Kauno technologijos universitetas

Informatikos fakultetas

Automatinių muzikos emocijų nustatymo metodų, tinkamų mobiliesiems įrenginiams, tyrimas

Baigiamasis magistro studijų projektas

Eimantas Morkūnas

Projekto autorius

Doc. dr. Mantas Lukoševčius

Vadovas

Kaunas, 2023

Kauno technologijos universitetas

Informatikos fakultetas

Automatinių muzikos emocijų nustatymo metodų, tinkamų mobiliesiems įrenginiams, tyrimas

Baigiamasis magistro studijų projektas

Programų sistemų inžinerija (6211BX011)

Eimantas Morkūnas

Projekto autorius

Doc. dr. Mantas Lukoševičius

Vadovas

Doc. dr. Svajūnas Sajavičius

Recenzentas

Kaunas, 2023

Kauno technologijos universitetas

Informatikos fakultetas

Eimantas Morkūnas

Automatinių muzikos emocijų nustatymo metodų, tinkamų mobiliesiems įrenginiams, tyrimas

Akademinio sąžiningumo deklaracija

Patvirtinu, kad:

1. baigiamąjį projektą parengiau savarankiškai ir sąžiningai, nepažeisdama(s) kitų asmenų autoriaus ar kitų teisių, laikydamasi(s) Lietuvos Respublikos autorių teisių ir gretutinių teisių įstatymo nuostatų, Kauno technologijos universiteto (toliau – Universitetas) intelektinės nuosavybės valdymo ir perdavimo nuostatų bei Universiteto akademinės etikos kodekse nustatytų etikos reikalavimų;

2. baigiamajame projekte visi pateikti duomenys ir tyrimų rezultatai yra teisingi ir gauti teisėtai, nei viena šio projekto dalis nėra plagijuota nuo jokių spausdintinių ar elektroninių šaltinių, visos baigiamojo projekto tekste pateiktos citatos ir nuorodos yra nurodytos literatūros sąraše;

3. įstatymų nenumatytų piniginių sumų už baigiamąjį projektą ar jo dalis niekam nesu mokėjęs (-usi);

4. suprantu, kad išaiškėjus nesąžiningumo ar kitų asmenų teisių pažeidimo faktui, man bus taikomos akademinės nuobaudos pagal Universitete galiojančią tvarką ir būsiu pašalinta(s) iš Universiteto, o baigiamasis projektas gali būti pateiktas Akademinės etikos ir procedūrų kontrolieriaus tarnybai nagrinėjant galimą akademinės etikos pažeidimą.

Eimantas Morkūnas

Patvirtinta elektroniniu būdu

Kauno technologijos universitetas

Informatikos fakultetas

Baigiamojo magistro projekto užduotis

Projekto tema	Automatinių muzikos emocijų nustatymo metodų, tinkamų mobiliesiems įrenginiams, tyrimas
Reikalavimai ir sąlygos (tikslinti pavadinimą pagal poreikį)	
Vadovas / Vadovė	
	(vadovo pareigos, vardas, pavardė, parašas) (data)

Morkūnas Eimantas. Automatinių muzikos emocijų nustatymo metodų, tinkamų mobiliesiems įrenginiams, tyrimas. Magistro studijų baigiamasis projektas / vadovas dr. Mantas Lukoševičius; Kauno technologijos universitetas, Informatikos fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Programų sistemos.

Reikšminiai žodžiai: muzikos nuotaikos nustatymas, muzikos informacijos išgavimas, mašininis mokymasis.

Kaunas, 2023. 36 p.

Santrauka

Darbe pristatoma muzikos emocijas nustatyti gebanti mobilioji programėlė, gebanti lokaliai atlikti dainų analizę, atvaizduoti gautus įverčius pasitelkiant spalvas bei sudaryti panašiomis nuotaikomis pasižyminčių dainų grojaraščius. Literatūros apžvalgoje pateikiami nuotaikos įvertinimo būdai, juos taikantys viešai prieinami duomenų rinkiniai bei mašininio mokymosi modeliai, tinkami automatiniam emocijų nustatymo uždaviniui spręsti. Darbo tyrimo dalyje aprašomi mašininio mokymosi metodai, naudoti ieškant optimalaus nuotaikų nustatymo modelio, kurie įvertinami darbo eksperimentinėje dalyje tikslumo bei greitaveikos atžvilgiais.

Morkūnas Eimantas. Research of music emotion estimation methods suitable for mobile devices. Master's Final Degree Project / supervisor dr. Mantas Lukoševičius; Informatics faculty, Kaunas University of Technology.

Study field and area (study field group): Software engineering.

Keywords: music emotion recognition, music information retrieval, machine learning.

Kaunas, 2023. 36 pages.

Summary

This work presents a mobile application designed to automatically evaluate music emotion. The application performs evaluation locally and can visualize the results using colors. The estimated emotions can be used to group similar songs into playlists. The thesis reviews approaches to emotion classification, datasets designed for continuous music emotion estimation, and solutions to the problem presented in other works. The investigation portion describes approaches that were tried in search of an optimal model for music emotion recognition. The experiments described in the last section evaluate the accuracy and performance of those approaches.

Turinys

Lentelių sąrašas

Paveikslų sąrašas

Santrumpų ir terminų sąrašas

CNN (angl. Convolutional Neural Network) – konvoliucinis neuroninis tinklas. Tinklas, apdorojantis n dimensijų įvestį dalimis slenkant per įvestis.

DEAM (angl. *Database for Emotional Analysis in Music*) – vienas iš duomenų rinkinių, pateikiančių parametrizuotų dainų emocijų įverčius laike.

GRU (angl. *Gated Recurrent Unit*) – viena iš rekurentinių neuroninių tinklų realizacijų.

LR (angl. *Linear Regression*) – tiesinė regresija.

LSTM (angl. *Long Short-Term Memory*) – viena iš rekurentinių neuroninių tinklų realizacijų.

MAE (angl. *Mean Absolute Error*) – klaidos apskaičiavimo būdas, išvedant vidurkį iš paklaidų modulių.

Mel skalė – garso dažnių skalė, kurioje garsai išsidėstę vienodu atstumu klausimo, tačiau ne dažnių, atžvilgiu.

ReLU (angl. *Rectified Linear Unit*) – funkcija, kuri grąžina parametro reikšmę jei jis yra teigiamas ir 0, jei neigiamas. Naudojama kaip neuronų aktyvacijos funkcija.

RMSE (angl. *Root Mean Squared Error*) – klaidos apskaičiavimo būdas, ištraukiant šaknį iš vidutinės kvadratinės klaidos.

RNN (angl. *Recurrent Neural Network*) – rekurentinis neuroninis tinklas. Tokio tinklo neuronai turi jungtis, kurios atliekant įvertinimą išsaugo dalį praeitos įvesties, taip sukuriant būseną, išliekančią tarp greta esančių įvesčių.

Įvadas

Muzikos emocijų atpažinimo sritis aktuali dainų kolekcijų valdymui, kuriose emocijos įvertis gali būti naudojamas kaip dar viena dimensija kategorizuojant muzikos kūrinius. Ši dimensija leidžia grupuoti dainas pagal intuityvius emocijų apibūdinimus, kurie gali palengvinti kūrinių paiešką, būti naudojami rekomendacijų algoritmuose. Srities aktualumas, kartu su tobulėjančiais mašininio mokymosi modeliais, lemia augantį susidomėjimą automatiniais muzikos emocijų nustatymo metodais. Dėl augančių skaičiavimo galimybių kasdieniniuose įrenginiuose, šių metodų pritaikymo potencialas taip pat didėja. Kadaise reikliais resursams laikyti mašininio mokymosi modeliai gali būti pritaikyti ir naudojami įprastuose mobiliuosiuose įrenginiuose.

Lokaliai įrenginyje atliekamas muzikos emocijų nustatymas gali būti patraukli alternatyva kliento-serverio architektūrai – minimizuojami paslaugos teikimo kaštai, pašalinamas poreikis perduoti didelį kiekį garso įrašo duomenų internetu. Tačiau mašininio mokymosi algoritmų taikymas mobiliuosiuose įrenginiuose kelia savų iššūkių – atliekamų skaičiavimų kiekį gali riboti baterijos resursai, įprastuose kompiuteriuose gerus rezultatus demonstruojantys metodai gali veikti neoptimaliai dėl kitokių procesorių architektūrų. Todėl norint taikyti automatizuotus muzikos emocijų atpažinimą mobiliuosiuose įrenginiuose prasminga ištirti, kurie metodai geba efektyviausiai išnaudoti prieinamus sistemos resursus tiksliausiam įverčiui gauti.

Darbo tikslas – nustatyti kurie automatiniai muzikos nuotaikos nustatymo metodai yra tinkamiausi taikyti mobiliuosiuose įrenginiuose.

Darbo uždaviniai:

1. Atlikti dalykinės srities analizę, apžvelgiant literatūroje pateikiamus automatinio muzikos emocijų nustatymo metodus.
2. Sukurti mobiliąją programėlę, gebančią atlikti muzikos emocijų analizę naudotojo įrenginyje bei vaizdžiai pateikti gautus rezultatus.
3. Išbandyti mašininio mokymosi modelius bei jų konfigūracijas, kurios būtų tinkamos muzikos emocijų nustatymui atlikti sukurtoje programėlėje.
4. Įvertinti įvairių muzikos emocijų nustatymo metodų tikslumą bei greitaveiką mobiliajame įrenginyje.

Dokumentas susideda iš keturių pagrindinių dalių – analitinės, projektinės, tiriamosios bei eksperimentinės. Kiekviena dalis atitinka vieną iš iškeltų uždavinių. Analitinėje dalyje pateikiami literatūros analizės rezultatai. Projektinėje dalyje apibūdinama sukurta dainų nuotaikų įvertinimą atliekanti sistema bei jos realizacijos detalės. Tiriamojoje dalyje aprašomi išbandyti modeliai ir jų variacijos, pateikiami apmokymo metu gauti tikslumo įverčiai. Eksperimentinėje dalyje pateikiami metodų tikslumo rezultatai vertinant apmokymui nenaudotus duomenis bei greitaveikos eksperimentai. Darbo gale pateikiamos išvados, apibendrinančios dokumente pateiktą informaciją, pastebėjimus bei rezultatus.

1. Analitinė dalis

Šioje dokumento dalyje apžvelgiami literatūroje pateikiami dainų nuotaikų įvertinimo sprendimai, jų taikomi algoritmai bei metodai. Taip pat įvertinamas dainų nuotaikų nustatymo realizavimui reikalingų duomenų prieinamumas.

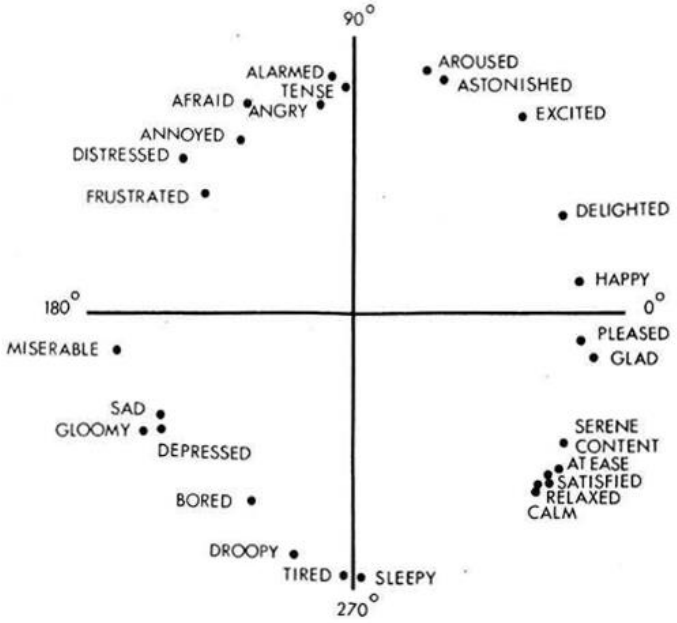
1.1. Nuotaikos įvertinimo būdai

Nuotaika gali atrodyti gan abstrakti bei subjektyvi koncepcija, tačiau psichologai siūlo aibę modelių, leidžiančių klasifikuoti emocijas. Galima išskirti du pagrindinius požiūrius į emocijų klasifikavimą – kategorinį bei dimensinį.

