Statistična analiza podatkov v časovnem zaporedju (EEG)

Jan Škorjanc

Vsebina

- Primerjava delovanja možganov zdravih ljudi z možgani epileptikov (EEG)
- Uporaba statističnih metod:
  - SVD
  - PCA
  - tSNE
- Klasifikatorja: k-NN in logistična regresija

Uvod v elektroencefalografijo

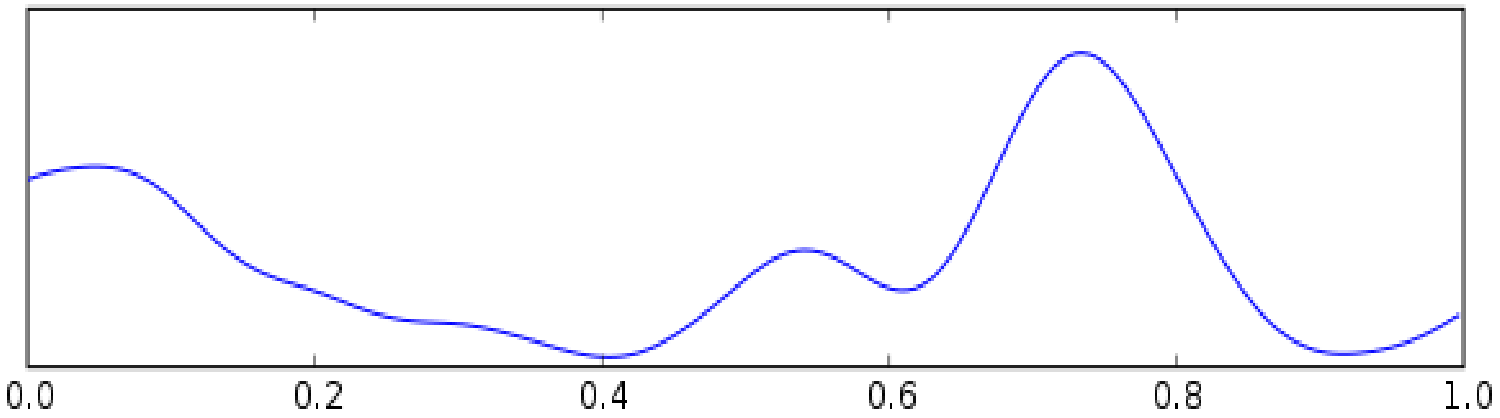
- Postopek merjenja možganske električne aktivnosti z uporabo elektrod
- Signal = vsota električnih potencialov možganskih celic
- Izmerjeni v mikrovoltih
- Čas meritve signalov: 20min
- Lahko v spanju ali budnem stanju

Uvod v elektroencefalografijo

- Diagnoza: epilepsija, možganska kap, tumor, uporaba drog, poškodbe glave…
- Ugotavljanje globine kome
- Pomen: razlikovati splošen možganski vzorec od vzorca, ki ga povzročajo motnje
- Meritve v milisekundah

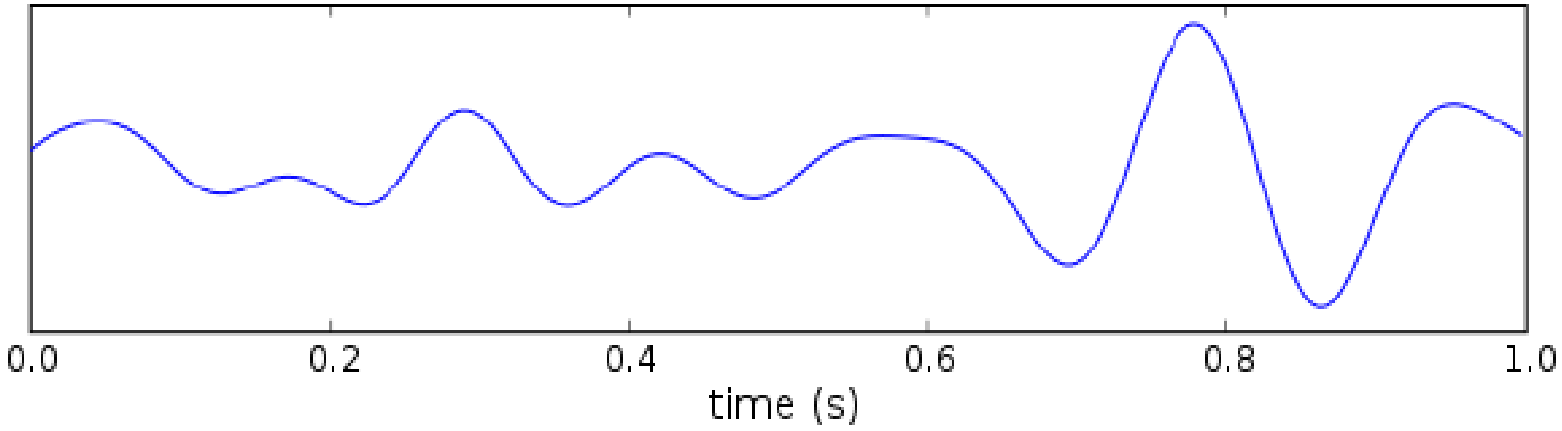
Vrste valov EEG

- Delta val (0,1Hz - 4Hz)
  - Največja amplituda
  - Najbolj počasna možganska aktivnost
  - Pogost pri spanju in globokem dihanju



