SVEUČILIŠTE U ZAGREBU
Fakultet
elektrotehnike i
računarstva

Explainability of Machine Learning Models in Prediction of Affective Disorders

FANI SENTINELLA-JERBIĆ

July 2024, Zagreb

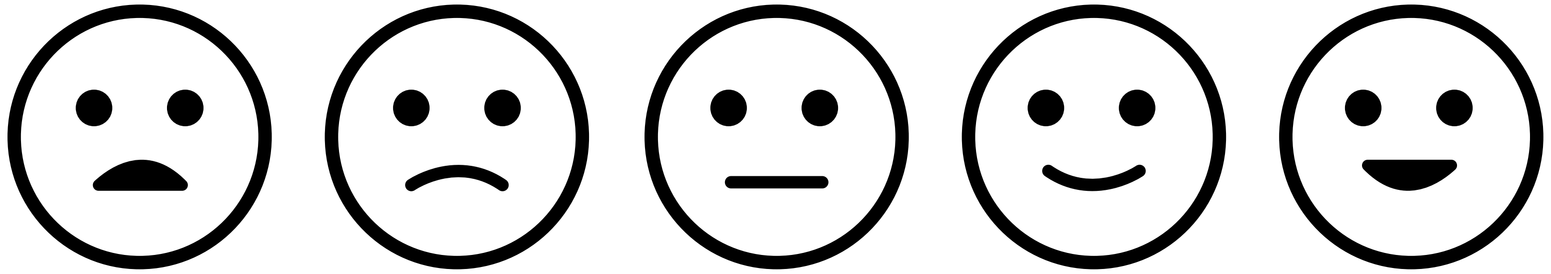
Overview

Overview

Introduction

Affective disorders

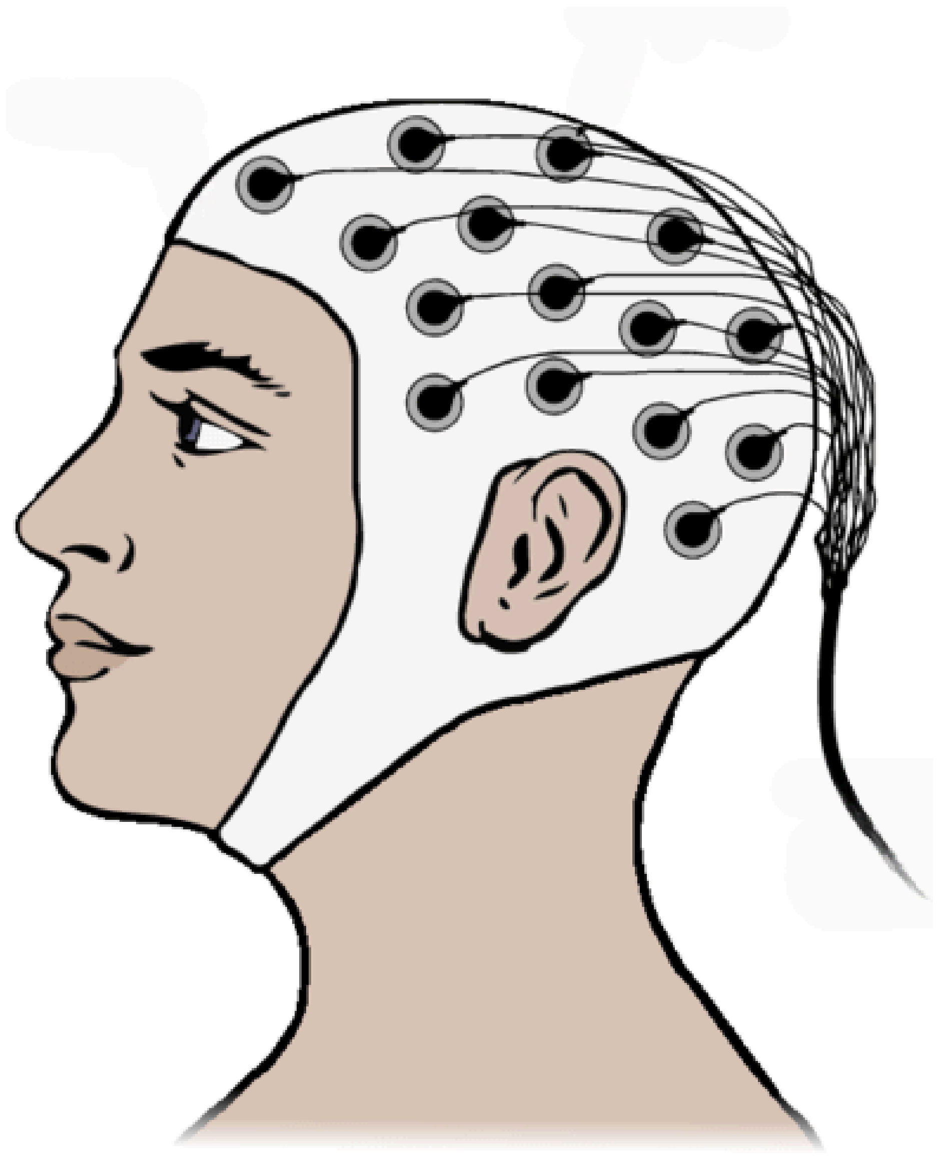
- Mental and behavioral disorders characterized by a shift in mood to either elation or depression
- Mainly diagnosed with patient interviews and questionnaires



