Poređenje metoda za klasifikaciju žanra muzike koristeći neuralne mreže

Autori:
Branko Grbić
Pavle Pađin
Mentor:
Milomir Stefanović

Apstrakt

Srpski

Cilj istraživanja je poređenje 4 različite metode (Fingerprint, spektar, MFCC i LPCC metode) za klasifikaciju žanra muzike koristeći neuralne mreže. Fingerprint metoda detektuje 10 komponenti spektra (frekvencija) najveće spektralne gustine signala. Spektar metoda podrazumeva korišćenje karakterističnih obeležja kao što su spektar signala i logaritam spektra signala. LPCC i MFCC su kepstralni koeficijenti, dobijeni pomoću kepstra signala na koje je primenjena Melova filterbanka (MFCC), odnosno linearne predikcije (LPCC). Korišćene su 2 baze koje sadrže audio fajlove, sačuvane u .wav formatu, gde je svaki audio fajl isečak iz pesme dužine 5 sekundi. Svaki audio fajl je označen žanrom muzike kome pesma iz audio fajla pripada. Prva baza je NATA, koja sadrži 6 različitih žanrova za klasifikaciju (klasična muzika, folk, haus, džez, R&B i rok), sa po 300 pesama po žanru. Druga baza je GTZAN, koja sadrži 10 žanrova za klasifikaciju (bluz, klasična muzika, kantri, disko, hiphop, džez, metal, pop, rege i rok), sa po 100 pesama po žanru. Najbolji rezultat nad NATA bazom je 95.9%, dobijen kombinacijom MFCC i Spektar karakterističnih obeležja, dok je nad GTZAN bazom dobijena maksimalna tačnost od 63.3% kombinujući Spektar i MFCC karakteristična obeležja. Ista tačnost (63.3%) nad GTZAN bazom postignuta je i kombinacijom Spektar, MFCC i LPCC karakterističnih obeležja.

English

Aim of the research is to compare several methods for creating features for neural network classification of music genres. Allocated samples of 5s in .wav format were used for comparing 4 different methods of digital audio signal processing: Fingerprint, Spectrum, MFCC and LPCC method. Two different databases were used - NATA and GTZAN. The first database, NATA, contains 6 different genres for classification, with 300 songs per genre. The second

database, GTZAN, contains 10 genres for classification, with 100 songs per genre. Fingerprint method is based on the Shazam application which detects 10 spectrum components (frequencies) biggest spectral density of a signal. Spectrum method combines features of the signal spectrum and the logarithm of the spectrum, while MFCC and LPCC methods use Mel scale and Linear Prediction scale, respectively. There were several predictions concerning the results of the research. First prediction was that the MFCC method would perform better than Spectrum method and Fingerprint method. Second prediction was that the Fingerprint method would be inferior to other methods. Third prediction was that MFCC and LPCC would have similar results over both databases. Fourth prediction was that the relative performance over both databases would be similar. Third and forth predictions were wrong, second prediction was correct, while first prediction was partially correct. The best result over the NATA database was 95.9%, obtained by a combination of MFCC and Spectrum methods, while over the GTZAN base the maximum accuracy of 63.3% was obtained by combining Spectrum and MFCC methods, as well as Spectrum, MFCC and LPCC methods.

Uvod

Automatska klasifikacija žanra pesme je deo mnogih kompleksnijih algoritama koje koriste kompanije kao što su Shazam, Spotify, Deezer i mnoge druge. Kod algoritma koji koristi Shazam aplikacija potrebno je odrediti žanr pesme kako bi se bliže odredila pesma. Spotify i Deezer koriste algoritme za preporučivanje muzike, kojima je neophodno da pesme budu označene žanrom kome pripadaju.

Problem klasifikacije žanra muzike je veoma izučavan u oblasti mašinskog učenja. Jedan pristup je objašnjen u referentnom radu [3], gde su korišćene konvolucione neuronske mreže kao klasifikatori. Mana konvolucionih neuronskih mreža je što su dosta vremenski zahtevnije za treniranje od klasičnih neuronskih mreža. Drugi pristup je korišćenje MFCC koeficijenata, kao što je prikazano u referentnom radu [10]. Spektar metoda je takođe korišćena u referentnom radu [1] za klasifikaciju žanra muzike (nad NATA bazom). Ideja ovog projekta bila je uporediti različite metode i njihove međusobne kombinacije radi nalaženja najbolje metode za klasifikaciju.

Digitalni audio signal je reprezentacija zvuka u obliku niza binarnih brojeva. Prema Furijeu, svaki signal se može razložiti kao zbir beskonačno mnogo sinusoida sa različitim amplitudama i fazama. U praksi, koristi se diskretna furijeova transformacija (DFT). DFT zahteva unos diskretnog signala, koji se dobija odabirom (semplovanjem) početnog signala, odnosno čitanjem vrednosti signala određenom frekvencijom - frekvencijom odabira.DFT pretvara funkciju iz vremenskog u frekvencijski domen. Izlaz DFT-a je spektar signala. Spektar signala je često korišćen kao karakteristično obeležje pri treniranju neuronskih mreža [1], jer sadrži informacije o prisutnim frekvencijama, a šumovi su manje izraženi. Logaritam baze 10 spektra signala dobija se primenom logaritamske funkcije na amplitude u spektru. Prednost logaritma ogleda se u isticanju razlika između malih vrednosti amplituda (koje se uglavnom nalaze na višim frekvencijama). U obradi signala je takođe široka primena filtera. Filteri su operatori koji imaju za cilj da delimično ili u potpunosti uklone neželjene, nasumične

komponente signala - šumove. Primeri filtera su high-pass filter (filter koji “propušta” samo više frekvencije), low-pass filter (propušta samo niže frekvencije), ...

Cilj istraživanja je bio izmeriti uspešnost različitih metoda za klasifikaciju žanra muzike. Odabrane su 4 metode: Fingerprint, Spektar, MFCC i LPCC metoda.

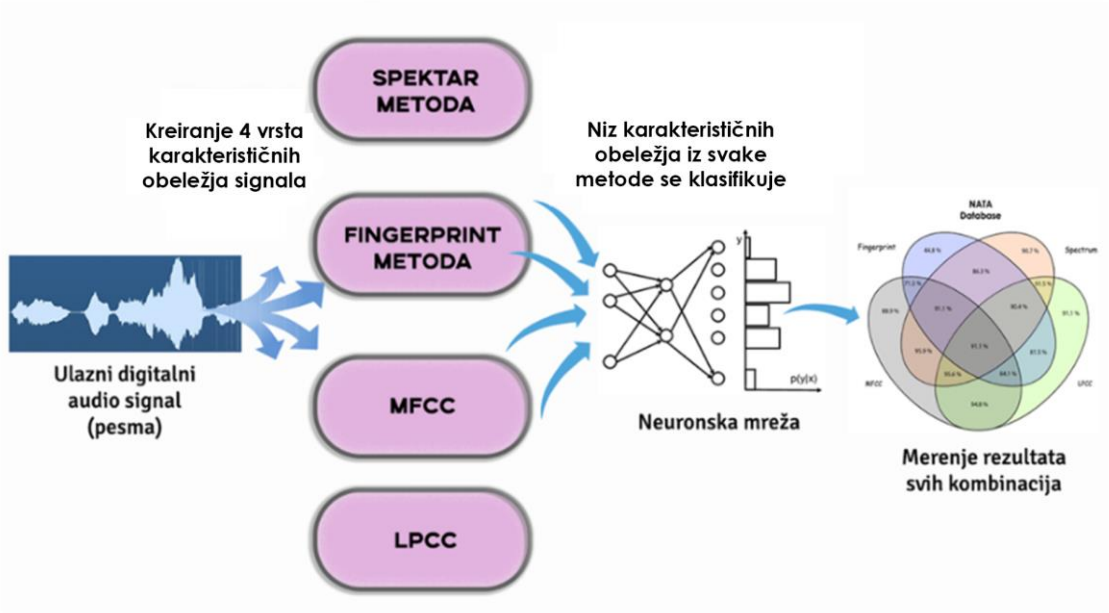
Kod MFCC i LPCC koeficijenata prva dva koraka su ista - signal se deli na manje delove dužine 25 milisekundi - prozore, na koje se potom primenjuje prozorska funkcija. MFCC koeficienti su potom dobijeni pomoću Mel filterbanke koja preciznije definiše razliku između frekvencija koji ljudi čuju i diskretne kosinusne transformacije koja signal pretvara u frekvencijski domen, dok se LPCC koeficijenti dobijaju pomoću kepstra i linearne predikcije signala.

Svaka od navedenih metoda ima za cilj da izdvoji neka svojstva signala, takozvana karakteristična obeležja. Ova karakteristična obeležja se prosleđuju kao ulaz neuralnoj mreži koja dalje na osnovu njih klasifikuje žanr muzike.

Pri treniranju je korišćena tensorflow biblioteka, kao i nprtool alatka u Matlab-u (Neuralna mreža za detekciju šablona odnosno obrasca).

Hipoteze

S obzirom da su MFCC koeficijenti veoma uspešni u prepoznavanju govora [5], može se pretpostaviti da će dati bolji rezultat od Spektar i fingerprint metode. Takođe, s obzirom da ne postoje rezultati za referenciranje fingerprint metode za klasifikaciju žanra muzike, pretpostavka je da će postići najniži rezultat. Prema drugom radu [12] gde su MFCC i LPCC metode postigle skoro isti rezultat u klasifikaciji nepravilnosti u izgovoru, pretpostavka je da će sličan rezultat dati i pri klasifikaciji žanra muzike. Sličan rezultat može se definisati tako da razlika u tačnosti dveju metoda ne bude veća od 5 procenata, nad bilo kojom od 2 baze. Poslednja hipoteza se odnosi na relativnu uspešnost metoda - pretpostavka je da će relativna uspešnost metoda biti slična nad obe baze. Relativna tačnost metode se definiše kao proizvod tačnosti metode nad nekom bazom sa brojem žanrova za klasifikaciju te baze (Ako baza sadrži m žanrova, a tačnost metode je p%, onda je relativna tačnost jednaka p * m). Relativna metoda daje sličnu relativnu tačnost ako se relativna tačnost nad obe baze razlikuje za manje od 50. Relativnu tačnost nije moguće savršeno definisati - mana ove konkretne definicije je što je maksimalna relativna tačnost bilo koje metode nad NATA bazom 600, te ako bi neka metoda postigla tačnost nad GTZAN bazom preko 65%, bilo bi nemoguće da relativna tačnost bude slična nad obe baze. U okviru istraživanja su ispitane sve mogućnosti kombinacija ulaznih karakterističnih obeležja za mrežu. Šema rada prikazana je na slici 1.1.



Slika 1.1: Šema rada istraživanja
Figure 1.1: Work scheme of the research

Baza podataka

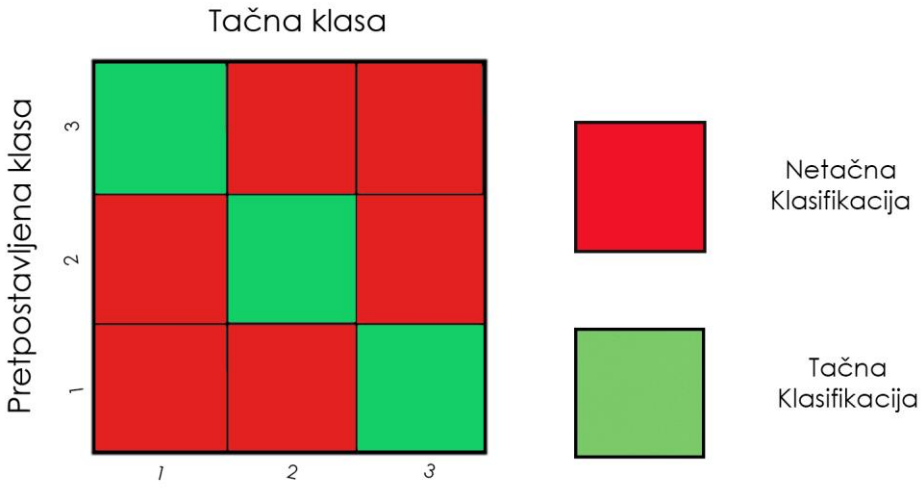
Predložene metode su poređene nad dve baze podataka. Prva baza podataka je GTZAN baza, sa 1000 audio fajlova u .wav formatu. Svaki audio fajl predstavlja isečak dužine 30 sekundi iz neke pesme. Svaka pesma (audio fajl) iz baze označena je jednim od 10 žanrova (bluz, klasična muzika, kantri, disko, hiphop, džez, metal, pop, rege i rok), tako da u okviru baze postoji tačno 100 pesama svakog od 10 žanrova. Druga baza je NATA baza, koja sadrži 1800 audio fajlova u .wav formatu. Svaki audio fajl je isečak dužine 5 sekundi iz neke pesme, označen jednim od 6 žanrova kome ta pesma pripada (klasična muzika, folk, haus, džez, R & B i rok). Postoji tačno 300 pesama u bazi svakog od 6 žanrova. Odlučeno je da baze budu pretprocesirane tako da svi audio fajlovi budu dužine 5 sekundi.

