https://doi.org/10.7494/978-83-67427-06-7_5

# Klasyfikacja emocji w muzyce filmowej z wykorzystaniem uczenia głębokiego

Tomasz Ciborowski[1*], Szymon Reginis[1], Dawid Weber[1,2], Adam Kurowski[1,2], Bożena Kostek[2]

[1] Katedra Systemów Multimedialnych, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska

[2] Laboratorium Akustyki Fonicznej, Wydział Elektroniki, Telekomunikacji i Informatyki, Politechnika Gdańska

[*] s165501@student.pg.edu.pl

## Streszczenie

Praca przedstawia zagadnienia związane z klasyfikacją emocji w muzyce filmowej. W artykule zaproponowano model emocji zawierający dziewięć stanów emocjonalnych, do których przypisany jest kolor zgodnie z teorią koloru w filmie. Kolejne kroki eksperymentu obejmowały wybór muzyki filmowej do testów (baza Epidemic Sound), przygotowanie założeń ankiety oraz modelu emocji wykorzystywanych w testach odsłuchowych, a także konstrukcję ankiety. Ankieta została zrealizowana za pomocą formularzy Google. Przeprowadzono również analizę statystyczną wyników uzyskanych w testach subiektywnych. Częścią pracy jest opracowanie aplikacji do automatycznej klasyfikacji emocji na podstawie muzyki filmowej. Aplikacja wykorzystuje splotową sieć neuronową, która klasyfikuje emocje na bazie 30-sekundowych fragmentów muzyki reprezentowanych przez spektrogramy w skali melowej. W końcowej części pracy pokazano przykład wyników klasyfikacji emocji w muzyce filmowej uzyskanych za pomocą sieci neuronowej, zweryfikowanych w sposób subiektywny.

## 1. Wstęp

Muzyka jest nieodłącznym elementem filmu, stanowi bowiem część przekazu artystycznego towarzyszącego obrazowi. Jej celem jest podbudowanie lub wywołanie emocji związanych z daną sceną, ale przede wszystkim może być ona wizytówką filmu, często określając jednoznacznie jego gatunek. Klasyfikacja emocji w muzyce filmowej – w uproszczeniu – obejmuje zagadnienia związane z analizą muzyki, kolorystyką obrazu, narracją filmu oraz przekazem emocji, które mają towarzyszyć projekcji filmu. Tak więc główne funkcje muzyki filmowej to opowiadanie, tworzenie atmosfery w scenie i wywoływanie emocji w widzach.

W literaturze bardzo popularnym wątkiem stał się temat rozpoznawania emocji w muzyce (*Music Emotion Recognition*, MER) [1]. W badaniach w obszarze MER wykorzystuje się metody uczenia maszynowego – zarówno algorytmy klasyczne [2, 3, 4], jak i uczenie głębokie [5, 6]. Klasyfikacja dotyczy fragmentów lub całych utworów muzycznych, ich reprezentacji parametrycznych lub reprezentacji 2D;

tj. spektrogramów, spektrogramów w skali melowej, cepstrogramów w skali melowej, chromagramów [7–9]. Wykorzystywane są różne bazy danych utworów muzycznych, również z podziałem na gatunki muzyczne czy emocje/nastrój. Proponowane są nowe modele emocji, a także wprowadzane są zmiany do modeli emocji znanych w psychologii. Również muzyczne portale streamingowe oferują użytkownikom możliwość dopasowania muzyki do własnych zainteresowań, stanu emocjonalnego czy nastroju [5, 10, 11].

Obecnie główny nurt MER wykorzystuje uczenie głębokie, w tym różne modele sieci neuronowych. Wykorzystuje się je z powodu ich własności i konstrukcji. Sieci te zbudowane są ze sztucznych neuronów wzorowanych na neuronach biologicznych [12]. Taki model jest w stanie symulować działanie procesów w ludzkim mózgu. Przez dziesięciolecia przeprowadzono wiele badań naukowych, aby zrozumieć, w jaki sposób muzyka wpływa na ludzi dzięki wzmacnianiu lub stymulowaniu określonych emocji [12–14]. Wpływ muzyki na człowieka jest dobrze znany, ale sposób, w jaki to się dzieje, jest wciąż niezbadany. Automatyczna klasyfikacja emocji z muzyki pojawia się zarówno w literaturze [15–18], jak i w technologiach [15]. Stanowi to tło badań prowadzonych w tym obszarze. Dlatego w rozdziale 2 przytoczono prace związane z tym nurtem, pokazując przykłady modeli emocji i ich klasyfikacji. Należy przy tym zauważyć, że wątek automatycznej klasyfikacji muzyki filmowej wydaje się być niszowy.

Celem pracy jest przedstawienie eksperymentów związanych z przeprowadzeniem testów subiektywnych określających emocje w muzyce filmowej. W szczególności dotyczy to automatycznej klasyfikacji emocji w muzyce filmowej. W rozdziale 2 zaprezentowany został przegląd literatury związany z MER. W rozdziale 3 przedstawiono założenia testów subiektywnych, mających na celu weryfikację zaproponowanego modelu emocji związanych z muzyką filmową. W tym celu przygotowano formularz ankiety internetowej, w której zadaniem respondentów/słuchaczy było przypisanie emocji i koloru do danego fragmentu muzyki filmowej. Model obejmuje dziewięć etykiet emocji, silnie związanych z psychologią koloru w filmie – energiczny, ekscytujący, wesoły, relaksujący, spokojny, depresyjny, smutny, straszny, agresywny. Wyniki ankiety posłużyły do odwzorowania emocji zawartych w bazie muzyki filmowej Epidemic Sound [19] z emocjami/kolorem w zaproponowanym modelu. Następnie w rozdziale 4 przedstawiono wyniki badań ankietowych wraz z analizą statystyczną sprawdzającą słuszność zaproponowanego modelu emocji. Kolejny rozdział zawiera odniesienie do skonstruowanej aplikacji oraz przykładów wyników rozpoznawania emocji w muzyce filmowej.

Ważnym wątkiem w tej pracy jest budowa modelu głębokiego przypisywania odpowiedniej emocji do zadanego pytania w formie sparametryzowanego fragmentu muzycznego. W tym celu stworzony został zbiór danych z muzyką filmową, zawierający 420 utworów muzycznych. Następnie przygotowywany został skrypt generujący spektrogramy w skali melowej na podstawie 30-sekundowych fragmentów i zapisujący je w postaci plików .png. Kolejnym krokiem było wybranie kilku modeli splotowych sieci neuronowych, które mogą pobierać reprezentacje 2D jako dane wejściowe. Spośród różnych dostępnych modeli sieci neuronowych proponowanych przez platformę Keras wybrano pięć spełniających powyższy warunek. Proces uczenia podzielony został na kilka etapów, w których sprawdzana była dokładność modeli. Ostatecznie do testów końcowych wybrano sieć charakteryzującą się najwyższą dokładnością. W rozdziale 5 zawarto szczegółowe informacje dotyczące klasyfikacji emocji oparte na uczeniu głębokim. Przedstawiono również aplikację do klasyfikacji emocji z wybranego utworu muzyki filmowej oraz pokazane przykłady jej działania. W końcowej części pracy przedstawiono wnioski oraz zarys dalszych prac.

## 2. Reprezentacja emocji i klasyfikacja

### 2.1. Reprezentacja emocji/nastroju

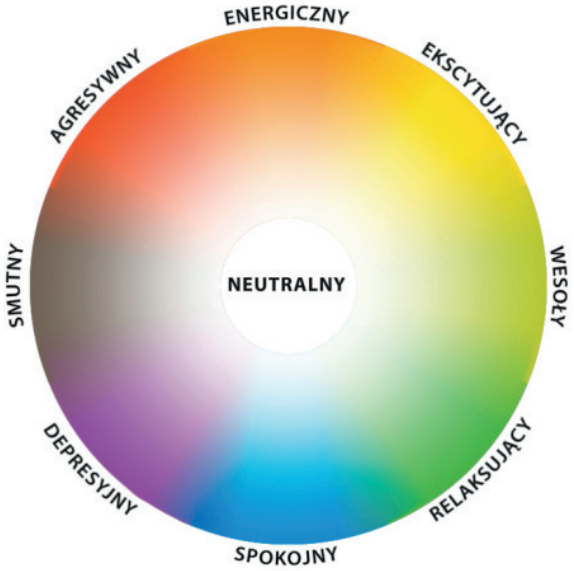
W literaturze dotyczącej klasyfikacji emocji/nastroju wyróżnia się dwa główne podejścia [2]:

- kategorialne (*categorical*) – emocje w modelach są opisywane za pomocą etykiet, które można przypisać do odpowiednich klas;
- wymiarowe (*dimensional*) – emocje są definiowane na podstawie ich lokalizacji w wyznaczonej przestrzeni, a każda emocja traktowana jest jako punkt w zdefiniowanej przestrzeni.

