SVEUČILIŠTE U ZAGREBU

**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

DIPLOMSKI RAD br. 2806

**IZVEDBA METODE STROJNOG UČENJA ZA RAZLIKOVANJE IZVORA AKUSTIČKIH EMISIJA U BILJNIM TKIVIMA NA UGRADBENOM SUSTAVU NISKE POTROŠNJE**

Renato Gracin

SVEUČILIŠTE U ZAGREBU

**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

DIPLOMSKI RAD br. 2806

**IZVEDBA METODE STROJNOG UČENJA ZA RAZLIKOVANJE IZVORA AKUSTIČKIH EMISIJA U BILJNIM TKIVIMA NA UGRADBENOM SUSTAVU NISKE POTROŠNJE**

Renato Gracin

**SVEUČILIŠTE U ZAGREBU**

**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

**Zagreb, 11. ožujka 2022.**

DIPLOMSKI ZADATAK br. 2806

Pristupnik: **Renato Gracin (0036508945)**

Studij: Računarstvo

Profil: Računalno inženjerstvo

Mentor: doc. dr. sc. Dinko Oletić

Zadatak: **Izvedba metode strojnog učenja za razlikovanje izvora akustičkih emisija u biljnim tkivima na ugradbenom sustavu niske potrošnje**

Opis zadatka:

Istražiti metode nenadziranog i polu-nadziranog strojnog učenja, primjenjive za razlikovanje vrste izvora ultrazvučnih akustičkih emisija u biljnim tkivima za prijenos vode. Odabrati metodu odabira značajki prikladnu za primjenu na zadanom skupu podataka i predložiti skup značajki signala. Odabrati metodu klasifikacije koja omogućava izvedbu na ugradbenom sustavu, vodeći računa o programskoj složenosti i memorijskom zauzeću. Izvesti odabrani algoritam ekstrakcije značajki i klasifikacije na ugradbenom procesoru porodice Arm Cortex-M STM32L4. Odabrati metodologiju i izvesti ispitno okruženje. Provjeriti točnost algoritma na snimljenom skupu podataka i trajanje izvođenja na odabranom ugradbenom sustavu. Procijeniti doprinos algoritma ukupnoj potrošnji ugradbenogsustava.

Rok za predaju rada: 27. lipnja 2022.

Sadržaj

[Uvod 1](#_Toc104971774)

[1. Pregled teme i prijašnjih doprinosa detekcije kavitacije 2](#_Toc104971775)

[1.1. Opis kavitacije kod biljaka 2](#_Toc104971776)

[1.2. Trenutna metodologija detekcije kavitacije 2](#_Toc104971777)

[2. Ekstrakcija značajki (Predobrada) akustičnih emisija 3](#_Toc104971778)

[2.1. Izbor i definiranje značajki akustičnih emisija 3](#_Toc104971779)

[2.2. Postupak izračuna matrice značajki akustičnih emisija 3](#_Toc104971780)

[2.3. Ekvilizacija sirovoj emisija 3](#_Toc104971781)

[2.4. Analiza i normalizacija skupa značajki 3](#_Toc104971782)

[3. Nenadzirano i polu-nadzirano strojno učenje 4](#_Toc104971783)

[3.1. Pregled metodologije strojnog učenja 4](#_Toc104971784)

[3.2. OPTICS algoritam i implementacija 4](#_Toc104971785)

[3.3. Analiza rezultata strojnog učenja 4](#_Toc104971786)

[4. Selekcija značajki 5](#_Toc104971787)

[4.1. Pregled metodologije selekcije značajki 5](#_Toc104971788)

[4.2. Analiza značajki pomoću PCA 5](#_Toc104971789)

[4.3. Implementacija selekcije značajki 5](#_Toc104971790)

[4.4. Indeksi validacije rezultata grupiranja (clustering-a) 5](#_Toc104971791)

[4.5. Analiza rezultata selekcije značajki 5](#_Toc104971792)

[5. Implementacija na ugradbenom računalnom sustavu 6](#_Toc104971793)

[5.1. Pregled ugradbenog računalnog sustava 6](#_Toc104971794)

[5.2. Nadogradnja algoritma klasifikacije akustičnih emisija 6](#_Toc104971795)

[5.3. Testiranje rada algoritma 6](#_Toc104971796)

[Zaključak 7](#_Toc104971797)

[Literatura 8](#_Toc104971798)

[Sažetak 9](#_Toc104971799)

[Summary 10](#_Toc104971800)

[Skraćenice 11](#_Toc104971801)

[Privitak 12](#_Toc104971802)

# Uvod

Preko 70% ukupne vode u svijetu koristi se u svrhu navodnjavanja agrikulture. Potreba za vodom raste porastom učestalosti i intenziteta suša u svijetu. Pronalaskom načina uštede vode u procesu navodnjavanja drastično bi utjecao na problem nestašice vode. Navedenom problematikom bavi se područje pametnog i preciznog navodnjavanja. Pametno navodnjavanje temelji se na proučavanju indikatora vodenog stresa biljke. Radi preciznije analize mijenja se pristup detekcije vodenog stresa od mjerenja deficita vode u tlu zbog velike ovisnosti o vrsti tla. Glavni indikatori vodenog stresa biljke je pojava kavitacija unutar ksilema biljke. Time fokus pametnog navodnjavanja prelazi na metode koje proučavaju hidrauličke, vizualne te auditivne indikatora kavitacije unutar biljke. Za detekciju kavitacije na terenu u stvarnom vremenu jedino su primjenjive akustične metode za koje nisu potrebni laboratorijski uvjeti. Akustične metode temelje se na analizi ultrazvučnih akustičnih emisija ili UAE koje biljka emitira u stanju vodenog stresa. Cilj akustičnih metoda je pronaći UAE uzrokovane kavitacijom. Unaprjeđivanje istraživanja UAE pokazuju na mogućnost prepoznavanje UAE nastala kavitacijom, međutim potrebna je dodatna analiza. Glavni temelji se rad je

Problem pametnog i preciznog navodnjavanja sve više raste na važnosti zbog povećanja nestašice vode. Istraživanja klimatskih promjena potvrđuju

kako Biljke emitiraju akustične emisije tijekom svih stadija svojega života koje daju informacijo o trenutnim fiziološkom stanju biljke.

U ovom radu analizirati će se akustične emisije biljaka emitirane prilikom dehidratacije u svrsi razvoja sustava pametnog navodnjavanja.

Glavni uzrok akustičnih emisija prilikom dehidratacije biljaka je proces embolizacije i kavitacije čijim djelovanjem nastaju akustične emisije. Prepoznavanjem akustičnih emisija uzrokovani kavitacijom omogućuje detekciju kada se biljka nalazi u procesu dehidratacije.

**Opiši ostale moguće uzroke u uvodu ili kasnije?**

Ovaj rad bavi se razvojem postupka klasifikacije uzroka akustičnih emisija kod biljaka prilikom dehidratacije pomoću metodologije strojnog učenja. Rad se nadovezuje na članak [2] u kojem se razvio senzorski čvor za akviziciju i jednostavnu klasifikaciju akustičnih emisija biljaka na terenu.

**Opis što će se odraditi u pojedinom poglavlju?**

**Mentor:**

1. **Motivacija – koji problem rješava i koji su izazovi za koje dajem rješenja**
2. **Jasno definiranje ciljeva – 2-3 rečenice**
3. **Moji doprinosi/pristup – 2-3 rečenice**
4. **Struktura paragrafa – što koje poglavlje opisuje**

# Pregled teme i prijašnjih doprinosa detekcije kavitacije

## Opis kavitacije kod biljaka

### Motivacija

Kod pametnog navodnjavanja biljaka potrebno je prepoznati kada je biljka u stanju vodenog stresa. Jedan od postupaka je određivanje vodnog potencijala tla iz kojeg biljka upija vodu. Zbog nepouzdanost navedeni način je nepouzdan jer izrazito ovisi o okolinskim faktorima. Radi uštede vode i optimalnog uzgoja biljaka pametno navodnjavanje se bazira mjerenjem internih proces biljke.

### Prirodni rad biljke

Efikasna transpiracija vode kroz ksilem biljke glavni je faktor prirodnog rada i razvoja biljke. Proces transpiracije uključuje transport vode kroz biljku te evaporaciju vode na površinskim organima biljke kao što su listovi i stabljika. Kroz transport vode biljka prenosi hranjive tvari do svih svojih organa. Transpiracijsko-adhezijsko-kohezijsko-tenzijska (TAKT) teorija objašnjava proces transporta vode kroz ksilem. Transport vode kroz biljku postiže se negativnim hidrostatskim tlakom koji povlači stupac vode ksilema prema gore. Hidrostatski tlak ili tenzija uzrokovana je ubrzanom evaporacijom vode iz lista biljke u svrsi hlađenja prilikom zagrijavanja biljke i njezine okoline. Zbog navedenih tenzija voda se nalazi u fizikalno nestabilnom stanju te kada se hidrostatski tlak vode izjednači s tlakom pare voda prelazi u plinovito agregatno stanje. Unatoč tome kohezivna i adhezivna svojstva vode i ksilema održavaju vodu u tekućem stanju kao neprekinuti stupac vode.

### Procesa i uzrok kavitacije kod biljaka

Tijekom izrazito velikih tenzija uzrokovanih evaporacijom vode moguće je formiranje i širenje mjehurića plinova unutar ksilema. Navedeni proces naziva se kavitacija ili embolija. Kavitacija je reverzibilna te anatomija ksilema je napravljena za sprječavanje širenja kavitacije tijekom tipičnih uvjeta rasta biljke. Tijekom razdoblja suše brzina evaporacije stvara mjehuriće koji potpuno sprječavaju protok vode kroz ksilem i uzrokuju ugibanje biljke. Kavitacija je prvi znak stanja vodenog stresa biljke. Prepoznavanje nastanka kavitacije služi za pravovremeno i efikasno navodnjavanje biljaka. [0]

## Trenutna metodologija detekcije kavitacije

### Pregled postojećih metoda detekcije i mjerenja kavitacije te usporedba s metodom analize akustičnih emisija

Glavne metode detektiranja kavitacije mogu se podijeliti na hidrauličke metode, vizualne metode te akustične metode. Hidrauličke metode temelje se na mjerenju postotka gubitka hidrauličke vodljivosti biljke ili PLC. Metoda se sastoji se od dva mjerenja hidrauličke vodljivosti provedenih na odrezanom uzorku biljke. Prvo mjerenje hidrauličke vodljivosti provodi se na uzorku u prirodnom stanju. Prilikom mjerenja potrebno je paziti da se ne uvedu dodatni mjehurići zraka u uzorak. Drugo mjerenje mjeri maksimalnu hidrauličku vodljivosti uzorka nakon uklanjaju mjehurići iz uzorka. PLC se računa kao omjer dvaju mjerenja. Zbog destruktivne prirode mjerenja te nedostatka automatiziranosti hidrauličke metode nisu primjenjiva za terensku detekciju kavitacija u stvarnom vremenu. Nadalje vizualne metode temelje se na vizualnoj identifikaciji mjehurića zraka unutar ksilema biljke. Kavitacija može se ne destruktivno identificirati pomoću nuklearne magnetske rezonanca (NMR) i rendgenske kompjuterizirane tomografije (µCT). Kvalitetniju analizu daje µCT zbog bolje prostorne razlučivosti, iako pri mjerenju postoji mogućnost oštećenje stanica biljke. Vizualne metode su najpouzdanije za detekciju kavitacije, međutim nisu primjenjive kao terenski i rasprostranjeni alat detekcije kavitacije. Zbog svoje ne destruktivne i pouzdane prirode koriste se kao referentno mjerenje kavitacije prilikom korištenja akustičnih metoda. [1.1]

Završno akustične metode temelje se na mjerenjima akustičnih emisija (AE) unutar ksilema biljke. Početno AE bile su smatrane uzrokom puknuća kavitacijskih mjehurića zraka unutar ksilema te time kao optimalno sredstvo prepoznavanja kavitacije. [] Daljnjim istraživanjima pokazano je kako emitiranje AE je aktivno tijekom prirodnog stanja biljke te u stanju nakon totalnog gubitka hidrauličke vodljivosti biljke. Oba navedena stanja biljke isključuju kavitaciju kao jedini uzrok AE. Time su identificirani dodatni uzroci AE: sužavanje stanične stijenke, formiranje pukotina, kapilarno djelovanje vlakna biljke (Rosner 2012.) i skupljanje tkiva kore. Navedeni uzroci AE proizvode mehaničke valove propagirane kroz tkivo biljke koje je moguće mjeriti piezo električnim senzorima postavljenim na površinu stabljike.

Za detekciju kavitacije na terenu u stvarnom vremenu jedino su primjenjive akustične metode za koje nisu potrebni laboratorijski uvjeti.

Ne destruktivnosti i prilagodljivost akustične metode predstavlja potencijalnu terensku metodu za ranu detekciju kavitacije kod biljaka.

Zaključci istraživanja:

Danjim istraživanjima identificiran je da je frekvencijski pojas pojavljivanja AE između 100 kHz i 1000 kHz.

Pokazana je korelacija između nastanka AE emisija i smanjenja hidrauličke vodljivosti biljke te vodenog potencijala biljke (Mapiranje AE na VC). Isto tako određene AE su uspoređene s direktnom vizualizacijom embolizma metodom µ-CT.

