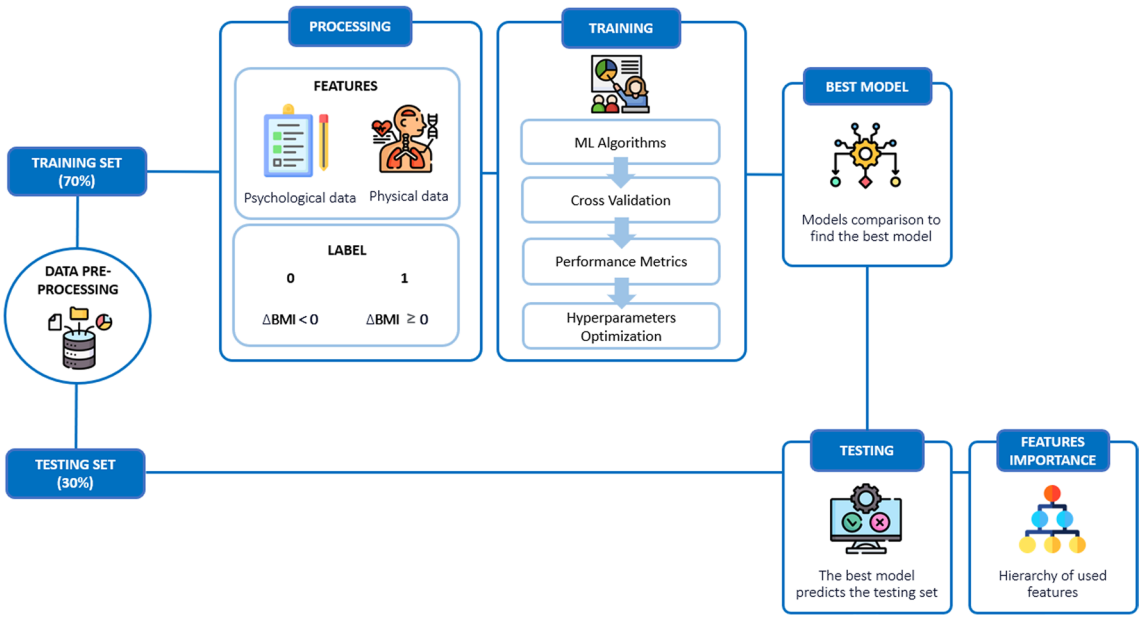

Fig. 1 Machine Learning pipeline. The figure graphically summarizes the machine learning steps

Based on optimization results and cross-validation processes, the two best-performing models were selected for final testing on the held-out test set. Importance features analyses were performed for the best model to identify the most important predictors. Finally, to enhance model interpretability, SHAP (SHapley Additive exPlanations) was utilized to understand the contribution of individual features to the model's predictions.

Model evaluation

The evaluation metrics considered for each phase of the training and testing set were accuracy (i.e., the ratio between the correctly classified samples and the total number of samples), precision (i.e., the ratio between correctly classified samples and all samples assigned to that class), recall (i.e., sensitivity or True Positive Rate, indicating the ratio between correctly classified positive samples and all samples assigned to the positive class), specificity (i.e., the ratio between correctly classified negative samples and all samples classified as negative), F1 score (i.e., harmonic mean of precision and recall), and Area Under the Curve—Receiver Operating Characteristic (AUC-ROC—i.e., ROC curve summary metric that reflects the ability to distinguish between classes, shows the overall diagnostic accuracy). These metrics were selected as they are the most widely used in binary classification in clinical research [43], including in EDs [18, 26]. Confusion matrix was also generated to provide a detailed breakdown of the model predictions. Finally, the area under the curve precision-recall (PR-AUC) was assessed to address potential imbalances between classes and obtain an additional metric of overall performance [44].

Model validation

A repeated stratified K-fold cross-validation was employed to evaluate the model effectively. This splits the data set into multiple folds to train and test the model iteratively, providing a more reliable performance measure than a single training-test split. The K-fold variant splits the data into K subsets, trains the model on K-1 folds, and tests it on the remaining fold, repeating this process K times. Stratification ensures that the proportion of positive and negative samples is consistent across folds, which is critical for unbalanced data sets because it preserves class distributions. It has been seen to be one of the most suitable cross-validation types for small clinical datasets with multi-characteristic subjects, ensuring a balanced trade-off between bias (errors from overly simplistic models that do not fit the data) and variance (errors from excessively complex models that overfit the data) [45, 46]. In addition, it has been seen that smaller K values (10, 5, 3) for small sizes offer a practical compromise, balancing computational efficiency and reliability of performance estimates [45].

Feature importance analysis

Feature importance analysis was performed using the feature_importances_ attribute from scikit-learn to specifically evaluate individual predictors' impact on the

Table 3 Evaluation metrics of the tenfold with 5 repeats cross-validation for the six algorithms

	Decision Tree	Random Forest	Gradient Boosting	Support Vector Machine	Logistic Regression	k-Nearest Neighbors
Accuracy	0.70 (0.18)	0.70 (0.15)	0.71 (0.18)	0.74 (0.09)	0.58 (0.22)	0.67 (0.19)
AUC-ROC	0.61 (0.23)	0.65 (0.27)	0.65 (0.27)	0.62 (0.30)	0.57 (0.31)	0.65 (0.26)
Precision	0.80 (0.19)	0.75 (0.14)	0.81 (0.18)	0.74 (0.09)	0.71 (0.21)	0.74 (0.16)
Recall	0.79 (0.23)	0.91 (0.15)	0.82 (0.20)	1.00 (0.00)	0.71 (0.26)	0.84 (0.20)
Specificity	0.43 (0.45)	0.14 (0.33)	0.44 (0.47)	0.00 (0.00)	0.24 (0.40)	0.18 (0.36)
F1-Score	0.77 (0.18)	0.81 (0.12)	0.79 (0.15)	0.85 (0.06)	0.69 (0.20)	0.78 (0.16)

The values are expressed as mean values and their standard deviation in brackets. AUC-ROC = Area Under the Curve—Receiver Operating Characteristic

best performing model and extract the 10 most influential features. These importance scores were computed as each tree's mean and standard deviation of the impurity decreased. Specifically, higher relative importance scores reflected features that are most relied on by the model to make predictions.

In addition, the SHAP (SHapley Additive exPlanations; [47]), Explainable AI (XAI) method was used to make the results more interpretable. SHAP was used because it assigns an importance value to each feature in the model output and is designed to be applied a posteriori to any type of ML model (Lundberg and Lee, [48]),the 20 most relevant parameters are extracted by default (Lundberg & Lee, [48]). In this context, Tree. Explainer algorithm was used since it is suitable for tree models such as Random Forest. SHAP values also provide a deeper understanding of the contributions of characteristics to individual predictions. SHAP summary plots were created for the overall model and separately for the two AN subtypes (0 = restrictive, 1 = binge-purge). This subtype-specific analysis was conducted solely as an additional post-hoc interpretability using SHAP values. No additional ML training was performed for the subtypes because the limited sample size precluded the development of separate ML models for each subtype.

Each patient was represented by a single point for each feature in the graph. The x-axis coordinate of each point was determined by the SHAP value, and the points were stacked along each feature to show their density. The features were sorted by the average absolute value of the SHAP values for each feature. Color was used to indicate the original value of a feature; red and blue indicate the high or low of the individual feature, respectively. The gray vertical line of the decision graph represents the baseline value of the model.

To compare and integrate the results of these two methods, it is essential to note that SHAP values provide a measure of how much each feature contributes to the model's output for a particular prediction, indicating both positive and negative impacts. In contrast, feature importance rankings generally focus on the overall importance of each feature in the model, providing only those that impacted positively. Consequently, discrepancies in the order of importance between the two methods may arise.

Table 4 Evaluation metrics of the tenfold with 5 repeats cross-validation for the two optimized models: Decision Tree and Random Forest

	Decision Tree	Random Forest
Accuracy	0.72 (0.20)	0.76 (0.15)
AUC-ROC	0.68 (0.23)	0.67 (0.23)
Precision	0.85 (0.17)	0.88 (0.12)
Recall	0.75 (0.26)	0.81 (0.23)
Specificity	0.60 (0.44)	0.60 (0.44)
F1-Score	0.77 (0.21)	0.81 (0.14)

The values are expressed as mean values and their standard deviation in brackets. AUC-ROC = Area Under the Curve—Receiver Operating Characteristic

ML algorithms training and testing were carried out in Python v. 3.10.12, using Google Colab v. 0.0.1a2. We used the following Python packages: *numpy, pandas, matplotlib, scikit-learn, seaborn, scipy.stats, scikit-optimize*, and *shap*.

Results

ML performance evaluation during training and testing to predict delta BMI

The six trained ML algorithms ten folds cross-validation performance is reported in Table 3. Notably, all models outperformed our benchmark for randomness, namely DC model (AUC-ROC = 0.50) and LDA (accuracy = 0.31; [49]). DC and LDA classification reports are reported as Supplementary Material.

The best hyperparameters were found with Bayesian Optimization, which obtained similar values to the other two methods and is the most robust technique. As shown in Table 4, the best-performing algorithms trained with the fitted hyperparameters were Decision Tree (DT) and Random Forewst (RF).