Jednym z przykładów reprezentacji z podziałem na kategorie emocji jest lista 66 deskryptorów zaproponowana przez Kate Hevner, które zostały przydzielone do ośmiu grup. Grupy są pokazane na okręgu – te o przeciwnych znaczeniach leżą naprzeciw siebie [20]. Jest to jedno z pierwszych badań nad ekspresją w muzyce.

Model emocji Russella [21, 22] jest przykładem modelu wymiarowego. Zakłada on, że emocje rozmieszczone są na płaszczyźnie, która zdefiniowana jest za pomocą dwóch wymiarów *V*/*A*: walencji (*valence*, *V*) i pobudzenia (*arousal*, *A*). Walencja jest zmienną dotyczącą emocji negatywnych i pozytywnych; energia jest zmienną spokojnych lub bardziej budujących emocji. Powstałe cztery regiony na płaszczyźnie można podzielić na entuzjazm, lęk, depresję i zadowolenie [23, 24]. Reprezentację modelu Russella w literaturze definiuje przedstawienie 28 deskryptorów opisujących emocje w 2D. Model Thayera jest najbardziej uproszczoną wersją dwuwymiarowego modelu emocji. Osie x i y wskazują odpowiednio poziomy energii i napięcia. Z kolei model Tellegena–Watsona–Clarka pokazuje wysoki stan pozytywny i wysoki stan negatywny jako dwa niezależne wymiary [25, 26].

Plewa i Kostek wprowadziły model rozmyty, w którym kolor płynnie przechodzi z jednego w drugi, co jest przedstawiane na palecie rozłożonej na okręgu [27, 28]. Ważną częścią tego modelu emocji jest etykieta odnosząca się do obszaru zwanego „neutralnym", zawartego w przekroju współrzędnych *x* i *y* (przedstawionego na rysunku 1). Model ten powstał na podstawie obserwacji (z przeprowadzonych testów subiektywnych), których głównym wnioskiem było stwierdzenie, że trudno jest określić intensywność emocji/nastroju jednoznacznie. Jednocześnie uczestnicy testu wskazywali, że również przejścia między emocjami powinny być rozmyte. Wszystkie te uwagi prowadzą do idei rozmycia granic w proponowanym modelu emocji/nastroju. Model z rozmytymi granicami i poziomem emocji/nastroju przedstawiono na rysunku 1.




**Rys. 1.** Zmodyfikowany model emocji z gradientem kolorów [27–29]

Istnieje wiele modeli emocji, które odnoszą się do klasyfikacji emocji z muzyki (MER) [3, 16, 17]. Autorzy prezentowanego w artykule badania postanowili jednak posłużyć się własnym opisem i przypisaniem emocji do kolorów. Zaproponowany model został wykorzystany do opracowania aplikacji mającej na celu automatyczne przypisywanie emocji muzyce filmowej na podstawie modelu głębokiego uczenia. Służyło to do porównania, czy badani, biorący udział w ankiecie, przypisują utwór muzyczny do emocji podobnie jak sieć neuronowa.

## 2.2. Klasyfikacja emocji

W literaturze można spotkać różnorodne podejścia do klasyfikacji emocji w muzyce. Zależą one głównie od przyjętych metod i algorytmów rozpoznawania, m.in. klasyfikacji opartej na całym utworze

muzycznym lub jego fragmencie, jak również zróżnicowanej reprezentacji sygnału audio. Wśród metod reprezentacji można wyróżnić przygotowane surowe próbki dźwięku, reprezentację 2D sygnału muzycznego (np. spektrogramy, cepstrogramy, chromagramy itp.) [6, 30, 31] lub postać parametryczną sygnału muzycznego, czyli wektor cech (np. wektor współczynników mel-cepstralnych lub parametry wykorzystujące standard MPEG-7) [13, 14, 16].

Podejście do treningu sieci neuronowej z danymi uzyskanymi z surowego, nieprzetworzonego dźwięku zaproponował Orjesek i współautorzy [16]. W algorytmie wykorzystano warstwy sieci splotowej połączone z warstwami rekurencyjnej sieci neuronowej. Autorzy pracy skupili się na klasyfikacji emocji na podstawie walencyjno-pobudzeniowego modelu emocji (*V*/*A*). Pierwsza z warstw sieci odpowiadała za wyodrębnienie cech własnych z wejściowego pliku audio przy użyciu 5-milisekundowego okna czasowego. W ten sposób uzyskano osiem map cech, które następnie zostały przetworzone i przygotowane dla rekurencyjnej części systemu. Aby zapobiec przeuczeniu sieci, zastosowano technikę dropout, która wyklucza niektóre neurony z wybranych warstw w kolejnych iteracjach uczenia się. Dzięki ostatniej warstwie sieci zostały zwrócone dwie wartości, czyli walencja i pobudzenie [16]. Do trenowania sieci wykorzystano bazę fragmentów muzycznych składającą się z 431 próbek (zestaw walidacyjny i treningowy), z których każda trwała 45 sekund. Zbiór testowy składał się z 58 utworów o średnim czasie trwania 243 sekundy. Dokładność predykcji sieci neuronowej mierzono wartością RMSE (pierwiastek z błędu średniokwadratowego; szczegóły w tabeli 1).

Inną metodą treningu sieci neuronowych jest użycie reprezentacji 2D; podejście to zostało wykorzystane przez M. Bilal Er [17]. Autorzy zaproponowali uczenie sieci neuronowych za pomocą chromagramów [17]. Opracowany algorytm ogranicza się do wyboru między czterema klasami emocji – radość, smutek, złość i relaks. W eksperymencie przetestowano dwie sieci neuronowe (VGG16, AlexNet) oraz dwa klasyfikatory (softmax, SVM) i wybrano kombinację, która dawała najlepsze wyniki w kontekście klasyfikacji emocji. Na potrzeby uczenia sieci neuronowej przygotowano dwa zbiory danych – pierwszy składający się ze 180 próbek audio o czasie trwania od 18 do 30 sekund, dostępnych w bazie Soundtracks oraz przygotowaną przez autorów bazę danych, która składała się z 400 próbek muzyki tureckiej. W celu przypisania etykiet przeprowadzono testy odsłuchowe, w których brało udział 13 osób. Kolejnym krokiem było wytrenowanie sieci neuronowej za pomocą zestawu 30-sekundowych chromogramów i wyodrębnienie czterech cech, które sklasyfikowano za pomocą dwóch przetestowanych klasyfikatorów. Zmierzono dokładność klasyfikacji emocji, która wyniosła około 89% [17].

W tabeli 1 przedstawiono przykłady badań dotyczących reprezentacji emocji/nastroju oraz ich automatyczną klasyfikację.

Ten krótki przegląd literatury z pewnością nie wyczerpuje badań przeprowadzonych w tym obszarze; ukierunkowany był jednak na wykazanie, że zwykle wszystkie elementy klasyfikacji emocji różnią się pomiędzy badaniami. Oznacza to, że modele emocji, reprezentacje dźwiękowe, metody, prezentacja wyników, a także cele aplikacyjne różnią się w badaniach związanych z klasyfikacją emocji.

