Exploring Depression Severity During Lockdown Through Explainable AI

Stefania Zinno*, Sayna Rotbei*, Giovanni Stanco*, Giorgio Ventre* Giordano D'Urso†, Alessio Botta*
*Department of Electrical Engineering and Information Technology, University 'Federico II' of Napoli, Italy
†Department of Neuroscience, Reproductive and Odontostomatological Sciences, Section of Psychiatry, University 'Federico II' of Napoli, Italy
{stefania.zinno, sayna.rotbei, giovanni.stanco, giorgio.ventre, giordano.durso, alessio.botta}@unina.it

***Abstract*—Depression and anxiety are among the most common responses to large-scale traumatic episodes (e.g. pandemics or armed conflicts), particularly when these involve prolonged restriction measures for the population. The goal of this study is to predict the extent of depressive symptoms, evaluated using the Beck Depression Inventory-II (BDI-II) scale, during the COVID-19 lockdown, starting from scores of multiple psychological scales collected prior to the lockdown. More importantly, we aim to identify the most influential features driving these predictions through eXplainable AI (XAI) techniques. To this end, we selected Gradient Boosting as our predictive model and applied SHapley Additive exPlanations (SHAP) to assess feature importance. We then generated Partial Dependence Plots (PDPs) on the top-ranking features identified by SHAP to further explore their impact on the model's output. Scores from Item 9, Item 3, and Item 1 of the BDI-II scales, regarding *Suicidal Thoughts*, *Sadness* and *Failure* emerged as key predictors in the model.**

***Index Terms*—Machine Learning, Explainable AI, depression**

I. Introduction

Environmental stressors are one of the causes that may trigger psychiatric symptoms in sensitive individuals, or exacerbate the severity of previously occurring conditions [1]. Lockdowns and limitations imposed because of the COVID-19 pandemic caused a variety of indirect psycho-social consequences, such as enforced isolation, change of established social behaviors, and worries about what lies ahead due to health and economic instability. Increased levels of depression were experienced by both the general population and psychiatric patients, along with a strong rise in suicidal thoughts and insomnia episodes [2]. In this scenario, it is crucial to understand which psychological characteristics to monitor to prevent the onset of depressive symptoms with the aim of improving the well-being of the population.

Machine Learning (ML) has been commonly used for finding complex correlations between variables and making accurate predictions across various domains, including healthcare, finance, networking and behavioral sciences [3], [4]. In the medical field in particular, ML has emerged as a powerful tool for managing data, allowing the discovery of patterns and predictors across many disciplines including health diagnostics [5], cancer prognosis [6], and diabetes management [7], [8].

In the first stages of the COVID-19 outbreak, ML was adopted by means of multiple logistic regression algorithms, a type of supervised learning, to determine predictors of anxiety and depression: a study among the Brazilian population indicated that women, younger individuals, and adults with fewer children were at greater risk of developing these symptoms [9]. ML has also been used to assess whether factors such as stress, sleep quality, and personality were related to depression levels measured by the Beck Depression Inventory-II (BDI-II) scale, an index of depression severity. Positive associations were found, and the models showed high accuracy in distinguishing between different depression levels [10]. ML has also been applied to recognize anxiety and depression through gait analysis. Fifty participants provided responses according to the BDI-II scale before walking in front of a Kinect sensor, able to record their gait. Algorithms such as Support Vector Machines (SVM), and deep neural network were adopted to predict anxiety and depression from gait features achieving high accuracies [11]. An SVM model based on acoustic features demonstrated strong capabilities in identifying individuals with major depressive disorder from healthy ones, with the same performance of an SVM model based on BDI-II scores. The model was trained on recording of patients discussing positive and negative life events [12]. Random Forest regressors were applied to data collected from cognitive and behavioral tests from 237 individuals with different anxiety and depression levels, evaluated by the State-Trait Anxiety Inventory (STAI-Y) and BDI-II scales. The analysis identified bias patterns unique to anxiety, as well as those common to both anxiety and depression [13].

Leveraging these approaches, this study aims to predict depression severity during a restriction period, from psychological scale scores collected before the lockdown. The goal is to identify specific instruments to detect higher risks of developing or increasing depression during pandemic related restrictions by analyzing the pre-lockdown clinical profiles of individuals.