Kategoriniai emocijų modeliai išskiria atskiras, nepersidengiančias emocijas [1]. Išskiriamos bazinės, universaliai suprantamos emocijos, tokios kaip džiaugsmas, liūdesys, pyktis ar baimė, kurias kombinuojant išvedamos sudėtingesnės antrinės bei tretinės emocijos. Taikant kategorinį emocijų nuotaikų modelį muzikos nuotaikos atpažinimui, dainos apibūdinamos viena arba keliomis emocijomis. Toks apibūdinamas intuityviai suprantamas, gali būti naudojamas tekstinėje paieškoje. Šių modelių trūkumas – sunkiai parenkamas nuotaikų kategorijų kiekis. Naudojant tik bazines emocijas nepadengiamas platus emocijų spektras, o pernelyg smulkios kategorijos gali būti nevienareikšmiškai suprantamos, persidengiančios. Didelis emocijų kategorijų skaičius taip pat gali gluminti žmones, atliekančius rankinį emocijų anotavimą, ar interpretuojančius nuotaikos nustatymo rezultatus.

Dimensiniai emocijų modeliai parametrizuoja emocijas keliomis dimensijomis. Nuotaika apibūdinama skaitiniais šių dimensijų įverčiais, o dimensijoms priskyrus ašis, emocijas galima išdėstyti erdvėje. Dažniausiai išskiriamos emocijos pozityvumo (angl. *valence*), energingumo (angl. *arousal*) bei dominavimo dedamosios. Vienas iš plačiausiai dainų nuotaikų apibūdinimui naudojamų dimensinių modelių – *James A. Russell* pristatytas apskritas emocijų modelis [2]. Šis modelis emocijas apibūdina pozityvumu ir energingumu bei jas išdėsto ratu dvimatėje erdvėje, priešingomis laikomas emocijas pozicionuojant skirtinguose apskritimo galuose (1.1 pav.). Modelio erdvė yra vientisa, todėl leidžia apibūdinti be galo didelę emocijų aibę, kurios negalima vienareikšmiškai išreikšti žodžiais. Tačiau skaitinius emocijų įverčius gali būti sunku interpretuoti, intuityviam duomenų fiksavimui, atvaizdavimui bei atrinkimui reikia specialių grafinių sąsajų.

Dainų nuotaikos įvertinimo sistemose kategoriniai bei dimensiniai modeliai gali būti kombinuojami. Pavyzdžiui, įverčiai, gauti taikant dimensinį modelį, gali būti grupuojami į kategorijas, kurioms priskiriami emocijų pavadinimai. Toks sprendimas leidžia išlaikyti platų emocijų spektrą, kuris aktualus taikant dainų kolekcijų apdorojimo algoritmus, išsaugant galimybę pateikti intuityvius nuotaikų pavadinimus, kurie gali būti naudojami tekstinėje paieškoje.



1.1 pav. Nuotaikų išsidėstymas dvimatėje erdvėje, parametrizuotoje pozityvumu (X ašis) bei energingumu (Y ašis), iliustracija iš [2].

Išskiriami du dainų nuotaikų įvertinimo būdai – statinis bei dinaminis. Statinio nuotaikų įvertinimo metu bandoma nustatyti vieną kūrinio perteikiamą emociją. Tačiau muzika kinta laike, kūrinio nuotaika gali būti nepastovi. Taikant statinį emocijų įvertinimą neatsižvelgiama į nuotaikų kaitą, bendras kūrinio įvertis gali būti tik netiksli aproksimacija. Dinaminio nuotaikos įvertinimo metu kūrinys suskirstomas į dalis ir vertinama kiekvienos dalies perteikiama emocija. Tokiu būdu gauti įverčiai leidžia užfiksuoti nuotaikų variacijas, todėl yra tikslesni. Dinaminio įvertinimo metu gaunamas žymiai didesnis kiekis informacijos, kurį gali būti sunku pritaikyti praktiškai, nes dažnu atveju naudojama bendras kūrinio ar jo reprezentatyvios dalies įvertinimas.

1.2. Automatinis muzikos emocijų nustatymas

Bendru atveju automatinis muzikos emocijų nustatymas susideda iš dviejų žingsniu – garso įrašo savybių iš audio signalo išgavimo bei gautų savybių susiejimo su emocijomis naudojant klasifikavimo arba regresijos modelį.

1.2.1. Garso signalo savybės

Garso signalo savybes galima suskirstyti į tembro (spektro), registro, dinamikos bei ritmo savybes. Tembro savybės apibūdina garso dažnių pasiskirstymą signale, kuris išreiškiamas spektro dažnių pasiskirstymo statistiniais įverčiais (pavyzdžiu, vidurkiu, dispersija, asimterija), mel skalės kepstrų koeficientais (angl. *Mel Frequency Cepstral Coefficients,* MFCC). Registro savybės apima tonaciją, dermę, natų ar akordų duomenis, neharmoniškumus. Dinaminės savybės leidžia įvertinti garso signalo energiją, kuri išreiškiama garsumu, garso slėgio lygiu, galios statistiniais įverčiais. Ritmo savybės apibūdina kūrinio tempą, jo pasikeitimus ir svyravimus, žemų dažnių raiškumą bei garsumą. Teigiama, kad didžiausią įtaką dainų emocijų įvertinimo tikslumui turi tembro savybės, o antros pagal svarbą yra registro savybės [3], [4].

Garso signalo savybių išgavimui naudojami audio signalo apdorojimui skirti įrankiai, tokie kaip *openSMILE, Essentia, MIRToolbox, jAudio, Marsyas*. Šie įrankiai leidžia apskaičiuoti žemo lygio savybes, kurios yra aiškiai apibrėžtos, objektyvios. Šios žemo lygio savybės gali būti panaudotos vidutinio lygio savybėms, tokioms kaip melodiškumui, tinkamumo šokiams ar darnumui, išvesti. Vidutinio lygio savybės, lyginant su žemo lygio savybėmis, yra subjektyvesnės, dažnai neturi aiškaus apibrėžimo, tačiau leidžia apibūdinant muzikos kūrinius žmonėms suprantamais aspektais. Tyrimais įrodyta, kad vidutinio lygio savybių taikymas gali pagerinti emocijų nustatymo tikslumą [5]. Vidutinio lygio savybės gali būti išgaunamos taikant euristinius algoritmus, mašininio mokymosi metodus, transformuojant kitas vidutinio lygio savybes, pavyzdžiui apskaičiuojant muzikos tekstūros savybes naudojantis automatiniu būdu nustatytas *MIDI* natas [6].

1.2.2. Muzikos nuotaikos atpažinimo metodai

Automatiniam muzikos nuotaikų atpažinimui galima taikyti tiesinius mašininio mokymosi modelius, tokius kaip tiesinė regresija arba atramos vektorių mašinas (angl. *Support Vector Machine,* SVM). Šie modeliai paprasti, greitai apmokomi bei pritaikius papildomas modifikacijas (pavyzdžiui išvesčių glodinimą) geba pasižymėti neblogu tikslumu [7]. Šie modeliai geba apdoroti tik vieną muzikinio kūrinio segmentą vienu metu. Prarandama segmento pozicijos laike informacija, kuri gali padėti tiksliau įvertinti įrašo nuotaiką. Sudėtingesni modeliai, pavyzdžiui rekurentiniai neuroniniai tinklai (angl. *Recurrent Neural Network,* RNN), geba atsižvelgti į tęstinę muzikos prigimti. Viena iš populiarių ir muzikos nuotaikų dažnai taikomų RNN architektūrų – ilgos trumpalaikės atminties (angl. *Long Short-Term Memory,* LSTM) architektūra. Šia architektūra paremti modeliai 2013-2015 metais vykusių *MediaEval* konkursų muzikos emocijos nustatymo užduotyje pasižymėjo geriausiu tikslumu [8].

2015 metų *MediaEval* konkurso muzikos emocijų atpažinimo užduotyje geriausiai pasirodė dvipusiais LSTM paremti modeliai, kurie įvestis apdoroja tiek jų pateikimo tvarka, tiek atvirkščia jai [9]. Geriausiai pasirodžiusį modelį sudarė 20 LSTM modelių ansamblis. Skirtingi modeliai naudojo skirtingą įvesčių segmento ilgį – 10, 20, 30 bei 60 sekundžių – ir buvo apmokyti naudojant skirtingus duomenų rinkinio poaibius. Kiekvieną LSTM modelį sudarė 5 sluoksniai, kurių kiekvieną sudarė 250 LSTM celių. Galutinis modelis buvo gan didelis, vertinama, kad jį sudarė daugiau kaip 12 milijonų apmokomų parametrų [10]. Tačiau modelių testavimo metu pavienis dvipusis LSTM modelio tikslumas ne daug nusileido ansambliui.

Literatūroje taip pat pateikiami konvoliuciniais neuroniniais tinklais (angl. *Convolutional Neural Network*, CNN) paremti nuotaikų nustatymo metodai. CNN sėkmingai taikomi vaizdų apdorojime, o jų atliekamas dimensijų mažinimas leidžia užfiksuoti duomenų poziciją bei pokyčius laike. Literatūroje pateikiami įvairios CNN variacijos, kurios skiriasi naudojamomis įvestimis. Kai kurie modeliai naudoja standartines garso signalo savybes, kiti taiko savybes, kurios gali būti išreikštos matricomis (pavyzdžiui, spektrogramas [11]), treti dirba su neapdorotu garso signalu tiesiogiai, tikintis, kad modelis savaime išmoks išvesti garso signalo savybes [12].

Vienas iš CNN taikančių metodų – *Malik et al.* pristatytas modelis, apjungiantis CNN bei RNN architektūras [10]. Šis modelis 1 minutės kūrinio segmento garso signalo savybes pateikia CNN sluoksniui, taikančiam dvimatį filtrą. Gautos išvestys toliau apdorojamos pilnai sujungtais bei RNN sluoksniais, kol galiausiai gaunami emocijų įverčiai. Teigiama, kad šis modelis pasižymi geresniu tikslumu nei *MediaEval* 2015 nugalėtojas naudojant tuos pačius apmokymo bei testavimo rinkinius,

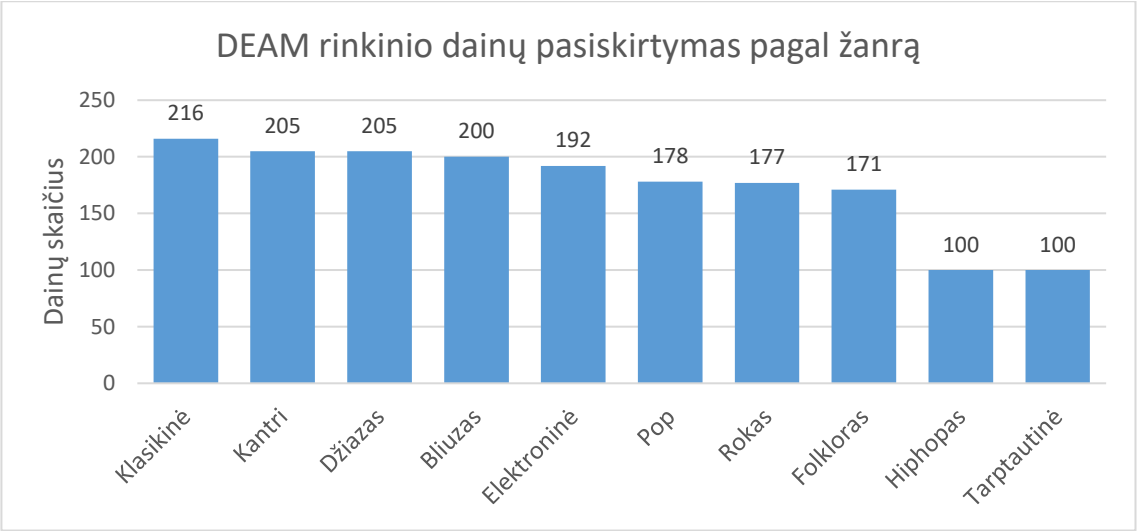
bei pasižymi gerokai mažesniu dydžiu – modelį sudaro apie 30 tūkst. apmokomų parametrų. Panašią architektūra naudoja *Orjesek et al.* aprašytas modelis [12], kuris skiriasi tik pirmuoju CNN sluoksniu, kuris įvestimi priima neapdorotą garso signalą ir taiko vienmatį 5 ms pločio filtrą. Pateiktuose rezultatuose modelio tikslumas pranoksta *Malik et al.* variaciją, o modelį sudaro tik apie 3,3 tūkst. apmokomų parametrų.

