Prepoznavanje instrumenata u zvukovnim zapisima primjenom dubokih neuronskih mreža

Odak, Ino

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 516

PREPOZNAVANJE INSTRUMENATA U ZVUKOVNIM ZAPISIMA PRIMJENOM DUBOKIH NEURONSKIH MREŽA

Ino Odak

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 516

PREPOZNAVANJE INSTRUMENATA U ZVUKOVNIM ZAPISIMA PRIMJENOM DUBOKIH NEURONSKIH MREŽA

Ino Odak

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

DIPLOMSKI ZADATAK br. 516

Pristupnik:	**Ino Odak (0036523560)**
Studij:	Računarstvo
Profil:	Znanost o podacima
Mentor:	izv. prof. dr. sc. Goran Delač

Zadatak: **Prepoznavanje instrumenata u zvukovnim zapisima primjenom dubokih neuronskih mreža**

Opis zadatka:

Istražiti i opisati modele zasnovane na dubokim neuronskim mrežama prikladne za klasifikaciju instrumenata u glazbenim zapisima. Poseban naglasak staviti na modele zasnovane na konvolucijskim neuronskim mrežama. Prikupiti prikladan skup zvukovnih zapisa u kojem je zabilježen zapis poznatog skupa instrumenata, primjerice skup podataka OpenMIC. Osmisliti i primjenski ostvariti sustav prepoznavanja instrumenata u zvukovnom zapisu. Opisati korišteni model i postupak učenja. Primjenom prikladno odabranih mjera za vrednovanje ispitati radna svojstva programskog ostvarenja.

Rok za predaju rada: 28. lipnja 2024.

Sadržaj

1. Uvod

1.1. Pozadina i motivacija

Područje pronalaženja glazbenih informacija (Music information retrieval - MIR) posljednjih je godina privuklo značajnu pozornost, potaknuto širenjem digitalne glazbe i sve većom potražnjom za učinkovitim i točnim metodama upravljanja, analize i razumijevanja velikih kolekcija zvučnih zapisa. Jedan od ključnih zadataka u MIR-u je klasifikacija glazbenih instrumenata unutar zvučnog zapisa, postupak koji podrazumijeva identifikaciju instrumenata prisutnih u određenom zvučnom snimku. Ovaj zadatak posebno je izazovan kod polifonog zvuka, gdje više instrumenata svira istovremeno, stvarajući složen i slojevit zvučni signal.

Motivacija za ovaj rad je višestruka. Prvo, točna klasifikacija glazbenih instrumenata ima značajne implikacije za muzikologiju i etnomuzikologiju, pružajući znanstvenicima alate za učinkovitu analizu glazbenih skladbi i izvedbi. Drugo, poboljšava mogućnosti sustava za preporuku glazbe omogućavajući personaliziranije i kontekstualno relevantnije prijedloge za slušatelje. Treće, podržava razvoj naprednih alata za uređivanje i produkciju zvuka, pružajući precizniju manipulaciju pojedinačnih zapisa instrumenata unutar miksa. Konačno, doprinosi širem području umjetne inteligencije i strojnog učenja predstavljajući bogatu i izazovnu problemsku domenu koja pokreće inovacije u prepoznavanju uzoraka i digitalnoj obradi signala.

Složenost klasifikacije glazbenih instrumenata u zvučnim zapisima proizlazi iz brojnih čimbenika. Akustične karakteristike različitih instrumenata često variraju te se mogu dodatno mijenjati različitim tehnikama sviranja, uvjetima snimanja i prisutnošću drugih instrumenata. Tradicionalne metode zvučne analize, poput spektralne analize i obrade u vremenskoj domeni, često nisu dovoljne za obuhvatiti sitne detalje potrebne za točno raspoznavanje. Nedavni napredak u strojnom učenju, posebice u dubokom učenju, rezultirao je novim mogućnostima u rješavanju ovih izazova učenjem složenih prikaza zvučnih signala koji mogu razlikovati različite instrumente čak i u gustim polifonim kontekstima.

1.2. Zvuk

Zvuk je mehanički val koji se širi kroz medij, obično zrak, kao rezultat vibracija iz izvora. Zvučne valove karakterizira nekoliko ključnih svojstava poput frekvencije, amplitude i timbra. Proučavanje zvuka uključuje razumijevanje njegovih fizičkih svojstava, kako ga percipira ljudski slušni sustav i kako se može analizirati i obraditi pomoću različitih tehnika.

Frekvencija zvuka odnosi se na broj vibracija ili ciklusa u sekundi, mjereno u Hercima (Hz). Određuje visinu zvuka, pri čemu se više frekvencije percipiraju kao više visine, a niže frekvencije kao niže visine. Ljudski sluh obično se kreće od 20 Hz do 20 000 Hz.

Amplituda je visina zvučnog vala i odgovara glasnoći ili jačini zvuka. Obično se mjeri u decibelima (dB). Valovi veće amplitude percipiraju se kao glasniji zvukovi, dok su valovi manje amplitude tiši.

Timbar, također poznat kao boja ili kvaliteta tona, razlikuje vrste proizvodnje zvuka, kao što su različiti glazbeni instrumenti ili glasovi. Na njega utječe harmonijski sadržaj zvuka i način na koji se on razvija tijekom vremena. Timbar je ono što nam omogućuje da razlikujemo klavir od violine, čak i ako sviraju istu notu pri istoj glasnoći.

Ljudski slušni sustav vrlo je složen i sposoban je razlikovati širok raspon zvukova. Na percepciju zvuka utječu različiti psihoakustički fenomeni, poput slušnog maskiranja, gdje glasniji zvukovi mogu otežati perceptivnost tiših, i učinak koktel zabave, koji se odnosi na ljudsku sposobnost da se usredotoči na jedan izvor zvuka u bučnom okruženju.

U kontekstu glazbe, zvuk poprima dodatne slojeve složenosti i značenja. Glazbeni zvukovi obično su organizirani u strukture kao što su melodija, harmonija i ritam, a proizvode ih različiti instrumenti i glasovi. Svaki instrument ima svoj jedinstveni zvučni profil, definiran njegovom fizičkom konstrukcijom i načinom na koji se svira. Razumijevanje ovih zvučnih profila ključno je za zadatke kao što je klasifikacija glazbenih instrumenata. U polifonoj glazbi, gdje više instrumenata svira istovremeno, interakcija njihovih zvukova stvara bogatu i složenu zvučnu teksturu. Analiza ove teksture zahtijeva napredne tehnike za određivanje doprinosa svakog instrumenta i identificiranje njihovih jedinstvenih karakteristika.

1.3. Pregled

U sklopu ovog diplomskog rada, izveden je i detaljno opisan sustav za klasifikaciju glazbenih instrumenata u polifonim zvučnim zapisima zasnovan na neuronskim mrežama. Istražene su suvremene metode digitalne obrade zvučnih signala, postupci izdvajanja korisnih i informativnih značajki iz neobrađenih zvučnih signala i objašnjeni su principi rada modernih klasifikacijskih algoritama u kontekstu klasifikacije zvučnih zapisa. Konačno, opisan je i vrednovan konačni sustav koji je davao razmjerno dobre rezultate.

Ostatak rada organiziran je na sljedeći način: Poglavlje 2 detaljno opisuje povijesni razvoj domene klasifikacije zvučnih zapisa, od metoda izdvajanja značajki do algoritama za klasifikaciju. Poglavlje 3 usredotočeno je na objašnjavanje zadatka klasifikacije glazbenih instrumenata unutar zvučnih zapisa. U ovom poglavlju opisuje se podijela glazbenih instrumenata u obitelji i daje se uvid u povijesni razvoj klasifikacije glazbenih instrumenata. Poglavlje 4 opisuje programsku izvedbu sustava za klasifikaciju zajedno s tehničkim pojedinostima. Čitav rad je zaključen u poglavlju 5.

2. Klasifikacija zvučnih zapisa

Klasifikacija zvuka osnovni je zadatak u područjima traženja glazbenih informacija (MIR) i obrade zvučih signala. Ovaj postupak podrazumijeva kategorizaciju zvučnih signala u unaprijed definirane razrede na temelju njihovih karakteristika, a ključan je za širok raspon primjena kao što su klasifikacija glazbenih žanrova, prepoznavanje govora, klasifikacija zvukova iz okoliša i klasifikaciju glazbenih instrumenata.

Važnost klasifikacije zvuka istaknuta je njezinim brojnim i praktičnim primjenama, na primjer, u uslugama prijenosa glazbe, algoritmi klasifikacije zvučnih zapisa pomažu organizirati i preporučiti glazbene zapise korisnicima na temelju žanrova, raspoloženja ili drugih specifičnih karakteristika. U području prepoznavanja govora, klasifikacija zvuka omogućuje identifikaciju izgovorenih riječi i fraza, omogućujući razvoj virtualnih pomoćnika i sustava prijepisa. Klasifikacija zvukova iz okoliša važna je za primjene kao što su nadzor i praćenje divljih životinja ili sustavi pametnih kuća, gdje prepoznavanje određenih zvukova može pokrenuti automatizirane reakcije.

Područje se značajno razvilo u proteklim godinama, a potaknuto je napretkom računalne snage, strojnog učenja i većom dostupnošću podataka. Rani pristupi zvučne klasifikacije oslanjali su se na ručno izrađivanje značajki i jednostavne statističke metode, dok moderne tehnike iskorištavaju duboko učenje za automatsko učenje složenih uzoraka iz neobrađenih zvučnih signala.

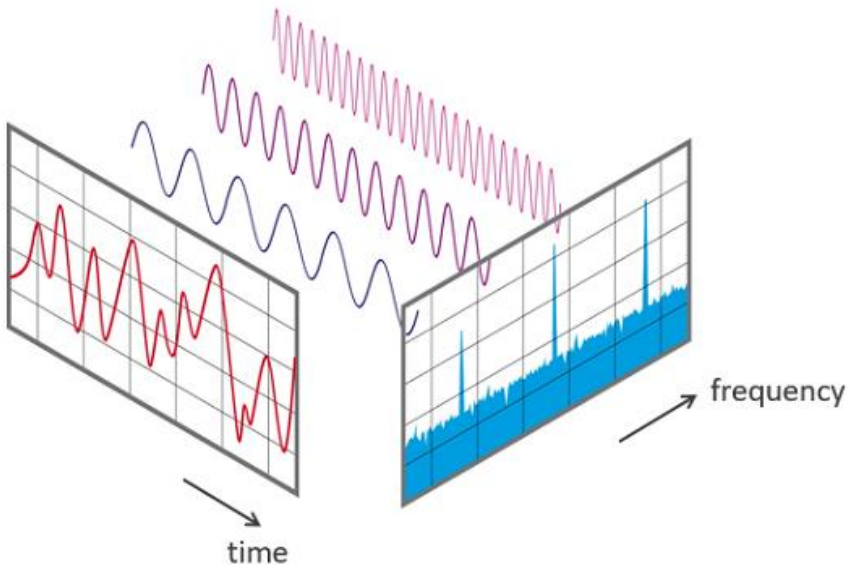
2.1. Povijesni pregled

Evolucija klasifikacije zvukova obilježena je s nekoliko ključnih faza. Svaku fazu karakteriziraju različiti tehnološki napreci i metodološki pristupi. Ovo poglavlje prati povijesni razvoj algoritama za klasifikaciju zvuka te ističe značajne prekretnice i otkrića.

2.1.1. Rane metode i pristupi

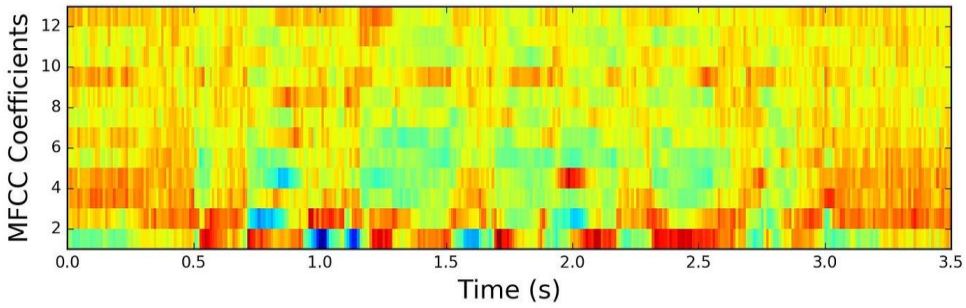
Rane metode klasifikacije zvučnih zapisa postavile su osnovu za napredne tehnike koje vidimo danas. Ovo je razdoblje obilježeno razvojem temeljnih koncepata i alata koji su omogućili početke analize i klasifikacije audio signala. Ključna područja istraživanja tijekom ovog razdoblja uključivala su tehnike obrade signala, metode ekstrakcije značajki i jednostavne algoritme za prepoznavanje uzoraka.

Područje klasifikacije zvuka započinje primjenom osnovnih tehnika obrade signala za analizu zvučnih signala. Cilj je bio transformirati neobrađeni valni oblik zvuka u prikaz koji se može lakše interpretirati i analizirati. Fourierova transformacija (FT) jedan je od prvih revolucionarnih alata koji je istraživačima omogućio rastavljanje zvučnog signala na sastavne frekvencije. Transformacijom signala vremenske domene u frekvencijsku domenu, Fourierova transformacija omogućila je uvid u spektralni sadržaj zvuka i postala standard za obradu zvučnih signala Kao što navode Oppenheim i Schafer (2009.), "Fourierova transformacija ključna je za frekvencijsku analizu signala, pružajući jasan prikaz spektralnih komponenti". Naprednija varijanta Fourierove transformacije, Fourierova transformacija kratkog vremena (STFT) omogućila je analizu audio signala tijekom vremena. Primjenom Fourierove transformacije na kratke preklapajuće segmente zvučnog signala, STFT proizvodi vremensko-frekvencijski prikaz, poznat kao spektrogram koji ispituje kako se spektralni sadržaj zvučnog signala mijenja tijekom vremena. Kao alternativa FT-u i STFT-u razvijena je Valićna transformacija koja pruža analizu zvučnog signala s više razlučivosti. Ova metoda rastavlja signal na komponente u različitim razmjerima i obuhvaća informacije o vremenu i frekvenciji s visokom preciznošću, što je osobito korisno za analizu prolaznih i nestacionarnih signala, poput glazbenih nota i govornih fonema.



Slika 1 - Prikaz signala u vremenskoj i frekvencijskoj domeni. Preuzeto s [https://www.nti-audio.com/en/support/know-how/fast-fourier-transform-fft]

Ekstrakcija značajki kardinalan je korak u klasifikaciji zvuka, a podrazumijeva transformaciju neobrađenih zvučnih podataka u skup značajnih karakteristika ili značajki. Rane metode bile su usmjerene na ručno izdvajanje značajki koje su opisivale važna svojstva zvučnog signala. Značajke vremenske domene izvedene su izravno iz neobrađenog valnog oblika zvučnog signala. Primjeri značajki vremenske domene su mjere poput stope prelaska nule (brzina kojom signal mijenja predznak), korijena srednje kvadratne energije (mjera snage signala) i momenata valnog oblika. Ove su značajke primitivne, ali pružaju vrijedne informacije o amplitudi i vremenskoj strukturi signala. Kao što su opisali Rabiner i Juang (1993.), "značajke vremenske domene jednostavne su, ali učinkovite u hvatanju osnovnih svojstava zvučnih signala, ključnih za početne zadatke klasifikacije". Značajke frekvencijske domene izvedene su iz spektralne reprezentacije zvučnog signala. Fourierova transformacija i njezine varijante olakšale su izdvajanje spektralnih značajki kao što su spektralno središte (središte mase spektra), spektralna propusnost (širina spektra) i spektralni odmotaj (frekvencija ispod koje određena sadržan postotak ukupne spektralne energije). Ove značajke bile su ključne za razlikovanje različitih vrsta zvukova na temelju njihovog frekvencijskog sadržaja. Cepstralnom analizom razvijeni su Mel-Frekvencijski Cepstralni Koeficijenti (MFCC) koji su postali široko korištena tehnika za izdvajanje značajki iz zvučnih zapisa. MFCC predstavljaju kratkoročni spektar snage zvučnog signala na Mel ljestvici, koja približno odgovara osjetljivosti ljudskog uha na različite frekvencije. To je učinilo MFCC vrlo učinkovitim u zadatcima poput klasifikacije govora i glazbe.



