Received 03 January 2026, accepted 15 May 2026, date of publication 30 September 2026
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.25.3.4840

Machine Learning Techniques for Diagnosis of Panic Disorder in Patients with a Psychiatric History of Anxiety

JORDAN URIEL LAQUI PACAYA, JAIR LENIN LUCAS BENAVIDES, YASIEL PÉREZ VERA
Universidad Nacional de San Agustín de Arequipa, Arequipa, Perú
(E-mail: jlaqui@unsa.edu.pe, jlucasb@unsa.edu.pe, yperezv@unsa.edu.pe)

Corresponding author: Jordan Uriel Laqui Pacaya (e-mail: jlaqui@unsa.edu.pe)

ABSTRACT Panic disorder is a significant anxiety-related condition that can severely impact patients' quality of life, especially among individuals with a prior psychiatric history of anxiety. Early and accurate detection is essential for timely intervention but remains challenging due to the multifactorial nature of the disorder. This study aimed to analyze different supervised machine learning approaches to identify the most effective way for diagnosing panic disorder in patients with a history of anxiety. A dataset containing clinical, psychological, and lifestyle variables was analyzed using PyCaret's automated machine learning workflow. Several classification algorithms, including logistic regression, decision trees, random forests, and advanced ensemble models, were trained and evaluated. The results demonstrated that boosting-based models, specifically LightGBM, XGBoost, and Gradient Boosting, achieved the highest performance, with accuracy exceeding 99% and AUC values close to 100%. Interpretability analysis identified key predictive factors such as moderate symptom severity, current stressors, regular exercise, and fear of losing control. These insights provide valuable information for clinical practice and future research, highlighting the potential of interpretable machine learning models as decision-support tools for early detection and personalized intervention. Further validation on extended and demographically varied cohorts is recommended to ensure the models' generalizability and ethical deployment in real-world settings.

KEYWORDS Machine Learning, Panic Disorder Diagnosis, Anxiety Disorder, Advanced Ensemble Models

I. INTRODUCTION

Generalized anxiety disorder (GAD) is a mental illness characterized by feelings of worry, fear, or restlessness. Although it is normal to feel anxious from time to time, if anxiety is excessive or persistent, it can become a disorder. Treatments can be psychotherapies or medications (typically antidepressants), and it is important to seek help to avoid symptoms interfering with day-to-day activities [1].

Individuals with GAD frequently experience nonstop excessive worry [1], typically without a specific trigger. GAD can impact fully on people's daily lives, compromising both emotional and physical health. It impacts people of all ages, and while each person's symptoms may be unique, its impact on QoL can be quite profound [2].

Panic disorder (PD) is an anxiety disorder that presents repeated, unexpected attacks of intense fear. An attack may peak in a matter of minutes. Patients often describe physical symptoms of chest pain, shortness of breath and dizziness [2]. This disorder can be caused by genetic, biological, psychological factors, or stressful situations.

PD impacts society by altering the quality of life of those who suffer from it. The fear of experiencing new episodes can lead people to avoid places or situations that they associate with previous experiences of panic, thus developing agoraphobia [3]. This condition affects their work and academic life, hindering their daily activities. PD can cause a loss of autonomy in the person, requiring support from others for daily activities [4].

Panic attacks are episodes of intense fear that last a short time, unlike anxiety disorders, which are less intense but more prolonged. Although both disorders share fear as an emotional component, PD is distinguished by the constant anticipation of future episodes. This generates persistent worry about when and where the subsequent attacks will occur [3].

A timely diagnosis of PD allows appropriate treatment to be established, which is important in managing the severity and

frequency of panic episodes and consequently obtaining a healthier and more functional lifestyle [4]. Without proper intervention, and when left untreated, panic episodes may escalate and develop into more debilitating experiences such as specific phobias, social withdrawal, and impairment in occupation or education [5]. Individuals with PD may use avoidance behaviors leading to therapeutic strategies that allow safe exposure to situations to diminish distress and prevent further impairment.

PD is a major mental illness. However, diagnosis can be difficult because there are numerous symptoms and confounding disorders with GAD, which causes many patients to be left undiagnosed for long periods of time [5]. Considering this issue, it would be beneficial to use machine learning techniques to diagnose this disorder because it will help healthcare professionals by giving them quicker assessments and better-quality assessments [6]. In addition, these techniques will help to predict the likelihood of the patient developing PD, leading to implementing preventative interventions or treatment guided by the patient's characteristics. Additionally, using machine learning techniques will make diagnosis more efficient, it will help reduce misdiagnosis and delayed diagnosis of PD or GAD, and help to enhance the quality of life of patients and help streamline the cost of treatment for the healthcare system.

The objective of this research is to analyze and evaluate machine learning models to determine the best way to diagnose PD in patients with a history of anxiety.

Unlike previous studies, which primarily focus on comparing standard machine learning algorithms on general-purpose datasets, this work introduces several distinctive elements. First, the analysis is specifically focused on patients with a psychiatric history of anxiety, a subgroup that has received limited attention in prior research despite its clinical relevance. Second, this study not only evaluates the predictive performance of multiple supervised learning models but also incorporates an interpretability analysis to identify the most influential variables associated with panic disorder.

Furthermore, the use of an automated machine learning framework (PyCaret) enables a systematic and reproducible comparison of models under consistent conditions, reducing potential bias in model selection. This integrated approach, combining performance evaluation, interpretability, and a clinically meaningful dataset, provides a more comprehensive perspective compared to traditional studies that focus solely on prediction accuracy. As a result, this research contributes not only to improving diagnostic performance but also to enhancing the understanding of the factors associated with panic disorder, which is essential for clinical decision-making. The main contributions of this study are:

- A comparative evaluation of multiple supervised machine learning models for panic disorder diagnosis.
- The use of a large-scale dataset specifically filtered for patients with a psychiatric history of anxiety.
- The application of an automated machine learning framework (PyCaret) to optimize model selection and performance.
- The integration of interpretability analysis to identify the most influential variables in the prediction.
- The provision of clinically relevant insights to support early detection and decision-making.

This paper consists of the following sections. Section I reviews related work. Section III details the materials and methodology of this study. Section IV presents the results and discusses each of them. Finally, Section V presents the conclusions of this research.

II. RELATED WORK

Recent studies have increasingly explored the application of machine learning techniques in mental health and clinical decision support systems. For instance, recent works have applied explainable machine learning models to improve the diagnosis of anxiety and depression, highlighting the importance of interpretability in clinical contexts. Additionally, advances in medical artificial intelligence have emphasized the need for robust and generalizable models, particularly when dealing with sensitive clinical data. These studies demonstrate a growing trend toward integrating predictive performance with interpretability and clinical applicability. Below are some related works on applying machine learning techniques to optimize disease diagnosis. These diseases include GAD, PD, cardiovascular disease (CVD), dementia, depression, and others. The collection of these studies was limited to research articles indexed in Scopus and MDPI and no older than five years.