Natūralios kalbos apdorojimo uždaviniuose pastaruoju metu geriausi rezultatai pasiekiami taikant transformerio tipo modelius. Šie modeliai naudoja dėmesio (angl. *attention*) mechanizmą duomenų pozicijai laike užkoduoti. *Chaki et al.* pabandė šį dėmesio mechanizmą pritaikyti muzikos emocijų nustatymui, kombinuojant jį su įprastu LSTM modeliu [13]. Darbe pateikiami rezultatai rodo, kad dėmesio mechanizmo taikymas leidžia pagerinti nuotaikų įvertinimo tikslumą lyginant su įprastais LSTM modeliais.

1.3. Duomenų rinkiniai, skirti muzikos nuotaikos nustatymui

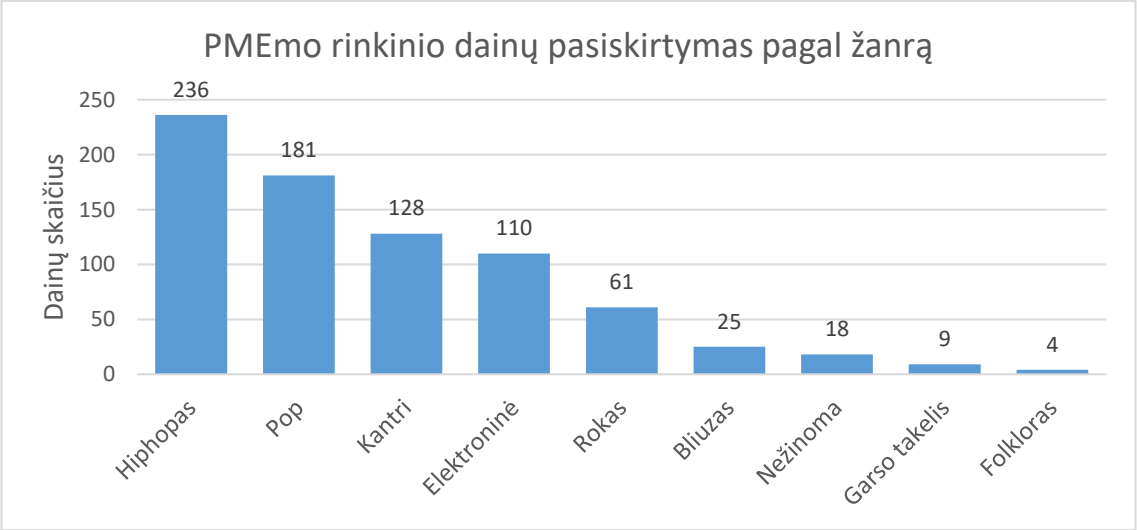
Nors muzikos nuotaikos srityje atlikta nemažai tyrimų, dauguma jų naudoja viešai neprieinamus duomenų rinkinius. Tai dažniausiai lemia autorinių teisinių keliami apribojimai, neleidžiantys platinti muzikinių kūrinių. Laimei, kai kurie muzikiniai kūriniai platinami taikant neapmokestinamų licencijų sąlygas, tačiau tokių dainų nėra daug, jos būna mažiau žinomos, prastesnės kokybės. Iš šių dainų sudaryti duomenų rinkiniai gali būti platinami ir naudojami rezultatų skirtinguose tyrimuose palyginimui.

Didžiausias viešai prieinamas dinaminiam muzikos emocijų nustatymui skirtas duomenų rinkinys – *Database for Emotional Analysis in Music* (DEAM) [14]. Duomenų rinkinį sudaro kūriniai, kurie buvo naudojami 2013-2015 metais vykusiuose *MediaEval* konkursuose dainų emocijų nustatymo uždaviniuose. Rinkinyje pateikiamos 1744 dainų 45 sekundžių trukmės iškarpos bei 58 pilnos dainos. Muzikiniai kūriniai ir jų iškarpos buvo gautos iš *freemusicarchive.org*, *jamendo* bei *medleyDB* šaltinių, kurie platina dainas, kurias galima naudoti neatlygintinai. Dainos apima įvairius muzikos žanrus, jų pasiskirstymas pateiktas 1.2 pav. Duomenų rinkinyje pateikiami dimensiniai nuotaikų įverčiai naudojant nuotaikos pozityvumo bei energingumo parametrus. Įverčiai pateikiami 2 Hz dažniu, pradedant nuo 15 muzikos įrašo sekundės. Dainų įverčiai buvo gauti naudojantis *Amazon Mechanical Turk* platforma, kurioje kiekvieną dainą įvertino ne mažiau kaip 10 žmonių. Įvertinimą atlikę darbuotojai buvo kruopščiai atrinkti, įsitikinant, kad pastarieji gerai supranta užduotį, geba pateikti kokybiškus rezultatus. Įvertinimo metu darbuotojai beklausydami muzikinių įrašų skalėje nuo -10 iki 10 parinkdavo pozityvumo bei energingumo įverčius bei juos koreguodavo dainai progresuojant. Į duomenų rinkinį taip pat įtrauktos dainų savybės, išgautos naudojant *openSMILE* įrankį. Šį savybių rinkinį sudaro 65 žemo lygio savybių vidurkiai bei standartiniai nuokrypiai – iš viso 260 skaitinių įverčių.



1.2 pav. *DEAM* rinkinio dainų pasiskirstymas pagal žanrą remiantis rinkinyje pateiktais dainų metaduomenimis.

Kitas dimensiam nuotaikų įvertinimui skirtas duomenų rinkinys – *PMEmo*. Duomenų rinkinys pateikia anotuotas populiariosios muzikos dainas [15]. Duomenų rinkinį sudaro 794 muzikinių kūrinių priedainių iškarpos. Dainos buvo parinktos remiantis 2016-2017 metais populiariausių dainų sąrašais. Dėl šios priežasties rinkinyje dominuoja hiphopo, pop muzikos žanrai, o mažiau populiarūs, tokie kaip klasikinė muzika, visai nepatenka (1.3 pav.). *PMEmo* rinkinį sudarančių kūrinių platinimą riboja autorinės teisės, todėl rinkinyje nepateikiamos pilnos dainos, tačiau įtrauktos iš pilnų dainų *openSMILE* įrankiu išgautos dainų savybės. Duomenų rinkinyje dainų emocijų anotavimui naudojamas dimensinis nuotaikų modelis, parametrizuotas nuotaikos pozityvumu bei energingumu. Duomenys buvo renkami naudojant programinę įrangą, kuri leido naudotojams pasirinkti dimensijų įverčius 9 balų sistemoje. Nuotaikų anotacijos pateikiamos 2 Hz dažniu, $[0; 1]$ rėžiuose. Anotavimą atliko Kinijos universitetų studentai bei kviestiniai kitataučiai, kuriems už atliktą nuotaikų nustatymą buvo mokama. Kiekvienos dainos emocijas įvertino ne mažiau kaip 10 žmonių, kiekvienas dalyvis įvertino po 20 dainų iškarpų.



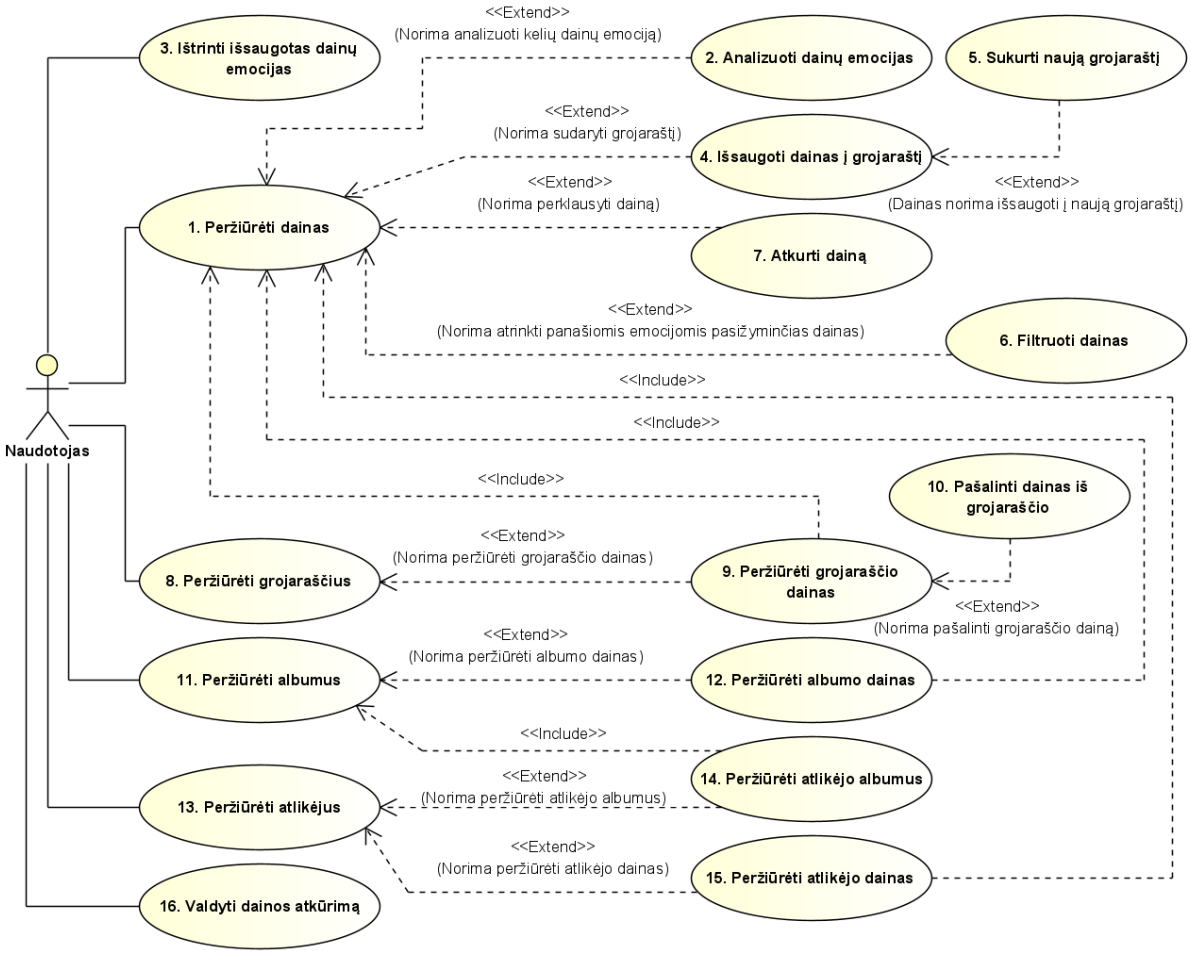
1.3 pav. *PMEmo* rinkinio dainių rinkinio pasiskirstymas pagal žanrą. Informacija apie žanrus gauta naudojantis *Spotify* API pagal rinkinyje pateiktus dainų metaduomenis. *Spotify* API pateikia tik atlikėjo atliekamų kūrinių, tačiau ne individualių dainų, žanrus, todėl galimi neatitikimai.