Table 5 Evaluation metrics of the testing set for the optimized Random Forest (RF) model

Random forest—test		
	Class 0 (negative Δ BMI)	**Class 1 (0 or positive Δ BMI)**
Accuracy	0.77	
AUC-ROC	0.72	
Precision	0.67	0.79
Recall	0.33	0.94
Specificity	0.33	0.33
F1-Score	0.76	0.86

The table shows the results for both binary classes. AUC-ROC = Area Under the Curve—Receiver Operating Characteristic

The testing phase was conducted exclusively on the two best-performing models, DT and RF, which were identified based on their superior performance during training. To assess their generalization ability, we evaluated these models on a previously unseen 30% hold-out test set from the original dataset. This test set was not used during training, ensuring that the evaluation reflects how well the models can classify new data in our binary classification. The results of the final best model—RF—are presented in Table 5: its results show an accuracy of 0.77, highlighting the model's ability to correctly classify most of the cases in the test set.

To evaluate the RF performance, its ability to predict class 1 and 0 were tested. The AUC-ROC of 0.72 indicated a fair discrimination without overfitting and outperforming the training curve (AUC-ROC = 0.67; Fig. 2a). The curves' shapes indicate that the model performs consistently better than random chance across various classification thresholds. Figure 2b presents the Precision-Recall (PR) curves for training and testing. The PR-AUC values for testing (0.88) and training (0.84) are notably high, indicating robust performance in balancing precision and recall.

The model correctly identified expected positives (precision = 0.79) and true positives (recall = 0.94). However, it showed difficulties in identifying true negatives (specificity = 0.33). Similarly, the confusion matrix (Fig. 3) showed that the model accurately classified 15 out of 16 positive cases (True Positives), but struggled with identifying negative outcomes (Class 0), misclassifying 4 out of 6 as positives (False Positives). Despite the 4 false positives, the high precision (0.79) is explained by the large number of true positives (15) relative to false positives (4). The high recall and precision for Class 1 suggest that the model is highly reliable in predicting positive BMI changes, which are critical for assessing treatment success. On the other hand, the high false positive rate and low specificity is given by the imbalance of the two classes, as the negative class is lower than the positive class [50].

Predictors importance

Initially, the importance of the features of the optimized RF model was evaluated using feature importance analysis, which showed the 10 most important model predictors for the training set. BUT-PST, EDI3-PA, and

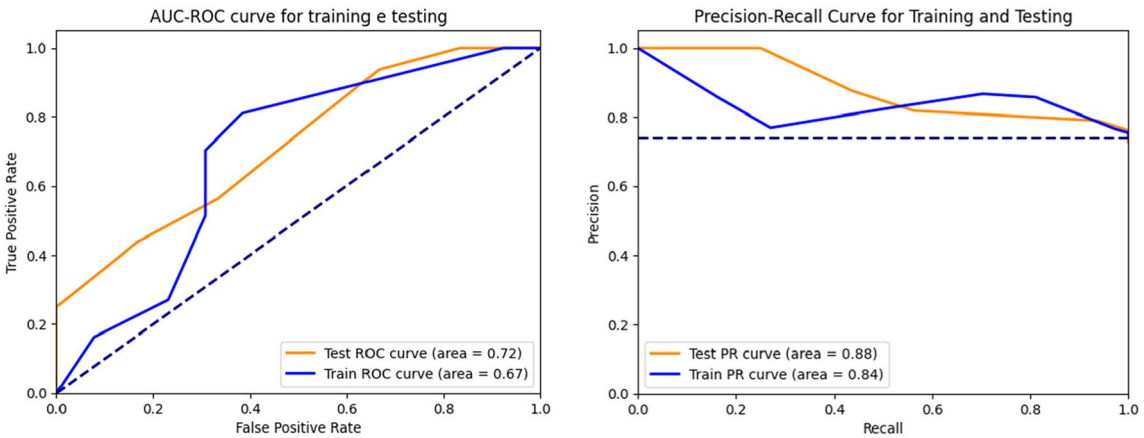

Fig. 2 Visual comparison of a performance model for training and testing set for Random Forest. The blue lines represent the training curves, the orange lines represent the testing curve, and the dashed lines of a random classifier. A higher curve represents a better performance of the model, to be acceptable it must exceed the dotter lines. The first figure represents the AUC-ROC curve (**a**). The AUC-ROC curve shows the relationship between the true positive rate and the false positive rate. The second figure represents the Precision-Recall (PR) curve (**b**). The PR curve shows the relationship between precision and recall. AUC-ROC = Area Under the Curve—Receiver Operating Characteristic

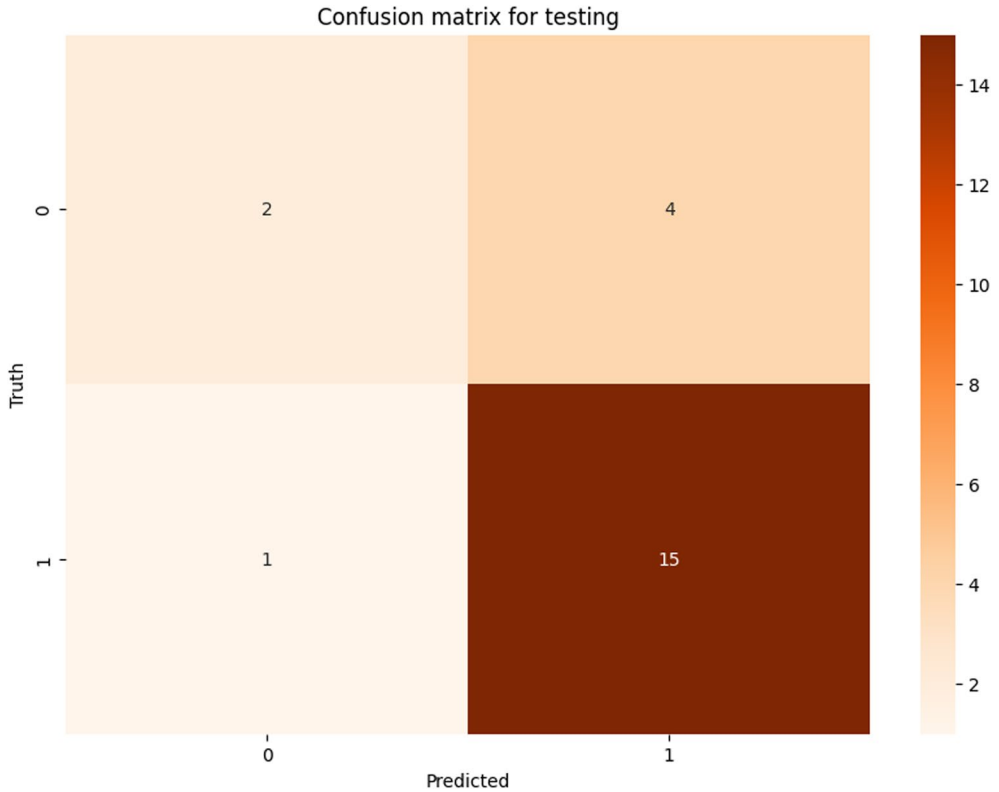

Fig. 3 Confusion matrix of testing set. The picture shows the four basic characteristics for evaluating a classification model. In this case it concerns the confusion matrix of the best model (Random Forest) for testing set. 1 indicated weight recovery or stability (ΔBMI ≥ 0) and 0 indicated weight loss (ΔBMI < 0). In the first row, the first cell indicates the true negatives (number of patients correctly classified that their delta BMI has worsened) and the adjacent cell the false positives (number of patients misclassified with improved or unchanged delta BMI, known as type I error), while the second row shows the false negatives (number of patients who were misclassified with improved or unchanged delta BMI but actually had negative delta BMI, known as type II error) and true positives (number of patients who were correctly classified as patients to whom delta BMI was positive or unchanged), respectively

EDI3-IPC resulted as the top three most influential parameters (Fig. 4).

To further understand the impact of features on individual predictions, the SHAP technique was applied. The overall SHAP summary plot (Fig. 5) highlighted BMI (admission) as the most impactful feature, in addition to EDI3-EDRC and EDI3-GPMC.

Comparing the Scikit-learn-based feature importance rank and the SHAP outputs, EDI3-PA, EDI3-IPC, FMPS-Concern Over Mistakes remained in the top 10 predictors, while BUT-PST and EDI3-A were still present but in lower positions.

A more detailed posteriori analysis was performed on the two subtypes of AN, i.e. restrictive (R) (Fig. 6a) and binge-purge (BP) (Fig. 6b). In both conditions, BMI at admission emerged as the most relevant predictor, as observed in the overall SHAP model. Specifically, low-to-medium initial BMI predicted an equal to zero or positive ΔBMI. No differences emerged when comparing the R and the overall SHAP model, whereas differences emerged when considering the BP model. Here, features showed different feature rankings despite variables being the same.