Introduction

Electroencephalography (EEG)

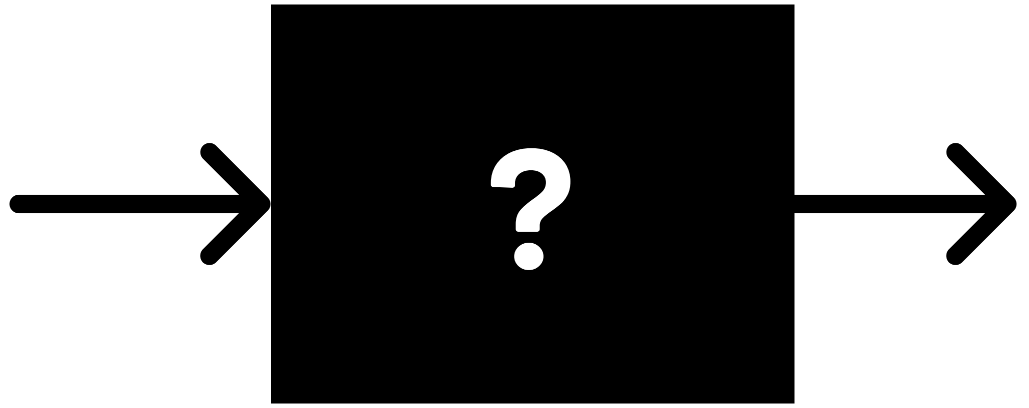
- Brain activity recording method
- Captures electrical signals fired by groups of neurons synchronizing



Introduction

Explainable Artificial Intelligence (XAI)

- Rising area of research aiming to increase trust and adoption of AI
- Focused on improving understanding of increasingly complex AI systems



Motivation

Brain Awareness Week 2022

Motivation

Objectives

1. Identify potential EEG biomarkers of depression
2. Compare different explainable AI methods

Overview

Literature Review

Identifying EEG Biomarkers of Depression with Novel Explainable Deep Learning Architectures (2024)

Input	raw EEG signal
Model	deep convolutional
Explainability	visualization of model internals
Findings	• β & δ power • brain-wide correlation • right hemisphere

Literature Review

An Explainable Assessment for Depression Detection Using Frontal EEG (2023)

Input	EEG extracted features from frontal lobes: Higuchi’s fractal dimension, sample entropy
Model	Decision Tree, LDA, k-NN, Random Forest, XGBoost
Explainability	feature importance
Findings	• complexity • high-frequency features

Depression detection based on analysis of EEG signals in multi brain regions (2023)

Input	EEG extracted features: Lempel-Ziv complexity, power spectral density
Model	Support Vector Machine
Explainability	subset evaluation
Findings	• temporal region • frontal, temporal, and central regions combined