Slika 2 - MFCC. Preuzeto s [https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd]

Nakon što su značajke izdvojene, sljedeći je korak bio klasificiranje audio signala na temelju tih značajki. Rani algoritmi za prepoznavanje uzoraka bili su relativno jednostavni, ali su postavili osnove za naprednije metode. Neki od algoritama koji su korišteni u ovom razdoblju

uključuju sustave zasnovane na pravilima, K najbližih susjeda (K-NN) i linearnu diskriminantnu analizu (LDA). Rane metode i osnove klasifikacije zvuka primijenjeni su na zadatke poput prepoznavanja ljudskog govora ili klasifikacije glazbenog žanra. Svaki od ovih zadataka predstavljao je jedinstven niz izazova zbog čega rezultati nisu bili zadovoljavajući.

2.1.2. Pojava strojnog učenja

Pojava strojnog učenja označila je prekretnicu u klasifikaciji zvučnih zapisa uvođenjem naprednih algoritama sposobnih za učenje obrazaca i predviđanje iz podataka. Ova promjena značajno je poboljšala točnost i učinkovitost na zadatacima klasifikacije zvukova:

Jedan od početnih koraka u primjeni strojnog učenja na zadatak klasifikacije zvučnih zapisa je razvoj efikasnih tehnika ekstrakcije značajki. Značajke služe kao ulaz u algoritme strojnog učenja, opisujući bitne karakteristike zvučnog signala. MFCC istakli su se kao posebno korisna metoda ekstrakcije značajki jer nude način predstavljanja spektra snage zvuka, oponašajući reakciju ljudskog uha na različite frekvencije. Davis i Mermelstein (1980.) naglašavaju: "MFCC su postali standard u ekstrakciji audio značajki, pružajući robusnu izvedbu u različitim zadacima klasifikacije". Korištene su i spektralne značajke poput chroma značajki, spektralnog centroida, propusnosti i odmotaja. Ove značajke opisuju frekvencijski sadržaj i razdiobu energije zvučnog signala i sadrže vrijedne informacije za klasifikaciju. Kombinacija spektralnih značajki s vremenskim značajkama, koje opisuju dinamiku audio signala tijekom vremena, pružila je kvalitetan prikaz zvučnih podataka.

Sljedeći korak ovog procesa primjena je klasičnih algoritama strojnog učenja za klasificiranje zvučnih signala. Nekoliko algoritama koji su odigrali ključnu ulogu u tom razdoblju uključuju strojeve potpornih vektora (SVM), stabla odluke i slučajne šume te modele Gaussovih mješavina (GMM). Ovi su algoritmi unaprijedili točnost klasifikacije u odnosu na osnovne metode i pristupe te su promijenili paradigmu područja klasifikacije zvučnih zapisa.

Razvoj strojnog učenja u klasifikaciji zvuka dodatno je bila uzrokovana povećanom dostupnošću velikih označenih skupova podataka. Ovi su skupovi osigurali istraživačima podatke potrebne za učenje i vrednovanje modela strojnog učenja. Obilje podataka modelima je omogućilo da uče uzorke iz različitih zvučnih primjera, poboljšavajući njihovu generalizaciju na situacije iz stvarnog svijeta.

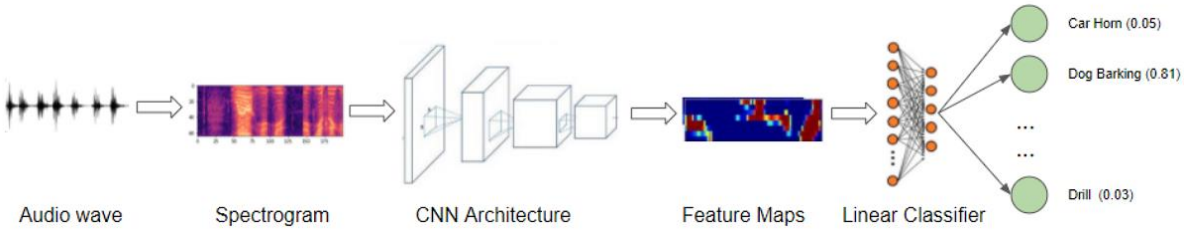
Kako bi utvrdili performanse modela strojnog učenja u klasifikaciji zvukova, istraživači su razvili različite metrike procjene i referentne vrijednosti poput točnosti, preciznosti, odziva i F1 mjere nudili su kvantitativne mjere uspješnosti klasifikacije. Mjerila i natjecanja, poput MIREX-ovih godišnjih evaluacija, potaknuli su suradnju i inovacije u području, pokrećući razvoj učinkovitijih i robusnijih metoda klasifikacije zvuka.

Integracija neuronskih mreža u audio klasifikaciju označila je značajan napredak i početak novog razdoblja klasifikacije zvučnih zapisa. Rani modeli neuronskih mreža, kao što su višeslojni perceptroni (MLP), primijenjeni su na zadatke klasifikacije zvučnih signala i demonstrirali potencijal ovog pristupa, ali bili su ograničeni računalnim resursima pa još nisu bili u mogućnosti pružiti zadovoljavajuće rezultate.

2.1.3. Revolucija dubokog učenja

Razvoj i napredak dubokog učenja i računalnog hardvera označio je revolucionarno razdoblje u području klasifikacije zvučnih zapisa, značajno poboljšavajući mogućnosti i točnost klasifikacijskih sustava. Modeli dubokog učenja, posebice konvolucijske neuronske mreže (CNN) i rekurentne neuronske mreže (RNN), omogućili su značajna poboljšanja automatskim učenjem složenih značajki iz neobrađenih zvučnih podataka.

Konvolucijske neuronske mreže (CNN) postale su standard u klasifikaciji zvuka zbog svoje sposobnosti učinkovitog obuhvaćanja lokalnih uzoraka u podacima. Izvorno razvijeni za obradu slike, CNN-ovi su prilagođeni za zvučne zadatke tretirajući vremensko-frekvencijske prikaze zvučnog signala, poput spektrograma, kao slike. Konvolucijski slojevi CNN-a mogli su automatski naučiti prostorne hijerarhije u tim spektrogramima, identificirajući značajke koje su bile relevantne za razlikovanje različitih audio klasa. Ovaj je pristup znatno nadmašio tradicionalne metode koje su se oslanjale na ručno izrađene značajke. Ovisno o prirodi podataka i zadatku klasifikacije, istraživači su koristili različite vrste CNN-a. Jednodimenzionalni CNN-ovi izravno su obrađivali neobrađene zvučne valove, modelirajući vremenske ovisnosti. Dvodimenzionalni CNN-ovi, koji su ujedno i najčešći tip za audio klasifikaciju,koristili su obrađene spektrograme i druge vremensko-frekvencijske reprezentacije. Trodimenzionalni CNN-ovi korišteni su za modeliranje prostorno-vremenskih uzoraka, posebno u primjenama koje zahtijevaju analizu zvučno-vizualnih podataka.



Slika 3 - Prikaz primjera podatkovnog cijevovoda u klasifikaciji zvučnih zapisa. Preuzeto s [https://towardsdatascience.com/audio-deep-learning-made-simple-sound-classification-step-by-step-cebc936bbe5]

Rekurentne neuronske mreže (RNN) su posebno prikladne za sekvencijalne podatke, što ih čini idealnim za zadatke koji uključuju podatke s temporalnom strukturom. RNN-ovi i njihove naprednije varijante poput mreže dugog kratkoročnog pamćenja (LSTM), modeliraju ovisnosti tijekom vremena, što je ključno za razumijevanje vremenske dinamike zvuka. LSTM-ovi su riješili problem nestajanja gradijenta u standardnim RNN-ovima, dopuštajući im da nauče dugoročne ovisnosti u sekvencijalnim podacima. Ova sposobnost modeliranja dugoročnih ovisnosti učinila je LSTM-ove vrlo učinkovitima za zadatke kao što je prepoznavanje govora, gdje su vremenski redoslijed i trajanje zvukova ključni za točnu klasifikaciju. Kombinacija CNN-a i RNN-a pokazala se moćnim pristupom u domeni klasifikacije zvukova. CNN-ovi su korišteni za izdvajanje prostornih značajki iz spektrograma, dok su RNN-ovi modelirali vremenske ovisnosti tih značajki. Ovakva hibridna arhitektura iskorištva prednosti obiju mreža, što je dovelo do značajnih poboljšanja u točnosti i preciznosti klasifikacije.

Mehanizmi pozornosti dodatno su revolucionirali modele dubokog učenja za zvučnu klasifikaciju. Nadahnuti ljudskim kognitivnim procesima, mehanizmi pozornosti i samopozornosti omogućili su modelima da se usredotoče na najrelevantnije dijelove ulaznih podataka, poboljšavajući njihovu sposobnost uočavanja važnih uzoraka. Inovacija transformer arhitekture, koja se oslanjaja na mehanizme samopozornosti, obilježila je ključan pomak u ovoj domeni. Ovi modeli, izvorno razvijeni za obradu prirodnog jezika, prilagođeni su za zadatke klasifikacije zvučnih signala. Mehanizam samopozornosti je omogućilo modelu da odredi važnost različitih dijelova zvučnog signala, poboljšavajući njegovu sposobnost da modelira dugoročne odnose i djeluje u okruženjima u kojima se različiti zvukovi preklapaju i interferiraju.

Prijenosno učenje pokazalo se kao još jedna od ključnih tehnika dubokog učenja, rješavajući problem manjka podataka i oznaka. Korištenjem modela predtreniranih na velikim i raznolikim skupovima podataka, istraživači su mogli fino podesiti ove modele za specifične zadatke klasifikacije zvuka, značajno poboljšavajući performanse i smanjujući potrebu za opsežnim označenim skupovima podataka. Modeli poput OpenL3 i VGGish, predtrenirani na velikim skupovima audio podataka, služili su kao podloga za razne zadatke klasifikacije zvuka. Fino podešavanje ovih modela na podacima specifičnim za zadatak omogućilo je istraživačima postizanje visoke točnosti s relativno malim količinama označenih podataka.

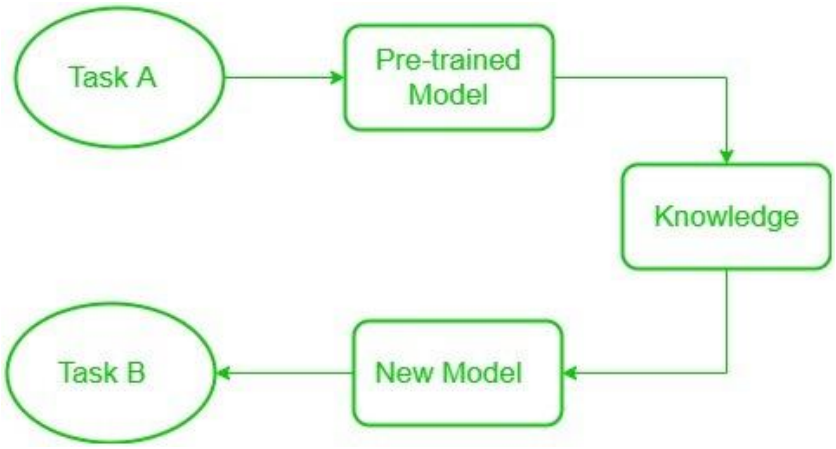
Kako bi dodatno poboljšali robusnost i generalizaciju modela dubokog učenja, istraživači su koristili različite tehnike augmentacije podataka. Ove tehnike uključivale su stvaranje dodatnih podataka za učenje primjenom transformacija kao što su pomicanje visine tona, vremensko rastezanje i dodavanje šuma izvornim zvučnim signalima kako bi se simulirali različiti uvjeti snimanja i varijacije.

2.2. Trenutno najuspješniji pristupi

Suvremenu zvučnu klasifikaciju karakterizira integracija modernih arhitektura dubokog učenja, inovativnih postupaka učenja i korištenja velikih, raznovrsnih skupova podataka. Ovi su napretci značajno poboljšali performanse i robusnost sustava klasifikacije zvuka u različitim primijenama.

Napredne arhitekture dubokog učenja i dalje su glavni fokus istraživanja u domeni. Konvolucijske neuronske mreže (CNN) i dalje su ključne zbog svoje sposobnosti učinkovitog modeliranja lokalnih uzoraka u vremensko-frekvencijskim prikazima audio signala, kao što su spektrogrami. Nedavni napredak uključuje dublje i složenije CNN arhitekture kao što su ResNet i DenseNet, koje imaju poboljšane mogućnosti izdvajanja značajki i zbog toga imaju veću točnost klasifikacije. Kao što su He et al. (2016.) opisuju: "ResNet olakšava obuku znatno dubljih mreža, omogućujući značajna poboljšanja u točnosti". Rekurentne neuronske mreže (RNN), posebno dugotrajna kratkoročna memorija (LSTM) izvrsna je u zadacima koji uključuju sekvencijalne podatke, kao što je prepoznavanje govora i analiza glazbe. Hibridni modeli koji kombiniraju CNN-ove za ekstrakciju značajki i RNN-ove za vremensko modeliranje pokazali su vrhunske performanse. Uvođenje transformer arhitekture, koja koristi mehanizme samopozornosti, dodatno je revolucioniralo ovo područje. Izvorno razvijeni za

obradu prirodnog jezika, transformeri su prilagođeni audio zadacima, omogućujući modelima da modeliraju dugotrajne ovisnosti učinkovitije od tradicionalnih RNN-ova. Vaswani et al. (2017.) navode: "Transformer arhitektura, sa svojim mehanizmom samopozornosti, pruža nov pristup modeliranju sekvenci, nadmašujući RNN-ove u mnogim zadacima". Arhitektura transformatora uspješno je primijenjena na zadatke kao što su prepoznavanje govora i generiranje glazbe, osiguravajući značajna poboljšanja performansi.



Slika 4 - Shematski prikaz prijenosnog učenja. Preuzeto s [https://www.geeksforgeeks.org/ml-introduction-to-transfer-learning/]

Inovativni postupci učenja također su odigrali važnu ulogu u unapređenju klasifikacije zvučnih zapisa. Prijenosno učenje, koje podrazumijeva fino podešavanje unaprijed naučenih modela na podacima specifičnim za zadatak, postaje je sve popularnije. Ovaj pristup omogućuje modelima da iskoriste znanje stečeno iz velikih skupova podataka, poboljšavajući performanse i smanjujući potrebu za velikim brojem označenih primjera. Kao što su primijetili Hershey et al. (2017.), "Prijenosno učenje s modelima kao što je VGGish omogućuje učinkovito korištenje unaprijed uvježbanih zvučnih reprezentacija, poboljšavajući performanse na određenim zadacima".

Tehnike samonadgledanog učenja omogućile su modelima da uče iz neoznačenih podataka stvaranjem pomoćnih zadataka, kao što je predviđanje maskiranih dijelova ulaza ili razlikovanje proširenih verzija istih podataka. Ovaj pristup smanjuje ovisnost o označenim podacima i pokazao je obećavajuće rezultate u domeni klasifikacije zvuka. Schneider et al. (2019.) naglašavaju: "Samonadzirano učenje otvara nove putove za korištenje velikih količina neoznačenih zvučnih podataka, značajno poboljšavajući performanse modela" . Kontrastivno

učenje, gdje modeli uče uspoređujući slične i različite zvučne segmente, najbolji je primjer potencijala samonadziranog učenja.