[0]

Indirektno mjerenje hidrauličke vodljivosti postiže se mjerenjem vodnog potencijala ksilema pomoću krivulja ranjivosti (VC) definiranih na temelju sorte ili vrste biljaka nad kojom se vrše mjerenja.

Alati za mjerenja vodnog potencijala:

1. Scholander tlačna komora
2. Psihorometri (mjere vlažnost zraka) za tlak pare stabljike – ne invazivno i ovisno o temperaturi okoline

Navedene metode nisu primjenjive za proizvodnju velikih razmjera na terenu.

In-situ -> mjerenja nad organizmom u prirodnoj okolini i stanju, u divljini

### Pregled napretka i eksperimenta u području klasifikacije akustičnih emisija prilikom dehidratacije biljke. Uključuje i kombinaciju ostalih metoda uz analizu AE.

[1]

### Detaljan opis rada [2] na koji se nadovezuje ovaj diplomski tad te metodologije kojom će se nadograditi. Kakvi senzori trebaju?

Korišteni skupovi podataka dobiveni su kontinuiranom akvizicijom UAE u eksperimentima sušenja vinove loze. Radi ubrzanja i kontrole procesa sušenja u većini eksperimenta korištena je klima komora, dok u eksperimentu 6 sušenje je odrađeno prirodnim putem. Proces akvizicije UAE sastojao se od akvizicije UAE piezo električnim senzorom Vallen VS600-Z1, obrade analognog signala Vallen pojačalom AEP3N, digitalizacije signala analogno digitalnim pretvornikom National Instruments USB-DAQ 6366 ADC te pohrana rezultata pomoću Matlab skripte. ADC je digitalizirao UAE s frekvencijom uzorkovanja 2 MHz te 16-bitnom rezolucijom. Odrađena je analiza dobivenih UAE na temelju ekstrakcije značajki vremena pojave pojedine UAE te frekvencije rezonantnih vrhova pojedine UAE. Unutar svih skupova podataka identificirano je slično grupiranje frekvencije rezonantnih vrhova pojedinih UAE u određenim frekvencijskim pojasevima. Izdvojene su 3 karakteristične grupe UAE pomoću značajke broja rezonantnih vrhova pojedinih UAE. UAE bez rezonantnih vrhova ili široko pojasne UAE (engl. *broadband*) čine u prosjeku 25% ukupnih UAE tijekom. Navedene UAE grupiraju se u frekvencijskom pojasu između 100 i 200 kHz. U prosjeku oko 15% UAE spada u grupu UAE s jednim rezonantnim vrhom (engl. *single peak*). Navedene UAE grupiraju su frekvencijskom pojasu oko 200 kHz. U prosjeku oko 60% UAE spada u grupu UAE s više rezonantnih vrhova (engl. *multiple peak*). Za navedene UAE vrh s najvećom magnitudom tipično se grupira oko 200 kHz, dok ostali vrhovi između 200 i 600 kHz. [4]

Ovaj rad želio bi raditi dodatnu analizu navedenih saznanja pomoću metoda strojnog učenja te metoda selekcije značajki u svrsi pronalaska novih značajki i grupacija na temelju kojih je moguće klasificirati UAE po izvoru nastajanja.

U tablici 1 prikazana su vremena trajanja, korišteni kultivar vinove loze te opis procesa sušenja za svaki eksperiment korištenih skupova podataka.

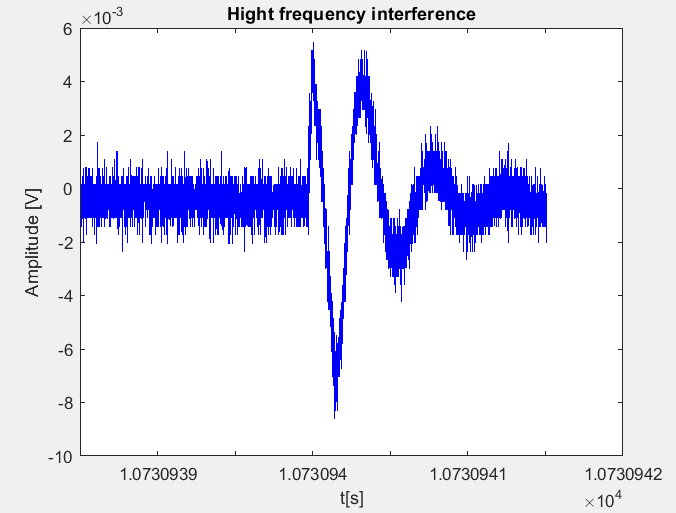
# Ekstrakcija značajki UAE

Ekstrakcija značajki (*engl. feature extraction*) je proces obrade sirovih podataka i pretvorbe u numeričke značajke uz očuvanje informacija o izvornom skupu podataka. Numeričke značajke se dovode na ulaz algoritma strojnog učenja koji klasificira skup podataka na temelju značajki. Zbog toga kvaliteta klasifikacije podataka iznimno ovisi o kvaliteti ekstrakcije značajki. Kvalitetna ekstrakcija značajki podrazumijeva izbor značajki relevantnih za domenu i problem koji se namjerava riješiti. [15] Skupovi podataka nad kojima se radi ekstrakcija značajki dobiveni su eksperimentima opisanim u [4]. Izabrano je 7 reprezentativnih skupova podataka prikazani u tablici 1. Ekstrakcija značajki odrađena je jednako na svim skupovima podataka te podrazumijeva nadogradnju na postojeću predobradu iz [4]. Glavni koraci ekstrakcije značajki su izdvajanje validnih emisija iz skupa podataka, ekvilizacija emisija te izračun vrijednost značajki za svaku emisiju. Navedeni koraci prikazani su na slici 1. U slijedećim poglavljima opisuje se implementacija ekstrakcije značajki u Matlabu koja će kasnije biti preslikana na ugradbeni računalni sustav.

## Izdvajanje validnih emisija iz skupova podataka

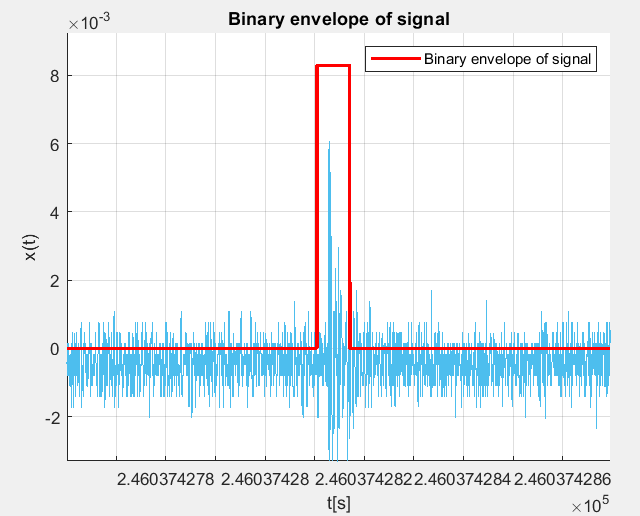
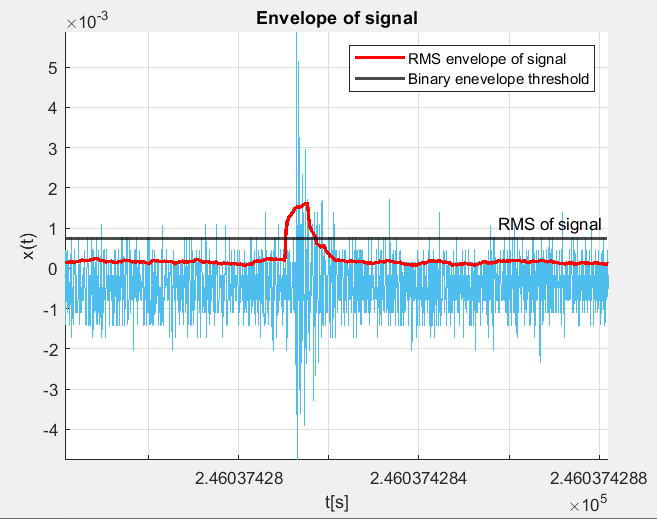
Izdvajanje potencijalnih UAE iz ukupnog skupa podataka na ugradbenom računalnom sustavu odrađuje akvizicijski sustav s postavljenim pragom od 2 mV za početak pohrane pojedine UAE. Međutim navedeni skupovi podataka dobiveni su kontinuiranom pohranom UAE te emisije su izdvojene po intervalima od 3 ms. Zbog toga prvi korak predobrade je odvojiti validne emisije od šuma za svaki vremenski prozoru od 3 ms.

Validacija emisija započinje izbacivanjem nisko frekvencijskih smetnji iz daljnje analize. Nisko frekvencijske smetnje karakterizirane su s maksimalnom magnitudom amplitudnog spektru ispod 100 kHz. U vremenskoj domeni manifestiraju se kao izobličenja spektra prikazana na slici 1.



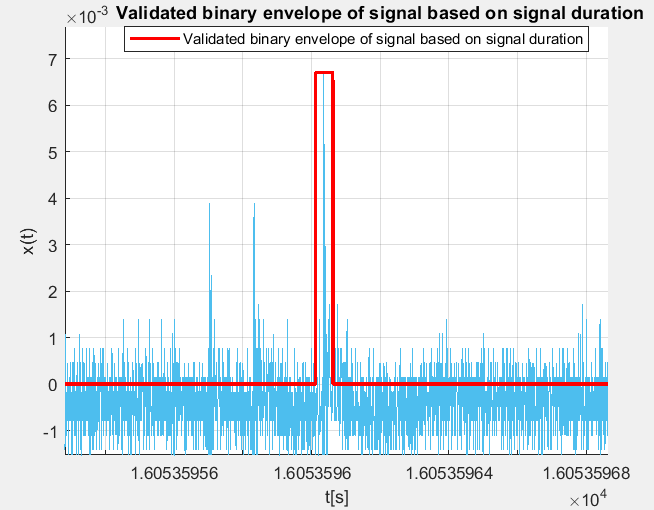
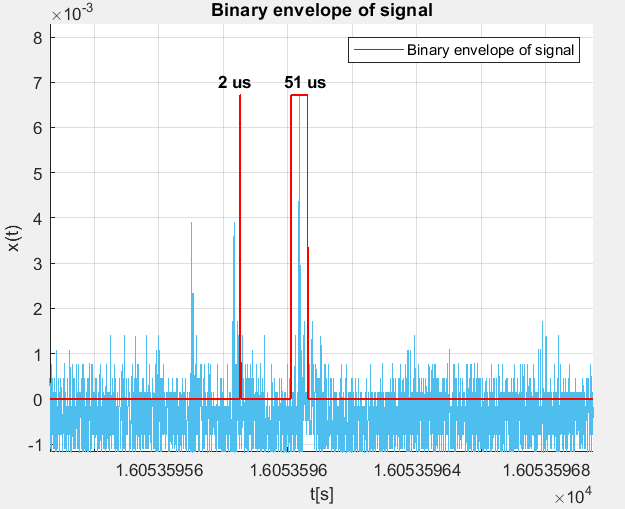
Slika 1 Emisija uzrokovana nisko frekvencijskom smetnjom

Kako bih se odvojila potencijalna emisija od šuma računa se RMS omotnica signala s prozorom od 100 točaka (50 us). Izračun RMS omotnice daje najveći uvid u amplitudu signala jer daje mjeru snage signala, dok stvara valni oblik koji je lako analizirati. [11] Nad dobivenim valnim oblikom, crvena krivulja na slici 2 lijevo, izdvajamo emisiju na temelju amplitudnog praga određen RMS-om čitave emisije, crni pravac na slici 2 lijevo. Time dobivamo binarnu omotnicu emisije prikazanu na slici 2 desno koja ima vrijednosti jedan u amplitudama potencijalne emisija te vrijednosti nula u amplitudama šuma.



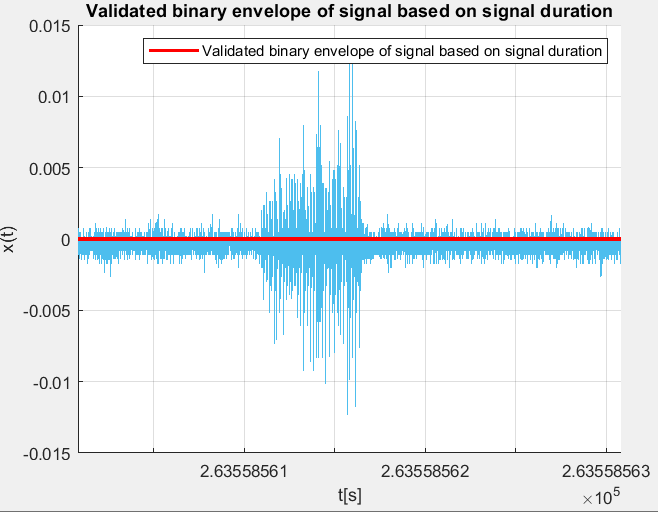
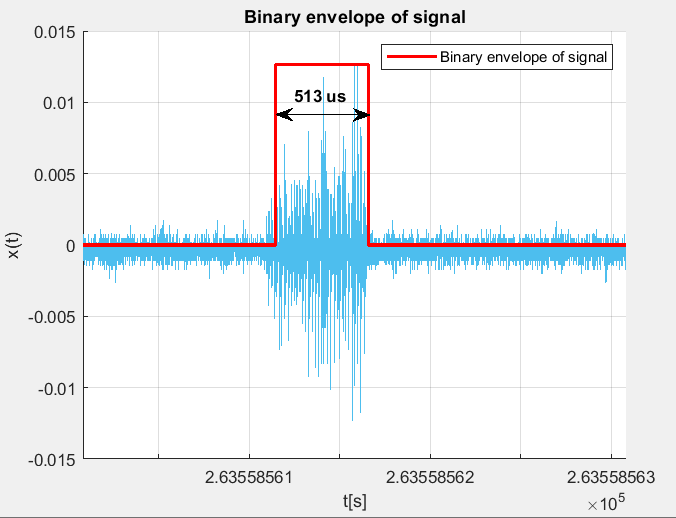
Slika 2 Prikaz izračuna binarne i RMS omotnice signala

Slijedeće radi se vremenska provjera ispravnosti emisija izdvojenih binarnom omotnicom. Emisije su identificirane kao prekratke ako imaju trajanje ispod 15 µs. Na slici 3 lijevo vidimo primjer izdvajanja prekratke emisije s trajanjem o 2 µs. Emisije su identificirane kao preduge ako imaju trajanje iznad 500 µs. Na slici 3 lijevo vidimo primjer izdvajanja preduge emisije s trajanjem od 513 µs. Preduge i prekratke emisije uzrokovane su smetnjama prilikom akvizicije. Navedene emisije se zanemaruju te formiraju se nove binarne omotnice koje uključuje samo validne emisije prikazane desno na slikama 3 i 4.

