



Mental health risk prediction in autoimmune patients in primary care: A class-sensitive machine learning approach

Mariachiara Di Cosmo ^a, Sara Campanella ^b ^{*}, Michele Bernardini ^c,
Adriano Mancini ^b, Lorenzo Palma ^b

^a BioRobotics Institute, Sant'Anna School of Advanced Studies, Viale Rinaldo Piaggio 34, Pontedera, 56025, Italy

^b Università Politecnica delle Marche, Information Engineering Department, Via Brecce Bianche 12, Ancona, 60131, Italy

^c Department of Theoretical and Applied Sciences, eCampus University, Via Isimbardi 10, Noverate, 22060, Italy

ARTICLE INFO

Keywords:

Autoimmune diseases
Decision support system
Electronic health record
Machine learning
Mental disorders
Primary care

ABSTRACT

Patients with autoimmune diseases are at increased risk of developing psychological conditions such as anxiety and depression. Timely detection of these mental health disorders is essential for effective intervention, yet remains difficult due to the overlap and subtle presentation of symptoms. In this study, we propose a machine learning (ML) framework for early prediction of anxiety and depression in autoimmune patients using routine clinical and demographic features in primary care. We develop an XGBoost-based model enhanced with custom loss functions specifically designed to handle class imbalance and semantic similarity between mental health outcomes. Experimental results demonstrate that the proposed penalty terms improve standard XGBoost performance in identifying minority classes and minimizing confusion between similar categories, by reaching a maximum statistically significant macro-Recall gain of 4.03% ($p < 0.05$) for the Logit Distance Exponential Penalty term. This work contributes to developing data-driven tools for mental health risk assessment, with potential applications in personalized care and digital screening for vulnerable clinical populations.

1. Introduction

For over a century, evidence has pointed to an epidemiological connection between autoimmune and mental health disorders (Benros et al., 2014). Recent studies confirm this association, showing that depression is more prevalent among individuals with autoimmune disorders (18.6%) rather than in the general population (10.5%) (Rakshasa-Loots et al., 2025). Many pathophysiological pathways appear to be involved in both autoimmune and anxiety-related disorders. Hormonal imbalances and neurotransmitter dysregulation are common features of anxiety and depression, which heighten feelings of dread and threat. A central player in this interaction appears to be the hypothalamic-pituitary-adrenal (HPA) axis, which regulates the body's stress response and links psychological stress to immunological dysfunction. Furthermore, oxidative stress, mitochondrial dysfunction, and pro-inflammatory cytokines have been identified as key contributors to symptomatic overlap between autoimmune and mental health disorders (Sedda et al., 2025). Beyond their clinical manifestations, exploring the connection between autoimmune diseases and depression/anxiety, and clarifying the mechanisms involved, could also support the hypothesis that immunological processes may play a causal role in the aetiology of depression (Raison et al., 2006): for instance, depressive symptoms in patients with rheumatoid arthritis may be associated with disease exacerbation (Euesden et al., 2017). Moreover, depression is frequently comorbid with anxiety and often

* Corresponding author.

E-mail address: s.campanella@pm.univpm.it (S. Campanella).

<https://doi.org/10.1016/j.smhl.2026.100634>

Received 9 September 2025; Received in revised form 30 November 2025; Accepted 10 January 2026

Available online 16 January 2026

2352-6483/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

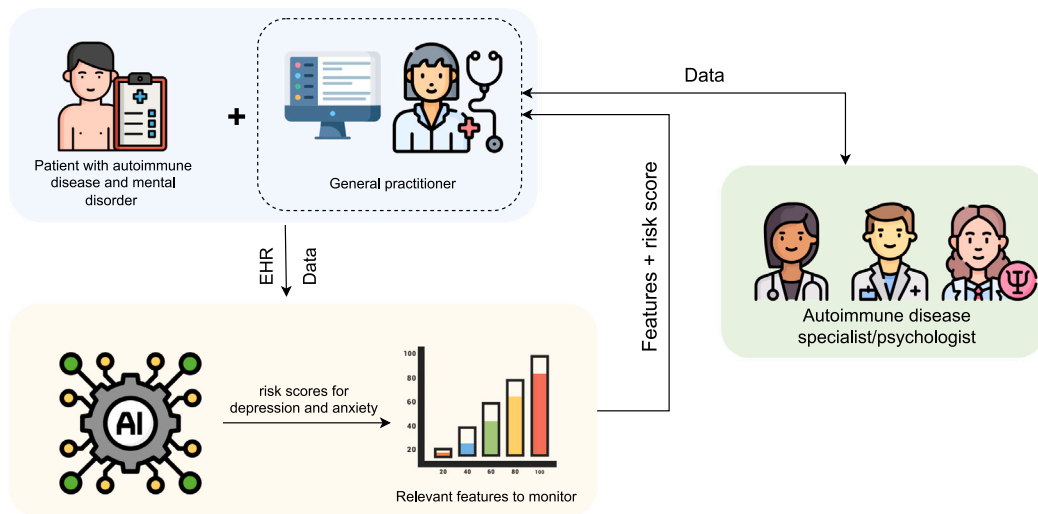


Fig. 1. Integration of the proposed ML framework into GPs' EHRs. The system ingests routine clinical and demographic data and returns class-specific risk scores for depression and anxiety-related disorders, supporting nuanced interpretation and early referral decisions.

misdiagnosed due to symptom overlap. Shared symptoms, such as mood swings or cognitive deficits, can be found in both mental health disorders and complicate accurate diagnosis (Wu et al., 2023). This highlights the importance of early identification strategies that can facilitate timely referral to appropriate care, especially in autoimmune patients. Recent advances in machine learning (ML) offer powerful tools to identify and track key features that connect mental health and autoimmune diseases (Nickson et al., 2023). Notably, ML techniques are well-suited to analyze high-dimensional, heterogeneous data to find predictive biomarkers and intricate patterns (Rajkomar et al., 2019). However, most research on the onset of depression and anxiety in autoimmune patients relies on self-reported questionnaires. Although easy to use, questionnaires lack clinical accuracy and temporal stability. Electronic health records (EHRs), instead, might be a more objective and reliable data source. Nevertheless, its use is limited by challenges, such as inconsistent documentation, case-ascertainment errors, system heterogeneity, missing or unstructured information, and stringent privacy constraints (Guo et al., 2025; Roehrs et al., 2019). Moreover, even though EHR data may provide more clinically sound insights into the beginning of depression, current ML studies often focus on a single autoimmune condition, limiting the generalizability to a broader patient population (Forrest et al., 2023; Maneeton et al., 2013; Yang et al., 2025). Such a narrow approach prevents general practitioners (GPs) from gaining a comprehensive understanding of a patient's overall health status, especially when multiple overlapping autoimmune conditions may contribute to the development of mental health disorders.

