Универзитет Св. Кирил и Методиј - Скопје
Скопје, Република Македонија
Факултет за електротехника и информациски технологии – Скопје

КЛАСИФИКАЦИЈА НА МУЗИКАТА ПО ЖАНРОВИ СО ПОМОШ НА МАШИНСКО УЧЕЊЕ

Ментор:
Проф. д-р Бранислав Геразов

Изработил:
Мартин Величковски 136/2014

Скопје 2020

I. ВОВЕД

Звукот (музиката) е представена во форма на аудио сигнал кој има свои параметри, како: фрекфенција, амплитуда (обично во логаритамска скала – децибели), ширина на опсегот и слично Со зголемувањето на музичките жанрови и поджанрови тешко може да се се направи класификација на песните што се слушаат денес. Со растот на базите на податоци за музика на Интернет – еден од начините за категоризација и организација на песните кои се засноваат на жанрот, кој пак е идентификуван со некои карактеристики на музиката како што се ритмичката структура, хармонската содржина и инструментацијата.

При постоење на можноста за автоматско класифицирање на музиката и обезбедување на специјални ознаки на истата присутна во библиотеката на корисникот, заснована врз жанрот, пожелно е да се користат услугите за аудио стримингот, како што се Spotify и iTunes.

Постојат повеќе обиди да се класифицира музиката со машинско учење.

Овие истражувања се базираат врз апликацијата на алгоритмите за машинско учење (ML) за да се идентификува и класифицира жанрот на дадена аудио датотека.

II. ИСТРАЖУВАЧКИ ПРИСТАП КОН ПРОБЛЕМОТ

Класификацијата на музичкиот жанр понекогаш се смета за субјективна материја. Жанровските ознаки на музичките фајлови на песните често се означени од уметникот или корисниците.

Истражувањето за машинско учење е на високо ниво со исклучителен успех во препознавањето на поставените задачи.

Целта на овие истражувања е да се обезбеди увид во класификацијата на музичките жанрови со користење на машинско учење, односно, надгледувано учење со примена на конволуциски невронски мрежи.

Успехот во препознавањето на звучната слика е она што ме инспирираше при изработката на овој проект, бидејќи класификацијата на музичките жанрови претставува проблем на истражување во сферата на музиката и дигиталното процесирање на аудио сигналите.

Овој истражувачки пристап се заснова на споредување на сличноста на карактеристиките на аудио сигналот, со цел да се имплементираат резултатите од истражувањата, бидејќи овој пристап бара дефинирање на сличноста со метриката, која се користи за мерење на сличноста помеѓу аудио сигналите.

Во истражувањата ги користам конвулуциските невронски мрежи, според системот на класификација на музички жанрови на Чои и други истражувачи. Тие споредиле изведба на неколку архитектури на CNN (конволуциски невронски мрежи) со CRNN(конволуциски рекурентни невронски мрежи) за класификација на музички жанрови.

Од резултатите при истражувањата заклучиле дека при CNN тренирањето се покажало како многу ефикасно, главно, кога се користат мел-спектрограми како аудио репрезентација во споредба со STFT. CNN се особено погодни за предвидување на високо ниво музички карактеристики како што се акорди и beat-ови, бидејќи тие овозможуваат хиерархиска структура која се состои од средни карактеристики на повеќекратни временски периоди.

Суровиот формат за аудио датотеки е обична бранова форма, како што обично се гледа во аудио уредувањето во софтвери како што се Audacity. Кога се рефлектираат како податоци, брановите форми се чуваат како еднодимензионални низи.

Истражувањата поврзани со аудиото обично бараат претходна обработка на сурови бранови, при што се потенцираат аудитивните карактеристики. Заедничките пристапи за пред-обработката вклучуваат брза Фуриерова трансформација (STFT) и Мел-спектрограм.

Основната механика на овие аудио методи за преработка се потпираат врз Фуриеовата трансформација, чија дефиниција следи:

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi} dx \qquad (1.1)$$

Во горенаведената равенка, f е оригинална функција на времето а x го претставува времето. $\hat{f}$ е трансформирана функција на фреквенција кадешто ξ ја претставува фреквенцијата.

Фуриеовата трансформација во голема мерка била инспирирана од проучувањето на фуриеовиот серијал, со која се разложува комплицираната функција од збир на едноставни бранови. Резултатот од фуриеовата трансформација е банки со повеќе фреквенции со соодветни магнитуди.

Во форма на податоци, ова би тгребало да биде еднодимензионална низа, кадешто информациите за фреквенцијата се содржат во индексите а информациите за магниудата се содржат во броевите. Популарниот начин на користење на фуриеовата трансформација се нарекува брза Фуриеова трансформација (STFT).

Примената на фуриоевата трансформација на мали прозорци од бранова форма ги комбинира резултатите во дводимензионална низа. Со соодветна стапка на земање примероци, долга аудио датотека може да се разложи на неколку парчиња и секоја може да се трансформира одделно.

Комбинираната матрица ја покажува врската време-фреквенција, со вредностите во секоја решетка што ја претставува магнитудата на одредена фреквенција за одредено време, а резултатот од STFT се нарекува спектрограм.

Со аудио податоците, особено кај музиката, популарната надградба на спектрограмот претставува употребата на мел-скала, наместо линеарно распоредена скала на фреквенција. Мел-скалата се заснова врз споредби извршени на теренот. Како што фреквенцијата се зголемува, така мел-интервалите бараат сè поголеми и поголеми фреквентни скокови. Формула за претворање *f* (Херц) во *m* (Мел) е

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \qquad (1.2)$$

Со мел-скалите, мел-спектрограмот нагласува пониски фреквенции и повисока компресија, приближна на човечката аудитивна перцепција.

II.1. Невронски мрежи

Историјата на невронските мрежи во вештачката интелигенција може да се проследи уште од 1940-тата година, но нивните перформанси станаа значајни дури во последните дваесет години.

Меѓу многуте модели на невронски мрежи, во овој проект се користат конволуциски невронски мрежи (CNN). Секаков вид на невронската мрежа се состои од неврони (јазли) и рабови. За време на фазата на учење, невроните од првиот слој се земаат како влезни, кои ги ставаме во функција за активирање, а резултатите се прикажани во следниот слој.

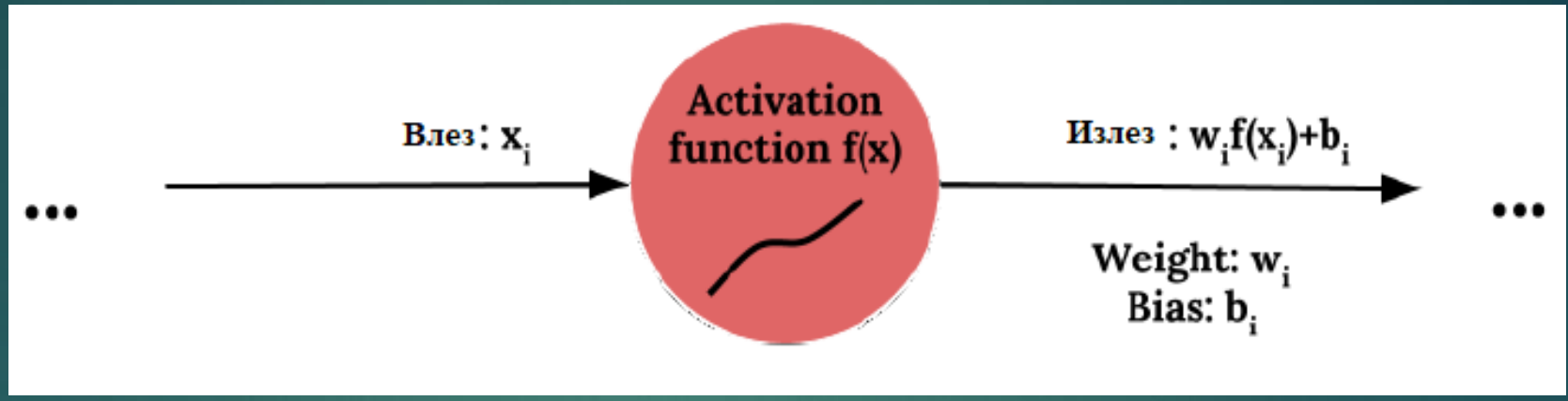
$$\sigma(x) = \frac{1}{1+e^{-x}} \qquad (1.3)$$

Функцијата за активирање обично ја мапира влезната вредност во одреден опсег за да се означи неговиот потенцијал за да го „активира“ следниот неврон. Функција за активирање на пример е сигмоид функција, прикажана во равенката 1.3, која го мапира влезот во опсег од [0, 1]. Како што влезот станува поголем, изводот од првиот ред на сигмоидната функција станува се помал.