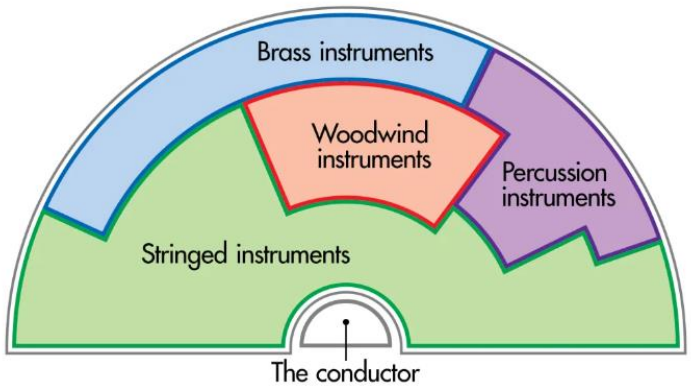
Budući smjerovi u istraživanju audio klasifikacije uključuju multimodalno učenje, kombiniranje audio podataka s drugim modalitetima poput videa i teksta kako bi se stvorili sveobuhvatniji i kvalitetniji sustavi klasifikacije. Poboljšanje objašnjivosti i interpretabilnosti također su ključni, kako bi učinili modele dubokog učenja transparentnijima i razumljivijima, čime se povećava povjerenje u ovakve sustave. Omogućavanje obrade u stvarnom vremenu još je jedno ključno područje, usmjereno na obradu podataka u stvarnom vremenu u područjima kao što su analiza glazbe uživo i obrada na uređaju za prijenosne uređaje. Osim toga, rješavanje pristranosti u modelima i osiguravanje da su uključivi i pošteni, posebno u primjenama koje utječu na različite skupine korisnika, ostaje važan cilj.

3. Klasifikacija glazbenih instrumenata u zvučnim zapisima

Klasifikacija glazbenih instrumenata jedna je od inačica zadatka klasifikacije zvučnih zapisa. Ovaj zadatak uključuje identifikaciju i kategorizaciju instrumenata prisutnih u zvučnom zapisu, postupak koji je neophodan za širok raspon primjena kao što su automatizirani prijepis glazbe, sustavi glazbenih preporuka i uređivanje zvučnih snimki. Područje klasifikacije glazbenih instrumenata usko je povezano s općenitim slučajem klasifikacije zvučnih zapisa. Klasifikacija zvuka obuhvaća velik broj zadataka, poput prepoznavanja govora i klasifikacije zvukova iz okoliša,dok se klasifikacija glazbenih instrumenata usredotočava na razlikovanje različitih glazbenih instrumenata unutar zvučnog signala. Glazbeni instrumenti proizvode zvuk koristeći različite mehanizme, kao što su žice ili membrane, od kojih svaki daje jedinstvene spektralne i vremenske karakteristike rezultirajućem zvučnom signalu. Točna klasifikacija instrumenata uključuje analizu ovih karakteristika i njihovo razlikovanje od karakteristika drugih instrumenata prisutnih u zvučnom zapisu.

3.1. Taksonomija glazbenih instrumenata

Taksonomija glazbenih instrumenata pruža strukturirani okvir za razumijevanje odnosa i sličnosti između različitih instrumenata. Zasniva se na različitim čimbenicima, uključujući metodu proizvodnje zvuka, akustična svojstva instrumenata i njihove timbralne karakteristike. Razumijevanje ovih odnosa bitno je za razvoj točnih modela klasifikacije, jer instrumenti unutar iste obitelji često dijele slične akustičke značajke, što njihovu diferencijaciju čini izazovnijom.



Slika 5 - Glazbeni orkestar. Preuzeto s [https://www.timeforkids.com/wp-content/uploads/2020/10/instruments5.jpg?w=1024]

Glazbeni instrumenti se obično grupiraju u obitelji na temelju njihovih mehanizama za proizvodnju zvuka. Primarne obitelji su gudački instrumenti, puhački instrumenti, limeni instrumenti i udarački instrumenti od kojih svaka obitelj ima različite karakteristike koje utječu na njihova akustična svojstva i boju tona.

Gudački instrumenti proizvode zvuk kroz vibraciju žica, koje se mogu trzati, gudati ili udarati. Visina zvuka određena je duljinom, napetosti i masom žice, dok tijelo instrumenta pojačava zvuk i dodaje rezonanciju. U ovu skupinu instrumenata spadaju violina, viola, violončelo, kontrabas, gitara, harfa i drugi. Gudački instrumenti često imaju bogat harmonijski sadržaj, s osnovnom frekvencijom koja je praćena brojnim prizvucima. Kao što Benade (1990.) primjećuje, "Vibracija žica stvara složene harmonike, koji su ključni za bogate timbralne karakteristike gudačkih instrumenata". Razlike u timbru između gudačkih instrumenata proizlaze iz čimbenika kao što su materijal žica (cijevo, čelik, najlon), oblik i veličina tijela instrumenta i način proizvodnje zvuka. Primjerice, gudala poput violine i violončela imaju glađi i dugotrajniji zvuk u usporedbi s trzalačkim instrumentima poput gitare, koji proizvode prolaznije i udarnije zvukove.

Puhački instrumenti stvaraju zvuk vibriranjem zračnog stupca unutar cijevi. Ova se vibracija može pokrenuti puhanjem preko usnika (kao kod flaute) ili korištenjem trske (kao kod klarineta, oboe i saksofona). Visina se kontrolira otvaranjem i zatvaranjem rupa duž tijela instrumenta, čime se mijenja duljina vibrirajućeg zračnog stupca. Na boju puhačkih instrumenata utječe oblik instrumenta (cilindričan ili stožast), materijal instrumenta (drvo, metal, plastika) i prisutnost jedne ili dvije trske. Na primjer, klarinet, sa svojim cilindričnim otvorom i jednim jezičkom, proizvodi zvuk bogate, rezonantne kvalitete i snažne osnovne frekvencije, dok oboa, sa svojim stožastim otvorom i dvostrukom jezičkom, ima svjetliji, prodorniji zvuk.

Limeni instrumenti proizvode zvuk vibracijom sviračevih usana na usniku, što zatim uzrokuje vibriranje stupca zraka unutar instrumenta. Visina tona se mijenja promjenom duljine cijevi pomoću ventila ili klizača. Ova skupina instrumenata uključuje trubu, trombon, francuski rog i tubu. Limeni instrumenti poznati su po svom snažnom zvuku i jakim harmonijskim prizvucima. Prema Campbellu i Greatedu (2004.), "Limena glazbala pokazuju jedinstvenu spektralnu kvalitetu zbog interakcije između vibracija usana svirača i rezonantnih frekvencija instrumenta". Materijal (obično mjed ili drugi metali) i oblik instrumenta (cilindrični ili stožasti

otvor) pridonose njihovom prepoznatljivom tonu. Na primjer, truba, sa svojim cilindričnim otvorom, proizvodi oštar i prodoran zvuk, dok francuski rog, sa svojim stožastim otvorom, ima topliji i mekši ton.

Udarački instrumenti zvuk stvaraju kroz vibraciju membrane ili čvrstog materijala kada ih se udari, trese ili struže. Udaraljke se dijele na tonske (npr. ksilofon, marimba, timpani) i beztonske (npr. mali bubanj, bas bubanj, činele). Zvuk udaraljki varira ovisno o materijalu (drvo, metal, koža), obliku i veličini instrumenta te načinu proizvodnje zvuka. Udaraljke s visokim tonom, poput marimbe, proizvode različite visine i imaju rezonantnu, melodijsku kvalitetu. Udaraljke bez tona, poput bubnja, proizvode složenije zvukove nalik buci s manje definiranim visinama. Rossing (2000.) navodi: "Udaraljke obuhvaćaju širok raspon akustičnih svojstava, od određenih visina melodijskih udaraljki do složenih spektara beztonskih instrumenata sličnih šumu".

3.2. Povijesna pozadina

Povijesna pozadina klasifikacije glazbenih instrumenata obilježena je razvojem skupova podataka i modela koji su postupno povećavali točnost i mogućnosti klasifikacijskih sustava.

U ranim fazama istraživanja klasifikacije glazbenih instrumenata dostupnost podataka bila je ograničena. Rani pristupi često su se oslanjali na male, ručno odabrane skupove zvučnih snimki kojima je nedostajala raznolikost i veličina potrebna za učenje i vrednovanje modela. Početni skupovi podataka uglavnom su bili usredotočeni na monofone snimke instrumenata i stvorili su osnovu za razvoj robusnijih klasifikacijskih modela. Jedan od prvih značajnih skupova podataka bila je RWC (Real World Computing) glazbena baza podataka, koja se pojavila 2002. godine. Baza podataka sadržavala je raznoliku zbirku uzoraka koji su pokrivali širok raspon instrumenata i tehnika sviranja. Analizama je utvrđeno kako je ovaj skup podataka sadržavao brojne pogrešne oznake, ali svejedno je poslužio kao vrijedan izvor za razvoj i vrednovanje klasifikacijskih algoritama. Još jedan skup podataka iz rane faze klasifikacije zvučnih zapisa bila je baza podataka Master Samples (MUMS) Sveučilišta McGill, koja je sadržavala kvalitetne snimke raznih glazbenih instrumenata. Baza podataka MUMS nudila je bogat skup izdvojenih zvukova instrumenata, koji su bili ključni za proučavanje akustičkih svojstava različitih instrumenata i razvoj metoda ekstrakcije značajki.

Slika 6 - Logo IRMAS skupa podataka. Preuzeto s [https://www.upf.edu/web/humanitats/noticies/-/asset_publisher/tclOfFgGUeZL/content/irmas-a-dataset-for-instrument-recognition-in-musical-audio-signals/3854030]

Kako je područje klasifikacije zvučnih zapisa napredovalo, javila se potreba za većim i raznovrsnijim skupovima podataka. Istraživači su shvatili da stvaranje kvalitetnog modela klasifikacije zahtijeva velik broj zvučnih snimki koje obuhvaćaju raspon instrumenata, žanrova i uvjeta snimanja. Ovo je dovelo do razvoja nekoliko opsežnih skupova podataka koji su značajno unaprijedili ovo područje. Jedan od ovih skupova podataka je IRMAS (Instrument Recognition in Musical Audio Signals). Skup podataka IRMAS široko je korišten u polju klasifikacije glazbenih instrumenata. Bosch et al. (2012.) opisuju skup podataka kao "raznoliku zbirku glazbenih isječaka pažljivo odabranih da izazovu i unaprijede područje prepoznavanja instrumenata u polifoničnim postavkama.". Razvijen je kao podrška istraživanju automatskog prepoznavanja instrumenata u stvarnim glazbenim snimkama. Skup podataka osmišljen je za rješavanje izazova klasifikacije instrumenata u polifonom zvuku, gdje više instrumenata svira istovremeno. IRMAS skup podataka sadrži glazbene isječke označene dominantnim instrumentom prisutnim u svakom isječku. Isječci pripadaju različitim glazbenim žanrovima, uključujući klasiku, jazz, rock, pop i druge što osigurava velike varijacije glazbenih stilova i konteksta. Svaki zvučni zapis u skupu podataka označen je dominantnim instrumentom, što ga čini prikladnim za zadatke nadziranog učenja. Raznolikost u žanrovima i uvjetima snimanja pomaže osigurati da su modeli trenirani na IRMAS skupu podataka robusni i da dobro generaliziraju na situacije iz stvarnog svijeta, a podijela podataka na podskup za treniranje i podskup za vrednovanje dodatno olakšava postupke treniranja ovih modela. Skup podataka OpenMIC (Open Music Instrument Classification) još je jedan važan izvor podataka u domeni klasifikacije glazbenih instrumenata. Prikupljen je s ciljem da osigura velik, raznolik i kvalitetan skup podataka za istraživače koji rade na klasifikaciji instrumenata u polifonoj

glazbi. OpenMIC sadrži više od dvadeset tisuća audio zapisa, svaki u trajanju od deset sekundi. Izvor za ovaj podatkovni skup je Free Music Archive (FMA), a zvučni zapisi iz skupa pokrivaju raspon žanrova, uključujući klasiku, elektroniku, jazz i druge. Dizajniran je na način da svaki zvučni zapis može sadržavati više instrumenata, pa samim time i više oznaka. Humphrey et al. (2018.) navode: "Shema označavanja s više oznaka OpenMIC-a točno odražava istodobnu prisutnost više instrumenata, predstavljajući realističan izazov za modele klasifikacije.". Oznake u skupu podataka OpenMIC prikupljaju se crowdsourcing-om i automatiziranim metodama kako bi se osiguralo da su one visoko kvalitetne. Ovakav pristup osigurava realističniji i izazovniji skup podataka za razvoj modela.

Uz razvoj velikih skupova podataka, napredak klasifikacijskih modela obilježen je značajnim napretkom u tehnikama izdvajanja značajki i algoritmima strojnog učenja. Raniji klasifikacijski modeli su se oslanjali na ručno izrađene značajke i jednostavne klasifikatore. Značajke kao što su MFCC, spektralni centroid i brzina prelaska nule, korištene su u kombinaciji s klasifikatorima poput LDA i modela Gaussovih mješavina. Iako su ovi modeli postigli umjeren uspjeh, oslanjanje na ručno izdvajanje značajki ograničavalo je njihove performanse i mogućnosti. Pojava strojnog učenja u kasnom 20. stoljeću označila je razmjeran napredak klasifikacijskih modela. Algoritmi poput stabala odluke, nasumičnih šuma i SVM-ova primijenjeni su na zvučne podatke, omogućujući naprednije i točnije prepoznavanje uzoraka. Početkom 21. stoljeća započinje revolucija u klasifikaciji glazbenih instrumenata s pojavom dubokog učenja. Modeli dubokog učenja, posebno konvolucijske neuronske mreže (CNN) i rekurentne neuronske mreže (RNN), pokazali su izvanredne performanse u izdvajanju značajki i učenju hijerarhijskih prikaza iz neobrađenih zvučnih signala. CNN-ovi su postali popularni zbog svoje sposobnosti da automatski izdvajaju značajke iz spektrograma i drugih vremensko-frekvencijskih prikaza zvučnog signala. Hijerarhijska struktura CNN-ova omogućila im je da modeliraju složene uzorke u podacima, značajno poboljšavajući točnost klasifikacije. RNN-ovi, pogodni su za modeliranje vremenskih ovisnosti, što ih čini prikladnima za zadatke koji uključuju sekvencijalne podatke, kao što je prepoznavanje glazbe i govora. Hibridni modeli koji kombiniraju CNN-ove za ekstrakciju značajki i RNN-ove za vremensko modeliranje pokazali su se posebno učinkovitima za klasifikaciju polifonog zvuka jer su iskoristili prednosti obiju arhitektura. Najaktualniji pristupi iskorištavaju prijenosno učenje, korištenjem modela koji su predtrenirani na velikim skupovima podataka. Prema Panu i Yangu (2010.), "Prijenosno učenje

nastoji poboljšati izvedbu ciljanog zadatka iskorištavanjem znanja iz srodnog izvornog zadatka.". Finim podešavanjem ovih modela za specifične zadatke moguće je postići značajne rezultate bez potrebe za velikim količinama podataka i oznaka. Modeli kao što su VGGish, temeljeni na VGG arhitekturi, i OpenL3, model za numeričku reprezentaciju zvuka, pokazali su izvrsne performanse u klasifikaciji glazbenih instrumenata.

4. Programsko ostvarenje sustava za klasifikaciju glazbenih instrumenata

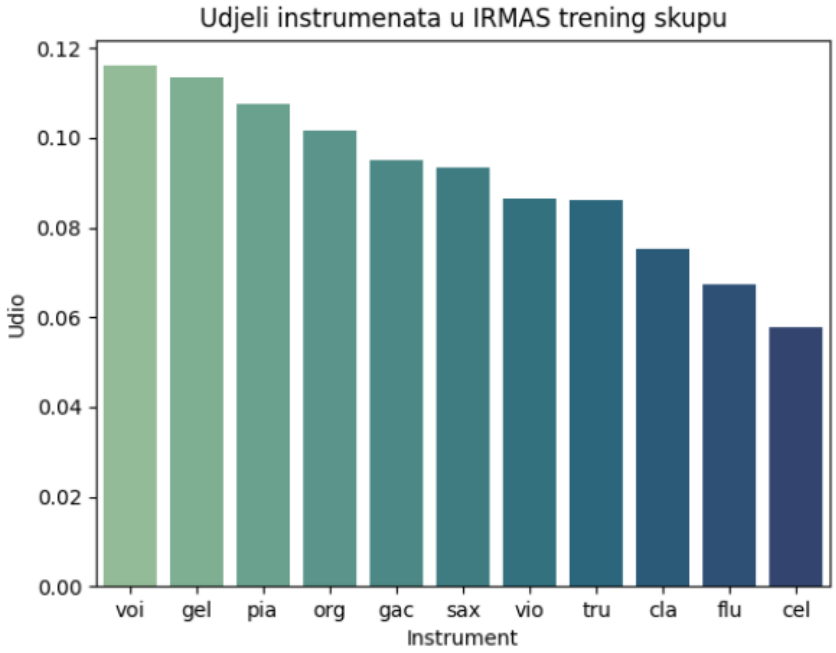
4.1. Pregled

Ovo poglavlje opisuje detalje implementacije sustava klasifikacije glazbenih instrumenata. Istraženi su koraci i korištene metodologije, od pripreme skupa podataka do izdvajanja značajki, arhitekture modela, postupaka obuke i evaluacije rezultata.