Building on these insights, we propose an ML framework to screen for mental health disorders – specifically depression and anxiety – in patients with autoimmune diseases, using clinical and demographic data from GPs' EHRs. Focusing on primary care data offers a novel and pragmatic approach, as GPs are often the first to manage complex multimorbidity. By leveraging routinely collected, structured information in GPs' settings, we aim to provide GPs with risk ratings for anxiety and depression, along with key features to monitor over time. Fig. 1 illustrates the proposed clinical workflow: EHR data from patients with a diagnosed autoimmune disease are used as input to the ML model to provide risk ratings for anxiety and depression and identify relevant features. These insights can be accessed by GPs, who could also share information with specialists such as psychologists and autoimmune disease experts. This process could provide support for the identification of mental health risk and timely referral to specialists, thereby improving the quality of care for autoimmune patients, a vulnerable and often under-recognized population. The main contributions of this work can be summarized as:

- We investigate an innovative and unexplored clinical scenario in primary care, assessing how clinical and demographic data collected by GPs can be used to screen for depression and anxiety within a broad patient population with autoimmune diseases.
- We propose a ML framework integration into the GPs' EHRs, providing feedback on patient mental health risk and supporting early referral to the appropriate specialist.
- We benchmark state-of-the-art ML models and also enhance the XGBoost model performance by introducing a set of custom penalty functions designed to improve discrimination between overlapping and minority classes (particularly, depression and anxiety), thus reducing prediction ambiguity.

2. Related works

Various studies have demonstrated the potential of ML to predict anxiety and depression across diverse populations (Baran & Cetin, 2025; Islam et al., 2024; Khoo et al., 2024). However, relatively few works leverage primary care EHRs from GPs, and

even fewer address mental health risk prediction specifically within autoimmune patient cohorts. In [Yang et al. \(2025\)](#), a CatBoost classifier is trained to recognize different levels of depression in 480 patients with connective tissue diseases, using comprehensive clinical data (comorbidities, labs, medications). The work ([Tennenhouse et al., 2020](#)) applied ML to patient-reported outcome measures in individuals with immune-mediated inflammatory diseases to predict major depressive disorder and anxiety disorders. However, most of these studies focus on either anxiety or depression ([Upadhyay et al., 2024](#)), and typically within narrowly defined autoimmune subgroups, rather than considering the full spectrum of autoimmune patients.

Moreover, in existing literature, the definitions of depression and anxiety vary widely, often relying on self-reported surveys, standardized questionnaires, clinical data, social media content, or lifestyle-related indicators ([Ahmed et al., 2022](#); [Upadhyay et al., 2024](#)). For instance, [Chahar et al. \(2024\)](#) proposed an XGBoost-Hippopotamus optimization algorithm to classify in a general population different levels of depression, anxiety and stress, defined based on PHQ-9 and GAD-7 questionnaire scores, from social and demographic features (e.g., age, gender, education). Even with improvements in recall through hybrid optimization techniques, they did not use clinical diagnoses or longitudinal EHR data.

Although these approaches showcase strong performance in their respective cohorts, they rarely generalize to real-world EHR settings or patient groups with heterogeneous autoimmune conditions. Moreover, standard ML objective functions may underperform in discriminating semantically adjacent and minority classes such as anxiety versus depression, given their high clinical overlap. In other domains, several approaches attempt to mitigate this by modifying model objectives in ML training. For example, [Mushava and Murray \(2024\)](#) explored a range of custom penalty functions tailored to reduce misclassification risk in an industrial context. However, none so far have tailored these approaches explicitly to structured primary care data in autoimmune patients, nor integrated with interpretable feature importance for clinical use.

In this work, we focus on routinely collected GP EHRs for a broad, heterogeneous autoimmune patient population. Moreover, we introduce custom penalty terms in the loss functions within XGBoost to reduce confusion between the closely related and underrepresented classes of depression and anxiety.

3. Materials and methods

We address a three-class classification task aimed at identifying mental health risks, specifically depression and anxiety-related disorders, in patients with autoimmune diseases, based on real-world data from general practice. The classes are defined as 0 (control), 1 (depression), and 2 (anxiety). Given the semantic and clinical similarity that could occur between depression and anxiety, our focus is on enhancing model's ability to distinguish between these classes by introducing custom penalty functions that explicitly discourage their confusion.

In Section 3.1, we describe the dataset, patient inclusion criteria, and preprocessing steps applied to extract meaningful predictors from primary care EHRs. Section 3.2 presents our classification model, based on XGBoost, along with the proposed penalty terms integrated into its objective function. Then, to assess the added value of our approach, we compare its performance to a wide range of baseline ML models, as reported in Section 3.3. Finally, we report details on the experimental setup, including training design, hyperparameter tuning, and evaluation metrics in Section 3.4.

3.1. Dataset and preprocessing

In this study, we used a real-world anonymized dataset, referred to as the FIMMG-6GP dataset ([Bernardini et al., 2021](#)), which includes structured EHR data collected from six GPs across a 10-year clinical history (2010–2019). The FIMMG-6GP dataset comprises a total of 14175 patients, with each GP contributing a portion of the total cohort. The FIMMG-6GP dataset, spanning both categorical and numerical variables, is structured in several fields, such as demographics, lifestyle, allergies, vaccinations, monitoring, exemptions, pathologies, exam prescriptions, lab tests, drugs, and hospitalization. To align with our study objectives and target population, the FIMMG-6GP dataset underwent a comprehensive preprocessing phase encompassing both patient selection and predictor extraction.

3.1.1. Patients selection

From the FIMMG-6GP dataset, we derived a case study focusing on subjects with autoimmune diseases. We extracted 925 patients diagnosed with various autoimmune diseases, identified based on their International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes in the pathologies field.

This specific cohort of patients is referred to as the AutoImmune-FIMMG-6GP dataset (AI-FIMMG-6GP dataset). The distribution of autoimmune diseases stored in the AI-FIMMG-6GP dataset is shown in [Table 1](#). For patients with multiple autoimmune diagnoses, for each patient we selected the earliest recorded diagnosis and corresponding pathology as the reference for temporal analyses and modeling. This approach ensures consistency in defining the starting point of the clinical history related to autoimmune conditions.

From the 925 autoimmune patients, we selected 299 patients who had been diagnosed with mental disorders based on ICD-9-CM codes recorded in the pathologies field. Among them, 187 patients were diagnosed with depression and 112 with anxiety. Patients diagnosed with both conditions were assigned to the depression class, due to its greater clinical severity. The remaining 626 patients without any mental disorder diagnosis are referred to as the control class. The statistics of the resulting patients of the AI-FIMMG-6GP dataset are shown in [Table 2](#).

Table 1

Distribution of autoimmune diseases in the AI-FIMMG-6GP cohort (n = 925). Patients may present multiple autoimmune diagnoses; for consistency, the first recorded diagnosis per patient was used as the reference date for this analysis.