Discussion

The primary aim of this study was to develop a ML model to predict weight recovery or stabilization in a cohort of AN inpatients. Results from this study showed ML good performance levels (accuracy = 0.77, AUC-ROC = 0.72, and PR curve = 0.88), similar to previous studies (AUC-ROC values between 0.49 and 0.93, and accuracy between 0.59 and 0.86; [18, 51], Haynos et al., [52]). Despite differences in outcome measures, predictor variables, and performance metrics, the results of this study remain within the performance range reported in

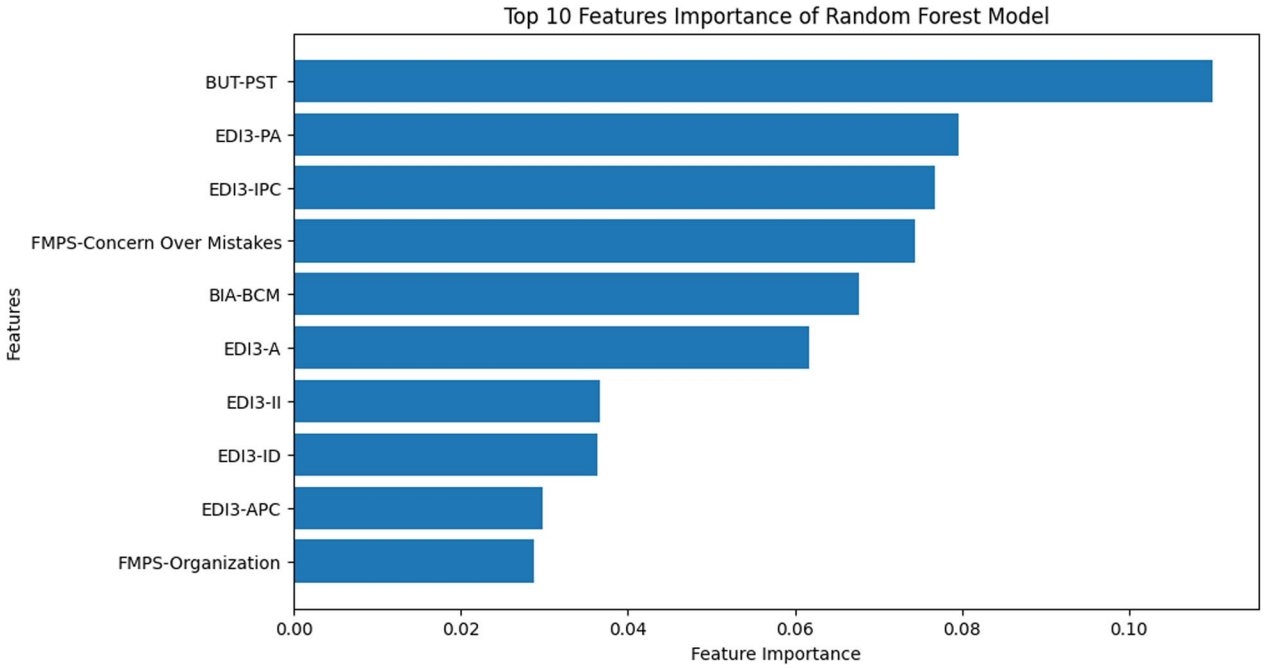

Fig. 4 Scikit-learn-based feature importance rank for predicting class 1 (0 or positive delta BMI). The features importance plot shows the 10 most important predictors of weight recovery and/or stability. BUT = Body Uneasiness Test, EDI3 = Eating Disorder Inventory 3, FMPS = Frost Multidimensional Perfectionism Scale, BIA = Bioelectrical impedance analysis. BUT-PST = Positive Symptom Total, EDI3-PA = Personal Alienation, EDI3-IPC = Interpersonal Problems Composite, BIA-BCM = Body Cell Mass, EDI3-A = Asceticism, EDI3-II = Interpersonal Insecurity, EDI3-ID = Interoceptive Deficits, EDI3-APC = Affective Problem Composite

the literature, highlighting the potential clinical utility of this model in the context of eating disorders.

The secondary aim of this study was to identify which variables better predicted weight recovery, as a proxy of successful treatment. Scikit-learn feature importance extraction [53] and SHAP values [54] were used to identify predictors at the global and individual levels respectively. Interestingly, results were consistent: overall predictors were also reflected at the individual level, thus reinforcing the robustness of results [55]. The subsequent section goes into more detail concerning the three most important weight stability and recovery predictors. To conclude, we critically discussed the implications of integrating ML into clinical practice.

This study represents a preliminary, proof-of-concept effort designed to highlight the potential of ML methods in the field of EDs research. The following sections will explore the model's most significant predictors to conclude with a critical evaluation of the advantages and challenges associated with applying ML in clinical settings.

Predicting weight recovery or stability

Body-self relationship

Scikit-learn features importance and SHAP analyses identified BUT score as a significant predictor of weight stability and recovery. In particular, the Positive Symptom Total (BUT-PST), Body Image Concerns (BUT-BIC), the overall Positive Symptom Distress Index (BUT-PSDI), and body-related concerns and dissatisfaction with own body image (BUT-A) emerged as critical factors. This aligns with early Bruch's [56] insights, according to which AN interventions should carefully consider how patients perceive and relate to their bodies to be effective in the long term.

A large amount of evidence showed that AN is fundamentally characterized by a distorted experience of one's body. This Body Image Disturbance goes beyond mere visual (mis)perception, representing a profound disconnection between the actual physical state and the subjective experience of their body at perceptual, emotional and cognitive levels [57, 58]. Specifically, research has shown that patients with AN report high levels of body dissatisfaction and body-related distress, which are reinforced by the internalization of unhealthy beauty standards [58]. Concerns still remain regarding perceptual alterations: while some studies reported individuals with AN significantly overestimate their body dimensions compared to healthy controls [59], others have found no such perceptual distortion [60]. Such inconsistency may partly stem from methodological differences across studies—e.g., different assessment tools ranging from explicit

Fig. 5 Summary plot of SHAP-calculation for the 20 highest-ranking features. Features are sorted by their mean absolute SHAP value in descending order with the most important variables at the top. Each dot corresponds to one patient in this study. This plot shows how the different variables of each patient affect the prediction of the RF model towards class 1 (0 or positive delta BMI). Positive SHAP values (to the right) indicate features that increase the likelihood of a positive BMI change, while negative SHAP values (to the left) indicate features that decrease this likelihood. Instead, the gradient of colors indicates positive (red) or negative (blue) values compared to the original feature values for each patient. The plot is based on the RF model with all features included for all 72 patients. BUT = Body Uneasiness Test, EDI3 = Eating Disorder Inventory 3, FMPS = Frost Multidimensional Perfectionism Scale, BIA = Bioelectrical impedance analysis, SCL90 = Symptom Checklist-90, PGWBI = Psychological General Well-Being Index questionnaire, BMI = Body Mass Index, EDI3-EDRC = Eating Disorder Risk Composite, EDI3-GPMC = General Psychological Disadaptation Composite, EDI3-IPC = Interpersonal Problems Composite, BIA-PA = Phase Angle, EDI3-MF = Maturity Fear, EDI3-PA = Personal Alienation, EDI3-OC = Obsession Composite, BUT-A = Avoidance Behavior, EDI3-IA = Interpersonal Alienation, EDI3-II = Interpersonal Insecurity, EDI3-A = Asceticism, EDI3-LSE = Low Self-Esteem, BUT-PST = Positive Symptom Total, BUT-BIC = Body Image Concerns, EDI3-DT = Drive for Thinness

questionnaires to implicit measures—as well as limited carefulness in capturing bodily experience complexity.

In fact, how we experience our body depends on both emotions and beliefs, but also on how we process and integrate information from inside (interoceptive) and outside our body (exteroceptive). Recent studies revealed deficits in patients with AN both in processing inner body information (Lucherini et al., [118]) and at the level of integration of exteroceptive and interoceptive information [4]. This favors the idea of perceptual alterations