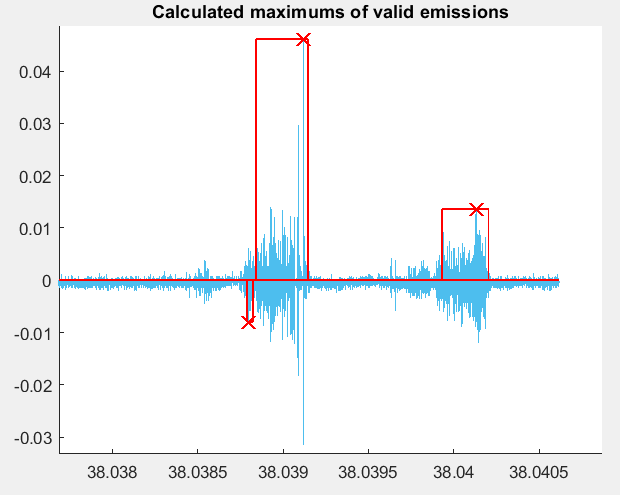


Slika 3 Prikaz izdvajanja prekratke UE

Slika 4 Prikaz izdvajanja preduge UAE

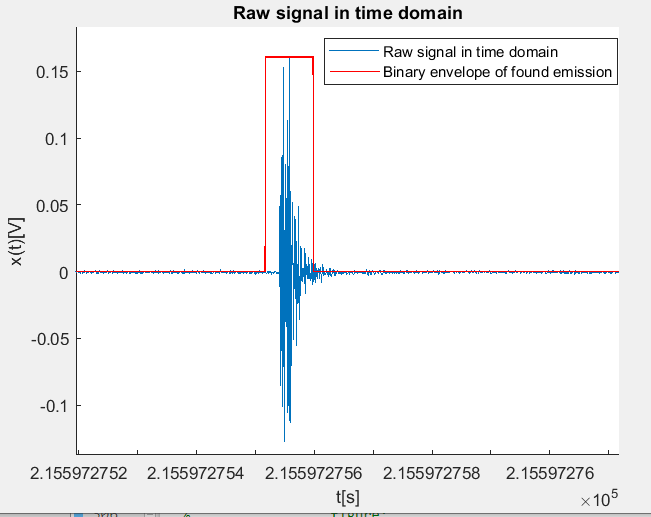
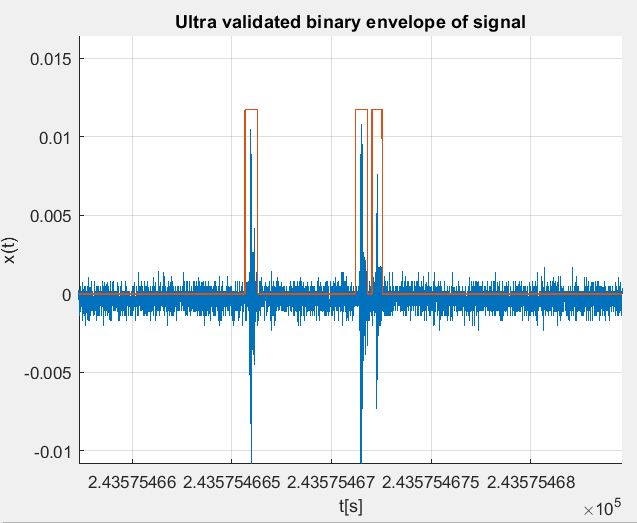


Amplitudno izbacivanje emisije radi se na temelju akvizicijskog praga od 2 mV. Gdje se emisije s maksimalnom amplitudom ispod 2 mV (lijeva emisija na slici 5) zanemaruju u daljnjoj analizi. Na slici 5 desno vidimo binarnu omotnicu nakon izbacivanja emisije na temelju maksimalne amplitude.



Slika 5 Izbacivanje emisije s maksimalnom amplitudom ispod 2 mV

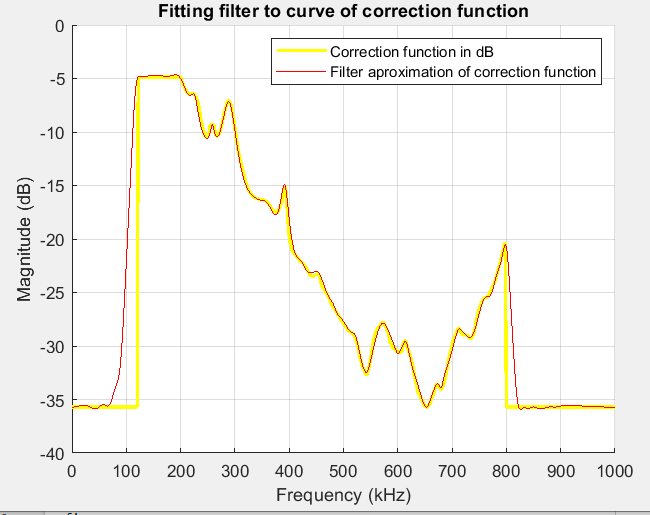
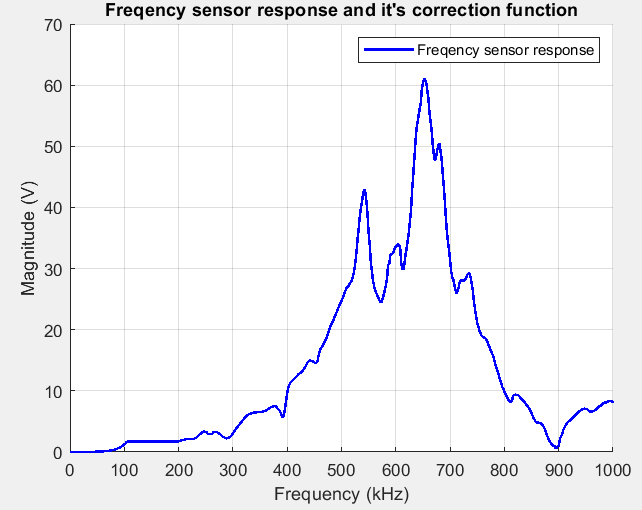
Završni emisije se izdvajaju pomoću validirane binarne omotnice te se proširuju nulama oko maksimalne amplitude emisije na fiksnu veličinu od 1024 točke ili 512 µs. Broj točaka je izabran kao potencija broja 2 kako bi računanje frekvencijskog spektra emisije pomoću brze fourierove transformacije (FFT) bilo efikasnije. [16] Isto tako kako bi uključio emisije svih trajanja ispod maksimalnog dozvoljenog od 500 µs. Prikaz tipično izdvojenih validnih emisija vidi se na slici 5 lijevo, dok tipična pojedina emisija produžena nulama vidi se na slici 5 desno.



Slika 6 Tipične validne emisije u vremenskoj domeni (lijevo) te pojedina izdvojena emisija proširena nulama (desno)

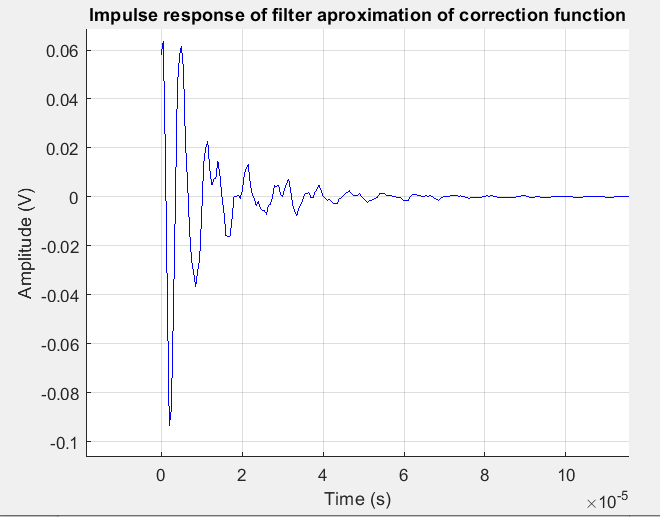
## Ekvilizacija sirovoj emisija

Jedan od najvažnijih koraka unutar navedenog postupka ekstrakcije značajki UAE je ekvilizacija sirovih ultrazvučnih emisija. Ekvilizacija predstavlja metodu kojom ispravljamo nelinearne promjene koji unosi senzor i obrada analognog signala prilikom akvizicije UAE. Za senzor korišteni pri akviziciji podataka definiran je amplitudni spektar frekvencijskog odziva senzora (slika 6 lijevo). Ekvilizacija amplitudnog spektra UAE postiže se množenjem s korekcijskom funkcijom definiranom recipročnim magnitudama frekvencijskog odziva senzora. Kako bi ekvilizirali UAE u vremenskoj domeni potrebno je formirati fazni spektar frekvencijskog odziva senzora u Matlabu. Pomoću alata za procesiranje signala u Matlabu [17] kreiranjem filtra određenog amplitudnog spektra u decibelima automatski se kreira prikladni fazni spektar. Na slici 6 desno prikazan je amplitudni spektar filtra u decibelima (crvena krivulja) koji aproksimira magnitude korekcijske funkcije (žuta krivulja).



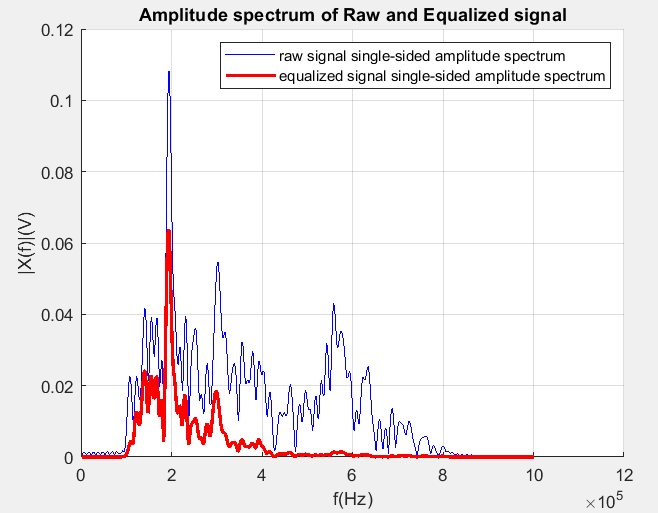
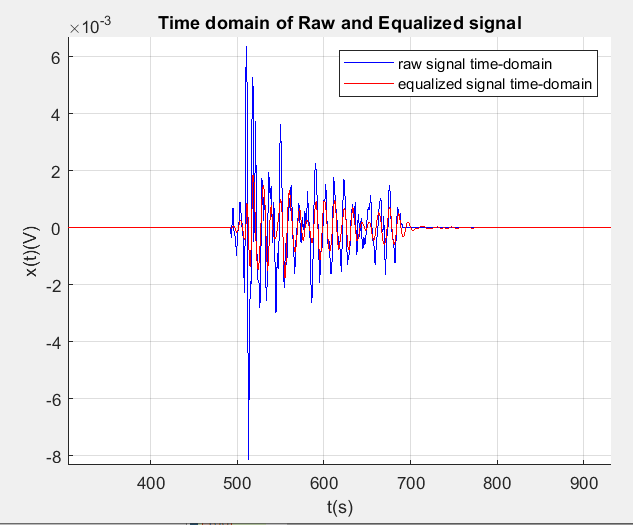
Slika 7 Amplitudni spektar frekvencijskog odziva senzora (desno) i njegova korekcijska funkcija (lijevo žuto) aproksimirana filtrom u Matlabu (lijevo crveno))

Iz navedenog filtra računa se impulsni odziv koji se koristi pri ekvilizaciji. Po teoremu konvolucije postupak konvolucije izvornog signala u vremenskoj domeni s impulsnim odzivom filtra ekvivalentan je množenju izvornog signala u frekvencijskoj domeni s magnitudama filtra u frekvencijskoj domeni. [18] Zbog toga postupak ekvilizacije obuhvaća konvoluciju amplituda UAE u vremenskoj domeni s impulsnim odzivom filtra koji aproksimira korekcijsku funkciju u decibelima prikazan na slici 7.