The study [7] focuses on the early detection of CVD using machine learning techniques, highlighting the importance of incorporating psychological factors to improve prediction performance. For this, five algorithms were used: Decision Tree (DT), Random Forest (RF), XGBoost, Support Vector Machine (SVM), and Deep Neural Network (DNN). This study included 375,145 people with normal health status and CVD. The machine learning model was able to predict CVD with an accuracy of 71.31% using only traditional risk factors. However, adding psychological aspects to the training data increased accuracy to 85.13%. Incorporating psychological factors into a machine learning model significantly improved prediction accuracy.

In the study [8], machine learning techniques were used to analyze electronic health record (EHR) data from American Indian and Alaska Native (AI/AN) populations. The goal was to create a two-year dementia risk prediction model for AI/AN persons who actively use the Indian Health Service (IHS). Four techniques were tested: logistic regression (LR), least absolute shrinkage and selection operator (LASSO), random forest (RF), and XGBoost. The generated models were based on the Dementia Risk Score (DRS), with performance evaluated using the Area Under the Curve (AUC). As a result, it was discovered that machine learning algorithms may accurately estimate the two-year dementia risk in AI/AN older person.

Conventional diagnosis of depression and anxiety involves several methods, including questionnaire-based approaches. These methods may lack accuracy because they depend on how the patient reports their emotional situation, which can be influenced by self-perception. Faced with this problem, in [9], machine learning techniques were evaluated to address these

limitations and improve diagnostic accuracy. The recursive elimination algorithm (RFE) was used for feature extraction from the dataset, and classification algorithms such as SVM, DT, RF, LR, and XGBoost were used to identify the most effective technique. The results were evaluated using performance metrics such as accuracy, precision, recall, and score. The best-performing algorithms were LR and XGBoost. In terms of effectiveness, RF stood out from SVM. The authors emphasize the high accuracy of the models used and how they can help clinicians identify the appropriate level of traditional or digital care before initiating treatment.

In clinical practice, the differentiation between PD and other anxiety disorders has been done primarily based on observable symptoms rather than in terms of more objective or quantifiable criteria. The study [10] aims to differentiate PD from other anxiety illnesses. To accomplish this, five algorithms were used: RF, Gradient Boosting Machine (GBM), SVM, Artificial Neural Network (ANN), and Regularized LR, with heart rate variability (HRV) as the baseline. The study included 60 people with Parkinson's disease and 61 participants with various anxiety disorders (ages 20 to 65). The findings revealed that Regularized LR had the highest accuracy (0.784), followed by ANN (0.730), SVM (0.730), GBM (0.676), and RF (0.649). The LR also fared well on other metrics, such as the F1 score (0.790), specificity (0.737), sensitivity (0.833), and Matthews' correlation coefficient (0.572). The study found that HRV, along with machine learning approaches, can be utilized to distinguish PD from other anxiety illnesses.

The study [11] presented a machine learning approach for anxiety detection based on perceived control data collected through a mobile application. Random Forest achieved the best performance and reached values of up to 88% ROC-AUC on a dataset comprising 401 samples. The results indicated that variables related to perceived control enabled the prediction of anxiety symptoms. This supported the use of behavioral data as a complementary diagnostic tool.

The study [12] reviews research on predicting anxiety disorders using machine learning approaches. The study expands its scope by analyzing studies on comorbidity in patients with anxiety. The tree ensemble technique with random subsampling and cross-validation was used to predict comorbidity in these patients. Machine learning models were applied to electronic medical records to improve the accuracy in predicting suicide attempts. The study proposes a model based on linguistic and acoustic patterns to classify individuals according to their level of suicide risk.

Another approach to studying anxiety disorder prediction is presented in [13]. This study explores using Natural Language Processing (NLP) to detect mental health status through text messages. Five classifiers were employed: LR, BernoulliNB, K-Nearest Neighbors (KNN), RF, and DT. LR performed the best, with 93% accuracy in classifying patient data.

In study [14], the use of combinations of polygenic scores for psychiatric disorders and intermediate phenotypes in discriminating between individuals with PD and healthy controls is explored. This study evaluates whether reliable discrimination can be obtained from polygenic scores for psychiatric disorders, calculated using large-scale genome-wide association studies. The authors used five classification methods: LR, neural networks, quadratic discriminant analysis, RF, and SVM to analyze 48 polygenic scores in 718 individuals with PD and 1,717 healthy controls. The study showed that increasing the number of polygenic score combinations improved the performance of the classifiers. SVM (AUC 0.617) and LR (AUC 0.600) were the classifiers with the highest accuracy.

Despite the significant advances reported in previous studies, several limitations remain. Many works rely on specific datasets with limited generalizability, lack interpretability, or do not focus specifically on panic disorder in patients with a psychiatric history of anxiety. Additionally, few studies combine high-performance models with interpretable insights for clinical decision-making. To address these gaps, this study proposes a comparative and interpretable machine learning approach focused specifically on panic disorder diagnosis in patients with prior anxiety history.

III. MATERIAL AND METHODS

This section presents the concepts, technologies, and tools used in this research and explains the proposed model's outline, implementation, and development.

A. THEORETICAL FOUNDATION

Below are the theoretical concepts necessary to fully understand the proposed approach. These concepts primarily encompass the machine learning concepts and algorithms used in this research.

Panic disorder (PD) is an anxiety disorder characterized by sudden episodes of fear. It is accompanied by physical symptoms such as chest pain, palpitations, breathing problems, dizziness, or abdominal discomfort. Misdiagnosis and inadequate treatment are common among non-psychiatric physicians. Despite this, these physicians' role is crucial in managing PD, as most patients initially seek medical care in general medicine centers [15]. Without adequate treatment, this disorder ultimately disrupts people's daily lives.

Machine learning (ML) is a branch of artificial intelligence (AI) that uses data and algorithms to replicate how humans learn, gradually improving their accuracy. These methods detect patterns and based on them, make predictions or decisions. The use of ML in mental health research is on the rise, especially as new, more complex types of data become available for analysis [16].

ML techniques encompass a variety of concepts, such as algorithms, methodologies, approaches, and practices. In other words, they are the general strategies and frameworks applied to address problems through algorithms. These techniques encompass the entire process, from model design and training to evaluation and deployment. ML techniques are classified as either "supervised learning," where predefined outputs are obtained, or "unsupervised learning," which seeks to identify hidden patterns in data that are not easily recognizable by humans [17].

Decision Tree (DT) is an algorithm used in AI and ML to make decisions or predictions in a clear and structured way. Its structure resembles a tree, with an initial root, intermediate

branches, and final leaves. It divides the data into parts according to specific characteristics, asking questions at each step. Internal nodes correspond to attribute evaluations, branches reflect the outcomes of these evaluations, and the final leaf indicates a category or outcome. It is a very popular tool in classification and prediction tasks due to its ease of interpretation. Thanks to its design, it helps to understand how a decision is reached step by step. Although it can become complex with extensive data, it is still a very useful and visual option [18].

Random forest (RF) is a supervised algorithm that utilizes multiple trained TDs. Each tree is trained on a random data subsampled, making it more generalizable. The final tree is combined by voting to get the final prediction. This method minimizes the chance of overfitting and improves accuracy for classification and regression tasks. RFs are more reliable than models developed using a single decision tree, as the predictions are based on a collection of decision trees [19].