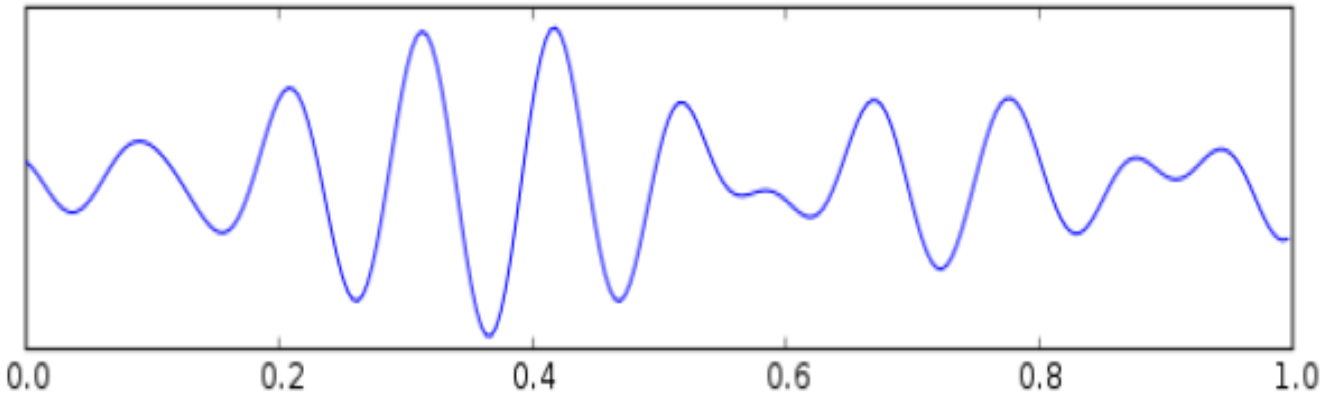
Vrste valov EEG

- Theta val (4Hz - 8Hz)
  - Značilen za mladostnike
  - Pojavljanje tudi v spanju