4.2. Skupovi podataka

Odabir i priprema skupova podataka ključni su za razvoj kvalitetnog sustava klasifikacije glazbenih instrumenata. U sklopu ovog diplomskog rada upotrijebljena su dva istaknuta skupa podataka: IRMAS i OpenMIC. Ovi skupovi podataka odabrani su kako bi se osigurala raznolika reprezentacija glazbenih instrumenata. Ovo potpoglavlje detaljno opisuje karakteristike svakog skupa podataka i objašnjava kako su korišteni za dobivanje jasnije i točnije slike performansi sustava.

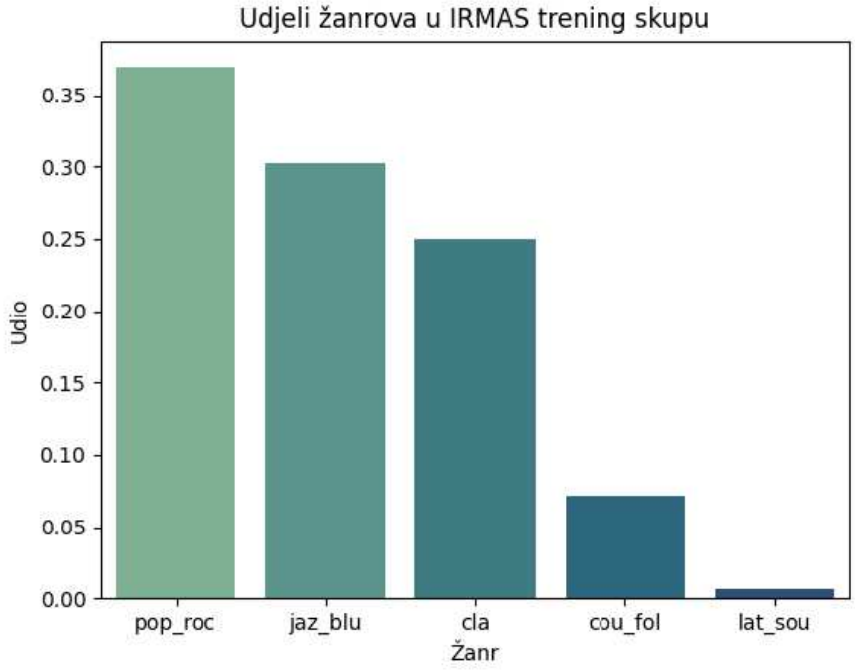
4.2.1. IRMAS skup podataka



Slika 7 - Razdioba glazbenih instrumenata u IRMAS skupu podataka

Skup podataka IRMAS (Instrument Recognition in Musical Audio Signals) dizajniran je posebno za zadatak prepoznavanja glazbenih instrumenata u polifonoj glazbi. Sadrži veliku

zbirku zvučnih snimaka označenih s prisutnošću raznih glazbenih instrumenata. Skup podataka sadrži zvučne zapise iz širokog raspona glazbenih žanrova, osiguravajući raznolikost u smislu instrumentacije i glazbenog stila. Skup podataka također sadrži i oznaku žanra, a sadrži klasiku, jazz, rock, pop i latino.



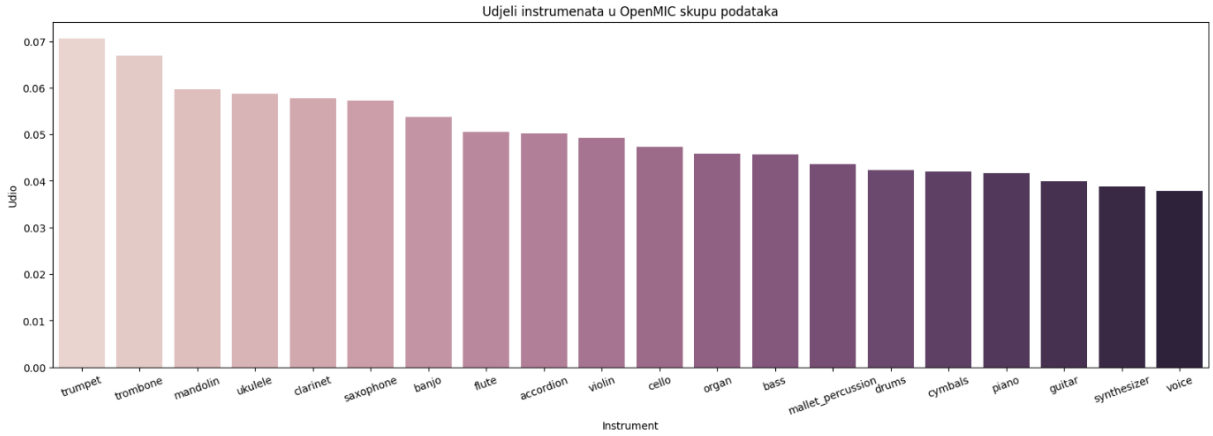
Slika 8 - Razdioba žanrova u IRMAS skupu podataka

Ova raznolikost osigurava da se model može generalizirati na različite glazbene stilove. Označeni instrumenti uključuju violončelo, klarinet, flautu, akustičnu gitaru, električnu gitaru, orgulje, klavir, saksofon, trubu, violinu i ljudski glas. Svaki zvučni zapis u skupu podataka sadrži oznaku dominantnog instrumenta koji svira za vrijeme trajanja zvučnog zapisa.Dodatno, svakom zvučnom zapisu pridružena je oznaka prisutnosti ili odsutnosti bubnjeva za vrijeme trajanja zapisa. Zvučni zapisi dostupni su u WAV formatu, s brzinom uzorkovanja od 44,1 kHz i 16-bitnom dubinom, što osigurava vjernost i točnost. Podatci su izvorno podijeljeni na podskup za treniranje i podskup za vrednovanje čime je olakšana obrada. Izvornu verziju ovog skupa podataka sastavio je Ferdinand Fuhrmann za potrebe svoje doktorske disertacije. Skup podataka kasnije je pročišćen i tako je nastao IRMAS skup podataka kakvog poznajemo danas. Skup podataka IRMAS korišten je kao primarni izvor za treniranje i vrednovanje sustava klasifikacije. Detaljne oznake i velika varijacija glazbenih

žanrova omogućili su sustavu uočavanje složenih uzoraka i prepoznavanje instrumenata u različitim glazbenim kontekstima.

4.2.2. OpenMIC skup podataka

Skup podataka OpenMIC dizajniran je za podršku klasifikaciji zapisa na kojima je prisutno više instrumenata i nudi drugačiju perspektivu uključivanjem izoliranih i miješanih zvučnih zapisa.Nastao je kao rezultat suradnje između Spotifyja, MARL-a Sveučilišta New York i Centra za podatkovnu znanost. Skup podataka sadrži 20 000 primjera glazbe s licencom Creative Commons dostupne na Free Music Archive. Svaki primjer je isječak od 10 sekundi koji je označen za prisutnost ili odsutnost 20 klasa instrumenata od strane anotatora na platformi za crowdsourcing.



Slika 9 - Razdioba glazbenih instrumenata u OpenMIC skupu podataka

Oznake su dane na razini vremenskog okvira, omogućujući preciznu identifikaciju kada svaki instrument svira unutar zapisa. Ovakva granularnost korisna je za modeliranje vremenske dinamike pojavljivanja instrumenata. Instrumenti u zvučnim zapisima svirani su različitim tehnikama i u različitim okruženjima što unosi dodatne varijacije u ovaj podatkovni skup. Prema Humphrey et al. (2018.), "OpenMIC nudi opsežan, žanrovski raznolik skup podataka koji odražava zamršenu prirodu polifone glazbe u stvarnom svijetu.". OpenMIC sadrži izolirane zapise (s jednim instrumentom) i polifone zapise (s više instrumenata). Ova dvojna priroda pomaže kod treniranja modela za uspješno djelovanje u jednostavnim i u složenim zvučnim scenarijima. Zvučni zapisi dostupni su u MP3 formatu, sa standardnom brzinom uzorkovanja od 44,1 kHz, čime se osigurava kompatibilnost s većinom alata za obradu

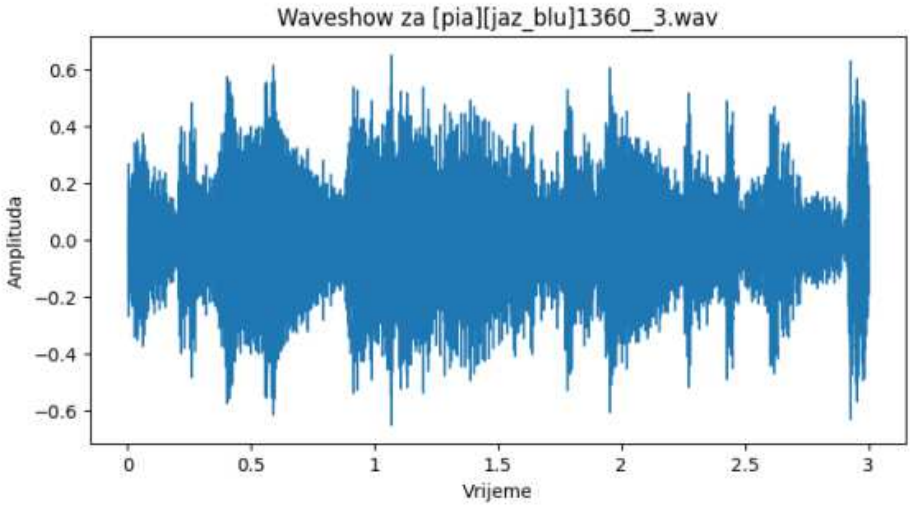
zvučnih signala. Skup podataka OpenMIC korišten je za treniranje i vrednovanje modela, kao i IRMAS skup podataka.

4.3. Generiranje značajki

U području klasifikacije glazbenih instrumenata korak generiranja značajki ključan je za transformaciju neobrađenih zvučnih zapisa u oblik koji se može učinkovito obraditi algoritmima strojnog učenja. Generiranje značajki podrazumijeva izdvajanje relevantnih karakteristika iz zvučnog signala koji opisuju bitne komponente glazbenog sadržaja. Ove značajke dizajnirane su za opisivanje različitih dimenzija zvuka, kao što su njegova spektralna, vremenska i harmonijska svojstva. Ovim postupkom, složena i visokodimenzionalna priroda sirovog zvuka koncentrirana je u kompaktniji i informativniji prikaz. Generiranje značajki čini poveznicu između sirovih podataka i modela, pružajući ulaze koji omogućuju postupke učenja i klasifikacije.

4.3.1. Spektrogram

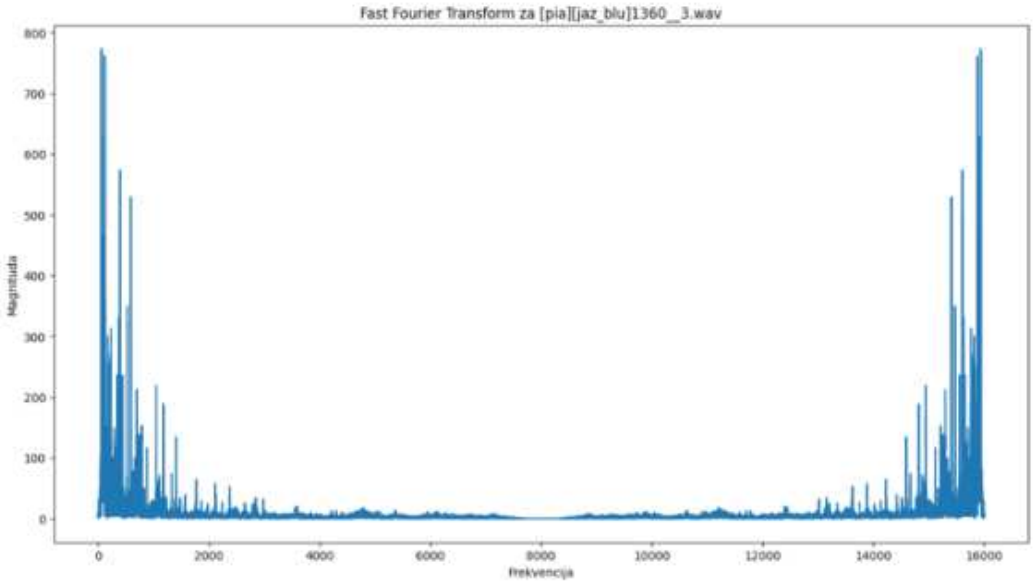
Standardni oblik zvučnog vala prikazuje oscilacije položaja čestica tijekom vremena, bilježeći vremenski razvoj zvučnog vala. Ovaj prikaz, iako koristan, može biti prilično složen, posebno kada se radi o zvučnim valovima koji nose zamršene glazbene informacije. Zvučni valovi povezani s glazbom posebno su složeni zbog svog bogatog harmonijskog sadržaja i dinamičkih promjena. Stoga je korisno predstaviti ovu zvučnu informaciju u obliku koji pruža uvid u frekvencijske komponente signala u bilo kojem trenutku.



Slika 10 - Zvučni val

U tradicionalnom prikazu u vremenskoj domeni, audio signal se vizualizira kao valni oblik amplitude koji se mijenja tijekom vremena. Ova metoda učinkovito pokazuje kako se zvuk razvija, ali ne daje informacije o frekvencijskom sadržaju signala. Slika 10 ilustrira tipičan valni oblik u vremenskoj domeni, gdje horizontalna os predstavlja vrijeme, a vertikalna os predstavlja amplitudu.

Da bismo razumjeli frekvencijske komponente složenog signala, koristimo se matematičkim procesom poznatim kao Fourierova transformacija (FT). Fourierova transformacija rastavlja zvučni val u određenom trenutku na zbroj jednostavnih sinusoidnih valova različitih frekvencija. Ova transformacija pomiče reprezentaciju zvučnog signala iz vremenske domene, koja se fokusira na njegovo širenje tijekom vremena, u frekvencijsku domenu, koja naglašava relativnu prisutnost čistih tonova na različitim frekvencijama (Slika 11).



Slika 11 - Rezultat Fourierove transformacije zvučnog vala

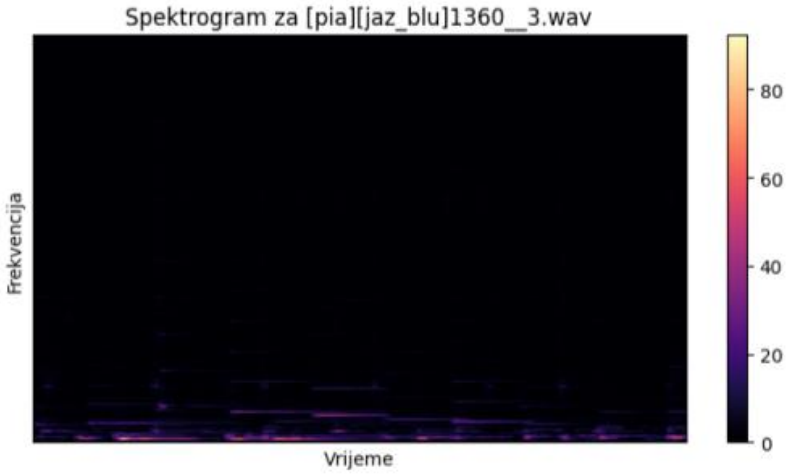
Iako frekvencijski spektar dobiven Fourierovom transformacijom pruža vrijedne informacije o frekvencijama prisutnim u zvučnom signalu, on implicitno pretpostavlja stacionarnost signala. To znači da ne uzima u obzir promjene tijekom vremena i analizira signal samo u jednom trenutku. Ovakav statički pogled neadekvatan je za analizu glazbe, gdje su vremenski razvoj, promjene u tonu, dinamici i tempu kardinalne karakteristike.