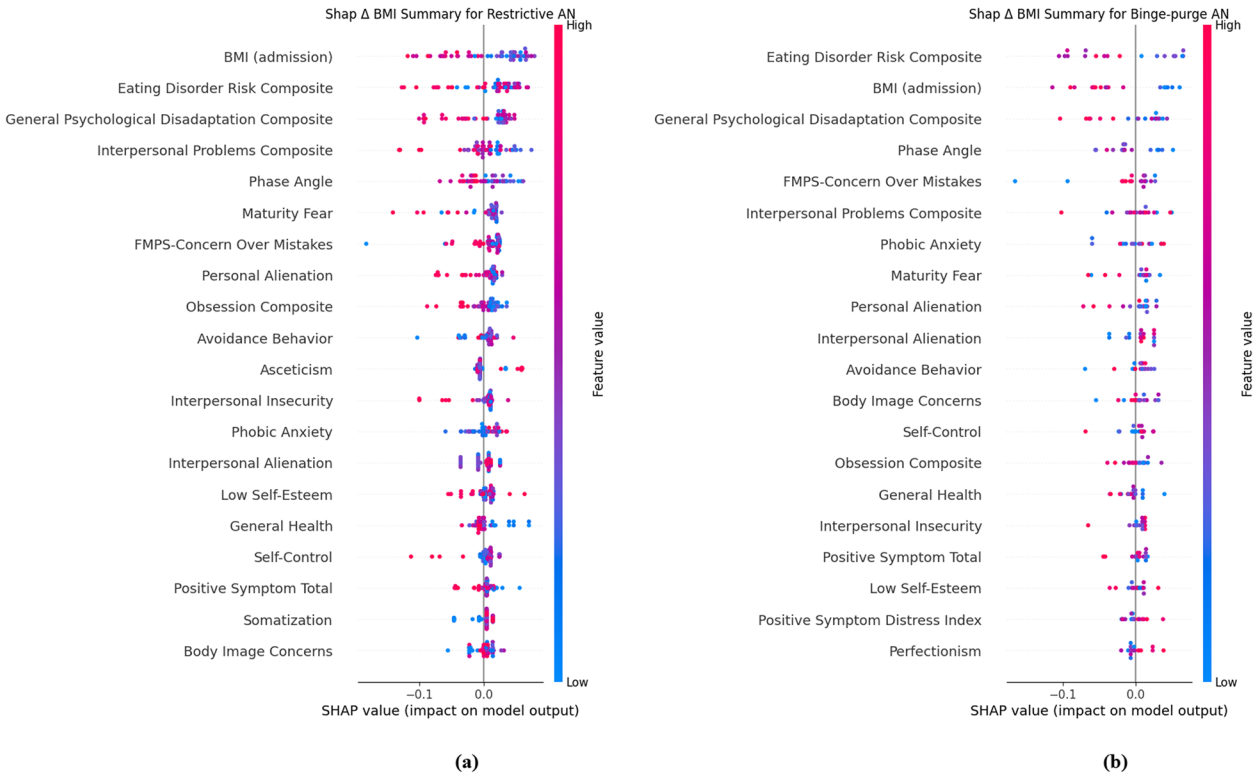

Fig. 6 **a**, **b** Summary plot of SHAP-calculation for the 20 highest ranking features for subtypes AN. Plot shown for restrictive AN (**a**) and binge-punge AN (**b**). Features are sorted by their mean absolute SHAP value in descending order with the most important variables at the top. Each dot corresponds to one patient in this study. Positive SHAP values (to the right) indicate features that increase the likelihood of a positive BMI change, while negative SHAP values (to the left) indicate features that decrease this likelihood. Instead, the gradient of colors indicates positive (red) or negative (blue) values compared to the original feature values for each patient. The plot is based on the RF model with all features included for all 72 patients. Body Uneasiness Test = BUT, Eating Disorder Inventory 3 = EDI3, Frost Multidimensional Perfectionism Scale = FMPS, Bioelectrical impedance analysis = BIA, Symptom Checklist-90 = SCL90, Psychological General Well-Being Index questionnaire = PGWBI, Body Mass Index = BMI, Eating Disorder Risk Composite = EDI3-EDRC, General Psychological Disadaptation Composite = EDI3-GPMC, Interpersonal Problems Composite = EDI3-IPC, Phase Angle = BIA-PA, Maturity Fear = EDI3-MF, Personal Alienation = EDI3-PA, Obsession Composite = EDI3-OC, Avoidance Behavior = BUT-A, Asceticism = EDI3-A, Interpersonal Insecurity = EDI3-II, Interpersonal Alienation = EDI3-IA, Low Self-Esteem = EDI3-LSE, Positive Symptom Total = BUT-PST, Body Image Concerns = BUT-BIC, Positive Symptom Distress Index = BUT-PSDI, Perfectionism = EDI3-P

that may reinforce body image disturbance, creating a self-reinforcing cycle of distorted body experience.

Further investigation of these perceptual mechanisms could reveal novel patterns and inform therapeutic approaches. For instance, impaired interoceptive awareness might create a cascade effect where disconnection from internal bodily signals leads to distorted cognitive interpretations—such as feeling bloated despite no physiological basis—ultimately contributing to body dissatisfaction.

Despite these insights into the multifaceted nature of body experience, standard intervention protocols in hospital settings tend to neglect the role of the body in AN, prioritizing the work on food-related behaviors and weight recovery through dietary management [58]. While this remains essential, evidence seems to suggest this focus alone may be insufficient for comprehensive recovery. From this perspective, Then, it is not surprising that, to date, there are no superior treatment options for AN. Established interventions—including Cognitive Behavioral Therapy (CBT) and Family-based therapy—show similar outcomes, with limited success in achieving clinically significant improvements [61]. Some authors have suggested that this therapeutic ceiling may stem from insufficient attention to body experience in conventional treatments [4, 58]. In response, an emerging line of research is exploring the integration of body image-centered protocols into AN treatment. These innovative approaches encompass multiple strategies: psycho-educational sessions (e.g., reasoning about body-checking behaviors, or social media use; [62, 63]), meditative practices (e.g., yoga, meditation,[64, 65]) or self-monitoring

journaling exercises following Cognitive Behavioral Therapy Enhanced approach (CBT-E,[66]). Initial findings suggest that addressing altered body-self relationships positively influences recovery outcomes. Notable initiatives such as the Body Project [67] and The Body Wise [119] have further advanced this approach by developing experimental body-focused protocols specifically for AN treatment.

Recently, technology-based treatments have been also proposed, including Mirror Exposure [106] and Body Swapping (Serino et al., [29]) in Immersive Virtual Reality (IVR) environments. Such approaches use multisensory technology to place patients in bodies different in shape and weight from real ones, with the primary goal of proposing behavioral experiments that would not be possible in physical reality. For instance, in IVR Mirror Exposure paradigms, patients embody a normal-weight body and look at themselves in a virtual mirror from a third-person perspective [106]. This experience was observed in the fear of gaining weight [106]. Similarly, during Body Swapping procedure patients are asked to embody a normal weight body (BMI = 18.5), but this time from a first-person perspective [110]. Interestingly, this experience has been observed to reduce body distortion, as the swap into a virtual body helped patients to re-establish a connection with their physical bodies (Serino et al., [29]). Recently, our research group proposed two different protocols in this direction: the study by Malighetti al et al. [68] used a combination of IVR body swapping and mirror exposure into different bodies while asking patients to recall positive, negative, and neutral autobiographical memories depending on the body size (overweight, underweight, and normal-weight respectively, whereas the study by Brizzi and colleagues [120] (pre-print asked patients to recall what each part of her bodies was able to do for them in terms of enjoying full activities (i.e., functional mirror exposure. Following this trend, Regenerative Virtual Therapy (RVT; Riva et al., [121] has been proposed. It is an innovative therapeutic approach that combines principles from neuroscience, psychology, and advanced technology to restructure faulty bodily self-representations. What makes this proposal particularly interesting is the combination of multisensory technologies to reshape both the inner and outer body self-perception, brain-stimulation techniques and mindfulness practices.

Drawing from Cognitive Behavioral Therapy (CBT) principles, Immersive Virtual Reality (IVR) approaches recognize the pivotal role of behavior in influencing emotional states and thought patterns. Just as psychological disorders often involve maladaptive behaviors that reinforce dysfunctional emotions and cognitions, IVR provides a unique platform for behavioral modification. By enabling controlled behavioral interventions in virtual environments, IVR facilitates bottom-up therapeutic change, where behavioral alterations can catalyze cognitive restructuring [69].

The findings from this study underscore the critical importance of integrating body-focused interventions into AN treatment protocols to optimize weight recovery and maintenance outcomes. Innovative technologies like IVR, when combined with traditional therapeutic approaches, offer unique opportunities to address the complex interplay between emotional, cognitive, and perceptual components of body experience.

Social and interpersonal factors

The feature importance analysis revealed two other two key predictors in addition to body-self relationship: the EDI3-PA (Personal Alienation) and EDI3-IPC (Interpersonal Problems). This aligns with previous studies showing that interpersonal difficulties and relationship style (e.g., Non-assertive and Friendly-submissive interpersonal style) influence AN symptomatology onset and maintenance [70–72]. Notably, such a result was observed both when considering qualitative data [73] and quantitative [74, 75].

The importance of interpersonal relationships for mental health is well stressed by Sullivan's interpersonal theory [76], according to which social interactions are fundamental for positive human functioning and dysfunctional relationships significantly increase psychopathology risk [77]. Moreover, they are the focus of AN cognitive interpersonal maintenance model proposed by Schmidt and Treasure [78], which stresses the role of relationships in pathology onset and maintenance. According to this theoretical framework, anorexic symptoms are maintained intrapersonally by beliefs about the positive function of the illness, and interpersonally by both positive and negative reactions elicited from close others by the physical presentation and behaviors associated with AN (Schmidt & Treasure, 2010,2013. Then, symptoms such as food avoidance might be understood as a strategy to address the need to avoid close relationships that might evoke intense negative emotions. This may be even more the case in Western cultures, where food and eating are inherently social activities: shared meals are often at the heart of social interactions, celebrations, and family gatherings, making eating more than just food and nutrition, but a vital part of social bonding and community.

Following this perspective, Lacan [79] proposed that AN should be understood as a manifestation of a profound struggle with desire and identity, intricately linked to the subject's relationship with the Other (i.e., namely the symbolic representation of societal norms, familial

expectations, and interpersonal relationships,[80]). That is, AN might be not merely a struggle with food, but it is a response to Other's demands and desires: food refusal can be then conceptualized as a form of resistance against the symbolic order and the expectations it entails [81]. Importantly, this might also, in some way, reflect their resistance to treatment and lack of trust in care personelle and programs [82].

In light of all these considerations, Cognitive Behavioral Therapy (CBT), the Maudsley Model of Anorexia Nervosa Treatment for Adults (MANTRA; [83, 84]) and Multi-family (MFT-AN,[85]) therapeutic approaches include activities specifically targeted to address social and interpersonal difficulties [111]. Using strategies such as role-playing and behavioral experiments on one hand, and identity exploration and group activities on the other hand, these approaches try to work on the social component of AN. The idea is to allow patients to explore alternative identities, practice and train social skills, and explore alternative behaviors and beliefs. In hospital settings, social activities include psychoeducation groups and peer-discussions. For example, the individuals enrolled in the study by Brusa et al. [27] participated in various rehabilitative activities with intrinsic relational and socializing value (i.e. group discussion, team-work, brainstorming ideas on specific topics, etc.).

The MEDverse proposal [86] represents a conceptual framework that merges VR and augmented reality (AR) technologies within the metaverse to create immersive therapeutic environments tailored for mental health treatment. For example, it is possible to envisage social activities in the metaverse. Such activities might range from role-playing to engaging in conversations about food and body image with virtual characters, behavioral experimentals while practicing eating in virtual social settings, to group therapy sessions with avatars representing other patients and therapists. The technology's principal advantage lies in its capacity to generate a profound sense of presence and embodiment [112], enabling patients to emotionally and cognitively engage with therapeutic content in a way that transcends traditional interventions.