Овие овде модели главно ја користат коригираната линеарна единица (ReLU) како функција за активирање. При тоа се има предност во намалувањето на проблемот со градиент на исчезнување, како и воведување на реткост во мрежата. Следува нејзината равенка.

$$f(x) = \begin{cases} 0 & if x > 0 \\ x & if x \geq 0 \end{cases} \qquad (1.4)$$

Механизмот за единечен неврон е прикажан на слика 2.



Слика 2. Еден неврон во невронска мрежа

Потоа земаме неврони од следниот слој со пондерирана сума на резултати од претходниот слој и ја повторуваме истата постапка, сè додека не го достигне крајниот излезен слој, кадешто бројот на невроните во последниот слој е еднаков на бројот на категориите, што значи дека секој излезен неврон претставува можност за категорија.

II .1.1. Конволуциски невронски мрежи

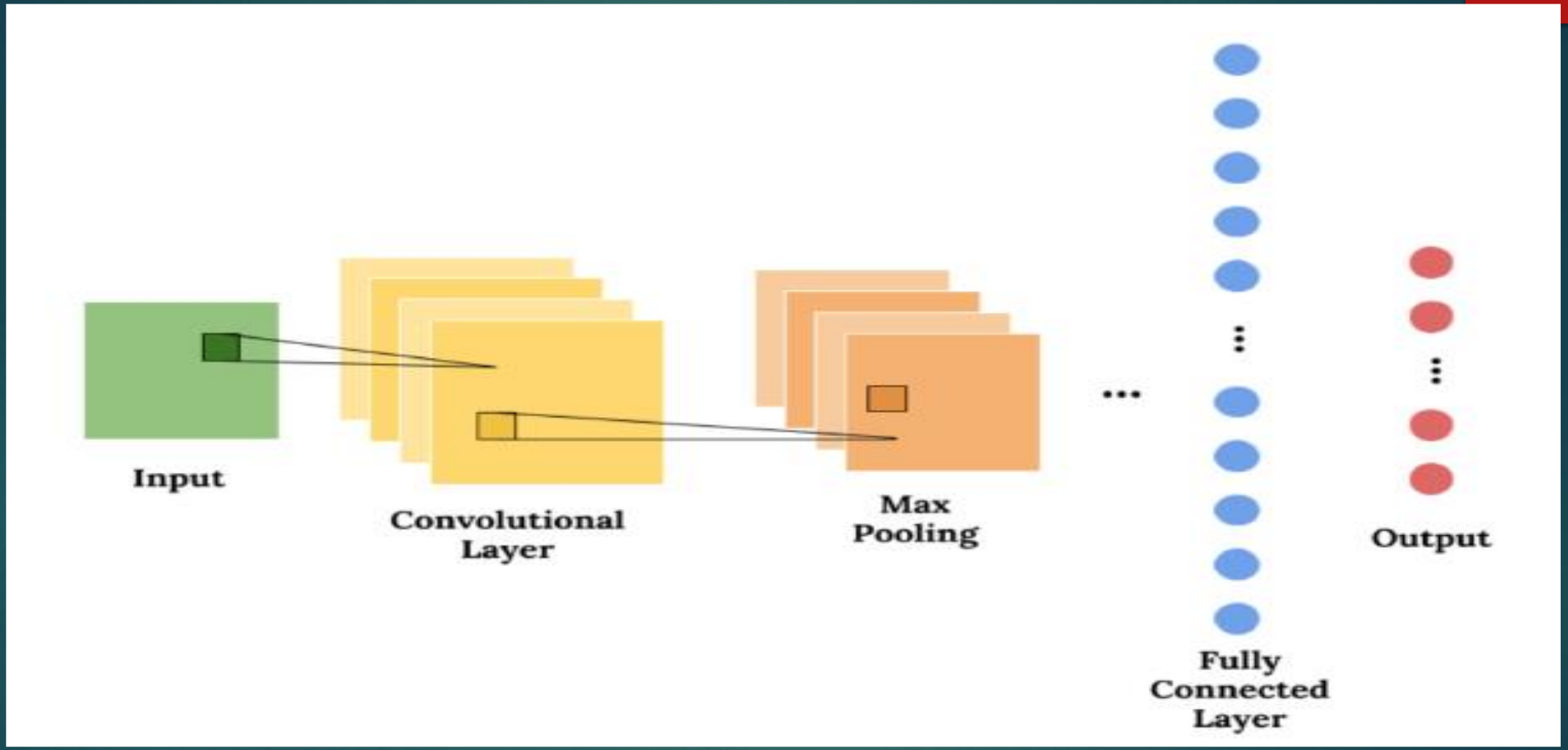
Традиционалните повеќеслојни модели на перцептронот се целосно поврзани и функционираат прилично добро во задачи за препознавање на слика. Сепак, тие не скалираат добро слики со висока резолуција поради ограниченоста на компјутерската моќ. Покрај тоа, повеќеслојните перцептрони не ја земаат во предвид просторната структура на визуелните обрасци, а со тоа и далечните пиксели можат да имаат исто влијание во препознавање на областа како поблизок пиксел.

Невронските мрежи го надминуваат овој проблем со спроведување на 3Д слоеви кои се поврзани само со мал регион од претходниот и филтрите во ист слој ги делат тежините и пристрасностите. Затоа, бројот на параметри во еден збиен слој е даден со формулата:

$$(n^2 \cdot x) \times 2 \quad (1.5)$$

каде n е страничната должина за еден мал регион, а x е бројот на филтрите во овој слој.

Општиот модел на CNN е прикажан на слика 1.3.



Слика 3. Општа форма на моделот CNN

II. 1.2. База на податоци на GTZAN

Базата на податоци што ја користев претставуваше база на податоци собрана од Г. Тзанетакис и П. Кук, која се нарекува GTZAN база на податоци. Оваа база на податоци беше собрана од различни извори, вклучувајќи лични CD-а, радио, снимање со микрофони и слично. Се состои од 100 аудио клипови со должина од 30 секунди за секој од десетте жанрови, во вкупна вредност од 1000 песни.

Десетте жанрови се блуз, класична музика, кантри, диско, хип-хоп, џез, метал, поп, реге и рок. Сите песни се моноаурални со брзина на земање примероци од 22050Hz.

Genre	Blues	Classical	Country	Disco	Hiphop	Jazz	Metal	Pop	Reggae	Rock
Size	100	100	100	100	100	100	100	100	100	100

Табела 1. GTZAN жанрови и број на аудо датотеки

GTZAN е широко користен во истражувањето за класификација на музичкиот жанр уште од објавувањето во 2002 година. Ја одбрав оваа база на податоци како почетна точка затоа што беше добро организирана и често цитирана од многу истражувачи, бидејќи обезбедува кредибилитет при користењето на референтна рамка за мрежните перформанси. Меѓутоа, имаше неколку недостатоци при користењето на оваа база на податоци.

Најограничувачкиот фактор беше неговата големина.

