Identifying Bipolar patients from controls, using post-mortem cerebellum gene expression data and fully Automated Machine Learning.

Identifying bipolar patients using machine learning and cerebellar genetic data

1. Georgios V. Thomaidis, MD 1, 2. Sotirios Michos PhD 2, 2. Konstantinos Papadimitriou MD 3, 3. Evangelos Chartampilas MD 4, Prof. Ioannis Tsamardinos PhD 2,6.

1. Greek National Health System, Psychiatric Department, Katerini General Hospital, Katerini, Greece 2. Department of Computer Science, University of Crete, Heraklion, Greece 3. Greek National Health System, Psychiatric Hospital of Thessaloniki, Thessaloniki Greece 4. Laboratory of Radiology, AHEPA General Hospital, University of Thessaloniki, Thessaloniki, Greece, 5. Gnosis DA, Heraklion, Greece.

Abstract

Objective: Complex machine learning classification algorithms using transcriptome data from post-mortem cerebellar tissue of bipolar patients and unaffected controls, have been recently included in pipelines for patient – control classification and identification of characteristic biomarkers. Transcriptomic profile differences between patients and controls, can provide useful information about the role of the cerebellum in the pathogenesis of bipolar disorder and mood deregulation and in normal mood regulation and physiology. User-friendly, fully automated machine learning algorithms, using data extracted from established repositories, could achieve extremely high classification scores and disease-related predictive biomarker identification, in very short time frames and scaled down to small datasets, thus facilitating research on mood disorders.

Method: An application of a fully automated machine learning platform, based on the most suitable algorithm selection and relevant set of hyper-parameter values, for classification between patients and controls and the production of models for biosignature selection, is presented. Transcriptome data used for the analysis were downloaded from the BioDataome preprocessed datasets database. The Dataome dataset, derived from the parent Gene Expression Omnibus GSE35974 (2013) and GSE35978 datasets, which have been originally produced from the cerebellar and parietal lobe tissue of deceased bipolar patients and unaffected controls, (from the Stanley Medical Research Institute’s Neuropathology Consortium and Array collections), using Affymetrix Human Gene 1.0 ST Array. Patient and control groups were closely matched for age and sex.

Results: Bipolar patients have been identified from controls based on the cerebellar transcriptomic profile with AUC 0.929 and Average Precision 0.955. Patients and Controls have been classified in two separated groups with no close-to-the-boundary cases. Using 6 of the...
characteristic features discovered during the selection process, 99.6% classification accuracy was achieved. The three biomarkers contributing most to the predictive power of the model (92.7%), are also deregulated in temporal lobe epilepsy.

**Conclusion:** The cerebellar transcriptome of bipolar patients has a discrete profile and can be used for further exploration of the role of this area in health and disease.

93% AUC and 96% Precision were achieved during classification between unaffected controls and patients with Bipolar Disorder.

**Introduction**

Bipolar disorder (BD) is a mood disorder characterized by unusual fluctuations of mood, thinking, activity and sleep patterns, classified in three major subtypes [1] and presented as a constellation of phenotypes, with a variety of cognitive and behavioral features [2]. It is a highly hereditary disease, running in families, with an early onset, unpredictable course and detrimental impact due to the great risk of fatal self destructive events, long term disability and great financial and social burden, despite existing pharmacological and psychotherapeutic treatment strategies [3]. For these reasons, the neuroanatomy [4] and neurobiology [5] of bipolar disorder are fields of intense research and of paramount importance for 45 million patients globally [6].

The main functional role of the cerebellum is related to modulation of movement [7]. Beyond this established role, research has linked the cerebellum to emotional, cognitive and affective processing and their disruption in mood disorders [7], [8]. Structural [9], [10], [11], [12], [13], [14] functional [7], [8], [14], [15], [16], [17], [18], neurotransmission [19], [20] metabolic [21], [22], [23] and transcriptomic [24], [25], [26] alterations in the cerebellum in BD point to a discrete role in the affected brain networks.

Machine learning, is now gradually used in psychiatry, in order to optimize genetic analysis’s results [27], [28], to highlight the most characteristic differences among groups of patients and normal controls and to confirm their importance for diagnostic classification between these groups. These complex classification algorithms, produce genetic signatures, both using living tissue, blood or saliva and postmortem (prefrontal cortex) [28] and SNP (5 studies) [27, Table 1.] and transcriptomic (2 studies) [28], [29] data. In this context, transcriptomic data analysis can contribute greatly in psychiatry [30], and data from the less explored area of the cerebellum, can add new and important signatures in the puzzle of bipolar disorder pathogenesis, progression and potentially treatment response and resistance. The current study is the first where specifically autoML and transcriptomic data from the cerebellum, were used for biomarker identification and patient classification.