Kako bi se zadržale informacije o vremenskom razvoju i frekvencijskom sastavu zvučnog signala, koriste se spektrogrami. Spektrogrami su vizualni prikazi koji pokazuju kako se spektar frekvencija razvija tijekom vremena. Izračun spektrograma uključuje uzorkovanje audio signala tijekom vremena s odabranom gustoćom vremenskih odsječaka. Za svaki vremenski odsječak primjenjuje se Fourierova transformacija, proizvodeći ono što je poznato kao kratkotrajna Fourierova transformacija (STFT).

STFT pruža niz frekvencijskih spektara, od kojih svaki odgovara određenom vremenskom intervalu, čime se hvataju i vremenski i frekvencijski aspekti signala. Ishod ovog procesa je skup podataka koji zadržava informacije o frekvencijskom sastavu tijekom vremena. Razlučivost spektrograma značajno ovisi o izboru gustoće vremenskog odsječka i trajanju svakog prozora koji se koristi u STFT-u.

Ucrtavanjem dobivenih podataka tako da je vrijeme na horizontalnoj osi, frekvencija na vertikalnoj osi, a intenzitet (amplituda) svake frekvencije u određenom trenutku označen bojom, dobivamo spektrogram (Slika 12).

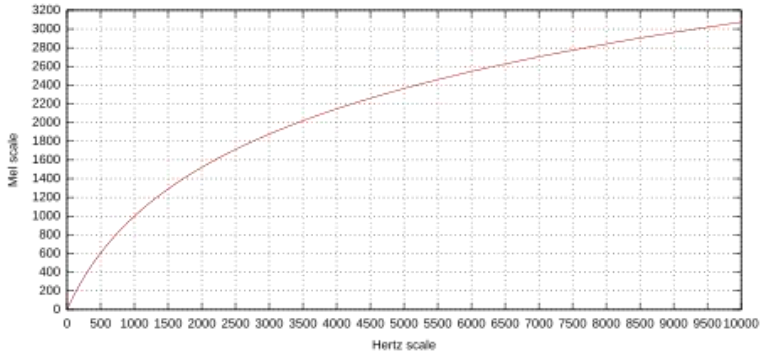
Spektrogrami su vrlo važni za analizu glazbe jer nude informativan pogled na promijene u frekvencijskom sadržaju signala tijekom vremena. Ovo omogućuje otkrivanje različitih glazbenih elemenata, kao što su note, akordi, ritmovi i dinamika, čineći spektrograme temeljnim alatom u procesu generiranja značajki za sustave klasifikacije glazbenih instrumenata.



Slika 12 - Spektrogram

4.3.2. Mel spektrogram

U kontekstu analize glazbenih zvukova, koji su specifičan podskup zvučnih signala strukturiranih da odgovaraju ljudskoj slušnoj percepciji, problem klasifikacije instrumenata idealno je prikladan za pristup temeljen na Mel ljestvici. Mel ljestvica je empirijski konstruirana ljestvica ljudske percepcije visine tona dizajnirana tako da jednake udaljenosti na ljestvici odgovaraju jednakim percipiranim razlikama u visini tona.

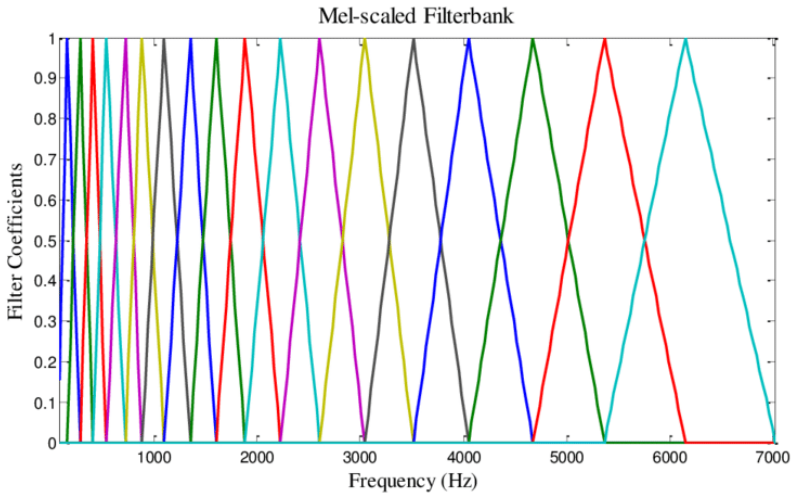


Slika 13 - Dijagram tona Mel ljestvice u odnosu na herc ljestvicu. Preuzeto s [https://en.wikipedia.org/wiki/Mel_scale]

Mel ljestvica dizajnirana je tako da odražava osjetljivost ljudskog uha na različite frekvencije. Za razliku od linearne frekvencijske ljestvice, koja jednako tretira sve frekvencije, Mel ljestvica je logaritamska, naglašavajući frekvencije koje su značajnije za ljudski sluh. Transformacija iz vrijednosti frekvencije (u Hercima) u vrijednosti Mel skale dana je sljedećom eksperimentalno određenom relacijom:

$$m(f) = 2595 \log(1 + \frac{f}{700})$$

Ova formula prevodi frekvencije u ljestvicu koja je više usklađena s ljudskom slušnom percepcijom. Slika 13 ilustrira odnos između vrijednosti frekvencije i njihovih odgovarajućih vrijednosti na skali mel.



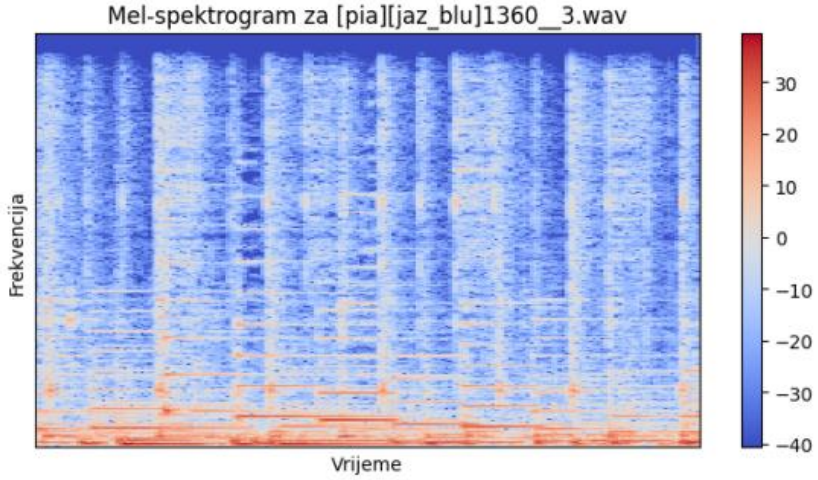
Slika 14 - Skup trokutastih filtera. Preuzeto s [https://www.researchgate.net/figure/Mel-filter-banks-basis-functions-using-20-Mel-filters-in-the-filter-bank_fig1_288632263]

Kako bismo izradili mel spektrogram, počinjemo sa standardnim spektrogramom i primjenjujemo skup filtara koji transformira frekvencijsku os u Mel ljestvicu. Ovaj postupak sastoji se od nekoliko ključnih koraka:

1. Kao što je opisano u prethodnom potpoglavlju, izračunavamo STFT audio signala kako bismo generirali spektrogram, koji predstavlja distribuciju frekvencija tijekom vremena.
2. Skup Mel filtara sastoji se od niza trokutastih filtara raspoređenih prema Mel ljestvici. Ti su filtri raspoređeni po frekvencijskom spektru na način koji odražava odnos između frekvencije i Mel ljestvice (Slika 14). Dizajn ovih filtara uzima u obzir osjetljivost ljudskog uha, s više filtara koncentriranih u nižem frekvencijskom rasponu gdje je ljudski sluh oštriji.
3. Svaki vremenski okvir spektrograma prolazi kroz skup Mel filtara. Ovaj postupak naglašava frekvencije prema Mel skali, transformirajući spektrogram u Mel spektrogram. Rezultat je vremensko-frekvencijski prikaz koji je bolje usklađen s ljudskom slušnom percepcijom.

Korištenje mel spektrograma posebno je povoljno za klasifikaciju glazbenih instrumenata iz nekoliko razloga. Naglašavanjem frekvencijskih komponenti koje ljudsko uho može percipirati Mel spektrogrami učinkovitije prikazuju nijanse glazbenih zvukova. Logan (2000.) navodi: "Mel spektrogrami su posebno učinkoviti u hvatanju suptilnosti u glazbenim zvukovima, što je ključno za razlikovanje instrumenata sličnog zvuka.". Dodatno, Mel spektrogrami pružaju

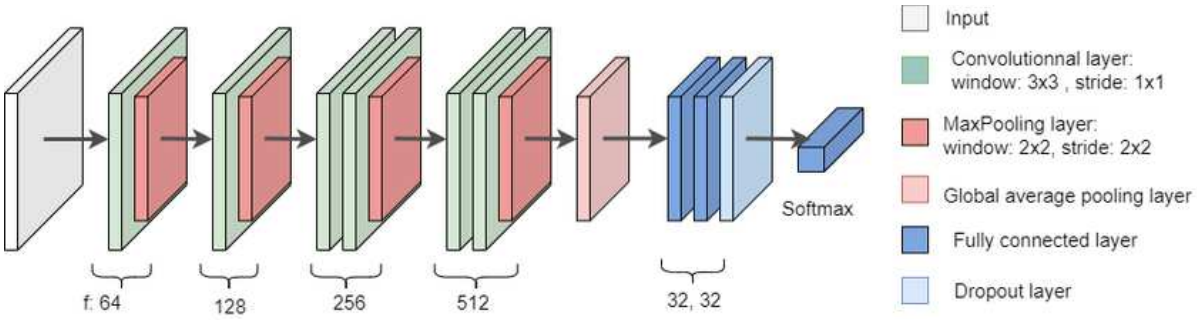
smislen prikaz zvučnog signala za modele strojnog učenja, budući da sadrže važne karakteristike zvuka koje su relevantne za ljudsku percepciju. Većina glazbenog sadržaja nalazi se unutar frekvencijskog raspona gdje je ljudsko uho najosjetljivije. Mel spektrogrami dizajnirani su za točnije hvatanje ovog raspona, što ih čini idealnim za analizu i klasifikaciju glazbenih instrumenata. Slika 15 prikazuje primjer Mel spektrograma, koji prikazuje kako se intenzitet različitih frekvencija razvija tijekom vremena, transformiran prema mel skali.



Slika 15 - Mel spektrogram

4.3.3. VGGish

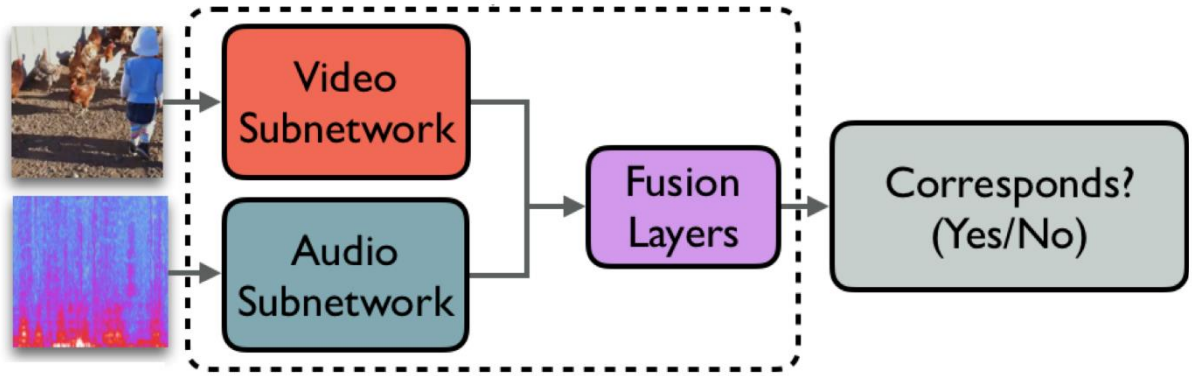
VGGish neuronska je mreža prilagođena zadatku zvučne analize predtrenirana na istome AudioSet skupu podataka koji se sastoji od 632 različitih zvučnih događaja odnosno razreda i više od 2 milijuna primjera. Mreža je zasnovana na prilagodbi karakteristične VGG konvolucijske arhitekture izvorno razvijenoj za analizu slika. Budući da je moguće izraditi slikovnu reprezentaciju zvučnog signala, ova arhitektura preinačena je kako bi bolje odgovarala slikama koje prikazuju vremensko frekvencijski sadržaj zvučnih signala kao što su spektrogrami ili MFCC. VGGish mreža konstruirana je s ciljem izdvajanja kompaktnih vektora značajki višega reda koji reprezentiraju spektralni sadržaj zvučnoga isječka. Glavni dio mreže prikazan na slici 16 sastoji se od 7 konvolucijskih slojeva, nakon čega slijede 2 potpuno povezana sloja. Prolaskom kroz više slojeva konvolucije, sažimanja te normalizacije grupe, dobivaju se konačni vektori značajki koji sadrže informacije o cjelokupnome zvučnom signalu. Ovim se postupkom realizira model s visokom prediktivnom moći za podatke iz različitih domena zvučne analize, ali i podatke različite kvalitete i tehnologije snimanja.



Slika 16 - Shematski prikaz VGGish arhitekture. Preuzeto s [https://www.researchgate.net/figure/VGGish-based-audio-baseline-models-architecture-The-size-of-the-feature-maps-f-of_fig4_374622903]

U sklopu ovog diplomskog rada VGGish mreža korištena je za generiranje vektora značajki iz Mel spektrograma koji su dobiveni obradom zvučnih signala.

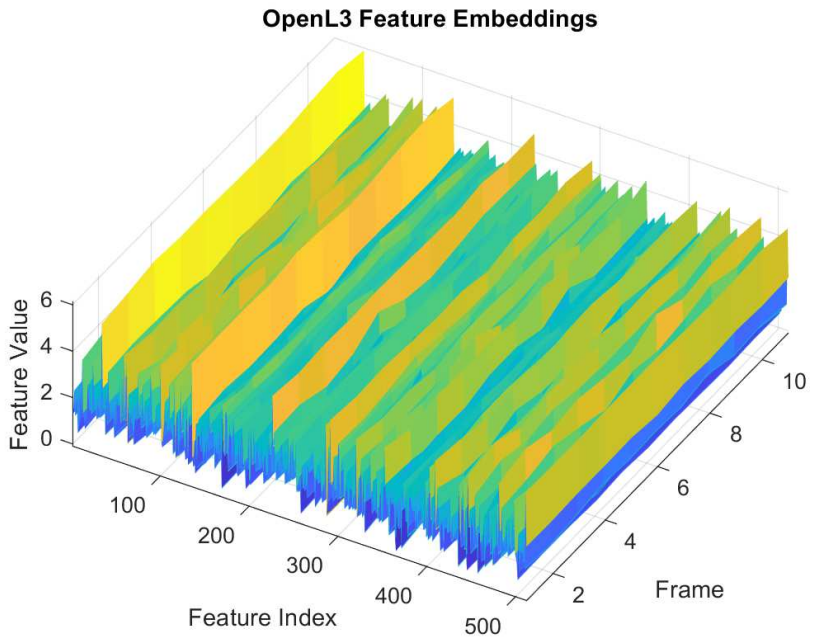
4.3.4. OpenL3



Slika 17 - Prikaz rada OpenL3 arhitekture. Preuzeto s [https://www.justinsalamon.com/news/openl3-competitive-open-deep-audio-embedding]

OpenL3 suvremeni je model neuronske mreže dizajniran za zadatak analize zvuka, s posebnim naglaskom na generiranje informativnih vektora značajki iz zvučnih zapisa. Model je predtreniran na velikom AudioSet skupu podataka,baš kao i VGGish model. OpenL3 koristi arhitekturu duboke konvolucijske neuronske mreže za obradu i analizu zvučnih signala, izdvajajući značajne prikaze značajki koji opisuju karakteristike zvuka. OpenL3 se temelji na modelu vektorskih reprezentacija Gledaj (Look),Slušaj (Listen), i Uči (Learn) (L3), koji integrira zvučne i vizualne informacije za učenje robusnih vektorskih reprezentacija. Za razliku od tradicionalnih modela koji se oslanjaju isključivo na zvučne signale, OpenL3 koristi multimodalni pristup treniranju,tako što u svojem postupku treniranja razmatra i zvuk i video te za cilj ima odgovoriti pripada li zadani zvuk zadanome videu.Arhitektura modela prikazana

je na slici 17 Ovakav pristup omogućuje OpenL3 modelu da vrlo dobro generalizira u raznim zadatcima analize zvuka. Mreža započinje s nizom konvolucijskih slojeva dizajniranih za obradu ulaznih spektrograma. Nakon konvolucijskih slojeva, slojevi sažimanja koriste se za smanjenje prostornih dimenzija vektora značajki. Nakon konvolucijskih slojeva i slojeva sažimanja, vektori značajki se spljoštavaju i prolaze kroz potpuno povezane slojeve koji generiraju konačan vektor značajki.