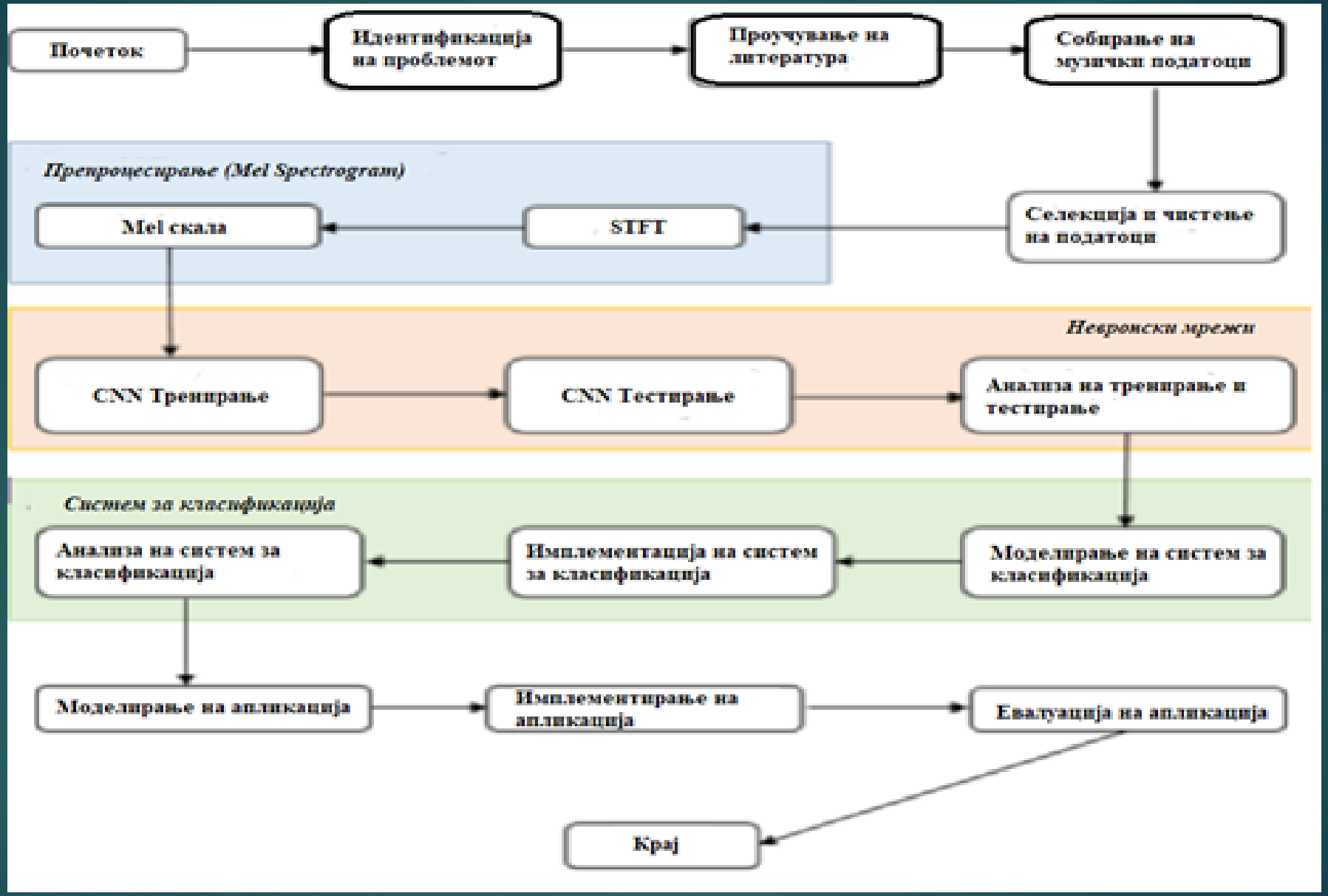
III. МЕТОДОЛОГИЈА ПРИ ГРАДЕЊЕ НА МОДЕЛОТ

Во овој дел се дадени детали за чекорите за обработка на податоци, проследени со описот на двата предложени пристапа кон овој проблем со класификацијата.

Првиот чекор беше да се најде база на податоци. Вториот чекор беше да се претворат суровите аудио-датотеки во форма со појасни музички карактеристики. Потоа тренирам невронска мрежа за претходно обработените податоци за класифицирање на жанровите. Имаше неколку компоненти во споменатата постапка, вклучително и изборот на базата на податоци, методите на пред обработка и структурите на невронската мрежа.

Во овие истражувања се следат резултатите со користење на Мел-спектрограми за аудио репрезентација и CNN за екстракција на карактеристиките на музички жанрови. Истражувањето се состои од неколку фази (, а тоа се:

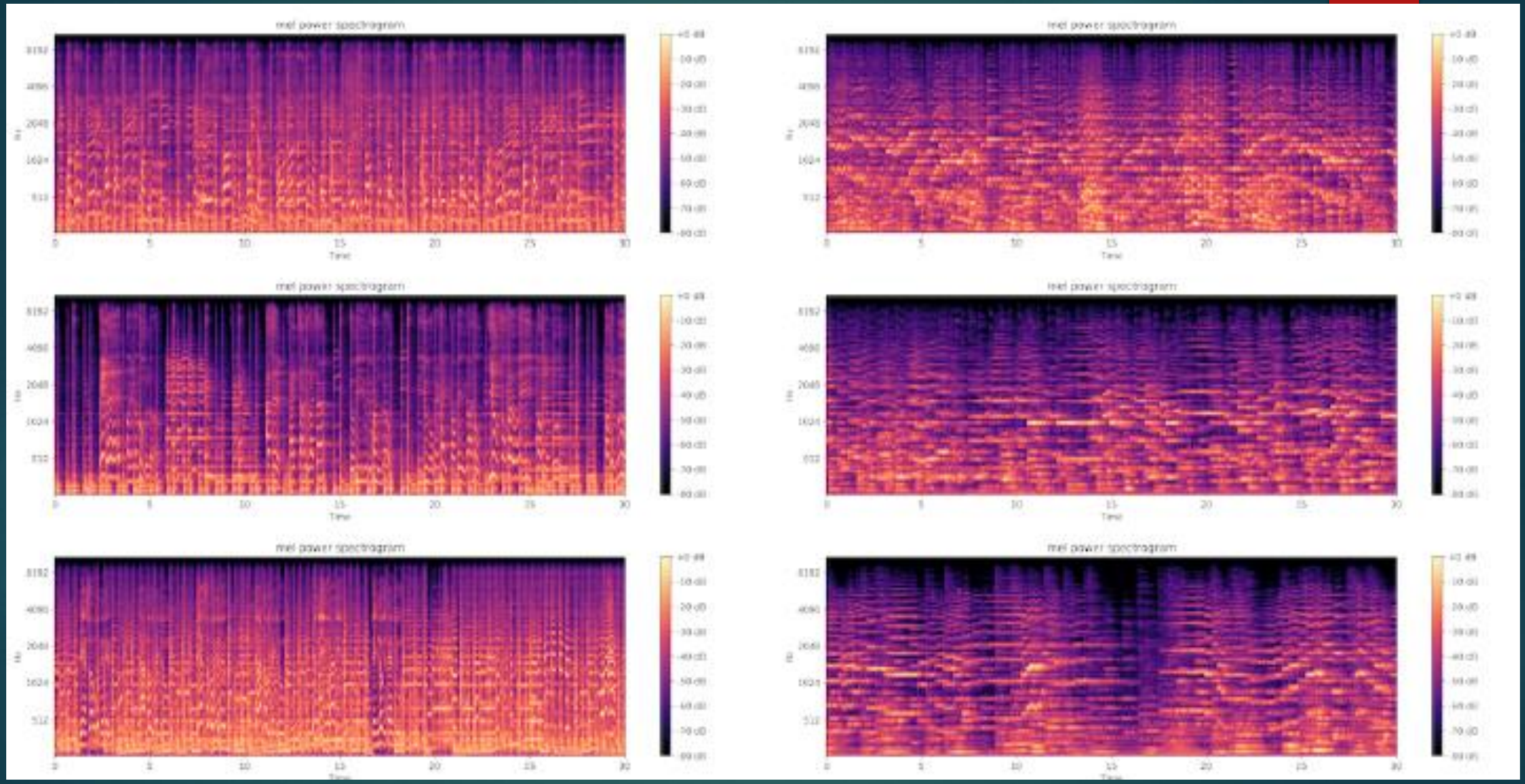
1. Почетна фаза (вклучува идентификација на проблемот и проучување на литература),
2. Собирање музички податоци,
3. Преработување на аудио датотеки,
4. Процесирање на Мел-спткторам,
5. Селекција и чистење на податоци,
6. Моделирање на невронска мрежа.



