# A Machine Learning Approach Toward the Detection of Psychiatric Conditions

Luke Profio

The University of Texas at Austin, Department of Computer Science
profio@utexas.edu

**Codebase**
**Dataset**

Psychiatric disorders are a leading cause of disability worldwide and are associated with elevated mortality, reduced quality of life, and substantial economic burden. Yet diagnosis of conditions such as schizophrenia, bipolar disorder, major depressive disorder (MDD), and anxiety disorders still relies largely on semi-structured interviews and symptom checklists, which are inherently subjective and can be influenced by recall bias, comorbidity, and differences in clinical training. Objective, biologically grounded measures that complement clinical judgement could help reduce diagnostic delay and misclassification, particularly in settings with limited access to specialists.

Neurophysiological and autonomic recordings—most prominently electroencephalography (EEG) and heart rate variability (HRV) derived from electrocardiography—provide rich, temporally resolved measures of brain and autonomic nervous system activity. Extensive evidence indicates that these signals differ systematically between psychiatric patients and healthy controls [3, 4, 19, 18]. Over the past two decades, machine learning (ML) and deep learning (DL) methods have increasingly been applied to EEG and related signals to classify psychiatric diagnoses, with many studies reporting high accuracies for disorders such as schizophrenia and depression [1, 2, 9, 8, 16, 13, 23, 21]. However, most prior work has focused on binary classification of a single disorder versus healthy controls, often on small proprietary datasets, and has offered limited insights into how models reach their decisions [2, 20, 22].

In this project, we investigate whether an ML pipeline can reliably distinguish between multiple psychiatric diagnoses using quantitative EEG features from the openly available *EEG Psychiatric Disorders Dataset* on Kaggle [26, 23]. We treat diagnosis as a multi-class prediction problem and systematically compare classical ML algorithms with neural network models under a consistent preprocessing and evaluation framework. Motivated by recent calls for interpretable and trustworthy AI in psychiatry [12, 22, 5, 25], we also explore feature importance patterns to identify which frequency bands and channels contribute most strongly to each diagnostic label. Although our experiments are not intended to produce a clinically deployable tool, they provide a detailed case study of what can be achieved using public EEG data and standard ML techniques, and they highlight practical considerations for future efforts aiming to translate neurophysiological biomarkers into real-world psychiatric decision support.

## 1 Introduction

Mental and substance use disorders represent a substantial and growing public health challenge. Global estimates suggest that hundreds of millions of individuals meet criteria for a psychiatric disorder at any given time, with major depressive disorder (MDD), anxiety disorders, bipolar disorder, and schizophrenia contributing heavily to years lived with disability and premature mortality. Beyond direct symptoms such as low mood, psychosis, or cognitive impairment, these conditions affect social functioning, occupational productivity, and physical health outcomes, including elevated risk for cardiovascular disease and suicide.

Current diagnostic practice in psychiatry is largely descriptive. Clinicians synthesize information from patient interviews, collateral reports, and observational assessments, mapping reported symptoms onto categorical criteria such as those defined in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM) or the *International Classification of Dis-*

*eases* (ICD). While these diagnostic systems provide a shared vocabulary, they do not incorporate biomarkers in the way that many other areas of medicine now do. The resulting diagnoses can be heterogeneous, with substantial symptom overlap across disorders and high rates of comorbidity. Two patients with MDD may share only a subset of symptoms, and individuals with bipolar disorder, schizoaffective disorder, or schizophrenia can present with overlapping mood and psychotic features. Misdiagnosis is common, particularly early in the course of illness, and treatment decisions may be based on incomplete or ambiguous information.

Objective measures that capture underlying brain and autonomic function have therefore attracted considerable interest as potential diagnostic aids. EEG records electrical activity at the scalp with millisecond temporal resolution, enabling analysis of frequency band power, phase synchrony, and large-scale functional networks. Numerous studies have reported group differences in EEG features across a range of psychiatric conditions. For example, atypical power in theta and gamma bands, altered alpha asymmetry, and disruptions in functional connectivity have all been associated with schizophrenia and mood disorders [1, 7, 24]. HRV, derived from beat-to-beat changes in heart rate, quantifies autonomic nervous system dynamics and has been consistently linked to emotion regulation and stress responsivity. Reduced HRV is a relatively robust marker across diagnoses including depression, anxiety, and post-traumatic stress disorder [3, 4, 19, 18]. These findings suggest that neurophysiological and autonomic signals may reflect transdiagnostic dimensions of psychopathology that are not fully captured by symptoms alone.

Machine learning provides a natural toolkit for extracting structure from such high-dimensional signals. Over the past decade, classical ML algorithms—such as support vector machines, random forests, $k$-nearest neighbours, and gradient-boosted trees—have been applied to engineered EEG features to classify patients with schizophrenia, depression, bipolar disorder, and obsessive–compulsive disorder (OCD) [1, 8, 11, 15]. More recently, deep learning approaches, including convolutional and recurrent neural networks, have shown promise in learning discriminative patterns directly from raw or minimally processed EEG time series and spectrograms [9, 10, 16, 13, 14, 23]. Systematic reviews summarise this literature and generally conclude that ML and DL models can achieve good performance in differentiating patients from healthy controls, especially when combined with careful feature engineering and cross-validation [1, 2, 21, 25].

Despite this progress, several important gaps remain. First, much of the literature focuses on *binary* classifications (e.g., schizophrenia vs. control), whereas real-world clinical practice requires distinguishing among multiple disorders that share overlapping features. Multi-class classification is more challenging and has been less extensively studied. Park et al. [9], for instance, demonstrated the feasibility of multi-class classification of major psychiatric disorders using resting-state EEG in a large clinical sample, but their dataset is not publicly available and their model details are tuned to a specific institutional recording protocol. Second, many studies rely on relatively small, idiosyncratic datasets, raising concerns about overfitting and limited generalizability. Meta-analyses of deep learning in psychiatric imaging emphasise heterogeneity in sample sizes, preprocessing pipelines, network architectures, and evaluation strategies, which complicates cross-study comparison and may inflate performance estimates [2]. Third, interpretability remains a major concern. Black-box models, while accurate, offer little insight into which features drive predictions, making it difficult for clinicians to trust or meaningfully integrate algorithmic outputs into their decision-making [12, 22, 20].

Recognizing these challenges, several recent reviews call for open datasets, standardised benchmarks, and explainable AI (XAI) methods in psychiatric ML [5, 21, 25, 22]. The *EEG Psychiatric Disorders Dataset* curated on Kaggle represents one step in this direction. The dataset is derived from a larger clinical EEG study and provides quantitative resting-state features for individuals with different psychiatric diagnoses as well as healthy controls [26, 23]. Because it is publicly available, it enables reproducible experiments and direct comparison of different models. However, prior published work using this dataset has primarily focused on deep learning approaches to binary classification of specific disorders [23], leaving open questions about how classical and deep models compare within a unified evaluation framework and how much diagnostic information is accessible in a multi-class setting.

The overarching goal of this project is to contribute to this evolving landscape by developing and analysing a transparent ML pipeline for the detection of psychiatric conditions using this open EEG dataset. Specifically, we pursue the following aims:

- To implement a consistent data preprocessing and feature scaling pipeline suitable for training multiple model families on the same inputs.
- To compare the performance of classical ML algorithms and neural network architectures on both binary and multi-class diagnostic classification tasks, using stratified cross-validation and held-out test evaluation.
- To explore model interpretability by examining feature importance estimates and analysing how different frequency bands and channels contribute to specific diagnostic predictions, building on work in interpretable EEG-based classification [12, 6, 7].
- To critically discuss the limitations of such models—including dataset biases, class imbalance, and the gap between offline performance and clinical deployment—in light of current debates around trustworthy AI in mental health [22, 20, 5].

By addressing these aims, we hope to provide a detailed and reproducible case study that illustrates both the promise and the pitfalls of using open EEG data and standard ML methods to build tools for psychiatric diagnosis.

## 2 Related Work

### EEG-based classification of schizophrenia and psychotic disorders

Schizophrenia and related psychotic disorders have been particularly prominent in EEG-based ML research. Traditional EEG studies identified abnormalities in oscillatory activity, such as increased theta and gamma power and altered alpha rhythms, while more recent work has emphasised connectivity and microstate dynamics. Rahul et al. conducted a systematic review of EEG-based automated schizophrenia classification, surveying both classical ML and DL approaches over two decades of research [1]. They report that accuracies in the range of 70–95% are common when distinguishing schizophrenia patients from controls, but note that studies often vary widely in sample size, recording protocol, and feature engineering, making direct comparison challenging.

Several representative empirical studies illustrate the diversity of approaches. Cukić et al. extracted non-linear features such as Higuchi fractal dimension and sample entropy from EEG signals and used them with conventional classifiers to detect depression, demonstrating that complexity measures can capture relevant signal differences [17]. Similar non-linear features have been applied to schizophrenia, where they may reflect disrupted temporal organisation of brain activity [11]. Vázquez et al. focused on interpretable classification, using directed connectivity measures (GPDC and dDTF) as input to random forests and demonstrating that beta and theta connectivity patterns in occipital and frontal regions were particularly informative for distinguishing schizophrenia patients from controls [12]. Their work is notable in that it explicitly links model features to neurophysiological hypotheses, rather than treating accuracy as the sole outcome.

Deep learning has enabled models to ingest raw or minimally processed time series and automatically learn hierarchical representations. Shoeibi et al. proposed CNN–LSTM architectures that combine convolutional layers for local feature extraction with recurrent layers to capture temporal dynamics in EEG, achieving high accuracies on public schizophrenia datasets [16]. Bagherzadeh et al. transformed effective connectivity matrices into image-like representations and used hybrid CNN–LSTM networks to classify schizophrenia, reporting that this connectivity-based representation improved performance relative to simpler feature sets [13]. Thakkar et al. constructed time–frequency representations via short-time Fourier transform and trained a deep residual network to identify delta-band alterations in posterior regions as key biomarkers for schizophrenia [14]. Uyanik et al. review such methods more broadly, concluding that deep architectures generally outperform classical models when sufficient data are available but require careful regularisation and interpretability strategies [21].

A complementary line of work examines EEG microstates, which are short-lived, quasi-stable scalp topographies thought to reflect fundamental building blocks of spontaneous brain activity. Pacchioni et al. review microstate-based machine learning for psychotic disorders, highlighting consistent findings of altered duration and occurrence of microstate classes C and D in schizophrenia [6]. Large cohort studies suggest that these microstate abnormalities may serve as endophenotypes, present in both patients and their unaffected relatives. Although microstate analysis is not directly applied in our work, the broader insight is that temporally structured EEG features beyond simple band power carry important diagnostic information, and that ML models can be designed to exploit such structure.