Iz GTZAN baze su izdvojeni isečci dužine 5 sekundi, od 13-te do 18-te sekunde, što je isto urađeno u radu [1]. NATA baza je ostala nepromenjena, jer su audio fajlovi odgovarajuće dužine.

Istraživanje

Rezultati svake od metoda za klasifikaciju žanrova prikazani su u vidu matrica konfuzije. Matrice konfuzije ilustruju uspešnost nekog klasifikatora, tako što se u kolonama nalaze stvarne vrednosti klasa, a redovi predstavljaju klase koje je predvideo klasifikator. Polja na glavnoj

dijagonali, dakle, pokazuju koliko puta je klasifikator tačno klasifikovao svaku od klasa. Prednost ovakvog predstavljanja rezultata pronalazi se ne samo u sveobuhvatnoj predstavi rezultata u jednoj slici, već i u mogućnosti da se vidi da li se neke klase često pogrešno klasifikuju više od ostalih (Slika 3.1).



Slika 3.1: Matrica konfuzije
Figure 3.1: Confusion matrix

Metod Rada

Fingerprint Metoda

Fingerprint metoda detektuje 10 komponenti spektra najveće spektralne gustine snage u digitalnom audio signalu trajanja od 5 sekundi. Kao referenca uzet je blog [4], u kom je izabrano 5 različitih frekvencionih opsega. U daljem radu je ispitivanjem utvrđeno da se odabirom 10 opsega postiže optimalan rezultat, gde veći prioritet imaju niske i srednje frekvencije jer je fundementalna karakteristika neke pesme sadržana u basu i bubnjevima. Sledeća lista prikazuje odabrane frekvencione opsege:

- 30 - 40 Hz
- 41 - 80 Hz
- 81 - 120 Hz
- 121 - 180 Hz
- 181 - 300 Hz
- 301 - 500 Hz
- 501 - 700 Hz

- 701 - 900 Hz
- 901 - 1200 Hz
- 1201 - 1500 Hz

Spektar signala se može dobiti diskretnom furijeovom transformacijom (DFT). DFT od ulaznog signala daje zavisnost spektralne gustine snage signala u zavisnosti od frekvencije. Kaže se da se DFT-om signal prebacuje u frekvencioni domen. Najbrži algoritmi za računanje DFT je brza Furijeova transformacija (FFT). FFT je bilo koji metod koji je definisan na sledeći način: Ako imamo signal x dužine N, onda je FFT definisana u obliku X[k]:

$$F(\omega) = \int_{-\infty}^{\infty} f(x) e^{-i\omega x}\, dx$$

Slika 4.1.1: Formula izračunavanja brze Furijeove transformacije (FFT)
Figure 4.1.1: Formula for finding Fast Fourier Transform (FFT)

Gde je **k = 0,1,...,N-1.**

Karakteristična obeležja Fingerprint metode se ekstraktuju tako što se iz signala izdvajaju spektralne gustine snage signala svake frekvencije koristeći FFT. Odatle se mogu detektovati frekvencije koje pripadaju najjačim spektralnim gustinama i time izdvojiti niz od 10 vrednosti koji se prosleđuje radi treniranja neuralne mreže.

Spektar Metoda

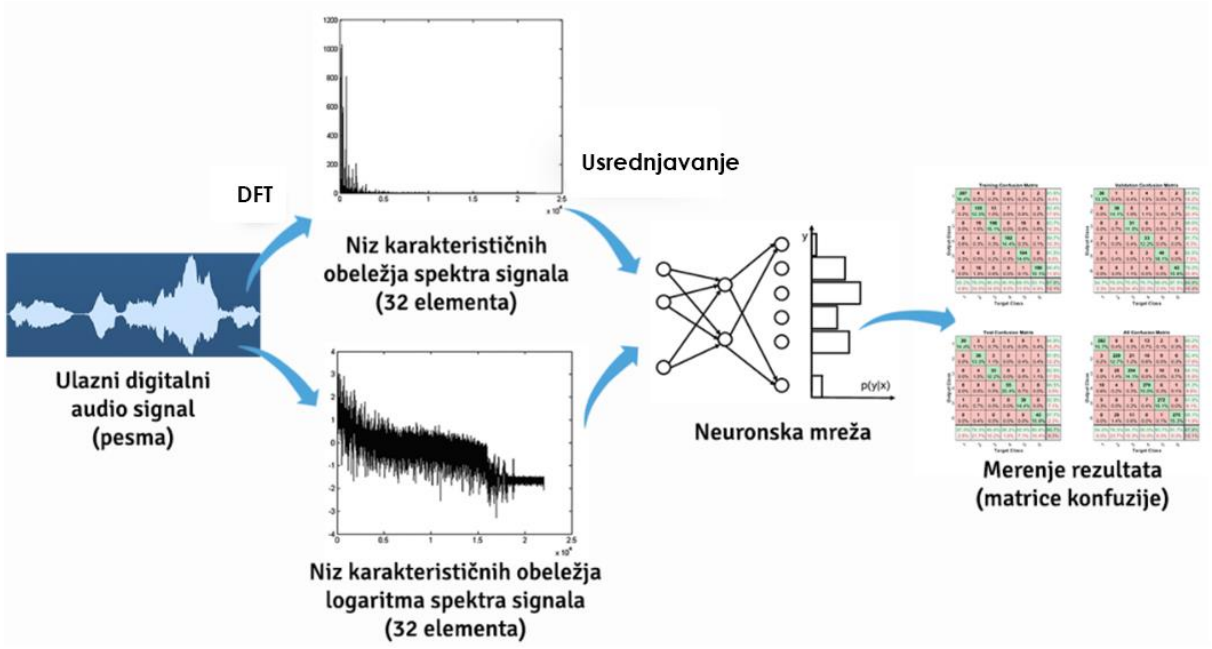
Spektar metoda koristi 3 vrste karakterističnih obeležja, takozvane setove karakterističnih obeležja. Ispitano je kombinovanje svih setova kako bi se utvrdilo koji setovi daju najbolji rezultat. Prvi set je spektar signala, koji se dobija primenom diskretne Furijeove transformacije (DFT). Spektar signala sadrži veliku količinu informacija o početnom signalu, a za razliku od početnog signala, sadrži mnogo manje nepotrebnih informacija (šumova), koje mogu da preokupiraju mrežu.
Korišćena je brza Furijeova transformacija. Celokupan spektar signala je podeljen na 32 segmenta i za svaki segment je uzeta prosečna vrednost spektralne gustine snage u tom segmentu (na ovaj način se značajno smanjuje veličina ulaza u mrežu). Dakle, prvo karakteristično obeležje signala za klasifikaciju je dobijeno spektrom signala koji sadrži niz od 32 decimalna broja, što predstavlja ulazne parametre za neuralnu mrežu.

Drugi set karakterističnih obeležja predstavlja logaritam spektra signala. Logaritamska funkcija pomaže isticanju razlika između nižih vrednosti amplituda, što se u praksi javlja kod visokih frekvencija. Slično kao kod spektra, vrednosti logaritma spektra su usrednjene na 32 segmenta i set karakterističnih obeležja takođe sadrži 32 ulazna parametra za mrežu.

Za treći set je izabran kepstar signala, što predstavlja inverznu Furijeovu transformaciju od logaritma spektra signala. Kepstar ima za cilj da razdvoji sporo promenljive komponente signala

i brzo oscilujuće kompomente [1]. Uzeto je prvih 20 vrednosti kepstra, kako su vrednosti kepstra veoma male.

Princip rada Spektar metode prikazan je na Slici 4.2.1:



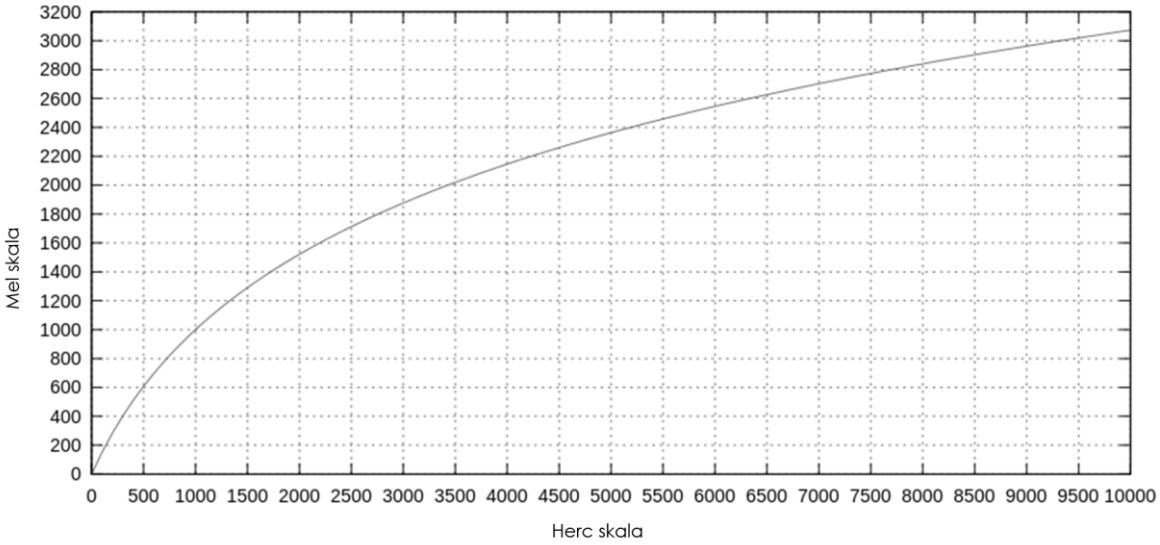
Slika 4.2.1: Princip rada spektar metode
Figure 4.2.1: Working principle of spectrum method

MFCC metoda

Pokazano je da ljudi razlike u frekvencijama zvuka ne čuju linearno [11]. Ispitivanjem je napravljena Mel skala koja je preciznije opisivala razliku između frekvencija koji ljudi čuju. Funkcija zavisnosti frekvencije na Mel skali od vrednosti frekvencije u hercima prikazana je formulom:

$$M(f) = 1125 * \ln(1 + f/700)$$

Gde je **f** frekvencija u hercima, a **M(f)** vrednost frekvencije na Mel skali, sa jedinicom mere Mel. Mel skala je logaritamska, s tim što je za vrednosti ispod 1kHz približno linearna.



Slika 4.3.1: Poređenje visine note mel i herc skale
Figure 4.3.1: Plots of pitch mel scale versus Hertz scale

MFCC metoda je motivisana ljudskim načinom percepcije zvuka i pokazuje najbolje rezultate pri prepoznavanju ljudskog govora [6], ali se pokazala uspešno i u klasifikaciji žanra muzike davajući najveći rezultat od 68,9% nad GTZAN bazom [8], kao i u prepoznavanju nepravilnosti u izgovoru reči sa tačnošću od 92.55% [12]. MFCC metoda se sastoji iz više koraka:

Prvi korak je primena filtera za isticanje viših frekvencija (pre-emphasis). Korišćenje ovakvog filtera pomaže balansiranju frekvencijskog spektra, kako viših frekvencija ima manje nego nižih. Ovaj filter ima ulogu high-pass filtera. Filter je definisan na sledeći način:

$$Y[n] = X[n] - b * X[n-1]$$

Za svako **n** iz vremenskog opsega, gde je **Y[n]** nova vrednost signala u **n**-tom trenutku, **X[n]** je početna vrednost, a **b** je koeficijent isticanja. Uzeta je vrednost b = 0.97 [9].

Sledeći korak je podela signala na frejmove od po 25 milisekundi - prozore, tako da je korak između dva susedna frejma 10ms (susedni frejmovi se preklapaju) [7]. Ovakav pristup nam pruža mogućnost da signal u okviru jednog frejma smatramo nepromenljivim, tj. da su celom dužinom frejma prisutne iste bazne frekvencije koje čine složeni signal. Potom se primeni prozorska funkcija na svaki od frejmova, koja ublažava takozvani aliasing efekat, tačnije nemogućnost razlikovanja susednih frejmova. Korišćena je Hamming prozorska funkcija. Hamming prozorska funkcija je fokusirana na smanjenje bočnih lukova, odnosno doprinosi blažim prelazima između susednih frejmova.