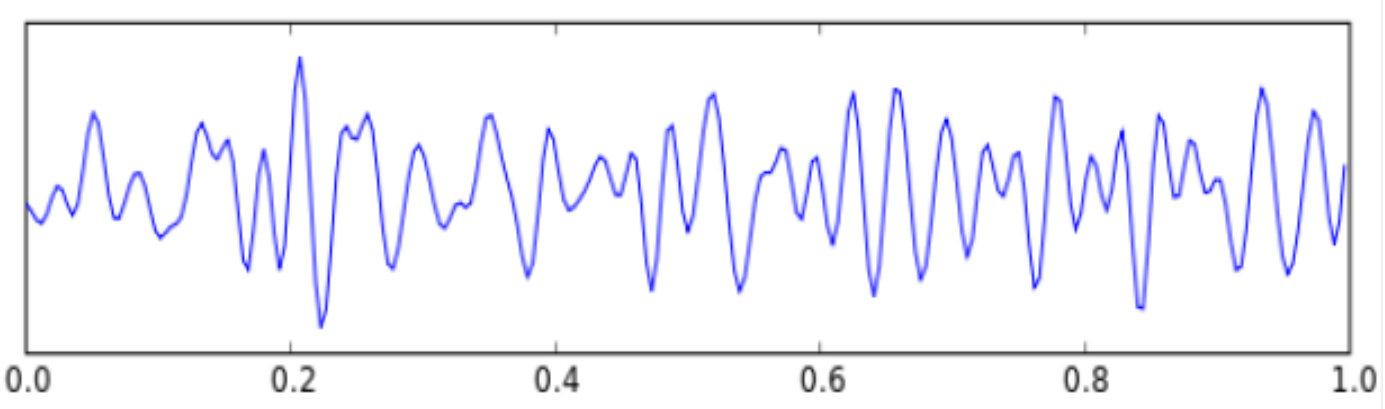
Vrste valov EEG

- Alfa val (8Hz - 15Hz)
  - Značilen za odrasle ljudi v budnem stanju



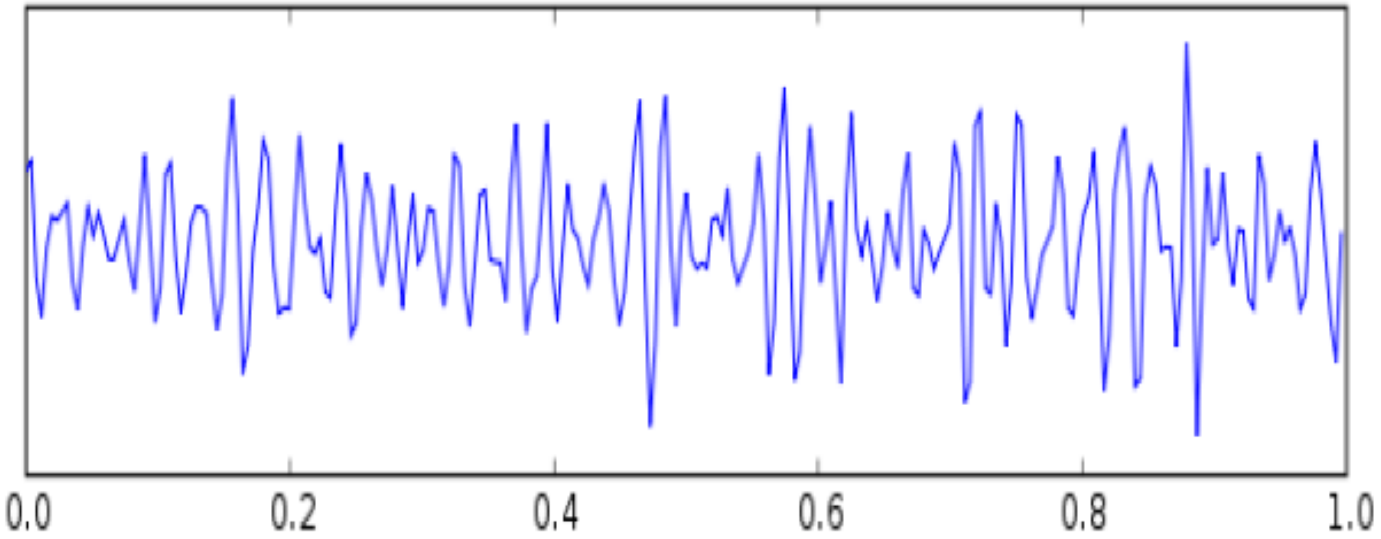
Vrste valov EEG

- Beta val (15Hz - 31Hz)
  - Pojavljanje pri govoru in reševanju nalog
  - Opazen na delu možganov namenjenem razmišljanju



Vrste valov EEG

- Gama val (>31Hz)
  - Pojavljanje pri opravljanju zahtevnih nalog, zaznavanjem z več čutili in visoki koncentraciji



Izbira podatkov

- 5 podatkovnih zbirk (A-E) ločenih glede na lastnost testiranca
  - A - zdrava oseba, odprte oči
  - B - zdrava oseba, zaprte oči
  - C - možganski tumor, merjenje na zdravem delu možganov
  - D - možganski tumor, merjenje na ne zdravem delu možganov
  - E - epileptiki

Izbira podatkov

- Uporaba enake naprave za vse
- Frekvenca zapisa: 173,61Hz
- Razlika dveh vrednosti je tako v razmaku približno 5ms

Opis metod

- SVD (Singular value decomposition)
  - Algoritem, ki poskuša zmanjšati rang matrike s pomočjo linearne kombinacije vektorjev
  - Ena izmed metod SVD je tudi PCA
- PCA (Principal component analysis)
  - Cilj metode je poiskati nekaj spremenljivk, ki vsebujejo največji del razpršenosti podatkov
  - Ostanejo podatki, ki nosijo največji del informacije
  - Problem je lahko izguba prevelikega dela informacije

Opis metod

- T-SNE
  - Algoritem strojnega učenja, ki je namenjen zmanjšanju dimenzije podatkov
  - Podatke dobro pretvori v dve dimenziji (graf)
  - Podatki, ki so si bližje na grafu so med seboj bolj povezani

Opis metod

- Logistična regresija
  - Regresija, pri kateri je odvisna spremenljivka kategorična (1,0; da/ne; epileptik/ni epileptik…)
  - Podobna linearni regresiji, vendar je transformirana z logistično funkcijo

Opis metod

- K-NN (k-nearest neighbors)
  - Algoritem izračuna razdaljo vseh instanc na podlagi razreda in klasificira v razred k najbližjih sosedov
  - K določi uporabnik in je ponavadi manjše celo število
  - Časovno zahteven vendar enostaven
  - Poznam več načinov izračuna razdalj
  - Uporabljena evklidska razdalja