2. Projektinė dalis

Šiame skyriuje pateikiamos projekto metu sukurtos programinės įrangos aprašymas, apimantis dainų nuotaikų nustatymą atliekančią sistemą, jos architektūrinius sprendimus, veikimo principus, bei mašininio mokymosi modelių parengimo bei pritaikymo mobiliesiems įrenginiams detales.

2.1. Dainų nuotaikos nustatymą atliekanti sistema

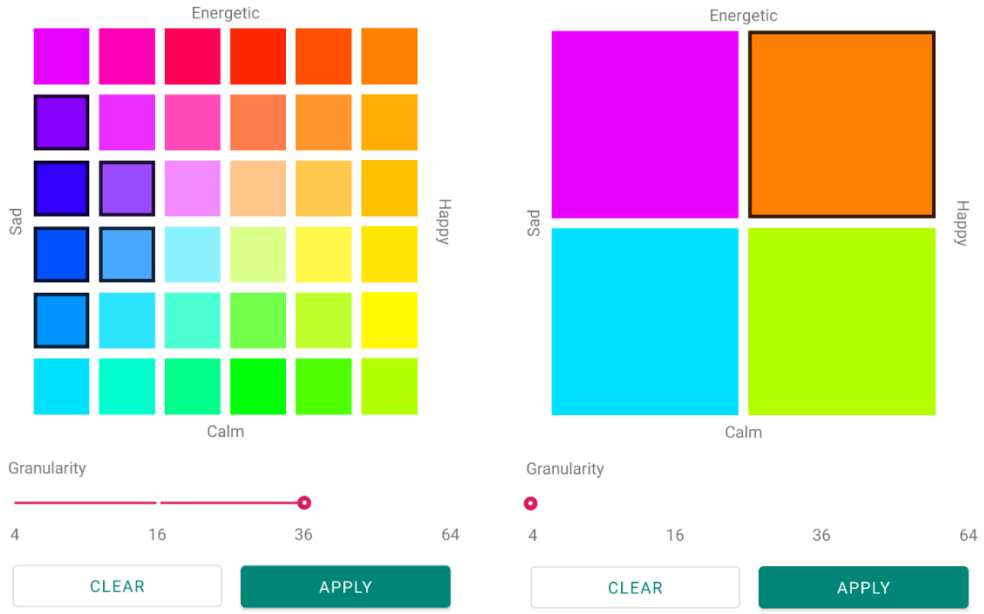
Buvo nuspręsta taikyti dinaminį dainų emocijų nustatymą taikant dimensinį modelį, kuriame emocija apibūdinama jos pozityvumu bei energingumu. Nuotaikų nustatymui buvo pasirinkta sukurti mobiliąją programėlę *Android* platformoje. Šios programėlės tikslas ne tik atlikti dainų emocijų analizę, tačiau ir praktiškai pritaikyti jos rezultatus, todėl sistemos funkcionalumas apima dainų atkūrimą atvaizduojant nustatytas emocijas ir jų kaitą laike, kūrinių filtravimą pagal pasirinktas nuotaikas. Visos sistemos atliekamos funkcijos pateiktos panaudojimo atvejų diagramoje (2.1 pav.).



2.1 pav. Dainų emocijas nustatančio grotuvo panaudojimo atvejų diagrama.

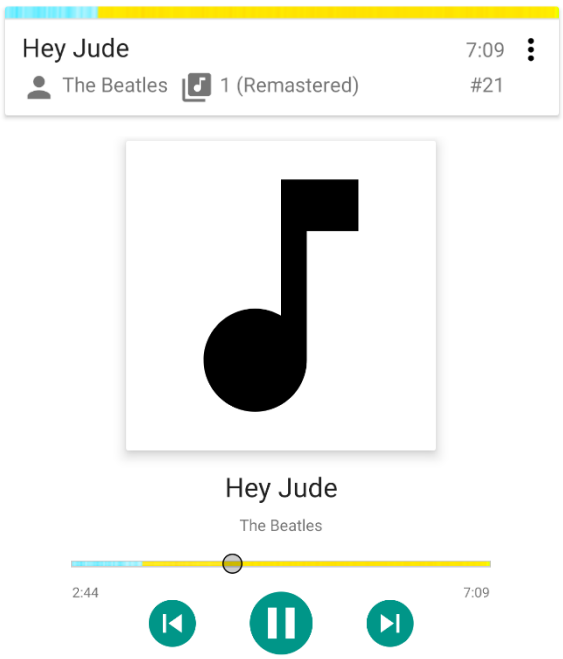
Dainų emocijų atvaizdavimui sistemoje buvo pasirinkta naudoti spalvas. Emocijų kodavimui spalvomis buvo pritaikytas modifikuotas *Adobe Color* naudojamo spalvų rato modelis, kuriame priešingomis laikomos spalvos išdėstomos priešingose apskritimo pusėse – kaip ir emocijos apskritajame emocijų modelyje. Emocijos pozityvumo kodavimui buvo nuspręsta naudoti geltoną spalvą teigiamoms emocijoms atvaizduoti ir mėlyną spalvą neigiamoms, liūdnoms emocijoms išreikšti. Energingumo kodavimui pasirinktos raudona ir žalia spalvos, kur energingos emocijos žymimos raudona spalva, o ramios – žalia. Neutralios emocijos, kurių parametrų įverčiai yra arti 0,

atvaizduojamos balta spalva. Spalvinį emocijų išsidėstymą galima pamatyti programėlės nuotaikų filtro konfigūravimo sąsajoje, kuri pateikiama 2.2 pav.



2.2 pav. Dainų filtro konfigūravimo sąsaja, leidžianti pasirinkti iš 36 (kairėje) ir 4 (dešinėje) emocijų kategorijas.

Taikomas dinaminis emocijų atpažinimas leidžia atvaizduoti emocijų kaitą laike. Sistemoje tai atliekama naudojant spalvotą juostelę, kurioje iš kairės į dešinę pateikiami emocijų įverčiai, užkoduoti spalvomis (2.3 pav.). Ši informacija pateikiama peržiūrint dainų, albumų ar atlikėjų sąrašus bei kūrinio atkūrimo valdymo sąsajoje. Pastarojoje nuotaikos įverčiai pateikiami dainos atkūrimo valdymo laike juostoje, kurioje matoma tuo metu atkuriama įrašo vieta bei jos emocija. Tai leidžia programėlės naudotojui įvertinti nuotaikų nustatymo algoritmo tikslumą perklausant dainą.

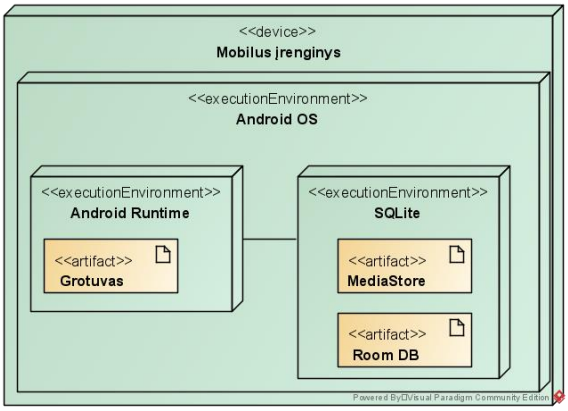


2.3 pav. Dainos nuotaikos įverčio peržiūra dainų sąraše (viršuje) bei atkuriant dainą (apačioje). Spalvota juosta nurodo dainos emocijas bei jų kaitą laike.

2.1.1. Išdėstymo vaizdas

2.4 pav. pateikiama sistemos diegimo diagrama. Iš jos matyti, kad sistema operuoja tik mobiliajame įrenginyje ir nekomunikuoja su išoriniais servisais.

Duomenys apie įrenginyje esančias dainas gaunami iš *Android* platformos *MediaStore* turinio tiekėjo, kuris duomenis saugo lokalioje *SQLite* duomenų bazėje. Sistemos įvertintų dainų emocijų įverčiai saugomi atskiroje lokalioje *SQLite* duomenų bazėje, kuri pasiekiama naudojant *Room* komponentą, atliekantį klasių objektų ir duomenų bazės duomenų apjungimą.

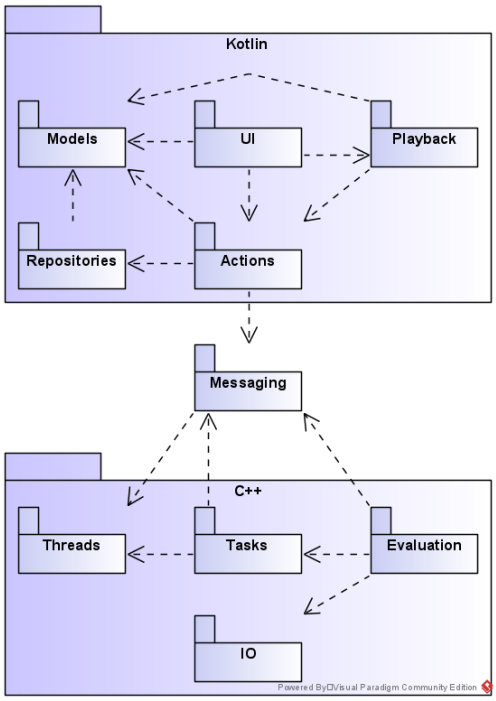


2.4 pav. Sistemos išdėstymo vaizdas.

2.1.2. Paketų diagrama

Sistemą galima išskaidyti į 10 paketų, iš kurių 5 realizuoti *Kotlin* programavimo kalba, 4 *C++* programavimo kalba, o vienas – abejomis (2.5 pav.). *Kotlin* kalba realizuoti grafinės sąsajos, manipuliacijos duomenimis, duomenų saugojimo sluoksniai bei dainų atkūrimą vykdantis servisas. *C++* kalba realizuotas nuotaikų įvertinimo algoritmas bei pagalbiniai komponentai darbui su daugeliu gijų, skaičiavimų valdymui bei audio failų skaitymui iš failų sistemos, dekodavimui. Komunikacijai tarp dviejų technologijų buvo realizuotas *Messaging* paketas, leidžiantis siųsti užklausas bei pranešimus *JSON* formatu tarp skirtingomis technologijomis realizuotų sistemos dalių.

Išskirstytą architektūra lėmė audio savybėms skirtų įrankių realizacijos technologijos. Dauguma įrankių realizuoti *C* arba *C++* kalbomis. Nors kai kurie įrankiai teikia ir *Java* sąsaja, dekoduoto audio signalo duomenų perkėlimas iš *Java* į *C++* reikalautų duomenų kopijavimo tarp *Java* virtualios mašinos ir sistemos aplinkų, kas turėtų neigiamos įtakos sistemos greitaveikai. Todėl buvo nuspręsta visą dainų įvertinimo logiką realizuoti *C++* kalba, minimizuojant tarp technologijų siunčiamų duomenų kiekį.