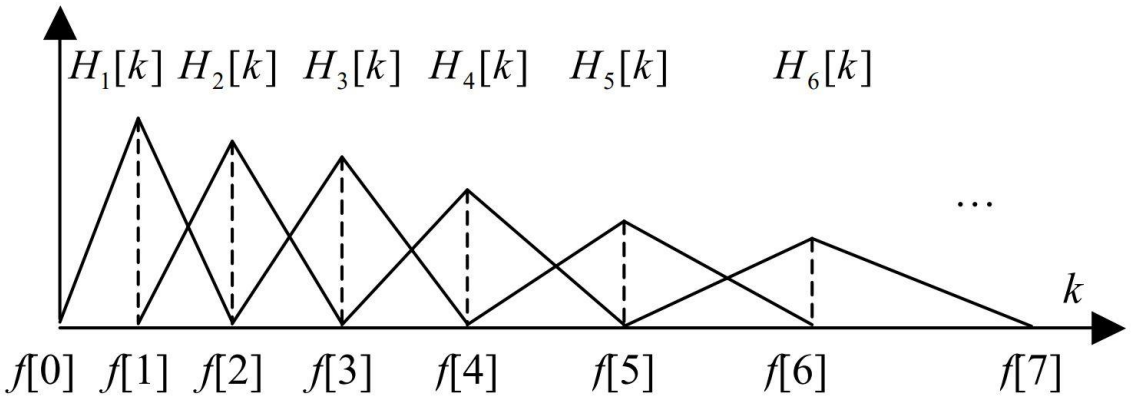
Nakon toga, nad svakim od frejmova se primeni DFT i dobija se spektar svakog frejma. Vrednosti amplituda su kvadrirane kako bi se dobio spektar snage.

Nakon toga su izabrane dve vrednosti frekvencije - 0 i Fs/2 (gde je Fs frekvencija odabiranja), čije su vrednosti preračunate u Mel skali. Između ove dve vrednosti, takođe u Mel skali, uniformno je raspoređeno 26 tačaka. Vrednosti frekvencija su potom pretvorene nazad u Herce.

Potom je nad spektrom svakog frejma primenjuje se trougaona filterbanka. Trougaona filterbanka sa **M** filtera definisana je kao:

$$H_m[k] = \begin{cases} 0, & k < f[m-1] \\ \frac{2(k-f[m-1])}{(f[m+1]-f[m-1])(f[m]-f[m-1])}, & f[m-1] \le k \le f[m] \\ \frac{2(f[m+1]-k)}{(f[m+1]-f[m-1])(f[m+1]-f[m])}, & f[m] \le k \le f[m+1] \\ 0, & k > f[m+1] \end{cases}$$

Gde je **m** redni broj filtera (**m = 1,2,...,M**), **f[m]** je **m**-ta uniformno raspoređena tačka (između 0 i Fs/2). Trougaona filterbanka ilustrovana je na slici 4.3.1.



Slika 4.3.2: Trougaona filterbanka [2]
Figure 4.3.2: Triangular filterbank [2]

Cilj filterbanke je da spektar signala svakog frejma izdeli u **M** različitih ,,korpi", gde je svaka korpa centrirana oko jedne od uniformno raspoređenih tačaka. Vrednost **m**-te korpe je izračunata kao težinska sredina frekvencija između **f[m-1]** i **f[m+1]**, tako da je težina središnje tačke (**f[m]**) najveća i linearno opada ka krajevima intervala, stvarajući oblik trougla.

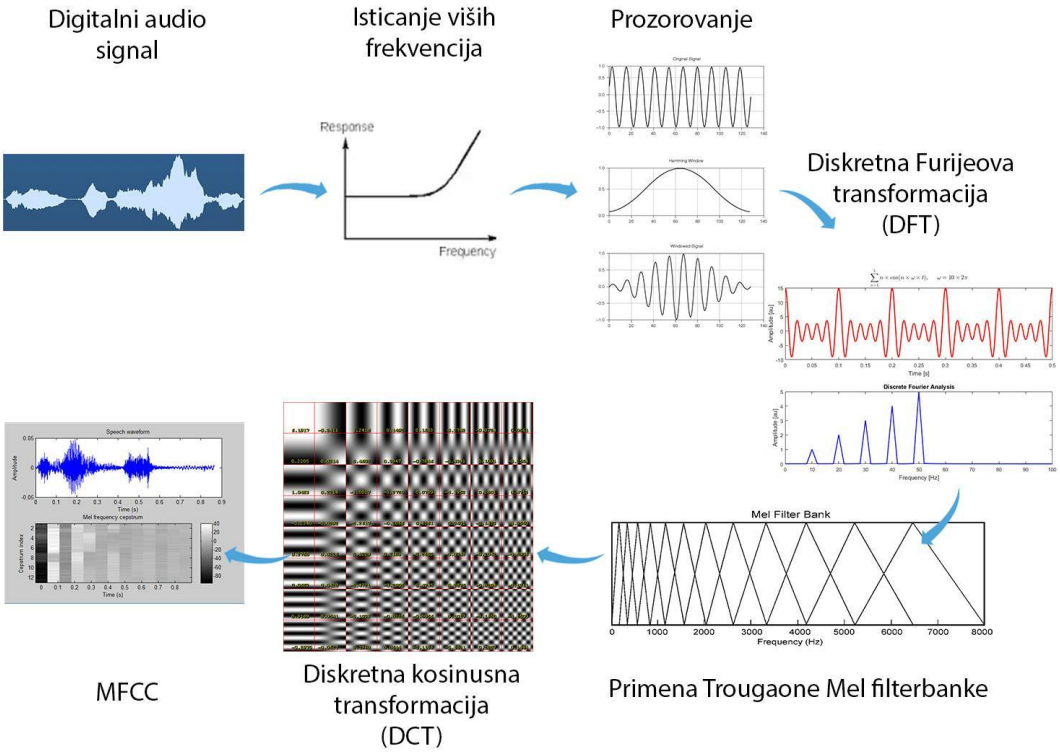
Filterbanka je u stvari niz filtera propusnika učestanosti i izdvaja energije spektra u svakoj datih **M** tačaka. Broj filtera uglavnom varira između 20 i 30. korišćeno je 26 filtera (**M = 26**).

Nakon toga vrednosti spektra su logaritmovane. Ovo je motivisano ljudskim načinom percepcije zvuka - ljudi zvuk približno Mel skali, koja je logaritamska.

Poslednji korak je primena diskretne kosinusne transformacije (DCT). S obzirom da se frejmovi preklapaju, spektri susednih frejmova su uzajamno povezani. DCT pomaže dekorelaciji podataka i samim tim smanjenju redudanse podataka. DCT je prikazana sledećom formulom:

$$c[n] = \Sigma_{m=0}^{M-1} L[m] \cos(\pi n(m + 1/2)/M) \quad 0 \leq n < M$$

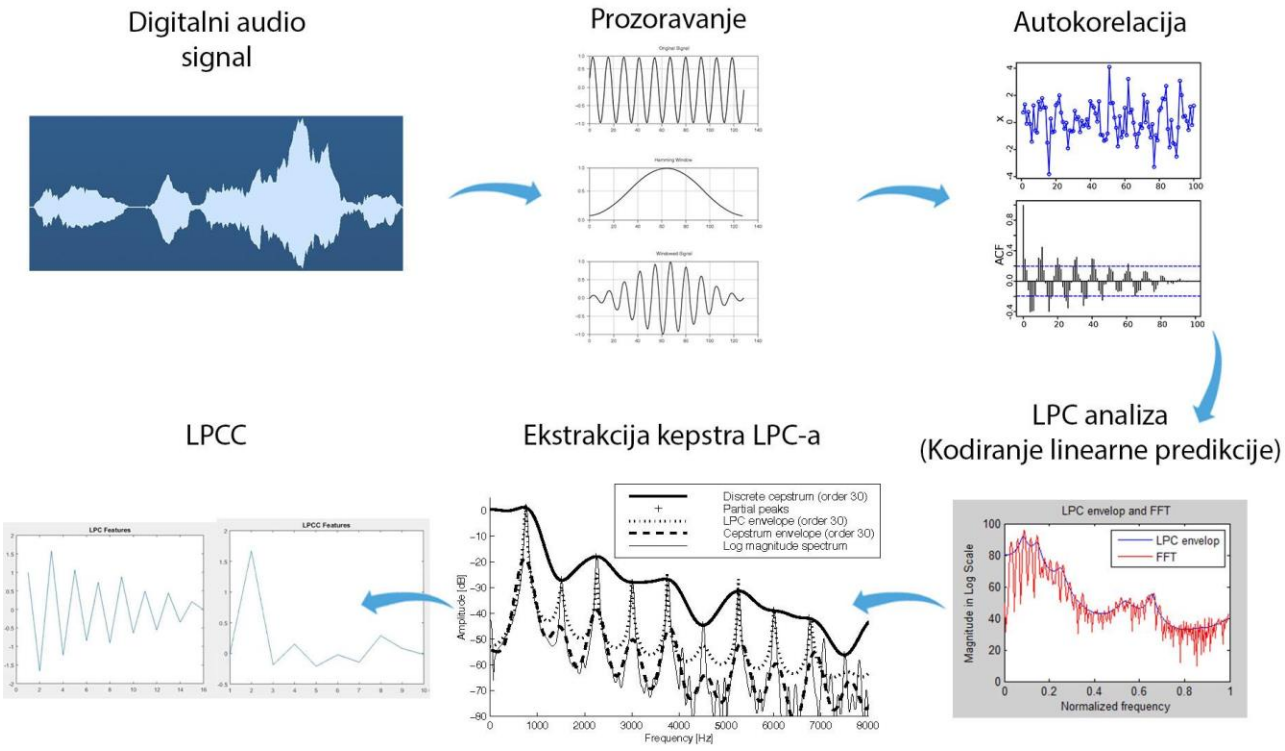
Gde je **L[m]** logaritam **m**-tog izlaza filterbanke nekog frejma.
Ovim putem je dobijeno 26 koeficijenata za svaki frejm, pa se koeficijenti celokupnog signala dobijaju sumiranjem vrednosti svakog od koeficijenta po frejmovima. U referentnom radu [16] je naznačeno da se u prepoznavanju ljudskog uglavnom uzima prvih 13 koeficijenata, međutim u istraživanju se pokazalo da korišćenje 20 koeficijenata daje najbolji rezultat u prepoznavanju žanra muzike.
Uzeto je prvih 20 koeficijenata , što ujedno predstavlja i veličinu ulaza u mrežu ove metode. Ceo proces dobijanja MFCC koeficijenata prikazan je na slici 4.3.2.



Slika 4.3.3: Proces ekstraktovanja MFCC koeficijenata
Figure 4.3.3: Process of extracting MFCC coefficients

LPCC metoda

Još jedna metoda koja je zastupljena u obradi signala jeste LPCC metoda koja predstavlja koeficijente kepstra koristeći linearnu predikciju. Prva 3 koraka u izdvajanju LPCC koeficijenata su ista kao kod MFCC metode - primenjuje se pre-emphasis filter, a potom se signal deli na frejmove i primenjuje se Hamming prozorska funkcija.
Posle izvršene autokorelacije signala, dobijaju se LPC koeficijenti, iz kojih se dobijaju LPCC koeficijenti tako što se primeni kepstar na dobijenim koeficijentima.



Slika 4.3.4. Proces ekstraktovanja LPCC karakterističnih obeležja
Figure 4.3.4: Process of LPCC feature extraction

Ovom metodom dobijen je veliki broj koeficijenata, pošto se radi o signalu trajanja od 5 sekundi. U referentnom radu [12] je uzeto prvih 21 koeficijenata, međutim ispitivanjem je dobijeno da 30 koeficijenata daju optimalan rezultat. Dakle, uzeto je prvih 30 koeficijenata za prosleđivanje neuralnoj mreži zarad ovog istraživanja.

Mreža

U istraživanju su se upoređivale ne samo metode nego neuralne mreže koje su upoređivale te metode, te je korišćena Keras iz Tensorflow biblioteke, ali i Matlab-ov nprtool. Korišćena je mreža sa jednim skrivenim slojem čiji je broj neurona zavisio od veličine niza prosleđenog kao ulaz. Od broja nizova koji su se koristili za klasifikaciju, 70% je korišćeno za

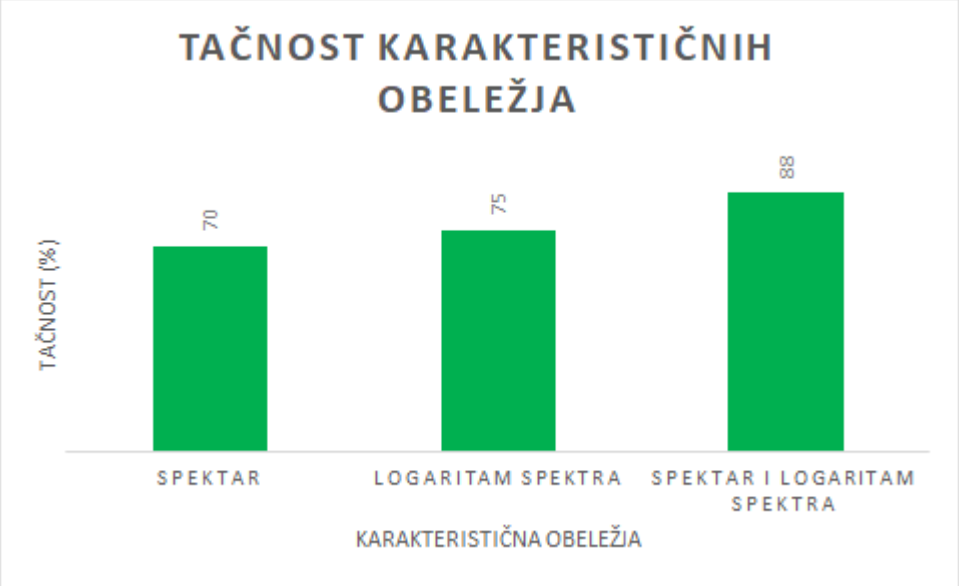
trening mreže, 15% za validaciju i 15% za test. Kao izlaz korišćena je softmax funkcija koja je u ovom slučaju predstavljala verovatnoću svakog žanra kao ciljnog, te je dobijen niz od 6 vrednosti pri radu sa NATA bazom, i niz od 10 vrednosti pri radu sa GTZAN bazom.