The geographical context in which we carried out our research adds value to this work, since Italy was heavily affected by the COVID-19 pandemic, and its population strongly perceived the strict lockdown measures. Leveraging this unique context, the aim of this work is twofold: first, to predict the level of depression during large-scale lockdown based on demographic data and scores from psychological scales; and second, to identify through eXplainable AI (XAI)

TABLE I: Demographic and medical history data of participants. For Education: P stands for Primary, M for Middle, H for High, U for University.

Gender		Age			Education				Health worker		Positive to COVID-19		Medical comorbities		Family issues due to lockdown		Family history of OCD	
Female	Male	[0; 25]	[26; 50]	[51; 99]	P	M	H	U	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
43	51	15	46	33	2	26	43	23	4	89	15	78	34	59	6	85	10	83

techniques the specific questionnaire items that contribute most to the prediction, thereby supporting informed clinical decision-making.

In this paper, Section II provides a detailed description of the data collection process, the clinical measures, the ML algorithms employed, and the principles behind SHAP-based explainable AI. Section III reports the performance metrics of the models along with the results of the SHAP and PDP analyses. Eventually, conclusions are drawn.

II. Experimental Setup

A. Participants

For this study, three groups of participants were involved. Alongside a cohort of 29 healthy subjects used as a baseline, two samples of patients from the University Hospital 'Federico II', in Napoli, Italy, were recruited: one sample of 46 patients with obsessive-compulsive disorder and another of 19 suffering from adjustment disorder. Patients, aged from 18 to 70 years, exhibited psychopathological stability at the beginning of the process. The selected healthy subjects had no prior history of psychiatric illness and possessed demographic characteristics (e.g., age, sex, education) closely matching those of the clinical samples. Collected data from the participants included demographic information and psychiatric medical history information. The total 94 participants were assessed at two distinct time points: at t_0, prior to the pandemic, and at t_1, during the lockdown. Table I reports the detailed demographic and medical history characteristics of participants. Further clinical characteristics for all three groups are provided in [14]. This campaign was approved by the Ethics Committee for Biomedical Activities of the affiliated University Hospital, as documented in protocol number 152/20, dated 22 April 2020.

B. Clinical Measures

Data collected from the campaign were linked with multiple psychiatric assessment scores, obtained from patients' medical records or interviews. In the following, details about the scales are discussed.

The *Yale–Brown Obsessive-Compulsive Scale (Y-BOCS)* considers both current and lifetime manifestations of obsessive-compulsive disorder symptoms. It is commonly adopted to evaluate both the presence and severity of symptoms. The test is in the format of a symptom checklist, taking into account obsessions and compulsions. Each symptom is first classified as present or absent and then rated by severity. The severity scale consists of 5 items related to obsessions and 5 to compulsions, each rated on a scale from 0 to 4. Symptoms are assessed over the previous week across multiple dimensions, including time spent, distress, resistance, interference, and control [15].

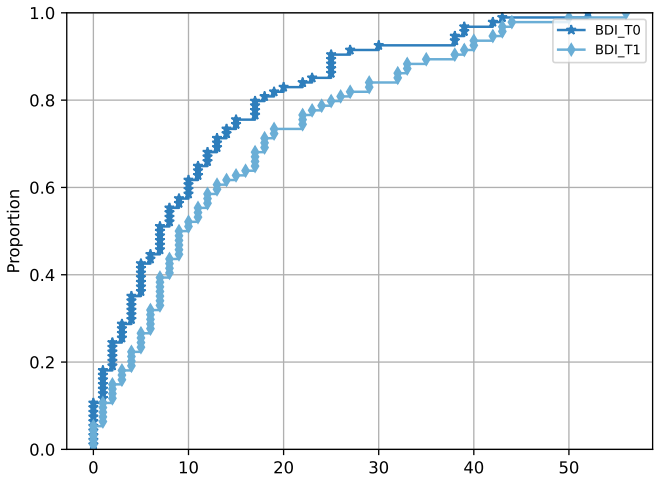


Fig. 1: CDF of BDI-II scores at t_0 (before lockdown) and t_1 (during lockdown): a rightward shift at t_1 indicates higher levels of depression during the lockdown period, compared to the baseline from the previous period.