Logistic Regression (LR) is a statistical method that is mainly used to estimate binary outcomes. LR essentially provides a relationship between predictors and categorical outcomes. LR relies on the sigmoid function to convert ratios into probabilities, which naturally fall within the boundaries of 0 and 1. LR is relatively easy to interpret, as the coefficient of the predictors help define probabilities as well as the degree of impact of risk factors/treatments on the outcomes of patients. LR, more generally, can be used to predict multi-category outcomes with multinomial and ordinal logistic regression. LR is efficient and scalable, and is possible with high dimensional data sources, further enhanced by feature selection methods [20].

Gradient Boosting (GB) algorithm is a machine learning method for developing predictive models through sequential ensembles of decision trees. Gradient boosting is a boosting algorithm that implies training many weak models in sequence, where each model attempts to fix the errors of the prior model. The result is an ensemble of weak models which, when aggregated, creates a single better and more accurate model [21].

Light gradient boosting (LightGBM) is an optimized implementation of the Gradient Boosting algorithm that includes regularization techniques with an excess number of optimized implementations. This algorithm, for example, has been shown to be efficient and scalable for relatively demanding tasks like Disease Detection [22]. LightGBM is noted for achieving high accuracy during predictive classification despite inputting many records of data; this quality is particularly advantageous for resource-intensive (computationally speaking) settings. While the performance may be faster in terms of classification time with large datasets, this could in fact mean a longer runtime depending on the way it is being trained [23].

XGBoost is an algorithm based on gradient boosting and uses DT-based models. It is used for both classification and regression. Its main advantage is speed and efficiency because it can work in parallel and optimize memory usage. Compared to traditional methods, XGBoost introduces a regularization term. In addition, XGBoost builds models sequentially to correct previous errors, thus avoiding overfitting in the models. On the other hand, it can handle missing data automatically, making it very practical in real situations [24].

B. TOOLS, LANGUAGES AND LIBRARIES

Below are the theoretical concepts about the technology necessary to implement the proposal model. These concepts primarily encompass the tools used in this research and the reasons for their selection.

Google Colab (GC) is a cloud platform that allows Python code to run from the browser without requiring prior configuration. It is particularly useful in Data Science (DS), ML, and AI. GC provides free access to GPUs and TPUs to optimize processing and calculations in the models. In addition, it can interact with Google Drive, which facilitates the storage and upload of files. GC offers fast access, automatic synchronization, and an optimized interface, enabling real-time collaboration between researchers, providing a fluid and efficient experience [25].

Python is a general-purpose, interpreted programming language used in fields such as AI, ML, Data Science, and many others. It has many specialized and optimized libraries for these fields. Python is characterized by its simple and readable syntax and large and active community. Over the last decade, Python has become extremely popular within the scientific computing community, which has led to the development of the most recent ML and Deep Learning (DL) libraries in this language [26].

PyCarte 3.0 is a Python library focused on automating ML workflows. PyCaret enables data scientists to perform comprehensive experiments quickly and efficiently. PyCaret and its ML capabilities integrate with Python-compatible environments such as Microsoft Power BI, Tableau, Alteryx, and KNIME. This provides a great advantage to users of these BI platforms as they can easily integrate PyCaret into their existing processes and apply ML [27].

C. DATASET

The dataset used in this study was obtained from the Kaggle platform [28] and originally collected from clinical records at Ayesha Clinic. It contains approximately 120,000 records with multiple clinical, psychological, and lifestyle variables related to panic disorder. The dataset includes features such as age, gender, family history, psychiatric history, current stressors, symptom severity, lifestyle factors, and other relevant attributes.

Although the dataset is publicly available, it is assumed that the labels were derived from clinical assessments at the source institution. However, detailed information about the data collection protocol and labeling procedures is not fully documented, which limits the ability to independently verify the reliability of the annotations.

Due to the medical nature of the problem, this represents an important limitation. The lack of direct control over data acquisition, diagnostic criteria, and labeling processes may introduce potential bias or inconsistencies in the dataset. Therefore, the results obtained in this study should be interpreted with caution.

To partially mitigate this limitation, standard preprocessing techniques and validation strategies were applied to ensure data consistency and robustness in model evaluation. Nevertheless, future work should include the use of clinically validated datasets and external validation cohorts to strengthen the generalizability and reliability of the proposed models.

Due to the nature of this study, the dataset will be limited to data from patients with a history of psychiatric anxiety. This psychiatric history constitutes 25% of the data, as shown in Fig. 1. Fig. 2 shows that the proportion of female and male patients is 50%, respectively. Fig. 3 shows the most common symptoms. Fear of losing control and dizziness is the most common, at approximately 20%, respectively.

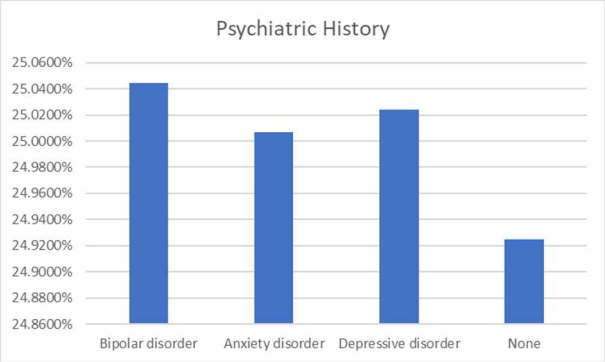

Figure 1. Percentage of psychiatric history in the dataset

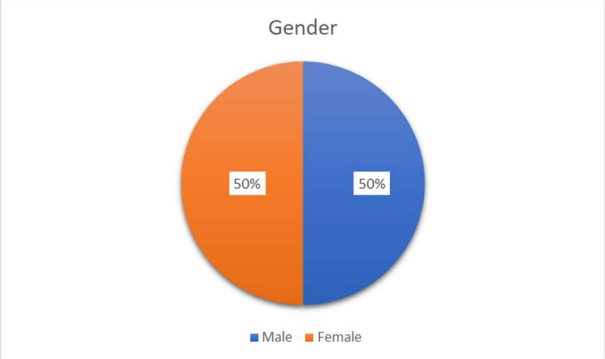

Figure 2. Percentage of psychiatric history in the dataset

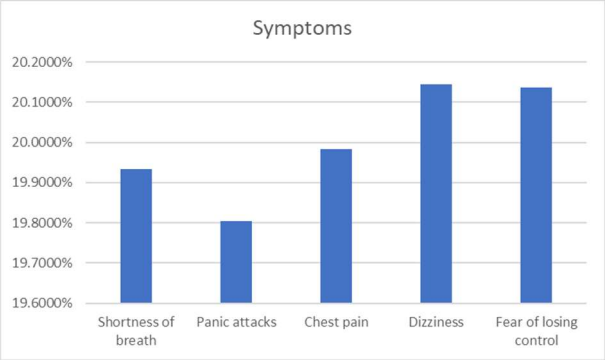

Figure 3. Percentage of psychiatric history in the dataset

D. PROPOSED METHODOLOGY

Below are the steps that form part of the methodology used in this research. Fig. 4 shows the general outline of this proposal.