Rezultati i diskusija

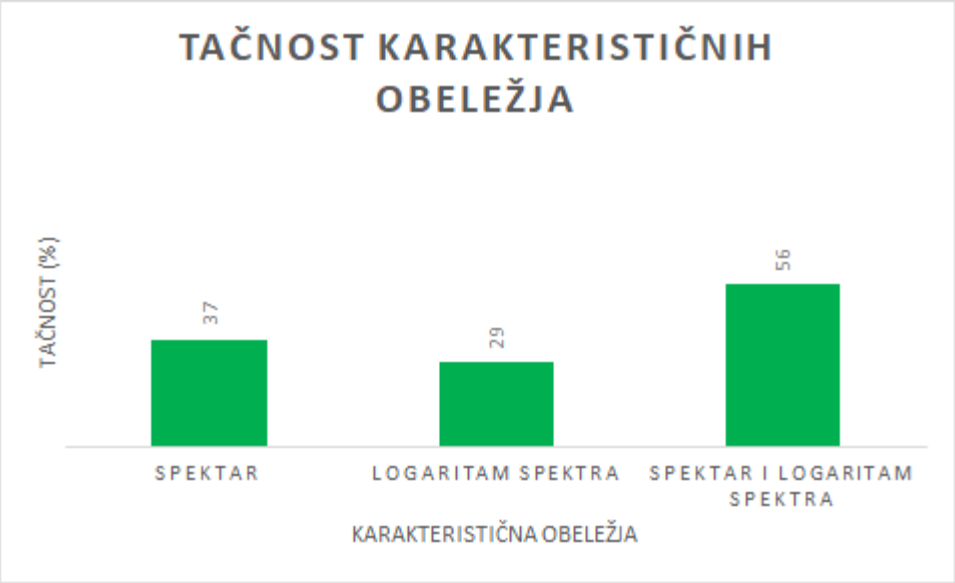
Merena je tačnost klasifikacije svih metoda posebno, kao i svaka kombinacija početnih metoda. Pod kombinovanjem dve ili više metoda smatra se da su ekstraktovani koeficijenti svih metoda i zajedno uneti u mrežu, gde je veličina ulaznog sloja mreže jednaka zbiru broja koeficijenata svih metoda posebno. Uspešnost neke metode posmatra se nad podacima za testiranje, ali su ispitivani rezultati i za validaciju i trening.

Rezultati - Spektar metoda

Spektar metoda ima 3 seta karakterističnih obeležja: spektar signala, logaritam spektra i kepstar signala.
Kepstar karakteristična obeležja nisu ispitivana kako je njihov unos u mrežu davao veoma nisku tačnost. Korišćena je neuronska mreža sa jednim skrivenim slojem koji sadrži 25 neurona i sigmoid aktivacionom funkcijom, koja je dala najbolji rezultat.
Rezultati Spektar metode mogu se videti na slici 5.1.1 za NATA bazu, tj. Na slici 5.1.2 za GTZAN bazu.



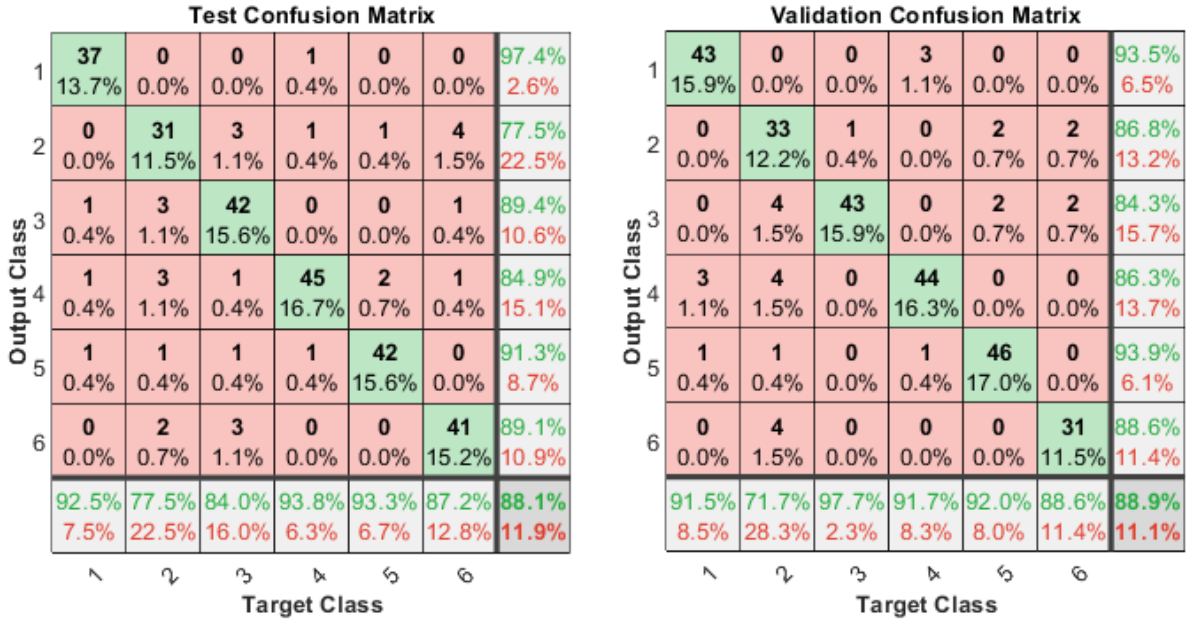
Slika 5.1.1: Tačnost Spektar metode nad NATA bazom
Figure 5.1.1: Accuracy of Spectrum method, when trained on NATA database



Slika 5.1.2: Tačnost Spektar metode nad GTZAN bazom
Figure 5.1.2: Accuracy of Spectrum method, when trained on GTZAN database

U poređenju sa referentnim radom nad NATA bazom, niža je tačnost za logaritam spektra (ref. 81%), a postignut je bolji rezultat kod spektra (ref. 66.5%), kao i kod kombinacije oba (ref. 51.8%). Manja odstupanja u tačnosti mogu se objasniti korišćenjem drugačijih alatki za treniranje mreže. Za GTZAN bazu ne postoje referentni rezultati.
Pošto je kombinacija spektra i logaritma spektra pokazala najbolji rezultat u okviru Spektar metode nad obe baze, samo ovaj set karakterističnih obeležja je korišćen pri kombinacijama sa drugim metodama.

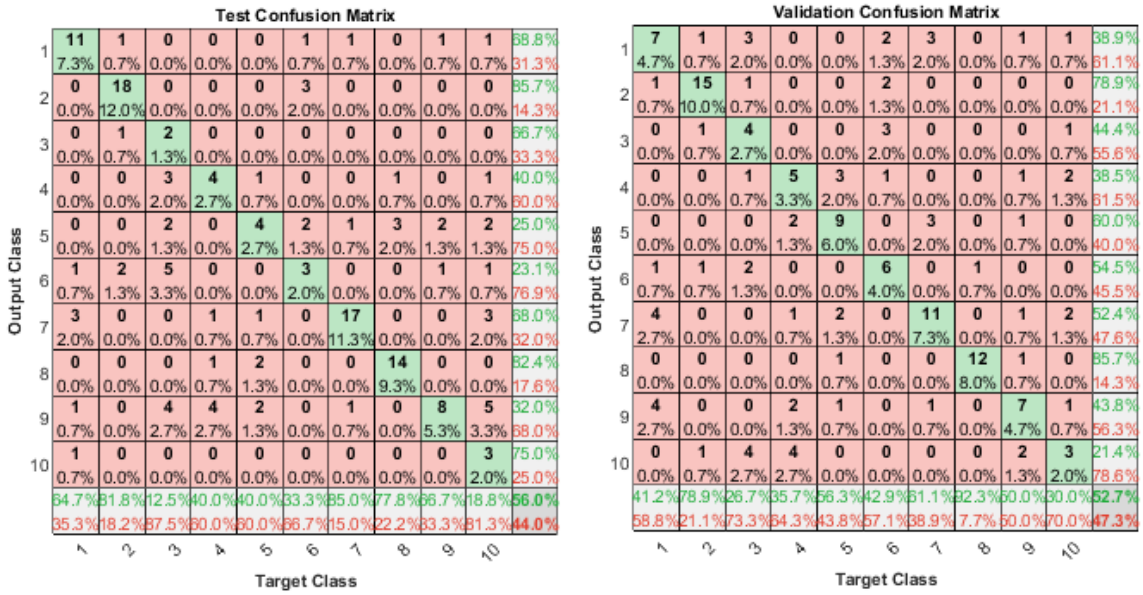
Nad NATA bazom na podacima za testiranje postignuta je tačnost od 88.1%, dok je na podacima za validaciju tačnost 88.9%. Na podacima za testiranje najniža tačnost postignuta je prilikom klasifikacije folk, haus i rok žanrova, dok su na podacima za validaciju u pitanju samo haus i rok. Na podacima za testiranje moguće je uočiti da se folk učestalo pogrešno klasifikovao kao haus, a takođe je haus učestalo pogrešno klasifikovan kao folk. Ovakvo mešanje zovemo simetrično mešanje. Folk i haus nemaju većih muzičkih sličnosti. Takođe je došlo do simetričnog mešanja folka sa rokom, između kojih takođe, nema većih muzičkih sličnosti. Na podacima za validaciju je takođe došlo do simetričnog mešanja folka i roka.
Matrice konfuzije za testiranje i validaciju su prikazane na slici 5.1.3.



Slika 5.1.3: Matrice konfuzije na podacima za testiranje i validaciju NATA baze
Figure 5.1.3: Confusion matrices on NATA database validation datasets

Nad GTZAN bazom na podacima za testiranje postignuta je tačnost od 56.0%, odnosno 52.7% na podacima za validaciju. Na podacima za testiranje najčešće mešane klase su kantri, disko, hiphop i rok, dok su na podacima za validaciju to klase bluz, kantri, disko, džez i rok. Na testiranju dolazi do simetričnog mešanja bluza i metala, ali oni nemaju većih muzičkih sličnosti. Na validaciji se javlja simetrično mešanje hiphopa i metala, kao i kantri muzike i džeza. Nijedan od ovih parova nema značajnijih muzičkih sličnosti.

Matrice konfuzije za testiranje i validaciju prikazane su na slici 5.1.4.

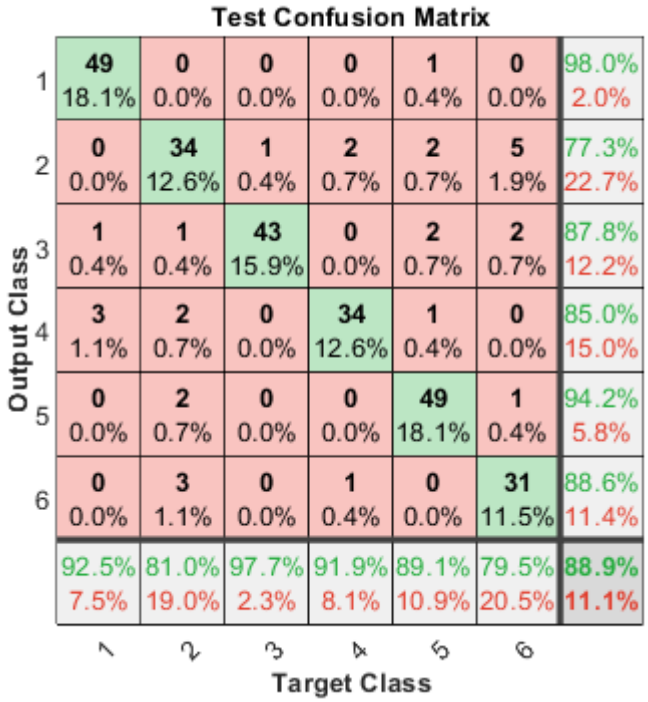


Slika 5.1.4: Matrice konfuzije na podacima za testiranje i validaciju GTZAN baze
Figure 5.1.4: Confusion matrices on GTZAN database testing and validation datasets

Rezultati - MFCC metoda

U okviru MFCC metode, najbolji rezultat je postignut kada neuronska mreža sadrži 50 neurona u skrivenom sloju, sa sigmoid aktivacionom funkcijom. Tačnost na podacima za testiranje nad NATA bazom je 88.9%, kao i na podacima za validaciju. Nad GTZAN bazom na podacima za testiranje postignuta je tačnost od 60%, dok je nad podacima za validaciju tačnost 51.3%.