The *Brown Assessment of Belief Scale (BABS)* is an instrument designed to assess how strongly individuals hold their beliefs and their delusionality. It is administered by clinicians and comprises seven items. The total score comes from the first six items, plus an additional item that is rated separately. Items are scored from 0 to 4, where 4 indicates poorer insights or more delusional thinking [16].

The *Beck Depression Inventory-II (BDI-II)* is widely adopted to assess the severity of depressive symptoms through a 21-item questionnaire. Symptoms are rated on a scale from 0 to 3, where 3 stands for higher levels of severity [17].

The *State-Trait Anxiety Inventory-Y (STAI-Y)* is a psychological assessment tool evaluating two forms of anxiety: state and trait. It comprises 40 self-report items that are rated on a 4-point Likert scale [18].

To support the prediction goal of our research, the four scales were administered to the research cohort at two distinct assessment points: prior to the COVID-19 pandemic (t_0) and during the Italian national lockdown (t_1), which occurred between March and June 2020. Owing to social distancing restrictions, follow-up assessments during the lockdown were administered through phone interviews by trained clinicians who were unaware of the participants' clinical conditions.

Pre-processing steps were performed prior to applying ML algorithms to the dataset, such as handling *NaN* values and

data standardization.

C. Prediction

In this study, pre-lockdown scores from the four scales are used as input features for predictive modelling. The target variable was the corresponding score from BDI-II scale, which measures depression during the lockdown period. By analyzing pre-lockdown scale values alongside demographic and medical history data, ML algorithms were trained to predict the psychological outcomes during the lockdown. Specifically, our goal is to predict depressive severity at t_1, during the lockdown, assessed using the BDI-II scale, by utilizing scores from other psychological scales collected at t_0, prior to the lockdown. The CDFs of the BDI-II scores mentioned above are shown in Fig. 1, illustrating the increase in BDI-II values at t_1 and highlighting the negative impact of the lockdown. Target BDI scores were categorized into two levels: scores between 0 and 19 were classified as low depression (*label = 0*), while scores between 20 and 63 were classified as high depression (*label = 1*). The individual raw scores of all scale items at t_0 were the input features used for the prediction models. Conversely, the clustered values of BDI-II scores served as the target variable to be predicted.

To predict individual outcomes on psychiatric scales, ten different ML classifiers were employed. Algorithms included ensemble methods such as Random Forest, AdaBoost, Gradient Boosting, and Extra Trees, as well as individual classifiers like Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Multi-Layer Perceptron (MLP), Stochastic Gradient Descent (SGD), and Logistic Regression. The algorithms were implemented in Python using the `scikit-learn` library [1].

In this study, to obtain more stable estimates and enhance the generalizability of model performance, the `K-fold` cross-validation technique was employed. All algorithms were also fine-tuned using `GridSearch` [2] to identify the best hyperparameter configuration.

D. Performance Metrics

The performance of the ML algorithms employed for the classification task is assessed using metrics including precision, recall, F1-score, and accuracy, able to reflect the model's ability to accurately predict depressive severity during the lockdown.

True positives (*TP*), false positives (*FP*), true negatives (*TN*), and false negatives (*FN*) are combined to calculate the *accuracy*, which measures the proportion of correct predictions made by the model out of all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

[1]https://scikit-learn.org

[2]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

TABLE II: Classification Scores for BDI-II at t_1

Classifier	Precision	Recall	F1 Score	Accuracy
Random Forest	0.88	0.87	0.86	0.87
SVM	0.89	0.88	0.87	0.88
KNN	0.76	0.75	0.73	0.75
Decision Tree	0.75	0.76	0.75	0.76
MLP	0.83	0.83	0.82	0.83
SGD	0.81	0.81	0.78	0.81
AdaBoost	0.87	0.86	0.85	0.86
Gradient Boosting	0.80	0.79	0.77	0.79
Extra Trees	0.88	0.88	0.88	0.88
Logistic Regression	0.90	0.89	0.89	0.89

Furthermore, *FP*, *FN*, and *TP* serve as the basis for calculating *precision* and *recall*, defined as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The *F1-score* is evaluated either by using both recall and precision or by using *TP*, *FP*, and *FN*, as shown below:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

E. Explainability Analysis: SHAP and PDP

Explainability analysis was conducted by means of SHapley Additive exPlanations (SHAP) and further explored with Partial Dependence Plots (PDPs). SHAP aims to explain the output of a ML model. Based on game theory, SHAP assigns an importance value to each feature, that measures the influence that the feature gives to the model's output [3].