1) Data collection

The original dataset contains 120,000 medical records from the Kaggle platform, which include multiple clinical, demographic, and mental health history variables. For this research, only records of patients with psychiatric histories related to anxiety disorders were selected, filtered using the Pandas library in a Google Colab-based runtime environment. After filtering, a subset of 30,000 records was obtained that met the inclusion criteria.

2) Data preprocessing

The dataset preprocessing phase included a data cleaning process to address the high proportion of missing values. To preserve the integrity of the dataset without reducing the sample size, imputation techniques were implemented: the arithmetic mean was used for numerical variables, while the mode (most frequent value) was applied for categorical variables. Additionally, outlier validation was performed; no occurrences were found in the dataset during this step. Subsequently, fundamental data transformations were performed to ensure statistical consistency and prepare the information for the subsequent stages of automatic model training.

3) Split

Once the cleaning and filtering phase of the original dataset was completed, 30,000 records were selected as the basis for the training and evaluation process. The set was divided into two parts: 80% for model training and the remaining 20% for final testing. The training subset was subdivided, allocating 80% for direct training of the algorithms and 20% for internal validation. This second partition allowed for a controlled comparison of the performance of multiple models before evaluating them with the independent test set, following a nested validation strategy that reinforces the robust selection of the final model.

4) Training

Pycaret facilitated the implementation, training, and validation of the models. The initial analysis performed with the Pycaret library allowed us to identify six algorithms with the best results in terms of metrics such as: Accuracy, Area Under the Curve (AUC), True Positive Ratio (Recall), Precision, High Precision and High Sensitivity (F1-score), Model Accuracy Indicator (Kappa), and the Matthews Correlation Coefficient (MCC). Additionally, each algorithm's main parameters were identified to improve accuracy, which are detailed in Table 1. Some of the most important are: 1) n_estimators, which defines how many individual models will be trained and combined in the final model; 2) max_depth, which determines the maximum depth of each decision tree in the model. 3) learning_rate, which defines and controls how a tree corrects the errors of the previous one. 4) n_jobs, to define how many processes will be executed in parallel.

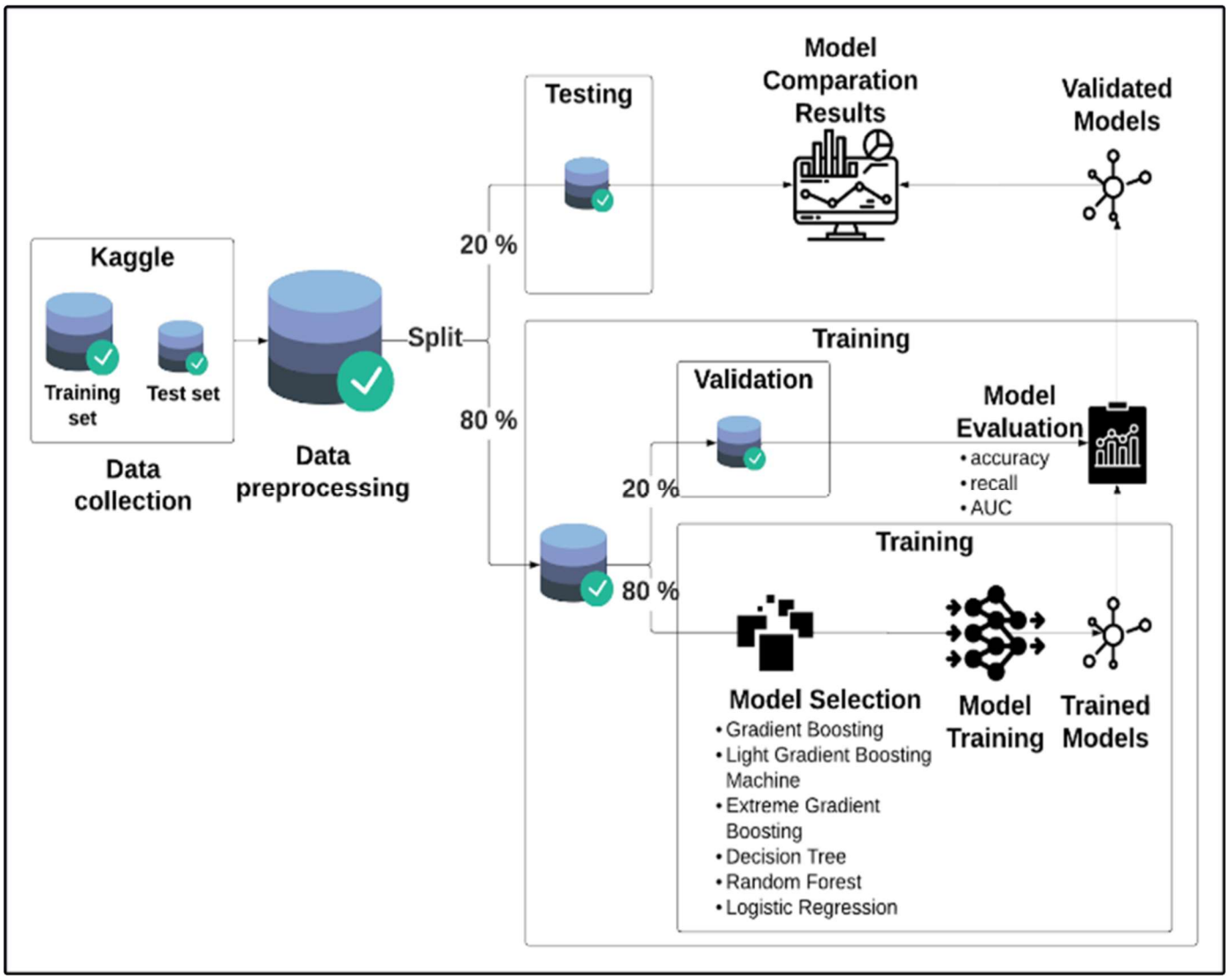

Figure 4. Proposed methodology steps

Table 1. Description of the parameters assigned to each algorithm.

Algorithm	Parameters
Gradient Boosting	n_estimators=160, learning_rate=0.03, max_depth= 5
LightGBM	learning_rate=0.1, max_depth= -1, num_leaves=31, n_estimators=100, min_child_samples= 100
XGBoost	learning_rate=0.3, max_depth=5, n_estimators=70, min_child_weight=5, subsample=1
Decision Tree	class_weight=none max_depth=none, max_features=none, max_leaf_nodes=none, splitter=best, min_weight_fraction_leaf=none
Random Forest	ccp_alpha=0, class_weight=none, criterion=gini, max_depth=none, max_features=sqrt, max_leaf_nodes=none, n_estimators=100, n_jobs=-1
Logistic Regression	max_iter=1000, fit_intercept=true, penalty=l2, C=6178

IV. RESULTS AND DISCUSSION

Below is a comparison of the performance of each algorithm evaluated, based on the most relevant classification metrics: Accuracy, AUC, Recall, Precision, F1-score, Kappa, and MCC. These indicators allow us to assess the predictive capacity of each model from different perspectives, including class balance and correlation between predictions and actual labels.