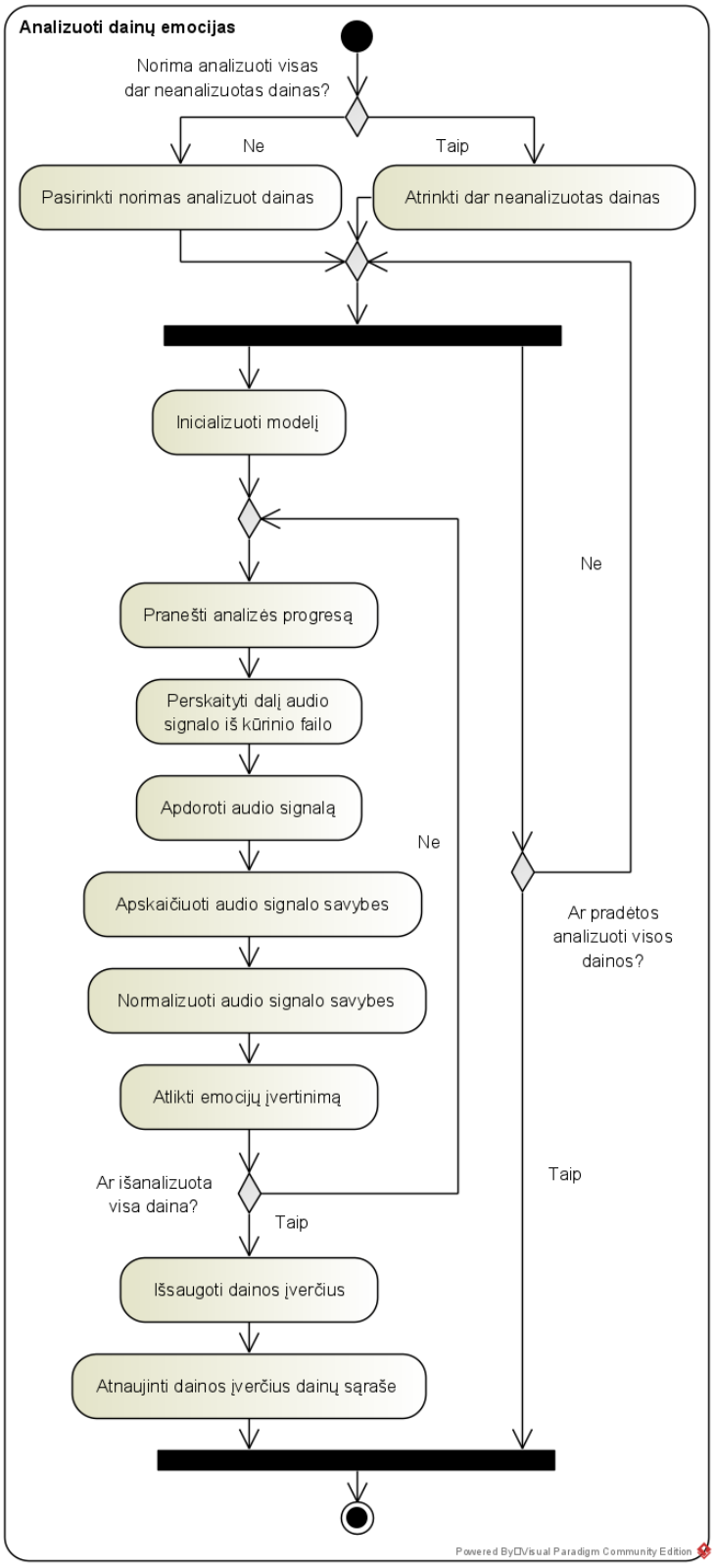
2.5 pav. Sistemos paketų diagrama.

2.2. Muzikos nuotaikų nustatymas

Muzikos nuotaikų nustatymo algoritmas pateikiamas 2.6 pav. pavaizduotoje sekų diagramoje. Analizuojant daugiau nei vieną dainą emocijų nustatymas atliekamas lygiagrečiai, viena gija atlieka vienos dainos vertinimą, tačiau gijų skaičius neviršija įrenginio procesoriaus branduolių skaičiaus. Taip siekiama išvengti per didelio gijų skaičiaus, kuris dėl dažno konteksto keitimo lėtintų bendrą sistemos veikimą.

Dainos nuotaikų įvertinimas atliekamas iteratyviai, dirbant tik su tiek garso signalo duomenų, kiek reikalauja taikomas mašininio mokymosi modelis. Kiekvienos iteracijos metu garso įrašo duomenys perskaitomi iš failo, atliekamas jų dekodavimas į audio signalą. Jeigu reikia, atliekamas pirminis signalo apdorojimas, kurio metu pakeičiamas įrašo diskretizavimo dažnis, kad šis atitiktų modelio įvestis, *stereo* kanalai apjungiami į vieną *mono* kanalą. Signalui taip pat pritaikomas garsumo suvienodinimo filtras, užtikrinantis tikslesnius garso signalo savybių įverčius.

Muzikos nuotaikų nustatymą atlieka mašininio mokymosi modeliai, kurie apmokomi *Python* ekosistemoje naudojant *TensorFlow* paketą. Apmokyti modeliai konvertuojami į *FlatBuffers* formatą, kuris naudojant kartu su *TensorFlow Lite* biblioteka leidžia atlikti nuotaikų įvertinimą mobiliajame įrenginyje.



2.6 pav. Dainų nuotaikos nustatymo algoritmo sekų diagrama.

3. Tiriamoji dalis

Šiame skyriuje pristatomas tyrimas, kurio metu buvo siekiama optimizuoti muzikos emocijų įvertinimo tikslumą išbandant įvairias modelių architektūras bei konfigūracijas. Tolimesniuose skyriuose pateikiami taikyti mašininio mokymosi modeliai, išbandytos įvesčių variacijos bei geriausią įverčių tikslumą apmokymo metu pasiekusios jų kombinacijos.

3.1. Mašininio mokymosi modelių apmokymo metodika

Modelių apmokymui buvo naudojamas *MediaEval 2015* duomenų rinkinys, kurį sudaro 431 daina. Šis *DEAM* duomenų rinkinio poaibis buvo pasirinktas dėl jo paplitimo literatūroje, kas leido tyrimo metu apmokytų modelių tikslumą lyginti su kituose moksliniuose darbuose pateiktais rezultatais.

Kiekviena išbandyta modelių ir jų įvesčių kombinacija buvo apmokyta taikant *k* imčių kryžminę validaciją (angl. *k-fold cross-validation*) naudojant 5 imtis. Modelio paruošimui skirti duomenys buvo padalinami į 5 dalis, iš kurių 4 dalys naudotos modelio apmokymui, o likusioji dalis – validacijai. Modelio apmokymas ir vertinimas buvo atliekamas 5 kartus, kas kartą naudojant vis kitą duomenų imtį validacijai atlikti. Validacijos metu klaida buvo apskaičiuojama pagal RMSE. Apmokymas buvo stabdomas, jei po tam tikro epochų skaičiaus validacijos klaida nebemažėjo. Sustabdžius apmokymą, modelio svoriai buvo atstatomi į geriausia validacijos klaida pasižymėjusios epochos gale buvusius svorius. Galiausiai, prieš pereinant prie sekančios duomenų imties, modelis buvo dar kartą įvertinamas naudojant validacijos duomenų rinkinį, apskaičiuojant RMSE ir MAE. Modelio konfigūracijos tinkamumas buvo vertinamas pagal 5 imčių klaidų vidurkį.

Duomenys į apmokymo ir validacijos aibes buvo skirstomi pagal dainas. Tai leido užtikrinti, kad įvestys gautos iš vienos dainos nepatektų ir į apmokymo, ir į validacijos duomenų rinkinius. Priešingu atveju, panašios įvestys abiejuose rinkiniuose galėtų lemti nereprezentatyviai žemą validacijos klaidą.

3.2. Modeliai, naudojantys garso įrašo savybes

Pirmoji bandytų modelių grupė įvestims naudojo garso įrašų savybes, gautas apdorojant garso signalą įvairiais algoritmais. Šiame skyriuje pateikiama informacijas apie naudotas savybes, jų išgavimą bei modelių apmokymo rezutlatus.

3.2.1. Garso įrašo savybių išgavimas ir apdorojimas

Savybių išgavimui buvo naudojama *Essentia* [16] biblioteka . Savybės buvo išgaunamos iš 44,1 kHz diskretizavimo dažnio signalo analizuojant nepersidengiančius intervalus. Intervalų ilgis, priklausomai nuo bandomos konfigūracijos, buvo $\{0{,}5;\ 1;\ 2\}$ sekundės. Kiekvienas intervalas buvo apdorojamas taikant 2048 verčių (angl. *sample*) langą su 1024 verčių persidengimu. Visų intervalo langų savybių įverčiai buvo agreguojami naudojant vidurkį bei standartinį nuokrypį.

Apdorojant signalą buvo išgaunama 17 savybių (3.1 lentelė), iš kurių buvo suformuojama 60 signalo skaitinių įverčių. Garso įrašo savybės buvo parinktos pagal *Essentia* bibliotekoje realizuotų algoritmų aibę bei garso įrašo savybių įtakos įverčių tikslumui analizę, pateiktą [4].

Gauti garso įrašo savybių įverčiai prieš pateikiant mašininio mokymosi modeliui buvo normalizuoti taikant. z-įverčio normalizavimą, kiekvieną įvestį transformuojant pagal formulę

$$z = \frac{x - \mu}{\sigma},$$

kur μ žymi savybės vidurkį apmokymui naudotuose duomenyse, o σ – standartinį nuokrypį.

3.1 lentelė. Dainų nuotaikų nustatymui naudotos garso signalo savybės.

Kategorija	**Savybė**	**Naudojamos reikšmės**
Tembro savybės	Spektro vidurkis	Vidurkis ir standartinis nuokrypis
	Spektro dispersija	Vidurkis ir standartinis nuokrypis
	Spektro asimetrija	Vidurkis ir standartinis nuokrypis
	Spektro ekscesas	Vidurkis ir standartinis nuokrypis
	Spektro plokštumas	Vidurkis ir standartinis nuokrypis
	Spektro svyravimas	Vidurkis ir standartinis nuokrypis
	Dažnis, žemiau kurio susikaupia 85% spektro energijos	Vidurkis ir standartinis nuokrypis
	Dažnis, žemiau kurio susikaupia 95% spektro energijos	Vidurkis ir standartinis nuokrypis
	Disonansas	Vidurkis ir standartinis nuokrypis
	MFCC	13 koeficientų, kiekvieno koeficiento vidurkis ir standartinis nuokrypis
	Signalo ženklo kitimo dažnis	Vidurkis ir standartinis nuokrypis
Registro savybės	Chromagramos vidurkis	Vidurkis ir standartinis nuokrypis
	Dermė	Minoras arba mažoras (0 arba 1)
	Neharmoniškumas	Vidurkis ir standartinis nuokrypis
	Tonacija	13 diskrečių dydžių, atitinkančių toną nuo A iki G# įskaitant pustonius
	Tristimulas	3 reikšmės, kiekvienos reikšmės vidurkis ir standartinis nuokrypis

3.3. Tiesine regresija ir rekurentiniais neuroniniais tinklais paremti modeliai

Parametrizuotų dainos emocijų nustatymas yra regresijos uždavinys, todėl vienas paprasčiausių galimų taikyti modelių – tiesinė regresija (LR). Šį modelį sudaro tik vienas neuronas, transformuojantis įvestis į išvestis. Validacijos metu LR pasiekė 0,235 energingumo bei 0,286 pozityvumo klaidas, apskaičiuotas pagal RMSE. Dėl savo paprastumo šio modelio tikslumas ribotas, tačiau jį galima laikyti atskaitos tašku – apatine tikslumo riba, žemiau kurios esančius modelius galima laikyti neperspektyviais. Tyrimo metu buvo išbandyti ir sudėtingesni tiesiniai modeliai (pavyzdžiui, neuroninis tinklas su keliais pilnai sujungtais neuronų sluoksniais), tačiau jų validacijos tikslumo įverčiai buvo tik neženkliai geresni nei tiesinės regresijos.

Geresnių rezultatų pavyko pasiekti naudojant architektūras paremtas rekurentiniais neuroniniais tinklais. Šios struktūros modeliams vienu metu buvo pateikiamos paskutinių {6; 10; 14} intervalų savybės, kurios modeliui suteikia daugiau kontekstinės informacijos. Taip pat buvo bandoma varijuoti individualaus intervalo ilgiu, išbandant {0,5; 1; 2} sekundžių intervalus. Rekurentiniams sluoksniams buvo bandyta naudoti LSTM, GRU celes, jų kombinacijas, *Dropout* sluoksnius, kurie atsitiktine tvarka įvestis pakeičia nuliais. Didžiausią tikslumą testavimo metu pavyko pasiekti naudojant 1 sekundės intervalus, vienu metu pateikiant 10 intervalų įvestis naudojant 3.2 lentelėje pateiktą modelio struktūrą. Šio modelio validacijos RMSE klaida buvo 0,230 energingumui ir 0,276 pozityvumui.