Slika 8 Impulsni odziv filtra koji aproksimira korekcijsku funkciju frekvencijskog odziva senzora

Ekvilizacija primjenjuje se na svakoj dobivenoj izdvojenoj validnoj emisiji. Odnos tipične sirove UAE (plava krivulja) i ekvilizirane UAE (crvena krivulja) prikazan je u vremenskoj domeni (lijevo) i amplitudnom spektru (desno) na slici 8.



Slika 9 Sirova (plavo) i ekvilizirana (crveno) emisija UAE u vremenskom (lijevo) i amplitudnom spektru (desno)

Prikaz karakteristične emisije:

## Izbor i definiranje značajki akustičnih emisija

Izbor reprezentativnih značajki UAE je iznimno bitan kako se čitavi postupak klasifikacije uzroka UAE temelji na grupiranju vrijednosti značajki. Zbog nedostatka informacija o značajkama prikladnim za distinkciju izvora UAE implementirani je izračun većine značajki korištenih u prijašnjim eksperimentima. Primarni skup značajki preuzet je iz literature [9] koja za računanje značajki koristi program AEwin [10]. Dodatno preuzeta je značajka WPF koja je bila uspješna kod razlikovanja uzroka UAE. [5] Završno izabrana je značajka broja rezonantnih vrhova u amplitudnom spektru UAE koja potencijalno izdvaja UAE izvore opisana u poglavlju 1.4.[2] Popis izabranih značajki uz prikladne mjerne jedinice i kratice prikazan je na slici 10.

Slika na kojoj se prikazuje stol

Opis je automatski generiran

Slika 10 Izabrane značajke uz prikladne mjerne jedinice i kratic

Iz vremenske domene izabrane su značajke definirane na temelju amplituda, energije, te trajanja karakterističnih dijelova UAE:

* **Maksimalna amplituda** (*engl. peak amplitude*)
  + Maksimalna apsolutna amplituda UAE izražena u voltima. (slika 10 crvena isprekidana linija)
* **Vrijeme uzdizanja** (*engl. rise time*)
  + Vrijeme od početka validne UAE *(engl. start time*) do pojave maksimalne amplitude UAE u vremenskoj domeni izraženo u sekundama. (slika 10)
* **Vrijeme pada** *(engl. fall time*)
  + Vrijeme od pojave maksimalne amplitude UAE do vremena kraja validne UAE *(engl. end time*) u vremenskoj domeni izraženo u sekundama. (slika 10)
* **Trajanje** (*engl. duration*)
  + Vrijeme trajanja validne UAE izraženo u sekundama. (slika 10)
* **Ukupni broj vrhova** (*engl. total counts*)
  + Ukupni broj amplituda UAE koje prelazi predodređeni amplitudni prag.
* **Broj vrhova prije maksimalne amplitude** (*engl. counts to peak*)
  + Broj vrhova validne UAE u vremenu prije maksimalne amplitude UAE uključujući maksimalnu amplitudu. (slika 10 crveni križići)
* **Broj vrhova poslije maksimalne amplitude** (*engl counts from peak*)
  + Broj vrhova validne UAE u vremenu poslije maksimalne amplitude UAE. (slika 10 crni križići)
* **Snaga signala** (*engl. signal strength*)
  + Integral apsolutnih amplituda UAE. Računa se po formuli (1) te izražena je u voltima po sekundi. (slika 10 plava površina)

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

* **Apsolutna energija** (*engl. absolute energy*)
  + Prava energija UAE izražena u džulima. Računa se po formuli (2) gdje 10 kΩ predstavlja referentni otpor tijekom trajanja UAE.

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

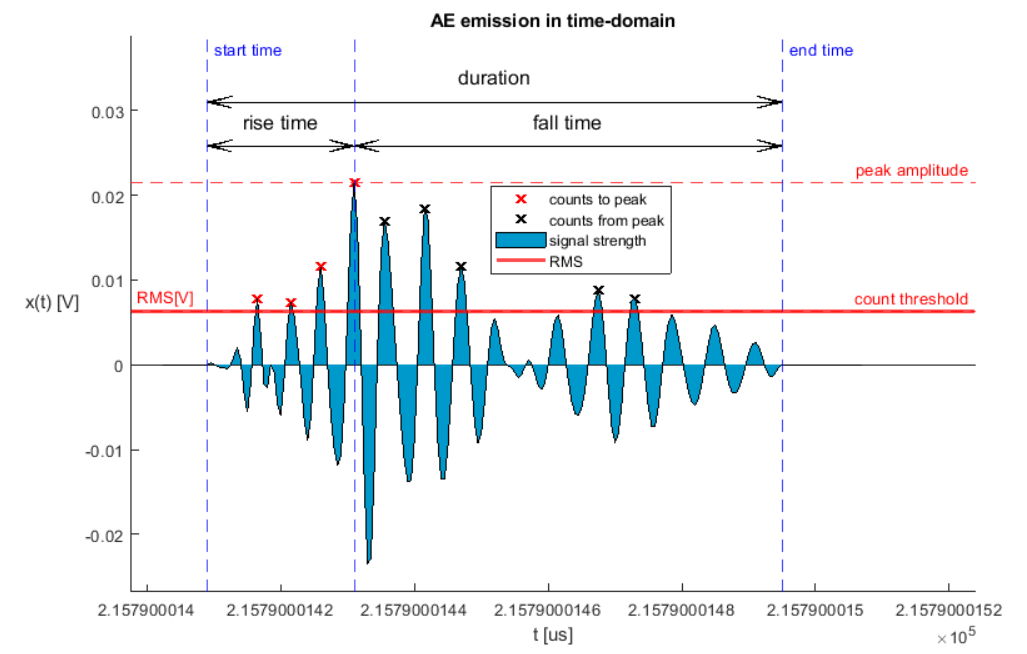
* **RMS**
  + Korijen iz srednje vrijednosti kvadrata amplituda validne UAE. Računa se po formuli (3) te izražena je u voltima. (slika 10 crveni pravac)

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

* **Srednja razina signala** (*engl. average signal level*) ili **ASL**
  + Korijen iz srednje vrijednosti amplituda validne UAE. Računa se po formuli (4) te izražena je u decibelima.

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

Prikaz vremenskih značajka na tipičnoj UAE može vidjeti na slici 10.



Slika 11 Prikaz vremenskih značajki na tipičnoj UAE

Računanje integrala na diskretnim vrijednostima UAE aproksimira se zbrojem površina pravokutnika između parova susjednih amplituda UAE ovisno o spektru. Visina pravokutnika je određen srednjom vrijednosti između amplituda, dok širina je određena vremenom između amplituda. Grubi prikaz frekvencijske informacije iz značajki vremenske domene daju značajke procjene karakteristične frekvencije vrhova prije i poslije maksimalnoj amplitude i čitavog signala:

* **Srednja frekvencija** (*engl. average frequency*)
  + Frekvencija pojavljivanja svih vrhova u validnoj UAE. Računa se po formuli (5) te izražena je u herzima.

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

* **Inicijacijska frekvencija** (*engl. initiaton frequency*)
  + Frekvencija pojavljivanja vrhova u vremenu uzdizanja validne UAE. Računa se po formuli (6) te izražena je u herzima.

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

* **Frekvencija odjeka** (*engl. reverberation frequency*)
  + Frekvencija pojavljivanja vrhova u vremenu pada validne UAE. Računa se po formuli (7) te izražena je u herzima.

|  |  |  |
| --- | --- | --- |
|  |  | (7) |

Za kvalitetan prikaz informacija frekvencijske domene značajke se računaju na jednostranom amplitudnom spektru UAE s oznakom *UAE\_fft\_abs(f)*. Frekvencijske značajke su definirane na temelju magnituda amplitudnog spektra UAE te frekvencija karakterističnih dijelova UAE:

* **Maksimalna magnituda** (*engl. peak magnitude*)
  + Maksimalna apsolutna magnituda amplitudnog spektra UAE izražena u voltima. . (slika 12 crveni pravac)
* **Frekvencija maksimalna magnitude** (*engl. peak frequency*)
  + Frekvencija maksimalne apsolutne magnitude amplitudnog spektra UAE izražena u herzima. . (slika 12 plava isprekidana linija)
* **Frekvencija centroida** (*engl. centroid frequency*)
  + Frekvencija centroida amplitudnog spektra UAE izražen u herzima. Računa se po formuli (8). . (slika 12 crvena isprekidana linija)

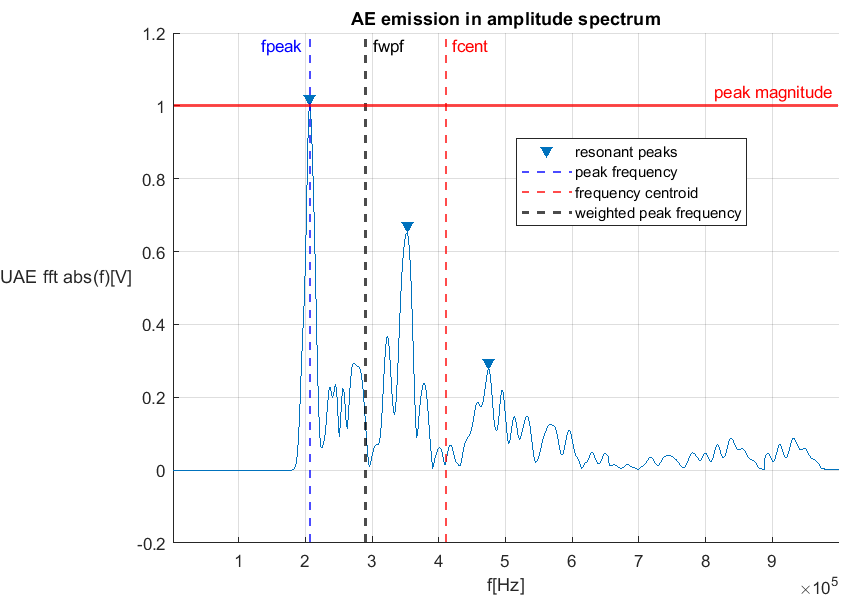
|  |  |  |
| --- | --- | --- |
|  |  | (8) |

Grupacije emisija na temelju značajke kao frekvencija maksimalne magnitude najčešće su povezane s rezonantnom frekvencijom senzora. Zbog toga ne čini se prikladnim koristiti samo takve značajke za razlikovanje izvora UAE. Kako bi izbjegli utjecaj senzora na rezultate izabrane su i značajke koje opisuju relativnu frekvenciju maksimalne magnitude te koje su pokazano kao dobri klasifikatori UAE po izvoru[4, 5]:

* **Težinska maksimalna magnituda** (*engl. weighted peak frequency*) ili **WPC**
  + Težinska maksimalna magnituda amplitudnog spektra UAE koristi se kombinacija informacija značajki PF i CF kako bi smanjila podložnost na utjecaj senzora.[2] Računa se po formuli (9) te izražena je u herzima. (slika 12 crna isprekidana linija)

|  |  |  |
| --- | --- | --- |
|  |  | (9) |

* **Broj rezonantnih frekvencijskih vrhova** (*engl. number of frequency peaks*)
  + Smanjuje utjecaj senzora na grupacije izborom vrhova određene širine te razlike u vrijednosti od susjednih magnituda. Pokazano kvalitetan za korištene skupove podataka. (slika 12 vrhovi označeni trokutima)



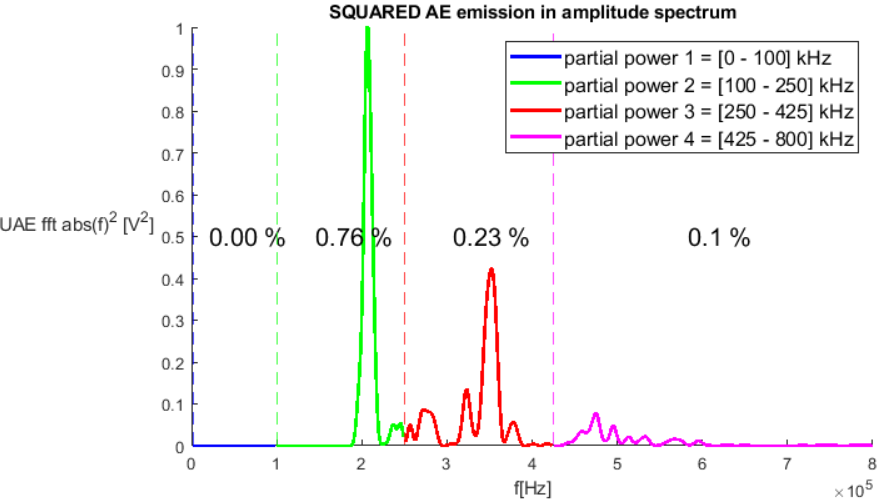
Slika 12 Prikaz frekvencijski značajki na tipičnoj UAE

Posljednje izabrane su frekvencijske značajke koje mjere kontribucije određenih frekvencijskih pojasa u ukupnom frekvencijskom spektru. Parcijalna snaga određenog frekvencijskog pojasa između frekvencija *f1* i *f2* određena je omjerom snage signala u određenom frekvencijskom pojasu i ukupne snagu frekvencijskog spektra. Računa se po formuli (10) te izražena je u postocima.