We applied a fully automatic machine learning (autoML) platform Just Add Data Bio (JADBio) platform [31] on public transcriptomic data from previous studies [25], [26], which analyzed the transcriptomic profiles of the cerebellum and parietal cortex of post mortem brain tissues.
Patient and control groups were homogenized by tissue sample location (cerebellum), psychiatric diagnosis, sex and age. The autoML system has a simple, user-friendly interface and has been created for direct application on low-sample, high-dimensional databases. It is automatically trained and evaluated (tested), in order to identify highly optimized predictive and classification models, using characteristic biosignature profiles. Provided a specific outcome (ex. diagnosis) and a well-defined set of features (ex. data from transcriptomic, biochemical, neuroimaging, psychometric or symptom intensity measurements), it can provide a minimal subset of predictors (biomarkers), selected from the features, leading to increased predictive power performance. In studies using binary classification (ex. between BD patients and controls, using transcriptomic data in the current study), the classification boundary is defined by the most statistically significant combination of biomarkers (the characteristic biosignature), which identifies patients from controls. System applicability has been tested for diagnostic classification and time to event prediction, producing robust classification, biomarker identification and prediction results (AUCs 85–95%) using data from oncology, neurology and psychiatry and in international evaluations [32-36].

Aims of the study

Scope of the analysis, is the selection of characteristic transcriptomic biosignatures of bipolar disorder in the cerebellum -using the autoML platform for optimal performance- and a primary interpretation of the results. Information on the features identified, could facilitate the discovery of the genetic networks leading to BD or are altered during disease course and their importance at the local and global brain network level.

Materials and Methods

Data acquisition

For this analysis, publicly available data have been used, from the online BioDataome database [37], which is constructed by uniformly preprocessed, disease-annotated omics data from GEO and RECOUNT databases, based on a uniform preprocessing pipeline [38]. We analyzed the BioDataome csv. which corresponds to the GEO dataset GSE35978, a. containing expression data from the human cerebellum (produced from GSE35974) and parietal cortex, b. from post mortem brain tissue samples, c. extracted from unaffected subjects and schizophrenic, bipolar and depressed patients, d. from the Stanley Foundation Brain Collection [39]. The expression data were obtained by microarray analysis using the "[HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array [Transcript (gene) version]". The dataset was initially used at the analyses by Chen C et al [25], [26].
Data Processing

a. Dataset selection and homogenization

Data have been downloaded in .csv format from the BioDataome database. The preprocessed file includes data for 144 samples from the cerebellum and 168 samples from the parietal cortex. The 144 cerebellum samples include unaffected subjects and patients with bipolar disorder, schizophrenia and depression (SI, information on GSE35974 and GSE35978). From the cerebellum group, all 50 unaffected subjects and 37 bipolar disorder patients (sex: females/males, age span: 20 - 70) were initially chosen (SI, Tables 1a, 1b). From the initial heterogeneous groups of affected / unaffected, a number of subjects were removed, and two new, smaller groups of affected / unaffected subjects were produced, matched for sex (female / male) and for age. In parallel, we aimed to exceed (as much as the samples allowed), the minimum threshold of 30 subjects per group, required for the machine learning analysis, without an impact on the age and sex between-sample matches and distributions (SI, Tables 2a, 2b). The final dataset includes these two groups, Group A with 35 bipolar patients (18 female +17 male) and Group B with 37 unaffected controls (19 female +18 male). The small size of available data excluded the possibility of a testing after the initial training; this is balanced by the extremely high AUCs produced during the initial (training) analysis. During the initial microarray analysis, a number of transcriptomes is used as controls [25], [26]. These have been identified, removed from the csv. of the analysis and the final datasheet (Diagnosed Subjects x Features) was produced. The datasets are 2D matrices (features/genes x diagnosis for any given subject, unaffected or patient).

b. Feature selection and biosignature construction

For the analysis, data were uploaded to JADBio version 1.4.14 (April 2021) and the binary classification (categorical) functionality of the platform was employed. The classification process is based on the Statistically Equivalent biosignatures (SES) method, with Support Vector Machines, Random Forest, and Penalized Linear Models algorithms. [31], [35]. For the given 2D matrices, the predicted outcome is diagnosis (Bipolar or Unaffected), and the metric chosen for optimization is the AUC. Preprocessing used Constant Removal Standardization.

Feature selection was performed using LASSO Feature Selection (penalty=0.0, lambda=5.509e-02). The analysis protocol followed has been a repeated 10-fold cross validation without dropping (max. repeats = 20), with 596 configurations, 5760 predictive models trained and 83440 predictive models omitted (total 89200). The chosen predictive algorithm uses Ridge Logistic Regression (with penalty hyper-parameter lambda = 10.0). The overall process applies the Bootstrap Bias Corrected Cross Validation, a protocol for algorithm hyper-parameter tuning during performance estimation and multiple tie adjustment [31], [35]. The technical analysis report is in SI-Appendix-1.
Results

1. Classification between BD patients and Unaffected Controls

The AutoML classification analysis produced a Ridge Logistic Regression model with high AUC for the positive class bipolar (93%), based on 25 characteristic biomarkers. AUC, and Average precision values and confidence intervals(CIs), ROC curve and main optimized classification threshold dependent metrics for Accuracy / Balanced Accuracy are shown in Image 1. The BD and HC groups are completely separated and coherent in the UMAP plot (Image 2).