Category (#Total, Perc%)	ICD-9-CM description	#
Systemic (#82, 8.86%)	Systemic lupus erythematosus	7
	Rheumatoid arthritis	54
	Polymyositis and dermatomyositis	5
	Systemic sclerosis	5
	Sjögren's syndrome	11
Hematologic (#22, 2.38%)	Schoenlein-Henoch purpura	8
	Pernicious anemia	4
	Kawasaki disease	1
	Horton's disease	4
	Churg-Strauss syndrome	2
Gastrointestinal (#94, 10.16%)	Ulcerative colitis	9
	Crohn's disease	60
	Celiac disease	20
	Primary sclerosing cholangitis	5
Skin (#249, 26.92%)	Psoriasis	194
	Vitiligo	10
	Dermatitis herpetiformis	3
	Pemphigus	6
	Alopecia areata	35
Endocrine (#427, 46.16%)	Type 1 diabetes	36
	Hashimoto's thyroiditis	338
	Autoimmune hypoparathyroidism	4
	Graves' disease (Basedow)	66
	Addison's disease	1
Neurological (#38, 4.11%)	Multiple sclerosis	24
	Devic's disease	1
	Myasthenia gravis	12
	Guillain-Barré syndrome	1
Musculoskeletal (#31, 3.35%)	Psoriatic arthropathy	26
	Ankylosing spondylitis	5
Other (#53, 5.73%)	Endometriosis	22
	Fibromyalgia	31

Table 2

Distribution of patients by GP and class labels (absolute values and percentages). Control: class 0; Depression: class 1; Anxiety: class 2.

GPs	Total	Control [0]	Depression [1]	Anxiety [2]
GP1	119	79 (66.4%)	27 (22.7%)	13 (10.9%)
GP2	155	100 (64.5%)	36 (23.2%)	19 (12.3%)
GP3	158	82 (51.9%)	42 (26.6%)	34 (21.5%)
GP4	173	128 (74.0%)	26 (15.0%)	19 (11.0%)
GP5	156	126 (80.8%)	22 (14.1%)	8 (5.1%)
GP6	164	111 (67.7%)	34 (20.7%)	19 (11.6%)
Total	925	626 (67.7%)	187 (20.2%)	112 (12.1%)

3.1.2. Predictors selection

To predict the onset of mental disorders in autoimmune patients, we defined the observational time window of interest as ranging from the beginning of the patient's clinical history until: i) the date of the first mental disorder diagnosis for the depression and anxiety classes, and (ii) the end of the clinical history for the control class. Once delimited, the observational time window of interest, categorical predictors were one-hot encoded (Rodríguez et al., 2018) and numerical predictors were aggregated by their average value within the time window. We selected a subset of features, retaining only those categorical with less than 98% zero values and those numerical with less than 95% missing values to reduce sparsity and discard features with limited informative value. ICD-9-CM codes utilized for labeling the mental disorders classes, or, in agreement with the GPs' consultation, Anatomical Therapeutic Chemical (ATC) codes of drugs field directly correlated with mental disorder labels, were excluded from the predictors' set. Missing values in numerical fields were replaced with a distinct out-of-range placeholder value (i.e., -999), to retain information about missingness while ensuring compatibility with models that do not handle missing values natively.

This filtering step reduced the final feature space to 239 predictors. The statistics of the resulting predictors of the AI-FIMMG-6GP dataset are shown in Table 3.

Table 3

Statistics of the 239 predictors grouped per field and their averaged missingness. Only for categorical features, missingness is referred as sparsity, defined as the percentage of zeros.

Field	#Predictors	Type	Missingness%
Demographics	3	Misc	
-Age: 56.86 ± 17.97 years	1	Num	0.00
-Gender: Males = 30.49%	1	Cat	0.00
-Education	1	Cat	88.32
Lifestyle (Smoking)	1	Cat	83.24
Vaccination	1	Cat	91.13
Hospitalization	1	Cat	87.57
Pathologies	34	Cat	97.75
Exams	89	Cat	95.63
Drugs	45	Cat	97.98
Exemptions	24	Cat	95.09
Monitoring	6	Num	
-Diastolic: mean, min, max	3	Num	77.30
-Systolic: mean, min, max	3	Num	77.30
Lab tests	35	Num	74.40

3.2. Model design and custom loss formulation

To tackle our three-class classification task, we adopted a gradient-boosted decision tree model (XGBoost). XGBoost is chosen for its robustness to missing data, ability to model non-linear relationships, and effectiveness in handling class imbalance, properties well-suited to real-world sparse EHRs (Stokanović et al., 2024).

To improve the model's ability to discriminate between the two minority and semantically similar classes - i.e., depression (1) and anxiety (2) - we modified the standard multi-class softmax objective by introducing a set of penalty terms $\mathcal{P}$, implemented based on Mushava and Murray (2024), Wang et al. (2020) and Joyce (2011), aimed at reducing prediction ambiguity by acting directly on the optimization process. Specifically, they adjust the gradients and, in some cases, the Hessians during training to encourage the model to better capture class-specific patterns. Each penalty is weighted by a scalar hyperparameter λ , which controls its relative contribution to the total loss: higher λ values increase the influence of the penalty, allowing the model to focus more aggressively on reducing confusion between the two classes. Additionally, each penalty may involve further hyperparameters, such as thresholds or margins, which are optimized during training.

Let $\mathbf{z}_i = (z_{i,0}, z_{i,1}, z_{i,2})$ be the unnormalized logits for sample i , and $p_{i,j} = \text{softmax}(\mathbf{z}_i)_j$ is predicted probability of i for class j . For each instance and class, the gradient and Hessian are computed as (Mushava & Murray, 2024):

$$g_{i,j}^{(0)} = w_{i,j} \cdot (p_{i,j} - \mathbb{1}_{j=y_i}), \quad h_{i,j}^{(0)} = w_{i,j} \cdot p_{i,j}(1 - p_{i,j}) \quad (1)$$

where $\mathbb{1}$ is the indicator function and $w_{i,j}$ are class-specific weights applied to account for class imbalance. Moreover, for all samples with class $y_i \in \{1, 2\}$, a penalty $\mathcal{P}$ contributes extra terms $\Delta g_{i,j}, \Delta h_{i,j}$ only to classes 1 and 2. The final gradients and Hessians are adjusted as:

$$g_{i,j} = g_{i,j}^{(0)} + \Delta g_{i,j}, \quad h_{i,j} = h_{i,j}^{(0)} + \Delta h_{i,j} \quad (2)$$

In the following, we describe each proposed penalty term. Unless otherwise specified, all penalties contribute to both $g_{i,j}^{(0)}$ and $h_{i,j}^{(0)}$ during training.

Class-Pair Similarity Penalty (CPS). It discourages simultaneous high softmax probabilities p for classes 1 and 2, reducing ambiguity in prediction.

$$\mathcal{P}_{\text{CPS}}(p) = \lambda \cdot p_1 \cdot p_2 \quad (3)$$

High-Confusion Threshold Penalty (HCT). It activates only when the combined probability exceeds a threshold τ , targeting high-confidence confusion cases.

$$\mathcal{P}_{\text{HCT}}(p) = \begin{cases} \lambda \cdot (p_1 + p_2 - \tau), & p_1 + p_2 > \tau, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Asymmetric Confidence Separation (ACS). It encourages the model to prefer one class over the other when probabilities are nearly equal. It contributes to $g_{i,j}^{(0)}$ only, with no second-order term.

$$\mathcal{P}_{\text{ACS}}(p) = \lambda \cdot (-|p_1 - p_2|) \quad (5)$$

Triplet-Style Class-Center Margin Loss (TCM). It applies a margin m separation between the true class and its most confusing counterpart. When the margin is violated, it contributes to $g_{i,j}^{(0)}$ only.