**Tabela 1**

Wyniki klasyfikacji emocji z muzyki na podstawie innych badań

| Autor | Metoda | Model emocji | Dane wejściowe | Wyniki |
|---|---|---|---|---|
| Orjesek R. i współpracownicy [16] | jednowymiarowa splotowa sieć neuronowa | walenyjno-pobudzeniowy | surowe dane foniczne | RMSE walencja – 0,123±0,003 pobudzenie – 0,116±0,004 |
| Bilal Er i współpracownicy [17] | dwie sieci neuronowe (VGG16, AlexNet) oraz dwa klasyfikatory: softmax, SVM | cztery klasy emocji – złość, smutek, radość, relaks | chromagramy | dokładność 89,2% |

**Tabela 1. (cd.)**

| | | | | |
|---|---|---|---|---|
| Yang Y.-H. i współpracownicy [18] | FKNN (sieci rozmyte; *Fuzzy k-Nearest Neighbor*) | walenyjno-pobudzeniowy | wektor parametrów | dokładność 70,88% |
| | FNM (sieci rozmyte; *Fuzzy Neural Network*) | walenyjno-pobudzeniowy | wektor parametrów | dokładność 78,33% |
| Bargaje M. [32] | Algorytm genetyczny + SVM | walencja – pobudzenie – głośność | wektor parametrów | dokładność 84,57% |
| Sarkar R. i współpracownicy [33] | transfer learning | cztery klasy: szczęśliwy, zły, smutny, neutralny | spektrogram w skali melowej | dokładność 77,82±4,06% |
| Pandeya i współpracownicy [5] | splot 2D/3D | sześć odrębnych klas: podekscytowanie, strach, neutralność, relaks, smutek, napięcie | spektrogram w skali melowej | dokładność 74%, miara F1 – 0,73 |
| Seo, Huh [34] | SVM/las losowy/ uczenie maszynowe/ *k*NN | klasy: szczęśliwy–zadowolony, podekscytowany–podniecony, smutny–zły, spokojny–znudzony przemapowane na model walencyjno-pobudzeniowy | wektor parametrów | dokładność 73,96%/69,01%/ 72,90%/70,13% dla $k = 5$ |

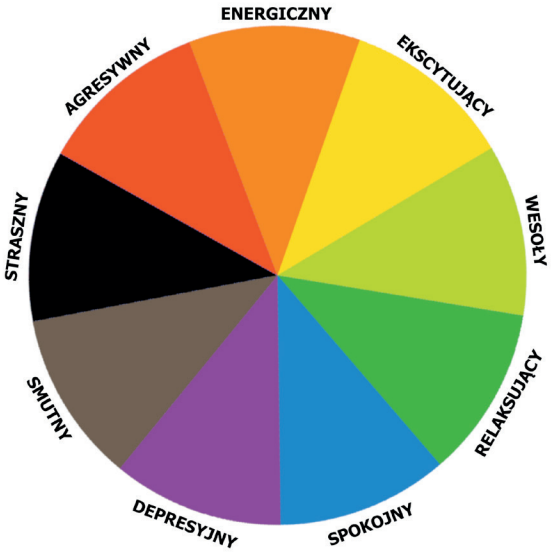
## 3. Testy subiektywne

Podstawowym celem ankiety było przemapowanie modelu emocji z wybranej muzycznej bazy danych (Epidemic Sound) do proponowanego modelu emocji. Przeprowadzono również testy subiektywne w celu wsparcia procesu etykietyzacji danych ze zbiorów używanych do uczenia i testowania sieci neuronowej. Główne założenia przeprowadzonej ankiety internetowej dotyczyły czasu jej trwania (maksymalnie 20 minut), czasu trwania próbki audio (przyjęto 15 sekund) oraz liczby odpowiedzi (ankieta powinna zawierać ponad 40 odpowiedzi). Proces konstrukcji ankiety został podzielony na trzy główne kroki, z których pierwszy polegał na wyborze modelu emocji towarzyszącemu klasyfikacji, a kolejny dotyczył wyboru bazy z utworami muzycznymi. Ostatnim krokiem było już faktyczne przygotowanie ankiety za pomocą narzędzia Google Forms.

### 3.1. Model emocji

Psychologia koloru w filmie jest wykorzystywana w celu wywołania w widzu konkretnych emocji i wrażeń, stąd istota włączenia koloru do modelu emocji. Wymienione modele, choć powszechnie stosowane w literaturze, nie nadawały się do wykorzystania w odniesieniu do muzyki filmowej ze względu na brak pełnego zestawu emocji, które mogą być wywołane przez muzykę w filmie.

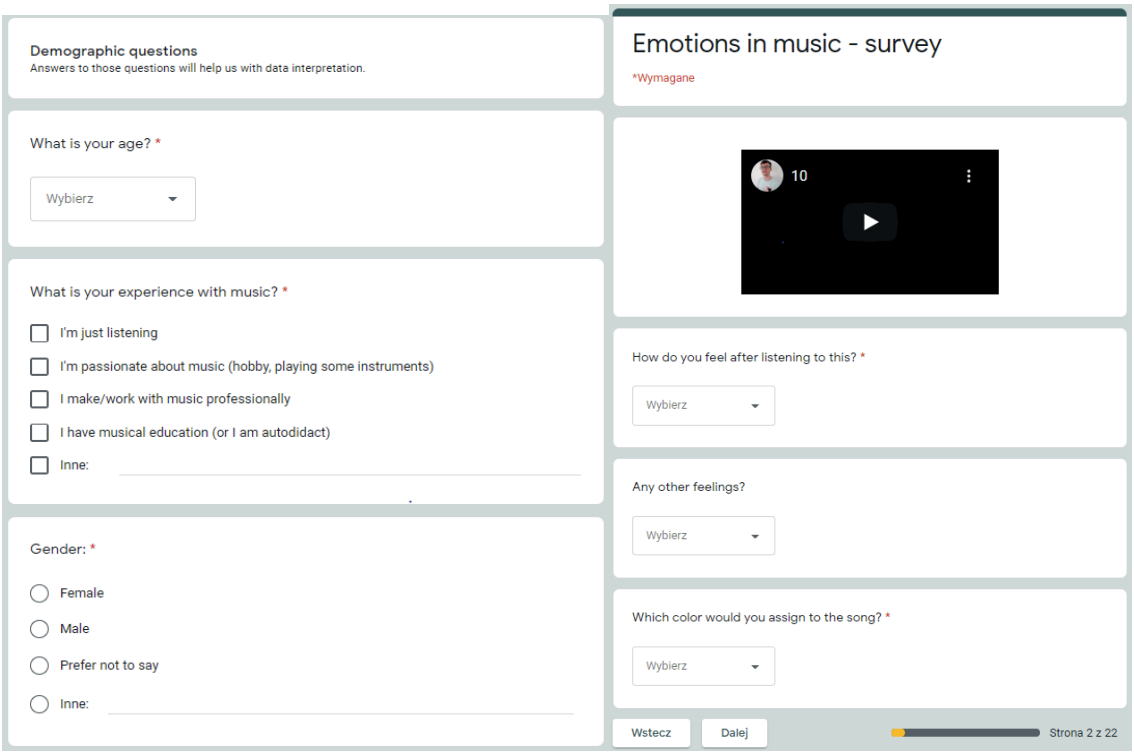
Przygotowując model emocji, uwzględniono dwa założenia – model nie powinien być zbyt skomplikowany, a jednocześnie dostosowany do tematyki problemu. Przegląd literatury pozwolił na wybór modelu, który byłby adekwatny do rodzaju problemu i pozwolił na skorelowanie emocji z kolorem [28]. Zaproponowany model (rys. 2) opiera się na modelu Plewy–Kostek (rys. 1). Wprowadzone modyfikacje obejmują rezygnację z gradacji natężenia emocji oraz z klasy „neutralny". Dodano klasę „straszny", przypisując jej kolor czarny.




**Rys. 2.** Zaproponowany model emocji

## 3.2. Przygotowanie ankiety

Głównym kryterium wyboru fragmentów muzyki filmowej była wysoka jakość (pliki .wav). Do eksperymentów wybrano utwory muzyczne z internetowej bazy Epidemic Sound [19]. Baza ta umożliwia wyszukiwanie utworów na podstawie emocji/nastroju. Z bazy danych wybrano utwory muzyczne z 19 różnych klas emocji/nastrojów; etykiety klas zostały następnie przemapowane do zaproponowanego przez autorów modelu emocji za pomocą stworzonej ankiety internetowej. Każdy ze znajdujących się w ankiecie utworów został wyłoniony w procesie selekcji polegającej na subiektywnym wyborze bazującym na odsłuchu kilkudziesięciu utworów z każdej z 19 klas – najlepszych reprezentantów danego nastroju.




