Automatic high throughput genotype classification pipeline with feature selection based on Rodents EEG Signals

NOVARTIS

Fangfang Sheng, Daniel J. Graziano, Eric J. Ma, Mei Xiao, Holger Hoefling and Michelle M. Sidor

Informatics Systems Department and Neuroscience Department, Novartis Institutes for Biomedical Research, Cambridge MA 02139

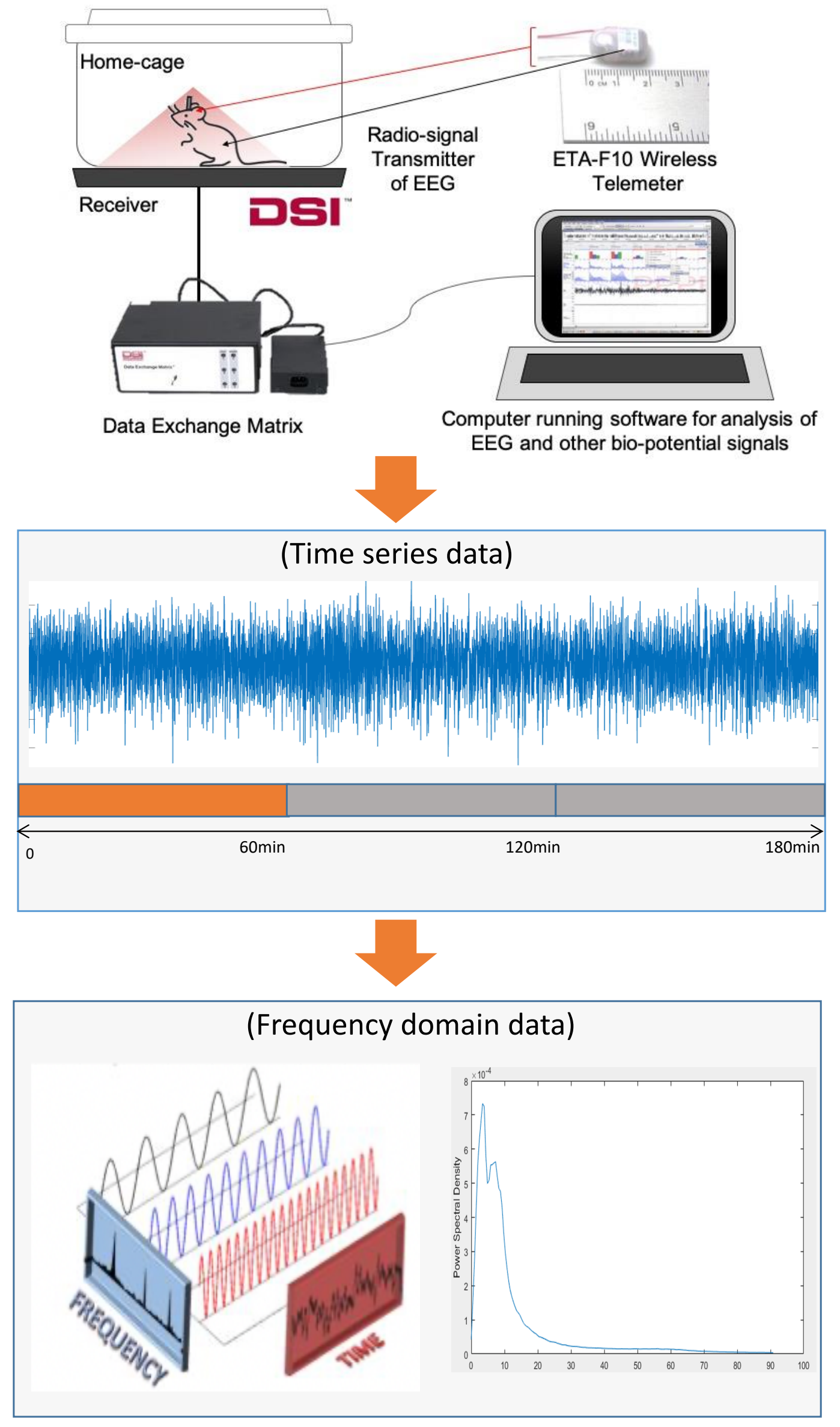
Introduction

- Electroencephalography (EEG) analysis has been an important tool in neuroscience with applications in neuroscience, neural engineering (e.g. Brain-computer interfaces, BCI's), and commercial applications.
- The availability of large EEG data sets and advances in artificial intelligence have both led to the deployment of machine learning architectures in the analysis of EEG signals and understanding the information it may contain for differentiate groups of objects.
- The difficulty in using EEG for research purposes, however, is that both feature selection and classifier training must be conducted across very noisy datasets.
- Despite the advances in artificial intelligence, the classification work is in most cases still performed manually. The robust automatic classification of these signals is an important step towards making the use of EEG more practical in many applications and less reliant on trained professionals.
- This study aims to build an automatic EEG signal-based genotype classification with feature selection and conduct experiments in various methods at each phase.