Overview

Dataset

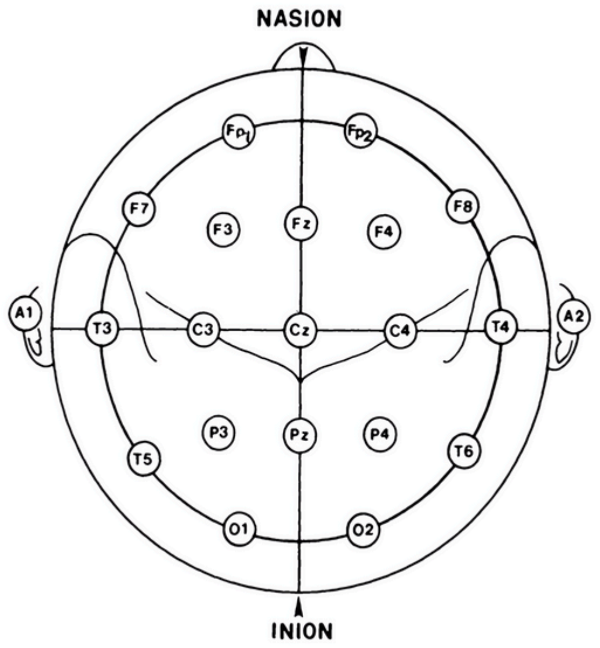
Composition

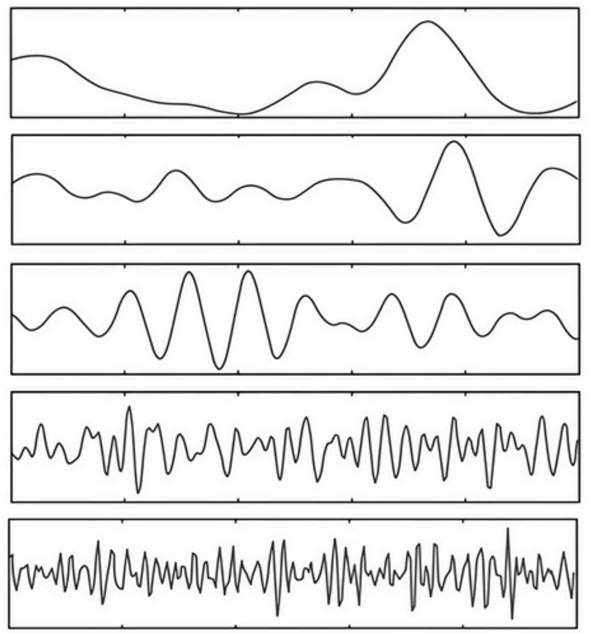
- **105 individuals**
  - 75 depressed
  - 30 healthy
- **representative train-test split**
  - 75 train examples
  - 30 test examples

Dataset

Features

19 electrodes x 5 brain waves x 6 extracted features = 570 features





absolute band power

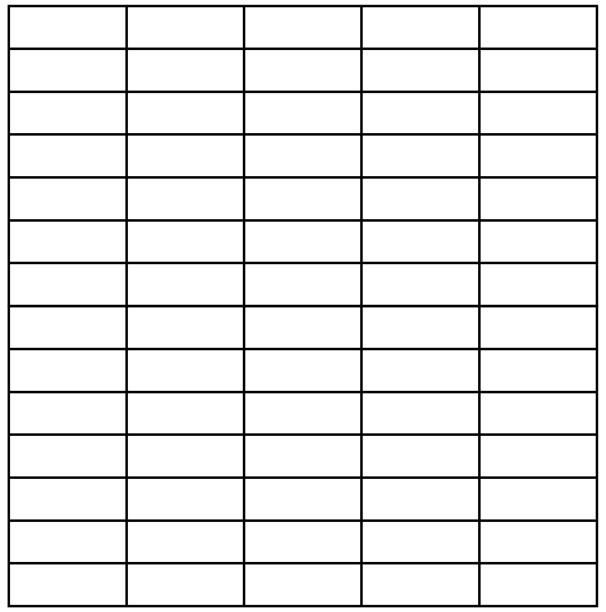
relative band power

spectral centroids

relative wavelet energy

wavelet entropy

Katz fractal dimension



Overview

Methodology

Approach

- **Feature Ranking with XAI Methods**
  - identify features that contribute most to the predictive power of a classification model