**Rys. 3.** Strona powitalna ankiety oraz przykładowa sekcja z pytaniami

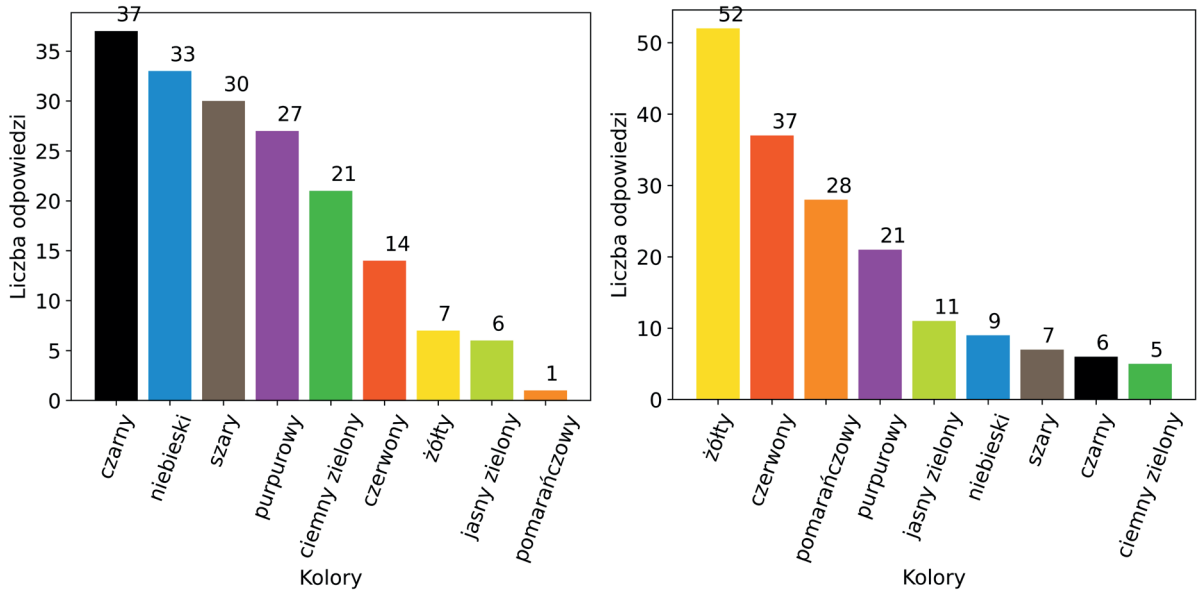
Ze względu na zbyt dużą liczbę utworów wybranych do przemapowania ankieta została skonstruowana – przy wykorzystaniu platformy Google Forms – w dwóch wariantach (zwanych dalej A i B); każdy wariant zawierał 21 sekcji, co miało pozwolić na jej wykonanie w założonym czasie 20 minut. W pierwszej części ankiety ankietowani odpowiedzieli na pytania dotyczące wieku, płci oraz swojego doświadczenia muzycznego. Następnie w drugiej jej części ankietowani odpowiadali na następujące trzy pytania dotyczące zaprezentowanego im fragmentu muzycznego (rys. 3):

– Jak się czujesz po odsłuchaniu tego utworu? Wybór jednej z dziewięciu emocji z zaproponowanego przez autorów modelu.
– Jakieś inne odczucia? (odpowiedź opcjonalna).
– Jaki kolor przypisałbyś/przypisałabyś do utworu? Wybór jednego z dziewięciu kolorów z zaproponowanego modelu emocji.

Ankietowani otrzymywali losowo wersję A lub B, które różniły się zawartymi fragmentami muzycznymi.

## 4. Testy subiektywne

W ankiecie wzięło udział 180 osób. Na rysunku 4 przedstawiono przykładowe wyniki zliczenia odpowiedzi dotyczące poszczególnych kolorów.




**Rys. 4.** Przykładowe wyniki zliczenia odpowiedzi przypisania emocji do kolorów

Zliczenia przypisania emocji z bazy Epidemic Sound do emocji z zaproponowanego modelu zostały znormalizowane do zakresu [0, 1] i przedstawione w formie tabeli (tab. 2). Rozwinięcia skrótów wykorzystywanych w tej tabeli przedstawiono w tabeli 3. W ten sposób uzyskano przemapowane etykiety, wykorzystywane dalej w procesie uczenia, walidacji oraz testowania algorytmu głębokiego.

**Tabela 2**

Zliczenia przypisania emocji z bazy Epidemic Sound do emocji z zaproponowanego modelu po normalizacji

| emocja | ang | cal | dep | ene | exc | hap | rel | sad | sca |
|---|---|---|---|---|---|---|---|---|---|
| ang | 0,233 | 0,0114 | 0,0057 | 0,6363 | 0,0739 | 0,017 | 0 | 0 | 0,0227 |
| dar | 0,0227 | 0,1705 | 0,0114 | 0,0568 | 0,2898 | 0,0057 | 0,0852 | 0,0511 | 0,3068 |
| dre | 0,017 | 0,3751 | 0,017 | 0,0739 | 0,1364 | 0 | 0,3011 | 0,017 | 0,0625 |

**Tabela 2. (cd.)**

| emocja | ang | cal | dep | ene | exc | hap | rel | sad | sca |
|---|---|---|---|---|---|---|---|---|---|
| epi | 0,0455 | 0,017 | 0,0114 | 0,5056 | 0,3693 | 0 | 0 | 0,0057 | 0,0455 |
| eup | 0,0455 | 0,0114 | 0,0057 | 0,6249 | 0,0739 | 0,2159 | 0,017 | 0,0057 | 0 |
| flo | 0,0057 | 0,4376 | 0,017 | 0,0398 | 0,142 | 0,0057 | 0,267 | 0,0852 | 0 |
| gla | 0,0227 | 0,1534 | 0 | 0,3808 | 0,017 | 0,1591 | 0,25 | 0,017 | 0 |
| hap | 0,0341 | 0,0909 | 0 | 0,3522 | 0,2045 | 0,2614 | 0,0398 | 0,0114 | 0,0057 |
| hop | 0,0194 | 0,2365 | 0,031 | 0,1124 | 0,0698 | 0,2364 | 0,2054 | 0,0891 | 0 |
| lai | 0,0227 | 0,3636 | 0,017 | 0,0795 | 0,0341 | 0,0625 | 0,4035 | 0,0114 | 0,0057 |
| mys | 0,0114 | 0,1193 | 0,0284 | 0 | 0,233 | 0,0227 | 0,0398 | 0,0966 | 0,4488 |
| rel | 0 | 0,5795 | 0,0227 | 0,0057 | 0,0114 | 0,0114 | 0,3068 | 0,0625 | 0 |
| rom | 0,0284 | 0,3581 | 0,0568 | 0,1534 | 0,0284 | 0,0795 | 0,2159 | 0,0795 | 0 |
| sad | 0,0057 | 0,267 | 0,1591 | 0,0114 | 0,0114 | 0,0114 | 0,1136 | 0,409 | 0,0114 |
| sen | 0,0185 | 0,237 | 0,1 | 0,0111 | 0,0222 | 0,0407 | 0,1667 | 0,3705 | 0,0333 |
| sex | 0,0227 | 0,3807 | 0,017 | 0,0455 | 0,0398 | 0,0284 | 0,4375 | 0,0284 | 0 |
| smo | 0,0398 | 0,3352 | 0,017 | 0,0114 | 0,0341 | 0,0682 | 0,4772 | 0,0114 | 0,0057 |
| sne | 0,017 | 0,0966 | 0,0341 | 0,0795 | 0,4546 | 0,017 | 0,0455 | 0,0398 | 0,2159 |
| sus | 0,0739 | 0,0284 | 0,0057 | 0,108 | 0,3693 | 0,0057 | 0 | 0,0284 | 0,3806 |