PDPs are a well-established method for analyzing the relationship between a set of input features and a specific target variable, while controlling the effects of other variables [19]. This technique allows us to visually represent and understand the impact that each single predictor has on the outcome, offering valuable insights into how features interact with the target variable. A PDP shows how the model's average prediction changes as a single feature varies, while all other features are held constant, typically at their average values.

III. Experimental Results

In this section, we first discuss the results from the classification task, which aims to predict depression at t_1, during the lockdown, using scores from other psychological scales collected before the pandemic at t_0, along with demographic variables.

As mentioned, `K-fold` cross-validation was implemented to guarantee a thorough and reliable evaluation of the model, and all algorithms were fine-tuned through `GridSearch` for optimal hyperparameter selection. In Table II, all metrics from

[3]https://shap.readthedocs.io/en/latest/

algorithm performance are reported. The classification performance is generally strong, with accuracy almost exceeding 80% across all models.

Gradient Boosting was then selected to conduct SHAP and PDP analysis, as it is well-suited for feature importance interpretation. In particular, it effectively supports SHAP values and PDPs, which are essential for our interpretability objectives. For the explainability analysis, we first generated a summary plot, shown in Fig. 2. In a SHAP summary plot, features are ranked by their average contribution to the prediction, with the top ones being the most influential. In our case, the top-ranking features indicating the most relevant variable in predicting the outcome were Item 9 from the BDI-II scale, followed by Item 1 and Item 3 from the same scale. The Items are as follows:

Item 9 from BDI-II scale

9. *Suicidal Thoughts or Wishes*

0. I have no suicidal thoughts.
1. I have suicidal thoughts, but I would not act on them.
2. I feel I would be better off dead.
3. If given the chance, I would not hesitate to kill myself.

Item 1 from BDI-II scale

1. *Sadness*

0. I do not feel sad.
1. I feel sad most of the time.
2. I am always sad.
3. I feel so sad or unhappy that I can't bear it.

Item 3 from BDI-II scale

3. *Failure*

0. I do not feel like a failure.
1. I have failed more than I should have.
2. When I look back on my life, all I can see is a series of failures.
3. I feel like a total failure as a person.

The three most relevant questions that potentially determine depression in a lockdown phase are related to *Suicide*, *Sadness*, and *Failure* before the pandemic. Moreover, SHAP values indicate whether a feature increases or decreases the prediction. High scores from Item 9 strongly contribute to the prediction of the positive class. This means higher scores related to Item 9, *Suicide*-related, lead to high levels of depression symptoms during a possible lockdown. Item 3 follows the same trend, where at higher scores of *Failure* are associated higher chances of depression symptoms during a lockdown. Item 1 is highly relevant, but not as much, to the prediction direction. These results highlight how Suicide, Sadness, and Failure are the most immediate and recognizable emotions related to the development and worsening of depressive symptoms caused by a lockdown. They are also additional evidence that social isolation can increase the link to depression.