Image 1. Using the best performing model option of in the platform, the AUC for the positive class bipolar is 0.929 (~93%) , with a 95% confidence interval (CI) between 0.868 - 0.977 and average precision 0.955, with a 95% CI 0.914 - 0.986. Accuracy has been calculated at 0.843, Precision at 0.906 and Specificity at 0.921 (full data in SI, Image 1). The classification threshold (0.61) has been optimized and determined for Accuracy / balanced accuracy. Classification as positive is performed when out-of-sample predicted probability is above this given threshold (0.61).
Image 2. Complete separation of BD patients from unaffected controls, in UMAP plots based on all the 25 selected biomarkers. In the Box Plot contrasting the cross-validated predicted probability of belonging to a specific class against the actual class of the samples, the medians are ~0.72 for the class “bipolar” and ~0.18 for the class “unaffected” (SI, Image 2).

2. Biomarker and Biosignature identification

The algorithm selected the most important 25, out of 28869 features (transcriptomes from the genes studied) in the original dataset for the reference signature, used for prediction of BD based on these samples. Inclusion of the 6 most important features (gene transcriptomes from RNU6-576P, MIR194-2, GDPS5, CARD16, RABGGTA, KREMEN2), achieves predictive performance (PP) 99.603%. Inclusion of the most important feature RNU6-576P, leads to 76.6% PP, inclusion of the first and second (MIR194-2) most important feature achieves 85.8% PP and
additional inclusion of the third most important feature (GDPD5) achieves 92.7% PP. The progressive feature inclusion plot for 6 most important of 25 identified features is presented in Image 3.

![Progressive Feature Inclusion Plot](image3.png)

The 21 known or probable functional roles of RNU6-576P, MIR194-2, GDPD5, in the nervous system and CNS diseases, were found after a thorough review of the relevant literature (using and the gene aliases from Gene Cards), which produced 36 articles, and are presented in SI-Appendix 2. The most consistent and important finding is that RNU6-576P, MIR194-2 and GDPD5, have been associated with epilepsy (but not until now with bipolar disorder). Both epilepsy and bipolar disorder are characterized by episodic functional deregulation in the CNS [40], co-occur [41], share common symptoms and precipitating factors [42], [43], their treatment with antiepileptics / mood stabilizers is partially overlapping [44], and potential pathophysiological links have been proposed recently [45], regarding aberrant neuronal excitation-inhibition related to ANK-3 gene expression. Finally, epileptiform EEG discharges are connected to progress and worse course of disease in BDII patients [46] and manic symptoms are more common in patients with temporal lobe epilepsy [42]. Significantly, alterations in RNU6-576P and MIR194-2 expression are connected to temporal lobe epilepsy [47], [48], [49], [50], [51], which shares the most common symptoms and pathways with Bipolar Disorders I and II [41] - [46]. Expression of gene RNU6-576P is the most overexpressed small non-coding mRNA in the hippocampus of patients with mesial temporal lobe epilepsy [47] and the most important identifying biomarker in the cerebellum of BD patients in this study.
Discussion:

Main Findings

The classification between the Bipolar and unaffected control groups was completed in <1 hour, with accuracy ~93% and without overlaps between the produced sets of individuals. Welsh t-test for the 6 most important genes, established that the differences in expression between patients with Bipolar Disorder and Unaffected Controls, are statistically meaningful (SI, Image 3). Classification using the JADBIO platform can be considered a reliable means and produces robust results, with potential research interest and physical meaning. The single most important identifier was by far the RNU6-576P small non-coding RNA, accounting for ~77% of total feature importance and also highly deregulated in temporal lobe epilepsy. The role of small non-coding RNAs and pseudogenes is a new area of intense research regarding their role in the onset of psychotic disorders, depression and bipolar disorder [48], [49]. The role of MIR-194-2 expression in epilepsy has been better studied and a constant pattern of down-regulation has been documented in various epilepsy studies [50]-[53].

Limitations of the study

The present study was based on a relatively small sample of patients with Bipolar Disorder Types I and II, with an increased analogy of deaths from suicide and was based on post-mortem tissue sampling. Genetic differences between patients with BD I and BD II have been suggested [54], [55], [56], [57], using family databases, but neuroimaging differences have not been confirmed [58]. The bipolar spectrum is highly heterogeneous, with many different biotypes and their probable neurobiological and functional variance [4], [59], [60], [61], [62]; different biotypes can be fully represented only in large samples [4]. The bipolar spectrum includes 7 of the 37 patients with bipolar disorder had committed suicide, a number close to known prevalence of death by suicide in BD. Suicide mainly occurs during the depressive state of the disease and – occasionally- during a manic episode and could be connected to certain patterns of gene expression [61], biomarker differences [63] and vary during an acute or prolonged depressive or manic or mixed episode) [64], including the cerebellum. Finally, the genetic characteristics of post-mortem brain tissue sampling could be divergent from the same characteristics of the living brain, in health and disease; still they remain one of the cornerstones of research on the neurobiology of the CNS and its disorders [65], [66], [67] [68].
References


38. “Download, preprocess, annotate and analyze omics data sets”. Kleanthi Lakiotaki. Computer Science Department, University of Crete. 2017-12-01


