SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Klasifikacija EEG signala za detekciju PTSP-a

SEMINAR
OTKRIVANJE ZNANJA U SKUPOVIMA PODATAKA

Iva Harbaš

Zagreb, veljača 2013.

Sadržaj

1. **Uvod**

Ljudski mozak je kompleksan sustav koji je jako dinamičan i čija aktivnost se može mjeriti na različite načine. Onaj koji je ovdje interesantan je EEG (elektroencefalogram) pomoću kojeg se mjere električni potencijali koje stvaraju neruroni i neuronske veze u mozgu. EEG signali se koriste u svrhu istraživanja različitih oboljenja među kojima i PTSP (posttraumatski stresni poremećaj). Dijagnosticiranje PTSP-a pomoću EEG signala nije toliko razvijeno kao npr. dijagnosticiranje i detekcija epilepsije, ali se u zadnje vrijeme sve više pažnje posvećuje tom problemu. Proučavanje električne aktivnosti mozga putem EEG-a je jedan od važnijih alata u dijagnozi. Prilikom snimanja EEG-a ne dobije se jedan signal, nego broj snimljenih signala ovisi o broju kanala koji se snima (može biti i do 128 kanala) i kao takav, jedan EEG sadži veliki broj informacija koje ponekad nije moguće obraditi vizualno. Abnormalnosti u EEG-u kod nekih ozbiljnih poremećaja su ponekad presuptilne da bi se uočile koristeći konvencionalne tehnike i upravo zbog toga se posvećuje velika pažnja korištenju računala za riješavanje problema i razvoja automatskih sustava koji bi za zadatak imali prepoznavanje promjena u EEG-u, donošenja zaključaka te postavljanje dijagnoza.

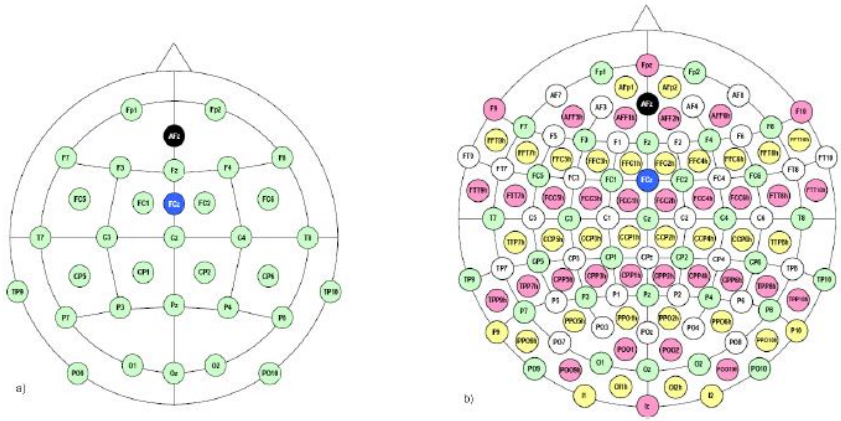
Cilj ovog rada je primjeniti znanja i metode iz područja otkrivanja znanja u skupovima podataka. Konkretno, za klasifikaciju signala koristiće se strojno učenje, te za određivanje kvalitete klasifikatora koristiće se metode krosvalidacije i izražavanje pogreške konfuzijskim matricama.

3. EEG

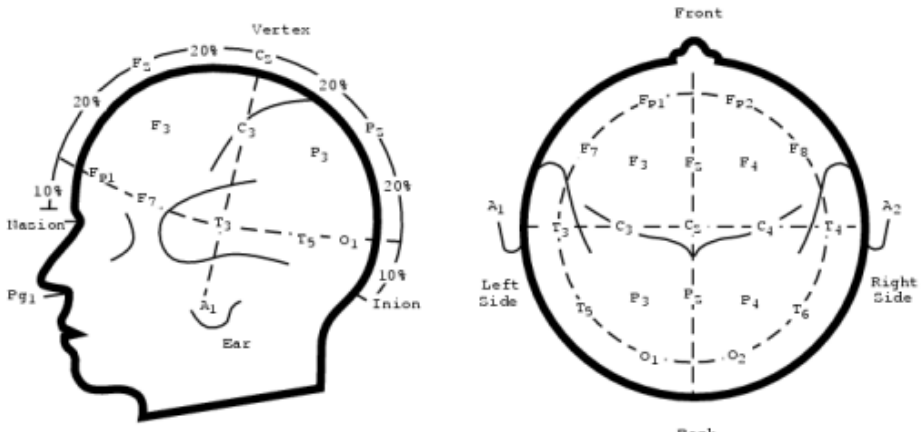
Elektroencefalografija (EEG) je proces snimanja električne aktivnosti duž skalpa. EEG mjeri promjene napona koje su posljedica protoka ionske struje među neuronima u mozgu. U kliničkom smislu, EEG se odnosi na snimanje sponatane moždane aktivnosti kroz neki vremenski period, najčešće od 20 do 40 minuta.

3.2. Snimanje EEG-a

EEG signali se snimaju preko elektroda koje se s provodljivim gelom ili pastom postavljaju na skalp. Broj elektroda koje se postavljaju na skalp nije fiksan, tj. EEG se može mjeriti pomoću manje ili više elektroda (slika 1.). Položaj elektroda na skalpu te njihovi nazivi su specificirani internacionalnim 10-20 sustavom. Raspored elektroda se može vidjeti na slici 2.



Slika 1.: Kapica sa rijeđe i sa gušće postavljenim elektrodama



Slika 2.: Raspored elektroda po sustavu 10-20

Prilikom snimanja EEG signala pacjent se podvrgava različitim procedurama koje mogu izazvati normalne i abnormalne EEG aktivnosti koje se nebi drugačije mogle primjetiti. Neke od precedura su hiperventilacija, fotostimulacija, zatvorene/otvorene oči i sl.

3.3. Valni oblici

U EEG-u se mogu uočiti četiri karakteristična signala po valnom obliku, veličini amplitude i frekvenciji. Oni se nazivaju α-valovi (alfa valovi), β-valovi (beta valovi), θ-valovi (theta valovi) i δ-valovi (delta valovi).

U tablici 1. prikazani su svi valni oblici sa osnovnim karakteristikama radi lakše usporedbe ([1]).

Tablica 1.: Valni oblici EEG signala i njihove osobine

Vrsta vala	Frekvencije	Amplituda	Lokacija	Normalno	Patološki
α-valovi	8 – 13 [Hz]	50 [μV]	Okcipitalna regija	-opušteno stanje -zatvorene oči -također povezano s kontrolom inhibicije	-koma
β-valovi	14 – 30 [Hz]	20 [μV]	Frontalna i parijetalne regije	-aktivno stanje -zaposleno ili anksiozno razmišljanje -aktivna koncentracija -otvorene oči	-uzimanje nekih lijekova kao što je benzodiazepin
θ-valovi	4 – 8 [Hz]	70 [μV]	Parijetalne i temporalne regije	-kod mlade djece -napetost kod starije djece i odraslih -kod odraslih pri emocionalnim stresovima	-fokalne subkortikalne lezije -metabolička encefalopatija -neke vrste *hydrocephalus*-a
δ-valovi	0,5 – 3,5 [Hz]	60 – 100 [μV]	Kora velikog mozga	-duboki san kod odraslih -kod beba	-subkortikalne lezije -difuzne lezije -metabolička encefalopatija

3.6. Korišteni EEG signali

EEG signali koji su korišteni u ovom radu su spremljeni u *.edf* formatu koji je u Matlab učitavan pomoću *EEGlab toolboxa* jer se pomoću njega EEG signal učita u jednu strukturu u kojoj su sadržane sve potrebne informacije kao što su frekvencija uzorkovanja, broj kanala, nazivi kanala, signali spremljeni u matrici, itd (signal se učitava pomoću funkcije implementirane u *toolboxu pop_biosig()*) i zbog toga je pogodniji za daljnju analizu i korištenje.

Korišteni *toolbox* je tipa *open source* i služi za analizu EEG signala. Da bi ga skinuli sa interneta najprije je potrebno otići na stranicu: http://sccn.ucsd.edu/~scott/ica.html, tamo popuniti obrazac u kojem objasnite za što vam treba *toolbox* i onda dobijete dozvolu da ga skinete na svoje računalo. Nakon što se skinuta *.rar* arhiva otpakira na računalu potrebno se unutar Matlaba pozicionirati u taj folder te u komandnom prozoru Matlaba ukucati „eeglab“.

Svi korišteni EEG signali su snimani ujutro, u ležećem položaju, u mirovanju, sa zatvorenim očima nakon što bi pacjent doručkovao. Birani su isječci bez artefakta, odnosno s ciljanim artefaktima. Svi pacjenti su bili muškarci, ne stariji od 55 godina. Većina je bila pod terapijom, ali isječci na kojima je bio evidentan utjecaj lijekova s posljedično promijenjenim EEG-om nisu korišteni u ovom istraživanju. Također se nisu koristili isječci pacijenata koji su imali neku neurološku bolest jer tada ne bismo mogli promatrati povezanost EEG-a i psihijatrijskog poremećaja.

Korištene elektrode su kositrene i pozicionirane su po klasičnom 10-20 sustavu. Mjerenje napona je bipolarno s ukupno 20 kanala, a aparat korišten za snimanje signala je Medialov, TG Valor T40 T64 T128, Nervus v3.x.

4. Klasifikacija EEG signala

Postupak klasifikacije se sastoji iz tri osnovna koraka:

1. predprocesiranje i odabir značajki
2. treniranje klasifikatora koristeći odabrane značajke
3. testiranje klasifikatora.

Za predprocesiranje značajki se koristi ANOVA (eng. *Analysis of Variance*) na način koji će biti opisan kasnije. Klasifikator koji je odabran za obavljanje klasifikacije u ovom radu je SVM (eng. *Support Vector Machines*) koji će također biti opisan u nastavku rada zajedno sa rezultatima testiranja.