- **Feature Subset Evaluation**
  - identify subsets of features that contribute most to the predictive power of a classification model

Methodology

Ideal outcomes

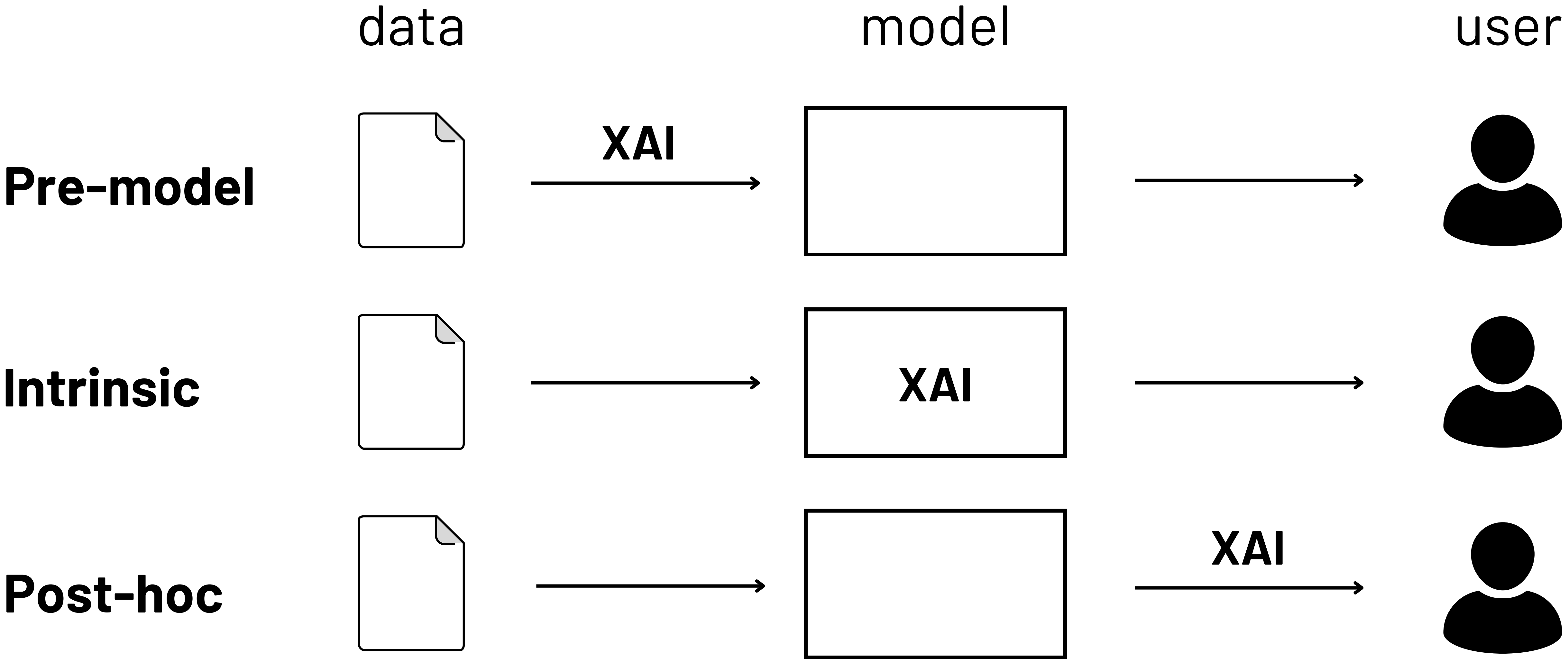
- **Feature Ranking with XAI Methods**
  - agreement across methods