3.2 lentelė. Tiksliausio apmokymo metu RNN modelis.

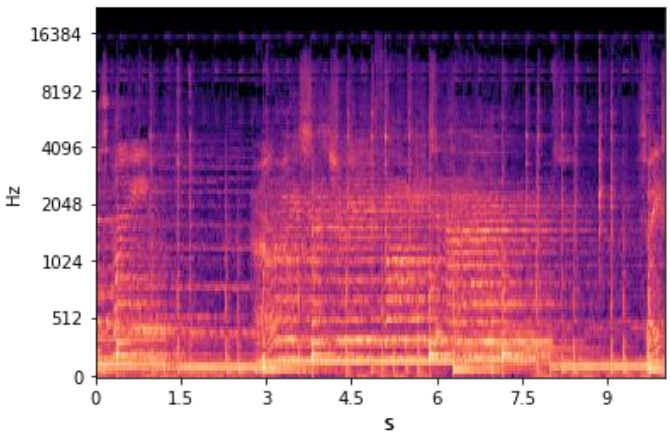
Sluoksnis	**Išvesties dydis**	**Parametrų skaičius**
LSTM (16 celės)	10 x 16	4928
Dropout (0,5)	10 x16	0
LSTM (12 celės)	10 x 12	1392
Dropout (0,25)	10 x 12	0
LSTM (8 celės)	8	672
Pilnai sujungtas (2 neuronų)	2	18
	Viso	7010

3.4. Modeliai, naudojantys garso įrašo spektrogramas

Garso įrašo savybės, išgautos taikant 3.2.1 aprašytą procesą, gali būti nereprezentatyvios. Apdorojimo metu prarandama dalis garso signalo informacijos, kuri gali būti naudinga atliekant nuotaikos įvertinimą. Šiame skyriuje aptariami modeliai, įvestims naudojantys mažai apdorotą audio signalą.

3.4.1. Spektrogramų sudarymas

Modelių parengimui buvo pasirinkta naudoti *mel* spektrogramas. Pastarosios sudaromos atliekant sparčiąją Furjė transformaciją (FFT) garso signalui ir gautas garso dažnių amplitudes padalinant į n diapazonų pagal *mel* skalę. Dažnių konvertavimui į *mel* skalę naudojami du algoritmai – *Slaney* ir HTK. Gautos diapazonų amplitudes reikšmes galima kelti antru laipsniu, taip gaunant signalo „galios“ įvertį. Amplitudžių nepakėlus kvadratu laikoma, kad spektrograma atitinką signalo „energiją“.



3.1 pav. 10 sekundžių garso signalo spektrogramos vizualizacija gauta naudojant *librosa* paketą. X ašis žymi laiką, Y ašis dažnius, o spalva priklauso nuo amplitudės – didesnė amplitudė lemia šviesesnę spalvą.

Tyrimo metu spektrogramos buvo sudaromos naudojant 2048 garso signalo vertes su 512 verčių poslinkiu. Bandymams buvo naudojamos įvairios spektrogramų sudarymo variacijos – naudoti *Slaney* bei HTK algoritmai, {128; 256; 512} dažnių diapazonai, 44,1 kHz ir 22 kHz diskretizavimo dažniai.

3.4.2. Konvoliuciniais neuroniniais tinklais paremti modeliai

Spektrogramų apdorojimui naudoti konvoliuciniai neuroniniai tinklai, dažniausiai taikomi apdorojant vaizdo duomenis. Kaip ir vaizdo apdorojimu atveju, tikimasi, kad tinklas iš mažai apdoroto garso signalo apmokymo metu sugebės pats išskirti signalo savybes, kuriomis naudojantis tiesiniai modelio sluoksniai gebės tiksliai įvertinti dainos emocijas.

Tyrimo metu be praeitame skyrelyje minėtų spektrogramos sudarymo parametrų buvo eksperimentuojama su spektrogramos ilgiu, išbandant {0,5; 1; 6; 10} sekundžių ilgius. Geriausias energingumo įverčio tikslumas (0,166 RMSE) buvo gautas taikant 128 diapazonų, 0,5 sekundės ilgio, 22 kHz diskretizavimo dažnio „galios“ spektrogramas, apskaičiuotas pagal *Slaney* algoritmą. Šio modelio struktūra pateikiama 3.3 lentelėje. CNN sluoksniai naudojo *ReLU* aktyvacijos funkciją.

3.3 lentelė. Tiksliausią energingumo įverti apmokymo metu pasiekęs CNN modelis (CNN-E).

Sluoksnis	**Išvesties dydis**	**Parametrų skaičius**
Conv2D (kernel 3x3, stride 1x1, filters 8)	126 x 20 x 8	80
Conv2D (kernel 3x3, stride 2x2, filters 8)	62 x 9 x 8	584
Conv2D (kernel 3x3, stride 1x1, filters 16)	60 x 7 x 16	1168
Conv2D (kernel 1x1, stride 1x1, filters 16)	60 x 7 x 16	272
Conv2D (kernel 3x3, stride 1x1, filters 16)	58 x 5 x 16	2320
Conv2D (kernel 1x1, stride 1x1, filters 24)	58 x 5 x 24	408
AvgPool2D (size 2x2)	29 x 2 x 24	0
Flatten	1392	0
Pilnai sujungtas (16 neuronų)	16	22288

Pilnai sujungtas (2 neuronai)	2	34
	Viso	27154

Geriausiu pozityvumo įverčiu (0,156 RMSE) pasižymėjo modelis, įvestims naudojęs 128 diapazonų, 6 sekundžių ilgio, 22 kHz diskretizavimo dažnio „galios“ spektrogramas, apskaičiuotas pagal *Slaney* algoritmą. Šio modelio struktūra pateikiama 3.4 lentelėje.

3.4 lentelė. Tiksliausią pozityvumo įverti apmokymo metu pasiekęs CNN modelis (CNN-P).

Sluoksnis	**Išvesties dydis**	**Parametrų skaičius**
Conv2D (kernel 3x3, stride 1x1, filters 8)	126 x 257 x 8	80
Conv2D (kernel 3x3, stride 2x2, filters 8)	62 x 128 x 8	584
Conv2D (kernel 1x1, stride 2x2, filters 16)	31 x 64 x 16	144
Conv2D (kernel 3x3, stride 1x1, filters 16)	29 x 62 x 16	2320
Conv2D (kernel 1x1, stride 1x1, filters 16)	29 x 62 x 16	272
Conv2D (kernel 3x3, stride 1x1, filters 24)	27 x 60 x 24	3480
Conv2D (kernel 1x1, stride 1x1, filters 24)	27 x 60 x 24	600
Conv2D (kernel 3x3, stride 2x2, filters 24)	13 x 29 x 24	5208
Conv2D (kernel 1x1, stride 2x2, filters 32)	7 x 15 x 32	800
Conv2D (kernel 3x3, stride 1x1, filters 32)	5 x 13 x 32	9248
Conv2D (kernel 1x1, stride 1x1, filters 32)	5 x 13 x 32	1056
AvgPool2D (size 2x2)	2 x 6 x 32	0
Flatten	384	0
Pilnai sujungtas (16 vienetų)	16	6160
Pilnai sujungtas (2 neuronai)	2	18
	Viso	29986

3.4.3. Konvoliucinius bei rekurentinius neuroninius tinklus apjungiantys modeliai

Nors spektrogramų X ašis žymi laiką, taikant CNN modelius ši laiko dimensija išplokštinama ir iš dalies prarandama. Kad ją panaudoti prasmingai, CNN pateikiamas įvestis galima skaidyti į trumpesnius segmentus, kuriuos, išsaugant eiliškumą, būtų galima apdoroti naudojant RNN sluoksnius. Tokioje architektūroje CNN sluoksniai naudojami audio signalo savybėms išgauti, o emocijų įvertinimas atliekamas naudojant RNN sluoksnį, gebantį įvertinti signalo tęstinumą laike.

Šios struktūros modeliai buvo išbandomi naudojant įvairius segmentų kiekius ($\{6; 10; 14\}$ sekundžių) bei jų ilgius ($\{0{,}5;\ 1;\ 2\}$ sekundės).

3.5 lentelė. Tiksliausią energingumo įvertį apmokymo metu pasiekęs CNN-RNN modelis (CNN-RNN-SE).

Sluoksnis	**Išvesties dydis**	**Parametrų skaičius**
Conv2D (kernel 3x3, stride 1x1, filters 8)	10 x 126 x 85 x 8	80
Conv2D (kernel 3x3, stride 2x2, filters 8)	10 x 62 x 42 x 8	584
Conv2D (kernel 1x1, stride 1x1, filters 16)	10 x 31 x 21 x 16	144
Conv2D (kernel 3x3, stride 1x1, filters 16)	10 x 29 x 19 x 16	2320
Conv2D (kernel 1x1, stride 1x1, filters 16)	10 x 29 x 19 x 16	272
Conv2D (kernel 3x3, stride 1x1, filters 16)	10 x 27 x 17 x 16	2320
Conv2D (kernel 1x1, stride 1x1, filters 24)	10 x 27 x 17 x 24	408
Conv2D (kernel 3x3, stride 2x2, filters 24)	10 x 13 x 8 x 24	5208
Conv2D (kernel 1x1, stride 2x2, filters 32)	10 x 7 x 4 x 32	800
Reshape	10 x 896	0
LSTM	16	58432
Pilnai sujungtas (2 neuronai)	2	18
	Viso	70602

Geriausią energingumo tikslumą (0,148 RMSE) apmokymo metu pasiekė modelis, įvestimi naudojęs 10 vienos sekundės ilgio spektrogramų, gautų taikant 128 diapazonų, 44,1 kHz diskretizavimo dažnį, HTK algoritmą ir „energijos“ įvertį. Šio modelio struktūra pateikta 3.5 lentelėje.

Geriausią pozityvumo tikslumu (0,154 RMSE) pasižymėjo modelis, naudojęs 10 sekundžių segmentų kiekį, 1 sekundės segmento ilgį, 512 diapazonus, 22 kHz diskretizavimo dažnį bei HTK algoritmą apskaičiuojant „energijos“ spektrogramas. Šio modelio struktūra pateikta 3.6 lentelėje.

3.6 lentelė. Tiksliausią energingumo įvertį apmokymo metu pasiekęs CNN-RNN modelis (CNN-RNN-SP).

Sluoksnis	**Išvesties dydis**	**Parametrų skaičius**
Conv2D (kernel 3x3, stride 1x1, filters 8)	10 x 510 x 42 x 8	80
Conv2D (kernel 3x3, stride 2x2, filters 8)	10 x 254 x 20 x 8	584
Conv2D (kernel 1x1, stride 1x1, filters 16)	10 x 127 x 10 x 16	144
Conv2D (kernel 3x3, stride 1x1, filters 16)	10 x 125 x 8 x 16	2320
Conv2D (kernel 1x1, stride 1x1, filters 16)	10 x 125 x 8 x 16	272
Conv2D (kernel 3x3, stride 1x1, filters 16)	10 x 123 x 6 x 16	2320
Conv2D (kernel 1x1, stride 1x1, filters 24)	10 x 123 x 6 x 24	408
Conv2D (kernel 3x3, stride 2x2, filters 24)	10 x 61 x 2 x 24	5208
Conv2D (kernel 1x1, stride 2x2, filters 24)	10 x 31 x 1 x 24	600

Reshape	10 x 744	0
LSTM (16 celių)	16	48704
Pilnai sujungtas (2 neuronai)	2	34
	Viso	60674

3.5. Modeliai, apdorojantys garso įrašo signalą

3.4 skyriuje aptarti modeliai įvestims naudojo mažai apdorotą garso signalą. Šiame skyriuje aptariamas kraštutinis variantas, kurio modeliams pateikiamas visai neapdorotas audio signalas. Tokie modeliai realizuoti naudojant vienmačius CNN sluoksnius, kurie, kaip ir praeitame skyriuje, apjungiami su RNN sluoksniais nuotaikos įvertinimui atlikti. Tyrimo metu buvo išbandytas vienmatis CNN, naudojantis tokią pačią konfigūraciją kaip [12] – signalas buvo apdorojamas 5 ms ilgio filtrais su 2.5 ms postūmiu. CNN įvestimi buvo 0,5 sekundės segmentai, kurių bendras kiekis vienos įvestiems metu sudarė $\{0{,}5;\ 1;\ 5;\ 10\}$ sekundžių. Geriausią energingumo įvertį apmokymo metu pasiekė modelis, naudojęs 1 sekundės duomenis (0,2073 RMSE, CNN-RNN-AE), o tiksliausiai pozityvumą vertino modelis, naudojęs 5 sekundžių duomenis (0,2242 RMSE, CNN-RNN-AP). 3.7 lentelėje pateikiama naudota modelio architektūra, kuri visais atvejais buvo identiška.