4.1 ANOVA

Struke koje svoje nove spoznaje pretežno izvode iz podataka u pravilu polaze od uzorka ispitanika. Na uzorku se izvode mjerenja i dobivaju se informacije u obliku aritmetičkih sredina, varijanci, proporcija i sličnog, a onda se dobiveno želi poopćiti na populaciju iz koje je uzorak uzet.

Testiranje hipoteze je statistički postupak kojim se određuje da li i koliko pouzdano raspoloživi podaci podupiru postavljenu pretpostavku. Testiranje hipoteza, odnosno testiranje značajnosti u osnovi je postupak kvantifikacije impresija o specifičnoj hipotezi.

ANOVA je računski postupak pomoću kojega se ispituju podaci određenoga pokusa, kroz procjenu otklona pojedinih srednjih vrijednosti od prosječne vrijednosti uzoraka uzetih iz nekog osnovnog skupa[3].

Kroz niz relativno jednostavnih izračunavanja potrebno je dobiti *F* vrijednost. Sinonim za izračunavanje *F* vrijednosti je *F*-test[1] ili grupni test za ispitivanje hipoteze pokusa.

F-testom se ispituje, pokusom postavljena, nulta hipoteza[2] da su aritmetičke sredine *k* osnovnih skupova ili tretmana međusobno jednake, odnosno, da u cjelini nema statistički značajne razlike. Cilj je ispitati odnos varijacija između uzoraka s varijacijama unutar uzoraka. Ako je taj odnos, tzv. empirijski *F-omjer*, statistički značajan zaključujemo kako promatrani uzorci ne pripadaju istoj populaciji, odnosno aritmetičke sredine se značajno razlikuju.

ANOVA se u ovom radu koristi za predprocesiranje značajki, tj. za pronalaženje onih značajki po kojima se dvije grupe signala (bolesni i zdravi) najviše razlikuju. ANOVA-om se testira nul-hipoteza da obje grupe značajki (značajka izračunata za zdrave signale i značajke izračunate za bolesne signale) pripadaju istoj populaciji što znači da će u ovom slučaju biti zanimljiv rezultat gdje se dobije niska vjerojatnost sličnosti tih grupa.

Cilj predprocesiranja značajki je smanjenje dimenzionalnosti vektora značajki koji se koristi u treniranju i testiranju sustava. Smanjivanje dimenzionalnosti je od velikog značaja jer svaki EEG signal ima 20 kanala, što zapravo predstavlja 20 signala. Tako bi se za jednu značajku jednog EEG signala dobio vektor od 20 značajki. Pored smanjivanja dimenzionalnosti, rezultatima ANOVA-e se može dobili bolji uvid u promjene EEG signala po kanalima. Tako npr. prilikom računanja snage beta pojasa (jedna značajka) primjenom ANOVA-e kao rezultat se dobije u kojem kanalu snaga beta pojasa pokazuje najveću razliku između normanih i PTSP EEG signala i taj kanal se smatra zanimljivim i njega ćemo koristiti za klasifikaciju.

---

[1] F-test je samo jedan od testova za provjeru hipoteza. Pored F-testa najčešće se koriste još T-test i χ^2 test

[2] *Nul-hipoteza,* H_0 (eng. *null hypothesis*) je pretpostavka o izostanku efekta, tj. da ne postoji razlika među uzorcima u populaciji od interesa (npr. nema razlike u aritmetičkim sredinama). To je hipoteza koja se testira, *hipoteza da nema razlike* (eng. *hypothesis of no difference*).

4.1.1. Implementacija

ANOVA je implementirana u Matlabu kao dio *Statistics Toolbox*-a i u sklopu istog postoje tri različite izvedbe i to:

- *One - way ANOVA* - ANOVA s jednim promjenjivim faktorom (karakteristikom)
- *Two – way ANOVA* - ANOVA s dva promjenjiva faktora
- *N – way ANOVA* - ANOVA s N promjenjivih faktora

Svrha *one-way* ANOVA-e je saznati da li podaci iz više skupina imaju zajedničku aritmetičku sredinu, tj. ustanoviti da li se skupine stvarno razlikuju u izmjerenoj karakteristici. *Two-way* ANOVA se razlikuje od *one-way* ANOVA-e u tome što su skupine u *two-way* ANOVA-i opisane sa dvije karakteristike. Analogno tome skupine u *N-way* ANOVA-i se opisuju sa *N* karakteristika.

U svrhu predprocesiranja značajki EEG signala za ovaj rad korištena je *one-way* ANOVA u kojoj se testiraju odabrane značajke (karakteristike) EEG signala za svaki kanal pojedinačno. Značajke su podjeljene u dvije grupe, značajke dobivene iz normalnih EEG-ova i značajke dobivene iz PTSP EEG-ova.

Implementirana funkcija u Matlabu koja realizira *one-way* ANOVA-u je[2]:

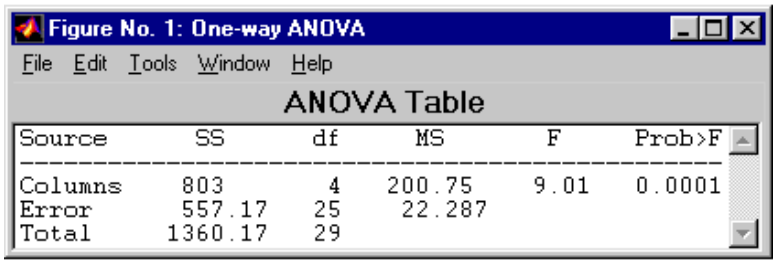
```
[p,tbl] = anova1(M);
```

gdje su:

- `M` – matrica sa dva stupca od kojih svaki predstavlja jednu skupinu (zdravi i bolesni) dok se u redovima matrice nalaze izračunate vrijednosti značajki
- `p` – vjerojatnost sličnosti dviju skupina
- `tbl` – standardna ANOVA tablica (tabica 2.) koja prikazuje izračunate vrijednosti sume kvadrata (eng. *Sum of Squares* (SS)), stupnjeve slobode (eng. *Degrees of Freedom* (df)), varijance, *F vrijednost* i *p*. Primjer dobiven iz Matlaba je prikazan na slici 3.

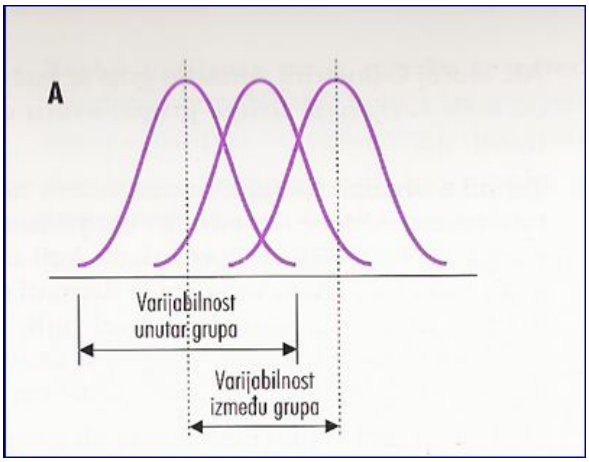
Tablica 2.: ANOVA tablica

Izvor varijacije	Zbroj kvadrata odstupanja (SS)	Stupnjevi slobode (df)	Sredine kvadrata odstupanja (MS)	F	p
Između uzoraka	$\sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2$	$k-1$	$S_A^2 = \frac{\sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2}{k-1}$	$\frac{S_A^2}{S_U^2}$	
Unutar uzoraka	$\sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2$	$n-k$	$S_U^2 = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}{n-k}$		
Ukupno	$\sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \bar{X})^2$	$n-1$			

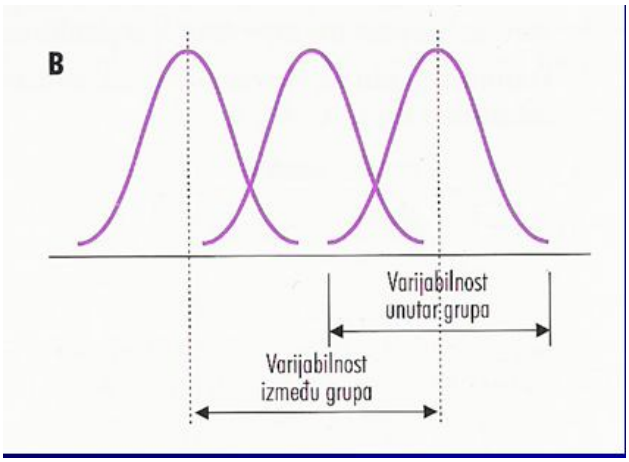


Slika 3.: ANOVA tablica dobivena Matlabom

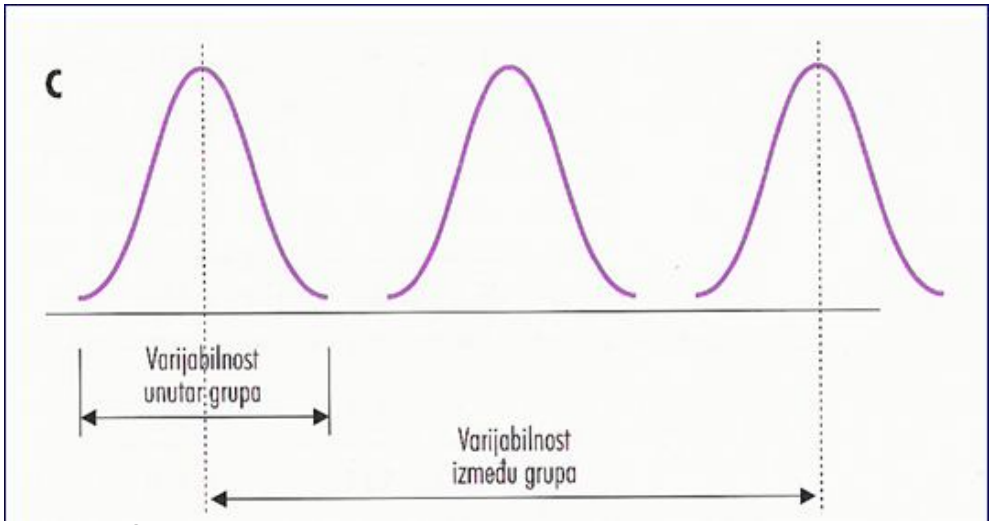
Grafički prikaz rezultata kad je varijabilnost između grupa manja od varijabilnosti unutar grupa je prikazan na slici 4. Za klasifikaciju EEG signala ovaj rezultat nije povoljan. Ono što se želi dobiti je veća varijabilnost između grupa kao što je prikazano na slici 5. To bi značilo da su grupe različite, tj. da postoji razlika između zdravih i bolesnih signala za zadanu značajku. Idealno bi bilo dobiti odnos kao na slici 6. gdje je varijabilnost između grupa puno veća od varijabilnosti unutar grupa.