Slika 18 - Polje značajki OpenL3 modela. Preuzeto s [https://www.mathworks.com/help/audio/ref/openl3embeddings.html]

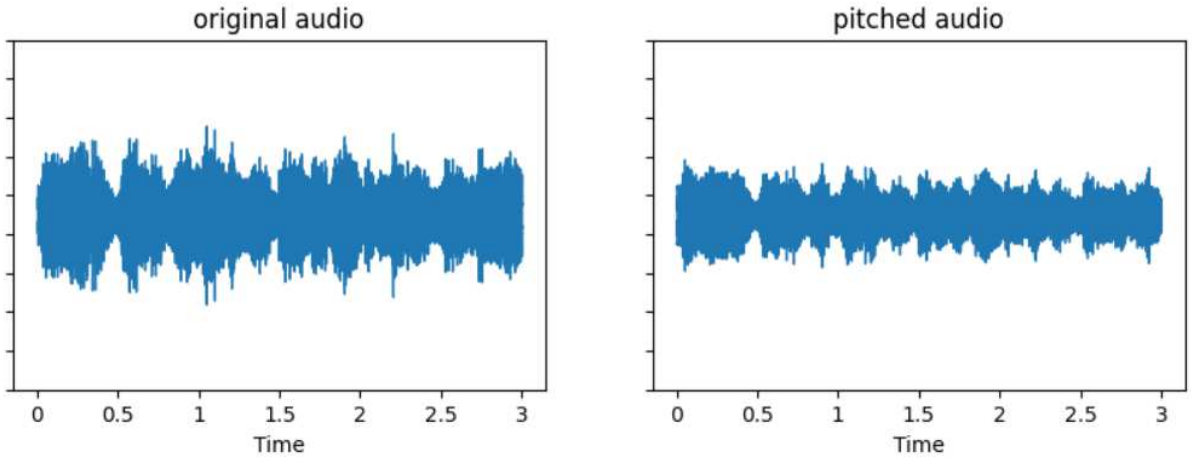
U sklopu ovog diplomskog rada OpenL3 mreža korištena je za generiranje vektora značajki iz Mel spektrograma koji su dobiveni obradom zvučnih signala, ali i za usporedbu s VGGish mrežom. Primjer značajki koje generira OpenL3 model prikazane su na slici 18.

4.4. Podatkovne augmentacije

Augmentacija podataka postupak je transformacije i variranja elemenata unutar postojećeg skupa podataka kako bi se proširio i obogatio u svrhu treniranja modela. Temeljna ideja je uvesti relativno nove, dodatne podatke kroz transformaciju postojećih, čime se postiže ne samo povećanje količine, već i raznolikost skupa za treniranje. Kao rezultat, dobiveni

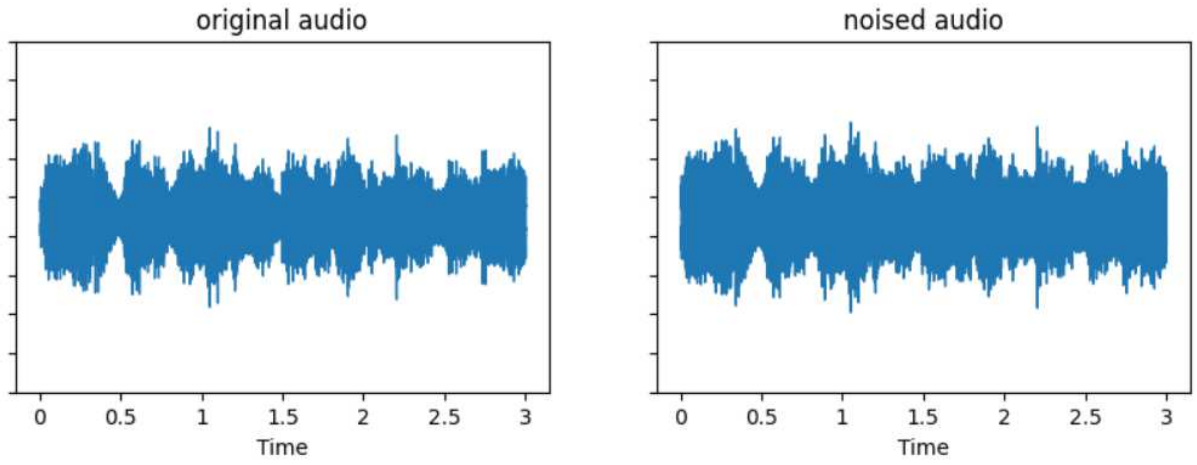
računalni modeli postaju precizniji, robusniji i bolje prilagođeni za točnu obradu različitih varijacaija podataka.

Jedna od korištenih metoda augmentacije je promjena visine tona. Ova tehnika simulira promjenu glazbene note, omogućujući modelu prepoznavanje glazbenih instrumenata u raznovrsnim glazbenim kontekstima. Prikaz ove transformacije dan je na slici 19. Mijenjajući visinu tona, modelu se omogućuje učenje identificiranja instrumenta bez obzira na note koje sviraju, poboljšavajući sposobnost generalizacije. Ova je metoda osobito korisna u scenarijima gdje se isti instrument može svirati u različitim tonalitetima, osiguravajući da model ne postane pristran prema određenoj visini.



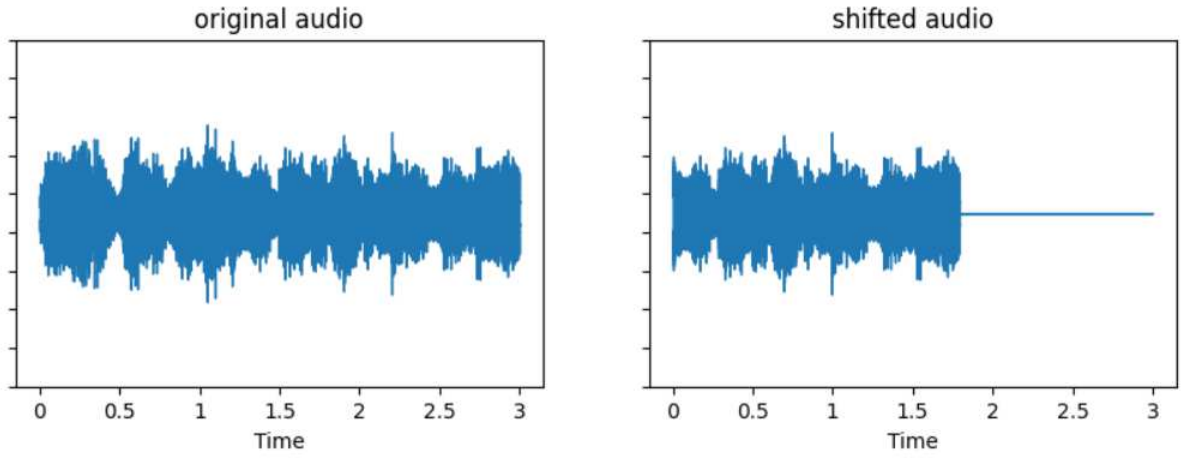
Slika 19 - Promjena visine tona

Dodavanje šuma podrazumijeva umetanje šuma u zvučne zapise. Najčešći oblik šuma koji se umeće u podatke je Gaussov bijeli šum, ali postoje i druge varijante poput ružićastog šuma ili crvenog šuma. Dodavanje šuma pomaže modelu da postane otporniji na varijacije u kvaliteti zvuka i različitosti u uvjetima snimanja. Ova tehnika je korisna pri treniranju modela za obradu zvučnih zapisa iz stvarnog svijeta, koji često sadrže pozadinsku buku i druge nesavršenosti. Izlaganjem modela zašumljenim podacima, on postaje bolji u razlikovanju ciljanih zvukova, robusniji i smanjuje se rizik od pretjeranog učenja na čistim, nezašumljenim podacima. Primjer zašumljivanja zvučnog signala prikazan je na slici 20.



Slika 20 - Zašumljivanje zvučnog signala

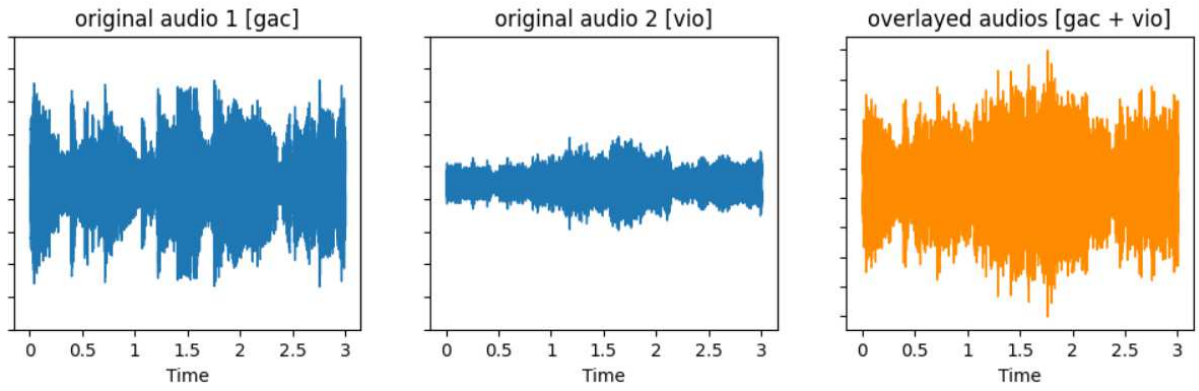
Vremenski pomak uključuje pomicanje audio zapisa unaprijed ili unatrag u vremenu. Ova metoda sprječava pretjerano prilagođavanje modela na određene vremenske položaje instrumenata unutar glazbenih skladbi. Uvođenjem vremenskih pomaka, model uči prepoznavati instrumente bez obzira na njihov položaj u vremenu. Ova augmentacija posebno je korisna u polifonoj glazbi, gdje više instrumenata može ući ili izaći iz skladbe u različito vrijeme, a slika 21 prikazuje kako ova transformacija djeluje. Vremenski pomaci pomažu modelu da postane fleksibilniji i sposobniji za rukovanje različitim vremenskim strukturama u glazbenim skladbama.



Slika 21 - Vremensko pomicanje zvučnog signala

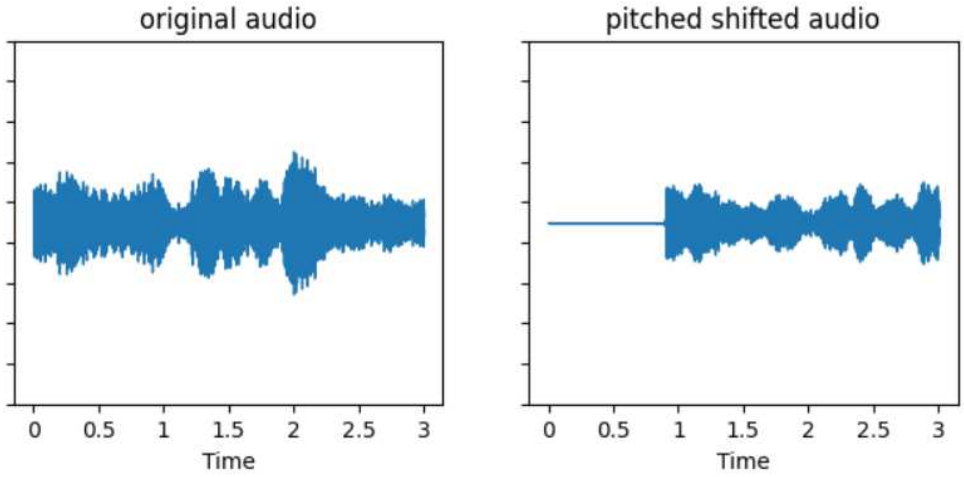
Tehnika preklapanja kombinira dva ili više zvučnih zapisa njihovim uprosječivanjem, stvarajući učinak istodobne reprodukcije. Ova je metoda izazovna jer nepravilno kombinirani zvukovi mogu rezultirati isječcima koji nisu glazbeno koherentni. Za učinkovito preklapanje, snimke se

moraju podudarati u tempu, glasnoći, dinamici i tonu. Kada se pravilno izvede, ova tehnika generira vrlo raznolike glazbene isječke koji obogaćuju skup podataka za treniranje. Međutim, pravilna implementacija zahtijeva značajnu količinu podataka i precizno usklađivanje pojedinih elemenata. Unatoč svojim izazovima, tehnika preklapanja može proizvesti jedinstvene i raznolike primjere za treniranje koji se mogu smatrati novima u vidu analize glazbe kao što je prikazano na slici 22.



Slika 22 - Preklapanje dvaju zvučnih signala

Kombiniranje metoda augmentacije može oponašati prirodne varijacije u glazbenim isječcima koje proizlaze iz različitih procesa snimanja i karakteristika izvedbe, a primjer kombiniranja transformacija prikazan je na slici 23. Na primjer, kombiniranjem pomaka visine tona i vremenskog pomaka mogu se stvoriti podaci koji odražavaju i tonske i vremenske varijacije. Ove kombinacije pomažu modelu da nauči identificirati instrumente u širokom rasponu uvjeta, povećavajući njegovu robusnost i općenito poboljšavajući njegove performanse.



Slika 23 - Primjer kombinacije dviju transformacija

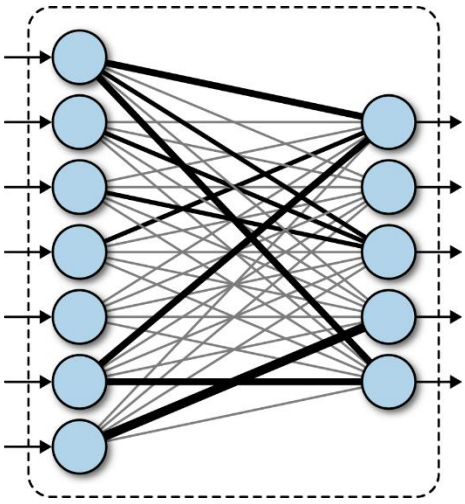
Učinkovitost navedenih metoda augmentacije ispitana je treniranjem i vrednovanjem različitih klasifikatora glazbenih instrumenata. U svojim izvornim oblicima ove metode nisu značajno poboljšale točnost identifikacije instrumenata. Iz ovog razloga, općenito su bile isključene iz cijevovoda konačnog modela kako bi se održala jednostavnost. Međutim, kroz iterativno usavršavanje, kombinacije ovih metoda pokazale su korisnim za razvoj robusnog i preciznog konačnog modela. Ovaj kompozitni pristup augmentaciji, koji spaja elemente pojedinačnih metoda, pridonio je uspjehu konačnog sustava klasifikacije povećavajući njegovu sposobnost generalizacije i točne klasifikacije širokog spektra glazbenih instrumenata.