Слика 4. Чекори при истражување на системот за класификација на музички жанрови

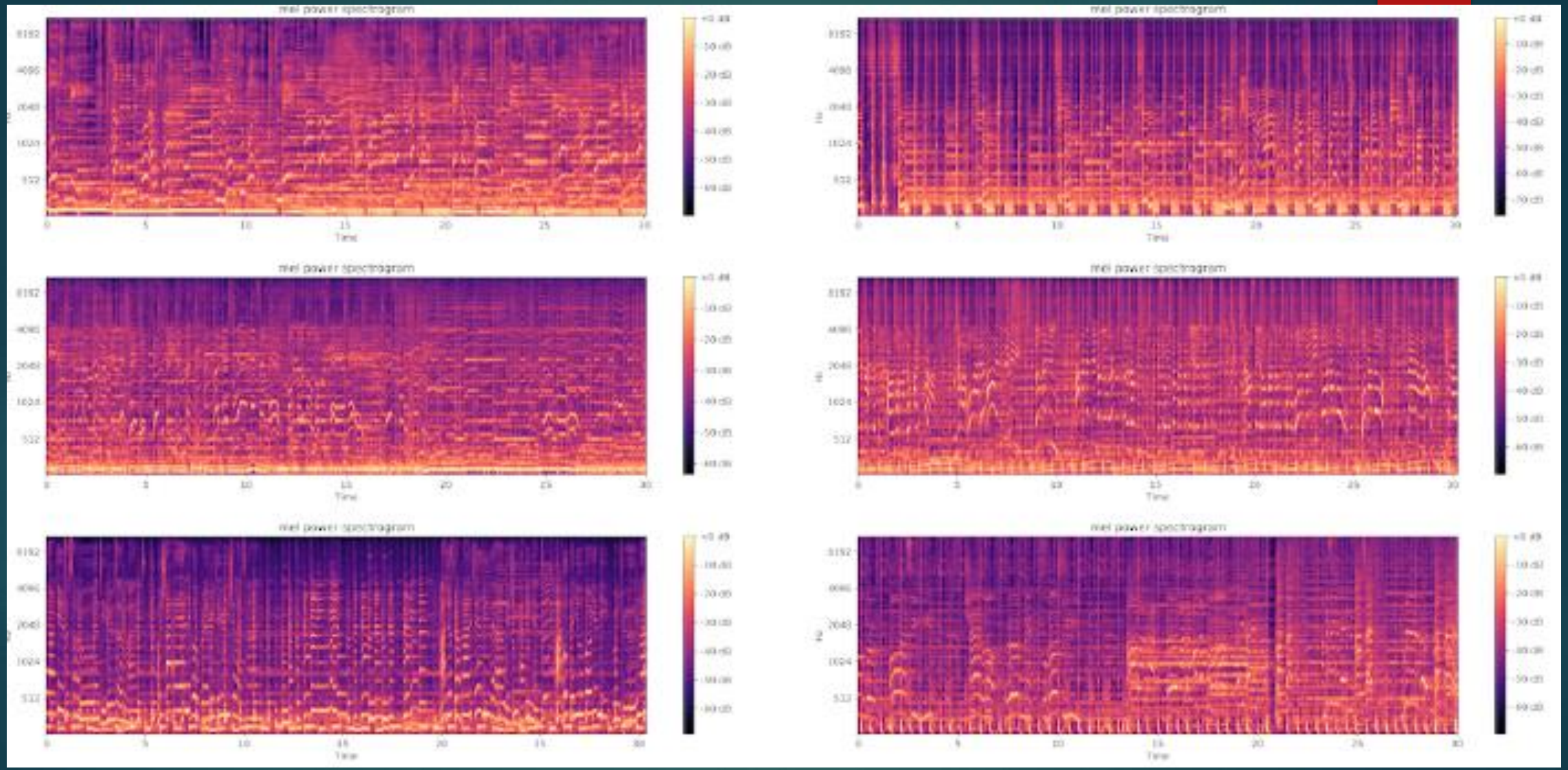
III.1. Пред-обработка на податоците

Она што ги разликува мел-спектрограмите од регуларните спектрограми е нивното фреквенциско растојание на у-оска. Мел-фреквенциското растојание подобро ја приближува скалата на слухот кај човекот, каде се потенцираат пониски фреквенции и се компресираат повисоки фреквенции. Ја користев библиотеката LibROSA за да се произведе мел-спектрограм за секоја песна. Преработените песни беа поделени на помали клипови како индивидуални семплови. Ја менувавме должината на ваквите примероци за да ја најдеме оптималната поделба, со цел да се добие визуелна перцепција на резултатите од мел-спектрограмот, па така по случаен избор од секој жанр одбрав три музички датотеки кои ги исцртав. Подолу се дадени графиконите за сите жанрови во базата на податоци GTZAN.