**Tabela 3**

Rozwinięcie skrótów wykorzystanych w tabeli 2

| | |
|---|---|
| ang | angry |
| cal | calm |
| dar | dark |
| dep | depressive |
| dre | dreamy |
| ene | energized |
| epi | epic |
| eup | euphoric |
| exc | excited |
| flo | floating |
| gla | glamorous |
| hap | happy |
| hop | hopeful |
| lai | laid back |
| mys | mysterious |
| rel | relaxed |
| rom | romantic |
| sad | sad |
| sca | scary |
| sen | sentimental |
| sex | sexy |
| smo | smooth |
| sne | sneaking |
| sus | suspense |

W celu wskazania istotnych statystycznie różnic w parach emocja–emocja z 19 wybranych emocji z bazy Epidemic Sound wykorzystano test chi-kwadrat. Wynik testu pomógł w późniejszym wyborze utworów użytych w procesach uczenia i testowania algorytmu uczenia maszynowego. Wyniki testu niezależności chi-kwadrat zostały przedstawione w trójkątnej macierzy (rys. 5). Znakiem „X" oznaczono pary emocji, pomiędzy którymi nie stwierdzono istotnych statystycznie różnic. Przyjęto, że poziom istotności α wynosi 0,001, ponieważ uznano, że standardowa wartość 0,05 nie jest wystarczająco dyskryminująca.

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| dar | - | | | | | | | | | | | | | | | | | |
| dre | - | - | | | | | | | | | | | | | | | | |
| epi | - | - | - | | | | | | | | | | | | | | | |
| eup | - | - | - | - | | | | | | | | | | | | | | |
| flo | - | - | X | - | - | | | | | | | | | | | | | |
| gla | - | - | - | - | - | - | | | | | | | | | | | | |
| hap | - | - | - | - | - | - | - | | | | | | | | | | | |
| hop | - | - | - | - | - | - | - | - | | | | | | | | | | |
| lai | - | - | X | - | - | - | - | - | - | | | | | | | | | |
| mys | - | X | - | - | - | - | - | - | - | - | | | | | | | | |
| rel | - | - | - | - | - | X | - | - | - | - | - | | | | | | | |
| rom | - | - | - | - | - | - | - | - | X | X | - | - | | | | | | |
| sad | - | - | - | - | - | - | - | - | - | - | - | - | - | | | | | |
| sen | - | - | - | - | - | - | - | - | - | - | - | - | - | X | | | | |
| sex | - | - | X | - | - | X | - | - | - | X | - | X | - | - | - | | | |
| smo | - | - | - | - | - | - | - | - | - | X | - | - | - | - | - | X | | |
| sne | - | X | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| sus | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| | ang | dar | dre | epi | eup | flo | gla | hap | hop | lai | mys | rel | rom | sad | sen | sex | smo | sne |

**Rys. 5.** Wyniki testu chi-kwadrat w przypadku 19 emocji z bazy Epidemic Sound

Do sprawdzenia poprawności powiązania koloru z emocjami w zaproponowanym modelu wykorzystano współczynnik korelacji *r* Pearsona. Test ten pozwolił na sprawdzenie siły i kierunku korelacji między badanymi zmiennymi. W celu przeprowadzenia porównania emocja–kolor zliczono wystąpienia emocji z zaproponowanego modelu w przypadku każdej z mapowanych emocji z bazy Epidemic Sound (X) [19], analogicznie postąpiono także odnośnie do kolorów (Y). Obie grupy liczebności (emocje i kolory) porównywano następnie parami na zasadzie każdy z każdym. Sprawdzono w ten sposób nie tylko słuszność przypisania kolorów do emocji w zaproponowanym modelu, ale również potencjalne poprawne połączenia emocja–kolor. Wyniki badań przedstawiono w tabeli 5. Wartości w tabeli odpowiadają współczynnikowi korelacji *r* Pearsona. Wartości wytłuszczone w tabeli 4 przedstawiają wartości współczynnika *r* dla kombinacji emocji i koloru zastosowanego w modelu.

Dla przypomnienia wartość *r* jest interpretowana w następujący sposób:

- [0,0; 0,3] – korelacja słaba,
- [0,3; 0,5] – korelacja umiarkowana,
- [0,5; 0,7] – korelacja silna,
- [0,7; 1,0] – korelacja bardzo silna,
- jeśli wartość *r* jest dodatnia, to wartości *x* i *y* są wprost proporcjonalne,
- jeśli wartość *r* jest ujemna, to wartości *x* i *y* są odwrotnie proporcjonalne.

**Tabela 4**

Współczynnik korelacji *r* Pearsona w przypadku emocji kolorów z zaproponowanego modelu

| Emocja | Kolor | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Czerwony | Niebieski | Purpurowy | Pomarańczowy | Żółty | Jasny zielony | Ciemny zielony | Szary | Czarny |
| Agresywny | **0,84** | −0,529 | −0,169 | −0,073 | −0,127 | −0,345 | −0,275 | −0,297 | 0,483 |
| Spokojny | −0,616 | **0,833** | −0,012 | −0,052 | 0,01 | 0,49 | 0,292 | 0,182 | −0,504 |
| Depresyjny | −0,246 | 0,366 | **−0,228** | −0,17 | −0,146 | 0,014 | 0,469 | 0,845 | 0,127 |
| Energiczny | 0,754 | −0,663 | −0,208 | **0,37** | 0,382 | −0,142 | −0,641 | −0,581 | −0,004 |
| Ekscytujący | 0,192 | −0,322 | 0,179 | −0,25 | **−0,273** | −0,307 | 0,084 | 0,059 | 0,396 |
| Wesoły | −0,161 | −0,252 | −0,222 | 0,711 | 0,945 | **0,654** | −0,435 | −0,291 | −0,388 |
| Relaksujący | −0,501 | 0,521 | 0,248 | 0,199 | 0,109 | 0,441 | **0,196** | −0,023 | −0,571 |
| Smutny | −0,235 | 0,391 | −0,255 | −0,186 | −0,105 | 0,057 | 0,502 | **0,864** | 0,182 |
| Straszny | −0,002 | −0,21 | 0,135 | −0,422 | −0,49 | −0,506 | 0,369 | 0,324 | **0,76** |

## 5. Automatyczna klasyfikacja emocji na podstawie muzyki filmowej

Implementacja algorytmu klasyfikacji emocji w muzyce filmowej została podzielona na kilka etapów. Pierwszym z nich było przygotowanie danych zebranych na podstawie wyników ankiety – pozwoliło to na wyselekcjonowanie utworów, na podstawie których wygenerowano zestaw reprezentacji 2D, czyli spektrogramów w skali melowej w przypadku procesu uczenia i testowania sieci. W kolejnym kroku przetestowano kilka sieci neuronowych, aby wybrać model głębokiego uczenia najbardziej odpowiedni do sklasyfikowania emocji. Ostatni krok opisuje proces budowy aplikacji umożliwiającej korzystanie ze skonstruowanego systemu klasyfikacji.

### 5.1. Wybór danych i środowiska programistycznego

Eksperymenty przeprowadzono w środowisku programistycznym Python, które współpracuje z bibliotekami Keras oraz Tensorflow. Ponieważ baza Epidemic Sound w większości przypadków przypisuje do każdego utworu po dwa deskryptory nastroju/emocji, zdecydowano się na jej przeszukiwanie względem par klas (w niektórych przypadkach rezultaty poszukiwań były niezadowalające, więc zdecydowano się na użycie pojedynczego deskryptora). Wykorzystane pary deskryptorów (oraz pojedyncze deskryptory) są następujące:

- angry,
- angry–dark,
- angry–epic,
- dark–mysterious,
- dark–suspense,
- dreamy–floating,
- epic–hopeful,
- epic–mysterious,
- euphoric–dreamy,
- happy–euphoric,
- happy–glamorous,
- happy–hopeful,
- happy–relaxing,
- hopeful–euphoric,
- hopeful–sentimental,
- hopeful–smooth,
- laid back–glamorous,
- laid back–sentimental,
- mysterious–floating,
- mysterious,
- mysterious–sneaking,
- mysterious–suspense,
- relaxing–laid back,
- relaxing,
- romantic–sad,
- romantic–sentimental,
- romantic–sexy,
- sad–dreamy,
- sad–relaxing,
- sad,
- sad–sentimental,
- sentimental–floating,
- sexy–laid back,
- smooth–laid back,
- sneaking–suspense.