|  |  |  |
| --- | --- | --- |
|  |  | (10) |

Za korištene skupove podataka na temelju prijašnjih analiza opisanih u [4] formiraju se karakteristične grupe u određenim frekvencijskim pojasevima. Na temelju očekivanih grupacija izabrane su parcijalne snage prikazane slikom 13:

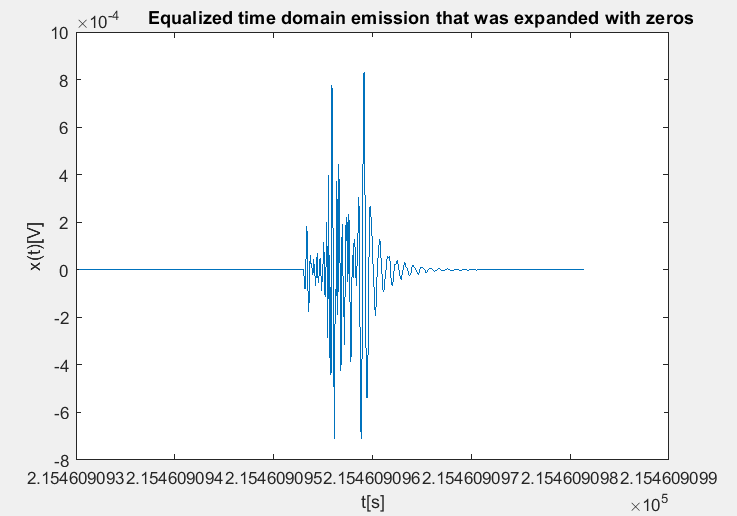
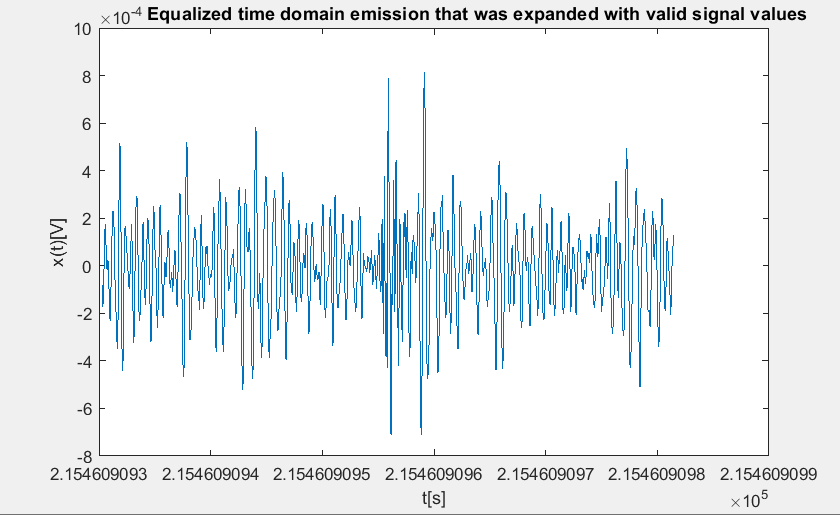
* **Parcijalna snaga** 1 (*engl. partial power 1*) (slika 13 plavo)
  + Parcijalna snaga u frekvencijskom pojasu između 0 i 100 kHz. Očekivano najmanji postotak zbog utjecaja ekvilizacije te senzorskog propusnog pojasa.
* **Parcijalna snaga** 2 (*engl. partial power 2*) . (slika 13 zeleno)
  + Parcijalna snaga u frekvencijskom pojasu između 100 i 250 kHz. Očekivano najviši postotak UAE kako obuhvaća primarnu grupaciju UAE.
* **Parcijalna snaga** 3 (*engl. partial power 3*) . (slika 13 crveno)
  + Parcijalna snaga u frekvencijskom pojasu između 250 i 425 kHz. Očekivana većina preostalih emisija koje obuhvaćaju drugu grupaciju.
* **Parcijalna snaga** 4 (*engl. partial power 4*) (slika 13 ljubičasto)
  + Parcijalna snaga u frekvencijskom pojasu između 450 i 800 kHz. Očekivano mali postoci UAE.



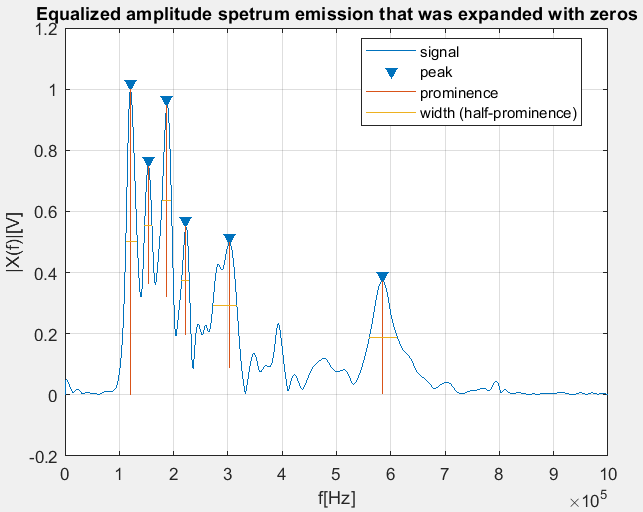
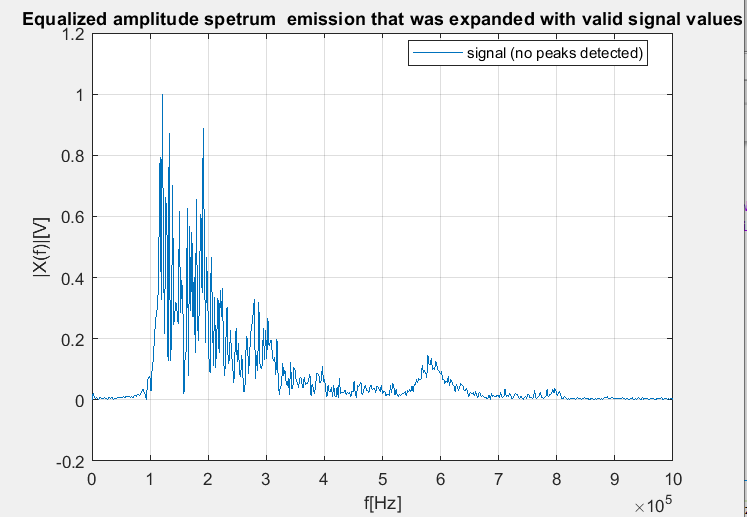
Slika 13 Prikaz parcijalnih snaga tipične UAE

## Analiza rezultata ekstrakcije značajki

Na slici 17 prikazana je statistička analiza izbačenih UAE na temelju kriterija opisanih u poglavlju 2.1. te raspodjela emisija u grupe po broja rezonantnih vrhova za svaki skup podataka. Usporedbom s rezultatima iz [4] uočavamo smanjenje broja ukupnih emisija zbog strožeg izbora validnih UAE. Isto tako uočavamo drastične promjene u broju UAE bez rezonantnih vrhova nastale zbog proširivanja UAE nulama (slika 16 desno) umjesto okolnim amplitudama signala koji mogu uključivati šum (slika 16 lijevo). Time rezonantni vrhovi frekvencijskoj domeni postaju izraženiji te frekvencijski spektar UAE se sastoji od jednog ili više rezonantnih vrhova. Na slici 17 lijevo vidi se tipična emisija bez rezonantnih vrhova opisana u [4]. Nakon promjene u proširivanje emisije nulama frekvencijski spektar pronalazi više rezonantnih vrhova. (slika 17 desno)

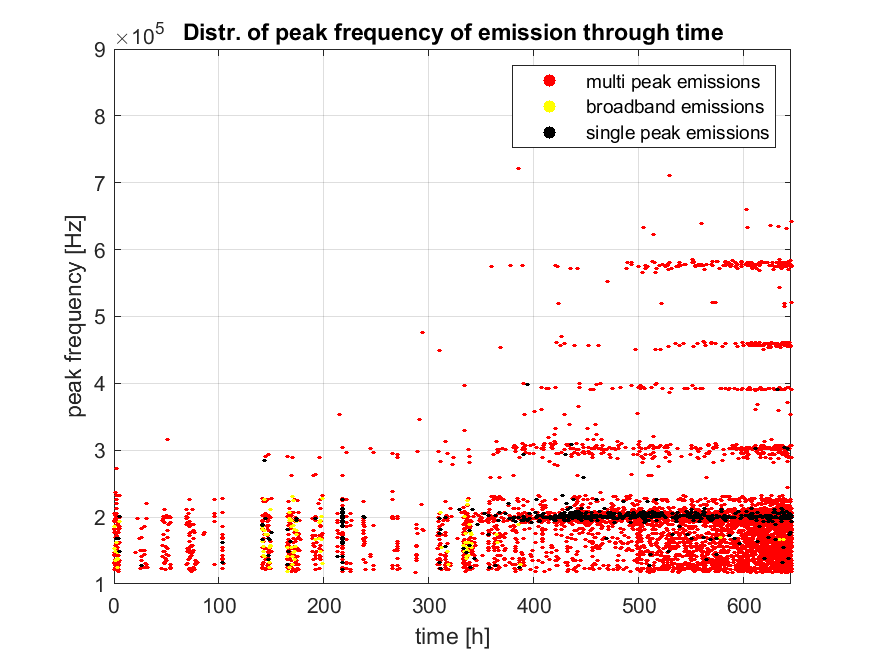


Slika 14 Pronađena emisija proširena okolnim amplitudama (lijevo) ili proširena nulama (desno) u vremenskoj domeni

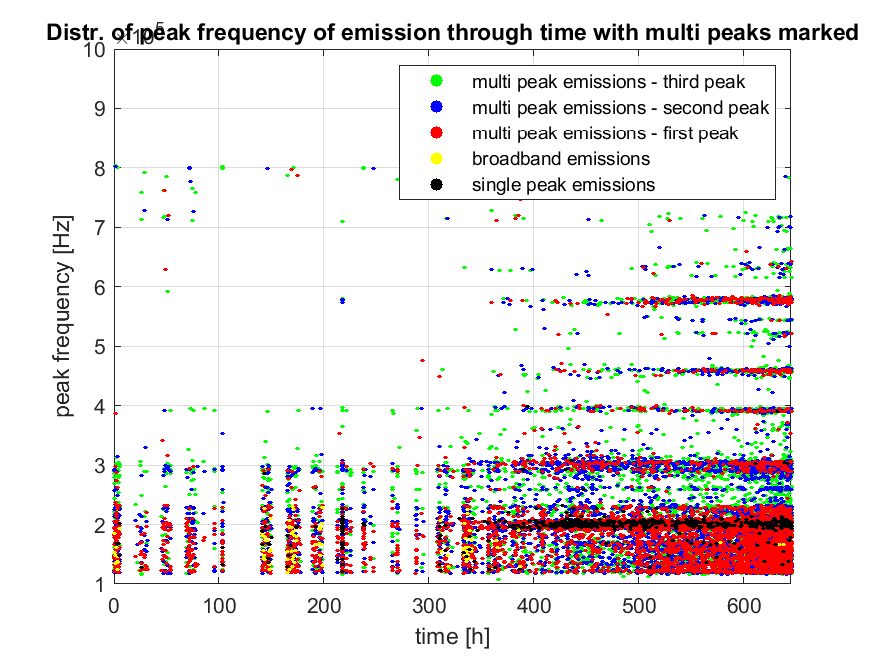


Slika 15 Pronađena emisija proširena okolnim amplitudama (lijevo) ili proširena nulama (desno) u frekvencijskoj domeni

Prikazom frekvencija maksimalne magnitude UAE u vremenu dolaska UAE uočljive su slične karakteristične grupacije iz [4]. Navedeno je prikazano slikom 18 samo za skup podataka dobiven iz eksperimenta 6 kako je najreprezentativniji. Uočljivo je kako UAE bez rezonantnih vrhova (žute točke) zamjenjuje UAE s više vrhova (crvene točke), dok UAE s jednim rezonatnim vrhom (crne točke) ostaju grupirane oko 200 kHz. Na slici 19. vidi se detaljniji prikaz UAE s više rezonantnih vrhova gdje su frekvencije prva tri vrha najvećih magnituda označene slijedno bojama crveno, plavo i zeleno.