- **Feature Subset Evaluation**
  - agreement across models

→ both would imply a reliable EEG biomarker

Methodology

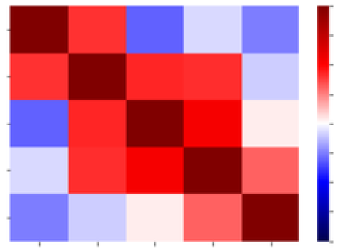
Feature Ranking with XAI Methods



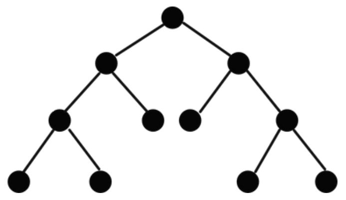
Methodology

Feature Ranking with XAI Methods

- Correlation with Diagnosis



- Decision Tree Importance

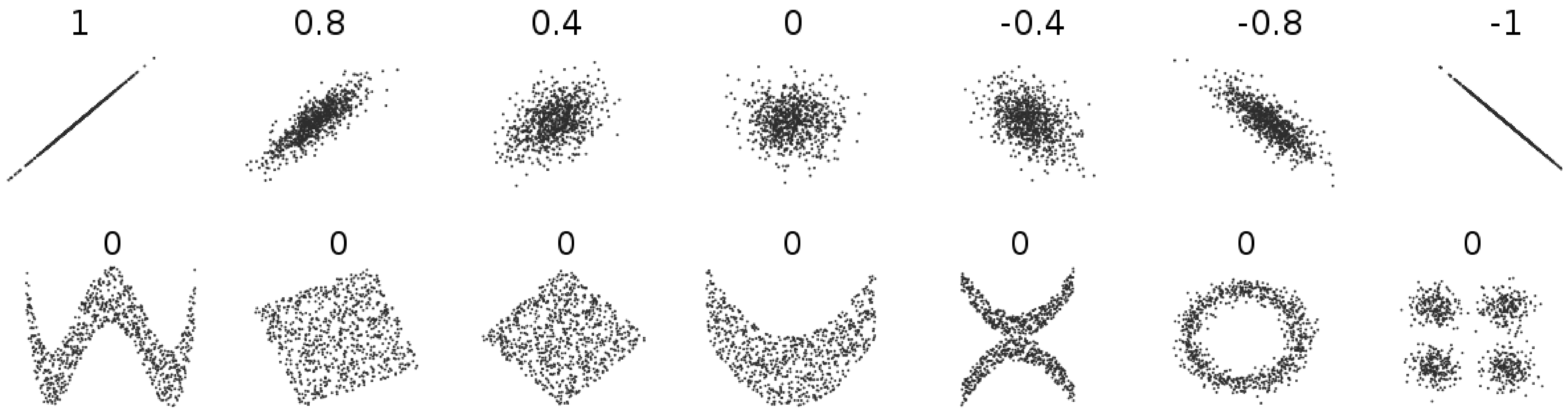


- Shapley Additive Explanations on SVM

Methodology

Correlation with Diagnosis

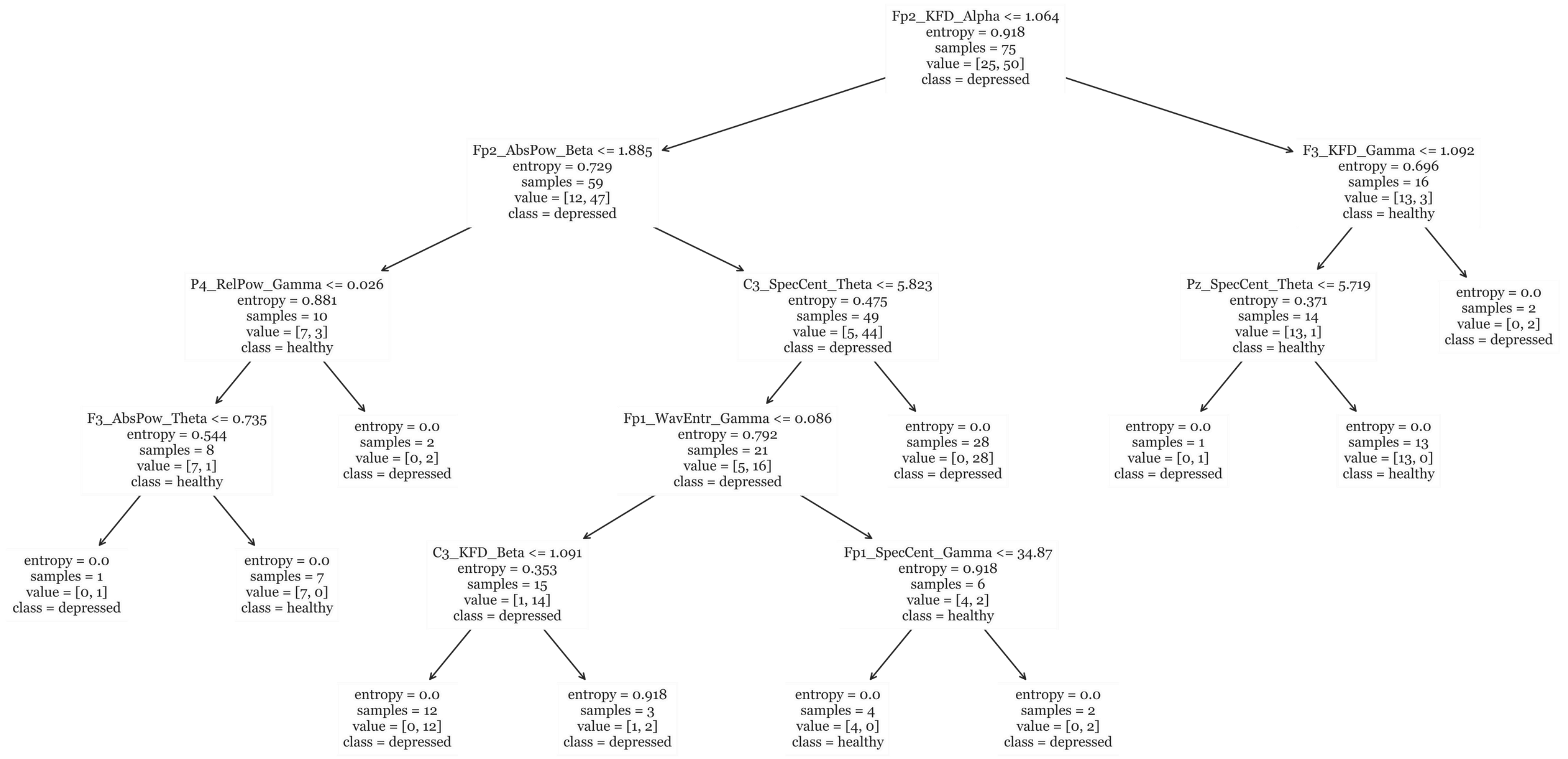
- Measure of the strength and direction of a relationship
- Assumes linear or monotonic relationship



Methodology

Decision Tree Importance

- A tree is built by recursively partitioning based on features that best separate the data into similar subsets



Methodology

Shapley Additive Explanations (SHAP)