### EEG and HRV markers in mood and anxiety disorders

EEG and HRV have also been extensively studied in mood and anxiety disorders, where abnormalities in affect regulation and autonomic balance are central. Liu et al. review EEG-based ML approaches for depression diagnosis, finding that many studies rely on power spectral density, frontal alpha asymmetry, and non-linear features such as entropy measures [8]. Safayari et al. used deep neural networks on resting-state EEG to classify depressed versus non-depressed participants, demonstrating that architectures with sufficient depth and careful regularisation can achieve strong performance even with relatively small datasets [10]. Cukić et al. while primarily focusing on depression, illustrate how fractal and entropy features can be combined with ML to capture subtle changes in signal complexity [17]; similar techniques have been explored for bipolar disorder and OCD [15].

HRV offers a complementary perspective on mood and anxiety disorders by quantifying autonomic nervous system activity. Jung et al. summarise evidence that reduced HRV, particularly in the high-frequency band, is associated with a range of psychiatric conditions, consistent with the notion of diminished vagal tone and impaired emotion regulation [3]. Ramesh et al. provide a systematic review of HRV in psychiatric populations, concluding that although effect sizes vary, HRV reductions are relatively robust in depression and several anxiety disorders [4]. Weber et al. extend this line of work to large population cohorts, using ECG-derived HRV profiles to identify individuals at elevated psychiatric risk [19]. Lee et al. explicitly investigate the relationship between HRV and EEG power in depressive and anxiety disorders, finding that specific patterns of heart–brain coupling emerge during cognitive tasks [18]. These findings motivate multimodal ML models that integrate EEG and HRV, although most ML studies to date have considered each modality separately.

### Multi-class and multimodal classification of psychiatric conditions

Moving beyond single-disorder vs. control contrasts, multi-class classification more closely mirrors the diagnostic challenges faced in clinical practice. Park et al. analysed resting-state EEG from 945 subjects and used machine learning to distinguish between several major psychiatric disorders and healthy controls, demonstrating that multi-class classification

is feasible but typically yields lower accuracies than binary setups [9]. They found that performance varied by disorder, with some diagnoses being easier to distinguish than others, likely reflecting differences in underlying neurophysiological signatures and symptom overlap.

Ahmed et al. introduced a deep learning framework for classifying psychiatric disorders using a large quantitative EEG dataset and later served as the basis for the Kaggle *EEG Psychiatric Disorders Dataset* [23, 26]. Their models achieve strong performance on multiple binary classification tasks and illustrate how publicly released data can catalyse further methodological work. Sarisik et al. used a different large cohort to compare EEG signatures of schizophrenia, depression, and healthy controls, highlighting both shared and disorder-specific spectral alterations, as well as the importance of controlling for age and medication status when training classifiers [7]. Kang et al. propose high-order brain network features for first-episode schizophrenia, pointing to the potential of graph-based representations for capturing subtle connectivity differences [24]. Together, these studies demonstrate that multi-class and multimodal classification are technically feasible, but they also underscore the importance of robust evaluation and careful handling of confounds.

Comprehensive reviews by Quaak et al. and Baydili et al. place these efforts in the broader context of AI for psychiatric diagnostics [2, 5]. Quaak et al. focus on deep learning applications across various neuroimaging modalities and emphasise heterogeneity in model architectures, data quality, and validation practices, calling for standardised benchmarks and reporting guidelines [2]. Baydili et al. survey AI methods applied to biological signals in psychiatry, including EEG and HRV, highlighting both technical challenges (e.g., noise, non-stationarity) and conceptual questions about how to align findings with clinical constructs [5]. Uyanik et al. similarly review automated detection of neurological and mental health disorders from EEG and argue that while ML offers considerable promise, reproducibility and dataset bias remain major concerns [21].

### Explainable AI and open datasets in psychiatric machine learning

Explainability and transparency have emerged as central requirements for deploying ML systems in clinical settings. Joyce et al. review explainable AI methods in mental health and argue that clinicians require more than global measures of model performance; they need case-level explanations, well-calibrated uncertainty estimates, and interfaces that support collaborative decision-making [22]. Ali et al. broaden this perspective, discussing issues of fairness, bias, and accountability in AI for mental health and emphasising the need for participatory design involving clinicians and patients [20]. In the specific context of EEG-based diagnosis, Vázquez et al. demonstrate how interpretable models built on connectivity features can reveal neurobiologically plausible patterns, thereby increasing trust in model outputs [12]. Zhao et al. in their review of multimodal EEG ML applications, similarly advocate for models that balance predictive accuracy with interpretability and clinical relevance [25].

Open datasets are a key enabler of this research ecosystem. The *EEG Psychiatric Disorders Dataset* on Kaggle, created by Dhekane based on the quantitative EEG study analysed by Ahmed et al., provides summary features for hundreds of individuals with various psychiatric diagnoses and healthy controls [26, 23]. Because the data and metadata are publicly available, researchers can develop and compare models under consistent conditions, facilitating replication and incremental improvement. Our work leverages this dataset to explore multi-class classification and model interpretability, aligning with broader efforts to build trustworthy, tools that complement—rather than replace—clinical expertise [5, 21, 22].

# 3 Methodology

## 3.1 Dataset

All analyses use the *EEG Psychiatric Disorders* dataset released on Kaggle [26], which in turn is derived from the quantitative EEG study of Ahmed *et al.* [23]. The raw file `EEG.machinelearing_data_BRMH.csv` contains resting state quantitative EEG features for adult participants recruited at a single site. Each row corresponds to one subject and includes demographic variables, summary EEG features and a primary diagnostic label. The dataset has 945 rows and 1149 columns, as reported in the first loading step of the notebook.

The main diagnostic label is stored in the column `main.disorder`. After stripping leading and trailing whitespace, the following seven categories are present, with observed counts from the notebook preprocessing step:

- Mood disorder (266)
- Addictive disorder (186)
- Trauma and stress related disorder (128)
- Schizophrenia (117)
- Anxiety disorder (107)
- Healthy control (95)
- Obsessive compulsive disorder (46)

These categories cover a broad range of common psychiatric conditions and a healthy comparison group. In line with prior EEG classification work that often starts with a binary distinction between patients and controls [1, 8, 9, 21], I derived a binary label `is_patient` that equals one for any row whose `main.disorder` is not "Healthy control", and zero otherwise.

## 3.2 Preprocessing and feature selection

The preprocessing step follows the pipeline implemented in the second notebook cell. First, duplicated rows are removed and any occurrences of positive or negative infinity are replaced with missing values. The code then identifies all numeric columns using `pandas` type inspection. A small set of numeric metadata columns is treated separately and not used as EEG features. This set includes the subject index `no.`, age, IQ, years of education and the binary label `is_patient`. All remaining numeric columns are treated as candidate EEG features.

Before modeling, the script removes features with a large fraction of missing values and features that are essentially constant. Specifically, any EEG feature with more than forty percent missing entries is dropped, and any column with at most one non-missing value is removed. The notebook output shows that this results in 1 140 usable EEG features, down from 1 141 numeric candidates, so only one high-missing column was discarded and no feature was flagged as constant. This is consistent with the fact that the dataset is already relatively clean in its published form [26, 23].

All models operate on this cleaned feature matrix. For exploratory visualization, the notebook applies median imputation with `SimpleImputer` followed by standardization with `StandardScaler`. A principal components analysis (PCA) is then fitted with up to 100 components, limited by the number of samples and features. The corresponding notebook printout reports that 100 components explain about 95.3 % of the total variance. PCA is used in two ways. First, it provides a low dimensional representation of the whole dataset for exploratory plots such as the cumulative variance curve, a correlation heatmap of high variance features and a t-SNE map of the first few components. Second, within the modeling pipelines, PCA acts as a compact linear transform that reduces the high dimensional quantitative EEG space to a more manageable representation, which is important when working with more than one thousand correlated variables [2, 5, 21].

## 3.3 Train–test split

To evaluate performance under realistic generalization conditions, the notebook splits the data into disjoint training and test sets. For all supervised models, the feature matrix is defined as $\mathbf{X} =$ `df[eeg_cols]`. The target vector is either the binary label `is_patient` or a one versus rest label derived from `main.disorder` as described below. The split is carried out using `train_test_split` with an 80 % training fraction and a 20 % test fraction, stratified on the label and with a fixed random seed of 42. For the binary setting, this yields 756 training subjects and 189 test subjects, with the same case control ratio in both sets.

All preprocessing steps that depend on the feature distribution, namely imputation, standardization and PCA, are contained inside scikit-learn `Pipeline` objects. The transformations are therefore fitted only on the training folds within cross validation and on the training split before final model evaluation. This avoids the common error of leaking information from the test set into model fitting [2, 21].

## 3.4 Binary classification pipeline

The first predictive task is to distinguish any psychiatric patient from a healthy control using the binary label `is_patient`. The modeling pipeline has four stages: median imputation, feature scaling, PCA and a final classifier. The imputer replaces missing values with the median of each feature in the training subset. The scaler standardizes each feature to zero mean and unit variance. The PCA step then projects the standardized features into a lower dimensional space with at most 100 components, as defined in the notebook. The number of components is also capped by the number of training samples and by the original feature dimension.

On top of this common preprocessing, a suite of classifiers is considered. The candidate models include logistic regression, a linear support vector machine, a radial basis function kernel SVM, $k$ nearest neighbours, a random forest, an extra trees ensemble, gradient boosting, histogram based gradient boosting and a multi layer perceptron. When the relevant libraries are available, tree boosting methods based on XGBoost, LightGBM and CatBoost are also included. Each classifier is given a modest but nontrivial hyperparameter grid. For example, the logistic regression explores three values of the regularization strength $C$, the linear SVM searches over three $C$ values, the random forest varies the number of trees, the maximum depth and the minimum number of samples per leaf, and the neural network adjusts the $\ell_2$ penalty. For the tree based models that support class weights, the code uses the built in “balanced” option so that the minority class (healthy controls) receives greater weight in the loss.

The model selection procedure is implemented as a series of grid searches. For each candidate model, the script clones the base preprocessing pipeline, inserts the classifier and runs a `GridSearchCV` with five fold stratified cross validation and ROC AUC as the scoring function. Any model that fails to converge or throws an error is skipped and assigned a missing score, which prevents one problematic search from terminating the entire run. The notebook output lists the best cross validated ROC AUC for each candidate. In this run, the tuned random forest achieves the highest mean cross validated ROC AUC of about 0.82, slightly ahead of extra trees and clearly ahead of linear models and the multilayer perceptron. Tree boosters such as XGBoost, LightGBM and CatBoost perform reasonably well but do not overtake the random forest in this particular configuration.

Given this result, the notebook performs a second focused tuning of the random forest with a two stage search. A randomized search explores a slightly wider hyperparameter space, followed by a narrower grid search around the most promising region. Both stages use five fold stratified cross validation and ROC AUC as the objective. The best estimator from this search, referred to as `best_rf_pipeline`, combines the same preprocessing steps with a random forest classifier tuned for AUC. This pipeline is then evaluated once,

without further adjustment, on the held out test set.