$$P_{TCM}(p, y) = \lambda \cdot \max(0, p_{\text{wrong}} - p_{\text{true}} + m) \quad (6)$$

Focal-Weighted Confusion Contrast (FWCC). It extends the class-pair penalty with a focal term that down-weights confident predictions.

$$P_{FWCC}(p) = \lambda \cdot p_1 \cdot p_2 \cdot (1 - \max(p_1, p_2))^\gamma \quad (7)$$

Local Entropy Minimization (LEM). It penalizes high entropy H over classes 1 and 2, affecting uncertain distributions and promoting confident predictions.

$$P_{LEM}(p) = \lambda \cdot H(p_1, p_2) \quad (8)$$

Kullback–Leibler Uniformity Penalty (KL-UP). It penalizes predictions close to a uniform distribution over classes 1 and 2.

$$P_{KL-UP}(p) = \lambda \cdot D_{KL}([p_1, p_2] \parallel [0.5, 0.5]) \quad (9)$$

Selective Soft-Confidence (SSC). It applies a contrastive penalty only when confidence in the true class p_y is below a threshold β .

$$P_{SSC}(p, y) = \begin{cases} \lambda \cdot p_1 \cdot p_2, & p_y < \beta, \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

Logit Distance Exponential Penalty (LDEP). It operates directly in the logit space, penalizing small logit differences between classes 1 and 2 (z_1 and z_2).

$$P_{LDEP}(z) = \lambda \cdot \exp(-|z_1 - z_2|) \quad (11)$$

3.3. Baseline models for comparison

To broadly assess the validity of our method, we compared its performance with an exhaustive selection of baseline models representative of the current state of the art in predictive modeling for structured, unimodal, medium-to-small real-world EHR datasets (Francesconi et al., 2025; Hollmann et al., 2025). Recent evidence shows that, in this setting, tree-based ensemble methods consistently outperform more complex deep learning architectures, while transformer-based models are generally unnecessary and computationally inefficient (Shwartz-Ziv & Armon, 2022). For this reason, our baseline selection focuses on the model families most commonly reported as competitive for tabular medical data, complemented by a set of lightweight neural architectures. Specifically, we benchmark our approach against the following classifier families:

- **Naive Bayes (NB):** Gaussian and Bernoulli NB models;
- **Nearest Neighbors:** k-Nearest Neighbors (KNN) and Nearest Centroid (NC);
- **Linear models:** Logistic Regression (LR), Ridge classifier (Ridge), Support Vector Classification (SVC) and Nu-Support Vector Classification (NuSVC);
- **Tree-based models:** Decision Tree (DT), Random Forest (RF), AdaBoost, LightGBM, CatBoost;
- **Neural Networks:** Multilayer Perceptron (MLP), 1D- Convolutional Neural Network (1D-CNN) and Tabular Prior- data Fitted Network (TabPFN) (Hollmann et al., 2025).

3.4. Experimental details

We adopted a stratified 5-fold Cross-Validation (5FCV), with a nested 3-fold CV for hyperparameter optimization. For XGBoost, we performed grid search over the following hyperparameters: learning rate (lr) in [0.05, 0.01, 0.1, 0.5], max_depth in [3, 5, 7, 10, 15, 25], number of estimators ($n_estimators$) in [50, 75, 100, 200, 300], and subsample fixed at 1.0. For the proposed penalized losses, we tune the penalty weight λ in [2, 5, 10, 20], the margin m in range [0.1, 0.9], the soft-confidence threshold β in [0.3, 0.8], the exponential coefficient γ in [1.0, 2.0] and the confusion threshold τ in [0.1, 0.9].

For state-of-the-art ML models, we tuned the following hyperparameters: $n_neighbors$ in [3, 5, 7, 9] for KNN; inverse of regularization strength (C) in [0.1, 1.0, 10] for LR; regularization strength ($alpha$) in [1.0, 2.0, 5.0] for Ridge; $kernel$ in ['linear', 'rbf'] for SVC and NuSVC, while C in [0.1, 1.0, 10] only for SVC and nu in [0.1, 0.3, 0.5] only for NuSVC; max_depth in [5, 15, 25], $n_estimators$ in [50, 75, 100], lr in [0.05, 0.1, 0.5] for all Tree-based ensemble models, while max_depth in [5, 15, 25] only for the DT classifier; $hidden_layers$ in [50, 75, 100] and L2 regularization term ($alpha$) in [0.0001, 0.001, 0.01] for MLP; filters = 16, kernel_size = 3, activation function = 'relu', units = 32, optimizer = 'adam', loss = 'sparse_categorical_crossentropy', epochs = 50, batch_size = 32 for 1D-CNN.

Table 4

Average and per-class performances of different classification models divided by families: Naive Bayes (NB), Nearest Neighbors, Linear models, Tree-based models and Artificial Neural Networks. All metrics are expressed as percentages. Recall_0, Recall_1, and Recall_2 are referred per class, while Recall_1_2 denotes the macro-average recall over classes 1 and 2.

Model	Precision	Recall	F1-score	Recall_0	Recall_1	Recall_2	Recall_1_2
Gaussian NB	66.63	66.63	66.58	100.00	58.82	41.07	49.95
Bernoulli NB	64.68	64.55	64.34	100.00	50.80	42.86	46.83
KNN	65.96	65.68	65.30	100.00	65.77	31.25	48.51
NC	67.07	60.56	58.57	100.00	20.83	60.87	40.85
LR	65.03	64.28	64.64	100.00	69.99	22.85	46.42
Ridge	64.98	64.84	64.92	100.00	76.61	17.91	47.26
SVC	60.02	64.79	59.89	100.00	89.87	4.51	47.19
NuSVC	62.76	62.94	62.81	100.00	62.03	26.79	44.41
DT	65.62	65.68	65.58	100.00	65.78	31.25	48.52
RF	67.96	67.69	67.05	100.00	76.27	26.79	51.53
AdaBoost	64.95	67.32	64.90	100.00	75.95	26.01	50.98
LightGBM	67.60	67.71	67.65	100.00	70.15	32.99	51.57
CatBoost	68.00	67.37	67.51	100.00	67.33	34.78	51.05
XGBoost	67.46	67.41	67.17	100.00	68.45	33.93	51.11
MLP	65.45	64.36	64.84	100.00	66.36	26.72	46.54
1D CNN	67.00	65.82	66.13	100.00	69.99	27.47	48.73
TabPFN	65.89	64.11	64.99	100.00	72.23	20.11	46.17

All models were trained and evaluated under the same configurations to ensure comparability. During training, we maximized the average recall_macro over classes 1 and 2 (i.e., Recall_1_2), defined as the average recall across these two clinically relevant classes.

To evaluate model performance, we calculated the macro-averaged F1-score (F1-score), macro-averaged precision (Precision), macro-averaged recall (Recall), Recall_1_2, and confusion matrices (CMs). Moreover, we examined the statistical significance of XGBoost with each custom penalty loss against the baseline with softmax cross-entropy loss using the Paired T-test (p -value < 0.05). Given the 5FCV test performance, we calculated the win rate (W), representing the instances where the competitor's p -value outperformed the baseline. We also measured the tie rate (T) and loss rate (L), indicating the percentage of times that the competitor was equal to or worse than the baseline, respectively.