Slika 4.: Varijabilnost između grupa < varijabilnost unutar grupa



Slika 5.: Varijabilnost između grupa > varijabilnost unutar grupa



Slika 6.: Varijabilnost između grupa >> varijabilnost unutar grupa

4.2. Značajke

Odabir značajki je najbitnija faza u procesu klasifikacije signala. Potrebno je naći one značajke koje dobro opisuju razlike između klasa koje se žele klasificirati. Tu nam u pomoć pristiže ANOVA, ali najprije se trebaju odabrati značajke na kojima će se ANOVA primijeniti.

U prvoj fazi su uzete sve značajke korištene za klasifikaciju EEG signala koje su spominjane u srodnoj literaturi. Zatim je slijedio proces eliminacije onih koje su bile računski (vremenski) zahtjevne. Nakon toga su eliminirane one značajke koje su specifične za druge bolesti (npr. epilepsija), a koje nisu pokazivale bitne razlike uspoređujući PTSP i normalne EEG-ove.

Nakon svih odabira i eliminacija došlo se do konačnog popisa značajki koje će se koristiti i to:

1. PSD srednja snaga α, β, δ i θ-pojasa
2. srednja snaga α, β, δ i θ-pojasa izračunata pomoću DWT koeficijenata
3. *Higuchieva* fraktalna dimenzija
4. *skewness*
5. *kurtosis*
6. *Hjortovi* parametri: aktivnost, mobilnost i kompleksnost.

4.2.1. PSD

PSD (eng. *Power Spectral Density*)[4] prikazuje kako je snaga nekog vremenskog signala raspoređena po frekvencijama. Drugim riječima, pomoću PSD-a se može vidjeti kolika je snaga signala na određenoj frekvenciji. To svojstvo je jako korisno kad se radi o EEG signalima jer se iz PSD-a mogu jasno očitati snage svih karakterističnih pojaseva EEG-a. Tako npr. pošto znamo da se α-valovi nalaze u frekvencijskom području od 8 Hz do 13 Hz iz PSD-a se može izračunati srednja snaga signala tog frekvencijskog pojasa i na taj način dobiti snaga α-pojasa. Analogno tome se računaju snage ostalih pojaseva EEG signala i upravo tako dobivene vrijednosti snage će se koristiti kao značajke EEG signala u procesu klasifikacije.

PSD se definira kao:

$$PSD(\omega) = F_T(\omega)F_T^*(\omega) \tag{1.}$$

gdje je $F_T(\omega)$ normalizirana Fourierova transformacija signala $f(t)$ koja se računa na slijedeći način:

$$F_T(\omega) = \frac{1}{\sqrt{T}}\int_0^T f(t)e^{-j\omega t}dt \tag{2.}$$

a $F_T^*(\omega)$ predstavlja konjugiranu Fourierovu transformaciju $F_T(\omega)$.

Koristeći navedene jednadžbe računaju se prosječne snage za sva četiri pojasa EEG-signala i te prosječne snage predstavljaju korištene značajke.

ANOVA i odabir značajki

Kao što je već spomenuto susrećemo se sa problemom velikog vektora značajki. Za svaki signal koji se sastoji od 20 kanala se izračuna prosječna snaga četiri pojasa što ukupno čini 80 izračunatih vrijednosti. Budući da će se koristiti još

značajki to bi vektor značajki učinilo još većim. Tu dolazi na red predprocesiranje značajki ANOVA-om.

Dakle, konstruira se matrica M (vidi poglavlje ANOVA) tako da se u prvi stupac matrice uvrste vrijednosti dobivene za snagu α-pojasa u prvom kanalu izračunate za zdrave EEG-ove, a u drugi stupac ide isto to samo što se računa za PTSP EEG-ove. Ova radnja se dalje ponavlja sa svaki kanal posebno, a kad se potroše svi kanali onda se cijeli postupak ponavlja za β-pojas. Ono što se dobije kao rezultat je vrijednost p za svaki kanal svake značajke. Vrijednost p predstavljasličnost dvije skupine (zdravih i bolesnih) što znači da nas zanimaju oni kanali gdje će vrijednost p biti najniža, tj. skupine se razlikuju.

Dobiveni rezultati se nalaze u tablici 3.

Tablica 3.: Razultati ANOVA-e

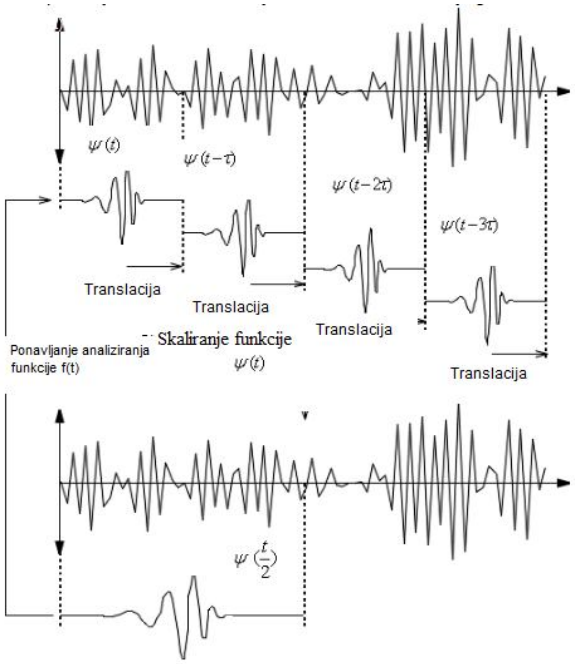
Značajka	PSD snaga α-pojasa	PSD snaga β-pojasa	PSD snaga δ-pojasa	PSD snaga θ-pojasa
Kanal	Vrijednost p	Vrijednost p	Vrijednost p	Vrijednost p
1	0,010779549	0,371136116	0,1868229	0,14376142
2	0,007604161	0,48317885	0,31433651	0,228254023
3	0,000110411	0,977935527	0,21123688	0,320246688
4	0,000469741	0,212278285	0,00237774	0,534677476
5	0,020787258	0,9160973	0,18758672	0,521787709
6	0,024423882	0,925467797	0,00152299	0,977214571
7	1,8931E-05	0,742513586	0,6906467	0,248651438
8	0,001167888	0,005971705	0,17370577	0,108749003
9	0,063115991	0,799064022	0,16246078	0,852341592
10	0,03066745	0,48151352	0,97587208	0,745847288
11	2,1813E-06	0,010540659	0,11481453	0,041568912
12	0,000673414	0,050129709	0,04118806	0,999217796
13	0,129883883	0,297825394	0,14067593	0,422901096
14	0,087179283	0,020403484	0,43944913	0,462010643
15	0,005084867	0,000542046	0,82026176	0,121701412
16	0,000102953	0,02256572	0,55793158	0,113940177
17	0,000292805	0,67865018	0,0108696	0,325894954
18	0,000470357	0,28071924	0,02556371	0,251498487
19	0,002279444	0,012096576	0,00082267	0,206534409
20	0,000641105	0,032760105	0,00675998	0,125276225

U tablici su zelenom bojom označene najniže, a crvenom najviše vrijednosti parametra p. Iz priloženog se vidi da će se koristiti slijedeće značajke:

1. PSD snaga α-pojasa u 11. kanalu
2. PSD snaga β-pojasa u 15. kanalu
3. PSD snaga δ-pojasa u 19. kanalu
4. PSD snaga θ-pojasa u 11. kanalu

4.2.2. DWT

Waveleti su funkcije koje mogu imati bilo kakav oblik, ali su vremenski ograničenef. Multirezolucijsko predstavljanje signala je osnovno načelo *wavelet* transformacije, koja za razliku od *Fourierove* transformacije signal prikazuje istodobno u vremenskoj i frekvencijskoj domeni. Signal se promatra u vremenskim intervalima i za svaki takav interval se računa spektar. *Wavelet* analiza je veoma slična *Fourierovoj*. *Fourierovom* analizom signal se predstavlja pomoću kosinusnih i sinusnih funkcija dok se kod *waveleta* prikazuje tzv. *wavelet* funkcijama. Sve *wavelet* funkcije generirane su iz iste funkcije, koja se zove osnovna ili *mother wavelet* funkcija, postupkom skaliranja i translacije koji je prikazan na slici 7.