### 3.5 One versus rest models for individual diagnoses

The second set of experiments examines whether the EEG features contain enough information to separate individual diagnostic groups from all other subjects. Rather than training a single multi class model, the notebook adopts a series of one versus rest classifiers, which is a standard tactic in medical machine learning when classes are imbalanced and some diagnoses are rare [2, 1]. For each unique value of `main.disorder`, a binary label $y^{(c)}$ is defined that equals one for rows with that diagnosis and zero otherwise.

To avoid extremely unstable estimates, the helper function in Step 18 skips any condition with fewer than five positive examples or fewer than five negatives. In this dataset, all seven diagnostic categories meet this minimum. For each condition, the script calls a helper that repeats the same model selection procedure as in the patient versus control setting. The features are again median imputed, standardized and projected into PCA space inside a pipeline. The candidate model family is similar, although in this revised version the code avoids additional package installation and will gracefully exclude boosters that are not available in the environment.

For each condition, the helper splits the data into an 80 % training set and a 20 % test set stratified on the one versus rest label. It then performs cross validated grid search over the classifier hyperparameters using ROC AUC as the selection metric. The name of the best performing classifier, the fitted pipeline and the corresponding cross validated AUC are returned. A separate evaluation function then applies this fitted pipeline to the held out test split and computes a set of metrics, including accuracy, precision, recall, $F_1$ score, balanced accuracy, ROC AUC and average precision. The classification report is also stored in case more detailed analysis is needed.

The resulting per condition metrics are collected into a `DataFrame` and printed in the notebook. This summary table is used directly in the Results section.

### 3.6 Evaluation metrics

The choice of evaluation metrics is guided by common practice in psychiatric EEG classification [1, 8, 9, 2, 21]. Because the positive and negative classes are usually imbalanced, the notebook reports both threshold dependent and threshold independent measures.

Accuracy, precision, recall, $F_1$ score and balanced accuracy are computed from the predicted labels at the default decision threshold of 0.5. Accuracy measures the proportion of correct predictions, precision measures the fraction of predicted positives that are true positives, recall measures the fraction of actual positives that are correctly identified and $F_1$ is their harmonic mean. Balanced accuracy is the mean of sensitivity and specificity and is more informative than accuracy when class prevalence is skewed.

To capture the quality of the ranking produced by predicted probabilities, the models also report the area under the receiver operating characteristic curve (ROC AUC) and the average precision, which is the area under the precision–recall curve. ROC AUC is less sensitive to class imbalance but can be optimistic in heavily skewed datasets. Precision–recall curves and average precision are often recommended for rare event detection because they focus attention on the positive class [1, 2]. In addition to numeric scores, the notebook generates confusion matrices, ROC curves and precision–recall curves for the tuned random forest and for the top one versus rest models.

### 3.7 Model interpretation

Although the present project is exploratory, interpretability is an important consideration for any future clinical application. In line with the broader literature on explainable AI in mental health [22, 20, 12, 25, 5], the notebook includes two types of inspections.

First, it uses PCA loadings to relate the model space back to the original EEG features. The tuned random forest pipeline exposes its PCA transform, which allows the projection of test set samples into principal component space. The code constructs two dimensional plots of the first two principal components and colours points by their true label and prediction outcome (true positive, true negative, false positive or false negative). This provides a geometric sense of how the model decision boundary sits in the reduced feature space.

Second, the notebook uses SHAP values to estimate feature contributions at the level of principal components. For the patient versus control classifier and for three selected one versus rest models (healthy control, mood disorder and addictive disorder), it calls `shap.TreeExplainer` on the tree based pipelines and computes SHAP values over the PCA transformed inputs. Summary bar plots and beeswarm plots are produced for each setting. A separate analysis cell then maps the principal components that receive the largest SHAP importance back to their original EEG features by examining PCA loadings, and visualizes these loadings for selected components. This interpretability layer doesn't make strong neurophysiological claims, but it provides an initial view of which groups of EEG features are most influential for the fitted classifiers, in line with previous interpretable EEG studies [12, 15, 6].

## 4 Results

### 4.1 Descriptive statistics and exploratory analysis

Psychiatric cases are shown to be the most prevalent. In the binary labeling scheme, 850 of 945 subjects are patients and 95 are healthy controls, which means only about ten percent of the sample is labeled as control. This class imbalance is important for interpreting all subsequent results, since a classifier that always predicts "patient" already achieves high nominal accuracy but has no clinical usefulness.

Exploratory plots based on the PCA representation provide a compressed view of the feature space. The cumulative variance curve shows that the first 20 components account for a substantial fraction of the variance and that 100 components reach approximately 95 % explained variance. A heatmap of pairwise correlations among the 30 highest variance features suggests that the quantitative EEG measures are strongly correlated in blocks that likely reflect shared frequency bands and scalp regions, which is expected for spectral and connectivity features derived from the same underlying recordings [7, 23]. A t-SNE map of the first 30 principal components illustrates the global structure of the sample in two dimensions. Although some local clustering is visible, there is no clear separation between patients and controls in this two dimensional embedding, which is consistent with the moderate ROC AUC obtained by the downstream classifiers. Histograms of the first principal component stratified by the binary label likewise show overlapping distributions rather than a simple shift.
*Supporting figures for this subsection are provided in Appendix 6.*

## 4.2 Patient versus control classification

The model selection stage compares a range of classifiers under the same preprocessing pipeline. The notebook output from the binary grid search reports the following approximate cross validated ROC AUC values on the training data. Logistic regression reaches about 0.69, the linear SVM around 0.68, the radial basis function SVM about 0.79 and $k$ nearest neighbours around 0.74. Among the tree based models, the random forest attains a mean cross validated ROC AUC of 0.815, extra trees about 0.813 and gradient boosting approximately 0.74. The histogram based gradient boosting model, the multilayer perceptron and the three boosting libraries (XGBoost, LightGBM and CatBoost) achieve mean AUC values between roughly 0.67 and 0.78. The highest observed cross validated AUC is obtained by the random forest with 800 trees, no maximum depth and a minimum of three samples per leaf.

The tuned random forest pipeline is then evaluated once on the held out test set of 189 subjects. At the default probability threshold of 0.5, the model predicts the patient class for every single test subject. As a result, the overall accuracy is 0.8995, the precision for the patient class is 0.8995 and the recall for the patient class is 1.0. The $\mathrm{F}_1$ score for the patient class is 0.9471. However, the recall for the control class is zero, since none of the 19 controls in the test set are correctly identified. The confusion matrix therefore has 170 true positives, 19 false positives, zero true negatives and zero false negatives. The balanced accuracy, which averages sensitivity and specificity, is exactly 0.5, indicating chance level performance when both classes are weighted equally.

Threshold independent metrics tell a slightly different story. The ROC AUC of the tuned random forest on the test set is approximately 0.64, and the average precision is roughly 0.94. The ROC curve lies above the diagonal but not by a large margin, which indicates that the model produces a probability ranking that is somewhat informative but far from perfect. The high average precision is driven in part by the high prevalence of patients in the test set. The precision–recall curve starts at high precision for the top ranked predictions but quickly drops as more subjects are considered. From a clinical perspective, the main conclusion is that the binary classifier distinguishes patients from controls better than chance, yet its performance is very sensitive to the choice of decision threshold and the severe class imbalance.
*Supporting figures for this subsection (e.g., ROC, PR, confusion matrix) are provided in Appendix 6.*

## 4.3 One versus rest performance by diagnosis

The one versus rest analysis shows that the difficulty of prediction varies markedly across diagnoses. Table 1 summarizes the results printed by the notebook helper for the seven conditions that met the minimum sample size requirement. Each row reports the best model family selected by cross validated ROC AUC, the cross validated AUC on the training data and three test set metrics, namely ROC AUC, $\mathrm{F}_1$ score for the positive class and average precision.

Table 1: One versus rest performance per diagnostic group on the held out test sets. $\mathrm{F}_1$ is the positive class $\mathrm{F}_1$ score at the default threshold.

| Condition | Best model | CV AUC | Test ROC AUC |
|---|---|---|---|
| Healthy control | ExtraTrees | 0.81 | 0.76 |
| Addictive disorder | XGBoost | 0.65 | 0.70 |
| Mood disorder | Gradient Boosting | 0.56 | 0.58 |
| Obsessive compulsive | Random Forest | 0.65 | 0.58 |
| Schizophrenia | $k$NN | 0.56 | 0.47 |
| Trauma and stress | $k$NN | 0.57 | 0.45 |
| Anxiety disorder | XGBoost | 0.54 | 0.45 |

Two patterns are immediately apparent. First, the cross validated AUC values on the training data tend to be higher than the corresponding test set AUC values, which suggests some degree of overfitting. This gap is particularly visible for the addictive disorder and obsessive compulsive disorder models. Second, many of the models achieve moderate test ROC AUCs in the range between 0.45 and 0.70 but have very low $\mathrm{F}_1$ scores on the positive class at the default decision threshold. In the healthy control and obsessive compulsive disorder settings, the best models never predict the positive class on the test set, which leads to an $\mathrm{F}_1$ score of zero despite nontrivial ROC AUC and average precision values. In the mood and addictive disorder settings, the models do make some positive predictions, but the recall remains low relative to the number of actual cases.

The average precision values from the notebook output reflect a similar compromise. For example, the addictive disorder model reaches an average precision of about 0.38, and the mood disorder model about 0.41, which indicates that the top ranked subjects are enriched for the positive class relative to chance but that a large fraction of cases would still be missed at practical thresholds. Models for schizophrenia, trauma and stress related disorder and anxiety disorder have lower ROC AUCs and modest average precision, which is consistent with

the finding that their per condition EEG signatures often overlap substantially with those of other disorders [7, 1, 2].
*Supporting figures for this subsection (per-condition curves and diagnostics) are provided in Appendix 6.*

### 4.4 Interpretability outputs

The SHAP analyses are primarily exploratory. For the patient versus control classifier, summary plots of SHAP values on the PCA components highlight a subset of components with larger contributions, while many components contribute little at the chosen operating point. The mapping from PCA loadings back to original EEG variables shows that these influential components load on combinations of spectral features distributed across multiple electrodes, which is consistent with the distributed nature of EEG alterations reported in previous psychiatric studies [1, 8, 6]. For the per condition models, SHAP on PCA space is computed for healthy control, mood disorder and addictive disorder, and bar plots of component importances are produced. These visualizations do not yet support strong neurobiological conclusions but demonstrate that the proposed pipeline can be connected to standard explainable AI tools [12, 22, 20].
*Supporting figures for this subsection (e.g., PCA scatter, SHAP plots, component loadings) are provided in Appendix 6.*

## 5 Discussion

The experiments in this project illustrate both the promise and the challenges of using quantitative EEG features to support psychiatric diagnosis. In the aggregate patient versus control setting, a reasonably tuned random forest achieves a mean cross validated ROC AUC of about 0.82 on the training data, which indicates that there is genuine discriminative signal in the features. However, when the model is evaluated on a held out test set, the ROC AUC drops to about 0.64 and the optimal tree ensemble collapses to a decision rule that predicts "patient" for every subject at the default threshold. This behavior reflects the severe class imbalance in the dataset. Because patients account for nearly ninety percent of the sample, the model can maximize accuracy and even maintain reasonable precision and recall for the majority class by ignoring the minority class entirely.