Furthermore, to evaluate the robustness of the proposed methodology, we performed a sensitive analysis, considering two complementary aspects: the model generalization across clinical practice and the model stability and interpretability in terms of feature importance. To assess model's ability to generalize across clinical practices, we performed a leave-one-GP-out (L1GPO) validation, iteratively training on all but one GP and testing on the held-out GP, in place of the 5FCV strategy. To evaluate feature importance, instead, we used both the XGBoost Gain strategy (Chen & Guestrin, 2016), which provides a global measure of each feature's contribution to model performance, and SHapley Additive exPlanations (SHAP), a model-agnostic method offering instance-level attributions. Using both approaches allows a more robust and method-invariant evaluation of the most influential predictors. By examining these distinct aspects, we can have a more robust and method-invariant assessment of both predictive performance and key feature relevance.

4. Results

Table 4 reports the performance of several ML models for the prediction of depression and anxiety onset in autoimmune patients. All ML models fully discriminate between control and mental disorder patients (i.e., Recall_0 = 100%). While some linear models achieved moderate overall results (e.g., LR: Recall = 64.28%, Ridge: Recall = 64.84%), they struggled to detect at least one of the mental disorder classes (e.g., LR: Recall_2 = 22.85%; Ridge: Recall_1 = 17.91%). On the contrary, the tree-based ensemble family consistently outperformed other model families in terms of Precision, Recall, and F1-score, and achieved the highest Recall_1_2 values ranging between 50.98% and 51.57%, thus demonstrating a better balance in detecting both mental disorder classes. These findings confirm XGBoost's robustness in handling structured clinical data with class imbalance, supporting its selection as the baseline model for further class-sensitive experiments.

Table 5 shows the average performance metrics for each penalized XGBoost variant. Among them, the model trained with the LDEP term achieved the best results, with an improved overall score and yielding the highest Recall_1_2 = 55.14%, with a substantial increase for standard XGBoost loss, in particular for true positives in class 2 ($\Delta TP_2 = +12$). These findings indicate a more balanced classification performance across all classes.

Table 6 shows the statistical comparison of the results in Table 5, with W-T-L indicating win, tie, and loss rates for each custom loss against standard XGBoost. These results confirmed LDEP integration into XGBoost training as the most effective approach for class-sensitive improvement.

Table 5

Average performance metrics [%] and confusion matrix counts of XGBoost models trained with different penalty loss terms. Best values are highlighted in bold. For the confusion matrix: TP_i is the number of correctly classified samples of class i and $CM_{i,j}$ the number of instances of true class i predicted as class j . Abbreviations in the table: Standard = Multi-class softmax cross-entropy, CPS = Class-Pair Similarity Penalty, HCT = High-Confusion Threshold Penalty, ACS = Asymmetric Confidence Separation, TCM = Triplet-Style Class-Center Margin Loss, FWCC = Focal-Weighted Confusion Contrast, LEM = Local Entropy Minimization, KL-UP = KL Uniformity Penalty, SSC = Selective Soft-Confidence, LDEP = Logit Distance Exponential Penalty; λ = penalty weight, m = margin, β = soft-confidence threshold, γ = exponential coefficient, τ = confusion threshold.

Class-Sensitive XGBoost	Precision	Recall	Recall_1_2	F1-Score	TP_0	CM_0_1	CM_0_2	CM_1_0	TP_1	CM_1_2	CM_2_0	CM_2_1	TP_2
Standard	67.46	67.41	51.11	67.17	626	0	0	1	128	58	2	72	38
CPS ($\lambda = 10$)	69.61	69.62	54.43	69.31	626	0	0	1	110	76	1	55	56
HCT ($\lambda = 20$, $\tau = 0.9$)	67.79	67.58	51.37	67.61	626	0	0	1	112	74	2	62	48
ACS ($\lambda = 5$)	64.38	63.97	45.96	63.88	626	0	0	1	102	84	2	68	42
TCM ($\lambda = 5$, $m = 0.2$)	69.68	69.53	54.29	69.41	626	0	0	1	123	63	2	62	48
FWCC ($\lambda = 2$, $\gamma = 2$)	69.47	69.33	53.99	69.29	626	0	0	1	120	66	2	61	49
LEM ($\lambda = 5$)	69.00	69.13	53.70	68.67	626	0	0	0	104	83	0	54	58
KL-UP ($\lambda = 2$)	65.39	65.10	47.65	64.99	626	0	0	0	105	82	1	67	44
SSC ($\lambda = 10$, $\beta = 0.6$)	68.26	68.18	53.15	67.94	626	0	0	1	117	69	2	63	47
LDEP ($\lambda = 5$)	70.41	70.10	55.14	70.09	626	0	0	0	123	64	1	61	50

Table 6

Comparison of XGBoost models trained with different penalty loss terms and XGBoost multi-class softmax cross-entropy loss (Standard) in terms of Recall_1_2 (average and standard deviation [%]), p-value, and Win-Tie-Loss (W-T-L) rates [%].

Class-Sensitive XGBoost	Recall_1_2	p_value	W-T-L
Standard	51.11 $\pm$ 7.74	–	–
CPS	54.43 $\pm$ 4.54	0.1159	60-0-40
HCT	51.37 $\pm$ 7.75	0.1777	60-20-20
ACS	45.96 $\pm$ 6.22	0.1947	40-0-60
TCM	54.29 $\pm$ 10.19	0.0870	80-0-20
FWCC	53.99 $\pm$ 8.82	0.0892	100-0-0
LEM	53.70 $\pm$ 1.07	0.0982	60-0-40
KL-UP	47.65 $\pm$ 4.94	0.3595	20-0-80
SSC	53.15 $\pm$ 5.40	0.2219	60-0-40
LDEP	55.14 $\pm$ 4.54	0.0374 ^a	80-0-20

^a Class-Sensitive XGBoost with p-value < 0.05.

Table 7

Average performance metrics [%] and confusion matrix counts of the XGBoost model trained with LDEP loss term using the Leave-one-GP-out (L1GPO) validation experimental procedure. Best values are highlighted in bold.

GP	Precision	Recall	Recall_1_2	F1-Score	TP_0	CM_0_1	CM_0_2	CM_1_0	TP_1	CM_1_2	CM_2_0	CM_2_1	TP_2
1	65.00	64.77	47.15	63.89	79	0	0	0	13	14	0	7	6
2	68.14	67.83	51.75	67.98	100	0	0	0	24	12	1	11	7
3	75.09	74.97	62.46	75.01	82	0	0	0	29	13	0	15	19
4	67.90	67.88	51.82	67.87	128	0	0	0	16	10	0	11	8
5	69.13	67.80	51.70	66.60	126	0	0	0	20	2	0	7	1
6	66.39	66.41	49.61	66.35	111	0	0	0	23	11	0	13	6
AVG	68.61	68.28	52.42	67.95	626	0	0	0	125	62	1	64	47

4.1. Cross-practice evaluation

Table 7 reports the performance of the XGBoost model with LDEP loss term under the L1GPO validation. Compared with the 5FCV, the average performance shows a slight decrease, with variability across GPs reflecting genuine heterogeneity in patient populations. Using the robust modified Z-score method (Iglewicz & Hoaglin, 1993), Recall_1_2 for GP3 is identified as a significant outlier compared with other GPs, indicating that it performs differently from the rest of the cohort. Notably, GP3 also exhibits the highest overall performance metrics and has the most balanced class distribution (51.9% controls, 26.6% depression, 21.5% anxiety; see Table 2). This may suggest that the variability in performance across GPs could be influenced by the class composition of each test set. Alternatively, GP3's higher performance may simply reflect better-quality clinical data collection. Overall, the model demonstrates consistent predictive behavior, with performance naturally depending on the representation of classes in the test set.