Slika 7. Postupak skaliranja i translacije

Za dobivanje značajki u ovom radu koristiće se diskretna wavelet transformacija DWT(eng. *Discreete Wavelet Transform*)[5]. DWT se definira prema izrazu

$$X[m,k] = \frac{1}{\sqrt{a_0^k}} \int_{-\infty}^{\infty} x(t)\psi(\frac{t}{a_0^k} - mT)dt \qquad (3.)$$

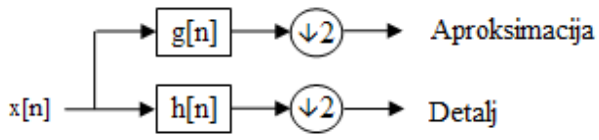
gdje su:

- $X[m,k]$ - frekvencijski sastav signala $x(t)$ u određenom vremenskom intervalu
- $x(t)$ - originalni signal
- a_0^k – skala (logaritamska podjela u skali)
- mT - pomak (translacija)

i gdje je $\psi_{m,k}(t) = \frac{1}{\sqrt{a_0^k}}\psi(\frac{t}{a_0^k} - mT)$ familija *wavelet* funkcija, pri čemu je $\psi(t)$ *mother wavelet*.

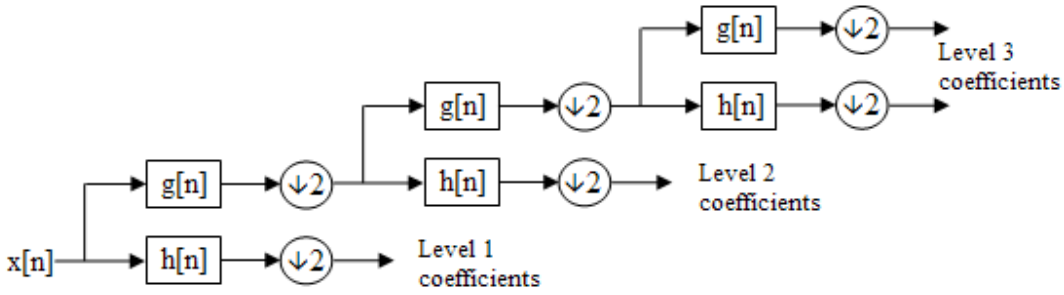
DWT signala se računa njegovim propuštanjem kroz niz filtera. Istovremeno se uzorci signala propuštaju kroz niskopropusni filter impulsnog odziva g i kroz visokopropusni filter impulsnog odziva h.

Izlazi iz niskopropusnog filtra se nazivaju aproksimacijama (eng. ca – *approximation coefficients*), a izlazi visokopropusnog filtra detaljima (eng. cd – *detail coefficients*). Ovaj postupak je prikazan na slici 8.



Slika 8.: Dobivanje koeficijenata aproksimacija i detalja

Ovakva dekompozicija se ponavlja da bi se povećala frekvencijska rezolucija. Taj postupak se prikazuje pomoću binarnog stabla koje se još naziva filtarskim slogom (slika 9.).



Slika 9.: Filtarski slog

Postoji puno familija *waveleta* i unutar svake familije postoje različite izvedbe tog *waveleta* i zbog toga je njihov odabir dugotrajan posao. Za EEG signale najboljom se pokazala familija *waveleta* pod imenom *Daubechies*[3] wavelet (u Matlabu skraćeno *db wavelet*). Skoro u svim srodnim radovima koji koriste DWT kao sredstvo ekstrakcije značajki koriste se upravo *Daubechijevi waveleti* [3] [8] [16].

Poznavanjem centralne frekvencije *waveleta* može se odrediti frekvencijsko područje i centralne frekvencije svih koeficjenata aproksimacije i detalja u svakoj razini dekompozicije. Pomoću tih poznatih podataka o waveletima lako se mogu odrediti koji koeficjenti odgovaraju kojim valnim oblicima EEG signala i upravo ti koeficjenti se koriste kao značajke.

ANOVA i odabir značajki

Dolazi do istog problema kao i kod računanja značajki PSD-om koji se rješava na isti način. Provedbom ANOVA-e dolazi se do slijedećih značajki:

1. DWT snaga α-pojasa u 11. kanalu
2. DWT snaga β-pojasa u 8. kanalu
3. DWT snaga δ-pojasa u 6. kanalu
4. DWT snaga θ-pojasa u 11. kanalu

[3] Dobili naziv po belgijskoj matematičarki Ingrid Daubechies koja je formulirala svoju familiju *waveleta* 1988. godine.

4.2.3. Higuchijeva fraktalna dimenzija

Izraz "fraktalna dimenzija"[7] odnosi se na necjelobrojnu ili fraktalnu dimenziju bilo kojeg objekta. FD analiza se često koristi u procesiranju biomedicinskih signala, kao što su EEG, HRV (eng. *heart rate variability*) i koračni intervali.

U fraktalnoj geometriji, fraktalna dimenzija D, je statistička veličina koja pokazuje koliko fraktalni objekt popunjava prostor i to na različitim skalama uvećanja. Postoji puno specifičnih definicija fraktalne dimenzije i nijedna od njih se ne tretira kao univerzalna.

Postoje različite metode i algoritmi za računanje fraktalne dimenzije, a za EEG signale koji se mogu promatrati kao vremenske serije koristi se *Higuchiev* algoritam izračunavanja fraktalne dimenzije. Fraktalna dimenzija, u ovom slučaju, predstavlja jednu moguću mjeru kompleksnosti signala.

Higuchijev algoritam za računanje fraktalne dimenzije D vremenske serije[7] :

HFD(s(j): vremenska serija duljine N, $k_{min}, k_{max} \in \aleph$*: doseg duljina intervala)*

for $k = k_{min}, \dots, k_{max}$

definiraj k *novih vremenskih serija,* $m = 1,2, \dots k$*:*

$$x_m^k = \left\{x(m), x(m+k), x(m+2k), \dots, x(m+\left\lfloor\frac{N-m}{k}\right\rfloor k)\right\}$$

izračunaj duljinu ovih krivulja:

$$L_m(k) = \frac{\sum_{i=1}^{\left\lfloor\frac{N-m}{k}\right\rfloor}|x(m+ik) - x(m+(i-1)k)|(n-1)}{\left\lfloor\frac{N-m}{k}\right\rfloor k}$$

definiraj $L(k) = mean\{L_m(k)|m = 1,2, \dots, k\}$

end

procjeni fraktalnu dimenziju kao gradijent pravca linearne regresije točaka $ln\big(L(k)\big)\ i\ ln(\frac{1}{k})\}$ *za* $k = k_{min}, \dots, k_{max}$.

ANOVA i odabir značajki

Kao značajka će se koristiti fraktalna dimenzija EEG signala izračunatog u trećem kanalu.

4.2.4. Skewness

Skewness je mjera asimetričnosti podataka oko njegove srednje vrijednosti. Ako je *skewness* negativna to znači da su uzorci signala više raspoređeni u lijevo u odnosu na srednju vrijednost, a ako je pozitivna onda su uzorci više raspoređeni u desnu stranu. *Skewness* za normalnu razdiobu (ili bilo koju drugo savršeno simetričnu razdiobu) je nula.

4.2.5. Kurtosis

Kurtosis je mjera koja pokazuje da li je signal izdužen (šiljast) ili pljosnat. *Kurtosis* za normalnu razdiobu iznosi tri. One razdiobe koje su izduženije imaju *kurtosis* veći od tri, dok one pljosnatije imaju *kurtosis* manji od tri.

ANOVA i odabir značajke

Rezultati ANOVA-e daju šesti kanal kao najbolji za računanje *kurtosisa*, a sedmi kanal za računaje *skewnessa*.

4.2.6. *Hjorthovi* parametri

Tri su *Hjortova* parametra koji opisuju signal x duljine N i to:

1. **Aktivnost**

Aktivnost je jednaka varijanci signala x:

$$ACTIVITY(x) = Var(x) = \sigma_x^2 = \frac{\sum_{i=1}^{N}(x(n) - \bar{x})^2}{N} \quad (4.)$$

gdje su:

- $\bar{x}$ - srednja vrijednost signala x
- σ_x - standardna devijacija signala x

2. **Mobilnost**

Mobilnost je mjera srednje frekvencije signala i definira se kao drugi korijen omjera aktivnosti prve derivacije signala i aktivnosti signala.

$$MOBILITY(x) = \sqrt{\frac{ACTIVITY(x')}{ACTIVITY(x)}} = \sqrt{\frac{Var(x')}{Var(x)}} = \frac{\sigma_x'}{\sigma_x} \quad (5.)$$

gdje je x' prva derivacija signala x.

3. **Kompleksnost**

Kompleksnost se definira kao omjer mobilnosti prve derivacije signala i mobilnost signala.

$$COMPLEXITY(x) = \frac{MOBILITY(x')}{MOBILITY(x)} = \frac{\frac{\sigma_x''}{\sigma_x'}}{\frac{\sigma_x'}{\sigma_x}} \quad (6.)$$

gdje je σ_x'' druga derivacija signala x.

Kompleksnost mjeri odstupanje signala od sinusnog oblika.