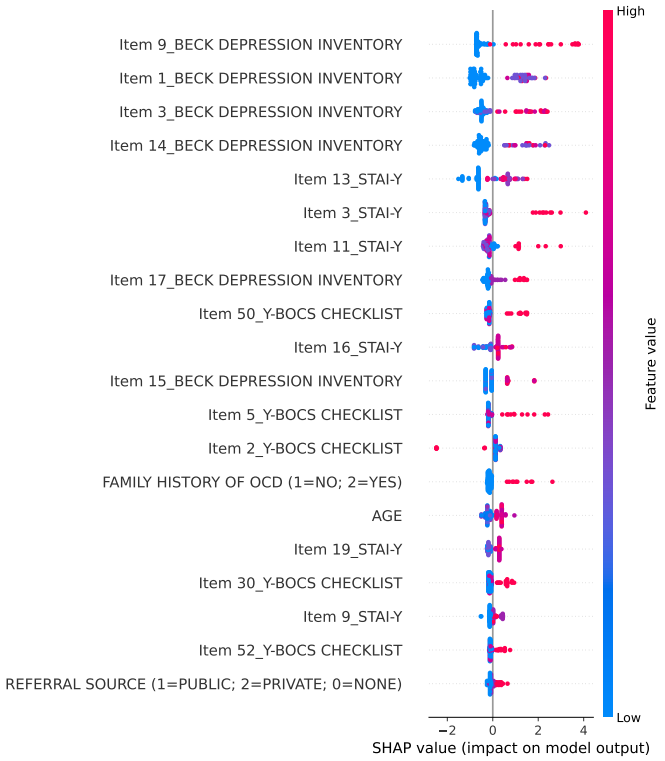
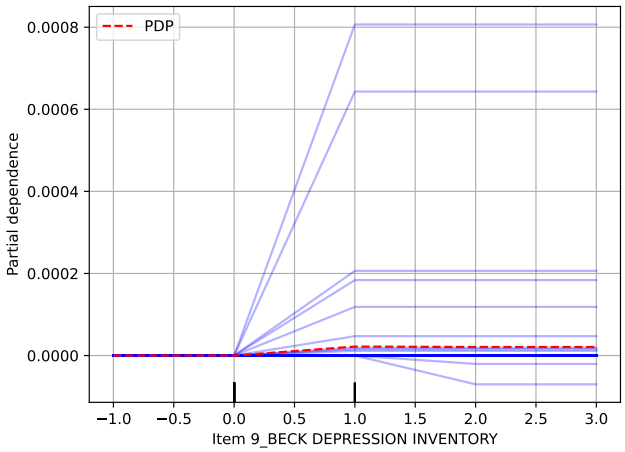


Fig. 2: SHAP summary plot of the Gradient Boosting model for the target variable *BDI-II*. Features are ranked by importance, highlighting the role of Item 9, 1 and 3 from BDI-II scale.

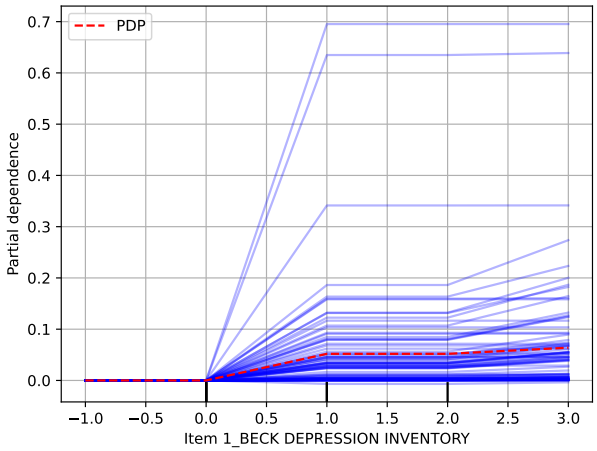
Since Item 9, 3, and 1 from BDI-II are the most relevant features identified from the SHAP analysis, we applied PDP to better interpret the model. The PDP for Item 9 (Fig. 3a) shows a largely flat trend, indicating that Item 9 has a hidden correlation with other variables and PDP is unable to show the influence the item has. The PDP for Item 1 (Fig. 3b) shows that the predicted probability of being classified as depressed increases when participants select option 0 or 1. For Item 3 (Fig. 3c), the prediction increases when participants select higher values. On the x-axis, in certain intervals the curve remains relatively constant, although it fluctuates slightly across different y-values. This suggests that after a certain level of failure or sadness is reached, further increases do not consistently influence the prediction the model's response becomes saturated.