3.7 lentelė. Modelių, naudotų emocijų nustatymui iš neapdoroto garso signalo, struktūra.

Sluoksnis	**Išvesties dydis**	**Parametrų skaičius**
Conv1D (kernel 5ms, stride 2.5 ms)	2 x 199 x 8	888
Pilnai sujungtas sluoksnis (16 neuronų)	2 x 199 x 16	144
Reshape	2 x 3184	0
Bidirectional GRU (16 vienetų)	16	153312
Pilnai sujungtas sluoksnis (2 neuronai)	2	34
	Viso	154378

3.6. Tikslumo tyrimo rezultatų apibendrinimas.

Bendri apmokymo metu gautų tikslumo įverčių rezultatai pateikti 3.8 lentelėje. Tiksliausius energingumo įverčius apmokymo metu pateikė CNN-RNN architektūros modelis, įvestims naudojęs spektrogramas. CNN architektūros modelis pasižymėjo geriausia pozityvumo paklaida.

Iš rezultatų taip pat matoma, kad CNN naudoję modeliai dažniausiai pasižymėjo geru vienos iš dimensijų įvertinimų, tačiau gan didele klaida vertinant kitą dimensiją. Panašu, kad trumpesnės įvestys labiau tinkamos energingumui nustatyti, o įvestys su daugiau duomenų lemia tikslesnį pozityvumo įvertį.

3.8 lentelė. Modelių tikslumo nustatant energingumą (E) bei pozityvumą (P) palyginimas remiantis apmokymo metu gautomis klaidomis. Mėlyna spalva išskirti geriausi rezultatai.

Modelis	**Validacijos klaida**			
	MAE		**RMSE**	
	E	**P**	**E**	**P**
LR	0,1940	0,2414	0,2348	0,2863
RNN	0,1831	0,2275	0,2297	0,2758
CNN-E	0,1658	0,1713	0,2067	0,2154
CNN-P	0,1782	0,1556	0,2192	**0,1934**
CNN-RNN-SE	**0,1480**	0,1668	**0,1859**	0,2088
CNN-RNN-SP	0,1603	**0,1539**	0,2006	0,1958
CNN-RNN-AE	0,1682	0,1903	0,2073	0,2269
CNN-RNN-AP	0,1836	0,1867	0,2223	0,2242

4. Eksperimentinė dalis

Šiame skyriuje aprašomi eksperimentai, kurie buvo atlikti 3 skyriuje aprašytų modelių kokybės įvertinimui. Eksperimentų rezultatai neturėjo įtakos tolimesniam modelių tobulinimui, tačiau leido nustatyti perspektyviausius modelis, kurie galėtų būti taikomi 2 skyriuje aprašytoje mobiliojoje programėlėje.

4.1. Modelių tikslumo įvertinimas

3 skyriuje atliktais tyrimais buvo bandomi surasti tiksliausiai muzikos emocijas įvertinantį modelį. Tyrimo metu priimti sprendimai rėmėsi apmokymo duomenų savybėmis ir galėjo lemti hiperparametrų, pritaikytų būtent tam duomenų rinkiniui, pasirinkimą. Siekiant objektyviau įvertinti modelių gebėjimą vertinti emocijas, eksperimentai šiame skyriuje buvo atlikti su duomenimis, kurie nebuvo naudoti parenkant hiperparametrus.

4.1.1. Modelių tikslumo įvertinimo metodika

Tikslumo vertinimas buvo atliktas naudojant *MediaEval* 2015 testavimo duomenų rinkinį, kurį sudaro 58 pilnos dainos, nenaudotos modelių apmokymo metu. Tikslumas buvo vertinamas naudojant mobiliems įrenginiams pritaikytus modelius. Tyrimo metu apskaičiuojamos dviejų tipų klaidos – MAE bei RMSE.

4.1.2. Modelių tikslumo įvertinimo rezultatai

4.1 lentelėje pateikti 3 skyriuje aprašytų modelių testavimo klaidos. Matyti, kad testavimo klaida gerokai didesnė nei validacijos klaida, kas gali indikuoti, jog modelių parametrai pernelyg priklausomi nuo apmokymo duomenų. Taip pat galima išskirti, kad nė vienas modelis nesugebėjo tiksliai įvertinti energingumo, paprastos tiesinės regresijos rezultatas pranoko sudėtingesniu modelius.

4.1 lentelė. Modelių tikslumo įvertinimo naudojant *MediaEval 2015* testavimo duomenis rezultatai. Mėlyna spalva išskirti geriausi rezultatai.

Modelis	**Testavimo klaida**			
	MAE		**RMSE**	
	E	**P**	**E**	**P**
LR	**0,194**	0,271	**0,234**	0,338
RNN	0,199	0,283	0,241	0,358
CNN-E	0,234	0,279	0,279	0,361
CNN-P	0,286	**0,193**	0,327	**0,241**
CNN-RNN-SE	0,208	0,283	0,260	0,360
CNN-RNN-SP	0,248	0,243	0,295	0,305
CNN-RNN-AE	0,208	0,275	0,249	0,351
CNN-RNN-AP	0,224	0,271	0,267	0,342

4.2 lentelėje pateikiami geriausiai pasirodę modeliai bei rezultatai, gauti kituose literatūros darbuose. Nuotaikos pozityvumo nustatymo atžvilgiu CNN-P modelio rezultatas gan solidus ir mažai

nusileidžia geriausią literatūroje įvertį užfiksavusiam CNN-RNN modeliui, dirbančiu su neapdoroto garso signalu. Nors energingumo nustatymas testavimo metu pasirodė prastokai, tiesinės regresijos rezultatas panašus į BLSTM-ELM modelio, kuris *MediaEval 2015* konkurse pasirodė geriausiai, rezultatą.

4.2 lentelė. Darbe pristatytų modelių (paryškinti mėlynai) tikslumo palyginimas su kitais literatūroje pateikiamais *MediaEval 2015* rezultatais. Mėlyna spalva išskirti darbe pristatyti modeliai.

Modelis	**Testavimo klaida (RMSE)**	
	E	**P**
LR	*0,234*	*0,338*
CNN-P	*0,326*	*0,241*
BLSTM-ELM [9]	0,234	0,308
CRNN_NB [10]	0,231	0,279
Raw audio CNN-RNN [12]	0,214	0,240

4.2. Modelių pritaikomumo naujiems duomenų rinkiniams įvertinimas

Siekiant įvertinti modelių tinkamumą spręsti dainos emocijų įvertinimo uždavinį ne tik vieno duomenų rinkinio rėmuose, buvo atliktas eksperimentas naudojant kitą duomenų rinkinį. Jo metu modeliai, sukonfigūruoti remiantis *MediaEval 2015* duomenų rinkiniu, buvo apmokyti naudojant *PMEmo* duomenų rinkinį. Gauti modeliai buvo testuojami naudojantis *PMEmo* testavimo aibe bei *MediaEval 2015* testų rinkiniu. Taip pat buvo patikrintas modelių, apmokytų naudojant MediaEval *2015* duomenis, tikslumas naudojant *PMEmo* testų rinkinį.

Prieš pradedant eksperimentą, *PMEmo* rinkinys buvo apdorotas. Iš rinkinio buvo pašalintos dainos, kurių įvertinimo vidutinis standartinis nuokrypis $\sigma > 0{,}35$ bet kuriai iš emocijos dimensijų. Taip iš 791 iškarpų buvo atrinktos 382 dainos. Likusių įrašų emocijų įvertinimai buvo perskaičiuoti rėžiuose $[-1; 1]$, kad šie atitiktų *DEAM* duomenų formatą. Duomenų rinkinys buvo padalintas į testavimo rinkinį, kurį sudarė 37 dainos, atrinktos išlaikant žanrų pasiskirstymą, bei 345 dainų modelio paruošimo rinkinį. Pastarasis buvo išskaidytas į validacijos (10%) bei apmokymo rinkinius (90%). Validacijos rinkinys buvo naudojamas apmokymui stabdyti kai validacijos klaida per paskutines 5 epochas nebemažėjo. Baigus apmokymą, modelio svoriai buvo grąžinami į svorius, su kuriais buvo pasiekta mažiausia epochos validacijos klaida.

Eksperimento rezultatai pateikiami 1 priede. Remiantis jais, buvo pastebėta:

- Naudojant *PMEmo* rinkinius apmokymui ir testavimui gautos klaidos panašios į klaidas, gautas naudojant tik *MediaEval 2015* rinkinį. Galima teigti, jog modelių struktūros yra bendrinės, nepriklausiančios nuo duomenų rinkinio.

- Modeliai apmokyti naudojantis *PMEmo* rinkiniu ir testuoti *MediaEval 2015* rinkiniu pasirodė geriau nei modeliai apmokyti *MediaEval 2015* rinkiniu ir testuoti *PMEmo* rinkiniu. Kai kuriais atvejais modeliai, apmokyti su *PMEmo* duomenimis, pasižymėjo mažesne testavimo klaida testuojant su *MediaEval 2015* rinkiniu nei tos pačios struktūros modeliai, apmokyti su *MediaEval 2015* duomenimis. Galima to priežastis – *PMEmo* duomenų rinkinys, nepaisant mažesnės žanrų aibės, pasižymi didesne įvairove.

- Modeliai testavimui naudojant tą patį duomenų rinkinį kaip ir apmokymui pasirodė geriau. Tai galėjo lemti duomenų rinkinio įverčių skirtumai, kilę dėl skirtingos anotavimo metodikos, vertintojų pasirengimo, kultūrinių skirtumų.

4.3. Modelių greitaveikos įvertinimas

Modelio sparta vertinant muzikos emocijas yra svarbus aspektas. Potencialūs programėlės naudotojai gali turėti didelį kiekį audio kūrinių, kurių ilgas apdorojimo laikas gali būti nepriimtinas. Taip pat ilgai trunkantis nuotaikų įvertinimas gali koreliuoti su didesniu energijos suvartojimu, kuris itin aktualus mobilių įrenginių savininkams.