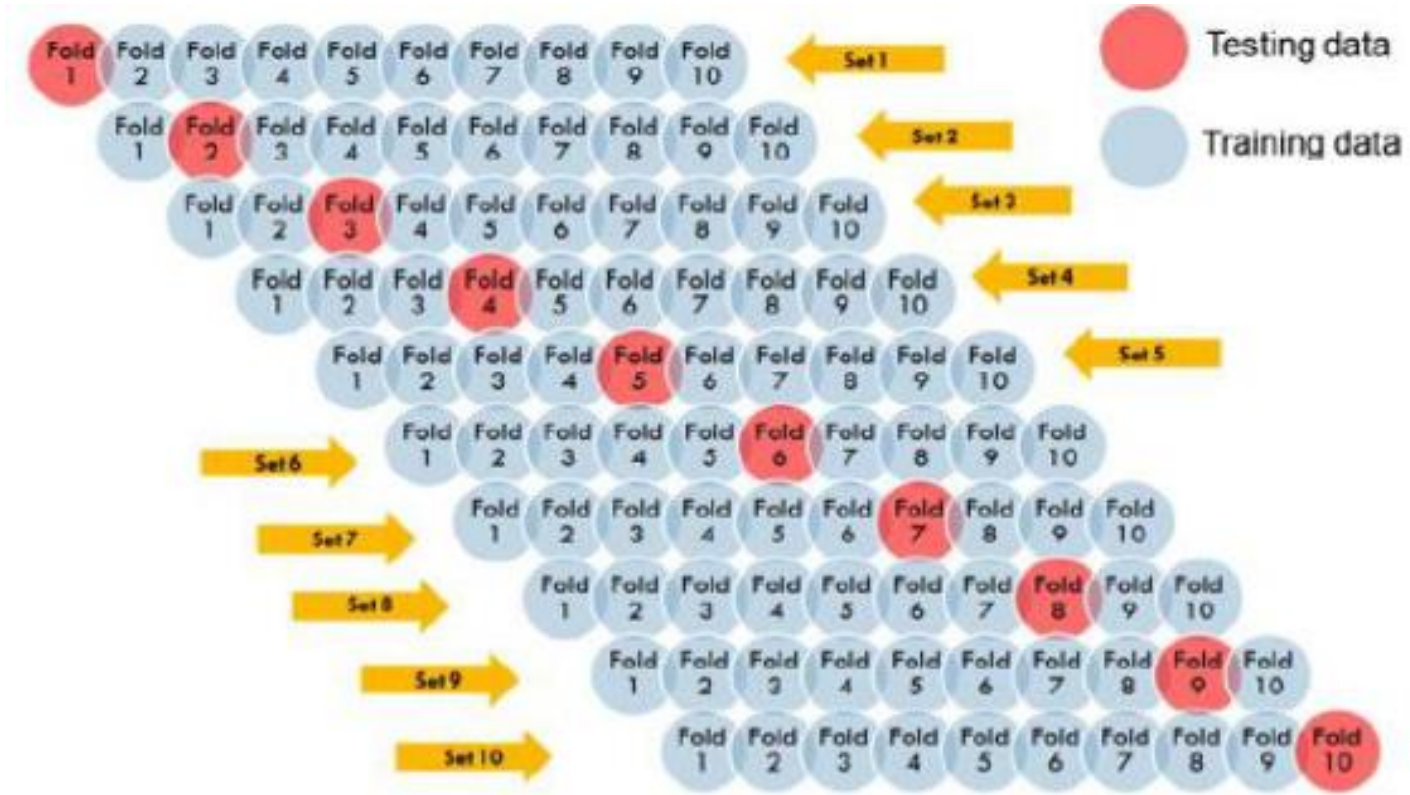
$$d(a,b) = \sqrt{(a_1 - b_1)^2 + \cdots (a_n - b_n)^2}$$

Opis metod

- K-kratno prečno preverjanje
  - Način vrednotenja algoritmov
  - Podatkovno zbirko razdeli na k enako velikih disjunktnih podmnožic
  - Učni model na k-1 podmnožicah
  - Testiranje na preostali množici
  - Rezulat predstavlja povprečje iteracij

Opis metod

- K-kratno prečno preverjanje



Uporabljena programska oprema

- Weka
  - Odprtokoden program, ki vsebuje zbirko algoritmov statistike in strojnega učenja
  - Napisan v javi
  - Podpira splošne algoritme za klasifikacijo, regresijo, razvrščanje v skupine in ostale metode podatkovnega rudarjenja.
  - Podpira formate datotek CSV, LibSMV in C4.5

Uporabljena programska oprema

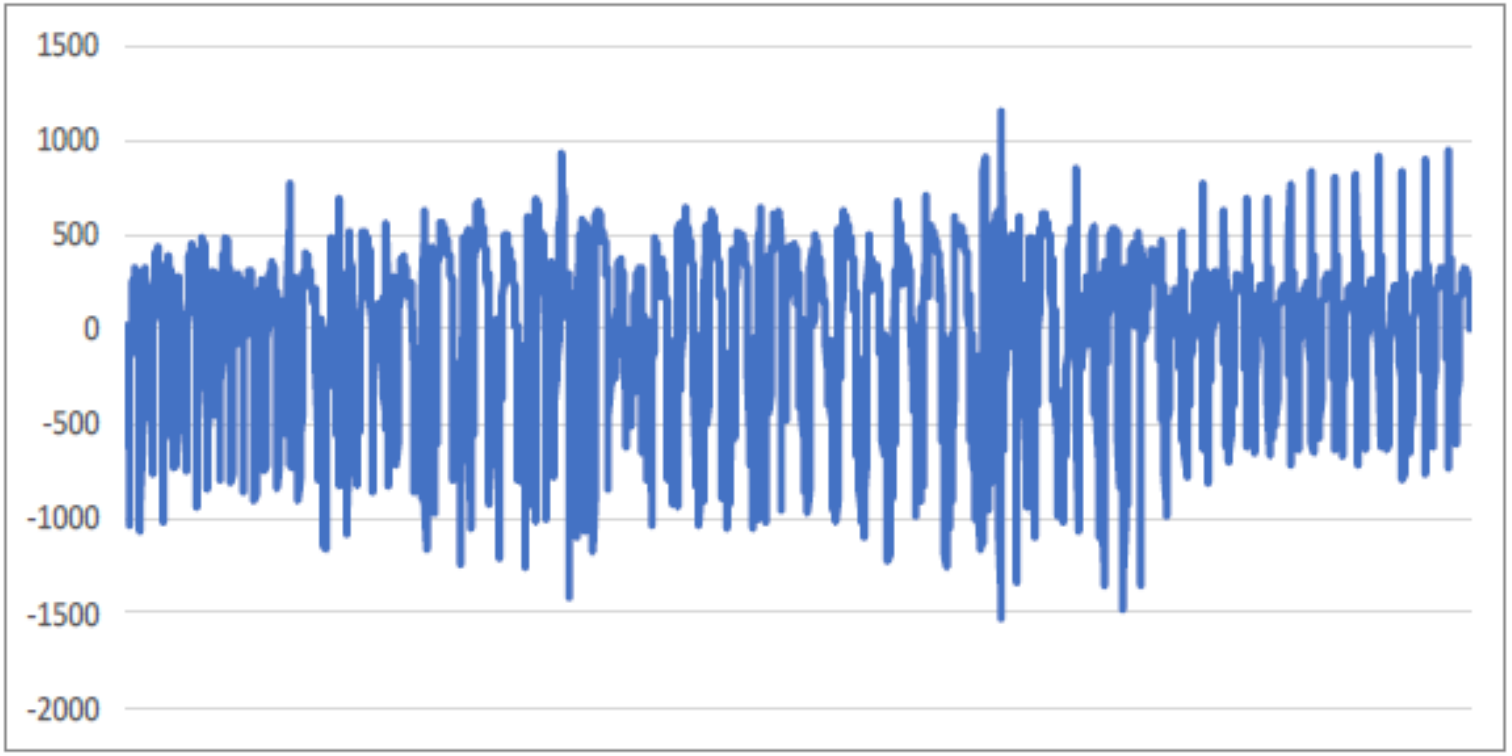
- R
  - Programski jezik namenjen statističnem programiranju in vizualizaciji
  - Eden bolj pogosto uporabljenih jezikov
  - Podpora z veliko paketi