ANOVA i odabir značajki

Koristiti će se slijedeće značajke:

1. Parametar ACTIVITY izračunat za 7. kanal
2. Parametar MOBILITY izračunat za 6. kanal

3. Parametar COMPLEXITY izračunat za 6. kanal

Konačno, odabrano je 14 značajki koje će činiti vektor značajki pomoću kojeg će se trenirati klasifikator. Tih 14 značajki je:

1. PSD snaga α-pojasa u 11. kanalu
2. PSD snaga β-pojasa u 15. kanalu
3. PSD snaga δ-pojasa u 19. kanalu
4. PSD snaga θ-pojasa u 11. kanalu
5. DWT snaga α-pojasa u 11. kanalu
6. DWT snaga β-pojasa u 8. kanalu
7. DWT snaga δ-pojasa u 6. kanalu
8. DWT snaga θ-pojasa u 11. kanalu
9. fraktalna dimenzija 3. kanalu
10. *skewness* u 7. kanalu
11. *kurtosis* u 6. kanalu
12. Parametar ACTIVITY u 7. kanalu
13. Parametar MOBILITY u 6. kanalu
14. Parametar COMPLEXITY u 6. kanalu

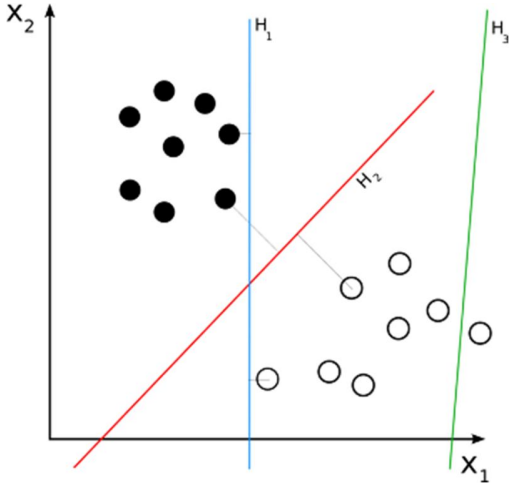
Navedene značajke daju najbolje rezultate kad se koriste zajedno. Gledajući ih pojedinačno ne uviđa se razlika između bolesnih i normalnih EEG-ova, ali zajedno SVM klasifikator uspjeva da nađe ravninu razdvajanja. Proučavanjem svake od 14 navedenih značajki pojedinačno primjetilo sa par zanimljivih stvari. Snage izračunate DWT-om u većini slučajeva, za sva četiri pojasa, poprimaju veće vrijednosti kad se računaju za normalne EEG-ove. Snage PTSP signala su niže. Isto to vrijedi i za PSD snage, a razlika je najizraženija za α-pojas (značajka broj 1.). Također je primjećeno da značajke pod rednim brojevima 12. i 14. poprimaju veće vrijednosti za normalne EEG-ove (ne u 100% slučajeva, ali u većini slučajeva), dok fraktalna dimenzija poprima malo veće vrijednosti za PTSP signale nego za normalne.

4.3 SVM

Klasifikacija podataka je čest zadatak strojnog učenja. Recimo da postoje dvije klase podataka i cilj je donijeti odluku kojoj klasi pripada novi podatak koji se dovede na ulaz. Pomoću SVM-a klasifikacija se vrši tako da se kreira *N*-dimenzionalna ravnina (eng. *hyperplane*) koja optimalno razdvaja podatke u dvije klase. SVM modeli su slični neuronskim mrežama koji po svojoj prirodi također spadaju u algoritme strojnog učenja. SVM model se najprije treba trenirati sa skupom podataka koji se naziva skup za treniranje. Zajedno sa skupom za treniranje se mora imati određeno *a priori* znanje kojoj klasi pripada svaki od uzoraka iz skupa za treniranje na osnovu čega se dobije trenirani model koji bi trebao novi uzorak (koji nije iz skupa za treniranje) svrstati u odgovarajuću klasu.

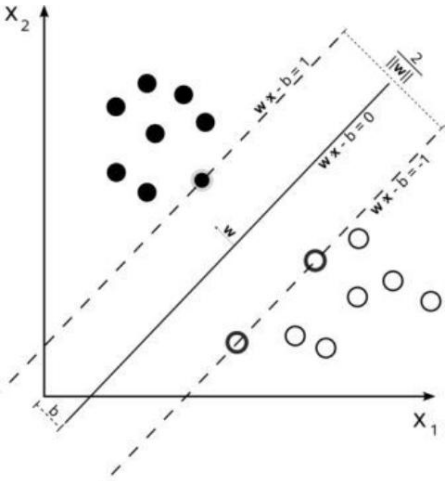
U slučaju SVM-a podatak se predstavlja i promatra kao *p*-dimenzionalan vektor (točka u prostoru) i želi se istrenirati model koji će dvije klase *p*-dimenzionalnih vektora moći razdvojiti *(p-1)*-dimenzionalnom ravninom. To se naziva linearnim klasifikatorom. Postoji nekoliko mogućih ravnina razdvajanja koje bi mogle klasificirati podatke, a jedna varijanta kojoj SVM teži je ona pomoću koje se postiže maksimalno razdvajanje odnosno maksimalna margina između klasa.

Dakle, SVM vrši klasifikaciju tako da konstruira *N*-dimenzionalnu ravninu koja optimalno razdvaja podatke u dvije kategorije (slika 10.)[8].



Slika 10.: Razdvajanje podataka u dvije klase

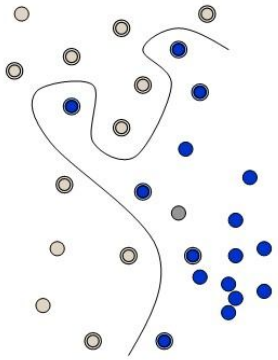
Postavlja se pitanje koja je razlika između tri pravca koji razdvajaju klase crnih i bijelih točaka. Vidljivo je da pravac H3 uopće ne razdvaja klase, tako da nam on nije interesantan, ali koja je razlika između pravaca H1 i H2 budući da oba odvajaju crne od bijelih točaka. Razlika je u tome da razdvajanjem klasa pravcem H2 se postižu maksimalne margine, odnosno udaljenosti od najbližeg člana pojedine klase do pravca H2. Stvar će možda biti jasnija uvođenjem još jedne slike (Slika 11.)



Slika 11.: Margine

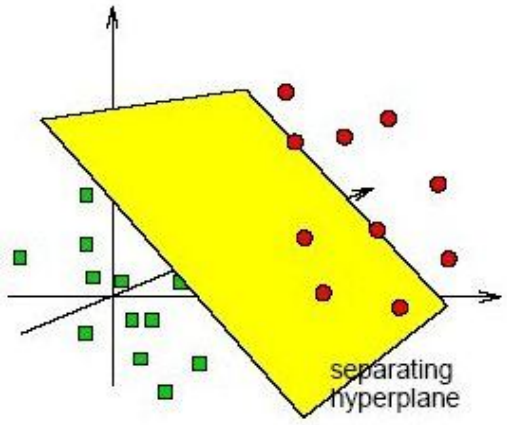
Na ovoj slici pomoćni pravci (vektori potpore – eng. *support vectors*) su nacrtani iscrtkanom linijom i oni prolaze kroz onog člana klase koji je najbliži središnjem pravcu (punom linijom) koji je zapravo klasifikator klasa.

Ovaj primjer je najjednostavniji koji se može pojaviti i najčešće se ne pojavljuje u stvarnim situacijama. Klasifikator na prethodne dvije slike je linearan, dok se mogu pojaviti i neki koji nisu linearni, npr. Slika 12.



Slika 12.: Nelinearan klasifikator

Klasifikator može biti i ravan ukoliko se klasifikacija vrši u tri dimenzije, npr. Slika 13.



Slika 13.: Ravan kao klasifikator

Najjednostavniji je linearni SVM pa ćemo njega detaljno opisati u nastavku da se uvede u svijet matematike koja se odigrava pri SVM učenju. Svi složeniji SVM-ovi od linearnog se zasnivaju na istom principu kao što će biti objašnjeno poslije.

Uzmimo skup za treniranje D s n točaka oblika

$$D = \{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \{-1,1\}\}_{i=1}^{n}$$

gdje

- y_i poprima vrijednost -1 ili 1 pokazujući time klasu kojoj pripada točka x_i, a
- svaki x_i je p-dimenzionlan realan vektor.

Cilj je pronaći maksimalnu marginu koja odvaja točke za koje vrijedi $y_i = 1$ ili $y_i = -1$. Bilo koja ravnina razdvajanja se može zapisati kao skup točaka x koje zadovoljavaju jednadžbu

$$\mathrm{w} \cdot \mathrm{x} - b = 0 \qquad (7.)$$

gdje

- $\cdot$ označava skalarni produkt vektora, a
- w je normalan vektor na ravninu razdvajanja.

Parametar $\frac{b}{\|w\|}$ predstavlja udaljenost ravnine razdvajanja od ishodišta.

Žele se odabrati parametri w i b takvi da se maksimizira margina, odnosno, udaljenost između paralelnih potpornih vektora, a da se pri tom odvajaju podaci.

Potporni vektori se mogu opisati jednadžbama:

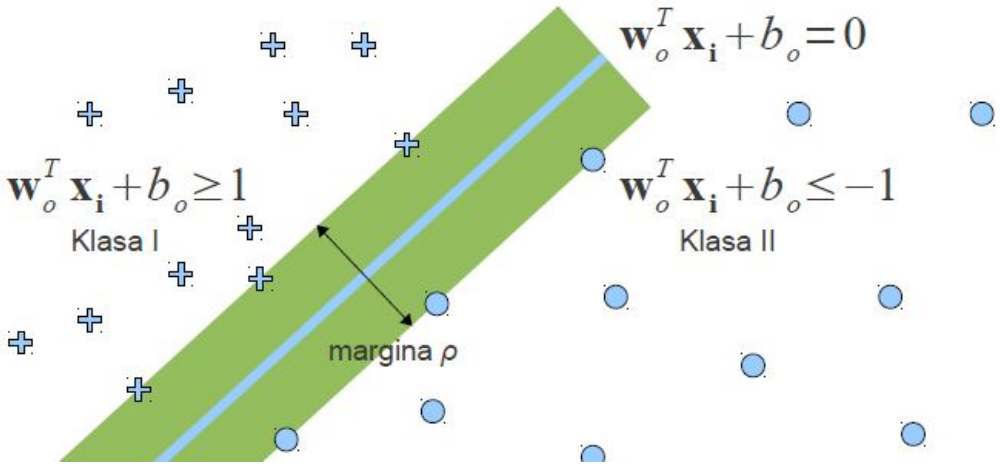
$$w \cdot x - b = 1 \quad (8.)$$

$$w \cdot x - b = -1 \quad (9.)$$

Ako su klase linearno separabilne onda se mogu odabrati dva potporna vektora tako da nema točaka između njih i onda maksimizirati udaljenost između njih. Ako se koriste jednadžbe 8. i 9. za opis vektora potpore dobije se da je udaljenost između njih, odnosno, da je širina margine $\frac{2}{\|w\|}$. Da bi maksimizirali širinu margine potrebno je minimizirati $\|w\|$. Također, da bi spriječili pojavu podataka (točaka) unutar margine dodaje se novi uvjet za svaki x_i (slika 14.)[10]:

$$w \cdot x_i - b \geq 1 \text{ za } x_i \text{ iz prve klase} \quad (10.)$$

$$w \cdot x_i - b \leq 1 \text{ za } x_i \text{ iz druge klase} \quad (11.)$$



Slika 14.: Klase podataka

Jednadžbe 10. i 11. se mogu zapisati kao:

$$y_i(w \cdot x_i - b) \geq 1 \quad (12.)$$

Kao što je spomenuto, maksimizacija margine ovisi o minimiziranju $\|\mathrm{w}\|$. To stvara problem jer je to operacija modula koja u sebi sadrži korjenovanje koje je računski zahtjevno. Zbog toga se primjenjuju razni postupci optimizacije.