Spośród dostępnych utworów muzycznych z bazy Epidemic Sound wybrano 420. W przygotowaniu danych wejściowych założono wykorzystanie reprezentacji 2D sygnału fonicznego, tj. spektrogramów

w skali melowej z logarytmiczną skalą amplitudy. Ta forma reprezentacji dźwięku łączy percepcyjną skalę częstotliwości i logarytmiczną skalę poziomu natężenia dźwięku, odzwierciedlając w ten sposób subiektywny sposób postrzegania dźwięku przez ludzi.

Następnie przygotowano skrypt generujący spektrogramy w skali melowej. Skrypt pozwala na ustawienie szerokości i wysokości reprezentacji 2D w pikselach, długości analizowanego okna sygnału w sekundach oraz kroku przesunięcia okna analizy w sekundach. Każdy utwór wczytywany jest z częstotliwością próbkowania *fs* = 22,05 kHz, a następnie na podstawie ustawianych parametrów generowane są spektrogramy w skali. Wybór takiej częstotliwości próbkowania służy głównie unifikacji przy przetwarzaniu próbek dźwięku oraz oszczędności mocy obliczeniowej przy generacji wielu tysięcy spektrogramów w skali melowej. Kolejnym krokiem było przypisanie etykiet do wygenerowanych spektrogramów w skali melowej. Etykiety zawierają informacje o stopniu przynależności każdej z dziewięciu klas emocji z zaproponowanego modelu do mapowanych emocji z modelu wykorzystywanego w bazie Epidemic Sound. Tabela 5 pokazuje przykładowe wygenerowane zbiory danych wraz z ich parametrami. Każdy z zestawów został przygotowany dla trzech następujących rozmiarów reprezentacji 2D: 224 × 224, 299 × 299 i 331 × 331.

**Tabela 5**

Parametry wygenerowanych zbiorów

| Długość okna analizy [s] | Długość kroku analizy [s] | Liczba obrazów w zbiorze |
|---|---|---|
| 30 | 2 | 29 495 |
| | 4 | 15 011 |
| | 5 | 12 045 |
| | 6 | 10 191 |
| | 8 | 7801 |
| | 10 | 6336 |
| 15 | 10 | 7007 |

## 5.2. Wybór sieci neuronowej

Przy wyborze modelu sieci neuronowej skupiono się przede wszystkim na sieciach splotowych, przyjmujących na wejściu reprezentacje 2D. Co więcej, wybrany model sieci powinien być modyfikowalny ze względu na zamierzone wykrywanie dziewięciu klas emocji – to założenie było szczególnie istotne w przypadku ostatniej warstwy gęstej architektury. Spośród sieci dostępnych na platformie Keras wybrano pięć modeli [35]:

- Xception – rozmiar obrazu wejściowego 299 × 299,
- VGG19 – rozmiar obrazu wejściowego 224 × 224,
- ResNet50V2 – rozmiar obrazu wejściowego 224 × 224,
- NASNetLarge – rozmiar obrazu wejściowego 331 × 331,
- InceptionV3 – rozmiar obrazu wejściowego 299 × 299.

Dokładność nauczonych modeli poddano ocenie przy wykorzystaniu zbioru testowego składającego się ze spektrogramów w skali melowej, wygenerowanych na podstawie utworów nieuwzględnianych w zbiorze uczącym i walidacyjnym. Miara dokładności proponowana przez moduł Keras – binarne porównywanie klas dla maksymalnych wartości na wyjściu i w etykiecie – okazała się niewystarczająca. Z tego względu – do celów czysto poglądowych – zaproponowano własny test. Opierał się on na:

- posortowaniu malejąco wartości zwróconych przez model,
- posortowaniu malejąco wartości etykiet w przypadku obiektu testowego,
- sprawdzeniu, czy trzy pierwsze emocje dla wartości z kroku 1 zawierają się wśród trzech pierwszych emocji dla wartości z kroku 2.

Przykładowo dla listy [a, b, c] z kroku 1 i listy [c, a, b] z kroku 2 dokładność wynosi 100%. Z kolei przy porównywaniu list [d, e, a] i [a, c, b] dokładność wynosiła już 33%. Wynik testu miał w założeniu pomóc przy wyborze najbardziej obiecujących modeli (na kolejnych etapach uczenia). Był on traktowany tylko i wyłącznie jako informacja, czy dany model sieci nauczony z uwzględnieniem wybranego zestawu hiperparametrów był wart uwagi. Wyniki testu miały wspomóc wybór najbardziej efektywnych modeli na kolejnych etapach uczenia sieci. W związku z tym niektóre modele i hiperparametry zostały odrzucone lub zmienione z powodu słabej wydajności na wszystkich etapach treningu.

Na pierwszym etapie uczenia i testowania sieci neuronowych niektóre modele były zbyt rozbudowane w kontekście możliwości sprzętowych, którymi dysponowali autorzy badania. W związku z tym próbowano zmniejszyć rozmiar wprowadzanych obrazów i przeprowadzić proces uczenia od nowa – udało się to tylko w przypadku modelu Xception po zmianie rozdzielczości z 299 × 299 na 224 × 224. W przypadku pozostałych dwóch sieci szukano alternatyw – VGG19 zamieniono na VGG16, a NASNetLarge na InceptionResNetV2. Proces uczenia modelu VGG16 wielokrotnie kończył się (podobnie jak w sytuacji poprzedniej, gdy zastosowano inną sieć) niepowodzeniem. Z tego względu zdecydowano się na odrzucenie tej architektury. We wszystkich przypadkach okazało się, że dla współczynnika uczenia $10^{-5}$ i $10^{-6}$ nie da się zminimalizować funkcji straty. Z tego powodu zrezygnowano z ich wykorzystania i ograniczono wektor współczynnika uczenia z [$10^{-3}$, $10^{-4}$, $10^{-5}$, $10^{-6}$] do [$10^{-3}$, $10^{-4}$]. W wyniku odrzucenia dwóch wartości współczynnika uczenia liczba przeprowadzonych procesów nauczania sieci spadła o połowę. Zaoszczędzony w ten sposób czas przeznaczono na przeprowadzanie uczenia z uwzględnieniem większej wartości parametru o nazwie „cierpliwość" (*patience*). Należy zauważyć, że przy ograniczonej liczbie epok strata, która jest wystarczająco niska, nie będzie już się zmniejszać (tj. dokładność nie wzrośnie). W tej sytuacji dalszy trening może być bezużyteczny. W tym przypadku stosuje się parametr, tzw. cierpliwość, który mówi, jak długo (ile epok) należy kontynuować trening po tym, jak strata przestała się zmniejszać. Sprawdzano, jak zmiana tej wartości wpływa na osiągniętą liczbę epok przez każdą z sieci neuronowych – 10 modeli osiągnęło maksymalną wartość 100 epok. Oznaczało to, że albo wybrana wartość cierpliwości jest zbyt duża, albo liczba epok jest zbyt mała. Przykładowe wyniki uczenia na II etapie przedstawiono w tabeli 6.