Postopek dela in rezultati

- Priprava podatkov
  - Zmanjšanje začetne podatkovne zbirke iz okoli 4000 (23 sekund) atributov na 173 + razredni atribut (1 sekunda)
  - Določanje vrednosti razredov glede na epilspsijo
  - A-D -> N (ni epileptik)
  - E -> Y (epileptik)

Postopek dela in rezultati

- Primer naključnega EEG v 23 sekundah



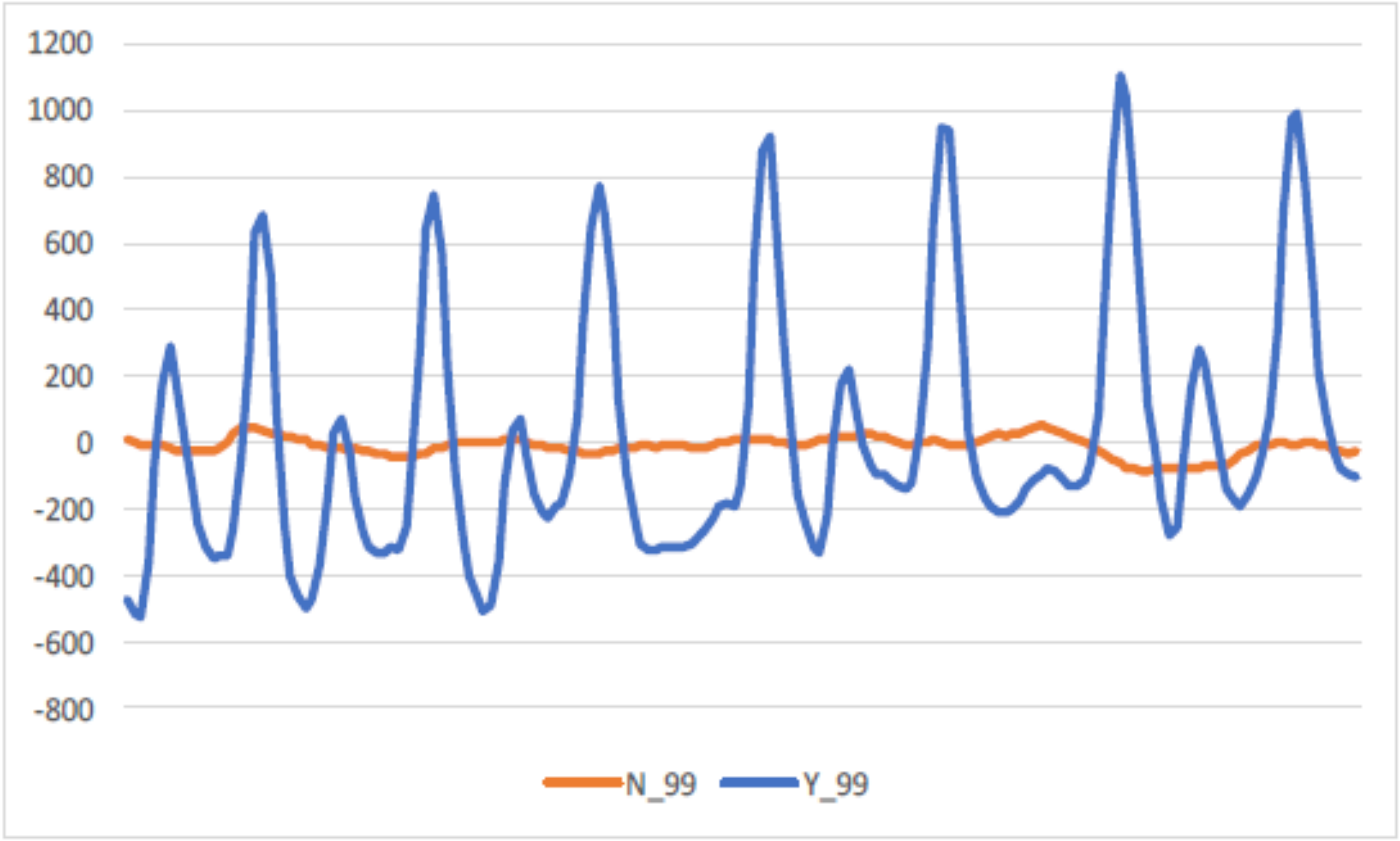
Postopek dela in rezultati

- Osnovni statistični podatki

	N	Y
Minimum	-889	-1885
Maksimum	2047	1624
Mediana	-8	-5
Povprečje	-9,25	-7,92
Standardni odklon	70,70	347,14