4.5. Metodologija

Metodologija sustava klasifikacije glazbenih instrumenata obuhvaća nekoliko faza, počevši od obrade zvučnih zapisa do izdvajanja značajki i na kraju klasifikacije. Ovo potpoglavlje detaljno opisuje cijeli process klasifikacije zvučnih zapisa, ističući korištenje VGGish arhitekture kao ekstraktora značajki ključnih za ovaj postupak.

Početni korak u postupku podrazumijeva obradu neobrađenih zvučnih datoteka kako bi se pretvorile u prikladan format za izdvajanje značajki. To se postiže pomoću knjižnice Librosa u Pythonu, koja pruža robusne alate za analizu i obradu zvuka. Ova knjižnica nudi gotova programska ostvarenja za većinu standardnih postupaka u domeni obrade zvučnih signala. Nakon učitavanja zvučnog zapisa u program, koristeći Librosa knjižnicu, ovaj se zvučni zapis ponovno uzorkuje kako bi se osigurala uniformnost i konzistentnost. Nakon što je zvučni zapis učitan i ponovno uzorkovan, pretvara se u Mel spectrogram. Mel spektrogram služi kao ulaz u fazu ekstrakcije značajki pomoću VGGish mreže.

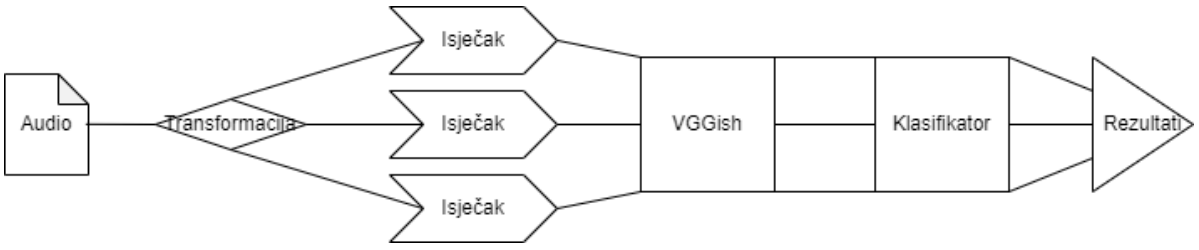
VGGish mreža obrađuje Mel spektrogram u segmentima od 96 milisekundi. Svaki segment se tretira neovisno kako bi se izdvojile lokalizirane značajke, odražavajući sposobnost mreže da uhvati sitne detalje i u vremenskoj domeni. Segmenti Mel spektrograma prolaze kroz više konvolucijskih slojeva. Ovi slojevi primjenjuju filtre za otkrivanje uzoraka kao što su rubovi, teksture i druge složene značajke. Slojevi sažimanja smanjuju prostorne dimenzije vektora značajki, koncentriraju informacije i smanju računsku složenost. Ovi se vektori značajki potom ulaze u potpuno povezani sloj koji na svojem izlazu daju konačnu vektorsku reprezentaciju zvučnog signala koja je pogodna za ulaz klasifikacijskih algoritama.



Slika 24 - Potpuno povezana neuronska mreža. Preuzeto s [https://www.oreilly.com/library/view/tensorflow-for-deep/9781491980446/ch04.html]

Posljednji korak ovog postupka je sama klasifikacija zvučnih zapisa ovisno o tome koji instrument je prisutan u zapisu. U sklopu ovog diplomskog rada odabrana je potpuno povezana neuronska mreža. Primjer ovog modela prikazan je na slici 24. Ovi modeli imaju sposobnost naprednog prepoznavanja različitih uzoraka u vektorima značajki što ih čini odličnim izborom u ovakvom okruženju. Potpuno povezana neuronska mreža na svoj ulaz prima vektore značajki isječaka izvornog zvučnog zapisa, a na izlazu daje vjerojatnosti prisutnosti svakog od instrumenata u tom isječku. Konačni rezultat klasifikacije dobije tako da se za svaki instrument uzme najveća vjerojatnost koja se pojavila u svim segmentima izvornog zvučnog zapisa. Kada za svaki instrument postoji vjerojatnost njegove prisutnosti odabere se granična mjera i ako je vjerojatnost veća od granične mjere taj se instrument smatra prisutnim u zvučnom zapisu, inače je odsutan.

Sljedeća shema ilustrira rad sustava, od neobrađenih zvučnih zapisa do konačnih klasifikacijskih odluka:



Slika 25 - Shematski prikaz rada klasifikacijskog sustava

4.6. Postupak treniranja

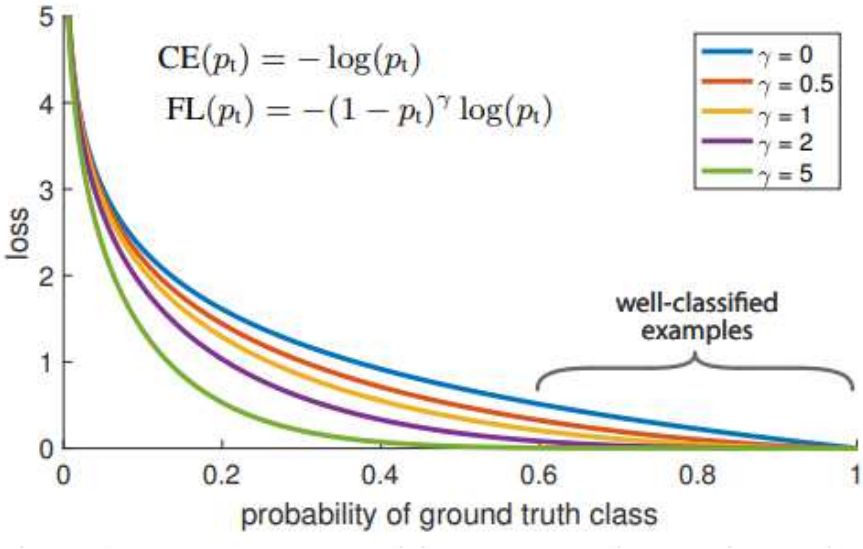
Postupak treniranja kritičan je korak u razvoju točnog i robusnog sustava klasifikacije glazbenih instrumenata. Ovaj postupak uključuje inicijalizaciju VGGish mreže, pažljivo upravljanje podacima, odabir prikladne funkcije gubitka i optimizaciju samog modela.

Proces obuke počinje inicijalizacijom VGGish mreže, koja je prethodno obučena na skupu podataka AudioSet. Inicijalizacija podrazumijeva učitavanje predtreniranih težina u mrežu. Ove težine služe kao početna točka, dajući modelu bogat skup naučenih značajki koje su korisne za općenite zadatke analize zvuka. Korištenje ovih prethodno istreniranih težina pomaže u postizanju boljih performansi i brže konvergencije tijekom postupka treniranja.

Jedan od izazova u treniranju modela klasifikacije je suočavanje s neuravnoteženim klasama, gdje neke klase imaju znatno više primjera od drugih. Kako bi se riješio ovaj problem, podaci unutar svake mini grupe odabrani su tako da imaju jednak broj primjera za svaki instrument. Ovakav uravnoteženi odabir osigurava da mreža dobije dovoljno podataka o klasama s manje primjera, sprječavajući da model bude pristran prema klasama s obilnijim brojem podataka. Kako bi se poboljšao postupak učenja, posebno za primjere koje je teško klasificirati, koristi se žarišni gubitak (Focal Cross Entropy Loss). Žarišni gubitak je modifikacija standardnog gubitka unakrsne entropije koji dodaje modulacijski faktor gubitku unakrsne entropije, naglašavajući učenje na teškim primjerima. Lin et al. (2017.) opisuju žarišni gubitak kao "dinamički skaliran gubitak unakrsne entropije, gdje faktor skaliranja opada na nulu kako se povećava povjerenje u ispravnu klasu, učinkovito usredotočavajući učenje na teške primjere.".Žarišni gubitak se definira kao:

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma}\log(p_t)$$

gdje p_t označava modelovu procjenu vjerojatnosti neke klase, α_t označava težinski faktor za neuravnoteženost razreda, a γ označava parametar usredotočavanja koji prilagođava magnitudu kojom se jednostavni primjeri ponderiraju.



Slika 26 - Žarišni gubitak

Žarišni gubitak zapravo smanjuje doprinos gubitku od lakih primjera i povećava interval u kojem primjer ima veliki gubitak. Ovo potiče model da se više usredotoči na teške primjere, što dovodi do bolje generalizacije i poboljšanih performansi u izazovnim zadacima klasifikacije. Slika 26 prikazuje kako funkcija gubitka varira u ovisnosti o navedenim parametrima.

U početku su sve težine VGGish mreže zamrznute, a treniraju se samo potpuno povezani slojevi. Ovaj je korak ključan za omogućavanje potpuno povezanim slojevima da nauče smislene prikaze bez ometanja predtreniranih konvolucijskih slojeva. Tijekom ove faze koristi se veća stopa učenja za brzo podešavanje težine potpuno povezanih slojeva, čineći ih korisnima i točnima za zadatak klasifikacije. Nakon nekoliko epoha treniranja potpuno povezanih slojeva, posljednji konvolucijski sloj VGGish mreže je odmrznut. U ovoj fazi, cijeli sustav se trenira uz manju stopu učenja. Ovaj postupak finog podešavanja omogućuje modelu da prilagodi slojeve ekstrakcije značajki kao odgovor na specifične karakteristike skupa podataka, dodatno poboljšavajući kapacitet i performanse mreže. Razlog zašto se cijela VGGish mreža ne odmrzava od samog početka ili zašto se ne odmrzavaju svi konvolucijski slojevi je ublažavanje rizika od pretjeranog treniranja (overfitting), do kojeg može doći kada model postane previše naučen na podatke za treniranje.

Optimizacijski algoritam korišten u ovom postupku treniranja je AdamW, varijanta standardnog Adam algoritma. AdamW uključuje smanjenje težine izravno u postupak optimizacije, što pomaže u sprječavanju pretjeranog treniranja kažnjavanjem velikih težina. Standardni Adam optimizator ažurira parametre na temelju prvog i drugog momenta gradijenata, koristeći sljedeće jednadžbe:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

$$\widehat{m}_t = \frac{m_t}{1 - \beta_1^t}$$

$$\widehat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

$$\theta_{t+1} = \theta_t - \delta \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \varepsilon}$$

gdje m_t označava procjenu prvog momenta, v_t označava procjenu drugog momenta, g_t označava gradijent u trenutku t, β_1 i β_2 su hiperparametri koji kontroliraju stopu opadanja procjena momenta, δ označava stopu učenja, a ε je mala konstanta za numeričku stabilnost.

AdamW modificira ovaj pristup odvajanjem izraza opadanja težina od ažuriranja gradijenta, što dovodi do stabilnije i učinkovitije regularizacije. :

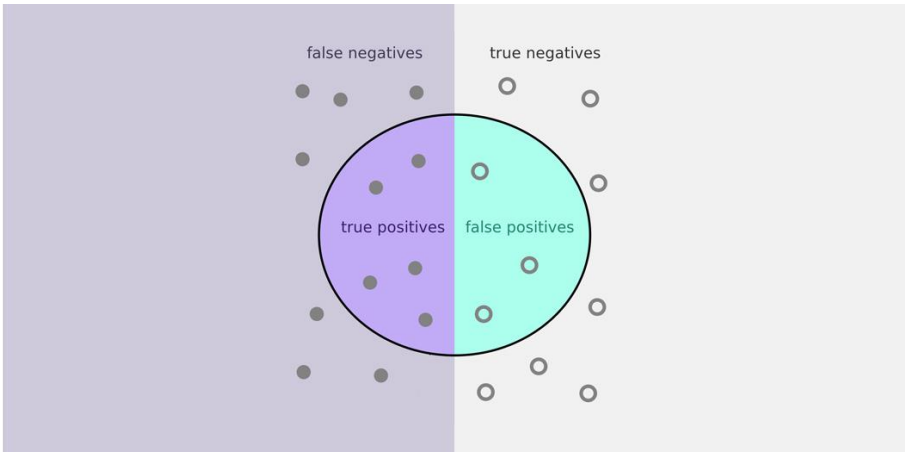
$$\theta_{t+1} = \theta_t - \delta \left(\frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \varepsilon} + \sigma \theta_t \right)$$

gdje σ označava koeficijent opadanja težina. Ova modifikacija osigurava da se opadanje težina primjenjuje na način koji je u skladu s L2 regularizacijom, osiguravajući bolju kontrolu nad složenošću modela i poboljšavajući generalizaciju. Loshchilov i Hutter (2019) navode: "Razdvajanjem opadanja težine od ažuriranja temeljenih na gradijentu, AdamW postiže bolja svojstva konvergencije i poboljšanu generalizaciju u usporedbi s tradicionalnim Adamom".

4.7. Vrednovanje i rezultati

4.7.1. Metrike

Odabir odgovarajuće metrike ključan je za razumijevanje i procjenu sustava klasifikacije, budući da pruža uvid u performanse sustava i pomaže identificirati područja za poboljšanje. U ovom dijelu objašnjene su tri ključne metrike: preciznost, odziv i točnost. Ove metrike nude različite perspektive o performansama sustava i zajedno daju precizniju procjenu valjanosti klasifikacijskog sustava. Slika 27 je ilustracija tipičnog klasifikacijskog okuženja.



Slika 27 - Tipičan primjer klasifikacijskog okruženja. Preuzeto s [https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/]

Preciznost, također poznata kao pozitivna prediktivna vrijednost, mjeri udio stvarnih pozitivnih predviđanja u odnosu na sva pozitivna predviđanja modela. Označava koliko je predviđenih pozitivnih instanci zapravo točnih. Visoka preciznost znači da model ima nisku stopu lažno pozitivnih rezultata. Izraz kojim se računa preciznost je:

$$Preciznost = \frac{TP}{TP + FP}$$

gdje TP označava broj stvarnih pozitivnih predviđanja, a FP označava broj lažnih pozitivnih predviđanja. Preciznost je osobito važna u kontekstima gdje je cijena lažno pozitivnih rezultata visoka, kao što je medicinska dijagnoza ili otkrivanje prijevare (Sokolova & Lapalme, 2009.).

Odziv, također poznat kao osjetljivost ili stvarna pozitivna stopa, mjeri udio stvarnih pozitivnih predviđanja u odnosu na sve pozitivne slučajeve. Odražava sposobnost modela da

ispravno identificira sve pozitivne instance. Visoki odziv znači da model ima nisku stopu lažno negativnih rezultata. Matematički izraz za odziv je:

$$Odziv = \frac{TP}{TP + FN}$$

gdje TP označava broj stvarnih pozitivnih predviđanja, a FN označava broj lažnih negativnih predviđanja. Odziv je ključan u situacijama gdje je pogreška na pozitivnim primjerima posebno skupa, kao što je prepoznavanje bolesti (Saito & Rehmsmeier, 2015.).