- Rooted in cooperative game theory
  - distribution of the total gain among players based on their contribution to the overall outcome

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!}(v(S \cup \{i\}) - v(S))$$

Methodology

Support Vector Machine

- Aims to find the best separation of classes
    - by maximizing the margin between the nearest data points of different classes, known as support vectors
- Uses the kernel trick for implicit mapping into high dimensional feature spaces where data becomes separable by a hyperplane
    - reduced interpretability

Methodology

Feature Subset Evaluation

- Domain-informed subsets
  - per electrode
  - per brain wave type
  - per feature extraction method

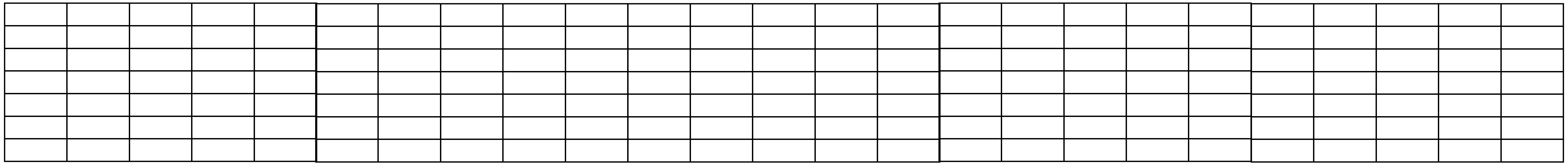
- Analyses-informed subsets
  - literature
  - hypothesis testing
  - feature ranking