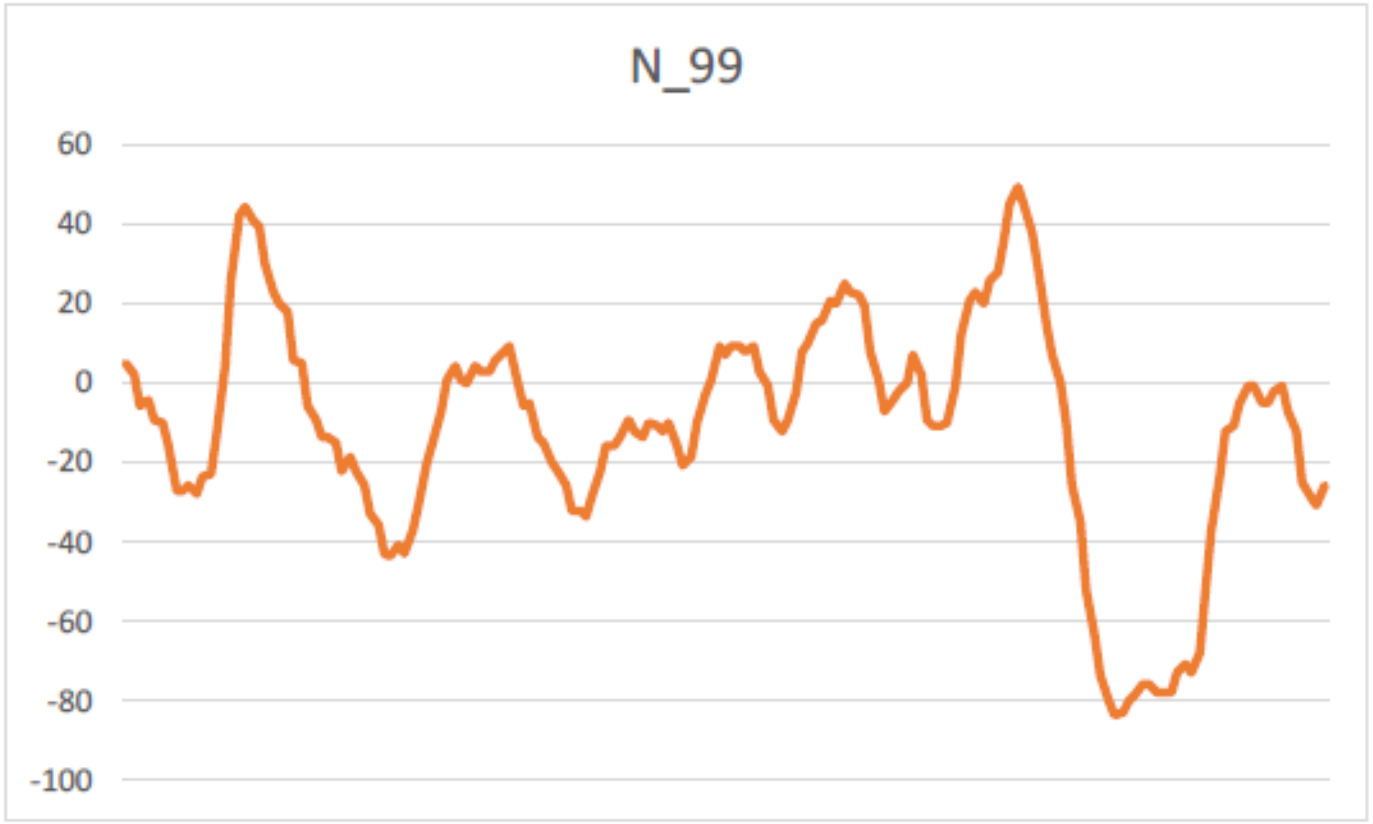
Postopek dela in rezultati

- Prikaz naključnega EEG v eni sekundi



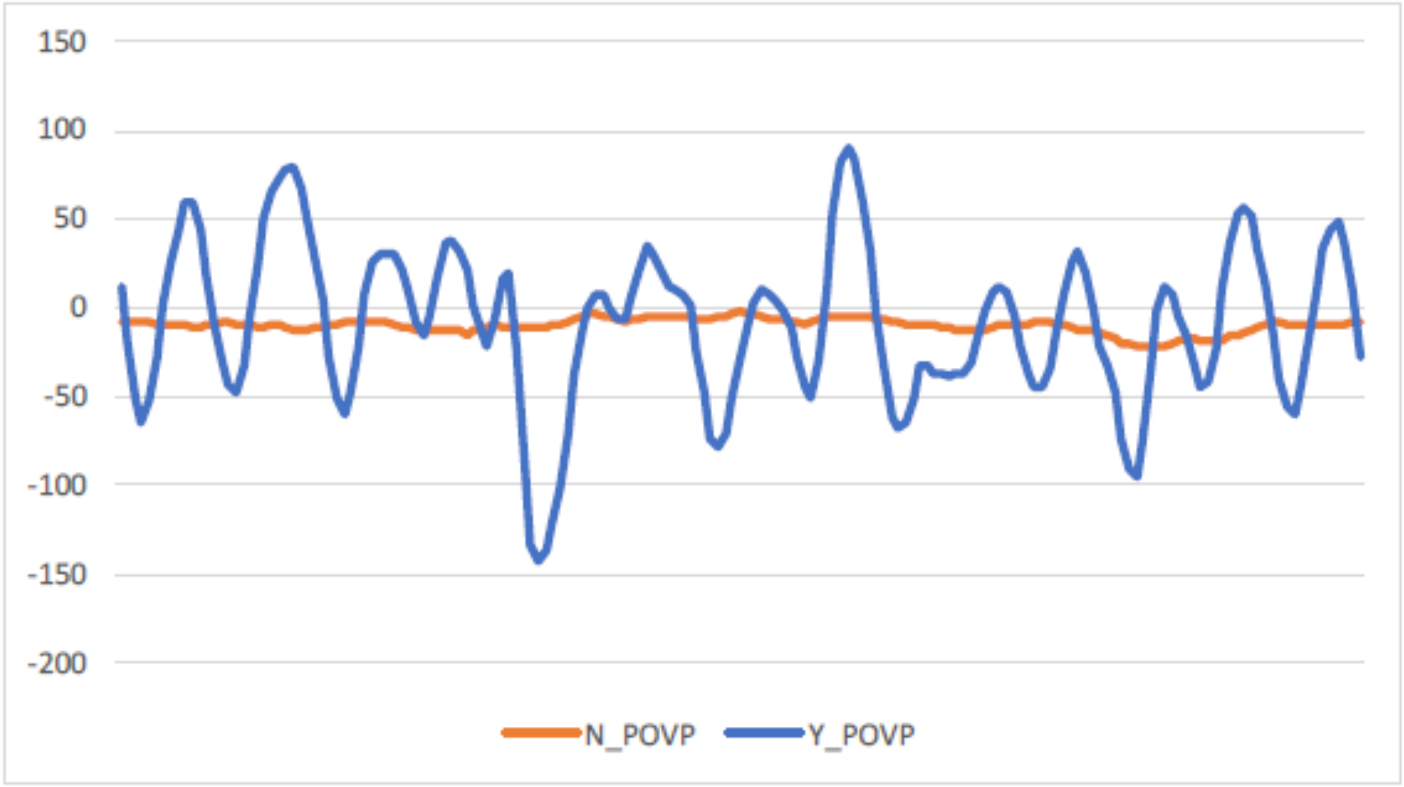
Postopek dela in rezultati

- Prikaz neepileptika EEG v eni sekundi



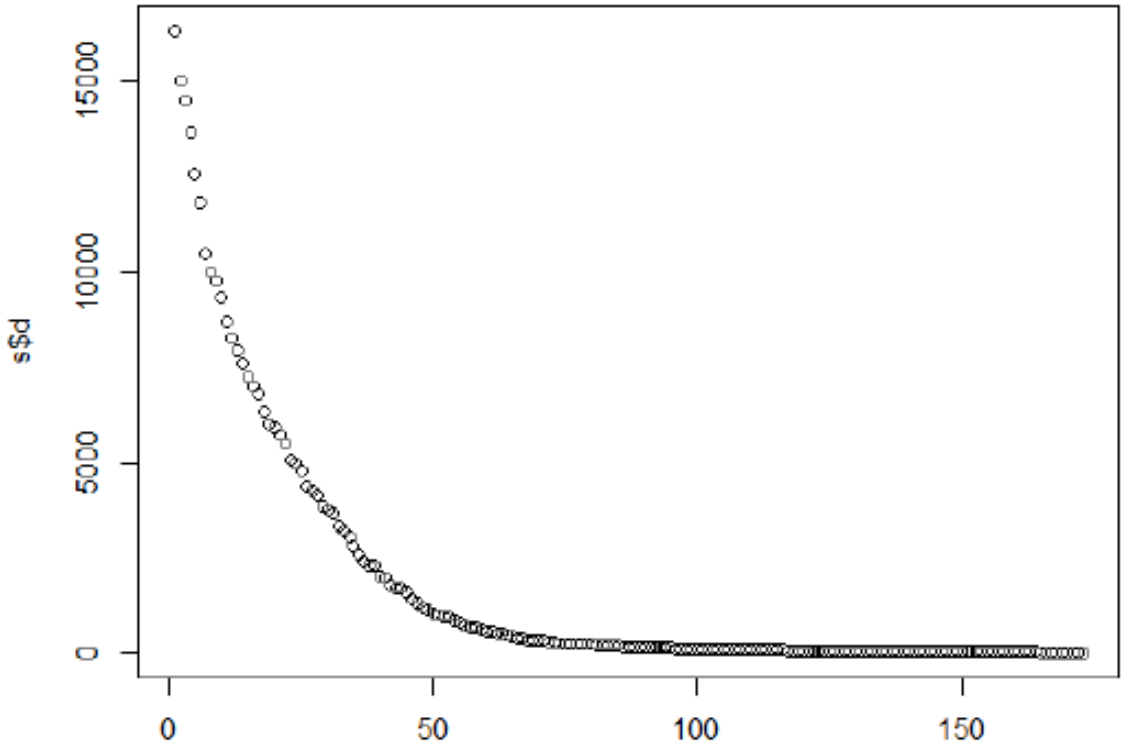
Postopek dela in rezultati

- Primerjava povprečnih vrednosti



Postopek dela in rezultati

- Pomembnost atributov po SVD metodi