****

Slika 16 Prikaz frekvencija maksimalne magnitude UAE u vremenu dolaska UAE s označenim UAE bez(žuto), s jednim (crno) i s više rezonantnih vrhova (crveno)

****

Slika 17 Prikaz frekvencija maksimalne magnitude UAE u vremenu dolaska UAE s označenim UAE bez(žuto), s jednim (crno) i s više rezonantnih vrhova za prvi 3 najvećih amplituda (#1 crveno, #2 plavo, #3 zeleno)

**Promatranjem distribucija značajki za više skupova podataka uočavamo sličnosti.**

**Opiši distribucije pojedniih bitnih značajki.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Eksperiment** | **Broj potencijalnih emisija** | **Nisko frekvencijske smetnje** | **Izbačene po trajanju** | **Izbačene po amplitudi** | **Broj validnih emisija** |
| **6** | 17073 | 8792 | 2421 | 0 | **6473** |
| **7** | 4969 | 434 | 602 | 2 | **3983** |
| **8** | **2338** | **189** | **400** | 0 | **1762** |
| **9** | **1005** | **17** | 2 | 0 | **1020** |
| **10 (1/2)** | **477** | **8** | 0 | 0 | **487** |
| **10 (2/2)** | **2161** | **25** | 4 | 0 | **2151** |
| **13** | **724** | **39** | **5** | 1 | **745** |
| **14 (2/3)** | **480** | **41** | **2** | **3** | **537** |
| **14 (3/3)** | **5608** | **110** | **2** | 4 | **5753** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Eksperiment** | **Single peak %** | **Multiple peak %** | **Broadband %** | **%** |
| **6** | **9.43** | **88.95** | **1.61** | **87.56** |
| **7** | 124 | 3855 | 4 | 105 |
| **8** | **14.25** | **85.75** | **0** | **94.41** |
| **9** | **15** | **84.9** | **0.1** | **97.38** |
| **10 (1/2)** | **9.65** | **90.35** | **0** | **97.87** |
| **10 (2/2)** | 9.57 | 90.33 | 0.1 | 98.05 |
| **13** | **15.84** | **83.76** | **0.4** | **93.22** |
| **14 (2/3)** | **64.52** | **35.25** | **0.23** | **99.05** |
| **14 (3/3)** | **4.01** | **95.6** | **0.39** | **67.46** |

## Skaliranje skupa značajki

Skaliranje je potrebna kako bi algoritam strojnog učenja svaku značajku gledao s jednakom važnosti. Skaliranjem značajke na različitim mjernim rasponima svode se na zajednički raspon vrijednosti uz očuvanje originalne distribucije. Izabrane značajke skalirati će se na vrijednosti od 0 do 1 izračunom prikazanim formulom (11) gdje X označava vrijednosti značajke.

|  |  |  |
| --- | --- | --- |
|  |  | (11) |

Određene značajke mogu sadržavati stršeće vrijednosti (*engl. outliers*) koje iskrivljuju distribuciju značajke. U navedenom slučaju potrebno je ispraviti originalnu distribuciju prije skaliranja. Konačni izbor minimalne i maksimalne vrijednosti značajke odrediti će se na temelju mjernih granice značajki. Na primjeru vremenskih značajki maksimum mjerne granice određen je maksimalnom veličinom emisije od 512 µs. Međutim za značajke kao maksimalna amplituda ili energija mjerna granica je neodređena te ovisi o fizikalnim svojstvima uzroka UAE koji su nepoznati. Za takve značajke fiksne granice određuju se na temelju ukupne minimalne i maksimalne vrijednosti izabranih skupova podataka poslije izbacivanja stršećih vrijednosti prikazane slikom 18.

Izbacivanje stršećih vrijednosti temelji se na izračunu raspona vrijednosti između prvog i trećeg kvartila ili IQR. Navedeni raspon između 25. i 75. percentila nije prikladan za distribucije izračunatih značajki jer identificira ispravne vrijednosti kao stršeće. Slika za neispravno.

Promjenom raspona IQR-a između 5. i 95. percentila dobivaju se zadovoljavajući rezultati. Slika za ispravni.

Pomoću modificiranog IQR-a računaju se maksimalne i minimalne vrijednosti svih značajki bez stršećih vrijednosti posebno za svaki korišteni skup podatak na temelju formule (12).

|  |  |  |
| --- | --- | --- |
|  |  | (12) |

Na slici 18. vidi se završni izbor maksimalnih i minimalnih vrijednosti pojedine značajke koji će koristiti za skaliranje značajki za svaki skup podataka.

|  |  |  |
| --- | --- | --- |
| **Značajka** | **Završni izbor granica** | **Zaokružene vrijednosti granica** |
| **rise time [s]** | [0.00e+00 , 2.47e-04] | [0 , 515 us] |
| **counts to [#]** | [1.00e+00 , 2.54e+02] | [0, 350] |
| **counts from [#]** | [5.00e+00 , 3.13e+02] | [0, 350] |
| **duration [s]** | [2.56e-04 , 5.12e-04] | [0 , 515 us] |
| **peak amplitude [V]** | [9.71e-05 , 1.08e-01] | [0 , 66 mV] |
| **average frequency [Hz]** | [1.15e+05 , 9.56e+05] | [0 , 1 MHz] |
| **rms [V]** | [1.55e-05 , 1.96e-02] | [0 , 11.4 mV] |
| **asl [dB]** | [-3.60e-05 , -1.26e-05] | [-0.36 , -0.126 udB] |
| **reverbation frequency [Hz]** | [2.38e+04 , 1.04e+06] | [0 , 1 MHz] |
| **initiation frequency [Hz]** | [7.55e+04 , 4.00e+06] | [0 , 4 MHz] |
| **signal strength [Vs]** | [1.35e-09 , 2.32e-06] | [0 , 1.27 uVs] |
| **absolute energy [J]** | [6.44e-18 , 3.67e-12] | [0 , 3.7 pJ] |
| **PP1 [%]** | [1.52e-07 , 1.59e-02] | [0 , 1] |
| **PP2 [%]** | [1.30e-02 , 1.00e+00] | [0 , 1] |
| **PP3 [%]** | [2.09e-04 , 9.58e-01] | [0 , 1] |
| **PP4 [%]** | [4.74e-07 , 9.57e-01] | [0 , 1] |
| **centroid frequency [Hz]** | [1.48e+05 , 5.35e+05] | [0 , 1 MHz] |
| **peak frequency [Hz]** | [1.15e+05 , 8.03e+05] | [0 , 1 MHz] |
| **amplitude of peak frequency [V]** | [2.28e-06 , 4.35e-03] | [0 , 2.79 mV] |
| **num of freq peaks [#]** | [0.00e+00 , 1.70e+01] | [0 , 10] |
| **weighted peak frequency [Hz]** | [1.36e+05 , 6.52e+05] | [0 , 1 MHz] |
| **total counts [#]** | [3.00e+01 , 3.41e+02] | [0, 350] |
| **fall time [s]** | [1.01e-04 , 4.29e-04] | [0 , 515 us] |

Pregledom distribucija dobivenih s skaliranjem s fiksnim globalnim parametrima za određene skupove podataka nije ispravilo distribuciju koje unose stršeće vrijednosti. Slika potrebna! Ovakav rezultat je očekivani kako globalni maksimumi najviše ovise o skupu podataka s najvećim vrijednostima. Bolje rješenje od fiksnih parametara bila bi normalizacija na varijancu (engl*. z-score*) gdje se vrijednosti značajke oduzimaju se sa srednjom vrijednosti i podijele sa standardnom devijacijom svake značajke prikazano formulom (13).

|  |  |  |
| --- | --- | --- |
|  |  | (13) |

U tome slučaju normalizirani podatci imaju srednju vrijednost 0 te standardnu varijaciju 1. Time će se sve značajke grupirati oko nule što otežava odvajanje grupa značajki te drastično mijenja originalne distribucije značajki. Zbog toga jedino primjenjivo rješenje je prije opisano skaliranje.

# Nenadzirano i polu-nadzirano strojno učenje

Ekstrakcijom značajki dobiva se skup značajki koji algoritam strojnog učenja treba grupirati po izvoru UAE. Izbor algoritma strojnog učenja najviše se bazira na označenosti ulaznog skupa značajki. Većina izabranih skupovi podataka je neoznačena, dok skup podataka dobiven eksperimentom 6 je djelomično označen. Unutar eksperimenta 6 faza sušenja nastupa nakon 350. sata izvođenja eksperimenta kako su uvjeti sušenja bili uzrokovana prirodnim uvjetima. Time poznato je da UAE unutar prvih 350 h nisu bile uzrokovane kavitacijom. Navedene UAE označene se kao ne uzrokovane kavitacijom. Ne i polu označenost korištenih skupova podataka ograničava izbor metoda strojnog učenja na nenadzirano učenje i polu-nadzirano učenje. Nenadzirano strojno učenje koristi metode strojnog učenja za analizu ili grupiranje neoznačenog skupa podataka. Nenadzirano učenje po vrsti zadatka može se podijeliti na grupiranje (*engl.* clustering), procjenu gustoće (engl. *density estimation*), otkrivanje novih ili stršećih vrijednosti (engl. *novelty/outlier detection*) te smanjenje dimenzionalnosti (engl*. dimensionality reduction*). Polu-nadzirano učenje koristi se kada je mali dio skup podataka označen, dok je većina ne označena. Grupiranja neoznačenog dijela podataka potpomognuto je poznatim oznakama za optimalnije odvajanje grupa. U ovom radu koristiti će se smanjenje dimenzionalnosti radi početne analize skupa značajki opisano u poglavlju 3.1. na temelju koje će se raditi grupiranje podskupa značajki opisano u poglavljima 3.2. i 3.3.

## Smanjenje dimenzionalnosti

Kod velikog broja značajki smanjenje dimenzionalnosti omogućuje lakši primjenu metoda strojnog učenja nad značajkama. U ovom radu smanjenje dimenzionalnosti omogućiti će početnu korelacijsku te vizualnu analizu skupa značajki koja će pomoći kod selekcije značajki. Isto tako omogućiti će izbacivanje određenih značajki iz skupa značajki. Metoda prikladna smanjenje dimenzionalnosti dobivenog ulaznog skup značajki je analiza principalnih komponenti ili PCA [3, 24].

PCA transformira ulazne značajke u jednak broj novih značajki koje nazivamo principalne komponente. Principalne komponente rangirane su po količini varijacije koje sadrže iz početnog skupa značajki. Prva principalna komponenta čini većinu varijance unutar skupa značajki. Druga principalna komponenta čini drugi najveću količinu varijance unutar skupa značajki te ne korelira s prvom principalnom komponentom. Ostale principalne komponente prate navedeni postupak. Pretpostavka PCA je da visoka varijanca značajke predstavlja veću količinu informacije koju značajki pridonosi skupu podataka. Na temelju navedene pretpostavke principalne komponente s niskom varijacijom se zanemaruju te posljedično smanjuje se dimenzionalnosti skupa značajki. Isto tako značajke koje visoko koreliraju s principalnim komponentama visoke varijacije su optimalne za korištenje pri grupiranju skupa podataka.

Preduvjeti za primjenu PCA jesu normalizirani ulazni skup značajki te postojanje primjetljivih korelacijskih odnosa između značajki. Normalizaciju je potrebno odraditi normalizacijom na varijancu opisanu formulom (13) kako PCA koristi standardne devijacije značajki za izračun principalnih komponenti. Korelacijski odnosi definiraju mogućnosti linearne kombinacije značajki u nove značajke kako bih se smanjila dimenzionalnost skupa značajki. [22] Analiza međusobnih korelacija značajki nad svim skupovima pokazuje konzistentno pojavljivanje 6 sličnih visoko pozitivno koreliranih grupa unutar skupa značajki prikazani slikom 19. Isto tako mogu se izdvojiti 3 visoko negativno korelacijska odnosa između grupa određenih značajki prikazani slikom 20.

Većina visoko korelacijskih grupa posljedica je sličnosti izračuna navedenih značajki. Iznimka je odnos 1 prikazan slikom 19 koji je uzrok karakterističnih svojstva UAE gdje se većina energije frekvencijskog spektra raspoređuje isključivo ispod 250 kHz u PP2 pojasu ili iznad 250 kHz u PP3 pojasu. Visoke energije u PP2 pojasu očituju se niskom srednjom energijom frekvencijskog spektra što se očituje u manjim vrijednostima WPC i frekvencije centroida.

Slika na kojoj se prikazuje stol

Opis je automatski generiran

Slika 18 Konzistetne visoko pozitivno koreliranih grupa unutar skupa značajki

Slika na kojoj se prikazuje stol

Opis je automatski generiran

Slika 19 Visoko negativni korelacijski odnosi između grupa određenih značajki unutar skupa značajki

Rasponi pozitivnih korelacija unutar visoko koreliranih grupa značajki iznose između 65% i 99%, dok rasponi negativnih korelacija između grupa značajki iznose između -50% do -90%. Navedeni rasponi prikazani zadnjim stupcem na slikama 19 i 20 dobiveni su izračunom maksimalne i minimalne korelacije između svih značajki za sve skupove podataka. Time su određeni najgori i najbolji slučaj korelacije između značajki koji ako poprimaju visoke vrijednost dokazuju konzistentnost korelacija. Koreleogrami za najveće i najmanje korelacije između značajki za sve skupove podataka vide se na slikama 22 i 23 slijedno.



Slika 20 Koreleogram najvećih korelacija između značajki za sve skupove podataka



Slika 21 Koreleogram najmanjih korelacija između značajki za sve skupove podataka

Kako korelacijski odnosi svih skupova podataka uključuju slične grupe značajki u nastavku PCA će se izvesti samo na najreprezentativnijem skupu podataka dobivenim eksperimentom 6. [4] Postupak PCA započinje izračunom kovarijance između svih značajki po formuli (14) koje čine u matricu kovarijance.

|  |  |  |
| --- | --- | --- |
|  |  | (14) |

Iz matrice kovarijance računa se dekompozicija na eigen vrijednosti te eigen vektore. [23] Eigen vektori predstavljaju principalne komponente, dok eigen vrijednosti predstavljaju iznos varijance čitavog skupa značajki koju određena principalna komponenta sadrži. [23] Slika 23 pokazuje kako prvih 6 principalnih komponenti sadrži 85% varijance unutar skupa značajki. Time daljnja analiza fokusirati će se na navedene principalne komponente.