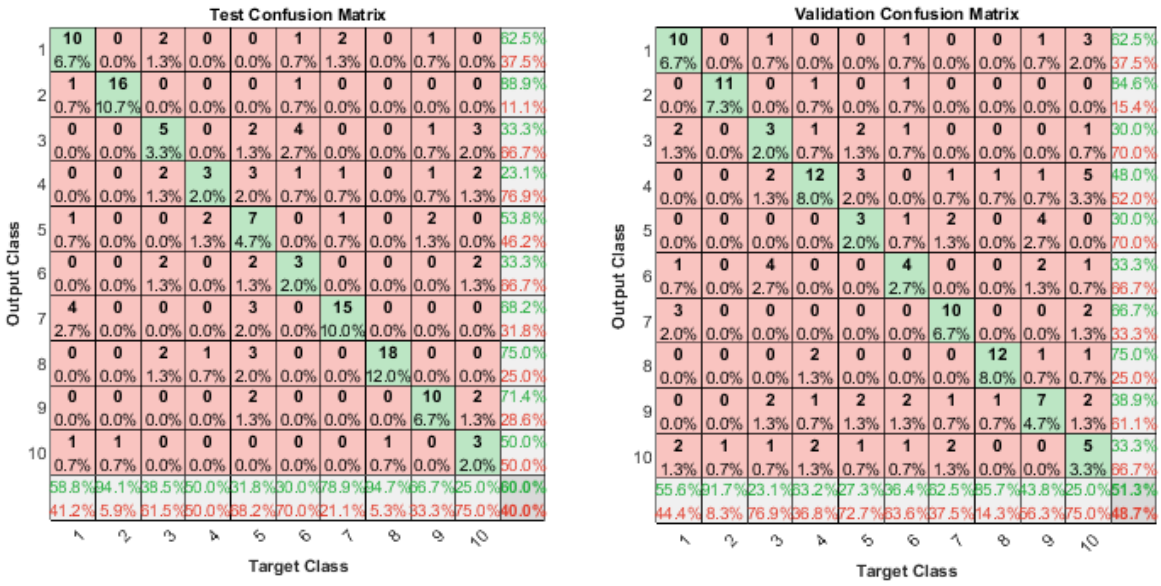
Nad NATA bazom na podacima za testiranje najviše su pogrešno klasifikovani folk, haus i džez. Najviše je džez pogrešno klasifikovan kao klasična muzika, a takođe se može primetiti da je često klasična muzika pogrešno klasifikovana kao džez. Ovo mešanje ima muzičkog značaja kako džez ima elemente klasične muzike. Ovakva opažanja važe i na podacima za validaciju.
Na slici 5.2.1 prikazana je matrica konfuzije za testiranje nad NATA bazom.



Slika 5.2.1: Matrica konfuzije na podacima za testiranje NATA baze
Figure 5.2.1: Confusion matrix on NATA database testing dataset

Nad GTZAN bazom na podacima za testiranje sa najmanjom tačnošću su klasifikovani hiphop, džez i rok, dok su na podacima za validaciju to kantri, hiphop, džez, rege i rok. Najčešće mešanje su bluz koji se pogrešno klasifikuje kao metal, kao i džez koji se pogrešno klasifikuje kao kantri. Ova mešanja su simetrična. Mešanje metala i bluza nema većeg muzičkog značaja, kao ni mešanje džeza i kantrija.

Na slici 5.2.2 su prikazane matrice konfuzije za testiranje i validaciju NATA baze.

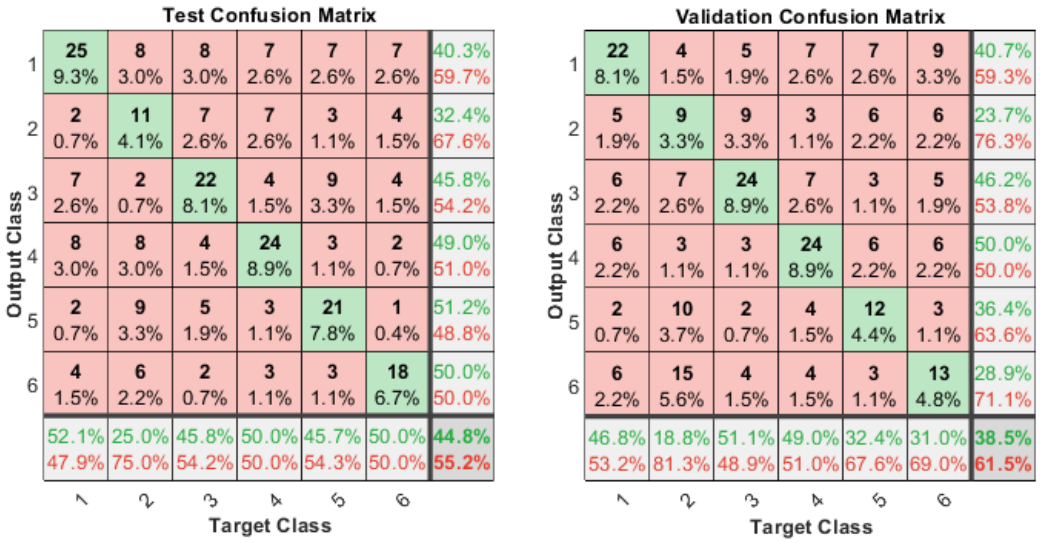


Slika 5.2.2: Matrice konfuzije na podacima za testiranje i validaciju GTZAN baze
Figure 5.2.2: Confusion matrices on GTZAN database testing and validation datasets

Rezultati - Fingerprint metoda

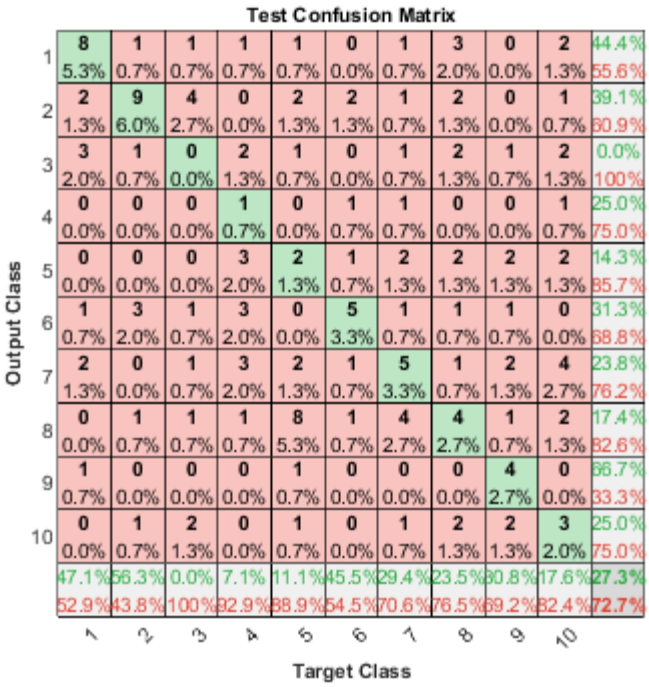
Fingerprint metoda je postigla najgori rezultat od svih metoda. Postignut je najbolji rezultat sa 50 skrivenih neurona i sigmoid aktivacionom funkcijom u skrivenom sloju. Nad NATA bazom postignuta je tačnost od 44.8% na testiranju, kao i 38.5% na validaciji. Nad GTZAN bazi postignuta je tačnost od 27.3% na testiranju, a 22.0% na validaciji.

Nad NATA bazom na podacima za testiranje postignuta je tačnost od 44.8% na podacima za testiranje, a tačnost od 38.5% na podacima za validaciju. Na podacima za testiranje ubedljivo najniža tačnost je postignuta pri klasifikaciji folk muzike, dok je nad podacima za validaciju najviše mešanja došlo kod folk, RnB i roka. Najčešće mešani parovi su folk koji se klasifikuje kao RnB, kao i RnB koji se pogrešno klasifikuje kao haus. Mešanje folka i RnB žanra nije simetrično, pa će se smatrati statističkom greškom. Mešanje haus i RnB žanra jeste simetrično, a muzički postoji sličnost u vidu učestalog pojavljivanja bubnjeva i basa. Na podacima za validaciju javlja se simetrično mešanje folka i RnB žanra, ali nema muzičkih sličnosti. Takođe se javlja veoma često mešanje folka sa rokom, koje je takođe simetrično. Muzička sličnost ova dva žanra je zanemarljiva.
Matrice konfuzije na podacima za testiranje i validaciju prikazane su na slici 5.3.1.



Slika 5.3.1: Matrice konfuzije na podacima za testiranje i validaciju NATA baze
Figure 5.3.1: Confusion matrices on NATA database testing and validation datasets

Nad GTZAN bazom na podacima za testiranje postignuta je tačnost od 27.3%, dok je na podacima za validaciju tačnost 22.0%. Na podacima za testiranje kantri žanr nijednom nije tačno klasifikovan, a jako malu tačnost postigli su i disko, hiphop i rok. S obzirom da je ovaj metod tek nešto bolji od nasumičnog odabira, dalja analiza je nepotrebna.
Matrica konfuzije za testiranje je prikazana na slici 5.3.2.

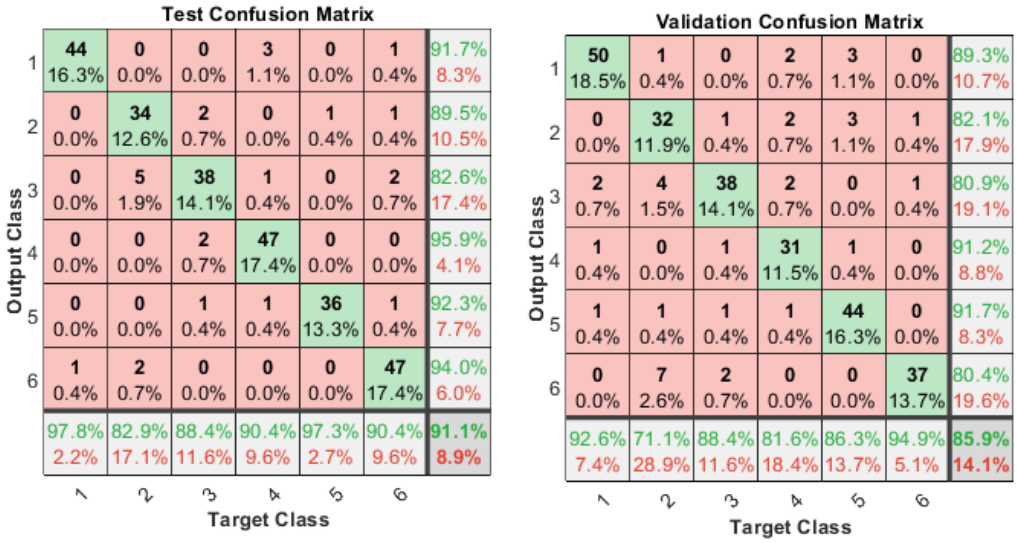


Slika 5.3.2: Matrica konfuzije na podacima za testiranje GTZAN baze
Figure 5.3.2: Confusion matrix on GTZAN database testing dataset

Rezultati - LPCC metoda

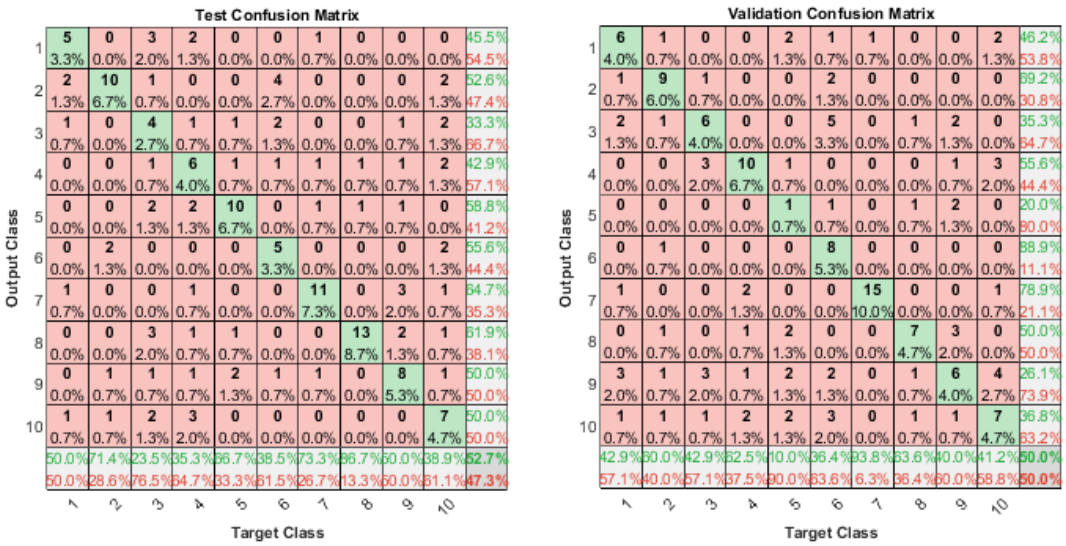
LPCC metoda je postigla najbolje rezultate nad NATA bazom (među zasebnim metodama), sa tačnošću na testu od 91.1%, dok je nad GTZAN bazom dobijeno 52.7% tačnosti. Korišćeno je 100 skrivenih neurona sa sigmoid aktivacionom funkcijom u skrivenom sloju.