This pattern echoes concerns raised in recent reviews about evaluation practices in EEG based psychiatric classification. Rahul *et al.* [1] and Quaak *et al.* [2] note that many published studies report high cross validated accuracies on small or imbalanced samples without a separate test set, and that these results may not generalize to new patients. The gap between the cross validated performance and the held out test performance in this study underscores this point. It also highlights the importance of reporting metrics such as balanced accuracy, class specific recall and precision–recall curves rather than relying on overall accuracy alone, a point that has been emphasized in broader work on explainable and trustworthy AI for mental health [22, 20].

The one versus rest analyses for individual diagnoses further emphasize the difficulty of fine grained prediction. Although some conditions, such as addictive disorder and mood disorder, yield moderate test ROC AUC values around 0.70 and 0.58, the actual $F_1$ scores for the positive class at the default decision threshold are low. For several diagnoses, including healthy control, obsessive compulsive disorder and trauma and stress related disorder, the tuned models never predict the positive class on the test set, even though their ROC AUCs suggest that the probability ranking contains useful information. This mismatch between ranking quality and thresholded predictions is a direct consequence of the combination of class imbalance and a default threshold that is not tailored to the clinical cost structure. In practice, one would need to adjust the decision threshold or use alternative decision rules to trade sensitivity against specificity, especially when the cost of missing a true case is high.

The heterogeneity of performance across diagnoses is also informative. The fact that addictive disorder and mood disorder reach somewhat higher ROC AUCs than anxiety disorder, schizophrenia and trauma and stress related disorder may reflect a combination of sample size differences and genuine variation in how strongly these conditions affect resting state quantitative EEG measures [7, 1, 8, 23]. However, given the modest size of the dataset and the reliance on precomputed summary features rather than raw waveforms, it would be premature to interpret these differences too strongly. Moreover, co-occurring conditions, medication effects and comorbid physical illnesses are not explicitly modeled here, even though they are known to influence both EEG and autonomic signals [3, 4, 19, 18].

Another limitation is the representation of the EEG itself. The Kaggle dataset supplies a rich set of spectral and connectivity features, but it does not expose the underlying raw time series. Deep learning approaches that operate directly on raw or minimally processed EEG have achieved impressive results in some settings [16, 13, 14, 23], particularly when trained on large cohorts. However, toolkits for deep learning on EEG also require careful regularization and data augmentation to avoid overfitting [2, 21]. In this project, the choice to work with precomputed features and relatively classical models was deliberate. It makes the pipeline easier to reproduce and interpret, but it may leave performance improvements on the table.

The interpretability analyses, while preliminary, show that the proposed pipeline can be made more transparent. By expressing the models in PCA space and applying SHAP at the component level, it becomes possible to identify which combinations of EEG features contribute most to the predictions. Mapping these components back to the original feature names provides a path toward linking machine learning outputs to more traditional neurophysiological findings, as in the work of Vázquez *et al.* on connectivity based schizophrenia markers [12] or the microstate based analyses reviewed by Pacchioni *et al.* [6]. To move beyond descriptive plots, future work would need to systematically compare these component level patterns across disorders and relate them to hypotheses about network level dysfunction.

Finally, it is important to note that this study does not attempt to replace clinical judgment or structured diagnostic interviews. Instead, it explores whether machine learning models trained on quantitative EEG features from an open dataset can recover meaningful structure in psychiatric diagnoses. The results suggest that there is signal worth pursuing, particularly for broad distinctions between patients and controls and for certain diagnostic contrasts, but they also reveal that performance is fragile when evaluated under realistic imbalance and held out test conditions. This aligns with the conclusions of recent surveys on AI in psychiatry, which outline open challenges [5, 20, 21].

# 6 Conclusion

This project developed and evaluated a machine learning pipeline for predicting psychiatric diagnoses from quantitative EEG features in an open dataset. After standard preprocessing, dimensionality reduction with PCA and extensive model selection, a tuned random forest achieved a cross validated ROC AUC of roughly 0.82 in distinguishing psychiatric patients from healthy controls. On a separate test set, the same model reached a ROC AUC of about 0.64 and high accuracy driven largely by the predominance of the patient class, with no correct identifications of controls at the default decision threshold. One versus rest models for individual diagnoses produced moderate ROC AUC values for some conditions, but $F_1$ scores for the positive class remained low, again reflecting class imbalance and limited sample size.

Taken together, these findings indicate that resting state quantitative EEG does contain information that is informative about psychiatric diagnosis, but that extracting clinically useful predictions is a nontrivial problem. Performance is strongly influenced by evaluation choices, such as how the data are split and which metrics are emphasized, and by the prevalence of each diagnosis. The work also demonstrates that even when using precomputed features and classical models, it is possible to connect fitted classifiers to interpretable summaries of feature importance through PCA loadings and SHAP values.

Several concrete directions follow from this exploratory study. One is to move from precomputed features back to raw EEG signals and to explore modern deep learning architectures under careful regularization, as in recent schizophrenia and depression work [16, 13, 14, 23]. Another is to integrate EEG with other biological signals, such as heart rate variability derived from ECG, which has shown promise as a transdiagnostic marker of autonomic regulation [3, 4, 19, 18]. A third direction is to design prospective studies in which models are trained on one cohort and evaluated on a truly independent sample, ideally across sites, to test robustness in more realistic settings [2, 1, 21]. Finally, future work should continue to incorporate explainability and fairness considerations from the outset, so that any eventual clinical decision support tools can be transparent, accountable and aligned with the needs of patients and clinicians [22, 20, 5].

Within these limitations, this study shows how an openly available EEG dataset can be used to build, benchmark and interpret a set of psychiatric classification models. The hope is that this kind of experimentation will serve as a bridge between the physiological characterizations available in modern neuroscience and the practical constraints of everyday psychiatric assessment.

# References


[1] Rahul, J., Sharma, D., Sharma, L.D., Nanda, U., & Sarkar, A.K. (2024). *A systematic review of EEG-based automated schizophrenia classification through machine learning and deep learning*. *Frontiers in Human Neuroscience*. `https://pmc.ncbi.nlm.nih.gov/articles/PMC10899326/`

[2] Quaak, M., van de Mortel, L., Thomas, R.M., & van Wingen, G. (2021). *Deep learning applications for classification of psychiatric disorders using neuroimaging data: systematic review and meta-analysis*. *Neuroscience & Biobehavioral Reviews*, 127, 659–673.

[3] Jung, W., et al. (2019). *Heart and brain interaction of psychiatric illness: A review*. *Frontiers in Psychiatry*, 10, 834.

[4] Ramesh, A., et al. (2023). *Heart rate variability in psychiatric disorders: A systematic review*. *Frontiers in Psychiatry*, 14, 112–119.

[5] Baydili, İ., et al. (2025). *Artificial intelligence in psychiatry: A review of biological signal-based methods*. *Diagnostics*, 15(2), 300.

[6] Pacchioni, F., et al. (2025). *Navigating the complexity of psychotic disorders: EEG microstate-based machine learning approaches*. *Diagnostics*, 15(5), 450.

[7] Sarisik, E., et al. (2024). *EEG-based signatures of schizophrenia and depression: A large-scale comparison*. *Frontiers in Neuroscience*, 18, 12061654.

[8] Liu, Y., et al. (2022). *Machine learning approaches for diagnosing depression using EEG*. *Frontiers in Psychiatry*, 13, 9375981.

[9] Park, S.M., et al. (2021). *Identification of major psychiatric disorders from resting-state EEG using a machine-learning approach*. *Frontiers in Psychiatry*, 12, 707581.

[10] Safayari, A., et al. (2021). *Depression diagnosis by deep learning using EEG signals*. *Biomedical Signal Processing and Control*, 68, 102744.

[11] de Miras, J.R., et al. (2023). *Schizophrenia classification using machine learning on EEG*. *Computers in Biology and Medicine*, 162, 107014.

[12] Vázquez, M.A., et al. (2021). *Interpretable machine learning for schizophrenia diagnosis using EEGs*. *Frontiers in Systems Neuroscience*, 15, 652662.

[13] Bagherzadeh, S., et al. (2022). *Detection of schizophrenia using hybrid deep learning on multichannel EEG*. *Computers in Biology and Medicine*, 146, 105731.

[14] Thakkar, H., et al. (2025). *Deep learning-based identification of EEG biomarkers for schizophrenia detection*. *Cognitive Neurodynamics*, 19(1), 215–230.

[15] Luján, M.Á., et al. (2022). *EEG-based schizophrenia and bipolar disorder classification using entropy and statistical features*. *Journal of Biomedical Engineering and Biosciences*, 9(2), 001–009.

[16] Shoeibi, A., et al. (2021). *Automatic diagnosis of schizophrenia in EEG signals using CNN-LSTM models*. *Frontiers in Neuroinformatics*, 15, 735.

[17] Cukić, M., et al. (2018). *EEG machine learning with Higuchi fractal dimension and sample entropy as features for depression detection*. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(8), 1586–1596.

[18] Lee, D., et al. (2022). *Associations between heart rate variability and brain EEG in depressive and anxiety disorders*. *Brain Sciences*, 12(2), 172.

[19] Weber, S., et al. (2025). *Electrocardiography-derived autonomic nervous system profiles in at-risk psychiatric cohorts*. *Nature Mental Health*, 2(1), 30–42.

[20] Ali, M., et al. (2025). *AI for mental health: A narrative review*. *Frontiers in Digital Health*, 3, 1215.

[21] Uyanik, H., et al. (2025). *Automated detection of neurological and mental health disorders from EEG: A review*. *WIREs Data Mining and Knowledge Discovery*, 15(3), e1508.

[22] Joyce, D.W., et al. (2023). *Explainable artificial intelligence for mental health: A systematic review*. *npj Digital Medicine*, 6, 212.

[23] Ahmed, Z., et al. (2024). *Psychiatric disorders from EEG signals through deep learning*. *Artificial Intelligence in Medicine*, 150, 102356.

[24] Kang, Y., et al. (2024). *High-order brain network feature extraction for first-episode schizophrenia*. *Frontiers in Psychiatry*, 15, 441.