A. ACCURACY

Accuracy is one of the most often used metrics for evaluating a classification model's performance, as it measures the proportion of correct predictions out of the total number of examined cases. This metric is especially useful when the dataset is balanced, or when the classes have a similar distribution, because high accuracy usually indicates strong overall model performance. Fig. 5 compares the scores for each algorithm in terms of accuracy. The best algorithm was LightGBM (0.9929), followed by XGBoost (0.9923) and Gradient Boosting (0.9923). Meanwhile, the algorithms with the worst scores were LR (0.9655), RF (0.9810), and DT (0.9827).

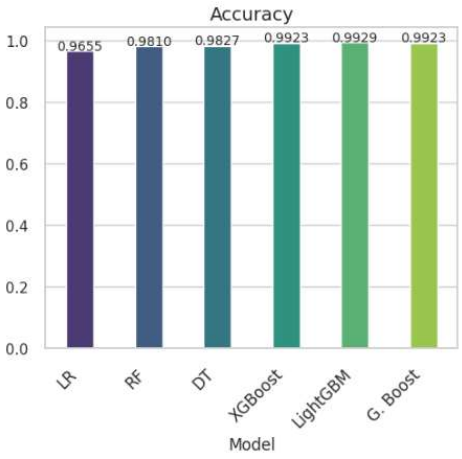

Figure 5. Scores of each algorithm in Accuracy

B. AREA UNDER THE CURVE (AUC)

This metric is widely used to assess the performance of classification models, particularly in cases where classes are imbalanced. This measure corresponds to the area beneath the ROC curve, which plots the true positive rate against the false positive rate across different threshold values. An AUC score of 1 implies that the model can differentiate between positive and negative classes, whereas 0.5 indicates chance performance. Fig. 6 compares the scores for each algorithm on the AUC metric. The best algorithms were Gradient Boosting (0.9994), XGBoost (0.9994), and LightGBM (0.9990). Meanwhile, the worst-scoring algorithms were RF (0.9940), DT (0.9955), and LR (0.9797).

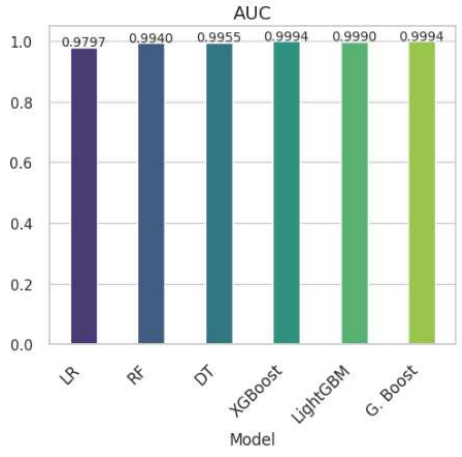

Figure 6. Scores of each algorithm in AUC

C. RECALL

It refers to the fraction of positive class instances that the model successfully detects. It is expressed as the ratio of true positives to the total number of actual positive cases, which includes both true positives and false negatives. Its importance lies in that it allows for the evaluation of the model's ability to correctly recover positive cases, which is especially critical in contexts where the omission of a positive instance is of high importance, such as in disease detection. Fig. 7 shows the comparison of the score of each algorithm for this metric. RF and XGBoost scored 1, followed by LightGBM (0.9268) and Gradient Boosting (0.9146). The worst algorithms were DT and LR with a score of 0.8049 and 0.5244, respectively.

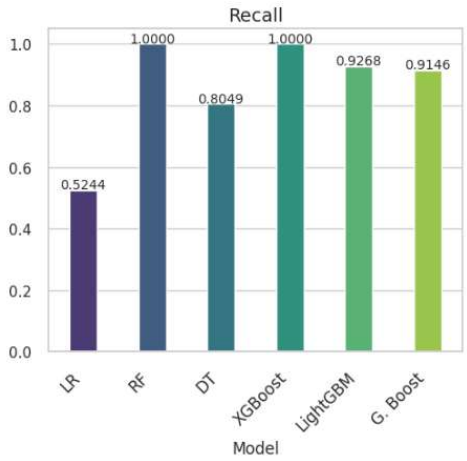

Figure 7. Scores of each algorithm in Recall

D. PRECISION

This metric is particularly important in scenarios where the cost of false positives is high. Precision is defined as the proportion of true positive predictions among all instances predicted as positive by the model, that is, the ratio of true positives to the sum of true positives and false positives. A high precision score indicates that the model makes relatively few false positive errors when predicting positive instances. Fig. 8 shows the comparison of each algorithm's score for this metric. LightGBM obtained the highest score (0.9268), followed by Gradient Boosting (0.9259). XGBoost (0.8632) and DT (0.8354) obtained moderate scores, while the worst-performing algorithms were RF and LR with scores of 0.7193 and 0.6935, respectively.

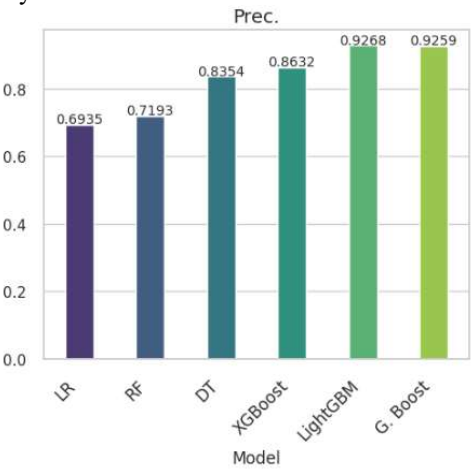

Figure 8. Scores of each algorithm in Precision

E. F1 SCORE

This metric is particularly useful in contexts where there is an imbalance between classes, and it is necessary to find a balance between Precision and Recall. This metric is calculated as the harmonic mean of Precision and Recall values. Fig. 9 shows the comparison of each algorithm's score for this metric. LightBGM obtained the highest score (0.9268), followed by XGBoost (0.9266) and Gradient Boosting (0.9202). RF and DT obtained a regular score of 0.8367 and 0.8199, respectively. The worst algorithm was LR (0.5972).

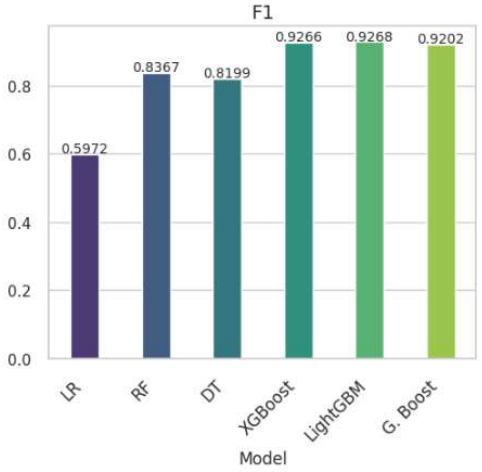

Figure 9. Scores of each algorithm in F1

F. KAPPA

This statistical indicator assesses the level of agreement between a classification model's predictions and actual classes. Its value ranges from -1 to 1, with 1 indicating perfect agreement, 0 representing agreement equal to chance, and negative values indicating agreement worse than random. Fig. 10 shows the comparison of each algorithm's score for this metric. LightBGM obtained the highest score (0.9231), followed by XGBoost (0.9225) and Gradient Boosting (0.9162). RF and DT obtained a fair score with 0.8269 and 0.8108, respectively. The worst algorithm was LR (0.5796).