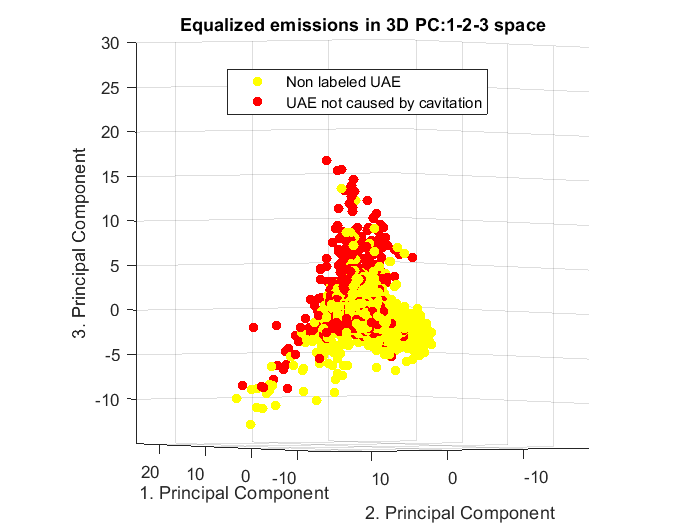
Slika 22 Prikaz postotka varijance koje prvih 10 komponenti unosi u skup značajki. Crveno su označene principalne komponente koje čine 85% varijance skupa značajki.

Korelogramom na slici 24 dobiva se aproksimirani uvid kombinacijom kojih značajki su dobivene određene principalne komponente. Značajke koje najviše definiraju određenu komponentu koristiti će se kao informacija potencijalno kvalitetnih značajki za korištenje kod grupiranja.

Prva principalna komponenta najviše korelira s energijskim značajkama opisanim grupom 3 na slici 18. Druga principalna komponenta snažno korelira s PP2 te opisuje odnos 1 na slici 19. Treća principalna komponenta visoko negativno korelira s energijskim značajkama te snažno korelira s grupama značajki 1 i 4 na slici 18 povezanim s vremenom i brojem vrhova prije maksimalne amplitude. Četvrta principalna komponenta prati odnos 3 na slici 19 te visoko korelira s grupom 2 pa opisuje vrijeme i broj vrhova poslije maksimalne amplitude. Peta principalna komponenta visoko korelira s PP1 i incijacijskom frekvencijom pa opisuje frekvencijski pojas prije maksimalne amplitude. Šesta principalna komponenta opisuje pojas PP3 s kojim snažno korelira. Kako duracija ima relativno mali utjecaj na principalne komponente možemo ju izbaciti iz skupa značajki. Zbog normalizacije s varijancom navedeni podaci u prostoru principalnih komponentna grupirani su bez vizualno distinktnih grupa. Navedeni podaci prikazani su u prostoru prve 3 principalne komponentne na slici 25. Označavanjem UAE emitiranih do 350 h, prikane na slici 25 crveno, ne uspijevaju se izdvojiti vizualno distinktne grupe u prostoru principalnih komponenti.



Slika 23 Koreleogram principalnih komponenti i originalnih značajki za skup značajki dobiven eksperimentom 6



Slika 24 UAE ne uzrokovane kavitacijom (crveno) i nepoznatog uzroka (žuto) u prostoru prve 3 principalne komponente

## Grupiranje skupa značajki

Grupiranje skupa je najvažniji korak prilikom određivanja izvora UAE. Grupiranje je postupak pronalaženja prirodnih grupa unutar skupa značajki na temelju pretpostavke postojanja navedenih grupa. Navedeni skup podataka te ugradbena implementacija algoritma grupiranja postavljaju određene zahtjeve koja metoda grupiranja treba zadovoljiti:

1. Deterministička pripadnost grupi – svaki podatak je grupiran u samo jednu grupu
2. Početno nepoznat broj grupa – grupiranje samo određuje prirodnih broj grupa
3. Robusnost svojstva grupe – podržano grupiranje grupa arbitrarnog oblika, veličine i gustoće
4. Otpornost na šum – grupiranje prepoznaje i odvaja šum od grupa
5. Neovisnost o redoslijedu i količini pristizanja podataka – grupiranje proizvodi slične grupe neovisno o redu i količini pristizanja podataka
6. Niska složenost – grupiranje je izvedivo na

Većina metoda baziranih na patricijskom i hijerarhijskom grupiranju dizajnirano je za pronalazak grupa sferičnog i konveksnog oblika. Metode grupiranja koje podržavaju grupe arbitrarnog oblika grupiraju ne temelju određivanja gustoće podataka te skraćeno se nazivaju DBCLA. DBCLA metode iz početnog nepoznatog broja grupa određuju prirodne grupe na temelju pronalasku područja veće gustoće prostora značajki. Isto tako DBCLA su otporne na šum koji je određen područjima niske gustoće prostora značajki. Glavna podjela DBCLA

Metode bazirane na mreži nisu prikladne kod grupa različitih gustoća kako gustoća svake ćelije se određuje po fiksnim parametrima. Izbor dimenzija ili granulacije ćelija izrazito ovisi o gustoći ulaznog skupu podataka. Kako je cilj metoda baziranih na mreži ubrzavanje komputacije algoritma zbog diskretizirane prostora u ćelije izračun prostora ćelija svaki put tijekom grupiranja unosi prevelike komuptacijske i memorijske zahtjeve.

Proučavanjem DBCLA metode uočava se novi zahtjev potreban zadovoljiti a to je osjetljivost na promjenu ulaznih parametara. Kako bih metoda grupiranja mogla raditi za više skupova podataka s različitom distribucijom gustoća u prostoru značajki potrebna je određena neovisnost na ulazne parametre.

Kod metoda grupiranja baziranih na mreži potrebno je prilagoditi veličinu ćelije mreže i prag gustoće. Postoji mogućnost krivog rezultata ako su grupe različite gustoće

Kvaliteta STING grupiranja ovisi granularnost najniže razine strukture mreže. Usprkos brzom izračunu rezultira netočnim grupama.

DBCLASD vrijeme izvođenja je dvostruko od DBSCAN algoritma

HDBSCAN – Iako je poboljšanje nad OPTICS algoritmom, uvodi komputacijski zahtjevno za ugradbenu verziju

Dsets-DBSCAN – samo radi za segmentaciju slika

Dynamic -density based clustering – nije primjeren za sekvencijalan rad kako je potrebno

OPTICS – postoji iskoristiva implementacija naizgled primjenjiva za naš problem

Metode bazirane na gustoćo podataka mogu se podijeliti po definiciji gustoće

1. Bazirana na točki –
   1. DBSCAN - ekstrakcija clustera arbitrarnih oblika i s šumom, međutim nemože detektirati clustere različitih gustoća
   2. CLIQUE – prikladan za više dimenzionalne podatke, međutim kvaliteta clustera nije garantirana
   3. GDBSCAN – mogućnost clusteriranja točaka i oblika u prostiru međutim nema heuristike za određivanje prikladnih parametara na ulazu algoritma
   4. Inc-DBSCAN – inkrementalni DBSCAN -senzitivan na promjenu parametar te nesmiju se brisati podaci, isto tako dodaje točku po točku što je komputacijski skupo te pamti sve podatke, dobar samo zato što može dodati jednu točku bez da se računa ponovo čitavi algoritam
   5. PDBSCAN (paralelan DBSCAN) – omgućuje pralelno izvođenje DBSCAN na više računala za brži izračun
   6. **OPTICS - Proizvodi prošireni poredak skupa podataka umjesto eksplicitne strukture grupiranja. Učinkovito ekstrahira intrinzičnu strukturu klastera. Adaptivan na promjene parametara. Međutim za visokodimenzionalne podatkovne prostore ne postoje indeksne strukture za učinkovitu podršku hiperprostora raspona upita. Složenost je O(n2) ili O(nlogn) ako se korsiti struktura drveta**
   7. SDBSCAN – ne detektira clustere varijabilne gustoće
   8. **SNN-DBSCAN – koristi koncept shared neighoirs neigbora gdje se ne gleda udaljenosti između točaka već samo koji su najbližu susjedi slični kod više točaka, zbog toga real time implementacija?**
   9. DBDC – distribuirano clusteriranje dok mi želimo sve na jednom sustavu
   10. IDBSCAN – nemože pronaći clustere varijabilne gustoće
   11. **ST-DBSCAN** – isto kao DBSCAN ali korsiti više epsilona za svaki tip ne spacijalnog podataka. Npr. za grupu podataka računa udaljenosti za prvi epsilonte razlike u temperaturama za drugi epsilon.
   12. **KIDBSCAN – Kmeans + IDBSCAN**
   13. CUDA-DClust – korsiti GPU za učinkvoitije izvođenje
   14. DVBSCAN – nemoguće automatkso određivanje parametra pa nemoguća primjena na drukčije datasetove – drukčije gustoće
   15. P-DBSCAN – za slike s oznakom geografske pozicije
   16. DADB – distribucijski algoritam gdje vilika potrošnja na spajanje distribuiranih rezultata
   17. MR-DBSCAN (MapReduce-DBSCAN) – paralelizam
   18. HDBSCAN - Runs is quadratic time complexity. Dobro procjenjuje epsilon
   19. Cludoop – sitribuirano klasteriranje na hadoop platformi
   20. DSets-DBSCAN – služi samo za segmentaciju slika
   21. **Dynamic density based clustering** – mora pamtiti sve postojeće točke ali može dinamički dodavati i micati točke
   22. **MBSCAN** – prevelika senzitivnost parametara, koristi matricu disimilariteta umjesto udaljenosti što je navodno efikasnije

Grid based metode se čine bolje jer podijele prostor na dijelove pa je lakše dinamički nadograditi.

Na temelju ovisnosti algoritama o puno parametara možemo izbaciti MBSCAN, ST-DBSCAN, IDBSCAN pa KIDBSCAN, DVBSCAN,

Neovisni na parametre su OPTICS, HDBSCAN, Dynamic density based clustering,

1. Bazirana na mreži
2. Vjerojatnosno po distribuciji
3. Ovisno o podacima

Metode bazirane na gustoćo podataka mogu se podijeliti po prilagodljivosti na vrijednosti parametra

1. Ovisne o parametru
2. Prilagođuju se parametru

Način izvršavanja

1. Serijalni
2. Paralelni
3. Distribuirani

Vrsta podataka:

1. Spacijalni – lokacija na površini zemlje
2. Ne spacijalni
3. Mutlimedijski – slike, grafovi, videi

OPTICS je bolji od DBSCAN i DENCLUE. [19].

Proučavanjem literature pokazalo da je OPTICS algoritam najprikladniji za korišteni skup podataka. Isto tako već je dokazano radi na našem skupu podataka te je implementira na ugradbenom računalnom sustavu.

Pregled metoda grupiranja baziranih na gustoći podataka: [3.3]

Slika na kojoj se prikazuje tekst

Opis je automatski generiran

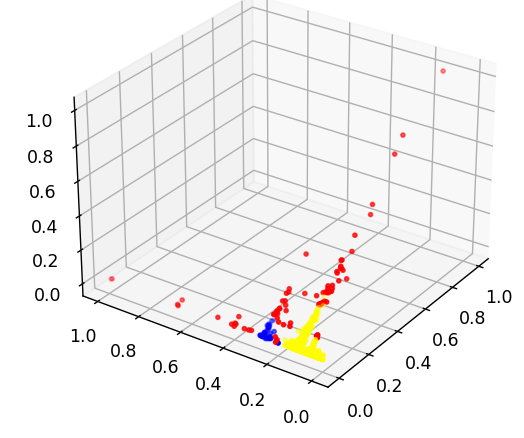
Metode grupiranja možemo podijeliti na temelju zadanih svojstva grupa podataka:

1. **Interna struktura grupe podataka** – postojanja interne strukture grupa podataka gdje grupe sadrže podgrupe. Kod particijskog grupiranja svaka grupa podataka je zasebna, dok kod hijerahijskog grupiranje svaka grupa može sadržavati podgrupe;
2. **Pripadnost podataka grupama** - kod čvrstog grupiranja svaki podataka pripada samo jednoj grupi, dok kod mekog grupiranja podaci mogu pripadati više grupa gdje pripadnost određenoj grupi je određena vjerojatnošću;
3. **Sličnosti podataka unutar grupa:**
   1. **Kvalitetno odvojene grupe** – svaki podatak u grupi je sličniji svakom drugom podatku unutar te grupe, nego podatcima izvan te grupe;
   2. **Bazirane na središtu grupe** – svaki podatak u grupi je sličniji središtu te grupe, nego središtima drugih grupa;
   3. **Bazirane na susjednim podacima unutar grupe** – svaki podatak u grupi je sličniji jednom ili više podataka unutar te grupe, nego podatcima izvan te grupe;
   4. **Bazirane na gustoću podataka** – raspoređenost podataka unutar grupe je veće gustoće, nego raspoređenost podataka između grupa. Primjenjuje se kod grupa arbitrarnog oblika i veličina te kod prisutnosti šuma ili stršećih vrijednosti;
   5. **Bazirane na konceptualnom svojstvu grupe** – grupe dijele zajedničko konceptualno svojstvo;
   6. **Bazirane na objektivnoj funkciji grupe** – grupe maksimiziraju ili minimiziraju određenu objektivnu funkciju. [3.2]