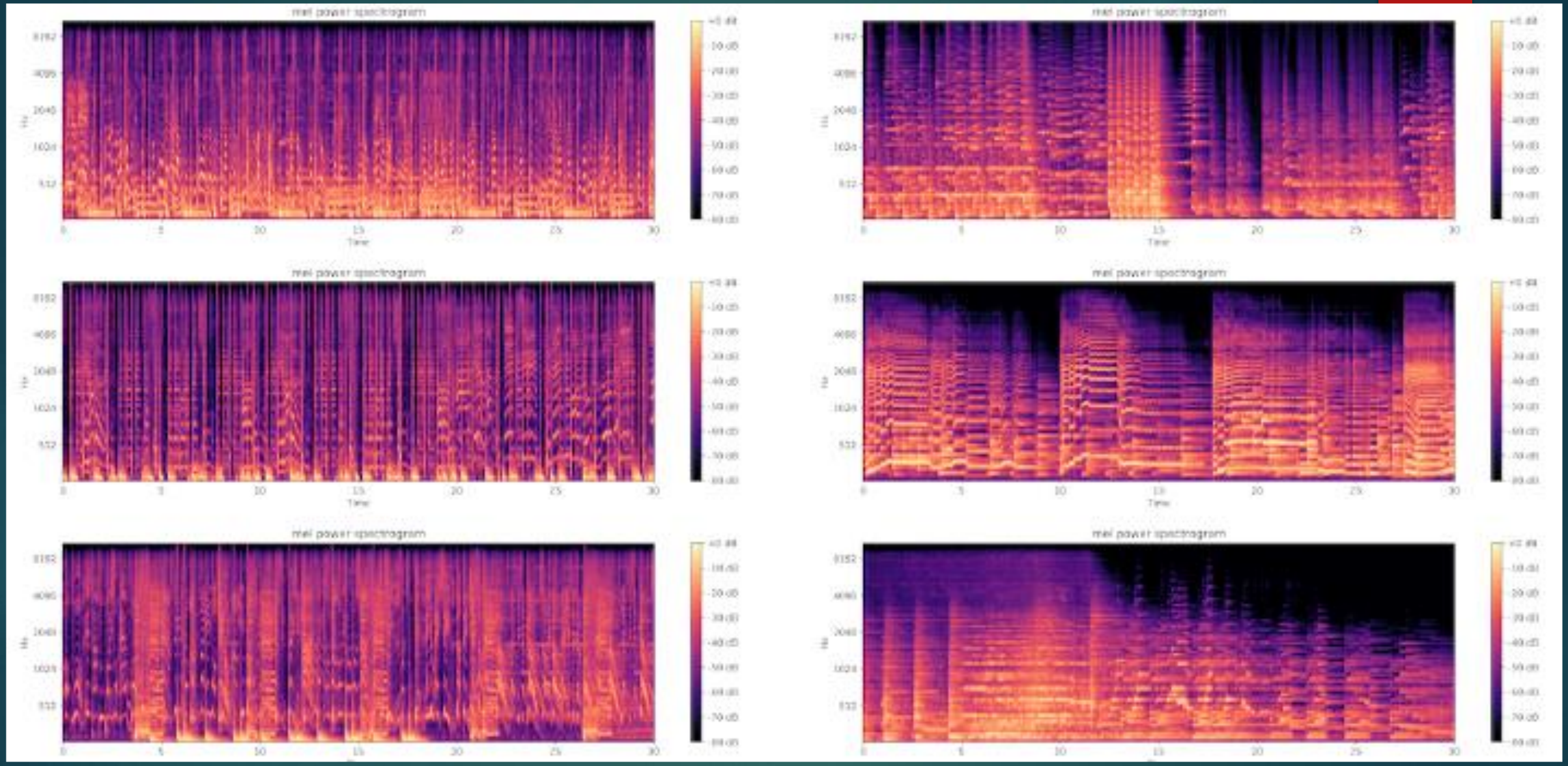
(а) блуз

(б) класична музика



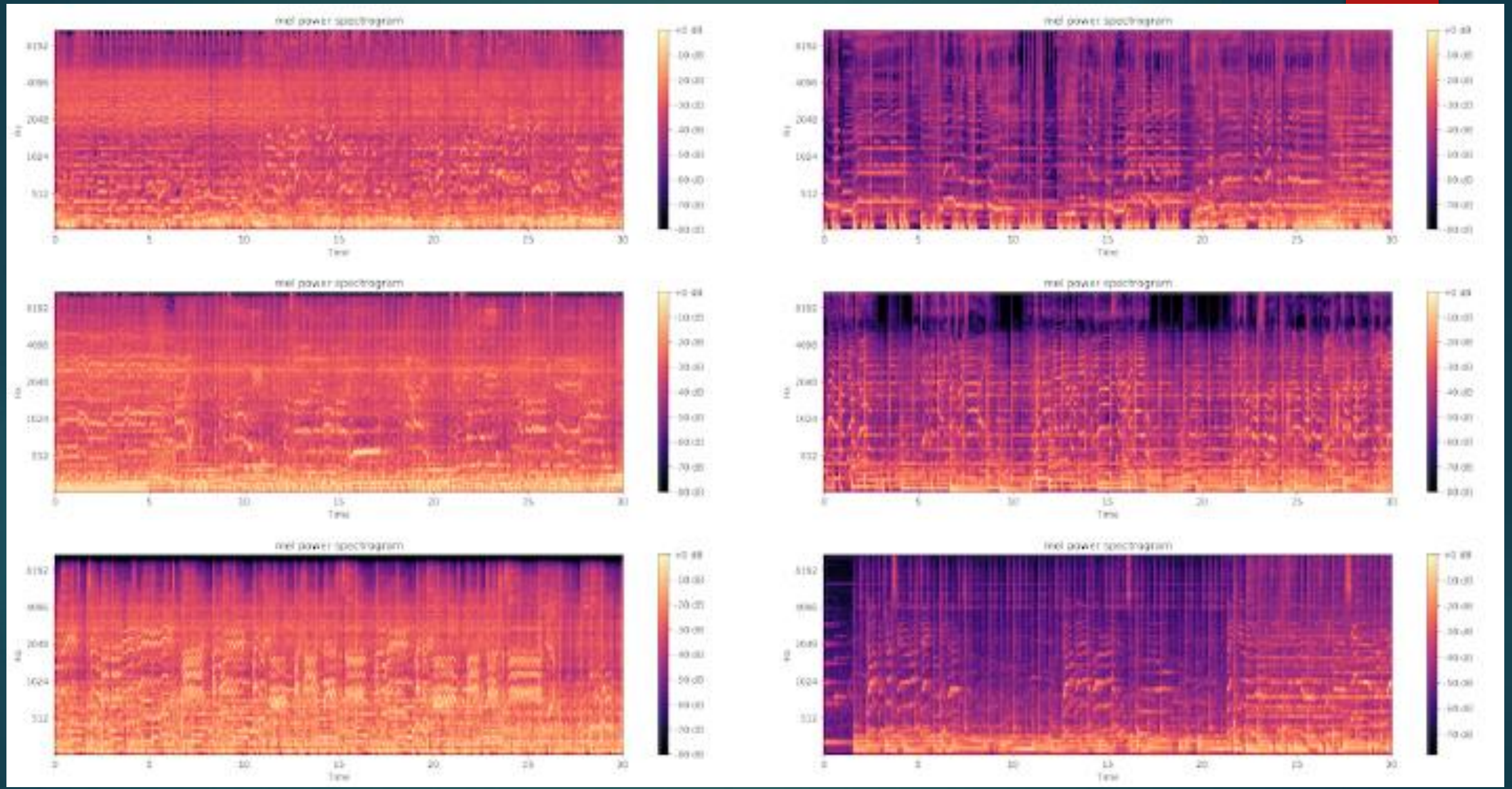
(в) Кантри

(г) Диско



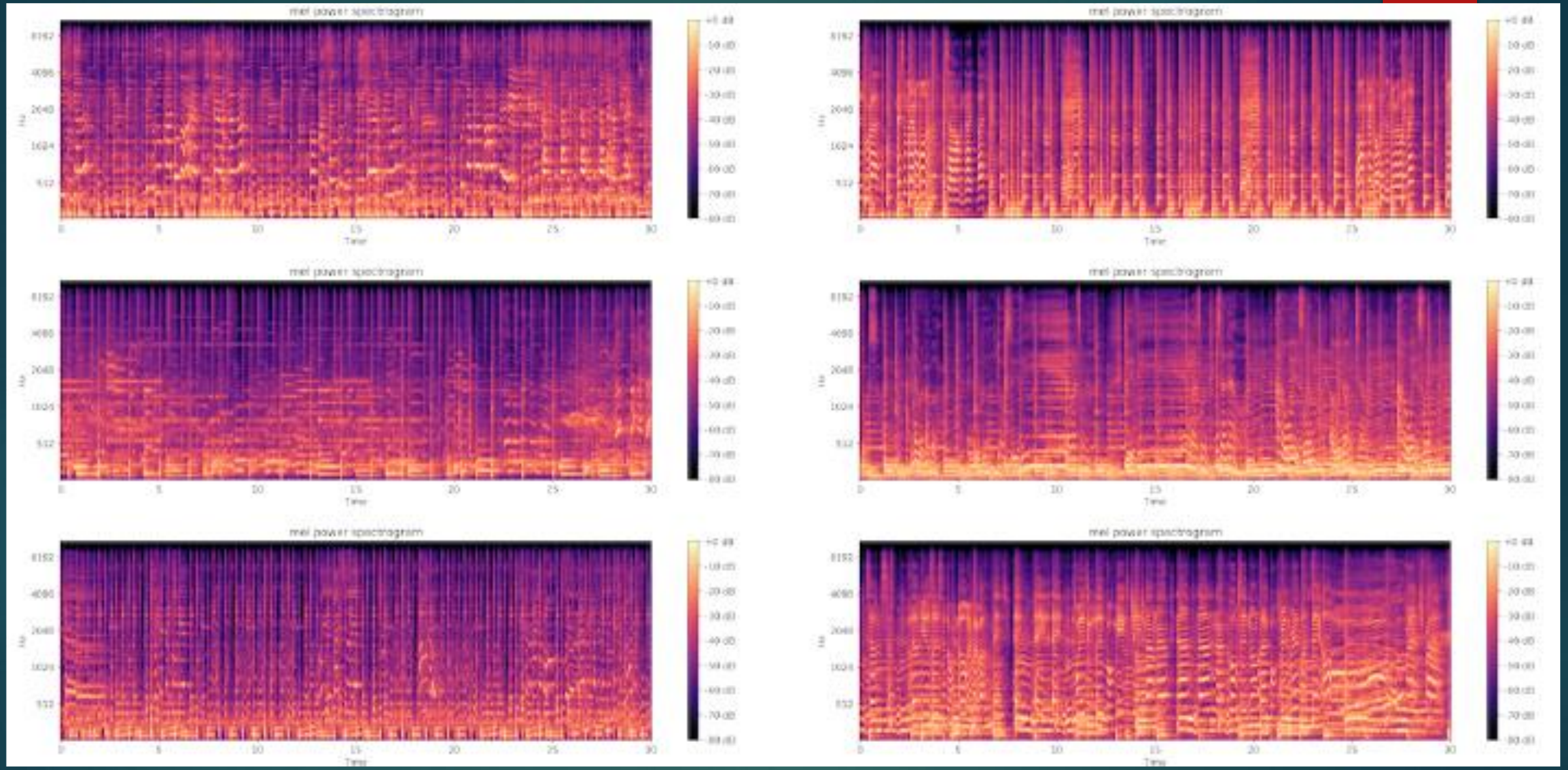
(д) хипхоп

(ѓ) џез



(е) Метал

(ж) Поп


mel power spectrogram
mel power spectrogram
mel power spectrogram
mel power spectrogram
mel power spectrogram
mel power spectrogram

з) Реге

(s) Рок

Сликите покажуваат сличности помеѓу жанровите, како и различности меѓу нив. Со разгледување на случајните примероци од мел-спектрограмите, може да се заклучи дека јасно се забележуваат различни карактеристики за неколку жанрови, но не за сите.

Музичките датотеки од жанрот класична музика имаат долги хоризонтални линии во спектрограмите, а музичките датотеки од жанрот метал низ фреквенцискиот спектар имаат тешка активност така што низ целиот мел-спектрограм се појавува светлина. Музичките датотеки од диско музиката, поради нивните постојани beat-ови, имаат долги вертикални линии со еднакви интервали на спектрограмите. Hiphop песните исто така се карактеризираат со вертикални линии низ целиот спектар на фреквенцијата, но интервалите не се еднообразни низ целиот тек на песната, што најверојатно се должи на фактот дека хип-хоп песните понекогаш го менуваат својот ритам од време на време. Иако би можеле да толкуваме некои визуелни одлики со музичко знаење за овие жанрови, сепак се забележува дека има некои сличности што не можат да се објаснат. Исто така, се забележува дека некои од жанровите изгледаат многу слични, како диско, рок и поп.