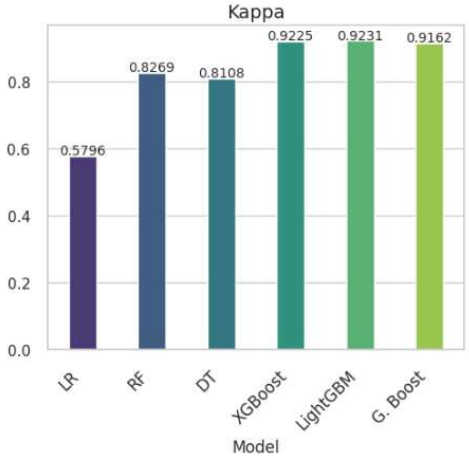

Figure 10. Scores of each algorithm in Kappa

G. MATTHEWS CORRELATION COEFFICIENT (MCC)

It is a metric known for its ability to offer a balanced evaluation of the performance of binary classification models. This metric considers true positives, true negatives, false positives, and false negatives. It is interpreted as a correlation between the model predictions and labels. Its value ranges from -1 to 1, where 1 represents a perfect prediction, 0 indicates random classification, and -1 indicates completely wrong performance (inverse predictions). Fig. 11 shows the comparison of each algorithm's score for this metric. XGBoost obtained the highest score (0.9253), followed by LightGBM (0.9231) and Gradient Boosting (0.9162). RF and DT obtained a regular score of 0.8396 and 0.8110, respectively. The worst algorithm was LR (0.5858).

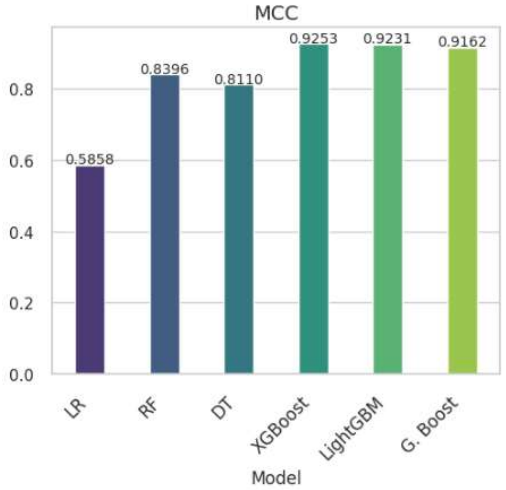

Figure 11. Scores of each algorithm in MCC

H. MODELS COMPARISON

After applying the proposed ML techniques, the algorithms with the best scores were identified, which can be visualized in Fig. 12. The quantitative results confirmed boosting models' superiority over more straightforward or traditional approaches. In particular, LightGBM, XGBoost, and Gradient Boosting achieved accuracy values greater than 99% and AUC close to 100%, highlighting their excellent discriminatory capacity. Furthermore, these models achieved a favorable balance between precision, recall, and F1-score, maintaining Kappa and MCC values above 0.91, indicating a high level of agreement between predictions and actual classification.

In contrast, algorithms such as LR and individual decision trees showed significantly lower performance metrics (e.g., a precision of only 0.69 and a recall of 0.52 for LR), highlighting the limitations of these models in capturing the complexity of the variables involved in PD. On the other hand, while achieving perfect recall, RF exhibited a lower precision (0.71), suggesting a tendency to overclassify positive cases, thus reducing its overall balance.

These findings support and reinforce the findings presented in [29], which highlight the suitability of boosting methods for addressing clinical problems. These methods are particularly appropriate for cases where risk factors are highly correlated. They are also useful when there is a need to maximize sensitivity without sacrificing accuracy or model generalization. Furthermore, the robustness of these models suggests that they can serve as a valuable tool to support medical decision-making, contributing to the early identification of patients at greater risk of developing PD.

Although the results obtained in this study show very high performance, with accuracy values exceeding 99% and AUC values close to 1, these results should be interpreted within the context of the dataset and experimental setup. Such high performance may be influenced by factors such as dataset characteristics and preprocessing strategies.

While a train-test split and internal validation strategy were applied to ensure consistency in model evaluation, these approaches may not fully capture the generalizability of the models to external data. Additionally, the use of a filtered subset focused on patients with a psychiatric history of anxiety may have resulted in a more homogeneous dataset, which can contribute to higher performance metrics.

Despite these considerations, the results remain consistent across multiple models and evaluation metrics, supporting the robustness of the findings within the scope of the study. Nevertheless, future work should incorporate more comprehensive validation strategies, such as k-fold cross-validation and testing on independent or diverse datasets, to further strengthen the generalizability and clinical applicability of the proposed approach.

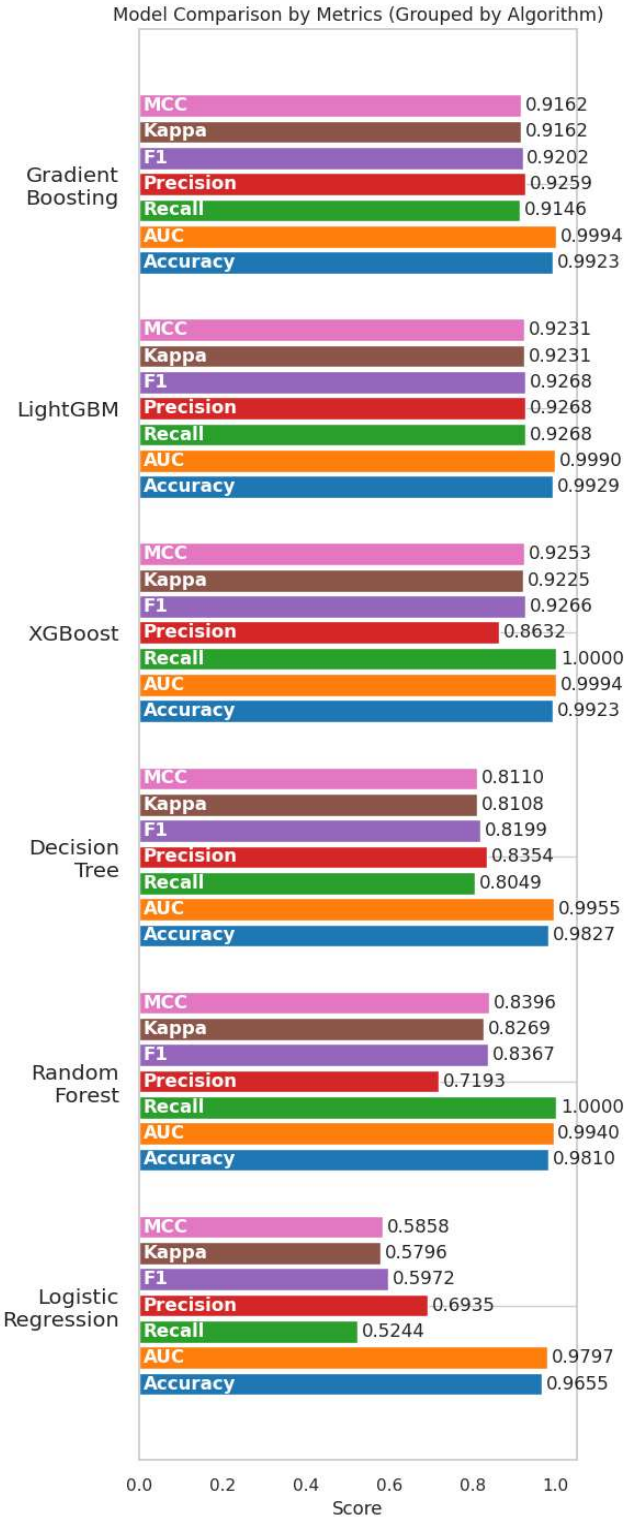

Figure 12. Model comparison by metrics

I. MOST INFLUENTIALS VARIABLES

Because the Gradient Boosting algorithm performed best across multiple evaluation metrics, including precision, recall, F1, and AUC, it was selected as the baseline model for further interpretability analysis. Specifically, the variables with the most impact on the model's predictions were identified, as illustrated in Fig. 13.