4.3.1. Modelių greitaveikos įvertinimo metodika

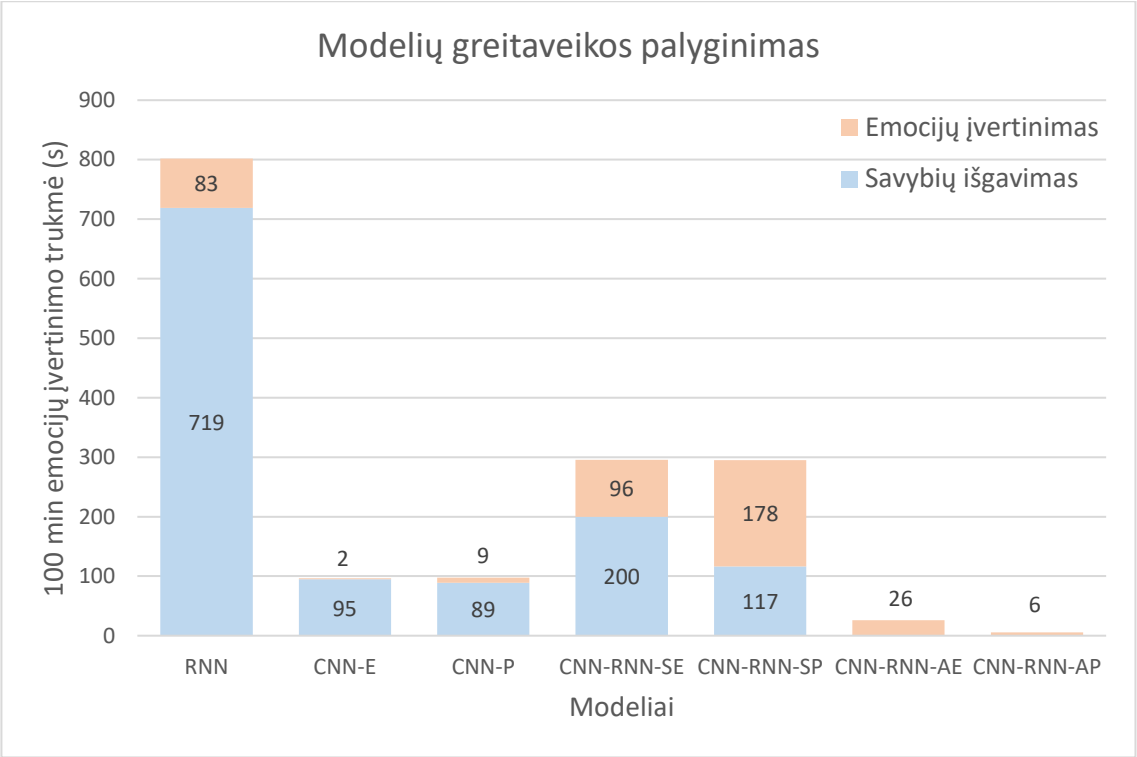
Greitaveikos įvertinimui buvo atlikti eksperimentai, kurių metu buvo analizuojamos vieno 100 minučių trukmės garso įrašo emocijos. Eksperimentų metu buvo fiksuojami atskirų nuotaikos nustatymo algoritmo žingsnių – garso signalo paruošimo, savybių išgavimo bei savybių konvertavimo į nuotaikų įverčius trukmės. Tyrimas buvo atliekamas su kiekvienu muzikos emocijų nustatymo metodu, kiekvienam metodui eksperimentą kartojant 3 kartus bei išvedant rezultatų vidurkį.

Greitaveikos eksperimentai buvo atliekamas fiziniame *Samsung Galaxy 10e* įrenginyje. Įvertinimui buvo naudota *TensorFlow Lite* versija nepalaikanti grafinės plokštės ar kitų neuroninių tinklų greitaveiką gerinančių sąsajų.

4.3.2. Modelių greitaveikos įvertinimo rezultatai

4.1 pav. pateikta diagrama su greitaveikos tyrimo rezultatais. Diagramoje pateiktos tik garso signalo savybių išgavimo bei emocijų įvertinimo trukmės, neįtraukiant audio failo skaitymo, dekodavimo bei pirminio apdorojimo trukmės remiantis prielaida, kad visiems metodams ši dedamoji turėtų būti pastovi. Remiantis rezultatais matyti, kad RNN modelis, naudojantis garso įrašo savybių išgavimą, pasirodė itin prasčiausiai. Sudėtingas audio signalo apdorojimas trunka per ne lyg ilgai lyginant su kitais, paprastesniais savybių išgavimo metodais.

CNN modelių greitaveika tenkina mobiliajai programėlei keliamus reikalavimus, vienos minutės apdorojimas trunka apie vieną sekundę. CNN ir RNN apjungiantys modeliai veikia apie 3 kartus lėčiau. CNN-RNN-SE modelio savybių išgavimo trukmę prailgina darbas su 44,1 kHz diskretizavimo dažnio signalu, kuris reikalauja apdoroti dvigubai daugiau duomenų nei 22 kHz signalas. CNN-RNN-SP emocijų įvertinimo trukmę galimai prailgino didelis įvesties apimtis – 512 diapazonų keturis kartus padidina duomenų kiekį lyginant su standartinėmis 128 diapazonų spektrogramomis.



4.1 pav. Greitaveikos įvertinimo rezultatai.

Modeliai, dirbantys su neapdorotu garso signalu pasirodė geriausiai dėl nenaudojamų įprastų savybių išgavimo algoritmų. Šių modelių emocijų nustatymo trukmė beveik nykstama lyginant su failo nuskaitymo ir dekodavimo trukme. Šie modeliai būtų optimalus variantas taikyti mobiliajame įrenginyje jeigu jei pasižymėtų geresniu tikslumu.

Išvados

1. Literatūros analizės metu buvo nustatyti metodai bei modelių architektūros, tinkamos dainų nuotaikų nustatymo uždaviniui spręsti. Remiantis analizės rezultatais, tyrimo metu buvo realizuoti 4 skirtingų architektūrų modeliai.
2. Atlikus dainų nuotaikos įvertinimo eksperimentą buvo nustatyta, kad geriausiai nuotaikos pozityvumą vertinantis modelis prilygo literatūros šaltiniuose pateikiamiems rezultatams, tačiau nė vienas iš bandytų modelių nepasiekė gero energingumo įvertinimo tikslumo.
3. Atlikus eksperimentą modelius paruošiant ir testuojant skirtingų šaltinių duomenų rinkiniais paaiškėjo, kad pasirinkti modeliai ir jų konfigūracijos universalios, neoptimizuotos konkrečiam duomenų rinkiniui. Taip pat buvo pastebėta, kad testavimui naudojant kitą duomenų rinkinį nei apmokymui, testavimo klaidos buvo didesnės. Tai gali indikuoti, kad įverčiai priklauso nuo duomenų surinkimo metodikos, vertintojų, ir nėra visiškai objektyvūs.
4. Atlikus greitaveikos tyrimus buvo nustatyta, kad savybių iš garso signalo išgavimas yra lėtoji nuotaikos įvertinimo algoritmo dedamoji, kurią minimizavus pasiekiama optimali greitaveika.
5. Atsižvelgiant į tikslumo bei greitaveikos rezultatus, mobiliojoje programėlėje būtų prasminga naudoti CNN-P ir CNN-RNN-AE modelių kombinaciją, naudojant pirmo modelio pozityvumo išvestį ir antro modelio energingumo išvestį. Tokia kombinacija veiktų sparčiai bei pasižymėtų geriausiu pozityvumo įverčiu bei priimtinu energingumo įverčiu.

Literatūros sąrašas

[1] P. Ekman, „An argument for basic emotions“, *Cognition and Emotion*, t. 6, nr. 3–4, p. 169–200, geg. 1992, doi: 10.1080/02699939208411068.

[2] J. A. Russell, „A circumplex model of affect“, *Journal of Personality and Social Psychology*, t. 39, p. 1161–1178, 1980, doi: 10.1037/h0077714.

[3] X. Yang, Y. Dong, ir J. Li, „Review of data features-based music emotion recognition methods“, *Multimedia Systems*, t. 24, nr. 4, p. 365–389, liep. 2018, doi: 10.1007/s00530-017-0559-4.

[4] Y. Song, S. Dixon, ir M. Pearce, „Evaluation of Musical Features for Emotion Classification“, pristatytas Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012, spal. 2012.

[5] A. Aljanaki ir M. Soleymani, „A data-driven approach to mid-level perceptual musical feature modeling“, *arXiv:1806.04903 [cs, eess]*, birž. 2018, Žiūrėta: 2021 m. vasario 28 d. [Interaktyvus]. Adresas: http://arxiv.org/abs/1806.04903

[6] R. Panda, R. Malheiro, ir R. P. Paiva, „Musical Texture and Expressivity Features for Music Emotion Recognition“, p. 9, 2018.

[7] A. Aljanaki, Y.-H. Yang, ir M. Soleymani, „Emotion in Music task: lessons learned“, p. 3.

[8] A. Aljanaki, Y.-H. Yang, ir M. Soleymani, „Developing a benchmark for emotional analysis of music“, *PLOS ONE*, t. 12, nr. 3, p. e0173392, kovo 2017, doi: 10.1371/journal.pone.0173392.

[9] X. Li *ir kt.*, „A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction“, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, kovo 2016, p. 544–548. doi: 10.1109/ICASSP.2016.7471734.

[10] M. Malik, S. Adavanne, K. Drossos, T. Virtanen, D. Ticha, ir R. Jarina, „Stacked Convolutional and Recurrent Neural Networks for Music Emotion Recognition“, *arXiv:1706.02292 [cs]*, birž. 2017, Žiūrėta: 2021 m. vasario 28 d. [Interaktyvus]. Adresas: http://arxiv.org/abs/1706.02292

[11] R. Sarkar, S. Choudhury, S. Dutta, A. Roy, ir S. K. Saha, „Recognition of emotion in music based on deep convolutional neural network“, *Multimed Tools Appl*, t. 79, nr. 1, p. 765–783, saus. 2020, doi: 10.1007/s11042-019-08192-x.

[12] R. Orjesek, R. Jarina, M. Chmulik, ir M. Kuba, „DNN Based Music Emotion Recognition from Raw Audio Signal“, *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, bal. 2019, p. 1–4. doi: 10.1109/RADIOELEK.2019.8733572.

[13] S. Chaki, P. Doshi, P. Patnaik, ir S. Bhattacharya, „Attentive RNNs for Continuous-time Emotion Prediction in Music Clips“, p. 12.

[14] M. Soleymani, A. Aljanaki, ir Y.-H. Yang, „DEAM: MediaEval Database for Emotional Analysis in Music“, p. 3.

[15] K. Zhang, H. Zhang, S. Li, C. Yang, ir L. Sun, „The PMEmo Dataset for Music Emotion Recognition“, *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ICMR ’18. New York, NY, USA: Association for Computing Machinery, birž. 2018, p. 135–142. doi: 10.1145/3206025.3206037.

[17] D. Bogdanov *ir kt.*, „Essentia: an audio analysis library for music information retrieval“, *Proceedings - 14th International Society for Music Information Retrieval Conference*, 2013.

Priedai

1 priedas. Modelių tikslumo įvertinimas apmokymui bei testavimui naudojant skirtingus duomenų rinkinius.

Modelis	Apmokymo rinkinys	Testavimo rinkinys	Testavimo klaida			
			MAE		RMSE	
			E	P	E	P
RNN	PMEmo	PMEmo	0,223	0,220	0,265	0,261
	PMEmo	MediaEval	0,284	0,262	0,321	0,325
	MediaEval	PMEmo	0,314	0,271	0,357	0,344
CNN-A	PMEmo	PMEmo	0,211	0,220	0,257	0,269
	PMEmo	MediaEval	0,211	0,280	0,274	0,348
	MediaEval	PMEmo	0,240	0,282	0,298	0,348
CNN-P	PMEmo	PMEmo	0,196	0,205	0,231	0,252
	PMEmo	MediaEval	0,202	0,203	0,242	0,239
	MediaEval	PMEmo	0,224	0,263	0,277	0,317
CNN-RNN-SE	PMEmo	PMEmo	0,186	0,203	0,235	0,260
	PMEmo	MediaEval	0,187	0,241	0,237	0,284
	MediaEval	PMEmo	0,230	0,260	0,288	0,321
CNN-RNN-SP	PMEmo	PMEmo	0,175	0,210	0,211	0,257
	PMEmo	MediaEval	0,223	0,247	0,286	0,293
	MediaEval	PMEmo	0,233	0,266	0,291	0,324
CNN-RNN-AE	PMEmo	PMEmo	0,217	0,228	0,264	0,273
	PMEmo	MediaEval	0,184	0,272	0,237	0,347
	MediaEval	PMEmo	0,224	0,261	0,277	0,306
CNN-RNN-AP	PMEmo	PMEmo	0,204	0,218	0,242	0,259
	PMEmo	MediaEval	0,179	0,276	0,236	0,345
	MediaEval	PMEmo	0,213	0,258	0,269	0,304