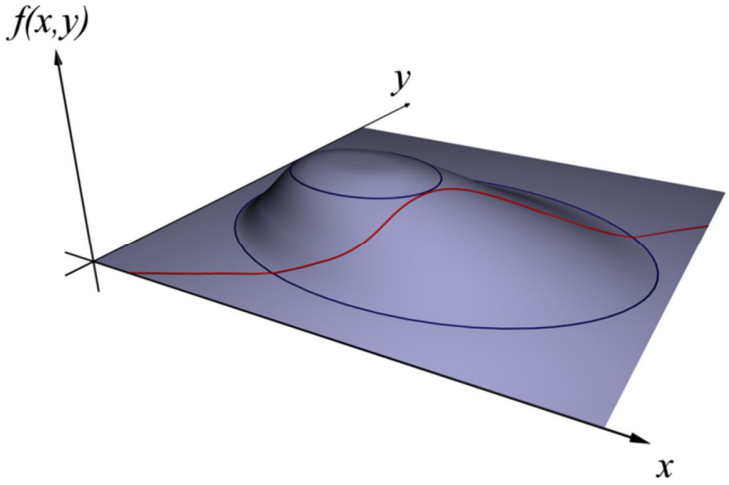
Problem korjenovanja se rješava substitucijom:

$$\|\mathrm{w}\| \rightarrow \frac{1}{2}\|\mathrm{w}\|^2$$

pri čemu se konačni rezultat neće promijeniti.

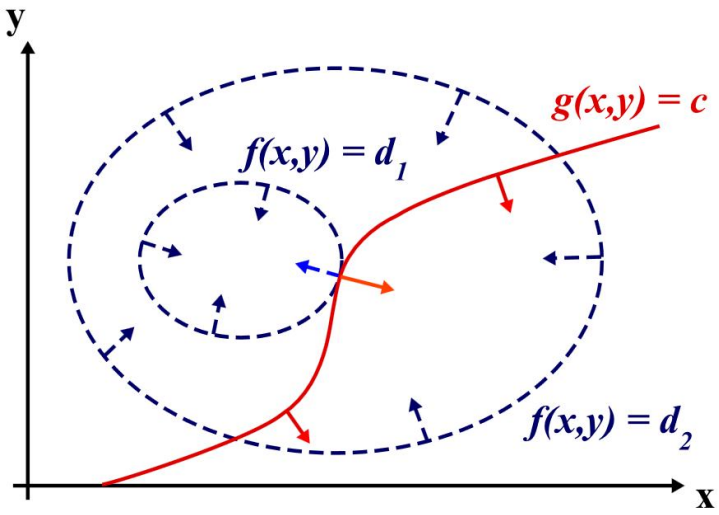
Da bi našli minimum od $\frac{1}{2}\|\mathrm{w}\|^2$ uvode se *Lagrangeovi* multiplikatori α.

U području matematičke optimizacije, metoda *Lagrangeovih* multiplikatora[9] pruža strategiju za pronalazak lokalnih maksimuma i minimuma neke funkcije. Tako npr. ako se traži maksimum funkcije $f(x, y)$ koja ima neko ograničenje (uvjet) koji može biti algebarska jednadžba kao npr. $g(x, y) = c$ kao na slici 15.



Slika 15.: Funkcija *f(x,y)* i uvjet *g(x,y)=c*

Na slici 16. je vidljivo da se konture funkcija f i g dodiruju kad su tangentni vektori konturnih linija paralelni.



Slika 16.: Konture funkcija *f* i *g*

Budući da je gradijent funkcije okomit na konture te funkcije može se reći da su gradijenti funkcija f i g paralelni. Stoga, traže se točke (x, y) gdje je $g(x, y) = c$ i

$$\nabla_{x,y} f = -\alpha \nabla_{x,y} g$$

gdje su:

- $\nabla_{x,y} f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right)$ - gradijent funkcije $f(x, y)$
- $\nabla_{x,y} g = \left(\frac{\partial g}{\partial x}, \frac{\partial g}{\partial y}\right)$ - gradijent funkcije $g(x, y)$

Konstanta α je potrebna jer iako su dva gradijenta paralelna njihove magnitude generalno nisu jednake.

Kad se svi ovi uvjeti ukomponiraju u jednu jednadžbu dobije se tzv. pomoćna funkcija:

$$\Lambda(x, y, \alpha) = f(x, y) + \alpha(g(x, y) - c) \qquad (13.)$$

Metoda *Lagrangeovih* multiplikatora se onda svodi na rješavanje jednadžbi:

$$\nabla_{x,y,\alpha}\Lambda(x,y,\alpha) = 0 \qquad (14.)$$

tj. rješavanje parcijalnih derivacija izjednačenih sa nulom.

Analogno prethodnom izvodu želi se dobiti pomoćna funkcija za konkretnu primjenu rješavanja optimizacijskog problema SVM-a. Uzimajući u obzir slijedeće:

- funckija koja se želi minimizirati je $\frac{1}{2}\|w\|^2$
- uvjet na tu funkciju je $y_i(w \cdot x_i - b) \geq 1$

slijedi:

$$\Lambda(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(w \cdot x_i - b) - 1] \qquad (15.)$$

Zatim se parcijalne derivacije izjednačavaju s nulom:

$\frac{\partial \Lambda}{\partial w} = 0$	$\frac{\partial \Lambda}{\partial b} = 0$	$\frac{\partial \Lambda}{\partial \alpha}$
$w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0$	$\sum_{i=1}^{n} \alpha_i y_i = 0$	$\sum_{i=1}^{n} [y_i(w \cdot x_i - b) - 1] = 0$
$w = \sum_{i=1}^{n} \alpha_i y_i x_i$		$\sum_{i=1}^{n} [y_i(w \cdot x_i - b)] = 1$

Samo nekoliko α_i će biti različiti od nule i upravo ti *Lagrangeovi* multiplikatori koji su različiti od nule automatski odabiru potporne vektore. Odgovarajući x_i su vektori potpore koji leže na margini i zadovoljavaju jednadžbu $[y_i(w \cdot x_i - b)] = 1$.

Za određivanje *Lagrangeovih* multiplikatora pravilo klasifikacije se zapisuje u dualnom obliku i otkriva da maksimum margine ovisi o vektorima potpore. Dobiveni izraz $w = \sum_{i=1}^{n} \alpha_i y_i x_i$ se uvrštava u jednadžbu 15. i dobije se tzv. dualni problem:

$$\mathcal{L}(\alpha) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i x_j - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i x_j$$
$$+\sum_{i=1}^{n}\alpha_i y_i b + \sum_{i=1}^{n}\alpha_i \quad = \left|\sum_{i=1}^{n}\alpha_i y_i\right| = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i x_j$$
$$= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j K(x_i, x_j) \qquad (16.)$$

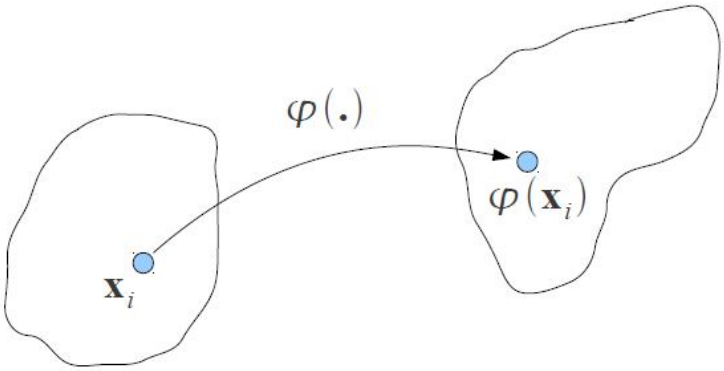
gdje se sa $K(x_i, x_j)$ označava *kernel* funkcija o kojoj će više biti rečeno u nastavku.