Točnost mjeri udio točnih predviđanja (i pravih pozitivnih i pravih negativnih) u ukupnom broju predviđanja. Pruža ukupnu procjenu performansi modela, ali može dovesti do pogrešnih zaključaka u slučajevima neuravnoteženosti razreda unutar skupova podataka. Izraz za točnost je:

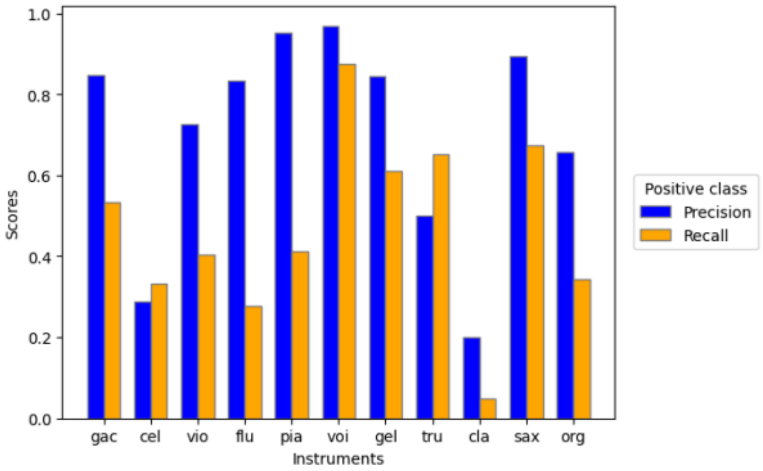
$$Točnost = \frac{TP + TN}{TP + FP + TN + FN}$$

gdje TP označava broj stvarnih pozitivnih predviđanja, FP označava broj lažnih pozitivnih predviđanja, TN označava broj stvarnih negativnih predviđanja, a FN broj lažnih negativnih predviđanja. Točnost daje opću ideju o izvedbi modela, ali sama po sebi možda neće biti dovoljna za procjenu modela koji se bave vrlo neuravnoteženim skupovima podataka (Jeni et al., 2013.).

Ove metrike zajedno daju informativnu procjenu izvedbe sustava klasifikacije. Svaka od metrika ističe različite aspekte klasifikacijskog problema kao što su točnost negativnih predviđanja, sposobnost identificiranja svih pozitivnih slučajeva i ukupna ispravnost predviđanja što ih zajedno čini vrlo učinkovitima u zadatcima procjene performansi klasifikacijskih modela.

4.7.2. Rezultati na IRMAS skupu podataka

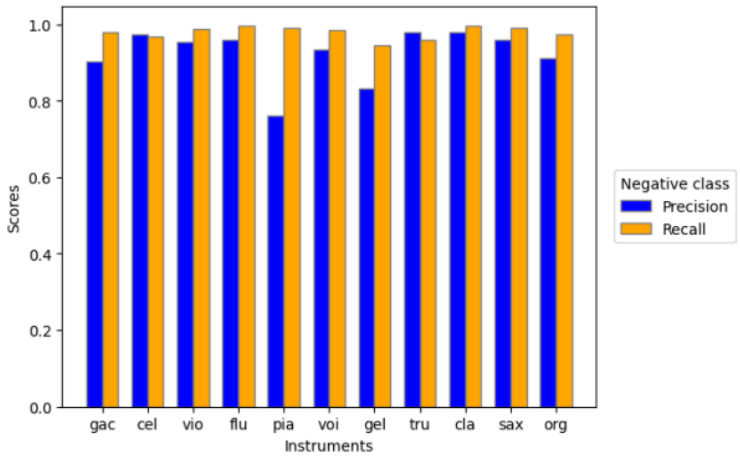
Kao referentna točka za procjenu performansi modela iskorišten je ranije opisani IRMAS skup podataka. Na IRMAS skupu podataka, konačni klasifikacijski sustav imao je sljedeće performanse u određivanju pozitivne klase:



Slika 28 - Metrike za pozitivnu oznaku na IRMAS skupu podataka

Rezultati ukazuju na činjenicu da se sustav lakše nosio s razredima za koje postoji veći broj primjera u skupu za učenje. Primjeri ovih razreda su ljudski glas (razred voi) i saksofon (razred sax). Dodatno, rezultati ukuzuju na gore performanse modela kod razreda s manjim brojem primjera kao što su clarinet (razred cla) i violončelo (razred cel). Ovakvi uzorci u rezultatima su očekivani s obzirom na to da većina modela uvelike ovisi o količini podataka i upravo količina podataka najviše korespondira s visokom točnošću. Zanimljivo je za primjetiti da većina razreda prati trend u kojemu je rezultantna preciznost veća od rezultantnog odziva. Ovo znači da je model razmjerno konzervativan pri pripisivanju pozitivne oznake nekom određenom primjeru. Konzervativnošću modela direktno je moguće upravljati mijenjanjem hiperparametara funkcije gubitka i optimizacijskog algoritma pa je moguće dobiti i drugačije rezultate ovisno o odabiru ovih hiperparametara.

Za negativnu klasu situacija je znatno drugačija:



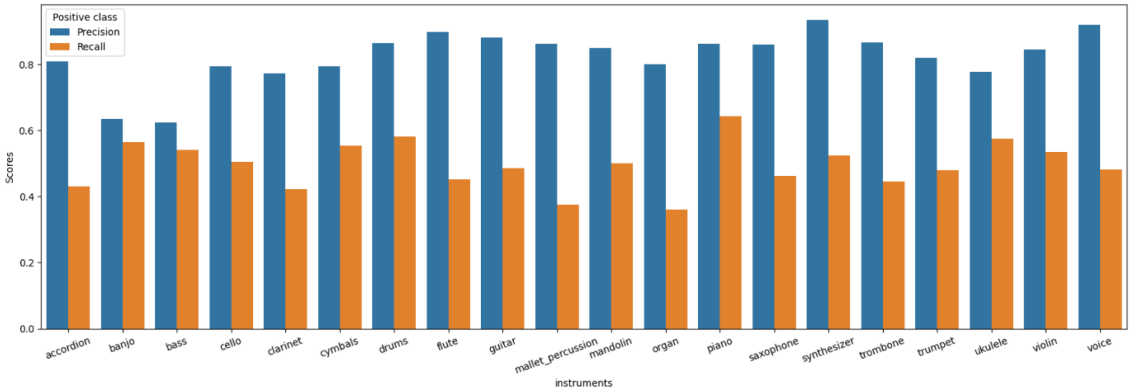
Slika 29 - Metrike za negativnu oznaku u IRMAS skupu podataka

Rezultati pokazuju da sustav s visokom točnošću pripisuje negativnu oznaku primjerima svih razreda. Kod negativne oznake, trend odnosa preciznosti i odziva drugačiji je nego kod pozitivne oznake. Dok je u situaciji pozitivne oznake preciznost bila veća od odziva s većinom razreda, u situaciji negativne oznake događa se obratno i uglavnom je odziv veća od dvaju mjera. Ovo znači da je model manje konzervativan pri pripisivanju negativne oznake nekom određenom primjeru.

Ukupna točnost sustava iznosi 91.78% što implicira da sveukupno gledano sustav ima relativno visoku mogućnost točnog raspoznavanja instrumenata u zvučnim zapisima.

4.7.3. Rezultati na OpenMIC skupu podataka

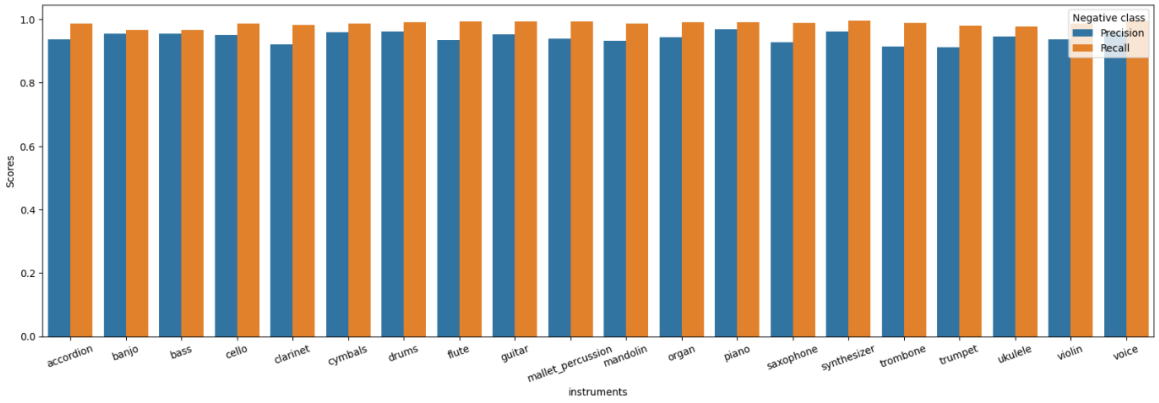
Kao dodatna referentna točka upotrijebljen je OpenMIC skup podataka kako bi se dodatno utvrdile performanse modela. Klasifikacijski sustav imao je sljedeće metrike za pozitivnu klasu:



Slika 30 - Metrike za pozitivnu oznaku u OpenMIC skupu podataka

Slično kao kod IRMAS skupa podataka, rezultati demonstriraju da sustav bolje klasificira one razrede za koje postoji veći broj primjera poput bendže (razred banjo) ili klavira (razred piano). Sustav ima slabije performanse kod primjera s instrumentima koji su manje zastupljeni. Ovo je vidljivo u metrikama razreda kao što su orgulje (razred organ) i harmonika (razred accordion). Kao što je slučaj i kod IRMAS skupa podataka, rezultati prate trend u kojem je preciznost veća od odziva, što implicira da je sustav konzervativan pri dodjeljivanju pozitivne oznake.

U odnosu na pozitivnu oznaku, negativna oznaka ima znatno drugačije rezultate, ali vrlo slične u usporedbi s performansama za IRMAS skup podataka:



Slika 31 - Metrike za negativnu oznaku u OpenMIC skupu podataka

Metrike demonstriraju visoku točnost rada klasifikacijskog sustava pri dodjeljivanju negativne oznake. Kao što je slučaj i s IRMAS skupom podataka, referentno na negativnu oznaku, trend odnosa odziva i preciznosti bitno je drugačiji u usporedbi sa situacijom pozitivne oznake. Sustav je znatno manje konzervativan pri dodjeljivanju negativne oznake slično kao i kod IRMAS skupa podataka.

Ukupna točnost sustava iznosi 93.56% što demonstrira da sustav zaista ima visoke mogućnosti točne klasifikacije i prepoznavanja instrumenata unutar glazbenih zvučnih signala.

5. Zaključak

Ovaj diplomski rad usredotočavao se na izazovni zadatak klasifikacije glazbenih instrumenata u polifonom zvuku, složen problem koji uključuje razlikovanje različitih instrumenata koji sviraju istovremeno. Kroz opsežna istraživanja, eksperimentiranje i analizu, ovaj rad pruža značajne uvide i inovacije u tom području. Zaključci izvedeni iz ovog rada sažimaju nalaze, doprinose i moguće buduće smjerove istraživanja i primjene.

Najvažniji doprinos ovog rada razvoj je i programsko ostvarenje robusnog klasifikacijskog sustava koji može precizno identificirati više glazbenih instrumenata unutar polifonih audio zapisa. Sustav sadrži napredne tehnike obrade signala i algoritme strojnog učenja, koji su pažljivo podešeni da se nose sa složenošću polifonih zvučnih tekstura. U sklopu rada opisana su i korištena dva dobro poznata skupa podataka, IRMAS i OpenMIC, za treniranje i vrednovanje sustava klasifikacije. Ovi skupovi podataka pružili su širok raspon glazbenih uzoraka, osiguravajući pritom da je sustav izložen različitim tonovima instrumenata, stilovima sviranja i uvjetima snimanja. Kako bi se dodatno poboljšala učinkovitost i robusnost klasifikacijskog sustava, korišteno je nekoliko tehnika augmentacije podataka. Sustavno su primjenjivane metode kao što su pomicanje visine tona, dodavanje šuma, vremensko pomicanje i preklapanje audio zapisa. Ove su tehnike pomogle u proširenju skupa podataka za treniranje i učinile model otpornim na različita izobličenja i varijacije zvuka.

Sustav klasifikacije pokazao je razmjerno visoku preciznost i odziv na skupu podataka IRMAS, što ukazuje na njegovu učinkovitost u ispravnom identificiranju i razlikovanju različitih glazbenih instrumenata. Ovake performanse sugeriraju da je model sposoban minimizirati lažne pozitivne i lažne negativne rezultate, što ga čini pouzdanim za praktične primjene u pronalaženju glazbenih informacija i analizi zvučnih zapisa. Performanse modela na OpenMIC skupu podataka, iako zadovoljavajuće, ističu izazove povezane s raznovrsnijim i složenijim glazbenim podacima. Nešto niža preciznost i odziv na ovom skupu podataka ilustriraju potrebu za daljnjim poboljšanjima u rukovanju varijacijama u glazbenim žanrovima, okruženjima snimanja i kvaliteti zvuka.

Nalazi iz ovog istraživanja nude nekoliko važnih implikacija i potencijalnih primjena. Razvijeni sustav klasifikacije može značajno poboljšati postojeće MIR sustave pružajući točnu i učinkovitu identifikaciju instrumenata samim time pomažući u zadacima kao što su glazbeno

indeksiranje, preporuke i automatizirano označavanje. Sustav je moguće integrirati u obrazovni sustav kako bi studentima glazbe i nastavnicima pomogao analizirati i razumjeti složene polifone glazbene skladbe, osiguravajući kvalitetnije i produktivnije učenje i predavanje. Točna klasifikacija instrumenata može pomoći u digitalnom očuvanju kulturne baštine i omogućiti detaljnu analizu tradicionalnih i povijesnih glazbenih zapisa.

Iako je ovaj rad doprinio polju klasifikacije glazbenih instrumenata u polifonom zvuku, nekoliko područja zahtijevaju daljnje istraživanje i razvoj. Buduća bi istraživanja trebala istražiti naprednije tehnike augmentacije podataka, kao što su generativne suparničke mreže (GAN) za generiranje sintetičkih zvučnih uzoraka, kako bi se dodatno poboljšala robusnost i performanse modela. Razvijanje sustava klasifikacije u stvarnom vremenu s niskim zakašnjenjem i visokom preciznošću proširilo bi praktične primjene ove tehnologije, posebno u stvarnim izvedbama i interaktivnim glazbenim sustavima. Buduća istraživanja trebala bi poboljšati sposobnost modela da generalizira na različite skupove podataka uključivanjem prijenosnog učenja i tehnika prilagodbe domene, osiguravajući dosljedne performanse u različitim glazbenim okruženjima. Stvaranje korisničkih sučelja i alata koji iskorištavaju mogućnosti klasifikacijskog sustava može olakšati šire korištenje u raznim područjima, od obrazovanja do glazbene industrije.

Zaključno, ovaj rad bavio se složenim problemom klasifikacije glazbenih instrumenata u polifonom zvuku, nudeći doprinos ovom području kroz razvoj robusnog sustava klasifikacije, učinkovitu upotrebu tehnika augmentacije podataka i informativne metrike procjene. Nalazi naglašavaju potencijal naprednih tehnika strojnog i dubokog učenja u poboljšanju analize zvuka i pronalaženja glazbenih informacija. Buduća istraživanja i razvoji nastavit će se graditi na ovim temeljima, težeći sve većoj točnosti, robusnosti i praktičnoj primjenjivosti u vidu glazbene tehnologije koja se neprestano razvija.

6. Literatura

Oppenheim, A. V., & Schafer, R. W. (2009). Discrete-Time Signal Processing. Pearson.

Rabiner, L., & Juang, B. H. (1993). Fundamentals of Speech Recognition. Prentice Hall.

Davis, S., & Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS).

Hershey, S., Chaudhuri, S., Ellis, D. P. W., et al. (2017). CNN Architectures for Large-Scale Audio Classification. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised Pre-training for Speech Recognition. arXiv preprint arXiv:1904.05862.

Benade, A. H. (1990). Fundamentals of Musical Acoustics. Dover Publications.

Campbell, M., & Greated, C. (2004). The Musician's Guide to Acoustics. Oxford University Press.

Bosch, J. J., Janer, J., Fuhrmann, F., & Herrera, P. (2012). A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals. Proceedings of the International Society for Music Information Retrieval Conference (ISMIR).

Humphrey, E. J., Bello, J. P., & LeCun, Y. (2018). Moving Beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics. Proceedings of the International Society for Music Information Retrieval Conference (ISMIR).

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering.

Logan, B. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. In Proceedings of the International Symposium on Music Information Retrieval (ISMIR).

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Perez, L., & Wang, J. (2017). The Effectiveness of Data Augmentation in Image Classification using Deep Learning. arXiv preprint arXiv:1712.04621.

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. Interspeech 2015.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE, 10(3), e0118432.

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing imbalanced data--recommendations for the use of performance metrics. IEEE Transactions on Affective Computing, 5(2), 161-170.