Nad NATA bazom na podacima za testiranje postignuta je tačnost od 91.1%, a na podacima za validaciju 85.9%. Na podacima za testiranje najniža tačnost je postignuta na folk i haus žanrovima, dok je na podacima za validaciju to slučaj na folk, džez i RnB žanrovima. Na podacima za testiranje najviše je folk mešan sa hausom, a takođe je haus mešan sa folkom. Ova dva žanra nemaju veće muzičke sličnosti. Na podacima za validaciju ubedljivo je najviše folk mešan sa rokom, ali s obzirom da mešanje nije simetrično, smatra se za statističku grešku.
 Na slici 5.4.1 prikazane su matrice konfuzije na podacima za testiranje i validaciju.



Slika 5.4.1: Matrice konfuzije na podacima za testiranje i validaciju NATA baze
Figure 5.4.1: Confusion matrices on NATA database testing and validation datasets

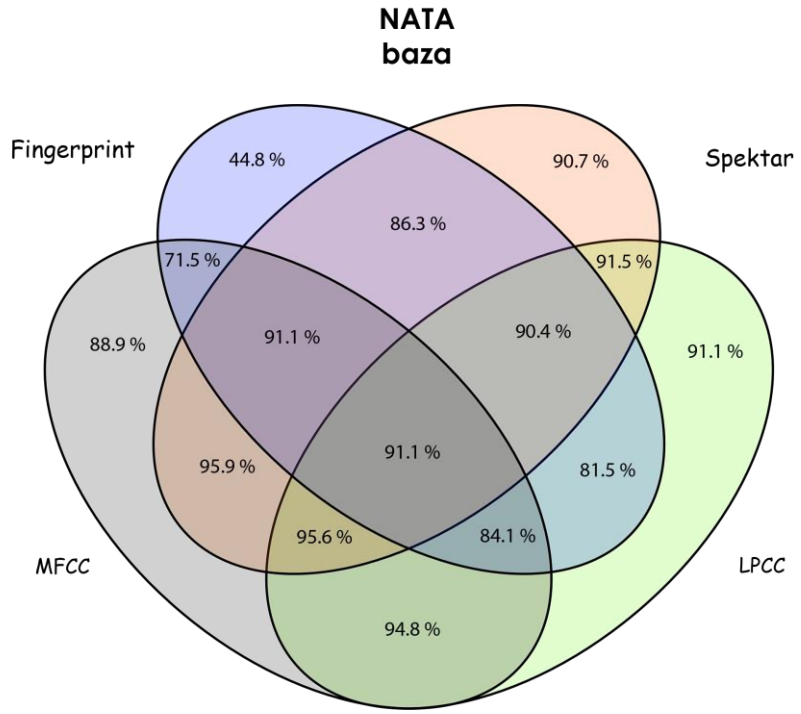
Nad GTZAN bazom na podacima za testiranje postignuta je tačnost od 52.7% na podacima za testiranje i tačnost od 50.0% na podacima za validaciju. Na podacima za testiranje najviše su mešani kantri, disko, rege i rok. Na podacima za validaciju najviše su mešani bluz, kantri, hiphop, džez i rok, što se značajno razlikuje od podataka za testiranje. Na testiranju najviše je džez mešan sa klasičnom muzikom, koji imaju sličnosti kao što je već rečeno. Na podacima za validaciju najviše je džez pogrešno klasifikovan kao kantri, ali s obzirom da kantri nijednom nije klasifikovan kao džez ovo mešanje se posmatra kao statistička greška.
Matrice konfuzije za testiranje i validaciju prikazane su na slici 5.4.2.



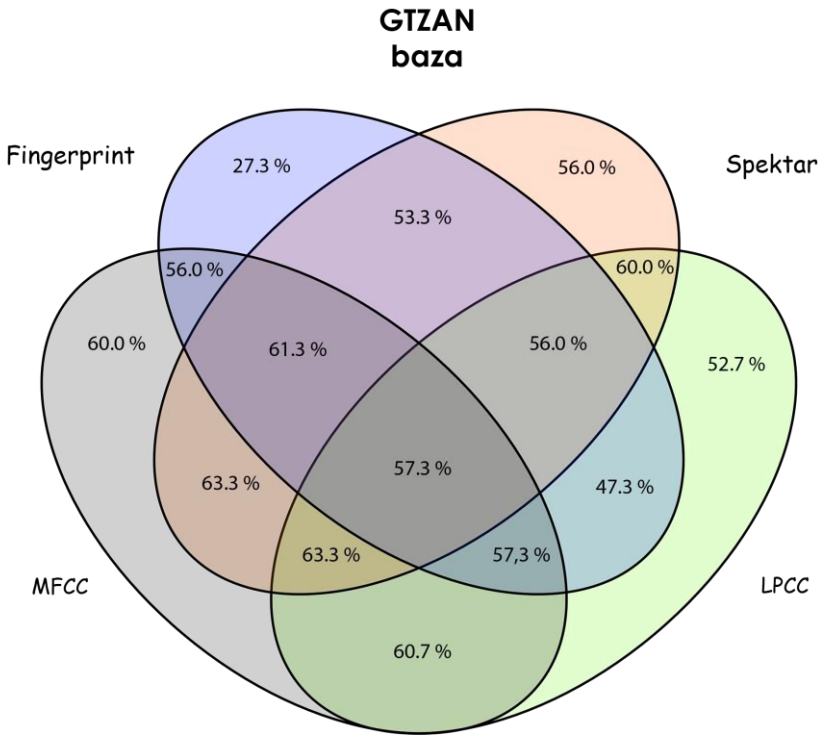
Slika 5.4.2: Matrice konfuzije na podacima za testiranje i validaciju GTZAN baze
Figure 5.4.2: Confusion matrices on GTZAN database testing and validation datasets

Rezultati - kombinovane metode

Kombinovanje metoda u nekim slučajevima je pokazalo bolju tačnost na podacima za testiranje nad obe baze.
Na podacima za testiranje, spoj MFCC i Spektar metode je dalo najveću tačnost nad NATA bazom (95.9%), dok spoj MFCC i Spektar metode, ali i spoj Spektar, MFCC i LPCC metode daju najveću tačnost nad GTZAN bazom (63.3%).
Rezultat kombinovanja 4 metode, na podacima za testiranje obe baze, prikazan je na dijagramima 5.5.1 i 5.5.2:



Slika 5.5.1. Kombinovani rezultati nad NATA bazom podataka
Figure 5.5.1. Combined results, when trained on NATA dataset



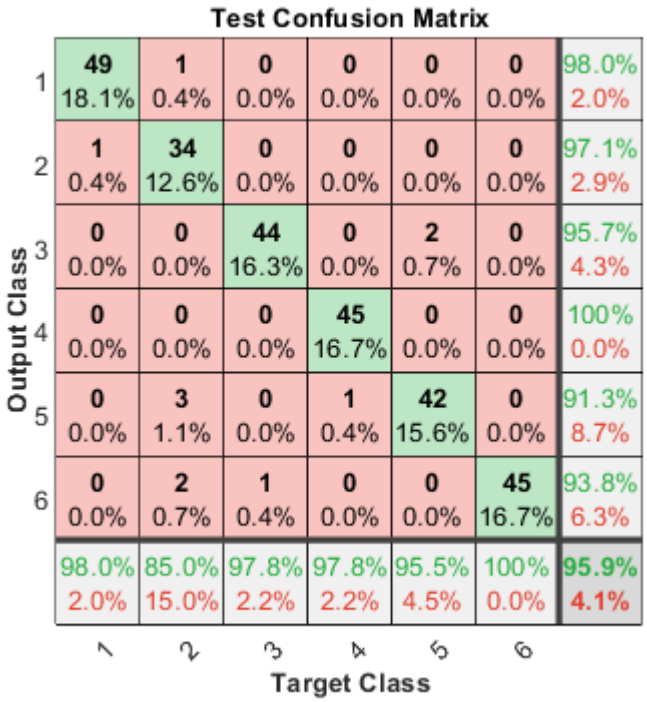
Slika 5.5.2. Kombinovani rezultati nad GTZAN bazom podataka
Figure 5.5.2. Combined results, when trained on GTZAN dataset

Sa dijagrama je moguće uočiti da kombinovanje Fingerprint metode sa bilo kojom drugom metodom ili kombinacijom metoda striktno smanjuje tačnost klasifikacije, što važi nad obe baze. U narednim odeljcima biće diskutovano o svim kombinacijama Spektar, MFCC i LPCC metoda.

Rezultati - Spektar + MFCC

Kombinacija Spektar i MFCC metode daje tačnost od 95.9% nad NATA bazom, odnosno 63.3% nad GTZAN bazom. Nad NATA bazom daje najbolji rezultat od svih metoda, dok nad GTZAN bazom ima isti rezultat kao i kombinacija MFCC, LPCC i Spektar metoda (što je ujedno i najbolji postignuti rezultat nad ovom bazom). Najbolji rezultat postignut je korišćenjem 100 neurona u skrivenom sloju sa sigmoid aktivacionom funkcijom.

Nad NATA bazom na podacima za testiranje postignuta je tačnost od 95.9%, a na podacima za validaciju 91.1%. Na podacima za testiranje, kao i na podacima za validaciju, najgori rezultat je postignut pri klasifikaciji folk žanra. Prisutna su značajnija mešanja folka sa RnB žanrom, folka sa rokom, kao i RnB sa hausom. S obzirom da nijedno od navedenih mešanja nije simetrično, smatraju se za statističke greške.
Na slici 5.6.1 prikazana je matrica konfuzije na podacima za testiranje NATA baze.

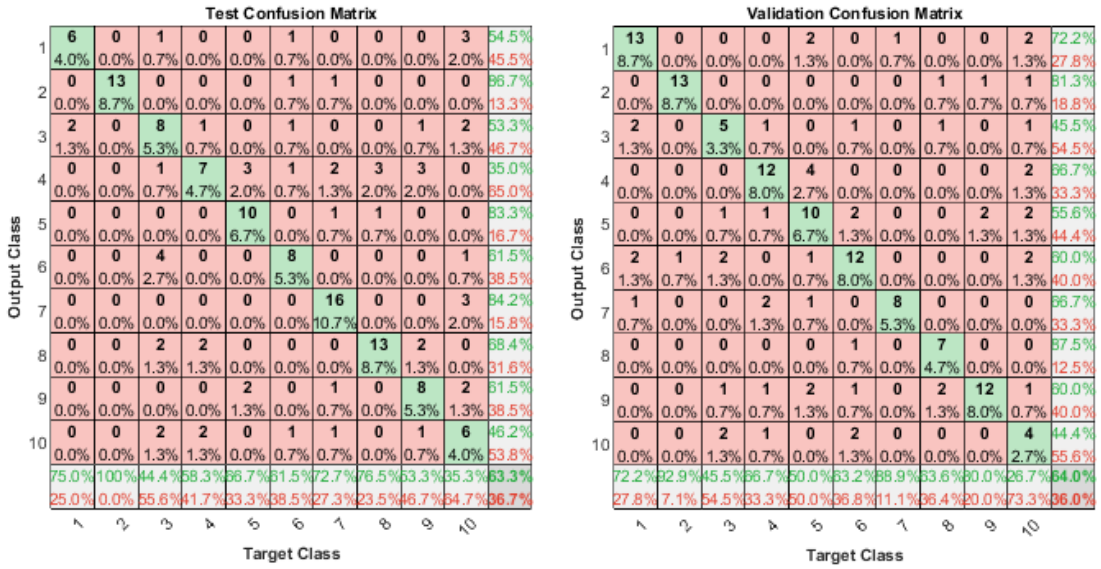


Slika 5.6.1: Matrica konfuzije na podacima za testiranje NATA baze
Figure 5.6.1: Confusion matrix on NATA database testing dataset

Na podacima za testiranje postignuta je tačnost od 63.3%, dok je na podacima za validaciju postignuta tačnost od 64.0%. Na podacima za testiranje se može uočiti da klasična muzika nije

nijednom pogrešno klasifikovana, dok su najgore klasifikovani kantri, rege i rok. Na podacima za validaciju najniža tačnost postignuta je pri klasifikaciji kantri, hiphop i rok pesama. Na podacima za testiranje najčešća simetrična mešanja su kantri i džez, kao i pop i disko. Kantri i džez nemaju većih sličnosti, dok pop i disko imaju značajnijih preklapanja u vidu pesama koje sadrže elemente oba žanra.