U slučaju da ne postoji ravnina razdvajanja koja može razdvojiti uzorke u dvije grupe bez pogreške uvodi se metoda mekih margina (eng. *soft margin*) koja odabire onu ravninu razdvajanja uz koju će razdvajanje klasa biti što bolje. Kad se radi o mekim marginama princip pronalaska *Legendreovih* multiplikatora je u suštini isti s tim da se dodaju neki novi uvjeti u prethodne izraze,a to je npr.: $y_i(\mathrm{w}\cdot x_i - b) \geq 1 - \varepsilon_i$ gdje je ε_i varijabla koja mjeri koliki je stupanj pogrešno klasificiranih podataka. Zatim, pomoćna funkcija poprima slijedeći oblik:

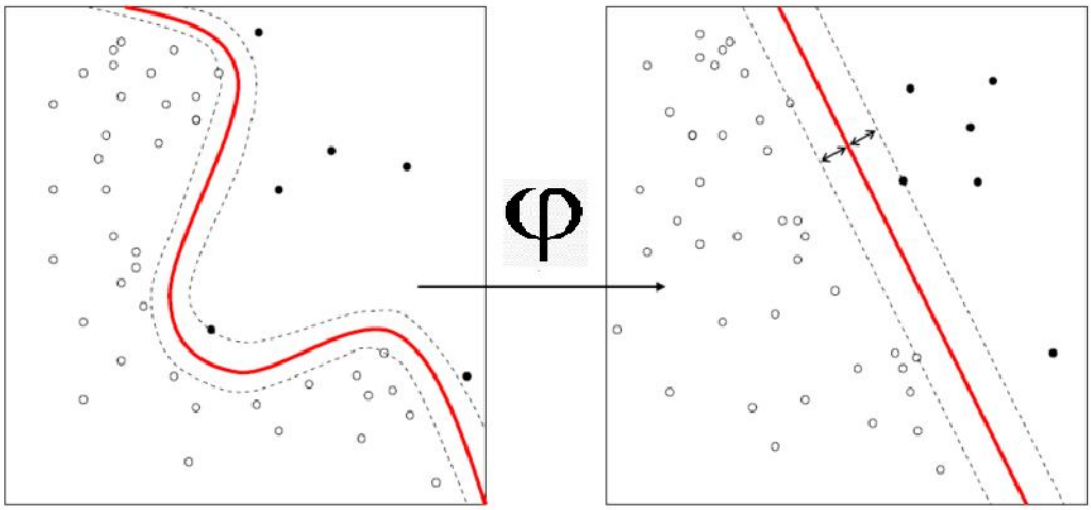
$$\Lambda = \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{i=1}^{n}\varepsilon_i - \sum_{i=1}^{n}\alpha_i[y_i(\mathrm{w}\cdot x_i - b) - 1 + \varepsilon_i] - \sum_{i=1}^{n}\beta_i\varepsilon_i \qquad (17.)$$

Daljnje rješavanje je analogno prethodnom.

Izvorni problem klasifikacije se može definirati u nekom konačnom dimenzionalanom prostoru, ali često se događa da klase koje se žele klasificirati nisu linearno separabilne u tom prostoru. Upravo zbog toga je predloženo da se originalni ulazni prostor mapira u prostor više dimenzionalnosti. Ta ideja potiče od *Coverovog* teorema koji kaže da prelaskom u višedimenzionalni prostor, raste vjerojatnost linearne separabilnosti. Osnovna ideja se sastoji iz dva koraka. Najprije se vrši nelinearno mapiranje ulaznog prostora u novi prostor značajki više dimenzionalnosti (slika 17.) i zatim se konstruira optimalna ravnina razdvajanja u tom novom (višedimenzionalnom) prostoru značajki (slika 18.).



Slika 17.: Mapiranje u novi prostor značajki



Slika 18.: Konstrukcija optimalne ravnine razdvajanja u novom prostoru značajki

Nelinearni klasifikatori se javljaju onda kad klase nisu linearno separabilne i nije moguće između njih povući lnearni klasifikator. Kreiranje nelinearnih klasifikatora se vrši korištenjem *kernel* funkcija u svrhu maksimiziranja margine razdvajanja. Algoritam pronalaska klasifikatora je isti kao i kod linearnih osim što se u ovom slučaju svaki skalarni produkt zamjeni s nelinearnom *kernel* funkcijom.

Dakle, jednadžba kojom se opisuje ravnina razdvajanja poprima novi izgled i to:

$$\sum_{j=1}^{m} w_j \varphi_j(x) + b = 0 \qquad (18.)$$

Iz ove jednadžbe se vidi da je jedina razlika u usporedbi sa prethodnom jednadžbom koja je opisivala linearni klasifikator ta da se umjesto x koristi $\varphi(\mathrm{x})$ što zapravo predstavlja vektor x u novom prostoru značajki.

Cijeli postupak određivanja klasifikatora i *Legendreovih* multiplikatora je identičan onom za linearni klasifikator uz substituciju $\mathrm{x} \rightarrow \varphi(\mathrm{x})$.

Prethodno, u jednadžbi 16. je spomenuta *kernel* funkcija $K(x_i, x_j)$ kao skalarni produkt vektora x_i i x_j. Budući da se *kernel* funkcije koriste samo u slučaju kad klasifikator nije linearan i prelazi se u novi prostor značajki onda je *kernel* funkcija direktno povezana sa transformacijom φ pri čemu je

$$K(x_i, x_j) = \varphi(x_i)\varphi(x_j) \tag{19.}$$

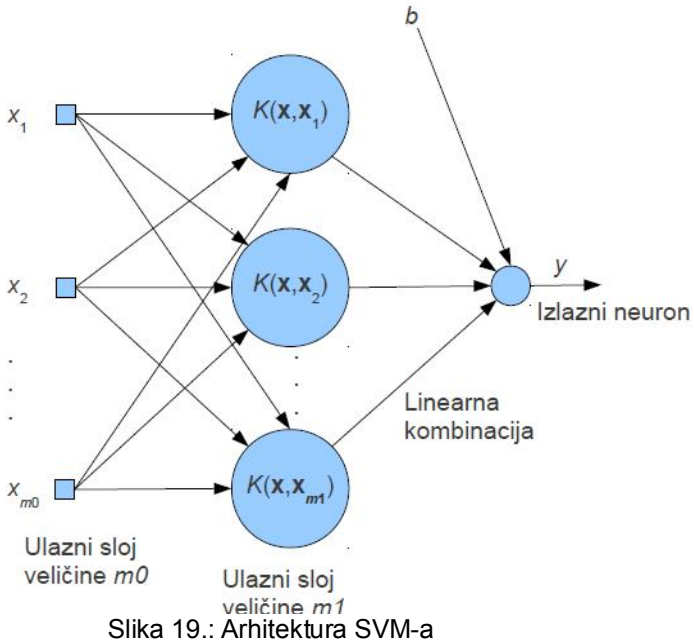
i tako dobivena funkcija K se naziva jezgrom unutarnjeg produkta. Postoji određena sloboda u izboru funkcije jezgre, ali je potrebno da zadovoljava *Mercerov* teorem koji kaže da je $K(x_i, x_j)$ simetrična funkcija jezgre definirana na zatvorenim intervalima od x_i i x_j ako se može rastaviti na slijedeći niz:

$$K(x_i, x_j) = \sum_{i=1}^{\infty} \alpha_i \varphi_i(x_i)\varphi_i(x_j) \tag{20.}$$

Korištenjem *kernelovih* funkcija se omogućava da algoritam pronađe ravninu razdvajanja sa maksimalnom marginom u transformiranom prostoru značajki. Neki od najkorištenijih *kernela* su:

- Polinomni homogeni *kernel* – $K(x_i, x_j) = (x_i \cdot x_j)^d$
- Polinomni nehomogeni *kernel* – $K(x_i, x_j) = (x_i \cdot x_j + 1)^d$
- Gaussov radijalni *kernel* - $K(x_i, x_j) = e^{\left(-\frac{1}{2\sigma^2}\|x_i - x_j\|^2\right)}$
- Hiperbolični *kernel* – $K(x_i, x_j) = \tanh(\varkappa x_i \cdot x_j + c)$

Konačno, arhitektura SVM-a se može prikazati dijagramom na slici 19.



Slika 19.: Arhitektura SVM-a

4.3.1. Implementacija

U Matlabu postoje implementirane funkcije za treniranje i testiranje SVM-a kao dio *Bioinformatics Toolbox-a*. Za treniranje se koristi funkcija *svmtrain()* kojoj se kao parametri zadaju signali za treniranje. Pored signala za treniranje, ukoliko se ne žele koristiti *defaultne* vrijednosti, moguće je zadati i dodatne parametre. Dodatni parametar koji se koristio u ovom radu je *'kernel_function'* pomoću kojeg se mijenja *kernel* funkcija koja se koristi u treniranju klasifikatora. Za treniranje klasifikatora najboljom se pokazala *kernel* funkcija *quadratic*[4] koja se opisuje izrazom:

$$K(x_i, x_j) = (1 + x_i \cdot x_j)^2 \qquad (21.)$$

Za testiranje dobivenog SVM klasifikatora koristi se funkcija *svmclassify()* kojoj se zadaju dva parametra od kojih je jedan trenirani klasifikator dobiven funkcijom *svmtrain()* i signal koji se želi klasificirati.

[4] Pored *quadratic* testirani su još linearni, polinomni i Gausov radijalni *kernel*.

5.Rezultati

Tip I pogreška (pogreška prve vrste) i pogreška tip II (pogreška druge vrste)[11] su precizni tehnički pojmovi koji se koriste u statistici za opisivanje određenih nedostataka u procesu testiranja gdje je točna nulta hipoteza pogrešno odbijena (tip I) ili kad je pogrešna nulta hipoteza prihvaćena (tip II). Ovi testovi se primjenjuju u testiranju klasifikatora koji se realizirao i koristio u procesu klasifikacije EEG signala.

U teoriji statističkog testiranja najprije je potrebno postaviti nultu hipotezu koja obično odgovara nekom uobičajenom stanju u prirodi kao npr. „osoba je zdrava“, „proizvod je čitav“, „optuženik je nevin“. Alternativna hipoteza je negacija nulte hipoteze kao npr. „osoba je bolesna“, „proizvod je slomljen“, „optuženik je kriv“. Ovdje se testiraju rezultati klasifikacije osoba oboljelih od PTSP-a na osnovu EEG signala tako da će nulta hipoteza za ovaj slučaj biti upravo $H_0 = "pacijent\ ima\ PTSP"$. Rezultat testa može biti negativan u odnosu na nultu hipotezu ili pozitivan. Ako rezultat testa odgovara stvarnosti onda je donesena ispravna odluka, a ako ne odgovara stvarnosti onda se desila pogreška. Kao što je već rečeno, razlikuju se dva tipa pogrešaka i to tip I i tip II.