**Tabela 6**

Przykładowe wyniki uczenia II etapu

| Parametr | Nazwa modelu | | | |
|---|---|---|---|---|
| | Xception | ResNet50V2 | Inception-ResNetV2 | InceptionV3 |
| Rozmiar paczki danych (liczba obrazów na wejściu sieci) | 32 | 16 | 32 | 64 |
| Współczynnik uczenia | $10^{-4}$ | | | |
| Dokładność – zbiór treningowy [%] | 99,82 | 98,32 | 98,33 | 96,89 |
| Dokładność – zbiór walidacyjny [%] | 96,66 | 96,37 | 94,48 | 94,63 |
| Dokładność [%] | 59,49 | 59,90 | 60,38 | 62,95 |
| Własny test [%] | 77,26 | 76,70 | 75,77 | 77,30 |
| Liczba epok | 100 | 100 | 100 | 83 |

Na trzecim etapie procesu uczenia zmniejszono wartość cierpliwości do 13, a każdy model trenowano z wartościami hiperparametrów, które na etapach poprzednich dawały najbardziej obiecujące wyniki dokładności klasyfikacji. Ponadto postanowiono sprawdzić, jak na dokładność wytrenowanej sieci splotowej wpłynie wykorzystanie innego zbioru uczącego. W tym celu zmodyfikowano dotychczas wykorzystywany zbiór uczący – wygenerowano zestaw składający się z 30-sekundowych fragmentów utworów i kroku 10 sekund (do tej pory krok wynosił 5 sekund) – oraz wygenerowano dodatkowy zestaw uczący i testowy składający się z 15-sekundowych fragmentów muzycznych i kroku 10 sekund. Każdy z modeli był trenowany dwukrotnie. Przykłady porównań wyników przeprowadzonych procesów treningu przedstawiono

w tabeli 7. W 14 przypadkach (łącznie było ich 16) okazało się, że sieci neuronowe nauczone zestawem składającym się z 15-sekundowych fragmentów utworów charakteryzowały się mniejszą dokładnością.

**Tabela 7**

Przykładowe wyniki uczenia na etapie III

| Parametr | Nazwa modelu | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Xception | | ResNet50V2 | | Inception-ResNetV2 | | InceptionV3 | |
| Rozmiar paczki danych (liczba obrazów na wejściu sieci) | 32 | | 16 | | 64 | | 64 | |
| Współczynnik uczenia | $10^{-4}$ | | | | | | | |
| Długość fragmentu w zbiorze uczącym | 30 | 15 | 30 | 15 | 30 | 15 | 30 | 15 |
| Dokładność – zbiór treningowy [%] | 97,47 | 97,85 | 91,16 | 95,57 | 98,95 | 97,42 | 96,52 | 93,25 |
| Dokładność – zbiór walidacyjny [%] | 82,49 | 79,29 | 64,15 | 72,26 | 76,43 | 73,13 | 87,8 | 59,96 |
| Dokładność [%] | 58,18 | 55,61 | 53,60 | 57,84 | 63,47 | 61,98 | 61,04 | 59,96 |
| Test własny [%] | 76,58 | 74,52 | 73,45 | 73,14 | 76,24 | 74,52 | 75,63 | 74,55 |
| Liczba epok | 100 | 99 | 19 | 40 | 69 | 48 | 97 | 29 |

Na ostatnim etapie uczenia wygenerowano dodatkowo zbiór uczący składający się z 30-sekundowych fragmentów – dla kroku 8 sekund. Uczenie przeprowadzono ponownie dwukrotnie w przypadku każdego z modeli:

– pierwszy raz dla wartości cierpliwości równej 20, zbioru uczącego o długości fragmentów 30 sekund i kroku 10 sekund (maksymalnie 150 epok);
– drugi raz dla wartości cierpliwości równej 15, zbioru uczącego o długości fragmentów 30 sekund i kroku 8 sekund (maksymalnie 150 epok).

Osiągane wartości dokładności nie rosły znacząco z etapu na etap – wobec tego postanowiono zakończyć poszukiwanie hiperparametrów na fazie IV i wybrać w niej cztery modele o najlepszych wynikach. Dokładność sieci na czwartym etapie przedstawiono w tabeli 8.

**Tabela 8**

Zestawienie najlepszych wyników w przypadku każdego z czterech modeli na IV etapie uczenia

| Parametr | Nazwa modelu | | | |
|---|---|---|---|---|
| | Xception | ResNet50V2 | Inception-ResNetV2 | InceptionV3 |
| Rozmiar paczki danych (liczba obrazów na wejściu sieci) | 64 | 16 | 64 | 16 |
| Współczynnik uczenia | $10^{-4}$ | | | |
| Zbiór uczący | Długość fragmentu = 30 s Przesunięcie = 10 s | | Długość fragmentu = 30 s Przesunięcie = 8 s | |
| Dokładność – zbiór treningowy [%] | 98,43 | 99,63 | 98,64 | 97,25 |
| Dokładność – zbiór walidancyjny [%] | 90,54 | 87,63 | 89,17 | 93,16 |
| Dokładność [%] | 60,86 | 59,90 | 61,11 | 61,66 |
| Test własny [%] | 76,58 | 75,57 | 77,90 | 78,71 |
| Liczba epok | 150 | 150 | 141 | 68 |

Ze względu na duże trudności z wybraniem metryki określającej dokładność modelu postanowiono nie ograniczać się do wyboru jednej. Obliczono średni błąd bezwzględny, błąd średniokwadratowy oraz podobieństwo kosinusowe. Otrzymane wartości miar dokładności zaprezentowano w tabeli 9.

**Tabela 9**

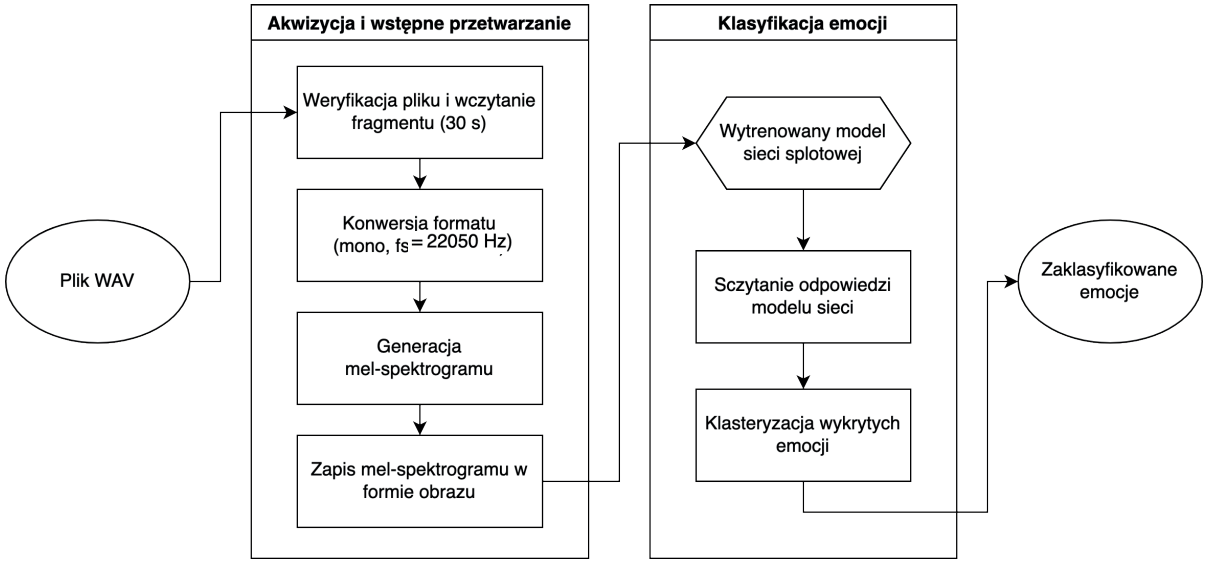
Porównanie miar dokładności w przypadku czterech najlepszych modeli na etapie VI uczenia

| Miara dokładności | Nazwa modelu | | | |
|---|---|---|---|---|
| | Xception | ResNet50V2 | InceptionResNetV2 | InceptionV3 |
| Podobieństwo kosinusowe | 0,8695 | 0,8720 | 0,8717 | 0,8892 |
| Średni błąd bezwzględny | 0,4735 | 0,4756 | 0,4573 | 0,4299 |
| Błąd średniokwadratowy | 0,0649 | 0,0632 | 0,0630 | 0,0542 |

Modelem sieci splotowej charakteryzującym się najmniejszymi błędami predykcji oraz największym podobieństwem przewidywanych wartości do oczekiwanych jest architektura InceptionV3. Dodatkową zaletą tego modelu jest jego rozmiar – 250 MB, czyli najmniej spośród czterech badanych.

## 5.3. Konstrukcja aplikacji klasyfikacji emocji w muzyce filmowej

Zaproponowany algorytm można podzielić na dwie główne części: etap akwizycji i wstępnego przetwarzania oraz etap klasyfikacji emocji. Uogólniony schemat blokowy algorytmu przedstawiony jest na rysunku 5.