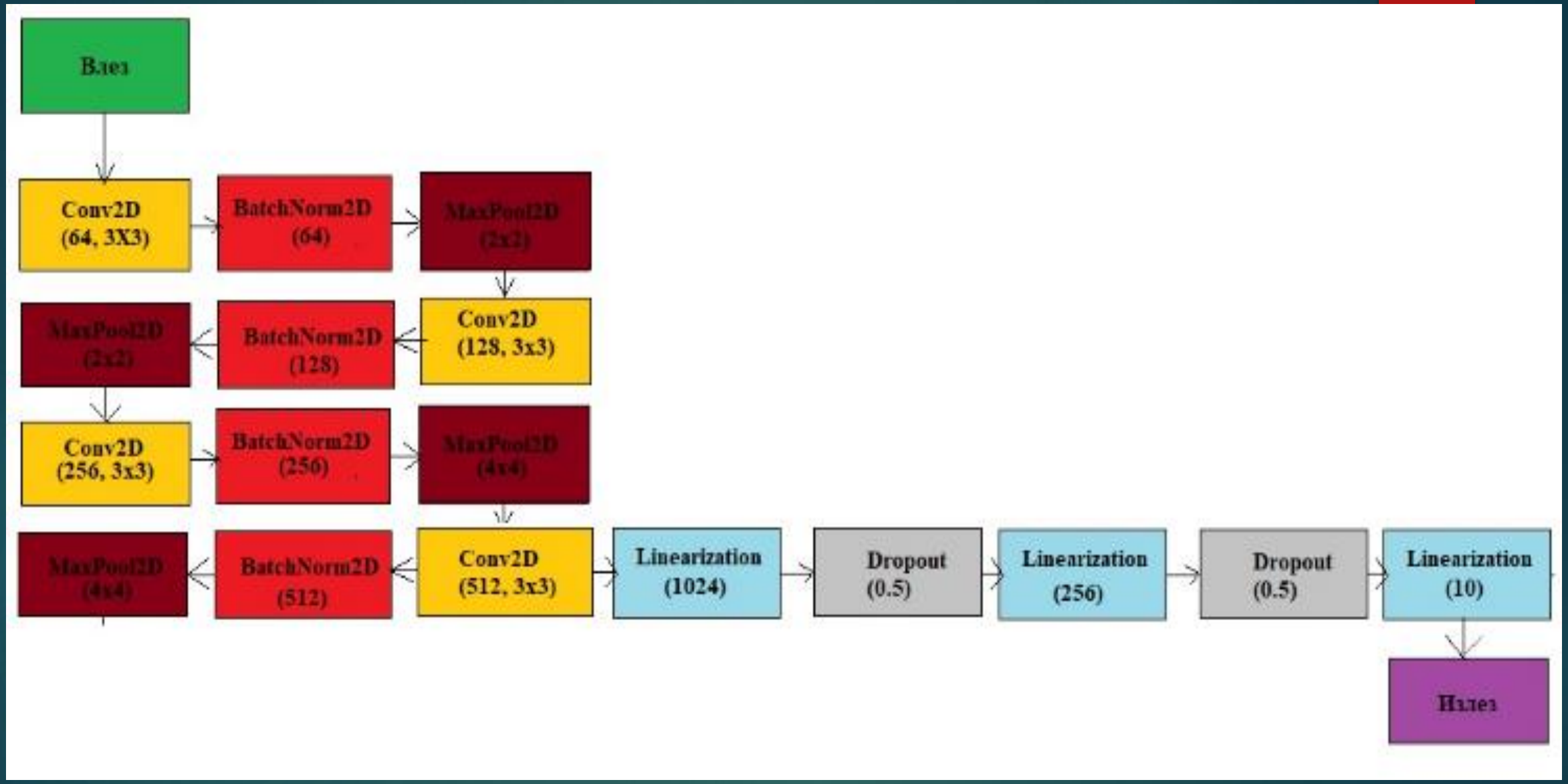
III.2 Примена на невронски мрежи во проектот

Невронските мрежи се суштината на овој проект.

Во овој случај користев конволуциска невронска мрежа, бидејќи моделите на конволуциските невронски мрежи (CNN) се добри за препознавање и класификација на сликата, па дури и поедноставните модели се во можност да дадат голема точност. Покрај тоа, следејќи ги нашите спектрограми, се чини дека потребен е само мал дел од песната за да се добијат информации за утврдување на нејзиниот жанр

III.2.1 Конволуциски невронски мрежи

Како што претходно беше кажано конволуциските невронски мрежи обично се користат за препознавање на слика и од секој примерок се очекува да има три димензии: висина, ширина и три канали во боја, сепак, нашите податоци немаа ниту канали во боја, ниту аудио еквивалент на нив (стерео влезни канали). Затоа, едноставно додадовме дополнителна димензија во нашите податоци за апроксимација на црно-белите слики. Моделот CNN се состоеше од четири групи на конволуциски слоеви проследени со BatchNormalization и MaxPooling слој кои меѓусебно се надоврзани, а потоа линеаризирани со Dropout стапка од 0,5, па повторно линеаризирани со Dropout стапка од 0.5 и на крајот е додадена уште една линеаризација, но без Dropout стапка. Овој модел е изработен во PyTorch (Деталите за нашиот модел се прикажани на Слика 5.



Слика 5. Модел на CNN

III.2.2. Оптимизатор RMSprop

Во овие истражувања е употребен RMSprop оптимизаторот, кој спаѓа во доменот на адаптивните методи за стапка на учење, на кои им се зголемува популарноста во последните години. RMSprop е необјавен алгоритам за оптимизација дизајниран за невронските мрежи, кој за прв пат беше предложен од Џефри Хинтон на предавањето број 6 од онлајн курсот „Невронски мрежи за машинско учење“.

Главната идеја на RMSprop е да се задржи подвижниот просек на квадратните градиенти за секоја тежина, а потоа градиентот го делиме со квадратниот корен од средниот квадрат. Затоа се нарекува RMSprop. Со математички равенки, правилото за ажурирање изгледа вака:

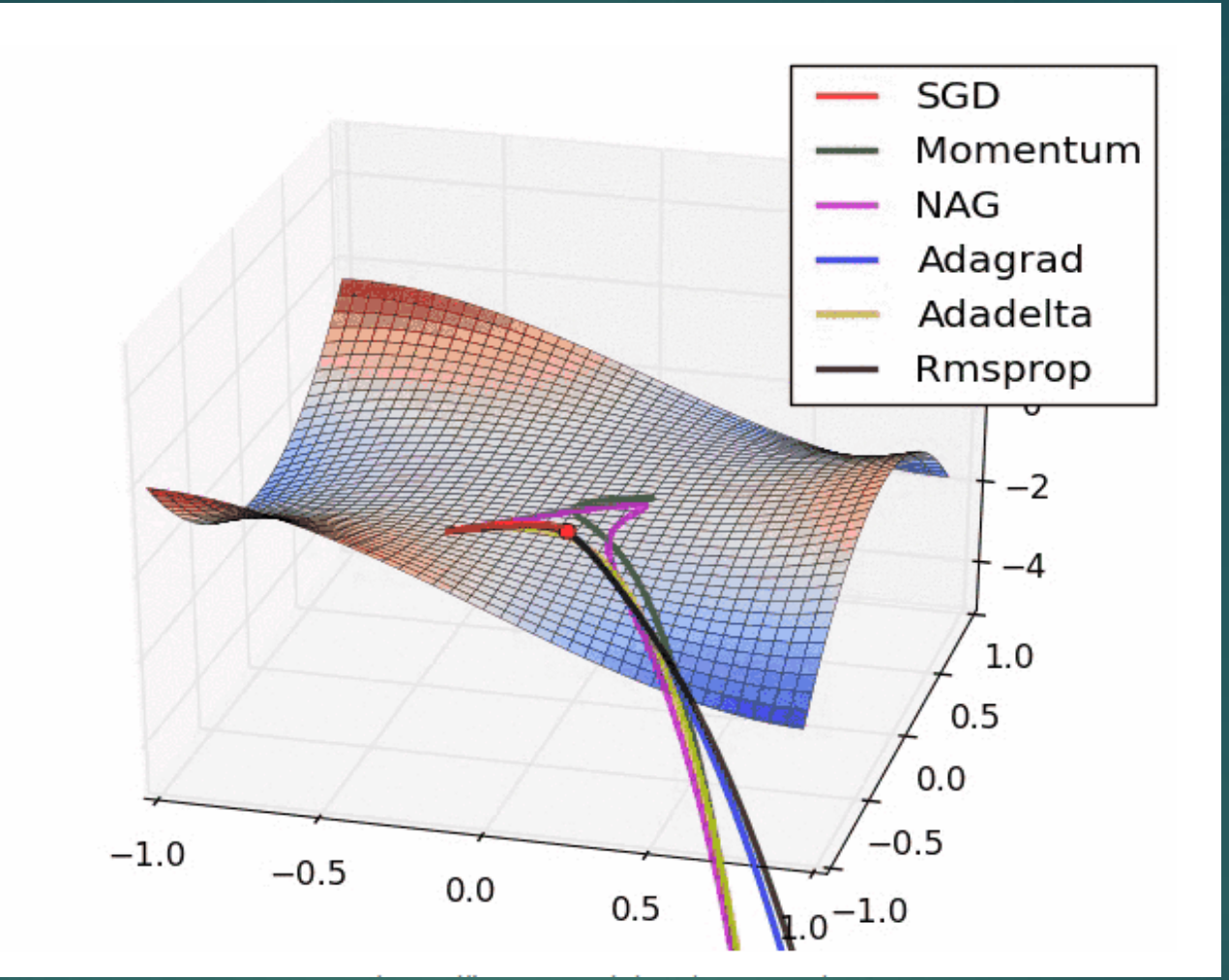
$$E[g^2]_t = \beta E[g^2]_{t-1} + (1-\beta)\left(\frac{\partial C}{\partial w}\right)^2 \quad (1.6)$$