Na slici 5.6.2 prikazane su matrice konfuzije na podacima za testiranje i validaciju GTZAN baze.



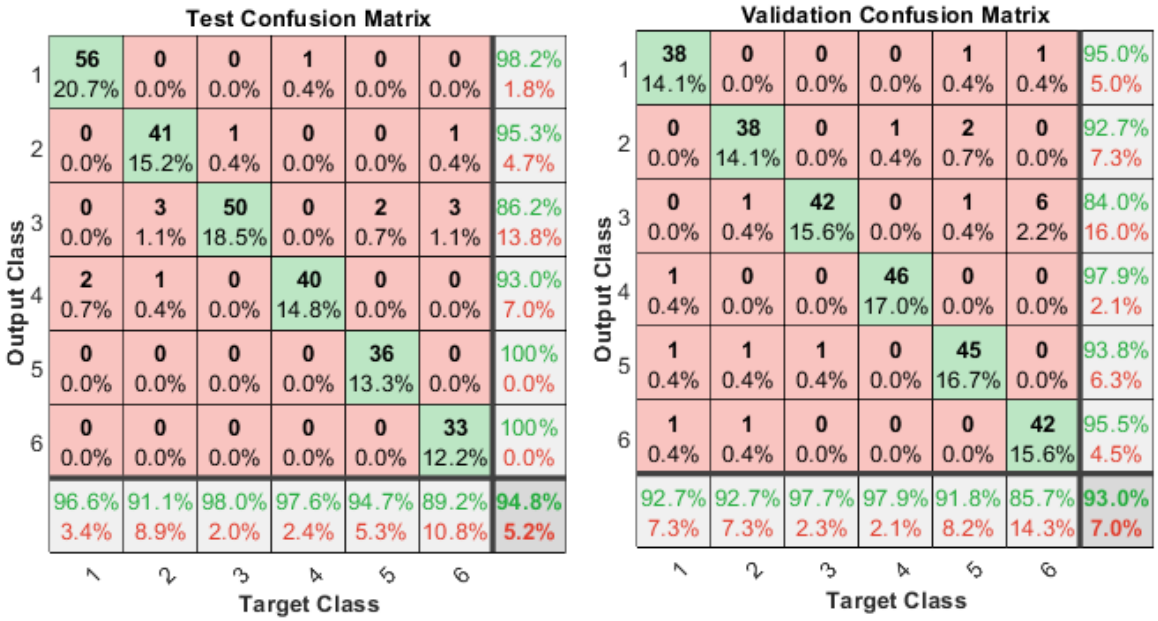
Slika 5.6.2: Matrice konfuzije na podacima za testiranje i validaciju GTZAN baze
Figure 5.6.2: Confusion matrices on GTZAN database testing and validation datasets

Rezultati - MFCC + LPCC

Kombinacija MFCC i LPCC karakterističnih obeležja postigla je tačnost od 94.8% nad NATA bazom, odnosno 60.7% nad GTZAN bazom. Najbolji rezultat postignut je sa neuronskom mrežom koja sadrži 100 neurona u skrivenom sloju sa sigmoid aktivacionom funkcijom.

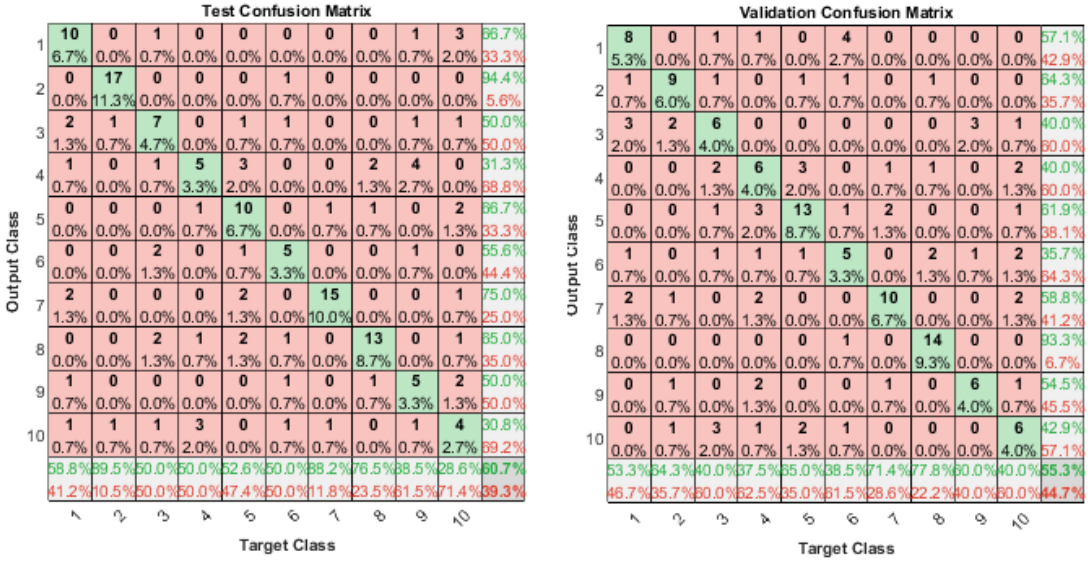
Na podacima za testiranje postignuta je tačnost od 94.8%, dok je na podacima za validaciju tačnost 93.0%. Na podacima za testiranje javlja se nesimetrična pogrešna klasifikacija roka kao haus, kao i RnB-ja kao haus. Takođe se javljaju simetrična mešanja folka i hausa, kao i klasične muzike i džeza. Folk i haus nemaju muzičke sličnosti, dok klasična muzika i džez imaju, kao što je već spomenuto.

Matrice konfuzije na podacima za testiranje i validaciju NATA baze prikazane je na slici 5.7.1.



Slika 5.7.1: Matrice konfuzije na podacima za testiranje i validaciju NATA baze
Figure 5.7.1: Confusion matrices on NATA database testing and validaton datasets

Nad GTZAN bazom na podacima za testiranje postignuta je tačnost od 60.7%, a na podacima za validaciju 55.3%. Najgori rezultati na podacima za testiranje dobijeni su prilikom klasifikacije rege i rok žanrova, što se razlikuje od rezultata na podacima za validaciju, gde su najgore klasifikovani kantri, disko, džez i rok. Na podacima za testiranje javljaju se simetrična mešanja roka i bluza, kao i hiphopa i disko muzike. Rok i bluz muzika koju danas znamo naizgled i ne liči, ali u njihovim počecima postoji dosta sličnih elemenata i stilova u tim žanrovima. Disko i hip hop nemaju značajnijih zajedničkih elemenata iako su se tokom godina javljala mešanja ovih žanrova.
Matrice konfuzije nad GTZAN bazom na podacima za testiranje i validaciju prikazane su na slici 5.7.2.

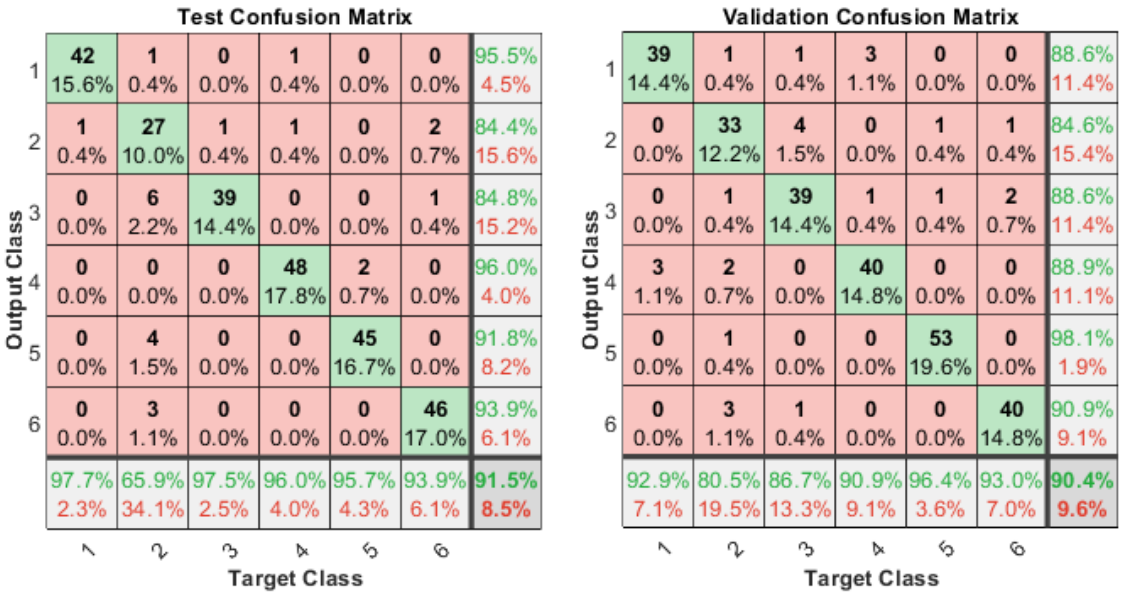


Slika 5.7.2: Matrice konfuzije na podacima za testiranje i validaciju GTZAN baze
Figure 5.7.2: Confusion matrices on GTZAN database testing and validation datasets

Rezultati - Spektar + LPCC

Kombinacija LPCC i Spektar karakterističnih obeležja postigla je tačnost od 91.5% nad NATA bazom, a tačnost od 60.0% nad GTZAN bazom. Optimalna tačnost postignuta je korišćenjem 100 neurona u skrivenom sloju sa sigmoid aktivacionom funkcijom.

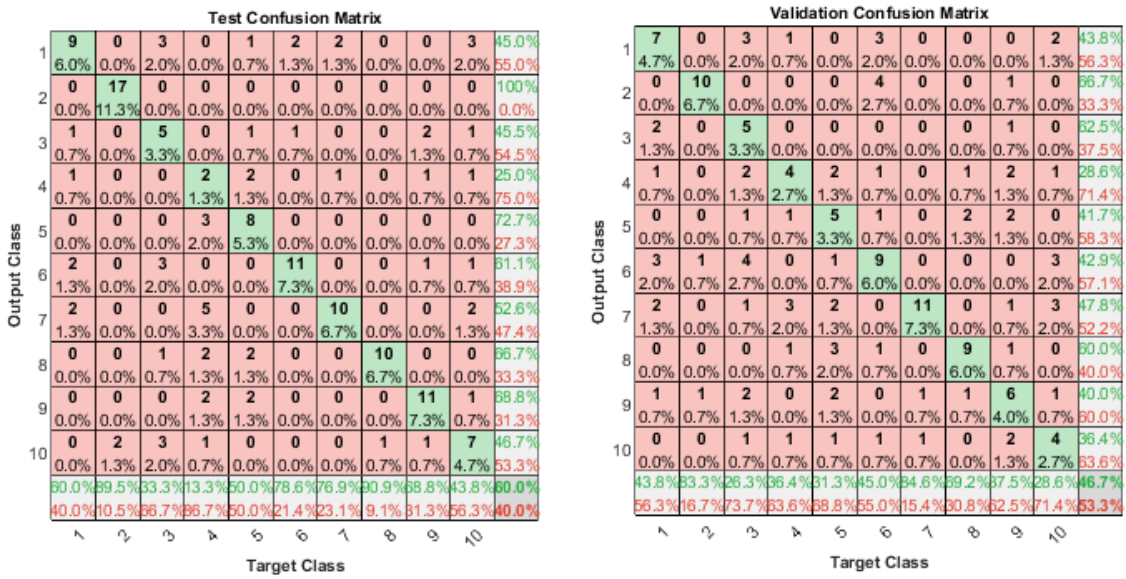
Nad NATA bazom na podacima za testiranje postignuta je tačnost od 91.5%, dok je na podacima za validaciju tačnost 90.4%. Na testiranju se folk izdvaja kao ubedljivo najgore klasifikovan žanr, što je slučaj i na validaciji, gde su folk i haus postigli najnižu tačnost.
Na slici 5.8.1 su prikazane matrice konfuzije na podacima za testiranje i validaciju NATA baze.