→ all on both DT & SVM ,
using F1-score as primary metric

Methodology

Main Limitations

- **dataset nature**
  - imbalanced
  - single train-test split
  - ~5:1 feature to instance ratio
- **multiple comparisons problem**



Overview

1 Introduction & Motivation
2 Literature Review
3 Dataset
4 Methodology
5 Results
6 Conclusion

Results

Feature Ranking

Correlation	Decision Tree	SHAP Values
T4_AbsPow_Delta	Fp2_KFD_Beta	O1_AbsPow_Delta
F8_AbsPow_Alpha	Fp2_AbsPow_Beta	**T3_AbsPow_Delta**
T3_AbsPow_Gamma	C3_SpecCent_Beta	Fp2_RelPow_Delta
T3_AbsPow_Delta	F3_KFD_Gamma	O1_WavEntr_Alpha
T4_AbsPow_Alpha	Fp1_WavEntr_Gamma	Fp2_RWE_Delta
F8_KFD_Alpha	Fp1_SpecCent_Gamma	T6_WavEntr_Alpha
F8_AbsPow_Gamma	Pz_SpecCent_Theta	P4_AbsPow_Alpha
T3_AbsPow_Theta	P4_RelPow_Gamma	Pz_SpecCent_Alpha
Cz_AbsPow_Delta	F3_AbsPow_Theta	Fp2_RWE_Theta
P3_AbsPow_Gamma	C3_KFD_Beta	Fp2_WavEntr_Theta

- poor agreement → different paradigms & assumptions

Results

Feature Subset Evaluation

- subsets with good predictive power:
    - P1, F7, C4, P3 and P4 electrodes
    - beta wave subset
    - left hemisphere with midline
    - decision tree important features
- results largely varied between Decision Tree and SVM
  → different paradigms & assumptions

Overview

Revisiting the Objectives

Identify potential EEG biomarkers of depression

- prefrontal, frontal, temporal, and parietal region
- left hemisphere combined with the midline
- beta alone or a subset of alpha, delta and gamma waves combined

Revisiting the Objectives

Compare different XAI methods

Method	Advantages	Disadvantages
Correlation with Diagnosis	simplicity, directionality, model independence	linearity or monotonicity assumption, feature interactions not considered
Decision Tree Importance	easy to compute, hierarhical information	instability, no directional information
SHAP Values	model-agnostic, robustness, directionality	computationally expensive

Conclusion

General Takeaways

- understanding ML models ≠ understanding depression
  - model explainability ~ human-computer interaction
  - depression explainability ~ biomarkers
- single train-test split
  - does not make sense for subset feature evaluation
  - introduce any variability to asses significance

References

S. Nagel, “*Towards a home-use BCI: fast asynchronous control and robust non-control state detection*,” Ph.D. dissertation, Universität Tübingen, 2019 **image from slide 5
C. A. Ellis, M. L. Sancho, R. L. Miller, and V. D. Calhoun, “*Identifying EEG Biomarkers of Depression with Novel Explainable Deep Learning Architectures*,” bioRxiv 2024.
F. Chen, L. Zhao, L. Yang, J. Li, and C. Liu, “*An Explainable Assessment for Depression Detection Using Frontal EEG*,” in Asian-Pacific Conference on Medical and Biological Engineering. Springer, 2023, pp. 377–383.
J. Yang, Z. Zhang, P. Xiong, and X. Liu, “*Depression Detection Based on Analysis of EEG Signals in Multi Brain Regions*,” Journal of Integrative Neuroscience, vol. 22, no. 4, p. 93, 2023.
S. M. Lundberg and S.-I. Lee, “*A Unified Approach to Interpreting Model Predictions*,” Advances in Neural Information Processing Systems, vol. 30, 2017

Thank you.

AUTHOR: FANI SENTINELLA-JERBIĆ
fani.sentinella.jerbic@gmail.com

MENTOR: PROF. DR. SC. MARIO CIFREK
mario.cifrek@fer.unizg.hr

ASSISTANT: EDA JOVIČIĆ
eda.jovicic@fer.unizg.hr