This heightened sense of immersion manifests through two crucial psychological mechanisms: first, the "place illusion", where patients genuinely feel present in the virtual environment, and second, the "plausibility illusion", where they process and respond to virtual events as if they were real [113]. These mechanisms, combined with the capability to induce body ownership illusions through multisensory integration, create a powerful therapeutic platform where patients can safely explore and challenge themselves.

Through carefully designed virtual social scenarios, patients can practice interpersonal interactions in a controlled environment that feels authentically engaging yet remains therapeutically safe. The technology facilitates graduated exposure to challenging social situations, from intimate family meals to broader social gatherings, allowing individuals to develop and refine their social skills while managing anxiety and eating-related behaviors simultaneously. The immersive nature of these experiences, combined with real-time feedback and professional guidance, creates a powerful platform for addressing social isolation, improving communication patterns, and rebuilding damaged relationships—elements that are often central to the maintenance of eating disorders but challenging to address in traditional therapeutic settings [87]. Moreover, the virtual environment can be populated with avatars representing various social roles (family members, peers, colleagues), enabling patients to work through specific interpersonal challenges while developing more adaptive social interaction. The underlying idea is that "exposure to one's problems and pain as they are experienced in the lives of the others facing you can bring about change of an unforeseen kind with far-reaching consequences" [88, 114].

The findings from this study underscore the critical importance of integrating social relationship work into AN treatment to optimize weight recovery and maintenance outcomes. Even in this context, technologies like IVR offer unique opportunities to address interpersonal challenges through direct experiential learning. The immersive nature of IVR enables patients to engage in social scenarios in real time, bypassing the limitations of traditional therapeutic approaches that rely heavily on mentalization and recall abilities. This is particularly valuable as individuals with AN often struggle with abstract cognitive processing and may benefit more from immediate, embodied experiences.

Machine learning in clinical practice: pros and cons

In the last few years, ML has emerged as a powerful analysis approach for predicting treatment outcomes and understanding complex patterns in patient data. ML indeed offers several advantages over traditional statistical approaches. First, the latter is primarily used to confirm or refute specific hypotheses and may be more susceptible to overfitting issues when dealing with complex and multidimensional datasets [89], while supervised ML focuses on prediction and provide robust internal validation of model performance through techniques such as k-fold cross-validation, which helps mitigate overfitting concerns and improves generalizability

[90]. Second, ML offers greater flexibility in handling diverse data types and model structures.

Supervised ML can be fine-tuned through various parameters and hyperparameters to optimize performance for specific feature types and research questions [91]. Furthermore, ML excels at processing high-dimensional data, allowing it to consider numerous clinical and nonclinical variables simultaneously [53], and it can identify complex, nonlinear relationships among variables without requiring predetermined assumptions [92].

A critical challenge in ML applications for clinical decision-making is the risk of misclassification. Misclassifying patients as likely responders when they are not (false positives) could result in the continuation of ineffective treatments, delaying more suitable interventions. Conversely, incorrectly predicting non-response (false negatives) could lead to the denial of beneficial treatments to patients, impacting their recovery trajectory. These risks underscore the need for rigorous model evaluation and can be mitigated through careful model selection, robust validation techniques, and post hoc explanation methods [93]. It is imperative to acknowledge that ML models should be regarded as decision support systems, wherein the ultimate decision rests with healthcare professionals to ensure patient-centered care.

In conclusion, the adaptability and flexibility of ML allow researchers to tailor their analyses to the unique characteristics of patient data, potentially uncovering new predictors, and to be able to use multidimensional data considering the multifaceted nature of the disorder. This is particularly useful in the AN context, where, as previously seen, multiple factors, physiological markers to psychological traits, and environmental influences can contribute to treatment outcomes. Finally, this data-driven approach has the potential to reveal unexpected connections and patterns in AN data, potentially leading to new insights into the etiology and treatment. In this regard, future studies could also consider non-clinical parameters (e.g., questionnaires on social aspects) and anamnestic factors (e.g., family history of mental health difficulties) to better understand their impact and possible links between them.

However, supervised ML also presents some important limitations. Many of them are like traditional statistics. Both methods can be affected by data quality issues, such as selection bias or measurement error, which can lead to misleading results or poor generalizability. This means that data might be not generalizable across different populations, such as individuals who seek help and those who do not [16]. Additionally, while ML often requires large datasets for optimal performance [94], traditional statistics can also benefit from larger sample sizes to increase statistical power and precision [95]. Lastly, a specific challenge for many ML models in some application areas, such as healthcare, is their "black box" nature [47]. Unlike traditional statistical methods, which often provide interpretable coefficients or odds ratios, complex ML algorithms can produce highly accurate predictions without offering clear explanations for their decision-making processes [96]. This lack of transparency can be problematic in clinical settings, where understanding the rationale behind predictions is crucial for trust and implementation. To address this limitation, the field of explainable AI (XAI) has emerged [97]. XAI aims to make the decision-making processes of ML models more understandable [98]. XAI techniques can provide explanations of model behavior and creation (global level) or individual predictions (individual level, e.g., individual patient), which can be implemented as early as during model creation or used later as posthumous explanations [47]. These insights are particularly important in clinical applications, where interpretability can have a significant impact on the acceptance and effective use of ML models by physicians and patients [99].

Limitations

This study is situated within the emerging field of research leveraging ML to evaluate treatment effectiveness in mental health (e.g., [100]). In recent years, ML techniques, particularly supervised learning models, have been increasingly used to enhance clinical decision-making and optimize treatment selection in psychiatric disorders, including depression, schizophrenia, and bipolar disorder [115, 116]. However, despite the growing interest in data-driven approaches, it is essential to recognize certain limitations that may impact the interpretation and generalizability of present findings.

First, the limited sample size and imbalance between the two groups being analyzed. However, our sample size aligns both with previous studies in the field [101] and with the minimal number of participants suggested by Figueroa et al. [117] for reliable ML evaluation. Additionally, several methodological strategies were implemented to address such possible limits (e,g., repeated stratified k-fold cross-validation, multicollinearity check, and precision-recall (PR) curve). Future research should validate the model on a larger and more heterogeneous sample to confirm its clinical utility. In this regard, we recommend researchers from different fields collaborate and combine large international datasets to fully exploit ML potential [102]. While we did not preregister this study due to its exploratory nature, adopting open science practices, such as data sharing and preregistration through platforms like the Open Science Framework, can enhance transparency, and reproducibility, as well as the collective effort to advance the field.

Second, we focused on change in BMI as an index of improvement based on Franket et al. [10]. Even though BMI is considered the primary target of AN hospitalization programs, other parameters might be considered as an index of successful treatment (e.g., psychological well-being). Additionally, we conceptualized weight stabilization ($\Delta BMI = 0$) at the same level as an improvement since, in severe conditions and short interventions, the most important aspect is that patients do not lose additional weight and start to stabilize their weight and then improve. This is based on Couturier and Lock [103] who reported that the mean time to AN remission for weight is 11.3 months.

Third, we did not have a follow-up measure as it is not foreseen in the rehabilitation program to have this information. However, the timeframe for assessing treatment success presents a challenge, with some research indicating that the clinical outcome of treatment for AN should be evaluated more than two years post-treatment completion [104]. Although this can be seen as a limitation, it reflects the reality of clinical settings: everyone is unique, just as the expression of their symptoms is, making it impossible to predict the rehabilitation path a person will undertake after discharge. This makes it difficult to maintain a connection with the person for potential follow-ups. Indeed, there are various contexts in which a person might find themselves after discharge, such as communities, hospitals, day hospitals, or returning home. This would make it challenging both to collect and to generalize follow-up data after discharge.

Additionally, in line with the research question, in this study, we focused on psychological and physical features. However, there might be other factors that might influence the treatment outcome, such as the presence of comorbidities, pharmacological treatment, and treatment duration, among others. Future research might consider additional parameters with larger datasets.

Lastly, the present study identifies predictors of treatment efficacy, but it does not explore how these predictors relate to specific therapeutic techniques or treatment outcomes. Thus, we encourage future research to explore ML use to match individuals with the most appropriate interventions based on their unique characteristics.

Conclusion

The ML potential in psychiatry has generated a great deal of enthusiasm in recent years. This study used ML to predict weight recovery in patients affected by AN attending a multidisciplinary intensive rehabilitation program starting from psychological and physical body parameters. Results revealed that body-self relationship and interpersonal difficulties play a pivotal role in weight restoring, suggesting that therapeutic interventions should focus more on psychological aspects, rather than purely nutritional interventions. If such sensitive topics could be difficult to tackle previously, technology offers new techniques through innovative techniques such as Body Swapping and activities: this would allow to address altered body experience and interpersonal difficulties respectively in immersive and engaging ways. This approach could also take into account the ego-syntonic typical of this disorder, i.e., the patients' difficulty in perceiving reduced food intake as a problem, and its complexity, which is often not centered exclusively on food. By acknowledging these factors, treatment efficacy can be maximized, ensuring that limited resources are strategically used to achieve meaningful clinical outcomes. We argue that this promising area of research could benefit from researchers from different fields working together to create large international datasets that could make a significant contribution to realizing the potential of ML, with the ultimate aim of helping as many people as possible.