  - Najbolj pomembnih prvih 50 atributov

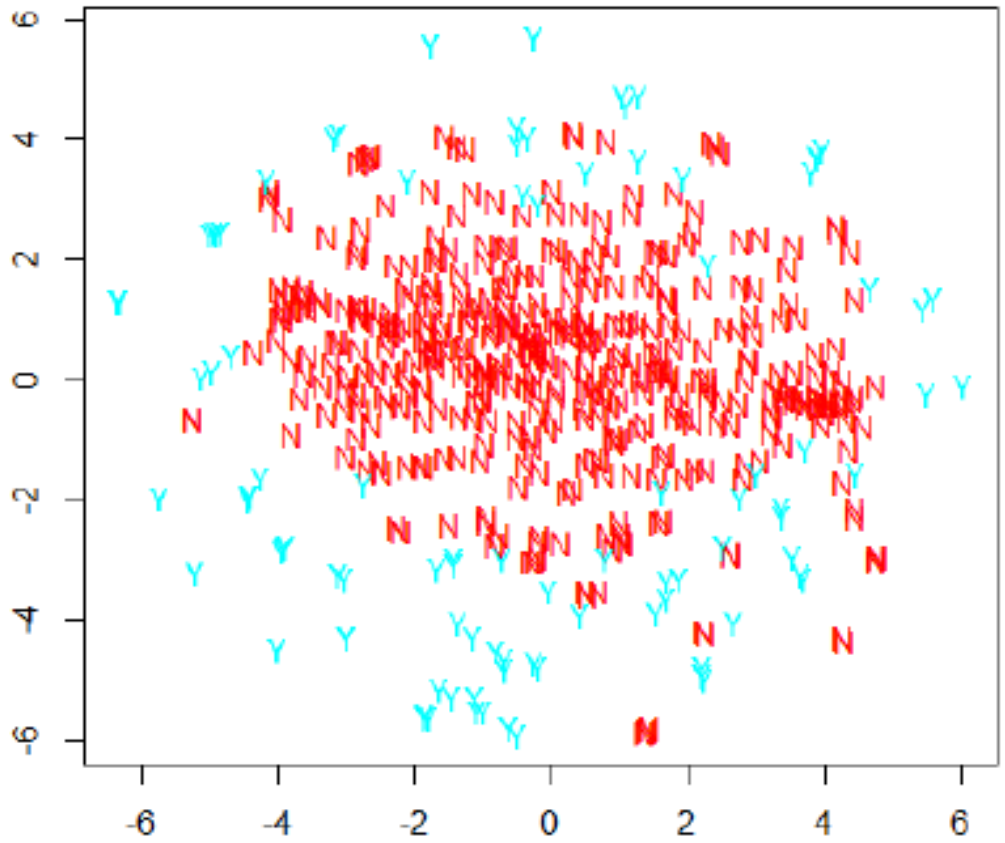


Postopek dela in rezultati

- PCA
  - Vhodni parameter vsebovanosti informacije 95% nam ponudi iz 173 atributov podatke zmanjša na 30
  - Uporaba algoritma v Weki

Postopek dela in rezultati

- T-SNE (2D prikaz podatkov)



Postopek dela in rezultati

- Rezultati klasifikacije (%)

	k-NN	Logistic regression
Originalni podatki	93,4	85,4
SVD	93.4	87.4
PCA	89,4	88
t-SNE	94,2	83,2

Ugotovitve

- Razlika med epileptiki in neepileptiki jasno vidna že iz osnovnih statističnih metod znotraj 1 sekunde
- Visoka uspešnost algoritmov za strojno učenje to potrdi
- Z uporabo k-NN algoritma je uspešnost okoli 93% nadpovprečno dobra

HVALA ZA POZORNOST ☺

Vprašanja