The analysis results show that the variables with the most significant relative weight in PD classification are severity at moderate intensity, low current stressors, regular exercise as a lifestyle factor, and fear of losing control as a relevant psychological symptom. These variables significantly contribute to class differentiation, suggesting their potential value as key indicators within the phenomenon studied.

These results reinforce the model's predictive capacity and provide a relevant interpretive perspective to guide clinical interventions or design more targeted prevention strategies. Furthermore, understanding the importance of these variables facilitates the generation of hypotheses for future research and supports decision-making in psychological contexts, where understanding the associated factors is essential.

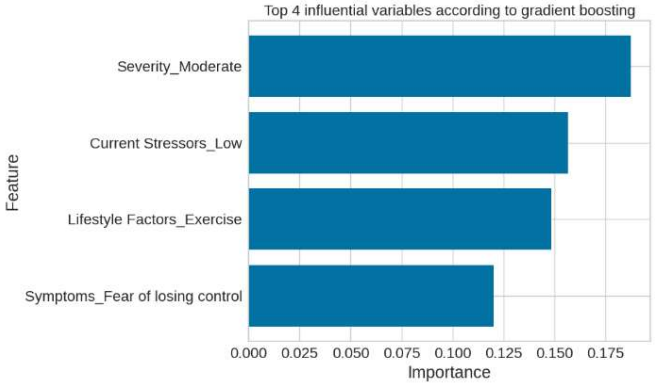

Figure 13. Most influential variables according to the Gradient Boosting

Compared to prior studies, this research goes beyond a simple comparison of machine learning algorithms by integrating both predictive performance and interpretability analysis within a clinically relevant context. While previous works have demonstrated the effectiveness of models such as Random Forest, Support Vector Machines, and boosting-based approaches, they often focus primarily on accuracy metrics without providing deeper insights into the factors driving model predictions.

In contrast, this study emphasizes the importance of interpretability, particularly in the context of mental health, where understanding the contributing variables is as important as achieving high predictive performance. The identification of key variables such as symptom severity, stress levels, lifestyle factors, and specific psychological symptoms provides valuable information that can support clinical reasoning and decision-making processes.

Additionally, the use of a filtered dataset focused on patients with a psychiatric history of anxiety allows for a more targeted analysis compared to broader datasets used in previous studies. This specificity enhances the relevance of the findings for real-world clinical applications. Although the high performance achieved by boosting models is consistent with existing literature, the integration of these results with interpretability and domain-specific data represents a meaningful contribution that distinguishes this work from

purely predictive or benchmarking studies.

V. CONCLUSIONS

Various machine learning techniques were applied and analyzed to diagnose panic disorder in patients with a history of psychiatric anxiety. Several predictive models and algorithms were used, such as decision trees, logistic regression, random forests, and other boosting-based methods.

Boosting-based algorithms such as LightGBM, XGBoost, and Gradient Boosting significantly outperformed traditional models. They excelled in key metrics, achieving accuracy above 99% and AUC values close to 100%. These results confirm the ability of boosting methods to identify complex patterns. Thus, they contribute to more accurate detection of patients at risk of developing PD.

Interpretability analysis identified variables with a high predictive impact. The most relevant variables were moderate symptom severity, low stress factors, regular exercise, and fear of losing control. These variables provide helpful information for better understanding the clinical phenomenon and open new avenues for preventive and personalized interventions.

Despite the encouraging results, these models must be validated in larger and more diverse populations. Their ethical, transparent, and regulated integration into clinical practice is also essential. This will allow for the development of data-driven decision-making support tools, strengthening prevention, early detection, and a comprehensive approach to PD in patients with a history of anxiety.

Overall, this work lays a solid foundation for future interdisciplinary research. It seeks to leverage artificial intelligence and data science as supporting tools for the diagnosis of complex psychological disorders such as panic disorder.

References

[1] R. J. McNally, "Anxiety sensitivity and panic disorder," *Biological Psychiatry*, vol. 52, no. 10, pp. 938–946, 2002. https://doi.org/10.1016/S0006-3223(02)01475-0.

[2] İ. Özdemir and E. Kuru, "Investigation of cognitive distortions in panic disorder, generalized anxiety disorder and social anxiety disorder," *Journal of Clinical Medicine*, vol. 12, no. 19, Art. no. 19, 2023. https://doi.org/10.3390/jcm12196351.

[3] P. P. Roy-Byrne, M. G. Craske, and M. B. Stein, "Panic disorder," *The Lancet*, vol. 368, no. 9540, pp. 1023–1032, 2006. https://doi.org/10.1016/S0140-6736(06)69418-X.

[4] W. J. Katon, "Clinical practice. Panic disorder," *The New England Journal of Medicine.*, vol. 354, no. 22, pp. 2360–2367, 2006. https://doi.org/10.1056/NEJMcp052466.

[5] A. Caldiroli *et al.*, "Clinical markers of panic and generalized anxiety disorder: Overlapping symptoms, different course and outcome," *Journal of Personalized Medicine*, vol. 13, no. 3, Art. no. 3, 2023. https://doi.org/10.3390/jpm13030491.

[6] T. Richter *et al.*, "Machine learning meta-analysis identifies individual characteristics moderating cognitive intervention efficacy for anxiety and depression symptoms," *NPJ Digital Medicine*, vol. 8, no. 1, pp. 1–15, 2025. https://doi.org/10.1038/s41746-025-01449-w.

[7] M. Dorraki *et al.*, "Improving cardiovascular disease prediction with machine learning using mental health data: A prospective UK biobank study," *JACC Advances*, vol. 3, no. 9, Part 2, p. 101180, 2024. https://doi.org/10.1016/j.jacadv.2024.101180.

[8] K. Ports *et al.*, "Machine learning to predict dementia for American Indian and Alaska native peoples: a retrospective cohort study," *The Lancet Regional Health – Americas*, vol. 43, 2025. https://doi.org/10.1016/j.lana.2025.101013.

[9] M. Marey, D. Salem, N. E. Rashidy, and H. ELBakry, "Enhancing the diagnosis of depression and anxiety through explainable machine learning methods," *Internal Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 15, no. 4, Art. no. 4, 33/30 2024. https://doi.org/10.14569/IJACSA.2024.0150481.