Declarations

Human ethics and consent to participate

The dataset for the present study was derived from the research by Brusa et al. [27], which was approved by the Ethics Committee of the involved Institution (Reference number: 2022_11_22_05). All participants were volunteers who gave informed written consent before participating in this study,in the case of participants under 18, parents signed the written consent.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40337-025-01265-3.

Additional file1 (DOCX 302 kb)

Acknowledgements
Google Colab v. 0.0.1a2 in Python version v. 3.10.12 was used for predictive analysis by machine learning. The authors take responsibility for the generated code.

Author contributions
GB = conceptualization, study conception and design, draft manuscript preparation; CP = data analysis and interpretation of results, draft manuscript preparation; ES = data analysis, writing; MB = data collection, draft manuscript preparation; FB = data collection, draft manuscript preparation; FS = data collection, writing; LM = supervision; GR = supervision. All authors reviewed the results and approved the final version of the manuscript.

Funding
This work was supported by the Italian Ministry of Health—Ricerca Corrente.

Data availability
The dataset and the code are available at https://github.com/chiara-pupillo/ML-AN.

Declarations

Conflict of interest

The authors declare no competing interests.

Author details

[1]Department of Psychology, Università Cattolica del Sacro Cuore, Largo Gemelli, 20121 Milan, Italy. [2]Humane Technology Laboratory, Università Cattolica del Sacro Cuore, Largo Gemelli, 20121 Milan, Italy. [3]Department of Computer Science, University of Pisa, Pisa, Italy. [4]Experimental Laboratory for Metabolic Neurosciences Research, I.R.C.C.S. Istituto Auxologico Italiano, 28824 Piancavallo, VCO, Italy. [5]"Rita Levi Montalcini" Department of Neurosciences, University of Turin, Turin, Italy. [6]U.O. di Neurologia e Neuroriabilitazione, Ospedale San Giuseppe, I.R.C.C.S. Istituto Auxologico Italiano, Piancavallo, VCO, Italy. [7]Applied Technology for Neuro-Psychology Laboratory, IRCCS Istituto Auxologico Italiano, 20149 Milan, Italy.

Received: 25 November 2024 Accepted: 14 April 2025
Published online: 02 June 2025

References

1. American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596
2. Walsh BT, Hagan KE, Lockwood C. A systematic review comparing atypical anorexia nervosa and anorexia nervosa. Int J Eat Disord. 2022;56(4):798–820. https://doi.org/10.1002/eat.23856.
3. Barakat S, Aouad P, Boakes RA, Brennan L, Bryant E, Byrne S, Maguire S. Risk factors for eating disorders: findings from a rapid review. J Eating Disorders. 2023. https://doi.org/10.1186/s40337-022-00717-4.
4. Brizzi G, Sansoni M, Di Lernia D, Frisone F, Tuena C, Riva G. The multisensory mind: a systematic review of multisensory integration processing in Anorexia and Bulimia Nervosa. J Eat Disord. 2023;11(1):204.
5. Strober M. Pathologic fear conditioning and anorexia nervosa: on the search for novel paradigms. Int J Eat Disord. 2004;35(4):504–8. https://doi.org/10.1002/eat.20029.
6. Bell C, Cooper MJ. Socio-cultural and cognitive predictors of eating disorder symptoms in young girls. Eating Weight Disorders-Stud Anorexia Bulimia Obesity. 2005;10:e97–100.
7. Garner DM, Garfinkel PE. Socio-cultural factors in the development of anorexia nervosa. Psychol Med. 1980;10(4):647–56. https://doi.org/10.1017/s0033291700054945.
8. Andrés-Pepiñá S, Plana MT, Flamarique I, Romero S, Borràs R, Julià L, Gárriz M, Castro-Fornieles J. Long-term outcome and psychiatric comorbidity of adolescent-onset anorexia nervosa. Clin Child Psychol Psychiatry. 2020;25(1):33–44. https://doi.org/10.1177/1359104519827629.
9. Auger N, Potter BJ, Ukah UV, Low N, Israël M, Steiger H, Healy-Profitós J, Paradis G. Anorexia nervosa and the long-term risk of mortality in women. World Psychiatry Off J World Psychiatric Assoc (WPA). 2021;20(3):448–9. https://doi.org/10.1002/wps.20904.
10. Frank GKW, Stoddard JJ, Brown T, Gowin J, Kaye WH. Weight gained during treatment predicts 6-month body mass index in a large sample of patients with anorexia nervosa using ensemble machine learning. Int J Eat Disord. 2024;57(8):1653–67. https://doi.org/10.1002/eat.24208.
11. Nagata JM, Golden NH. New us preventive services task force recommendations on screening for eating disorders. JAMA Intern Med. 2022;182(5):471. https://doi.org/10.1001/jamainternmed.2022.0121.
12. Boltri M, Brusa F, Apicella E, Mendolicchio L. Short- and long-term effects of Covid-19 pandemic on health care system for individuals with eating disorders. Front Psych. 2024;15:1360529. https://doi.org/10.3389/fpsyt.2024.1360529.
13. Kazdin AE, Fitzsimmons-Craft EE, Wilfley DE. Addressing critical gaps in the treatment of eating disorders. Int J Eat Disord. 2017;50(3):170–89. https://doi.org/10.1002/eat.22670.
14. Mehler PS, Anderson K, Bauschka M, Cost J, Farooq A. Emergency room presentations of people with anorexia nervosa. J Eat Disord. 2023;11(1):16.
15. Fichter MM, Quadflieg N, Crosby RD, Koch S. Long-term outcome of anorexia nervosa: results from a large clinical longitudinal study. Int J Eat Disord. 2017;50(9):1018–30.
16. Fardouly J, Crosby RD, Sukunesan S. Potential benefits and limitations of machine learning in the field of eating disorders: current research and future directions. J Eat Disord. 2022;10:66. https://doi.org/10.1186/s40337-022-00581-2.
17. Sajno E, Bartolotta S, Tuena C, Cipresso P, Pedroli E, Riva G. Machine learning in biosignals processing for mental health: a narrative review. Front Psychol. 2023;13:1066317. https://doi.org/10.3389/fpsyg.2022.1066317.
18. Ghosh S, Burger P, Simeunovic M, Maas J, Petkovic M. Review of machine learning solutions for eating disorders. Int J Med Inf 105526 (2024).
19. Jankowsky K, Krakau L, Schroeders U, Zwerenz R, Beutel ME. Predicting treatment response using machine learning: a registered report. Br J Clin Psychol. 2024;63:137–55. https://doi.org/10.1111/bjc.12452.
20. Sajjadian M, Lam RW, Milev R, Rotzinger S, Frey BN, Soares CN, Parikh SV, Foster JA, Turecki G, Müller DJ, Strother SC, Farzan F, Kennedy SH, Uher R. Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis. Psychol Med. 2021;51(16):2742–51.
21. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, Corlett PR. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry. 2016;3(3):243–50. https://doi.org/10.1016/S2215-0366(15)00471-X.
22. Grassi M, Rouleaux N, Caldirola D, Loewenstein D, Schruers K, Perna G, Dumontier M. A novel ensemble-based machine learning algorithm to predict the conversion from mild cognitive impairment to Alzheimer's disease using socio-demographic characteristics, clinical information, and neuropsychological measures. Front Neurol. 2019;10:756. https://doi.org/10.3389/fneur.2019.00756.
23. Lavagnino L, Amianto F, Mwangi B, D'Agata F, Spalatro A, Zunta-Soares GB, Soares JC. Identifying neuroanatomical signatures of anorexia nervosa: a multivariate machine learning approach. Psychol Med. 2015;45(13):2805–12. https://doi.org/10.1017/S0033291715000768.
24. Arold D, Bernardoni F, Geisler D, et al. Predicting long-term outcome in anorexia nervosa: a machine learning analysis of brain structure at different stages of weight recovery. Psychol Med. 2023;53(16):7827–36. https://doi.org/10.1017/S0033291723001861.
25. Merhbene G, Puttick A, Kurpicz-Briki M. Investigating machine learning and natural language processing techniques applied for detecting eating disorders: a systematic literature review. Front Psych. 2024;15:1319522. https://doi.org/10.3389/fpsyt.2024.1319522.
26. Sandoval-Araujo LE, Cusack CE, Ralph-Nearman C, Glatt S, Han Y, Bryan J, Hooper MA, Karem A, Levinson CA. Differentiation between atypical anorexia nervosa and anorexia nervosa using machine learning. Int J Eat Disord. 2024;57(4):937–50. https://doi.org/10.1002/eat.24160.
27. Brusa F, Scarpina F, Bastoni I, Villa V, Castelnuovo G, Apicella E, Mendolicchio L. Short-term effects of a multidisciplinary inpatient intensive rehabilitation treatment on body image in anorexia nervosa. J Eating Disorders. 2023;11(1):178.
28. Flanagin A, Pirracchio R, Khera R, Berkwits M, Hswen Y, Bibbins-Domingo K. Reporting use of AI in research and scholarly publication—JAMA Network Guidance. JAMA (2024).
29. Serino S, Chirico A, Pedroli E, Polli N, Cacciatore C, Riva G. Two-phases innovative treatment for anorexia nervosa: the potential of virtual reality body-swap. Annu Rev CyberTherapy Telemed. 2017;15:111–5.
30. Kaplan AS, Walsh BT, Olmsted M, Attia E, Carter JC, Devlin MJ, Pike KM, Woodside B, Rockert W, Roberto CA, Parides M. The slippery slope: prediction of successful weight maintenance in anorexia nervosa. Psychol Med. 2009;39(6):1037–45. https://doi.org/10.1017/S003329170800442X.
31. Makhzoumi SH, Coughlin JW, Schreyer CC, Redgrave GW, Pitts SC, Guarda AS. Weight gain trajectories in hospital-based treatment of anorexia nervosa. Int J Eat Disord. 2017;50(3):266–74. https://doi.org/10.1002/eat.22679.
32. Toppino F, Longo P, Martini M, Abbate-Daga G, Marzola E. Body mass index specifiers in anorexia nervosa: anything below the "extreme"? J Clin Med. 2022;11(3):542. https://doi.org/10.3390/jcm11030542.