Keywords

EEG signals; Classification algorithm; Feature Selection Algorithms; Automatization; Performance evaluation

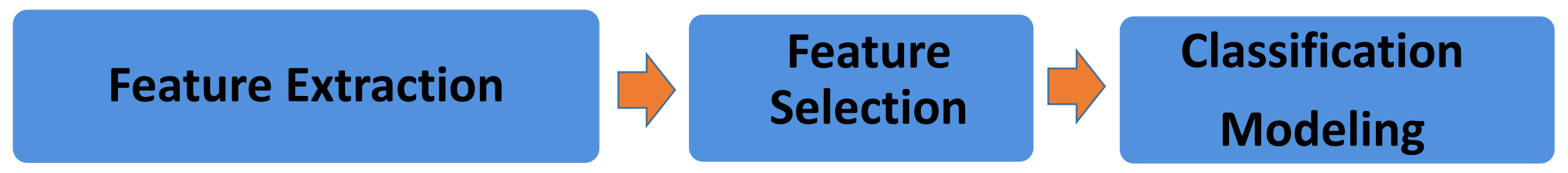
Frequency labeling



- A High-Throughput EEG analysis pipeline for Neurophysiological Monitoring Combined with Optogenetic Capabilities in Awake-Behaving Rodents has been successfully built in NIBR NS EEG lab.

Figure 1. Data Sciences International (DSI) wireless radiofrequency telemetry platform was used to obtain EEG recordings. EEG time-series data is comprised of multiple frequency components. Custom script was written in Python to deconvolve data from the time-domain into the frequency domain using the Fast Fourier Transform (FFT).

Machine Learning Pipeline



Feature Extraction → Feature Selection → Classification Modeling

Figure 2. The **machine learning workflow** has been divided into three essential stages including: feature extraction, feature selection and classification.

Feature Extraction

Time segmentation means manually splits the phenotyping data into a fixed length using the sliding window approach in a long temporal phenotype dataset. Here, the window length is the smallest temporal period of any emerging phenomenon that users can specify. We used 60min windows here.

Binned and unbinned full spectrum power spectral density are calculated. There are two ways to generate features from the relative EEG power spectral density, traditional frequency bin definitions ("Binned Spectrum = Delta, Theta, Alpha1/2, Sigma, Beta1/2, Low, Med, High Gamma, Gamma) and an unbinned approach (Full Spectrum = 2Hz resolution). Each of the frequency bins is treated as a single parameter and tested for its ability to separate groups; 53 such parameters (frequency bins) were available.

Feature Selection

ANOVA F-value filter method used to check the means of two or more groups that are significantly different from each other. To get F value, we would compare the variance between the groups and variance within the groups.

$$F\ value = \frac{SSB/_{df_B}}{SSW/_{df_w}}$$

Where SSB was the Between the Sum of Squares, SSW was the Within Sum of Squares, df was the degrees of freedom.

Jax-based logistic regression with lasso attention uses the LASSO method on logistic regression model and exclude the features with coefficient equal to zero as is a powerful method for feature selection. We applied a two layers Neural Network structure.

XGBoost is short for Extreme Gradient Boosting and is an efficient implementation of the stochastic gradient boosting machine learning algorithm. We imported XGBClassifier from xgboost package.

Recursive Feature Elimination(RFE) recursively removes features, builds a model using the remaining attributes and calculates model accuracy. Features are ranked by the model's feature importance attributes, requires the optimal number of features which is achieved by series of experiments.

Classification model

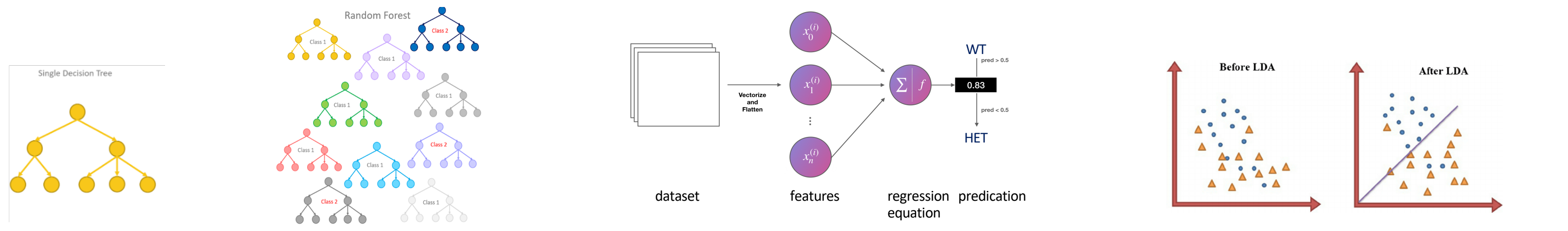


Figure 3. Decision Tree, Random Forest, Logistic Regression and Linear Discrimination Analysis are state of the art algorithms in Machine Learning for binary classification task.