[10] K.-S. Na, S.-E. Cho, and S.-J. Cho, "Machine learning-based discrimination of panic disorder from other anxiety disorders," *Journal of Affective Disorders*, vol. 278, pp. 1–4, 2021. https://doi.org/10.1016/j.jad.2020.09.027.

[11] P. Azaglo, P. van de Ven, and J. Nelson, "Machine learning models for anxiety detection and prediction using perceived control data," *Procedia Computer Science*, vol. 248, pp. 78–88, 2024. Elsevier. https://doi.org/10.1016/j.procs.2024.10.366.

[12] T. Kotsilieris, E. Pintelas, I. E. Livieris, and P. Pintelas, "Predicting anxiety disorders and suicide tendency using machine learning: A review," *Internal Journal of Medical Engineering and Informatics*, vol. 12, no. 6, p. 599, 2020. https://doi.org/10.1504/IJMEI.2020.111040.

[13] J. Chung and J. Teo, "Mental health prediction using machine learning: Taxonomy, applications, and challenges," *Applied Computational Intelligence and Soft Computing.*, vol. 2022, no. 1, p. 9970363, 2022. https://doi.org/10.1155/2022/9970363.

[14] K. Ohi *et al.*, "Discrimination between healthy participants and people with panic disorder based on polygenic scores for psychiatric disorders and for intermediate phenotypes using machine learning," *The Australian and New Zealand Journal of Psychiatry*, vol. 58, no. 7, pp. 603–614, 2024. https://doi.org/10.1177/00048674241242936.

[15] N. Manjunatha and D. Ram, "Panic disorder in general medical practice-A narrative review," *Journal of Family Medicine and Primary Care*, vol. 11, no. 3, p. 861, 2022. https://doi.org/10.4103/jfmpc.jfmpc_888_21.

[16] K. K. Park, M. Saleem, M. A. Al-Garadi, and A. Ahmed, "Machine learning applications in studying mental health among immigrants and racial and ethnic minorities: an exploratory scoping review," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 298, 2024. https://doi.org/10.1186/s12911-024-02663-4.

[17] E. Fountzilas, T. Pearce, M. A. Baysal, A. Chakraborty, and A. M. Tsimberidou, "Convergence of evolving artificial intelligence and machine learning techniques in precision oncology," *NPJ Digital Medicine*, vol. 8, no. 1, pp. 1–19, 2025. https://doi.org/10.1038/s41746-025-01471-y.

[18] L. A. Shoaib, S. H. Safii, N. Idris, R. Hussin, and M. A. H. Sazali, "Utilizing decision tree machine learning model to map dental students' preferred learning styles with suitable instructional strategies," *BMC Medical Education*, vol. 24, no. 1, p. 58, 2024. https://doi.org/10.1186/s12909-023-05022-5.

[19] E. S. Mohamed, T. A. Naqishbandi, S. A. C. Bukhari, I. Rauf, V. Sawrikar, and A. Hussain, "A hybrid mental health prediction model using Support Vector Machine, Multilayer Perceptron, and Random Forest algorithms," *Healthcare Analytics*, vol. 3, p. 100185, 2023. https://doi.org/10.1016/j.health.2023.100185.

[20] K. Taha, "Machine learning in biomedical and health big data: a comprehensive survey with empirical and experimental insights," *Journal of Big Data*, vol. 12, no. 1, p. 61, 2025. https://doi.org/10.1186/s40537-025-01108-7.

[21] K. Yoneda, T. Seki, Y. Kawazoe, K. Ohe, N. Takahashi, and on behalf of the N. R. N. of Japan, "Immediate postnatal prediction of death or bronchopulmonary dysplasia among very preterm and very low birth weight infants based on gradient boosting decision trees algorithm: A nationwide database study in Japan," *PLOS ONE*, vol. 19, no. 3, p. e0300817, 2024. https://doi.org/10.1371/journal.pone.0300817.

[22] T. O. Omotehinwa, D. O. Oyewola, and E. G. Dada, "A Light Gradient-Boosting Machine algorithm with Tree-Structured Parzen Estimator for breast cancer diagnosis," *Healthcare Analytics*, vol. 4, 2023. https://doi.org/10.1016/j.health.2023.100247.

[23] S. Ogawa, "Predicting dropout from cognitive behavioral therapy for panic disorder using machine learning algorithms," *Journal of Clinical Medicine Research*, vol. 16, no. 5, pp. 251–255, 2024. https://doi.org/10.14740/jocmr5167.

[24] H. Yuan, K. Zhao, Y. Yan, L. Wan, Z. Tian, and X. Chen, "Long tunnel group driving fatigue detection model based on XGBoost algorithm," *Journal of Traffic Transportation Engineering English Edition*, vol. 12, no. 1, pp. 167–179, 2025. https://doi.org/10.1016/j.jtte.2023.02.008.

[25] J. Llerena-Izquierdo, J. Mendez-Reyes, R. Ayala-Carabajo, and C. Andrade-Martinez, "Innovations in introductory programming education: The role of AI with Google Colab and Gemini," *Education Sciences*, vol. 14, no. 12, Art. no. 12, 2024. https://doi.org/10.3390/educsci14121330.

[26] S. Raschka, J. Patterson, and C. Nolet, "Machine learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence," *Information*, vol. 11, no. 4, Art. no. 193, 2020. https://doi.org/10.3390/info11040193.

[27] A. T. Olanipekun and D. Mashao, "An improved prediction of transparent conductor formation energy using PyCaret: An open-source

machine learning library," *Journal of Applied Data Science*, vol. 5, no. 4, pp. 1914–1924, 2024. https://doi.org/10.47738/jads.v5i4.202.

[28] "Panic Disorder Detection Dataset," 2025. [Online]. Available at: https://www.kaggle.com/datasets/muhammadshahidazeem/panic-disorder-detection-dataset.

[29] P. Cruz-Gonzalez *et al.*, "Artificial intelligence in mental health care: A systematic review of diagnosis, monitoring, and intervention applications," *Psychological Medicine*, vol. 55, p. e18, 2025. https://doi.org/10.1017/S0033291724003295.

JORDAN URIEL LAQUI PACAYA, completed a bachelor's degree in Systems Engineering at the National University of Saint Agustine, Arequipa, Peru, in 2022. He has extensive experience in mining software development and currently works in the automotive software field. His research interests include Artificial Intelligence and Graphics Computing.

JAIR LENIN LUCAS BENAVIDES, Bachelor's degree in Systems Engineering at the National University of Saint Agustine, Arequipa, Peru. He has experience in Software Development and is currently working in Software Quality Assurance and Software Testing. His research interests include Artificial Intelligence, Big Data and Cloud Computing.

YASIEL PÉREZ VERA, is a researcher in the Department of System Engineering and Informatics of the National University of Saint Augustine, Arequipa, Peru. He is also working as a Head of the Software Engineering Program at La Salle University, Arequipa, Peru. He obtained his Project Management Master's and Bachelor's Computer Engineering Degree in the University of Informatics Science, La Havana, Cuba. His research interests include Machine Learning, Cloud Computing and Databases. He is a certified Cisco DevNet Associate.