33. Maurel L, MacKean M, Lacey JH. Factors predicting long-term weight maintenance in anorexia nervosa: a systematic review. Eat Weight Disord. 2024;29:24. https://doi.org/10.1007/s40519-024-01649-5.
34. Cuzzolaro M, Vetrone G, Marano G, Garfinkel PE. The Body Uneasiness Test (BUT): development and validation of a new body image assessment scale. Eating Weight Disorders EWD. 2006;11(1):1–13. https://doi.org/10.1007/BF03327738.
35. Clausen L, Rosenvinge JH, Friborg O, Rokkedal K. Validating the eating disorder inventory-3 (EDI-3): a comparison between 561 female eating disorders patients and 878 females from the general population. J Psychopathol Behav Assess. 2010;33(1):101–10.
36. Sarno I, Preti E, Prunas A, Madeddu F. SCL-90-R. Symptom checklist-90-R. Giunti Organizzazioni Speciali: Firenze (2011).
37. Frost RO, Marten P, Lahart C, Rosenblate R. Frost Multidimensional Perfectionism Scale (FMPS) [Database record]. APA PsycTests (1990). https://doi.org/10.1037/t05500-000
38. Batista GEAPA, Monard MC. An analysis of four missing data treatment methods for supervised learning. Appl Artif Intell. 2003;17(5–6):519–33. https://doi.org/10.1080/713827181.
39. De Amorim LB, Cavalcanti GD, Cruz RM. The choice of scaling technique matters for classification performance. Appl Soft Comput. 2023;133:109924.
40. Graf R, Zeldovich M, Friedrich S. Comparing linear discriminant analysis and supervised learning algorithms for binary classification—a method comparison study. Biom J. 2024;66(1):2200098.
41. Gholamy A, Kreinovich V, Kosheleva O. Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation. Int J Intell Technol Appl Stat. 2018;11(2):105–11.
42. Yu T, Zhu H. Hyper-parameter optimization: a review of algorithms and applications (2020). arXiv:2003.05689.
43. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep. 2022;12:5979. https://doi.org/10.1038/s41598-022-09954-8.
44. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432.
45. Tougui I, Jilbab A, El Mhamdi J. Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. Healthc Inf Res. 2021;27(3):189–99.
46. Wilimitis D, Walsh CG. Practical considerations and applied examples of cross-validation for model development and evaluation in health care: tutorial. JMIR AI. 2023;18(2):e49023. https://doi.org/10.2196/49023.
47. Guidotti R, Monreale A, Pedreschi D, Giannotti F. Principles of explainable artificial intelligence. Explainable AI Within the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications, 9–31 (2021).
48. Lundberg S. A unified approach to interpreting model predictions (2017). arXiv:1705.07874.
49. Dhamnetiya D, Goel MK, Jha RP, Shalini S, Bhattacharyya K. How to perform discriminant analysis in medical research? Explained with illustrations. J Lab Phys. 2022;14(04):511–20.
50. Thölke P, Mantilla-Ramos YJ, Abdelhedi H, Maschke C, Dehgan A, Harel Y, Kemtur A, Mekki Berrada L, Sahraoui M, Young T, Bellemare Pépin A, El Khantour C, Landry M, Pascarella A, Hadid V, Combrisson E, O'Byrne J, Jerbi K. Class imbalance should not throw you off balance: choosing the right classifiers and performance metrics for brain decoding with imbalanced data. Neuroimage. 2023;277:120253. https://doi.org/10.1016/j.neuroimage.2023.120253.
51. Espel-Huynh H, Zhang F, Thomas JG, Boswell JF, Thompson-Brenner H, Juarascio AS, Lowe MR. Prediction of eating disorder treatment response trajectories via machine learning does not improve performance versus a simpler regression approach. Int J Eat Disord. 2021;54(7):1250–9.
52. Haynos AF, Wang SB, Lipson S, Peterson CB, Mitchell JE, Halmi KA, Crow SJ. Machine learning enhances prediction of illness course: a longitudinal study in eating disorders. Psychol Med. 2021;51(8):1392–402. https://doi.org/10.1017/S0033291720000227.
53. Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
54. Orlenko A, Moore JH. A comparison of methods for interpreting random forest models of genetic association in the presence of non-additive interactions. BioData Mining. 2021;14:9. https://doi.org/10.1186/s13040-021-00243-0.
55. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst. 2021;32(11):4793–813. https://doi.org/10.1109/TNNLS.2020.3027314.
56. Bruch H. Eating disorders: obesity, anorexia nervosa, and the person within. London: Routlege & Kegan; 1974.
57. Brizzi G, Sansoni M, Riva G. The BODY-FRIEND project: using new technology to learn about how people with anorexia feel about their bodies. Cyberpsychol Behav Soc Netw. 2023;26(2):141–3.
58. Artoni P, Chierici ML, Arnone F, Cigarini C, De Bernardis E, Galeazzi GM, Pingani L. Body perception treatment, a possible way to treat body image disturbance in eating disorders: a case–control efficacy study. Eating Weight Disorders-Stud Anorexia Bulimia Obesity. 2021;26:499–514. https://doi.org/10.1007/s40519-020-00875-x.
59. Sattler FA, Eickmeyer S, Eisenkolb J. Body image disturbance in children and adolescents with anorexia nervosa and bulimia nervosa: a systematic review. Eating Weight Disorders-Stud Anorexia Bulimia Obes. 2020;25:857–65.
60. Provenzano L, Porciello G, Ciccarone S, Lenggenhager B, Tieri G, Marucci M, Bufalari I. Characterizing body image distortion and bodily self-plasticity in anorexia nervosa via visuo-tactile stimulation in virtual reality. J Clin Med. 2019;9(1):98.
61. Gan JKE, Wu VX, Chow G, Chan JKY, Klainin-Yobas P. Effectiveness of non-pharmacological interventions on individuals with anorexia nervosa: a systematic review and meta-analysis. Patient Educ Couns. 2022;105(1):44–55.
62. Cash TF, Hrabosky JI. The effects of psychoeducation and self-monitoring in a cognitive-behavioral program for body-image improvement. Eat Disord. 2003;11(4):255–70.
63. Mahon C, Hevey D. Pilot trial of a self-compassion intervention to address adolescents' social media-related body image concerns. Clin Child Psychol Psychiatry. 2023;28(1):307–32.
64. Rizzuto L, Hay P, Noetel M, Touyz S. Yoga as adjunctive therapy in the treatment of people with anorexia nervosa: a Delphi study. J Eat Disord. 2021;9:1–12.
65. Sala M, Levinson CA, Kober H, Roos CR. A pilot open trial of a digital mindfulness-based intervention for anorexia nervosa. Behav Ther. 2023;54(4):637–51.
66. Danielsen YS, Ärdal Rekkedal G, Frostad S, Kessler U. Effectiveness of enhanced cognitive behavioral therapy (CBT-E) in the treatment of anorexia nervosa: a prospective multidisciplinary study. BMC Psychiatry. 2016;16:1–14.
67. Stice E, Presnell K. The body project: promoting body acceptance and preventing eating disorders. Oxford: Oxford University Press; 2007.
68. Malighetti C, Chirico A, Serino S, Cavedoni S, Matamala-Gomez M, Stramba-Badiale C, Riva G. Manipulating body size distortions and negative body-related memories in patients with Anorexia Nervosa: a virtual reality-based pilot study. Ann Rev CyberTherapy Telemed (2020).
69. Di Natale AF, Pizzoli SFM, Brizzi G, Di Lernia D, Frisone F, Gaggioli A, Riva G. Harnessing immersive virtual reality: a comprehensive scoping review of its applications in assessing, understanding, and treating eating disorders. Curr Psychiatry Rep. 2024;26:470–86.
70. Carter JC, Kelly AC, Norwood SJ. Interpersonal problems in anorexia nervosa: social inhibition as defining and detrimental. Personal Individ Differ. 2012;53(3):169–74.
71. Jones A, Lindekilde N, Lübeck M, Clausen L. The association between interpersonal problems and treatment outcome in the eating disorders: a systematic review. Nord J Psychiatry. 2015;69(8):563–73.
72. Martini M, Marzola E, Musso M, Brustolin A, Abbate-Daga G. Association of emotion recognition ability and interpersonal emotional competence in anorexia nervosa: a study with a multimodal dynamic task. Int J Eat Disord. 2022;56(2):407–17. https://doi.org/10.1002/eat.23854.
73. Cardi V, Mallorqui-Bague N, Albano G, Monteleone AM, Fernandez-Aranda F, Treasure J. Social difficulties as risk and maintaining factors in anorexia nervosa: a mixed-method investigation. Front Psych. 2018;9:12.
74. Carfagno M, Barone E, Arsenio E, et al. Mediation role of interpersonal problems between insecure attachment and eating disorder