Decision trees (DT) is generated in the first stage of the classification. Data is applied one by one to the tree to undertake a classification process in the second stage.

Random forest algorithm (RF) combines multiple trees with multi variables, each of which can be trained with different training clusters instead of generating a single decision tree. Different training clusters are generated from the original training set by using bootstrap and random feature selection.

Logistic Regression(LR) is a classification algorithm used to assign observations to a discrete set of classes. Logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to different genotypes.

Linear Discrimination Analysis (LDA) estimates a linear function of the variables that acts as a classification rule to predict group membership.

Testing and Evaluation

To derive a more accurate estimate of model prediction performance , we adopted the **5-fold cross-validation** combining **averaging 1000 times results**. To speed up the process, we analyze datasets and train models on HPC systems with Dask, a parallel computing library that integrates well with the existing Python software ecosystem.

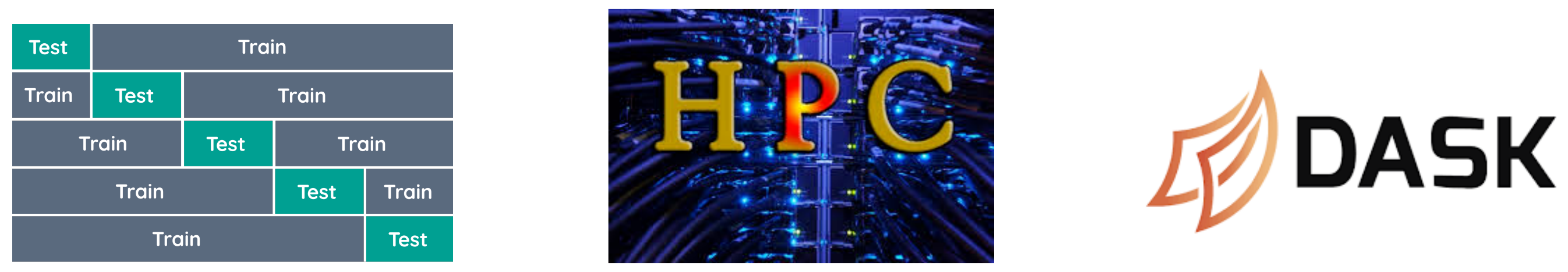


Figure 4. The goal of cross-validation is to test the model's generalization and flag problems like overfitting or selection bias. NIBR-Dask enables the training process and **automatic** classification pipeline of rodent genotype.

Accuracy and **stable score** are two metrics to assess different machine learning pipelines' performance. They were calculated using following formulas:

$$Classification\ accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$Stable\ Score = \frac{\sum_{a=1}^{num\ trials} Intersection\ rate}{num\ trials}$$

Results

Feature Selection	Classifier Model	Accuracy	Stable Score	Selected Features
RFE_SVM	Decision Tree	0.93	1.00	0.5-2Hz,4-6Hz, 8-10Hz, Peak frequency Broadband, theta/gamma ratio
Lasso score	Logistic Regression	0.90	N/A	Spectral Edge Density, 12-14Hz, 18-20Hz, Low Gamma, Z ratio
RFE_SVM	Random Forest	0.89	0.78	0.5-2Hz,4-6Hz, 8-10Hz, Peak frequency Broadband, theta/gamma ratio
RFE_Logistic Regression	Linear Discrimination Analysis	0.87	1.00	0.5-2Hz, 8-10Hz, Peak frequency Broadband, theta/beta ratio, theta/gamma ratio
XGBoost	Logistic Regression	0.80	1.00	14-16Hz, 24-26Hz, 28-30Hz, Low Gamma, theta/gamma ratio

Table 1. Partial experimental results showed above. The **RFE feature selection and Decision tree classifier** had the best performance.

Summary and Future work

1. This work provided a comprehensive survey of automatic EEG-based signal processing techniques applied to rodent genotype identification. The analysis procedure has been divided into three essential parts including: feature extraction, feature selection and classification.
2. The models were able to reach state of the art classification performance (accuracy: 93%, F1-score: 89%) on the sample 3 hours phenotyping dataset.
3. The study offers valuable information for researchers to find out which modelling methods have been used for certain EEG schemes and discusses their performances and efficiency. Moreover, an automatic classification workflow with feature selection was developed that could be easily implemented to differentiate classification problem using EEG signals.