$$w_t = w_{t-1} - \frac{n}{\sqrt{E[g^2]_t}} \frac{\partial C}{\partial w}$$

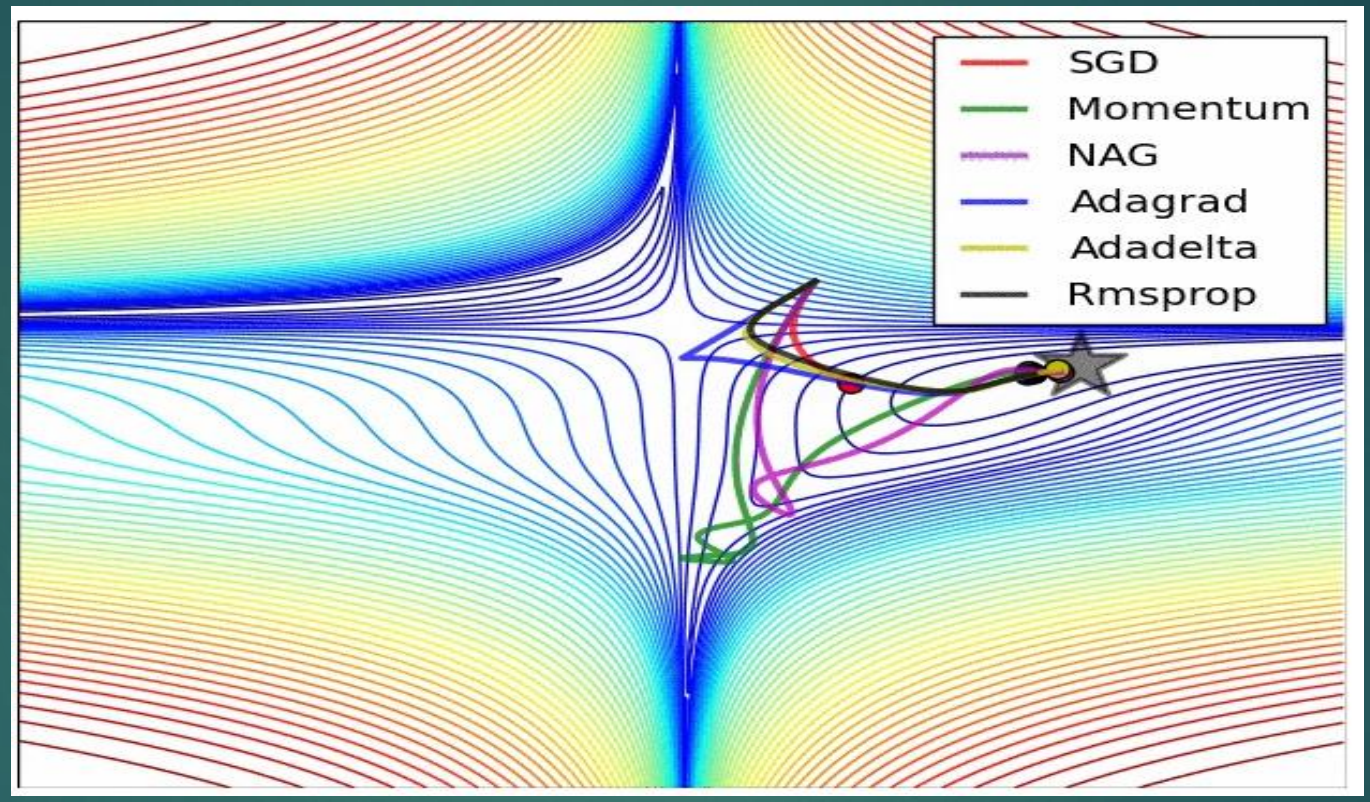
E [g] - подвижен просек на квадратни градиенти. dC / dw - градиент на функцијата на трошоците во однос на тежината.
n - стапка на учење. Просечен параметар што се движи бета (добра стандардна вредност - 0,9)

Од горенаведената равенка може да се забележи прилагодување на стапката на учење со тоа што се дели квадратниот корен од квадратниот градиент. Но, бидејќи имаме проценка на градиентот на тековната мини-серија, треба да го користиме подвижниот просек на истиот. Стандардна вредност за подвижниот просечен параметар што се користи е 0,9, кој работи многу добро за повеќето апликации. Со RMSprop ние сè уште ја одржуваме проценката на квадратни градиенти. Наместо да дозволиме таа проценка постојано да се акумулира во текот на тренингот, ние го одржуваме подвижниот просек на истата.

Резултатите добиени при истражувањето од овие прекрасни визуелизации за различни алгоритми за оптимизација, покажуваат како тие се однесуваат во различни ситуации.



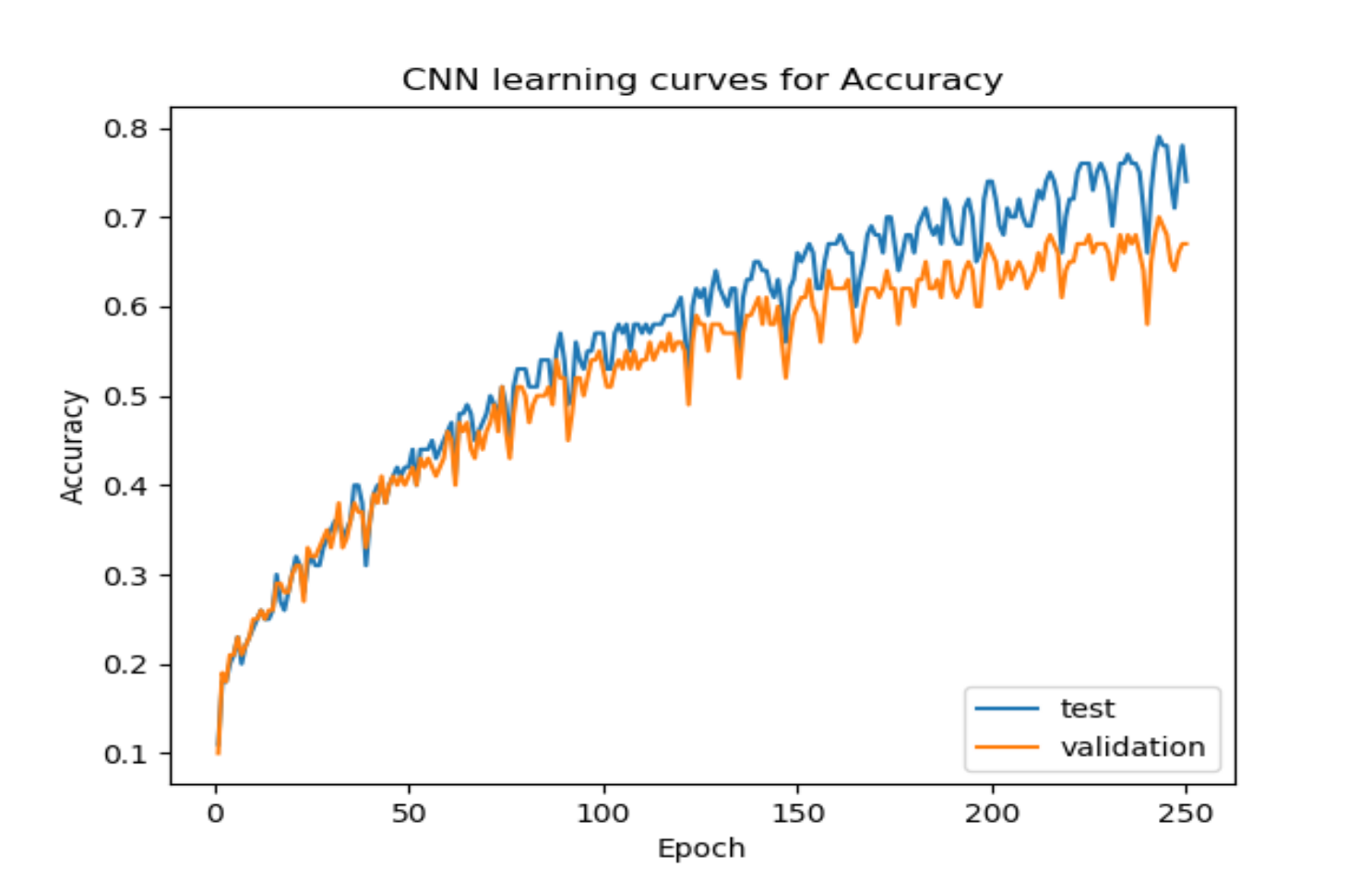
Анимацијата 1 (https://imgur.com/a/Hqolp#NKsFHJb).