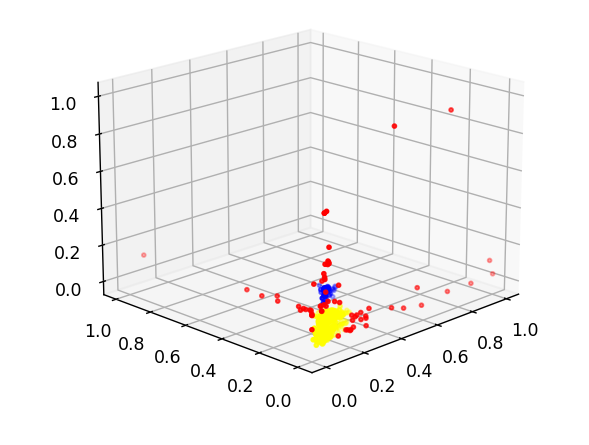
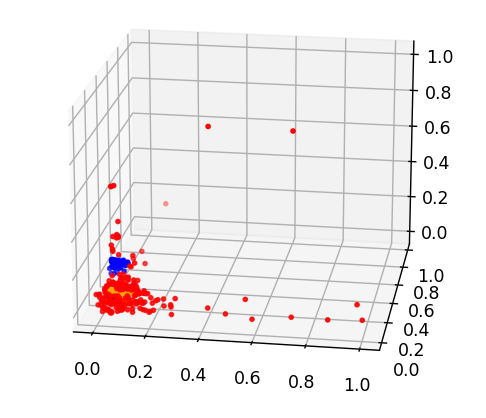
Svrha grupiranja je nalaženje prirodnih grupa u skupu podataka na temelju. Za navedeni skup podataka grupe će odgovarati izvorima AE emisija kod vodenog stresa biljaka. Zbog toga potrebno je grupirati u podatke u zasebne grupe gdje svaki podataka pripada jednoj grupi te grupe mogu biti arbitrarnog oblika, veličine i gustoće. Isto tako potrebno je prepoznati i odvojiti emisije od šuma. Iz navedenih zahtjeva vidi se kako je potrebno koristiti meke particijske metode grupiranja bazirane na gustoći podataka (DBCLA). DBCLA ovisi o određivanju gustoće određenih područja skupa podataka. Cilj DBCLA je pronaći grupe na različitim razinama gustoće uz odgovarajuće filtriranje šuma.

HDBSCAN služi za smanjenje broja ulaznih parametara u DBSCAN i mogućnost clusteriranja clustera različite gustoće. Iako za većinu slučajeva HDBSCAN relativno dobro clusterira, za određene slučajeve je potrebno manualno podesiti min\_points i epsilon parametre što ga čini ekvivalentom OPTICS-algoritmu što se tiče korištenja u svrsi automatskog clusteriranja.

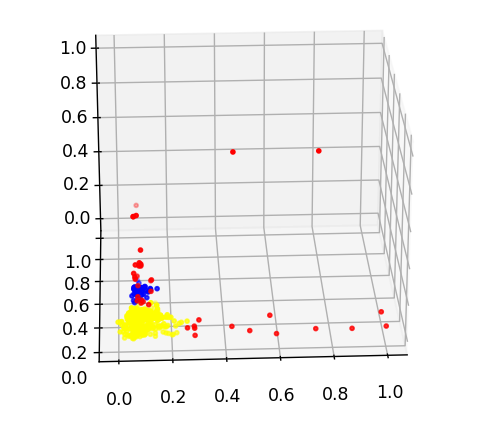
Na donjoj slici HDBSCAN je kvalitetno clusteriro bez podešavanja epsilon parametra.



Na slici dolje lijevo prikazano je clusteriranje bez podešenog epsilona već je HDBSCAN sam izračunao optimalni, dok na slici dolje desno prikazano je clusteriranje s podešenim epsilonom. Vidi se kako HDBSCAN nije prikladno odredio epsilon za prikazani podskup značajki. Zaključujemo da HDBSCAN nije prikladan za naš skupo podataka kako unosi dodatnu složenost, a daje približno jednake rezultate kao OPTICS algoritam. Isto tako mora primiti jednake parametre na ulazu kako bih došao do optimalnog rezultata što ga čini jednakim OPTICS-u prilikom selekcije značajki.



Ipak daje dobre rezultate ako mijenjamo odnos min\_points i min\_cluster\_size.



Opisuje općeniti proces ne-nadziranog strojnog učenja te uspoređuje ga s polu-nadziranim strojnim učenjem.

SSDBSCAN konceptualno za svaku ne labeliranu UAE dodjeljuje se labela od UAE koji je najbliži iz perspektive gustoće povezivosti (*engl. density connectivity*) ako se navedena UAE ne nalazi u clusteru s različitom labelom.

Polu-nadzirano strojno učenje za density based clustere nije prikladno ako labele se ne podudaraju strukturom gustoće clustera.

without violating the label consistency assumption (Assumption 5). This assumption requires that objects with different labels have to reside in disjunct density-based clusters following Definition 4 (w.r.t. the same mpts, but possibly different values).

Isto tako kako nije poznati ukupni broj labela …[20]

Korišteni skup podataka iz eksperimenta 6 jedini je primjenjiv za polu-nadzirano strojno učenje. U početnih 350 h eksperimenta navedenog skupa podataka nije došlo do vodenog stresa. Zaključujemo da emitirane UAE kada biljka nije u stanju vodenog stresa nemogu biti uzrokovane kavitacijama. Time labelira se jako mali dio skupa podataka kao UAE ne kavitacijskog uzrok. Proučavanjem semi-supervised metoda baziranih na grupiranju po gustoći zbog izabranih skupova podataka zaključujemo da nema prikladne metode polu-nadziranog strojnog učenja za tako izabrani skup podataka.

**Polu-nadzirano strojno učenje nije primjenjivo real-time već samo za selekciju značajki, zbog toga je potrebno fokusirati se na metode polu-nadzirane selekcije značajki.**

## OPTICS algoritam i implementacija

Opisuje OPTICS algoritam strojnog učenja na temelju [4] te implementaciju u Matlabu.

Nakon OPTICS izračun računa se gradient clustering na temelju OPTICS reachability plot. Prvo reachability plot sortira po dobivenom redoslijedu točaka te postavljaju se vrijednosti svih outlier točaka na 1024. Time će te točke sigurno biti točke infleksije te će odvajati zasebne clustere. Samo zbog toga unutar gradient clusteringa OPTICS otkriva sve clustere manje gustoće od epsilon i Nmin što mi se čini jednako kao DBSCAN. W parametar udaljenosti između točaka je optimalno da bude epsilon koljeno kako bi infleksija točke bila izračunato ispravno s obzirom na vrijednosti reachability distance-a. Zbog toga moguće je postaviti epsilon na Inf i možda otkriti clustere. Međutim gradient clustering onda otkriva previše manjih clustera unutar velikih koje spaja u jedan cluster. Isto tako teško prepoznaje kraj clustera ako je povećanje reachability distance-a podjednako i naglo padne, već to gleda samo kao početak novog clustera.

## Analiza rezultata grupiranja

Analizira rezultate strojnog učenja za optimalni podskup značajki na temelju znanja iz fiziologije bilja i procesa kavitacije.

# Selekcija značajki

## Pregled metodologije selekcije značajki

Definira pojam selekcije značajki unutar metodologije strojnog učenja.

Opisuje pregled općenitih postojećih metoda selekcije značajki podijeljenih na:

1. Filter - uključuje metode redukcije dimenzionalnosti kao **PCA**
2. Wrapper
3. Embedded
4. Hybrid

Spominje koja selekcija značajki će se koristiti u ovom radu i na temelju koje literature.

## Analiza značajki pomoću PCA

Opisuje analizu značajki pomoću PCA,

## Implementacija selekcije značajki

Detaljno opisuje metodu selekcije značajki koja se koristi u ovom radu na temelju [3].

## Indeksi validacije rezultata grupiranja (clustering-a)

Kratko opisuje općenite indekse i opće density-based indeksa.

Analizira detaljno izabrani indeks DBCV. [4]

Pokazuje rezultate zašto je dobar DBCV za naš skup podataka.

## Analiza rezultata selekcije značajki

Analiza rezultate selekcije značajki na temelju kojih bira podskup značajki koji dobro grupira podatke u distinktne skupine.

# Implementacija na ugradbenom računalnom sustavu

## Pregled ugradbenog računalnog sustava

Opisuje koji se URS koristi za implementaciju iz [2] .

Opisuje postojeći algoritam na URS-u te potrebna nadogradnju.

## Nadogradnja algoritma klasifikacije akustičnih emisija

Opisuje kako je implementirana nadogradnja i koji su bili problemi prilikom implementacije.

## Testiranje rada algoritma

Provjera točnost algoritma na snimljenom skupu podataka i trajanje izvođenja na odabranom ugradbenom sustavu.

Procjena doprinosa algoritma ukupnoj potrošnji ugradbenog sustava.

# Zaključak

Na kraju rada piše se kratak zaključak, duljine do najviše jedne stranice.

# Literatura

[0] B. Lazarević i M. Poljak, Fiziologija bilja. Zagreb, Sveučilište u Zagrebu Agronomski fakultet, 2019., str. 28 [Online] Dostupno na: <https://urn.nsk.hr/urn:nbn:hr:204:366622>

[1] L. Vergeynst, “Investigation and application of the acoustic emission technique to measure drought-induced cavitation in woody plants,” Ghent University. Faculty of Bioscience Engineering, Ghent, Belgium, 2015. [**http://hdl.handle.net/1854/LU-6925217**](http://hdl.handle.net/1854/LU-6925217)

[2] D. Oletić , S. Rosner , M. Zovko, V. Bilas, 2020., Time-frequency features of grapevine’s xylem acoustic emissions for detection of drought stress. Computers and Electronics in Agriculture. Volume 178, Studeni 2020, 105797

[3] Lidewei L. Vergeynst, Markus G.R. Sause, Niels J.F. De Baerdemaeker, Linus De Roo, Kathy Steppe, Clustering reveals cavitation-related acoustic emission signals from dehydrating branches, *Tree Physiology*, Volume 36, Issue 6, June 2016, Pages 786–796, <https://doi.org/10.1093/treephys/tpw023>

[4] [DBCV.pdf (lmu.de)](https://www.dbs.ifi.lmu.de/~zimek/publications/SDM2014/DBCV.pdf)

[5] Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). OPTICS. ACM SIGMOD Record, 28(2), 49–60. doi:10.1145/304181.304187

(Rosner 2012) Rosner, Sabine. (2012). Acoustic Emission Related to Drought Stress Response of Four Deciduous Broad-Leaved Woody Species. Journal of Acoustic Emission. 30. 11-20.

[2.1] – De Baerdemaeker, N.J., Stock, M., Van den Bulcke, J., De Baets, B., Van Hoorebeke, L., Steppe, K., 2019. X-ray microtomography and linear discriminant analysis enable detection of embolism-related acoustic emissions.

[2.2] P.A. Corporation, AEwin PCI-2 Based AE System User’s Manual. Mistras Group Inc., REV 3 (April), (2007) 1–312.

[2.3] [Amplitude Analysis: Root-mean-square EMG Envelope - Delsys](https://delsys.com/amplitude-analysis-root-mean-square-emg-envelope/)

[3.1] [SU-2020-19-Grupiranje[1].pdf (unizg.hr)](https://www.fer.unizg.hr/_download/repository/SU-2020-19-Grupiranje%5b1%5d.pdf)

[3.2] [Microsoft PowerPoint - ClusteringAnalysis.pptx (wmich.edu)](https://cs.wmich.edu/alfuqaha/summer14/cs6530/lectures/ClusteringAnalysis.pdf)

[3.3] [Sci-Hub | A survey of density based clustering algorithms. Frontiers of Computer Science, 15(1) | 10.1007/s11704-019-9059-3](https://sci-hub.se/10.1007/s11704-019-9059-3)

# Sažetak

# Summary

# Skraćenice

# Privitak

Privitak je također opcionalno poglavlje (u dogovoru s mentorom).

Sadržaj koji se stavlja u privitak je, općenito, nešto što je, kao cjelinu, prikladno izdvojiti iz sadržaja samog rada.

Mogući primjer je tehnička dokumentacija vezana uz diplomski rad - npr. električka i položajna shema sklopa, sastavnica, predložak tiskane veze, plan bušenja, ispis programa s detaljnim opisom.

Drugi primjer uključuju upute za korištenje rezultata rada (softvera ili hardvera), detaljni ispisi mjerenja čiji su rezultati sažeto ili grafički prikazani u radu. Ako se radi o softveru, uobičajeno je navesti podatke o platformi na kojoj se izvodi (npr., karakteristike uređaja i operacijskog sustava te pomoćnog softvera), kao i upute za instalaciju.

U privitku nemojte koristiti stilove razine Heading, već samo (nenumerirani) stil Podnaslov.

Na primjer:

Instalacija programske podrške

Upute za korištenje programske podrške

.

.

.

Ostali savjeti

U izborniku Tools - opcija Language postavite Croatian kao *default* jezik. Na kraju provedite strojnu provjeru teksta (*spell checking*, ako ga imate ugrađenoga), ali svakako pažljivo i pročitajte vlastiti tekst.

Ako nemate ugrađeni *spelling checker* za hrvatski jezik, možete se poslužiti Hascheckom (izgovara se Hašek, a ime dolazi od kratice za Hrvatski akademski *spelling checker*), dostupan putem poveznice <https://ispravi.me/>.