Slika 5.8.1: Matrica konfuzije na podacima za testiranje i validaciju NATA baze
Figure 5.8.1: Confusion matrices on NATA database testing and validation datasets

Nad GTZAN bazom na podacima za testiranje postignut je rezultat od 60.0%, dok je na podacima za validaciju tačnost 46.7%. Na podacima za testiranje nanjiža tačnost postignuta je pri klasifikaciji kantri, disko i rok žanrova. Na podacima za validaciju, s obzirom na veliku razliku u tačnosti, očekivano je da ima više klasa sa niskim procentom tačnosti. To se ispostavlja kao tačno - najniži procenat tačnosti postiže se prilikom klasifikacije kantri, disko, hiphop, rege i rok žanrova. Na podacima za testiranje dolazi do nesimetričnog mešanja disko žanra sa metalom, što se uzima kao statistička greška. Na podacima za validaciju dolazi do simetričnog mešanja bluza i kantri muzike, koji nemaju većih muzičkih sličnost.i

Matrice konfuzije GTZAN baze na podacima za testiranje i validaciju prikazane su na slici 5.8.2.



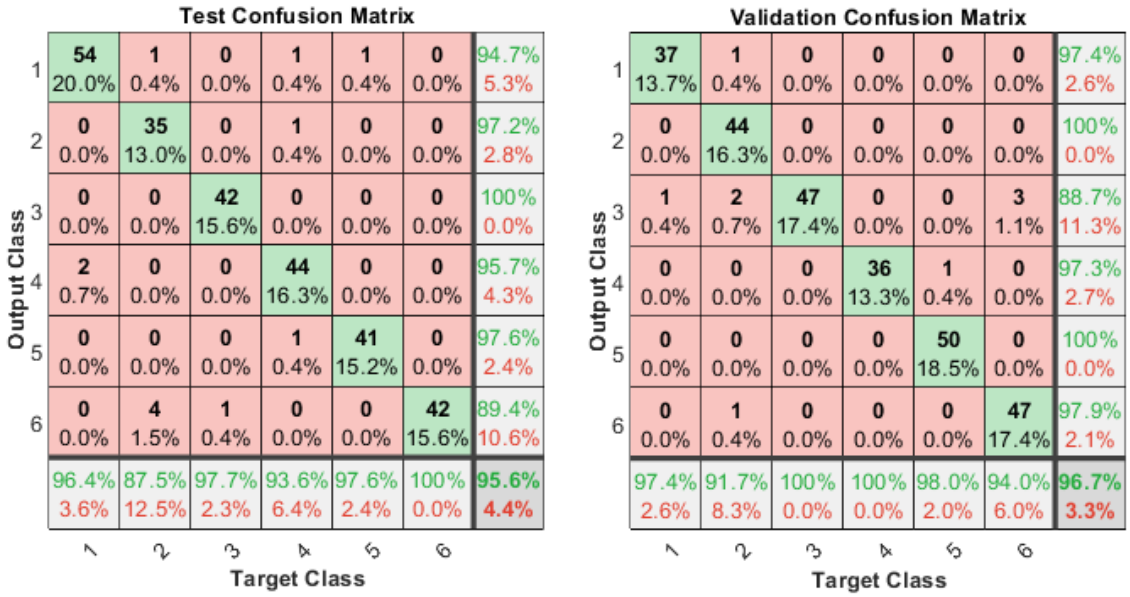
Slika 5.8.2: Matrice konfuzije na podacima za testiranje i validaciju GTZAN baze
Figure 5.8.2: Confusion matrix on GTZAN database testing and validation datasets

Rezultati - MFCC + LPCC + Spektar

Kombinacija MFCC, LPCC i Spektar karakterističnih obeležja na testiranju je postigla tačnost od 95.6% nad NATA bazom, odnosno 63.3% nad GTZAN bazom. Optimalna tačnost postignuta je korišćenjem 130 neurona u skrivenom sloju sa sigmoid aktivacionom fukncijom.

Nad NATA bazom na podacima za testiranje postignuta je tačnost od 95.6%, a na podacima za validaciju je tačnost 96.7%. Može se uočiti da i na podacima za testiranje i na podacima za validaciju postoje klase koje su bezgrešno klasifikovane i to rok kod testiranja, a haus i džez kod validacije. Na podacima za testiranje najgore je klasifikovan folk žanr, kao i na podacima za validaciju. Nije došlo do simetričnog mešanja ni na podacima za testiranje, ni na podacima za validaciju.

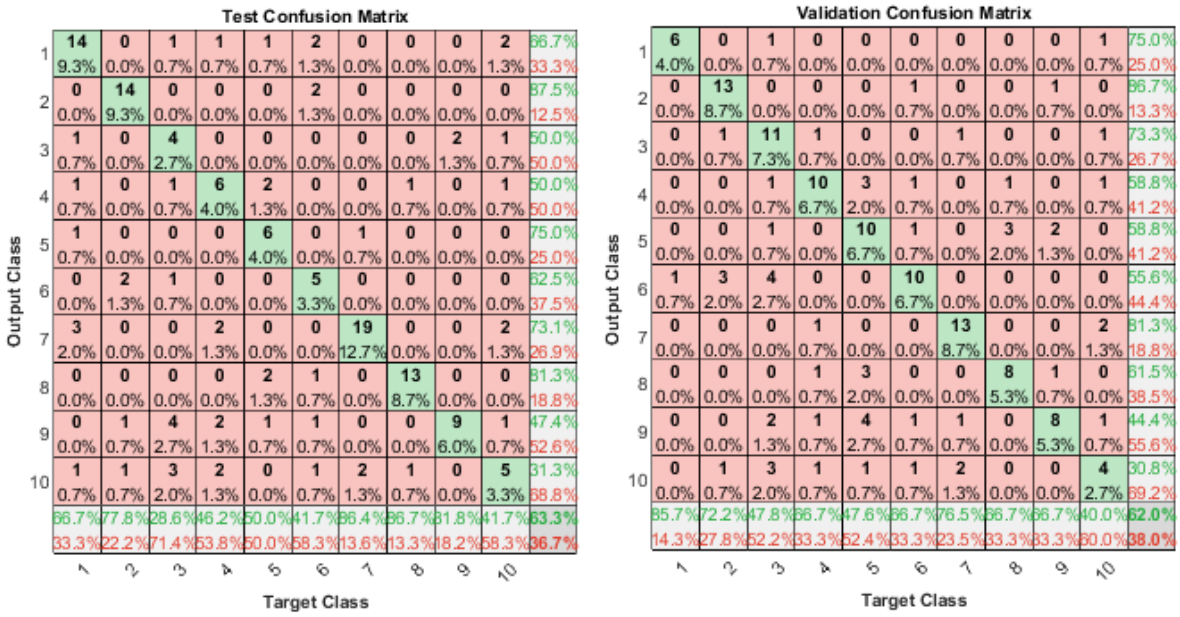
Na slici 5.9.1 prikazane su matrice konfuzije NATA baze redom na podacima za testiranje i validaciju.



Slika 5.9.1: Matrice konfuzije na podacima za testiranje i validaciju NATA baze
Figure 5.9.1: Confusion matrices on NATA database testing and validation datasets

Nad GTZAN bazom na podacima za testiranje tačnost klasifikacije iznosi 63.3%, a na podacima za validaciju 62.0%. Na testiranju najniža tačnost je ostvarena prilikom klasifikacije kantri, džez i rok pesama. Na validaciji su slični žanrovi dostigli slabiju tačnost - u pitanju su kantri, hiphop i rok. Na testiranju se može uočiti simetrično mešanje kantri i rege žanra. Rege i kantri žanrovi dele sličnost u vidu korišćenja akustične gitare, ali u vidu ritma i melodije nemaju većih sličnosti. Na validaciji se uočava simetrično mešanje hiphopa i regea, koji nemaju muzičke sličnosti.

Matrice konfuzije na podacima za testiranje i validaciju prikazane su na slici 5.9.2.



Slika 5.9.2: Matrice konfuzije na podacima za testiranje i validaciju GTZAN baze
Figure 5.9.2: Confusion matrices on GTZAN database testing and validation datasets

Zaključak

Predmet istraživanja je bilo poređenje tačnosti 4 različite metode za klasifikaciju žanrova muzike, kao i kombinacija tih metoda. Metode su testirane na dve različite baze - NATA bazi i GTZAN bazi. Prednost korišćenja dve baze u odnosu na jednu najviše se ogleda u mogućnosti ispitivanja sličnosti između relativnih tačnosti nad obe baze. Najveća postignuta tačnost nad GTZAN bazom je 63.3%, što su postigle kombinacija Spektar i MFCC karakterističnih obeležja, kao i kombinacija Spektar, MFCC i LPCC karakterističnih obeležja. Najveća postignuta tačnost nad NATA bazom je 95.9%, što je postigla kombinacija Spektar i MFCC karakterističnih obeležja.

Istraživanjem su ispitane sve hipoteze. Hipoteze istraživanja odnose se na rezultate na podacima za testiranje, ne uzimajući u obzir podatke za treniranje i validaciju.
Hipoteze istraživanja su sledeće:

1. MFCC metoda će dati bolji rezultat pri klasifikaciji žanra muzike od Spektar i Fingerprint metoda, nad obe baze
2. Fingerprint metoda će postići najgori rezultat nad obe baze

3. MFCC i LPCC će dati sličan rezultat nad obe baze, gde dve metode daju sličan rezultat ako se njihove tačnosti ne razlikuju za više od 5 procenata
4. Relativna tačnost metoda će biti slična nad obe baze

Istraživanjem je pokazano da Fingerprint metoda postiže najniži rezultat, što potvrđuje drugu hipotezu. Prva hipoteza se može smatrati delimično tačnom, iako je Spektar metoda postigla bolji rezultat od MFCC metode nad NATA bazom. Spektar metoda je uspešnija od MFCC metode za samo 1.2% (što se može smatrati statistički zanemarljivim), a MFCC metoda je pokazala bolju uspešnost pri mešanju sa drugim metodama nego Spektar metoda. Treća hipoteza je oborena, kako je nad GTZAN bazom MFCC metoda dostigla tačnost od 60.0%, a LPCC svega 52.7%. Četvrta hipoteza je takođe oborena, s obzirom da relativna tačnost nije slična nad obe baze za skoro polovinu ukupnog broja metoda.

Na osnovu ovog istraživanja, moguće je izdvojiti i dodatne zaključke:

1. Povećavanjem broja ulaznih parametara optimalno je povećanje i broja neurona u skrivenom sloju, zarad dobijanja najveće tačnosti za tu metodu, osim u slučaju Fingerprint metode
2. Sveukupno najbolji rezultat postigla je kombinacija Spektar i MFCC karakterističnih obeležja, jer je nad obe baze ostvarila najveću tačnost
3. Nad NATA bazom žanr koji je najčešće pogrešno klasifikovan je folk
4. Nad GTZAN bazom žanr koji je najčešće pogrešno klasifikovan je rok

Zarad daljeg istraživanja trebalo bi ispitati performanse ovih metoda nad većom bazom. Takođe, potrebno je istražiti kako se menja uspešnost kombinovanih metoda pri promeni dužine isečaka pesama (npr. dužine od 2 ili 10 sekundi). S obzirom da su ispitivane isključivo mreže sa jednim skrivenim slojem, potrebno je analizirati promenu tačnosti metoda pri povećanju broja skrivenih slojeva.

Zahvalnost

Zahvaljujemo se mentoru Milomiru na velikom strpljenju i pomoći pri radu na ovom istraživanju. Takođe zahvalnost na podršci bismo dali Bojanu Rošku za podsticaj i pomoć pri ideji projekta.

Literatura

[1] Natalija Todorčević. 2012. Analiza karakterističnih obeležja pri klasifikaciji muzike. Istraživačka stanica Petnica.

[2] Huang, X., Acero, A., Hon, H., 2001. Spoken Language Processing: A guide to theory, algorithm, and system development. Prentice Hall, Upper Saddle River, NJ, USA (pp. 314-315).

[3] An evaluation of Convolutional Neural Networks for music classification using spectrograms, Yandre M.G.Costa, Luiz S.Oliveira, Carlos N.Silla Jr. (2017).

[4] How does Shazam work? Music Recognition Algorithms, Fingerprinting, and Processing

[5] Neural network based recognition of speech using MFCC features, Pialy Barua,Kanij Ahmad, Ainul Anam Shahjamal Khan, Muhammad Sanaullah (2014)

[6] N. Dave. 2013. Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition. International journal for advance research in engineering and technology

[7] Davis, S. Mermelstein, P. (1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences

[8] Tom LH. Li , Antoni B. Chan ; Genre Classification and the Invariance of MFCC Features to Key and Tempo

[9] Beigi, Homayoon (2011). Fundamentals of Speaker recognition.

[10] G. Kour, N. Mehan. 2015. Music Genre Classification using MFCC, SVM and BPNN. International Journal of Computer Applications (0975 – 8887) Volume 112 – No. 6.

[11] A Scale for the Measurement of the Psychological Magnitude Pitch, S. S. Stevens, J. Volkmann, and E. B. Newman

[12] O. C. Ai, M. Hariharan, S. Yaacob, L. S. Chee. 2011. Classification of speech dysfluencies with MFCC and LPCC features.