[25] Zhao, S., et al. (2024). *Systematic review of machine learning for multimodal EEG data in clinical applications*. *Frontiers in Neuroscience*, 18, 708.

[26] Dhekane, S. (2022). *EEG Psychiatric Disorders Dataset*. *Kaggle*. https://www.kaggle.com/datasets/shashwatwork/eeg-psychiatric-disorders-dataset

[27] Michel, C.M., & Koenig, T. (2018). *EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks*. *Brain Topography*, 31(1), 95–99. doi:10.1007/s10548-017-0627-8

[28] Thibodeau, R., Jorgensen, R.S., & Kim, S. (2006). *Depression, anxiety, and resting frontal EEG asymmetry: A meta-analytic review*. *Journal of Abnormal Psychology*, 115(4), 715–729. doi:10.1037/0021-843X.115.4.715

[29] Stam, C.J. (2014). *Modern network science of neurological disorders*. *Human Brain Mapping*, 35(10), 4595–4620. doi:10.1002/hbm.22531

[30] Delorme, A., & Makeig, S. (2004). *EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis*. *Journal of Neuroscience Methods*, 134(1), 9–21. doi:10.1016/j.jneumeth.2003.10.009

[31] Lundberg, S.M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. In *Advances in Neural Information Processing Systems* (NeurIPS), 4765–4774. https://arxiv.org/abs/1705.07874

[32] van der Maaten, L., & Hinton, G. (2008). *Visualizing Data using t-SNE*. *Journal of Machine Learning Research*, 9, 2579–2605. https://www.jmlr.org/papers/v09/vandermaaten08a.html

[33] Jolliffe, I.T., & Cadima, J. (2016). *Principal component analysis: a review and recent developments*. *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202. doi:10.1098/rsta.2015.0202

[34] Saito, T., & Rehmsmeier, M. (2015). *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets*. *PLOS ONE*, 10(3), e0118432. doi:10.1371/journal.pone.0118432

[35] He, H., & Garcia, E.A. (2009). *Learning from imbalanced data*. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. doi:10.1109/TKDE.2008.239

[36] Breiman, L. (2001). *Random Forests*. *Machine Learning*, 45(1), 5–32. doi:10.1023/A:1010933404324

[37] Friedman, J.H. (2001). *Greedy function approximation: A gradient boosting machine*. *Annals of Statistics*, 29(5), 1189–1232. doi:10.1214/aos/1013203451

[38] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research*, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

[39] Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD), 785–794. doi:10.1145/2939672.2939785

[40] Ke, G., Meng, Q., Finley, T., et al. (2017). *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. In *Advances in Neural Information Processing Systems* (NeurIPS), 3146–3154. https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[41] Dorogush, A.V., Ershov, V., & Gulin, A. (2018). *CatBoost: unbiased boosting with categorical features*. *arXiv preprint* arXiv:1810.11363. https://arxiv.org/abs/1810.11363

# Appendix: Experimental Figures

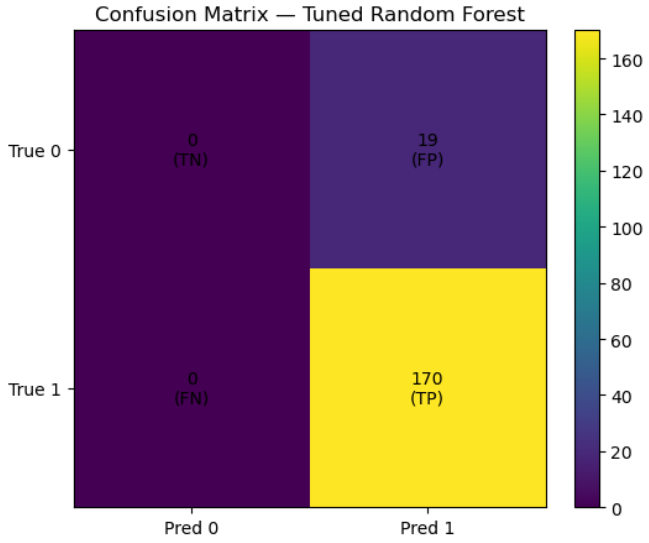



Figure 1: `Code2_11_0.png`

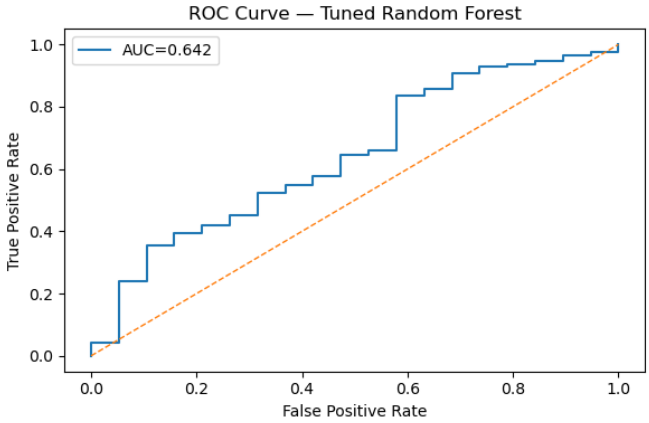



Figure 2: `Code2_12_0.png`

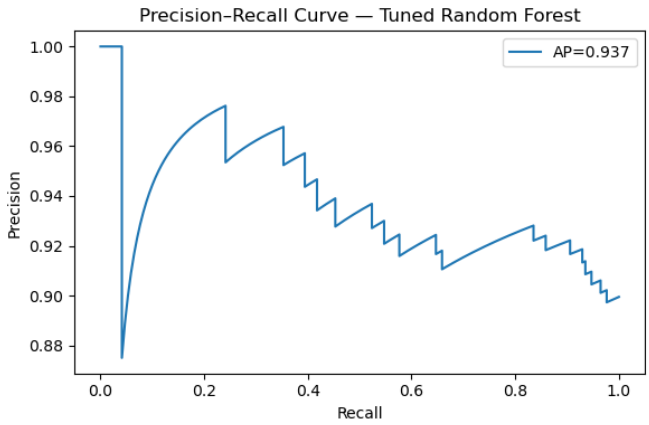



Figure 3: `Code2_12_1.png`

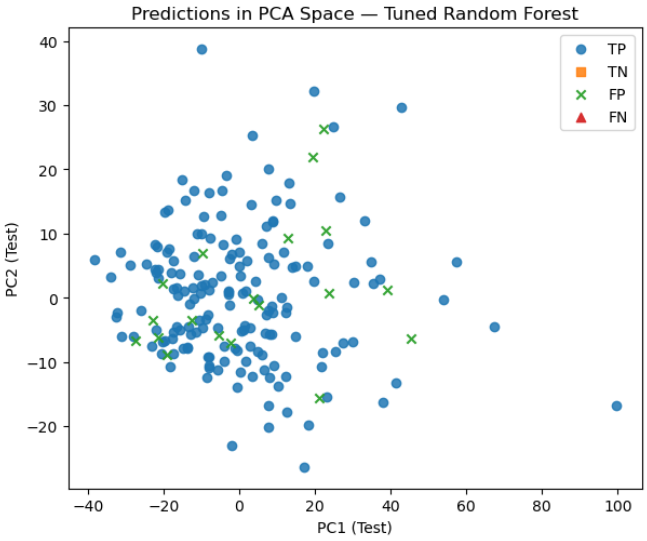



Figure 4: `Code2_13_0.png`

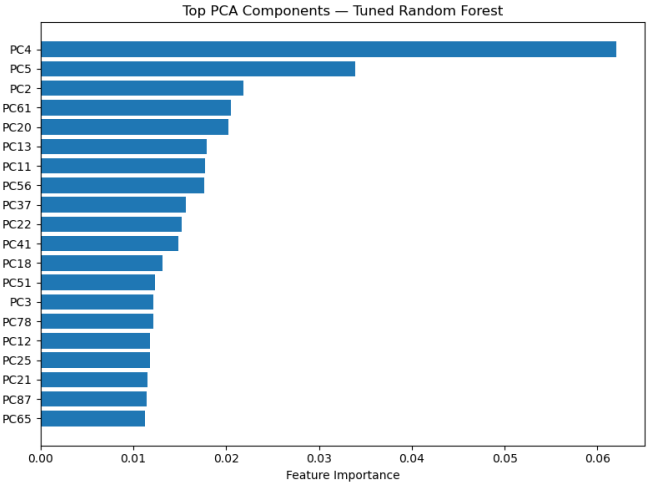
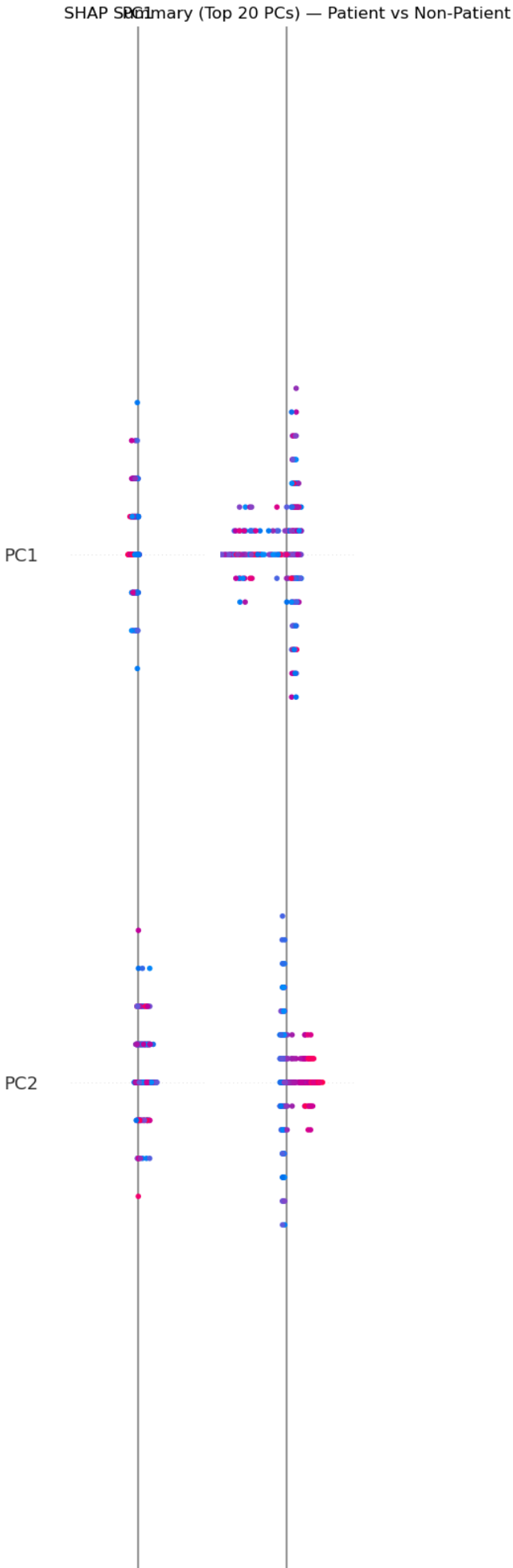
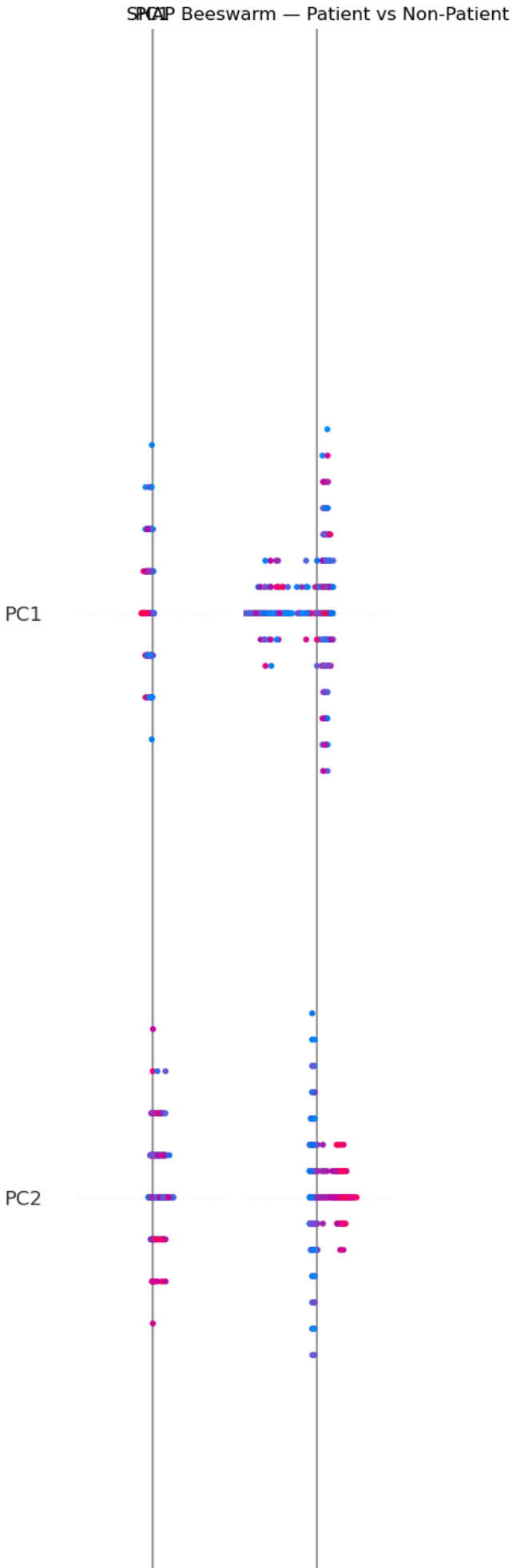