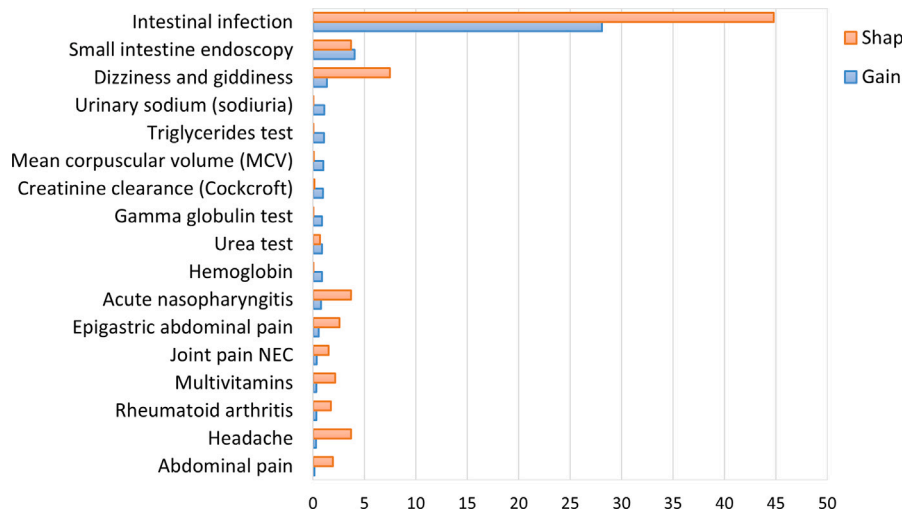


Fig. 2. Top 10 features ranked by XGBoost gain and mean absolute SHAP values for the XGBoost model with LDEP term (best-performing). Both metrics are normalized and expressed as percentages.

4.2. Feature importance analysis

Fig. 2 reports the top-10 ranked by XGBoost gain-based feature importance and by SHAP, both normalized and expressed in percentages for the XGBoost model trained with LDEP loss term. Across both methods, intestinal infection consistently emerges as the dominant predictor, with SHAP assigning an even greater relative contribution than gain. All remaining features (e.g., small intestine endoscopy, dizziness and giddiness, urinary sodium, triglycerides, and hematological indices) show substantially lower and broadly comparable importances. SHAP places slightly more emphasis on symptom-based features such as headache and abdominal pain, but the overall hierarchy remains highly consistent across the two approaches, confirming the strong concentration of model signal in gastrointestinal features. We emphasize that this analysis is exploratory and intended to assess model behavior, rather than to draw clinical conclusions.

5. Discussion

This section examines the results from both clinical and implementation perspectives. Clinically, we evaluate the implications of the model's predictive performance and potential integration into clinical care. From the technical standpoint, we assess model robustness and interpretability considering issues such as data imbalance, generalizability and scalability.

5.1. Clinical impact

Analyzing the features that most influence the onset of mental disorders is essential for prevention and more accurate and personalized monitoring (see Fig. 2). The intestinal infection ICD-9-CM code was the most significant feature. This implies that autoimmune diseases, anxiety, or depression are strongly correlated with gastrointestinal infections. This discovery is consistent with emerging data about the gut-brain axis and the possible involvement of intestinal inflammation in immunological and neuropsychiatric dysregulation (Doenya et al., 2025; Suganya & Koo, 2020). Although gastrointestinal autoimmune illnesses are only the third most prevalent diagnostic category (see Table 1), these features stand out as significant in the model. This implies that gastrointestinal issues are similarly distributed among all autoimmune conditions and might affect a patient's general health, even if they are not the most common.

The second most relevant feature was endoscopy of the small intestine, a diagnostic exam frequently prescribed to look into persistent gastrointestinal symptoms. Its significance could be a reflection of the frequency of gastrointestinal problems in this patient group or of the efforts made by doctors to exclude pathological illnesses such as chronic gastritis or celiac disease. Given the very high importance attributed to the gastrointestinal feature pointed out from Fig. 2, we explored whether this effect might partially reflect class imbalance rather than a genuine clinical mechanism. To assess this, we conducted a sensitivity analysis by re-estimating the models after removing the dominant gastrointestinal feature. The overall ranking remained largely stable, with most top predictors unchanged. The only exception was Basedow disease, which became the highest-ranked feature; however, all its cases occurred exclusively within the depression group, suggesting that its apparent importance may be driven by class imbalance rather than a robust clinical association. Notably, the subsequent key predictors closely aligned with those from the original models, supporting the stability and reliability of the overall feature ranking. The comparatively low value of all other characteristics suggests that a small number of important indicators may be disproportionately important in defining people who have both autoimmune and

mental health disorders. Generic lab tests (such as routine blood test values) and non-specific drugs like ciprofloxacin or paracetamol, which are often administered but not very discriminative in this clinical setting, were the main less significant characteristics. This highlights the need for a more robust preprocessing phase to filter out low-informative features and focus on clinical variables that are more suitable for this case study. By refining this selection, the model can achieve greater accuracy and clinical relevance.

These considerations further reinforce the importance of integrating predictive modeling into clinical practice to support an early intervention and a rigorous follow-up. The framework in Fig. 1 is designed to support GPs in identifying early risk factors of mental disorders in patients with autoimmune diseases. The goal is to enhance primary care physicians' ability to monitor the emotional and psychological well-being of this specific patient population, who are often at increased risk of mental health comorbidities. In addition, offering GPs information on the percentage of falling into the anxiety and depression categories gives them a valuable tool to differentiate between these two often overlapping conditions. This distinction, which can be challenging due to their similar clinical presentations, can help streamline the diagnostic process, thereby reducing the workload for both GPs and the broader healthcare system.

In a personalized treatment perspective, wearable devices could be used to continuously monitor physiological and behavioral parameters, such as sleep patterns, heart rate variability, and activity levels. This represents a promising direction for future development since these data can not only provide valuable information for GPs during follow-up but can also serve as additional input features for ML models. More patients will be needed for future research to have a more representative and balanced dataset, which will allow for gender studies to be carried out. Furthermore, to build more effective and discriminating models of the three classes, it will be required to refine the initial features to find the most significant ones. Additionally, it would be interesting to use the same framework in a different time frame, such as after COVID-19 (i.e., since 2019), to see how GPs' perspectives and attention to anxiety and depression have evolved and to verify the emergence of additional factors that might help in the detection of mental disorders' onset.