IV. Conclusion

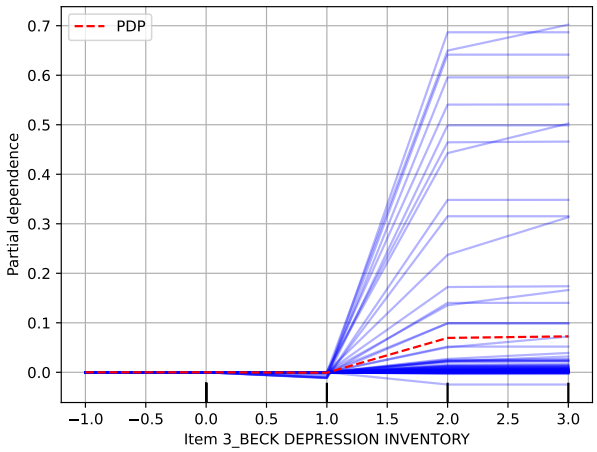
This paper presents an experimental study that aims to predict the Beck Depression Inventory (BDI-II) scores during the lockdown. The prediction leverages demographic information and scores from multiple psychological scales collected



(a) PDP related to the *Suicidal Thoughts* score



(b) PDP related to the *Sadness* score



(c) PDP related to *Failure* evaluation

Fig. 3: Partial Dependence Plots for selected BDI-II items at t_0. A noticeable increase in model response is observed for low-to-moderate scores of both Item 1 and Item 3 for BDI-II scale at t_0.

prior to the pandemic. The study was conducted in Italy, a country strongly impacted by the COVID-19 pandemic restrictions, in particular at the University Hospital 'Federico II' in Napoli, where complete standardized psychological questionnaires were administered to a selected population of patients both before and during the COVID-19 lockdown. The collected dataset has a limited sample size of 94 samples. To assess and confirm the generalizability of the outcome of this research, we plan to run other campaigns to collect more samples and to further validate our approach. The predictive performance was strong, with accuracy almost exceeding 80% in all cases. Gradient Boosting was selected as the best-performing model, and eXplainable AI (XAI) techniques such as SHAP and Partial Dependence analysis was applied. The analysis revealed that certain items such as Item 1, 3, and 9 from the BDI-II played a significant role in the prediction. These findings suggest that specific questionnaire items could serve as important indicators to monitor in order to anticipate depressive outcomes in similar future scenarios.

Acknowledgments

This work was partially supported by project cyberHuman, part of the SERICS program (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU and by European Union through the ADAPTO project, part of the RESTART program, NextGenerationEU PNRR, CUP E63C2 2002040007, CP PE0000001.

References

[1] Ladislav Kesner and Jiří Horáček. Three challenges that the COVID-19 pandemic represents for psychiatry. *The British Journal of Psychiatry*, 217(3):475–476, 2020.

[2] Andrea Fiorillo, Gaia Sampogna, Vincenzo Giallonardo, Valeria Del Vecchio, Mario Luciano, Umberto Albert, Claudia Carmassi, Giuseppe Carrà, Francesca Cirulli, Bernardo Dell'Osso, et al. Effects of the lockdown on the mental health of the general population during the COVID-19 pandemic in Italy: Results from the COMET collaborative network. *European Psychiatry*, 63(1), 2020.

[3] Nicola Pasquino, Stefania Zinno, Federica Cotugno, and Sofia Petrocelli. A comparative approach of unsupervised machine learning techniques for lte network parameter clustering. In *2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6, 2020.

[4] Stefano Avallone, Nicola Pasquino, Giorgio Ventre, and Stefania Zinno. Experimental characterization of long term evolution multiple input multiple output performance in urban propagation scenarios. In *2018 Workshop on Metrology for Industry 4.0 and IoT*, pages 254–259, 2018.

[5] Sayna Rotbei, Gennaro Esposito Mocerino, Muhammad Salman Haleem, Leandro Pecchia, and Alessio Botta. Frequency and uncertainty driven deep learning approach to segment electrocardiogram signals for effective heart parameters estimation. In *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–8, 2024.

[6] Sayna Rotbei, Luigi Napolitano, Stefania Zinno, Paolo Verze, and Alessio Botta. Predicting patient sexual function after prostate surgery using machine learning. In *2023 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE, 2023.