Figure 5: `Code2_14_0.png`





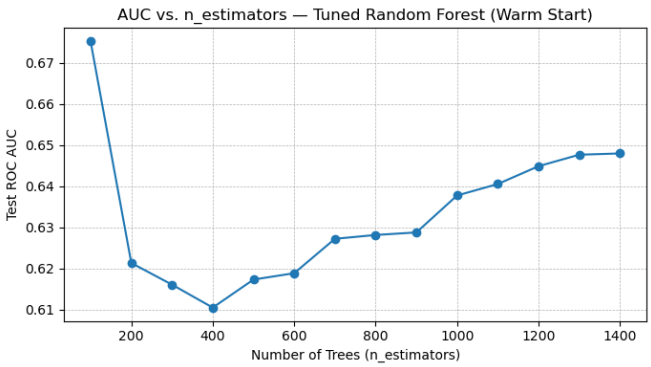




Figure 8: `Code2_16_0.png`

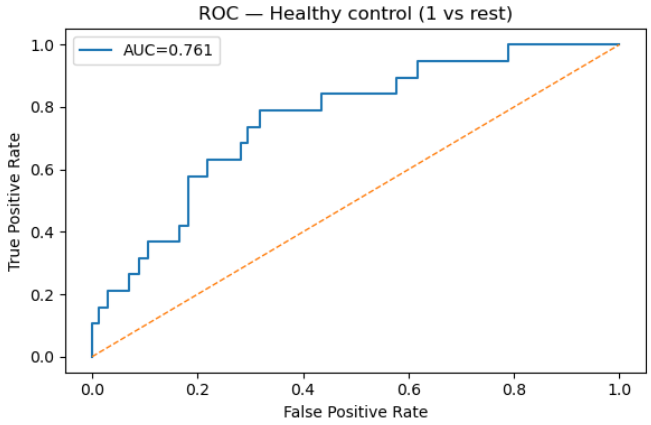



Figure 9: `Code2_19_0.png`

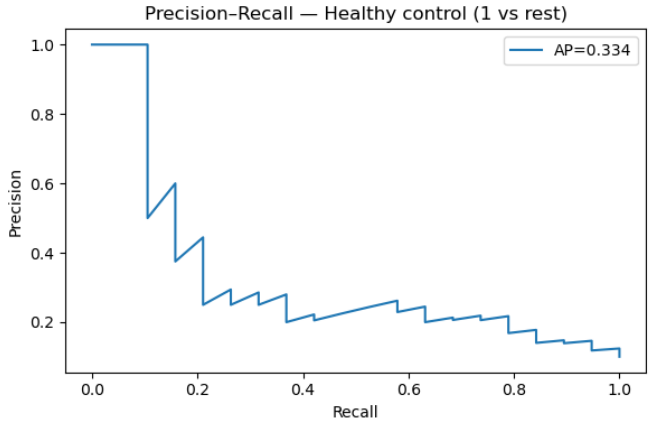



Figure 10: `Code2_19_1.png`

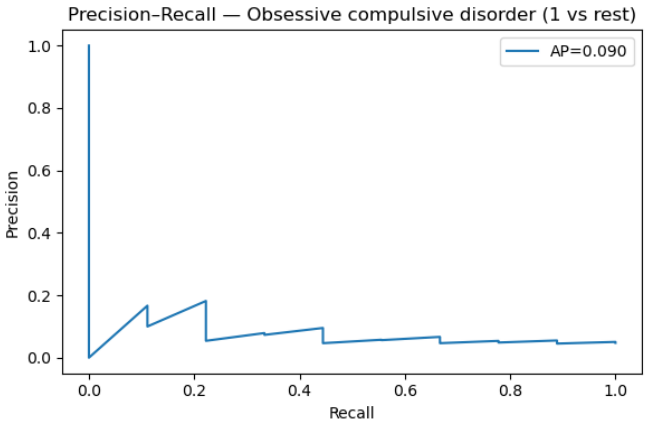



Figure 11: `Code2_19_10.png`

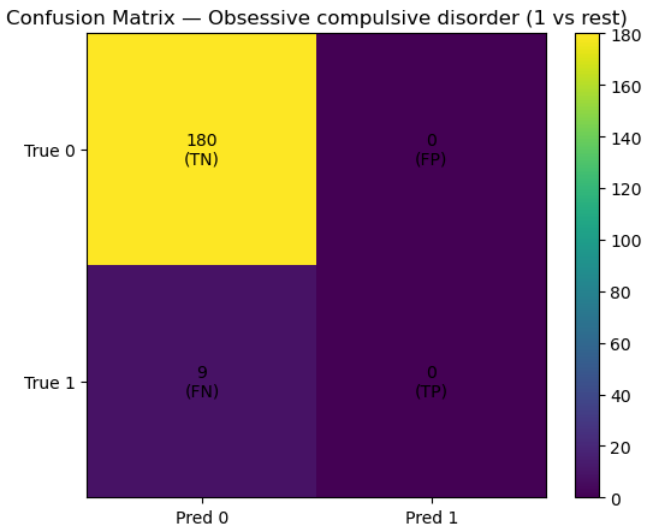



Figure 12: `Code2_19_11.png`