psychopathology. Eat Weight Disord. 2024;29:43. https://doi.org/10.1007/s40519-024-01673-5.
75. Ung EM, Erichsen CB, Poulsen S, et al. The association between interpersonal problems and treatment outcome in patients with eating disorders. J Eat Disord. 2017;5:53. https://doi.org/10.1186/s40337-017-0179-6.
76. Sullivan HS. The psychiatric interview (No. 506). New York: WW Norton & Company; 1954.
77. Brown GW, Harris T. Social origins of depression: a study of psychiatric disorder in women. London: Routledge; 2012.
78. Schmidt U, Treasure J. Anorexia nervosa: Valued and visible. A cognitive-interpersonal maintenance model and its implications for research and practice. Br J Clin Psychol. 2006;45:343–66.
79. Recalcati M. L'ultima cena: anoressia e bulimia. Pearson Italia Spa (2007).
80. Cosenza D. A Lacanian reading of anorexia. Abingdon-on-Thames: Taylor & Francis; 2023.
81. Abínzano R. Algunos basamentos antropológicos y filosóficos de la concepción de anorexia mental de Jacques Lacan. Affectio Societatis. 2021;18(35):4.
82. Abbate-Daga G, Amianto F, Delsedime N, De-Bacco C, Fassino S. Resistance to treatment and change in anorexia nervosa: a clinical overview. BMC Psychiatry. 2013;13:1–18. https://doi.org/10.1186/1471-244X-13-294.
83. Allison S, Warin M, Bastiampillai T, Looi JC, Strand M. Recovery from anorexia nervosa: the influence of women's sociocultural milieux. Australas Psychiatry. 2021;29(5):513–5. https://doi.org/10.1177/10398562211010796.
84. Loomes R, Bryant-Waugh R. Widening the reach of family-based interventions for anorexia nervosa: autism-adaptations for children and adolescents. J Eating Disorders. 2021. https://doi.org/10.1186/s40337-021-00511-8.
85. Baudinet J, Eisler I, Konstantellou A, Hunt T, Kassamali F, McLaughlin N, Schmidt U. Perceived change mechanisms in multi-family therapy for anorexia nervosa: a qualitative follow-up study of adolescent and parent experiences. Eur Eating Disorders Rev. 2023;31(6):822–36. https://doi.org/10.1002/erv.3006.
86. Cerasa A, Gaggioli A, Marino F, Riva G, Pioggia G (2022) The promise of the metaverse in mental health: the new era of MEDverse. Heliyon, 8(11).
87. Treasure J, Schmidt U. The cognitive-interpersonal maintenance model of anorexia nervosa revisited: a summary of the evidence for cognitive, socio-emotional and interpersonal predisposing and perpetuating factors. J Eat Disord. 2013;1:13. https://doi.org/10.1186/2050-2974-1-13.
88. Schlapobersky J. From the couch to the circle: Group-analytic psychotherapy in practice. Abingdon-on-Thames: Routledge; 2016.
89. Chekroud AM, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, Cohen Z, Belgrave D, DeRubeis R, Iniesta R, Dwyer D, Choi K. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry Off J World Psychiatric Assoc (WPA). 2021;20(2):154–70. https://doi.org/10.1002/wps.20882.
90. Wang SB. Machine learning to advance the prediction, prevention and treatment of eating disorders. Eur Eating Disorders Rev J Eating Disorders Assoc. 2021;29(5):683–91. https://doi.org/10.1002/erv.2850.
91. James G. An introduction to statistical learning (2013).
92. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
93. Yang Y, Lin M, Zhao H, Peng Y, Huang F, Lu Z. A survey of recent methods for addressing AI fairness and bias in biomedicine. J Biomed Inf. 2024. https://doi.org/10.1016/j.jbi.2024.104646.
94. Ravindran AC, Kokjohn SL. Evaluation of the sample size requirements of machine learning models used in engine design and research. Int J Engine Res. 2023;24(7):2973–90. https://doi.org/10.1177/14680874221137185.
95. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14(5):365–76. https://doi.org/10.1038/nrn3475.
96. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206–15. https://doi.org/10.1038/s42256-019-0048-x.
97. Gupta J, Seeja KR. A comparative study and systematic analysis of XAI models and their applications in healthcare. Arch Comput Methods Eng 1–26 (2024).
98. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI—explainable artificial intelligence. Sci Robot. 2019;4(37):eaay7120.
99. Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, Ranjan R. Explainable AI (XAI): core ideas, techniques, and solutions. ACM Comput Surv. 2023;55(9):1–33.
100. Del Fabro L, Bondi E, Serio F, et al. Machine learning methods to predict outcomes of pharmacological treatment in psychosis. Transl Psychiatry. 2023;13:75. https://doi.org/10.1038/s41398-023-02371-z.
101. Cerasa A, Castiglioni I, Salvatore C, Funaro A, Martino I, Alfano S, Donzuso G, Perrotta P, Gioia MC, Gilardi MC, Quattrone A. Biomarkers of eating disorders using support vector machine analysis of structural neuroimaging data: preliminary results. Behav Neurol. 2015;10:e924814. https://doi.org/10.1155/2015/924814.
102. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. 2016;18(12):e323.
103. Couturier J, Lock J. What is recovery in adolescent anorexia nervosa? Int J Eat Disord. 2006;39(7):550–5. https://doi.org/10.1002/EAT.20309.
104. Gowers S, Smyth BP. The impact of a motivational assessment interview on initial response to treatment in adolescent anorexia nervosa. Eur Eat Disord Rev. 2004;12(2):87–93. https://doi.org/10.1002/erv.555.
105. Dupuy HJ. "The psychological general well-being (PGWB) Index," in Assessment of Quality of Life in Clinical Trials of Cardiovascular Therapies. In: Wenger N editor. New York: Le Jacq1984. 170–83
106. Ferrer-Garcia M, Porras-Garcia B, Miquel H, Serrano-Troncoso E, Carulla-Roig M, Gutiérrez J. The way we look at our own body really matters! Body-related attentional bias as a predictor of worse clinical outcomes after a virtual reality body exposure therapy. Annu Rev Cybertherapy Telemed. 2021;19:99.
107. Khalil SF, Mohktar MS, Ibrahim F. The theory and fundamentals of bioimpedance analysis in clinical status monitoring and diagnosis of diseases. Sensors (Basel). 2014;14(6):10895–928. https://doi.org/10.3390/s140610895.
108. Serino S, Baglio F, Rossetto F, et al. Picture interpretation test (PIT) 360°: an innovative measure of executive functions. Sci Rep. 2017;7:16000. https://doi.org/10.1038/s41598-017-16121-x.
109. Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate data analysis. 7th Edition. New York; Pearson. 2010.
110. Serino S, Polli N, Riva G. From avatars to body swapping: The use of virtual reality for assessing and treating body-size distortion in individuals with anorexia. J Clin Psychol. 2019;75(2):313-322. https://doi.org/10.1002/jclp.22724.
111. Koskina A, Schmidt U. Who am I without anorexia? Identity exploration in the treatment of early stage anorexia nervosa during emerging adulthood: a case study. Cogn Behav Therapist. 2019;12:e32.
112. Riva G. Neuroscience and eating disorders: The allocentric lock hypothesis. Med Hypotheses. 2012;78(2):254–257.
113. Slater M. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philos Trans R Soc Lond B Biol Sci. 2009;364(1535):3549-57.
114. Lipsitz JD, Markowitz JC. Mechanisms of change in interpersonal therapy (IPT). Clinical Psychol Rev. 2013;33(8):1134–1147. https://doi.org/10.1016/j.cpr.2013.09.002.
115. Koutsouleris N, Hauser TU, Skvortsova V, De Choudhury M. From promise to practice: towards the realisation of AI-informed mental health care. Lancet Digit Health. 2022;4(11):e829-e840. https://doi.org/10.1016/S2589-7500(22)00153-4.
116. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, Marsh J, DeVylder J, Walter M, Berrouiguet S, Lemey C. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. J Med Internet Res. 2021;23(5):e15708. https://doi.org/10.2196/15708.
117. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. BMC Med Inform Decis Mak. 2012;12:8. https://doi.org/10.1186/1472-6947-12-8.
118. Lucherini Angeletti L, Innocenti M, Felciai F, Ruggeri E, Cassioli E, Rossi E, Rotella F, Castellini G, Stanghellini G, Ricca V, Northoff G. Anorexia

nervosa as a disorder of the subcortical-cortical interoceptive-self. Eat Weight Disord. 2022;27(8):3063-3081. https://doi.org/10.1007/s40519-022-01510-7.
119. Mountford VA, Brown A, Bamford B, Saeidi S, Morgan JF, Lacey H. BodyWise: evaluating a pilot body image group for patients with anorexia nervosa. Eur Eating Disorders Rev J Eating Disorders Assoc. 2015;23(1):62–7. https://doi.org/10.1002/erv.2332
120. Brizzi G, Boltri M, Guglielmini R, Castelnuovo G, Mendolicchio L, Riva G. Therapeutic immersion: a single-subject study on virtual reality multisensory experiences for mitigating body disturbance in anorexia nervosa. Eat Weight Disord. 2025;30:32. https://doi.org/10.1007/s40519-025-01740-5.
121. Riva G, Serino S, Di Lernia D, Pagnini F. Regenerative virtual therapy: the use of multisensory technologies and mindful attention for updating the altered representations of the bodily self. Front Syst Neurosci. 2021;15:749268. https://doi.org/10.3389/fnsys.2021.749268.

Publisher's Note