[7] Sayna Rotbei, Pablo Matías Soler, Beatriz Merino-Barbancho, Hania Tourab, Arturo Corbatón Anchuelo, Luis Picazo García, Ricardo Mesanza Forés, Laura Mariel Matus, Ricardo Muñoz Albert, Aitor Odiaga Andicoechea, Raquel Piñero Panadero, María Ángeles Díez San Martín, Ainhoa Burzaco Sanchez, Rosana Soriano Barrón, Andrea Irimia, Esther Ruescas Esculano, Mireia Cramp Vinceixo, Fahd Beddar Chaib, Giuseppe Fico, and Alessio Botta. Prediction of glycemic

event in emergency section patients using machine learning. In *2024 IEEE International Conference on E-health Networking, Application & Services (HealthCom)*, pages 1–4, 2024.

[8] Sayna Rotbei, Wei Hsuan Tseng, Beatriz Merino-Barbancho, Muhammad Salman Haleem, Luis Montesinos, Leandro Pecchia, Giuseppe Fico, and Alessio Botta. Evaluating impact of movement on diabetes via artificial intelligence and smart devices systematic literature review. *Expert Systems with Applications*, page 125058, 2024.

[9] Stephen X Zhang, Hao Huang, Jizhen Li, Mayra Antonelli-Ponti, Scheila Farias de Paiva, and José Aparecido da Silva. Predictors of depression and anxiety symptoms in brazil during COVID-19. *International Journal of Environmental Research and Public Health*, 18(13):7026, 2021.

[10] Graziella Orrù, Rebecca Ciacchini, Anna Conversano, Ciro Conversano, and Angelo Gemignani. Beyond the hot flashes: how machine learning is uncovering the complexity of menopause-related depression. *CNS Spectrums*, 30(1), 2025.

[11] Milad Shoryabi, Ahmad Hajipour, Afshin Shoeibi, and Ali Foroutannia. Recognition of anxiety and depression using gait data recorded by the kinect sensor: a machine learning approach with data augmentation. *Scientific Reports*, 15(1), 2025.

[12] Felix Menne, Felix Dörr, Julia Schräder, Johannes Tröger, Ute Habel, Alexandra König, and Lisa Wagels. The voice of depression: speech features as biomarkers for major depressive disorder. *BMC Psychiatry*, 24(1), 2024.

[13] Thalia Richter, Shahar Stahi, Gal Mirovsky, Hagit Hel-Or, and Hadas Okon-Singer. Disorder-specific versus transdiagnostic cognitive mechanisms in anxiety and depression: Machine-learning-based prediction of symptom severity. *Journal of Affective Disorders*, 354:473 – 482, 2024.

[14] Giordano D'Urso, Alfonso Magliacano, Sayna Rotbei, Felice Iasevoli, Andrea de Bartolomeis, and Alessio Botta. Predicting the severity of lockdown-induced psychiatric symptoms with machine learning. *Diagnostics*, 12(4):957, 2022.

[15] Wayne K Goodman, Lawrence H Price, Steven A Rasmussen, Carolyn Mazure, Roberta L Fleischmann, Candy L Hill, George R Heninger, and Dennis S Charney. The Yale-Brown obsessive compulsive scale: I. development, use, and reliability. *Archives of general psychiatry*, 46(11):1006–1011, 1989.

[16] Jane L Eisen, Katharine A Phillips, Lee Baer, Douglas A Beer, Katherine D Atala, and Steven A Rasmussen. The brown assessment of beliefs scale: reliability and validity. *American Journal of Psychiatry*, 155(1):102–108, 1998.

[17] AT Beck, RA Steer, GK Brown, et al. Beck Depression Inventory. San Antonio. *TX: The Psychological Corporation*, 1996.

[18] CD Spielberger. State-trait anxiety inventory for adults (STAI-AD)[database record]. apa psyctests, 1983.

[19] Mauro Pettorruso, Roberto Guidotti, Giacomo d'Andrea, Luisa De Risio, Antea D'Andrea, Stefania Chiappini, Rosalba Carullo, Stefano Barlati, Raffaella Zanardi, Gianluca Rosso, et al. Predicting outcome with intranasal esketamine treatment: A machine-learning, three-month study in treatment-resistant depression (esk-learning). *Psychiatry Research*, 327:115378, 2023.