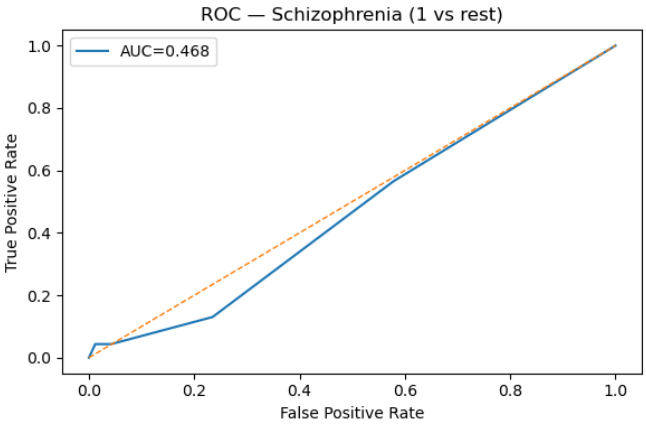



Figure 13: `Code2_19_12.png`

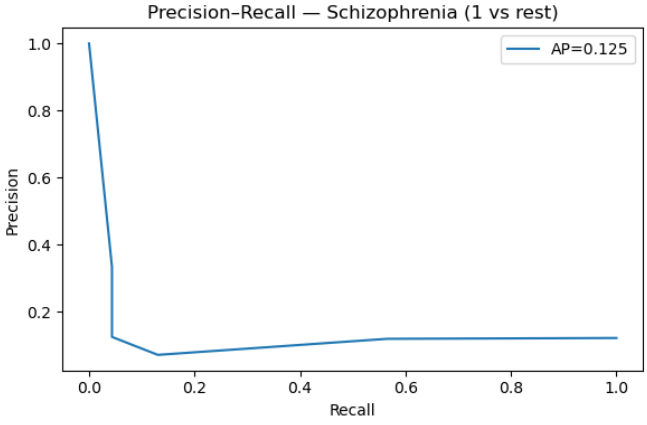



Figure 14: `Code2_19_13.png`

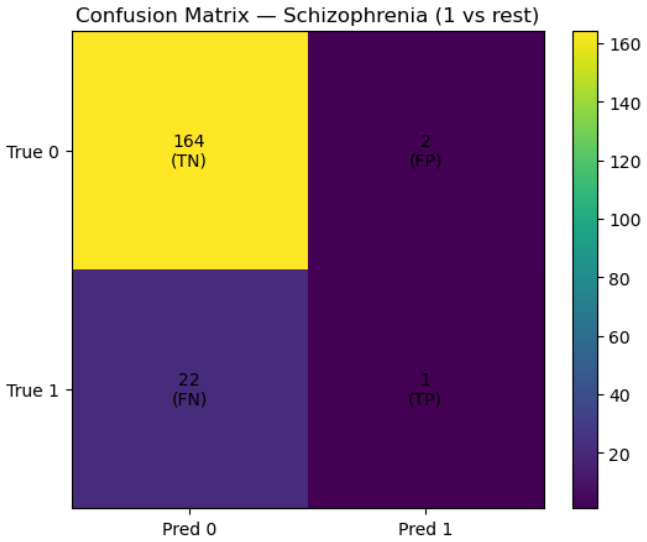



Figure 15: `Code2_19_14.png`

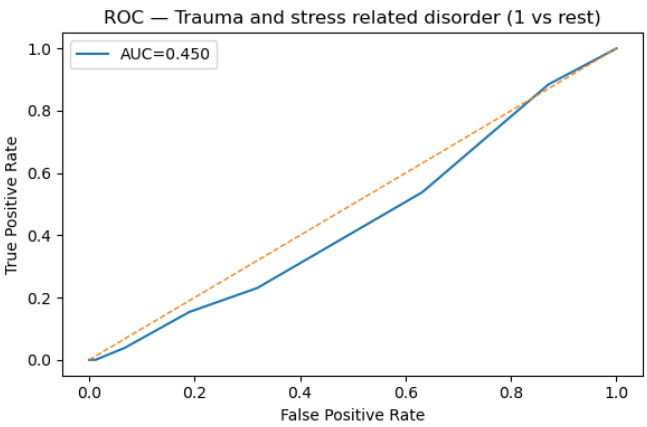



Figure 16: `Code2_19_15.png`

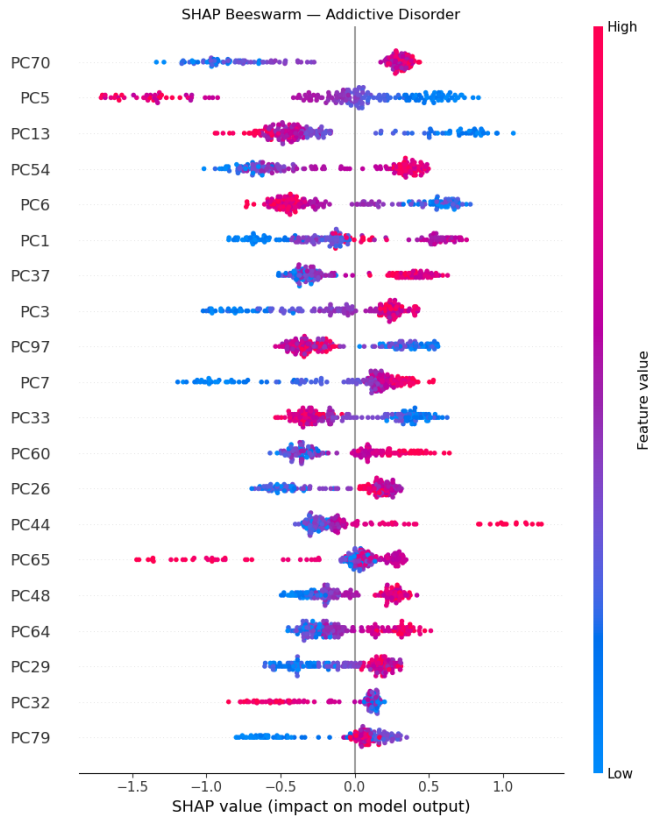
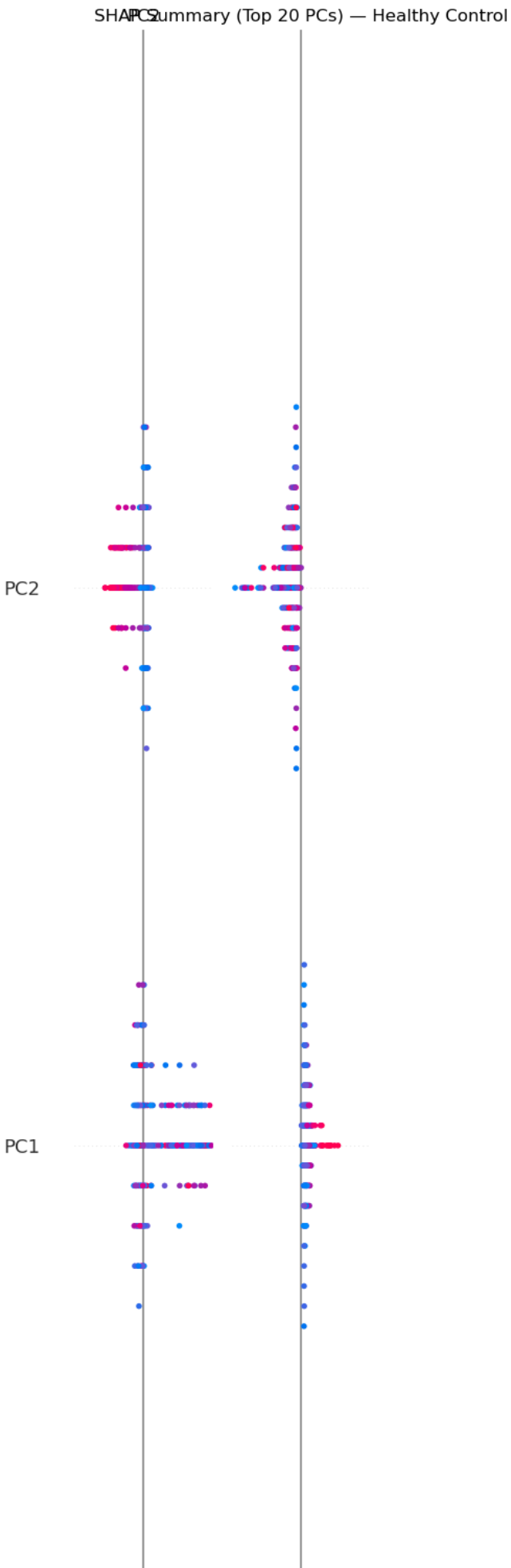
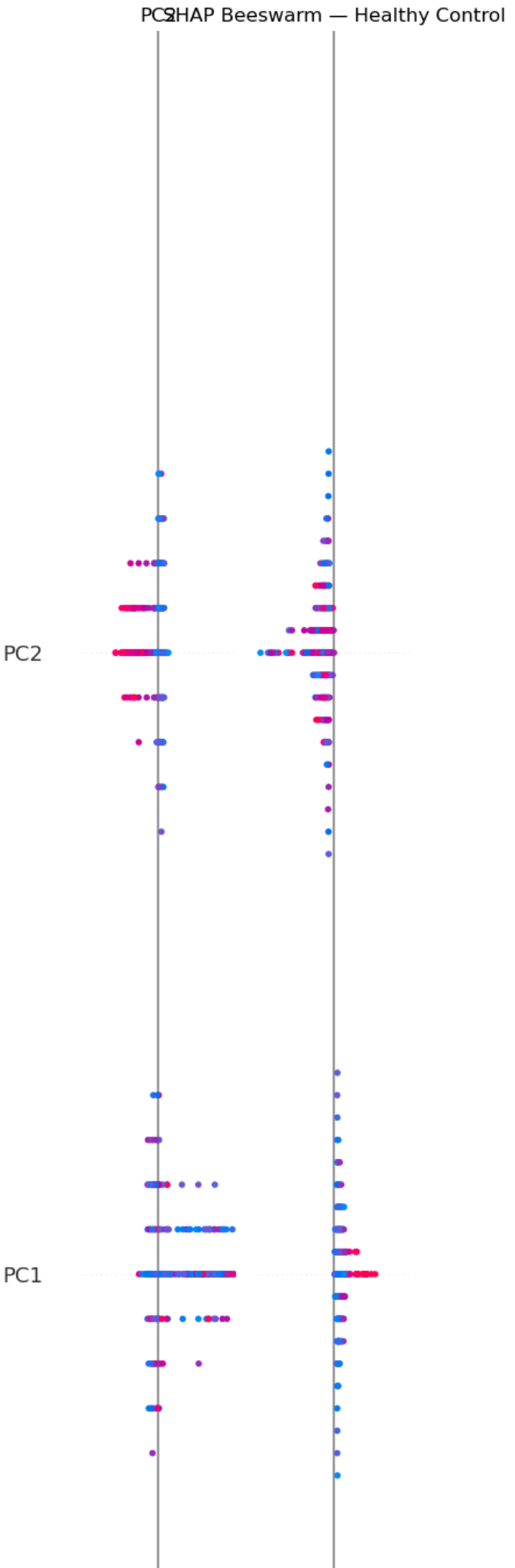