5.2. Machine learning impact

We leveraged state-of-the-art ML models to discriminate autoimmune patients with and without mental health disorders. The control class was found to be linearly separable from the two mental disorder classes; however, anxiety and depression often overlapped, resulting in frequent misclassification across all model, likely due to both clinical ambiguity and their status as minority classes.

To address this, we incorporated custom loss functions into the XGBoost model and focused on discriminating between the two previously misclassified mental disorder classes. XGBoost belongs to the ensemble-tree family, which, among all models tested, showed more robust performance in distinguishing mental disorder subtypes. We demonstrated that modifying XGBoost's objective function using class-sensitive penalties substantially improved the model's ability to separate anxiety from depression. Different penalty terms produced distinct effects, often enhancing the recognition of one class (either depression or anxiety), while increasing misclassification of the other. In particular, improving the prediction of the anxiety class, the most challenging to detect, frequently came at the cost of reduced performance for depression, as observed with penalties such as CPS, ACS, LEM, and KL-UP. Therefore, the most substantial improvements were achieved when a balance between the recognition of both classes was maintained, as in the case of MTPM and LDEP. Since only the LDEP term led to a statistically significant improvement over the baseline, while also further enhancing anxiety detection compared to MTPM, it can be considered the most effective approach for this specific setting. However, this balance might be highly task-dependent and not generalizable across different datasets or clinical contexts, where class distributions, symptom presentation, or diagnostic overlaps may vary. The sensitivity analysis on cross-practice robustness, based on the L1GPO validation, provides important insights into the model's ability to generalize across different clinical practices. Importantly, Table 7 shows that predictive stability is largely driven by the inherent heterogeneity of the practices. Each GP represents a distinct clinical micro-environment, with its own demographic, behavioral, and comorbidity profile. When the model is trained on data from some practices and tested on another with a markedly different case mix, misclassifications are more likely—not because of a modeling weakness, but because of genuine shifts in the underlying patient population. The slight decrease in average performance relative to the standard 5FCV aligns with this interpretation: given the small dataset and the limited representation of minority classes in several practices, even shifting only a few instances in the confusion matrix (sometimes as few as 2–3 patients) can lead to large variations the terms of average recall or F1-score. This effect represents a limitation of the dataset rather than an intrinsic instability of the model itself. Overall, the L1GPO analysis confirms that achieving cross-practice generalization in heterogeneous, small-scale clinical settings is inherently challenging, but it also requires interpreting performance in relation to the underlying population structure and to account for real-world variability across sites. These findings reinforce the need to increase both the size and diversity of the dataset as essential to stabilize class distributions, reduce site-driven variability, and fully leverage the model's potential across heterogeneous clinical environments.

5.3. Limitations and future directions

This study presents limitations which are important to recognize and discuss. Firstly, while the inclusion of the LDEP penalty term improves the model's ability to discriminate minority classes, overall predictive performance remains moderate, reflecting the inherent difficulty of detecting mental-health conditions in heterogeneous and imbalanced patient populations. This is also supported by cross-practice validation, in which performance variability across GPs reflects not only differences in demographic, behavioral, and comorbidity profiles, but also potential label and selection biases arising from misdiagnosis or under-recognition of anxiety

and depression. This underscores that variability in model performance can be driven both by inherent population heterogeneity and by limitations in the reliability of recorded diagnoses. These challenges carry ethical implications (Giovannola & Tiribelli, 2023): models trained on retrospective routine-care data inherit eventual biases present in the dataset and in diagnostic practices, potentially perpetuating misdiagnosis and under-recognition of anxiety and depression. Moreover, leveraging a 10-year longitudinal EHR dataset from 6 GPs inevitably introduces challenges typical of real-world primary care data, such as missing values, heterogeneity, and class imbalance, all of which affect model performance. While we consider this heterogeneity a strength, because it captures the true variability of routine clinical practice, it also adds an additional layer of difficulty for model robustness and generalizability.

Looking ahead, future work should focus on expanding the dataset to improve class balance and integrate richer clinical information, including secondary-care data (Smith et al., 2021), to reduce misclassification.

Beyond dataset curation, methodological developments should explicitly address both ethical and clinical challenges. Model design should remain grounded in realistic clinical scenarios and aligned with the state of the art, aiming for reliable, interpretable, and ethically responsible applications. Future work could explore multimodal composite models combining structured EHR data with free-text notes (Garriga et al., 2023), uncertainty-aware frameworks (Campagner et al., 2025) to quantify prediction confidence and support cautious clinical decisions, and federated learning approaches (Joshi & Joseph, 2025) to enable collaborative model training across sites while preserving data privacy and accounting for heterogeneous populations.

6. Conclusion

This study explores the primary risk factors for anxiety and depression in patients with autoimmune disorders leveraging longitudinal EHR data from primary care by evaluating multiple loss functions and machine learning algorithms. Our results align with current literature, both from a methodological point of view, with tree-based models still outperforming others in handling class imbalance and high missingness, and from a clinical point of view, emphasizing the importance of in-depth gastrointestinal monitoring and the brain-gut axis, particularly in patients with complex comorbidities. The critical role of primary care data, routinely collected by GPs, is often underestimated; however, it can offer unique insights into the complex relationship between autoimmune disorders and mental health conditions. Responsible application of machine learning in this context holds great promise for supporting clinical decision-making, while simultaneously highlighting the ongoing need for interdisciplinary research to advance understanding, guide ethically informed practices, and refine predictive models to fully harness the potential of primary care data.

CRediT authorship contribution statement

Mariachiara Di Cosmo: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sara Campanella:** Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Michele Bernardini:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Adriano Mancini:** Writing – review & editing, Supervision, Resources, Project administration, Conceptualization. **Lorenzo Palma:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Conceptualization.

Ethical approval

The Ethical Committees of the University approved the experimental study and its guidelines as a clinical non-interventional (observational) study. EHR data are anonymous and their use, detention and conservation are regulated by an agreement between the authors and data owners. The entire process is inside the EU GDPR.

Declaration of competing interest

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

Data availability

Data will be made available on request.