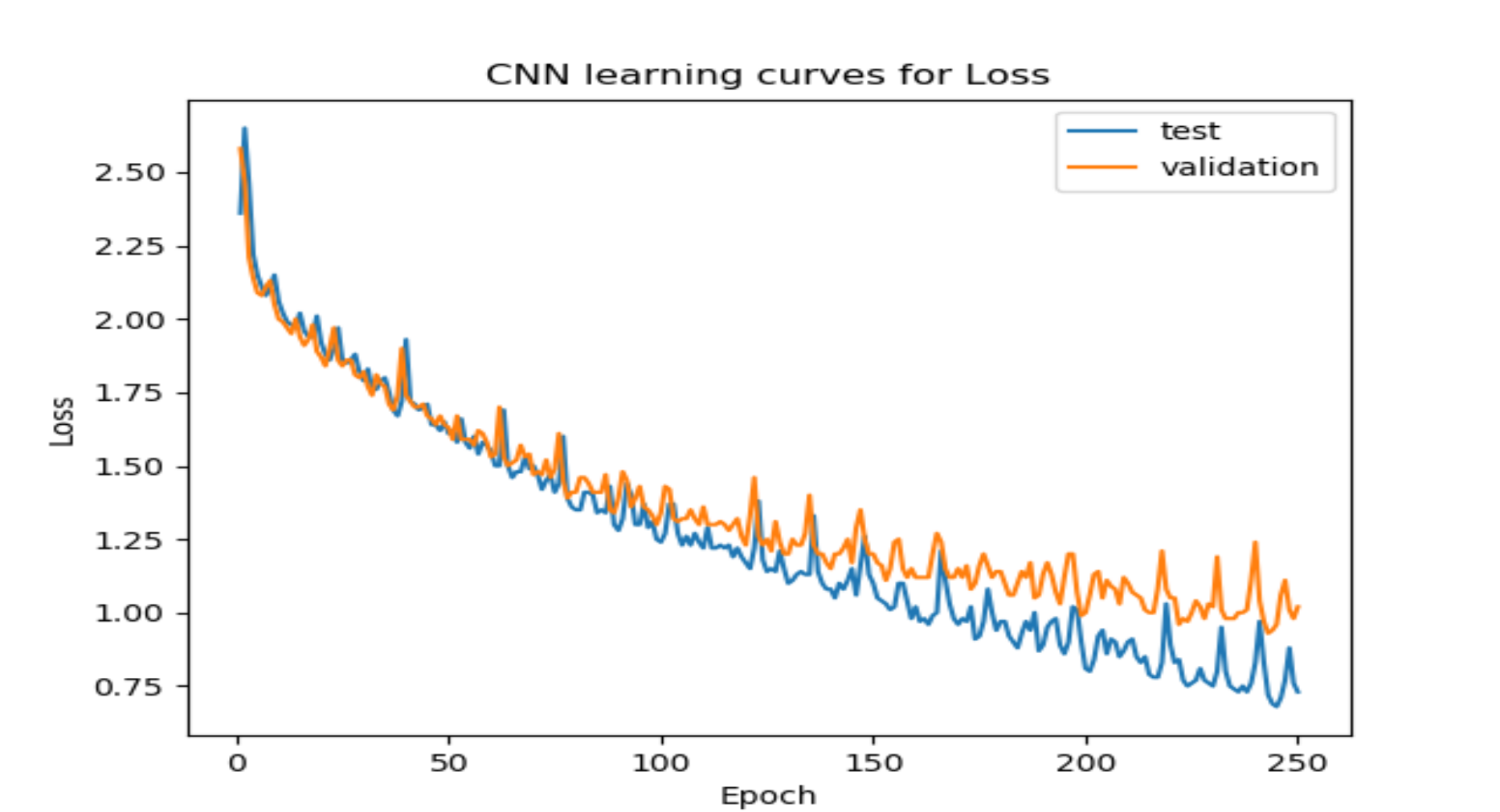
Анимацијата 2 (https://miro.medium.com/max/1240/0*o9jCrrX4umP7cTBA)

III.3. Агрегација

По завршувањето на тренирањето, го зачувавме моделот од невронската мрежа со својата структура и тежина. Потоа holdout групата (не се користи за време на тренинг или тестирање) од истата база на податоци се доставува до моделот за да се предвидат жанровите. Исто така, исцртани се и криви на учење за точност на тест и точност на валидација, како и загуба при тест и загуба при валидација. (Види ги кривите на слика 8 и 9)



Слика број 8.
Криви на учење за точност

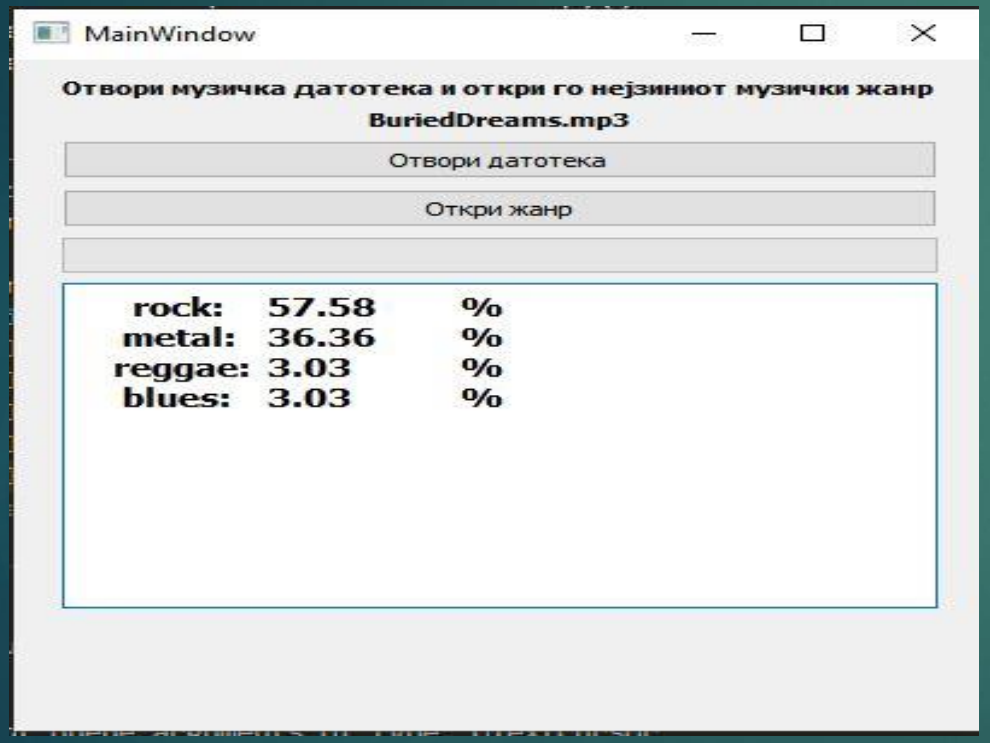


Слика број 9. Криви на учење за загуба

Од тренирањето на моделот добиен е резултат за тест на точност кој изнесува 65%.

IV. ЗАКЛУЧОК

Постојат многу варијабли при утврдување на точноста на класификацијата. Во овој семинарски труд одбрав да ги анализирам ефектите од: методот на пред-обработка, должина на примерок, број на жанрови, избор на база на податоци и структура на невронска мрежа. Изгледот на апликацијата може да се види подолу:



https://www.youtube.com/watch?v=kxJfwDP65Bs

Моделот реализиран за оваа проектна задача е релативно едноставен и неоптимизиран за конкретниот проблем. Овој модел има слаба способност за генерализација. За да се добие вистински оптимално решение за овој проблем, потребно е дополнително да се оптимизираат или пак да се примени сосема различен пристап.

Врз основа на нашите анализи, можеме да предложиме идно истражување со додавање на други музички одлики со цел да се подобри точноста на системот за класификација на музичките жанрови

Ви благодарам на вниманието