Figure 17: `Code2_20_10.png`





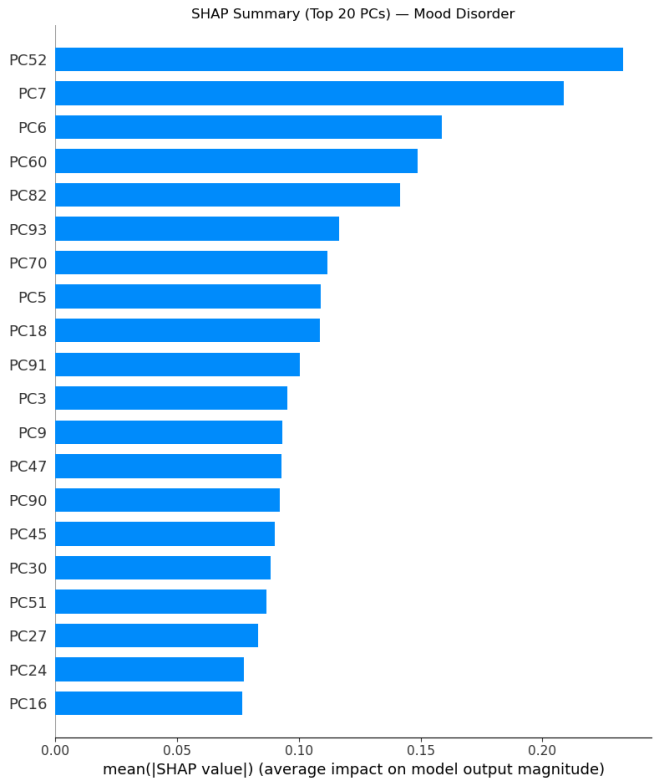




Figure 20: `Code2_20_6.png`

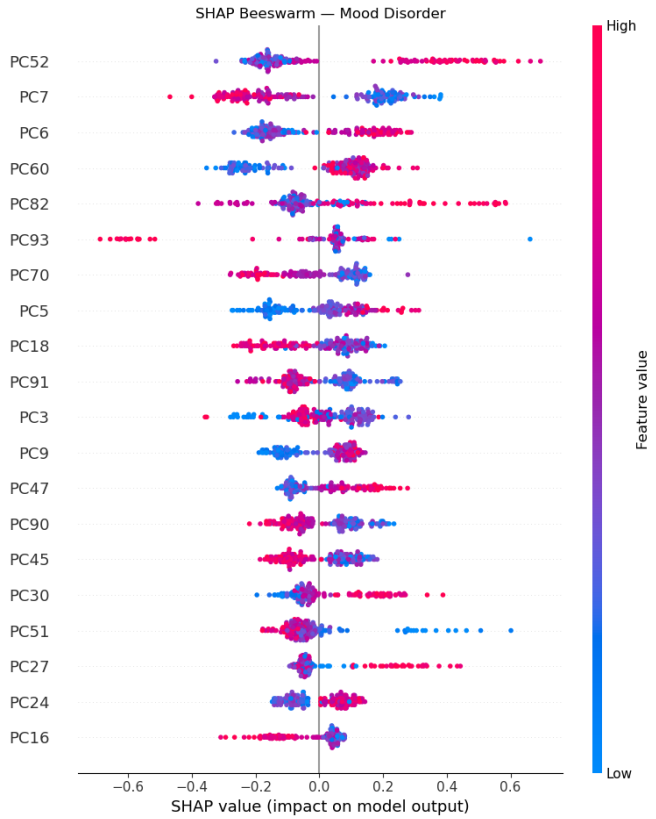



Figure 21: `Code2_20_7.png`

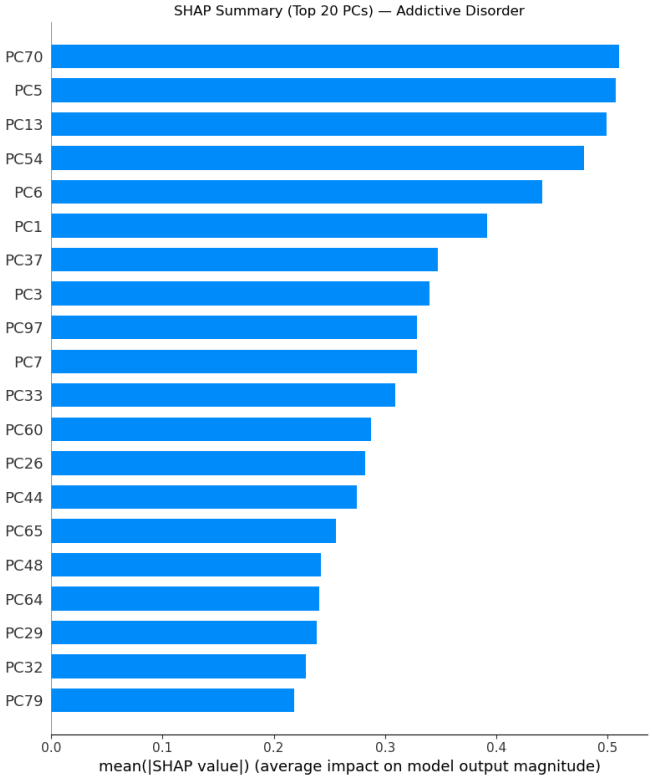



Figure 22: `Code2_20_9.png`

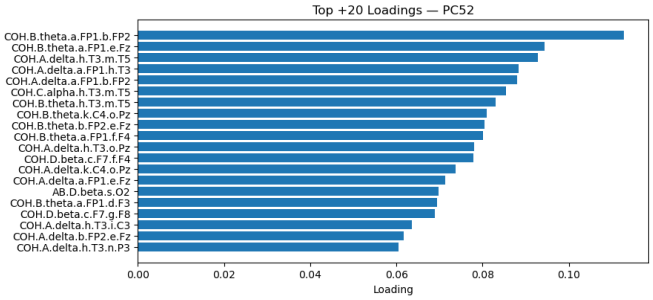



Figure 23: `Code2_21_10.png`

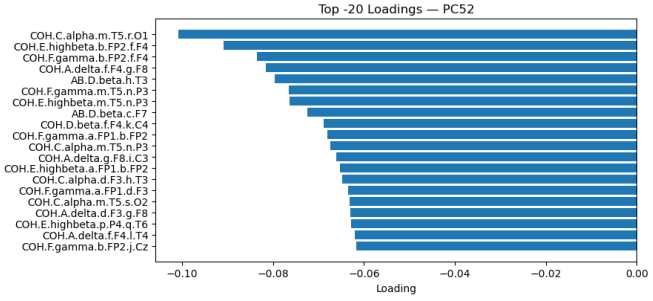



Figure 24: `Code2_21_11.png`

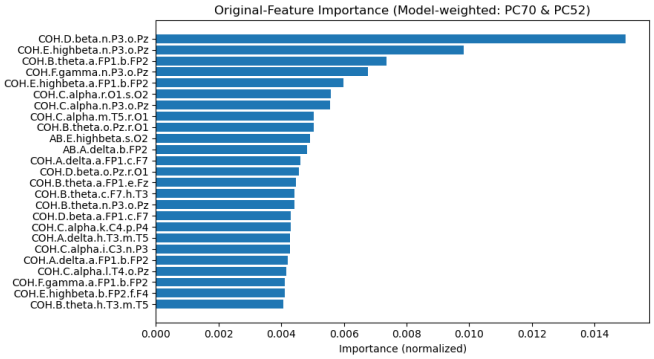



Figure 25: `Code2_21_14.png`

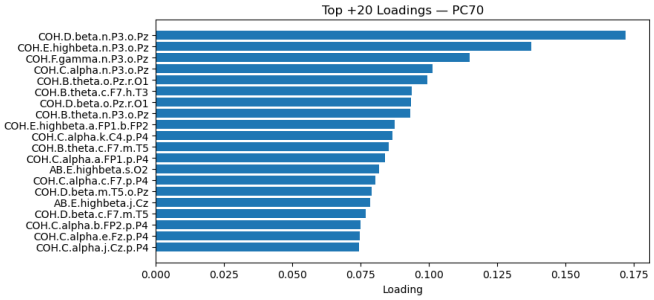



Figure 26: `Code2_21_4.png`

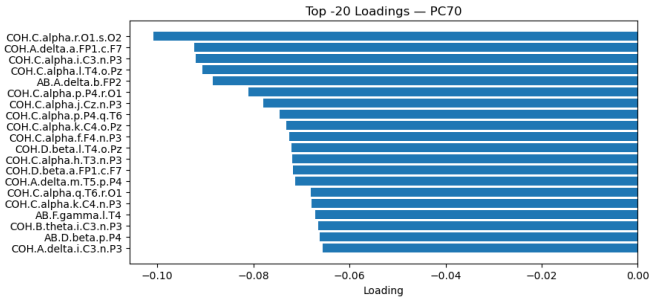



Figure 27: `Code2_21_5.png`

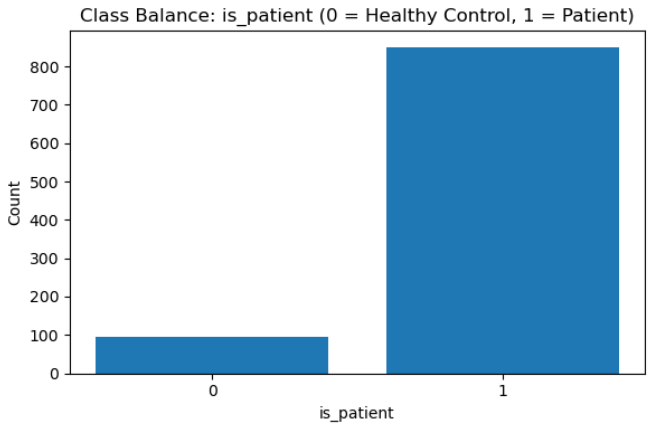



Figure 28: `Code2_3_0.png`

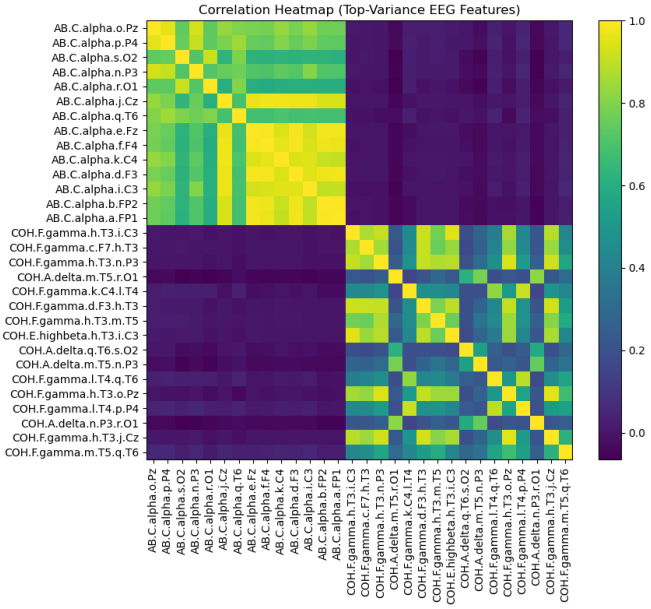



Figure 29: `Code2_4_0.png`

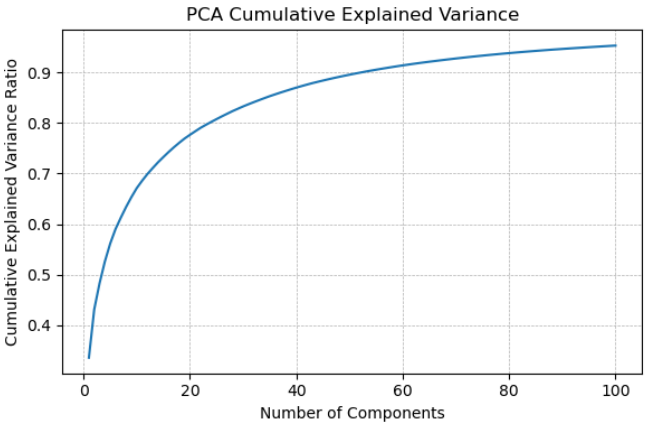



Figure 30: `Code2_5_0.png`

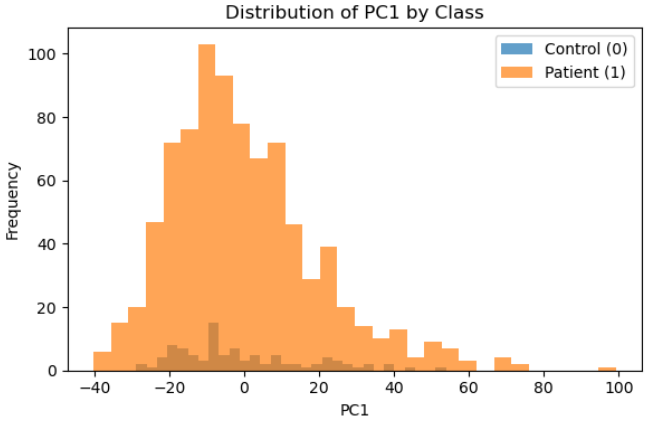



Figure 31: `Code2_7_0.png`