Pogreška prve vrste se javlja kad je nulta hipoteza istinita, ali je odbijena. Pogreška prvog tipa se još naziva lažno pozitivna (eng. FP – *False Positive*) i odlika dobrog klasifikatora bi bila da je FP što manji.

Pogreška druge vrste se javlja kad je nulta hipoteza pogrešna, ali se prihvaća kao istinita. Pogreška drugog tipa se još naziva lažno negativna (eng. FN – *False Negative*) i za dobar klasifikator će vrijediti da je FN, ujedno kao i FP, što niži, tj. da što manje bolesnih osoba klasificira kao zdrave i da što manje zdravih osoba klasificira kao bolesne. Naravno, idealno bi bilo da su FN i FP jednaki nuli, tj. da nema pogrešnih klasifikacija, ali to se rijetko dešava u praksi.

Pored FN i FP još se klasificiraju dvije vrste pogrešaka koje zapravo označavaju točan rezultat i to:

- TN (eng. *True Negative*) – kad se zdrava osoba klasificira kao zdrava
- TP (eng. *True Positive*) – kad se bolesna osoba klasificira kao bolesna

Dobar klasifikator će imati visoke vrijednosti za TN i TP, tj. veliki postotak točno klasificiranih uzoraka.

Sve ove pogreške čine konfuzijsku matricu koja se koristi za razlikovanje pogrešaka. Standardni način računanja pogreške preko učestalosti pogreške (broj_pogrešaka/broj primjera) je vrlo grub jer npr. U medicini, konkretno u ovom slučaju kad se radi o dijagnosticiranju PTSP-a puno većom i ozbiljnijom pogrškom se smaatra ako se neke bolesne osobe klasificiraju kao zdrave. Zbog toga se u medicini koriste pogreške FP i FN zajedno sa mjerama osjetljivosti (eng. sensitivity) i specifičnosti(eng. specificity) koje su prikazane u tablici 4.

Visoka osjetljivost u dijagnostici bolesti znači visok postotak ispravnog klasificiranja pacijenata koji imaju bolest. Dok visoka specifičnost znači visok postotak klasificiranja pacijenata koji nemju bolest. U stvarnosti je jako teško, skoro nemoguće, napraviti klasifikator koji ima visoku i specifičnost i osjetljivost, ali se vrše istraživanja i testiranja koja poboljšavaju rezultate i polako nas približuju nekom idealnom klasifikatoru.

U tablici 4.[12] je prikazan primjer konfuzijske matrice sa dodatnim mjerama točnosti koje će se računati u ovom radu:

Tablica 4.: Pogreške

		PTSP	Zdrav	
Rezultat klsifikatora	PTSP	Točan rezultat True Positive - TP	Tip I False Positive – FP	PPV[5] $= \frac{TP}{TP+FP}$
	Zdrav	Tip II False Negative – FN	Točan rezultat True Negative - TN	NPV[6] $= \frac{TN}{FN+TN}$
		$Osjetljivost = \frac{TP}{TP+FN}$	$Specifičnost = \frac{TN}{FP+TN}$	

[5] PPV (eng. *Positive Predictive Value*) – Prediktivna vrijednost pozitivnih
[6] NPV (eng. *Negative Predictive Value*) – Prediktivna vrijednost negativnih

Bitan postupak u procesu analize podataka je priprema podataka za analizu. Kad se radi klasifikacija podataka potrebno je imati skupinu podataka za učenje, tj.za treniranje klasifikatora za koje se mora a-priori znati u koju skupinu spada(npr.koji su bolesna, a koji su zdravi) i skupinu podatka za testiranje klasifikatora kojima je također potrebno znati u koju skupinu spada. Pored ova dva skupa podataka postoji i treća, a to su novi podaci s nepoznatom klasifikacijom koje bi naš klasifikator nakon treniranja trebao moći klasificirati u odgovarajuću skupinu.

Dalje se postavlja pitanje kako podijeliti skup podatak koji imamo na raspolaganju u skupinu za učenje i skupinu za treniranje. Odgovor na to pitanje ovisi o veličini populacije, tj.o broju podataka s kojim raspolažemo. U ovom slučaju na raspolaganju imamo 57 EEG signala i to je mala populacija uzoraka i za tako malu skupinu podataka najboljim postupkom treniranja klasifikatora se smatra *Bootstrap* postupak (postupak samopodizanja) koji se sastoji iz toga da se skup za učenje generira tako da se slučajno odaberu podaci za učenje, a oni neizvučeni će služiti za testiranje. Očekivani postotak signala u skupu za testiranje je 36,8% što u ovom konkretnom slučaju znači 21 EEG signal. Ostalih 36 EEG signala će činiti skup za treniranje klsifikatora. *Bootstrap* postupak se sastoji iz slučajnog odabira 36 signala za treniranje nakon čega se provede testiranje. Nakon testiranja svi signali se vraćaju u populaciju odakle se ponavlja postupak slučajnog odabira signala za učenje. Na ovaj način neki signali će biti izvučeni nekoliko puta, a neki uopće neće biti izvučeni. Za dobru procjenu potrebno je taj postupak ponoviti oko 100 puta i izračunati srednju pogrešku.

Bootstrap postupak je primijenjen na klasifikator u ovom radu i postupak je ponovljen 100 puta. U nastavku su u tablici 5. prikazani dobiveni rezultati.

Tablica 5.: Rezultati *Bootstrap* krosvalidacije

Mjera	Iznos
FP	14,38%
FN	5,38%
TN	85,62%
TP	94,62%

$Specifičnost$	85,62%
$Osjetljivost$	94,62%
$False\ Positive\ Rate = 1 - Specifičnost$	14,38%
$False\ Negative\ Rate = 1 - Osjetljivost$	5,38%
PPV	91,45%
NPV	90,73%
$False\ Discovery\ Rate = \frac{FP}{FP+TP}$	8,55%
$Likelihood\ Ratio\ Positive = \frac{Osjetljivost}{1-Specifičnost}$	6,6
$Likelihood\ Ratio\ Negative = \frac{1-Osjetljivost}{Specifičnost}$	0,063
$Pogreška\ (Error\ rate)$	8,81%
$Točnost = \frac{TP+TN}{TP+TN+FN+FP}$	91,19%

Iz priloženog se vidi da su dobivene zadovoljavajuće karakteristike klasifikatora kao što su niske vrijednosti za FP i FN, a visoke vrijednosti za TP, TN, *Specifičnost* i *Osjetljivost* te je čak i sveukupna točnost klasifikacije jednaka 91,19% što je vrlo zadovoljavajući rezultat.

Iz svih provedenih testiranja se naučilo da se kombiniranjem različitih značajki mogu dobiti bolji rezultati klasifikacije nego koristeći jednu značajku ili značajke koje su po svojoj prirodi slične. Isto tako se naučilo da preveliko miješanje značajki ne mora nužno dati bolje rezultati i opet se dolazi do zaključka da je odabir značajki najbitniji i ujedno najteži dio klasifikacije signala.

6. **Zaključak**

U ovom radu je opisana razvijena metoda za klasifikaciju EEG signala. Cilj je bio primjeniti novostečena znanja iz područja dubinske analize podataka, te izvršiti testiranje odabranog klasifikatora. Pored testiranja izračunate su različite mjere koje pokazuju kvalitetu klasifikatora, njegovu osjetljivost i specifičnost za konkretnu primjenu dijagnosticiranja PTSP-a.

Koristeći kombinaciju različitih značajki i SVM klasifikatora *Bootstrap* postupkom se ustanovila ukupna točnost klasifikacije od 91,19% s tim da je osjetljivost klasifikatora 94,62% što smatramo dobrim rezultatom.

Bez prvođenja ikakvog postupka krosvalidacije se dobivaju značajno drugačiji rezultati što upućuje na važnost primjene odgovarajućih metoda krosvalidacije (u ovom radu *Bootstrap*) i primjene tehnika za dubinsku analizu podataka.

7. Literatura

[1] „*Electroencephalography*“. http://en.wikipedia.org/wiki/Electroencephalography. Svibanj 2012.

[2] „*ANOVA*“. *http://www.mathworks.com/help/toolbox/stats/bqttcvf.html* . Svibanj 2012.

[3] Bašić, B. D. „*Analiza varijance: ANOVA, Analysis of variance, Analysis of means using variance*“. Materijal za kolegij “Statističko učenje”. Fakultet elektrotehnike i računarstva, Zagreb, 19.4.2005.

[4] “*Spectral density*”. http://en.wikipedia.org/wiki/Power_spectral_density . Svibanj 2012.

[5] „*Discrete wavelet transform*“, travanj 2012. http://en.wikipedia.org/wiki/Discrete_wavelet_transform . Lipanj 2012.

[6] Seršić, D. Materijal za kolegij “Napredne metode digitalne obrade signala”. Fakultet elektrotehnike i računarstva.

[7] „*Fraktali III dio - fraktalna geometrija i dimenzije*“. 2008. http://nauka.adsoglasi.com/fizikamatematika/fraktali3geometrija.php . Lipanj 2012.

[8] *“Support vector machine*”. http://en.wikipedia.org/wiki/Support_vector_machine . Lipanj 2012.

[9] *“Lagrange multiplier*”. http://en.wikipedia.org/wiki/Lagrange_multiplier . Lipanj 2012.

[10] Lončarić, S., Subašić, M. “*Neuronske mreže: Stroj s potpornim vektorima (SVM)”.*Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, materijali za kolegij “Neuronske mreže”. Lipanj 2012.

[11] “Type I and type II errors”, svibanj 2012. http://en.wikipedia.org/wiki/Type_I_and_type_II_errors . Lipanj 2012.

[12] “Sensitivity and specificity” http://en.wikipedia.org/wiki/Sensitivity_and_specificity