References

- Ahmed, A., Aziz, S., Toro, C. T., Alzubaidi, M., Irshaidat, S., Serhan, H. A., Abd-Alrazaq, A. A., & Househ, M. (2022). Machine learning models to detect anxiety and depression through social media: A scoping review. *Computer Methods and Programs in Biomedicine Update*, 2, Article 100066.
- Baran, F. D. E., & Cetin, M. (2025). AI-driven early diagnosis of specific mental disorders: a comprehensive study. *Cognitive Neurodynamics*, 19(1), 70.
- Benros, M. E., Eaton, W. W., & Mortensen, P. B. (2014). The epidemiologic evidence linking autoimmune diseases and psychosis. *Biological Psychiatry*, 75(4), 300–306.
- Bernardini, M., Romeo, L., Frontoni, E., & Amini, M.-R. (2021). A semi-supervised multi-task learning approach for predicting short-term kidney disease evolution. *IEEE Journal of Biomedical and Health Informatics*, 25(10), 3983–3994.
- Campagner, A., Biganzoli, E. M., Balsano, C., Cereda, C., & Cabitza, F. (2025). Modeling unknowns: A vision for uncertainty-aware machine learning in healthcare.
- Chahar, R., Dubey, A. K., & Narang, S. K. (2024). Multiclass classification of mental health disorders using XGBoost-HOA algorithm. *SN Computer Science*, 5(8), 1167.
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd international conference on knowledge discovery and data mining* (pp. 785–794).
- Doeniyas, C., Clarke, G., & Cserjési, R. (2025). Gut–brain axis and neuropsychiatric health: recent advances. *Scientific Reports*, 15(1), 3415.
- Euesden, J., Danese, A., Lewis, C. M., & Maughan, B. (2017). A bidirectional relationship between depression and the autoimmune disorders—new perspectives from the national child development study. *PLoS One*, 12(3), Article e0173015.
- Forrest, I. S., Petrazzini, B. O., Duffy, Á., Park, J. K., O’Neal, A. J., Jordan, D. M., Rocheleau, G., Nadkarni, G. N., Cho, J. H., Blazer, A. D., et al. (2023). A machine learning model identifies patients in need of autoimmune disease testing using electronic health records. *Nature Communications*, 14(1), 2385.
- Francesconi, A., di Biase, L., Cappetta, D., Rebecchi, F., Soda, P., Sicilia, R., Guarrasi, V., Alzheimer’s Disease Neuroimaging Initiative, et al. (2025). Class balancing diversity multimodal ensemble for Alzheimer’s disease diagnosis and early detection. *Computerized Medical Imaging and Graphics*, 123, Article 102529.
- Garriga, R., Buda, T. S., Guerreiro, J., Iglesias, J. O., Aguerri, I. E., & Matic, A. (2023). Combining clinical notes with structured electronic health records enhances the prediction of mental health crises. *Cell Reports Medicine*, 4(11).
- Giovanola, B., & Tiribelli, S. (2023). Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI & Society*, 38(2), 549–563.
- Guo, Y., Huang, Y., Wang, Y., Liu, Y., Jing, S., Shan, T., Miao, Y., & Li, B. (2025). MDP-GRL: Multi-disease prediction by graph-enabled representation learning. *IEEE Journal of Biomedical and Health Informatics*, 1–14. <http://dx.doi.org/10.1109/JBHI.2025.3584916>.
- Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirmmeister, R. T., & Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045), 319–326.
- Iglewicz, B., & Hoaglin, D. C. (1993). *Volume 16: how to detect and handle outliers*. Quality Press.
- Islam, M. M., Hassan, S., Akter, S., Jibon, F. A., & Sahidullah, M. (2024). A comprehensive review of predictive analytics models for mental illness using machine learning algorithms. *Healthcare Analytics*, 6, Article 100350.
- Joshi, H., & Joseph, S. (2025). Standardization and interoperability: Federated learning’s impact on EHR systems and health informatics. *Advances in Health Information Science and Practice*, 1(1).
- Joyce, J. M. (2011). Kullback–Leibler divergence. In *International encyclopedia of statistical science* (pp. 720–722). Springer.
- Khoo, L. S., Lim, M. K., Chong, C. Y., & McNaney, R. (2024). Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches. *Sensors*, 24(2), 348.
- Maneeton, B., Maneeton, N., & Louthrenoo, W. (2013). Prevalence and predictors of depression in patients with systemic lupus erythematosus: a cross-sectional study. *Neuropsychiatric Disease and Treatment*, 999–804.
- Mushava, J., & Murray, M. (2024). Flexible loss functions for binary classification in gradient-boosted decision trees: An application to credit scoring. *Expert Systems with Applications*, 238, Article 121876.
- Nickson, D., Meyer, C., Walasek, L., & Toro, C. (2023). Prediction and diagnosis of depression using machine learning with electronic health records data: a systematic review. *BMC Medical Informatics and Decision Making*, 23(1), 271.
- Raison, C. L., Capuron, L., & Miller, A. H. (2006). Cytokines sing the blues: inflammation and the pathogenesis of depression. *Trends in Immunology*, 27(1), 24–31.
- Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358. <http://dx.doi.org/10.1056/NEJMra1814259>.
- Rakshasa-Loots, A. M., Swiffen, D., Steyn, C., Marwick, K. F., & Smith, D. J. (2025). Affective disorders and chronic inflammatory conditions: analysis of 1.5 million participants in our future health. *BMJ Mental Health*, 28(1), Article e301706.
- Rodríguez, P., Bautista, M. A., Gonzalez, J., & Escalera, S. (2018). Beyond one-hot encoding: Lower dimensional target embedding. *Image and Vision Computing*, 75, 21–31.
- Roehrs, A., da Costa, C. A., da Rosa Righi, R., Rigo, S. J., & Wichman, M. H. (2019). Toward a model for personal health record interoperability. *IEEE Journal of Biomedical and Health Informatics*, 23(2), 867–873. <http://dx.doi.org/10.1109/JBHI.2018.2836138>.
- Sedda, S., Cadoni, M. P. L., Medici, S., Aiello, E., Erre, G. L., Nivoli, A. M., Carru, C., & Coradduzza, D. (2025). Fibromyalgia, depression, and autoimmune disorders: An interconnected web of inflammation. *Biomedicine*, 13(2), 503.
- Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81, 84–90.
- Smith, D., Willan, K., Prady, S. L., Dickerson, J., Santorelli, G., Tilling, K., & Cornish, R. P. (2021). Assessing and predicting adolescent and early adulthood common mental disorders using electronic primary care data: analysis of a prospective cohort study (ALSPAC) in southwest England. *BMJ Open*, 11(10), Article e053624.
- Stokanović, S., Đukić, D., & Miljković, N. (2024). Robustness of XGBoost algorithm to missing features for binary classification of medical data. In *2024 23rd international symposium INFOTEH-JAHORINA* (pp. 1–6). IEEE.
- Suganya, K., & Koo, B.-S. (2020). Gut–brain axis: role of gut microbiota on neurological disorders and how probiotics/prebiotics beneficially modulate microbial and immune pathways to improve brain functions. *International Journal of Molecular Sciences*, 21(20), 7551.
- Tennenhouse, L. G., Marrie, R. A., Bernstein, C. N., Lix, L. M., et al. (2020). Machine-learning models for depression and anxiety in individuals with immune-mediated inflammatory disease. *Journal of Psychosomatic Research*, 134, Article 110126.
- Upadhyay, D. K., Mohapatra, S., & Singh, N. K. (2024). An early assessment of persistent depression disorder using machine learning algorithm. *Multimedia Tools and Applications*, 83(16), 49149–49171.
- Wang, C., Deng, C., & Wang, S. (2020). Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *Pattern Recognition Letters*, 136, 190–197.
- Wu, Y., Wang, L., Tao, M., Cao, H., Yuan, H., Ye, M., Chen, X., Wang, K., & Zhu, C. (2023). Changing trends in the global burden of mental disorders from 1990 to 2019 and predicted levels in 25 years. *Epidemiology and Psychiatric Sciences*, 32, Article e63.
- Yang, L., Jin, Y., Lu, W., Wang, X., Yan, Y., Tong, Y., Su, D., Huang, K., & Zou, J. (2025). Application of machine learning in depression risk prediction for connective tissue diseases. *Scientific Reports*, 15(1), 1706.