**Rys. 5.** Schemat blokowy skonstruowanego systemu

Pierwszym krokiem jest weryfikacja wczytywanego pliku. System przyjmuje na wejściu wyłącznie pliki o rozszerzeniu .wav, jednakże przed dalszym przetwarzaniem sprawdzana jest jego kompatybilność z biblioteką Librosa oraz jego długość (utwór musi trwać przynajmniej 30 sekund). W przypadku spełnienia wymagań wczytany zostaje 30-sekundowy fragment, który jest następnie konwertowany do formatu mono, a częstotliwość próbkowania ustawiana jest na wartość $fs$ = 22,05 kHz. Następnie, przy wykorzystaniu biblioteki Librosa, na podstawie sygnału fonicznego generowany jest spektrogram w skali melowej z amplitudą w skali liniowej. Wówczas dokonywana jest operacja przeskalowania amplitudy do skali logarytmicznej (decybelowej), w przypadku której wartością odniesienia jest maksimum pierwotnego spektrogramu w skali melowej. Kolejnym krokiem jest zapis wygenerowanych danych w postaci reprezentacji 2D. W tym celu wartości spektrogramu w skali melowej przedstawiane są w postaci mapy termicznej, w której każdy piksel (w przestrzeni barw RGB, Red, Green, Blue) odpowiada jednej wartości. Na wejściu wytrenowanego modelu sieci neuronowej InceptionV3 podawany jest obraz o rozmiarach 299 × 299. Po jego przetworzeniu na wyjściu sieci – w warstwie softmax – pojawia się dziewięć wartości, które reprezentują stopień przynależności przetwarzanego utworu do każdej emocji z zaproponowanego modelu. Następnie wartości wyjściowe są klasteryzowane w celu wyodrębnienia zbioru maksymalnie trzech najbardziej pasujących do utworu emocji. W tabeli 10 przedstawiono wyniki klasyfikacji systemowej w przypadku dziesięciu różnych utworów muzyki filmowej.

**Tabela 10**

Wyniki systemu klasyfikacji emocji z muzyki filmowej

| Utwór | Początek fragmentu [s] | Wybór 1 | | Wybór 2 | | Wybór 3 | |
|---|---|---|---|---|---|---|---|
| | | Emocja | [%] | Emocja | [%] | Emocja | [%] |
| Main Theme | 0 | energiczny | 36,59 | ekscytujący | 17,26 | wesoły | 15,29 |
| African Rundown | 0 | energiczny | 35,32 | ekscytujący | 29,05 | straszny | 16,04 |
| Dumbledore's Farewell | 15 | spokojny | 29,64 | relaksujący | 23,2 | smutny | 18,39 |
| Time | 10 | ekscytujący | 27,0 | straszny | 20,68 | spokojny | 17,46 |
| Who is She | 42 | energiczny | 30,77 | wesoły | 18,53 | ekscytujący | 17,19 |
| Seizure of Power | 20 | energiczny | 48,18 | agresywny | 16,46 | ekscytujący | 15,29 |
| Flesh-16374 | 0 | straszny | 40,16 | ekscytujący | 28,09 | spokojny | 11,11 |
| Auschwitz-Birkenau | 0 | spokojny | 30,39 | smutny | 18,38 | relaksujący | 17,31 |
| Discombobulate | 20 | energiczny | 48,29 | ekscytujący | 22,32 | agresywny | 12,69 |
| Ragnar Says Goodby to Gyda | 0 | spokojny | 25,32 | smutny | 20,08 | relaksujący | 16,96 |

Predykcja, którą uzyskuje system, jest zadowalająca w zestawieniu z wynikami nieformalnego testu subiektywnego. Analizując wyniki działania aplikacji, można zauważyć, że dalsze prace powinny skoncentrować się na stworzeniu nowej ankiety w celu weryfikacji większej liczby fragmentów muzycznych dzięki ich subiektywnej ocenie pod kątem emocji wywoływanych u respondentów i porównaniu ich z wynikami uzyskanymi przez aplikację.

## 6. Podsumowanie

Problem klasyfikacji emocji w muzyce, zapoczątkowany w 1936 roku przez Kate Hevner, to stale rozwijająca się dziedzina nauki. Trudność w wykrywaniu emocji w muzyce wiąże się z subiektywną naturą zagadnienia. Rozwój badań dotyczących tych kwestii był niewątpliwie inspiracją do stworzenia odrębnej dziedziny nauki. Wiele badań na ten temat pokazuje mnogość możliwych podejść i rozwiązań przy użyciu różnych algorytmów klasyfikacji, różnych baz danych roboczych i różnych danych wejściowych do algorytmów.

W ramach przygotowania aplikacji przypisującej dany fragment muzyki filmowej do emocji z zaproponowanego modelu powstała baza utworów muzycznych, która została wykorzystana w procesie uczenia i testowania modelu sieci neuronowej. Etykietyzacja utworów muzycznych została przeprowadzona na podstawie testów subiektywnych oraz analizy statystycznej wyników ankiety. Wyniki analizy statystycznej odpowiedzi uzyskanych w procesie ankietowym wykazały akceptowalność zaproponowanego modelu emocji dotyczącego muzyki filmowej, składającego się z dziewięciu klas (energiczny, ekscytujący, wesoły, relaksujący, spokojny, depresyjny, smutny, straszny i agresywny).

Wybrany model splotowej sieci neuronowej InceptionV3 bardzo dobrze radzi sobie z problemem klasyfikacji emocji w muzyce filmowej, a uzyskane wyniki dokładności klasyfikacji można uznać za zadowalające.

Cały projekt ma charakter eksperymentalny – na każdym etapie prac można wprowadzić zmiany, które mogą wpłynąć na dokładność klasyfikacji algorytmu. Dalsze plany obejmują wprowadzenie poprawek do modelu emocji oraz wykorzystanie spersonalizowanej sieci neuronowej. Ponadto można zastosować inne metody doboru hiperparametrów wykorzystywanych w procesie uczenia. Planowane jest również wykorzystanie różnych architektur sieci neuronowych (wraz ze zróżnicowaną reprezentacją danych) oraz porównanie otrzymanych wyników z rezultatami uzyskanymi przy użyciu splotowej sieci neuronowej. Jak już wspomniano, należy dodatkowo sprawdzić system klasyfikacji emocji w muzyce

filmowej dzięki stworzeniu nowej ankiety w celu zweryfikowania większej liczby fragmentów muzycznych. Pozwoli to na sprawdzenie, czy istnieje rzeczywiście związek pomiędzy odpowiedziami osób ankietowanych a wynikami uzyskanymi przez system.

## Bibliografia


[1] M. Barhet, G. Fazekas, M. Sandler, *Music Emotion Recognition: From Content- to Context-Based Models*, w: M. Armaki, M. Barthet, R. Kronland-Martinet, S. Ystad (eds.), *From Sounds to Music and Emotions*, Lecture Notes in Computer Science, vol. 7900, Springer, Berlin–Heidelberg, 2013, s. 228–252, https://doi.org/10.1007/978-3-642-41248-6_13, dostęp 12.10.2021

[2] J. Grekow, *From Content-based Music Emotion Recognition to Emotion Maps of Musical Pieces*, Studies in Computational Intelligence, vol. 747, Springer, Cham, 2018

[3] P. Dwivedi, *Using CNNs and RNNs for Music Genre Recognition*, w: *Towards Data Science*, https://towardsdatascience.com/using-cnns-and-rnns-for-music-genre-recognition-2435fb2ed6af, dostęp 12.10.2021

[4] Z. Xiao, D. Wu, X. Zgang, Z. Tao, *Music mood tracking based in HCS*, w: *2012 IEEE 11th International Conference on Signal Processing*, Bejing, China, 2012, s. 1171–1175

[5] Y.R. Pandeya, B. Bhattarai, J. Lee, *Deep-Learning-Based Multimodal Emotion Classification for Music Videos*, Sensors 2021, vol. 21(14), 4927, https://doi.org/10.3390/s21144927

[6] M. Malik, S. Adavanne, K. Drossos, T. Virtanen, D. Ticha, R. Jarina, *Stacked convolutional and recurrent neural networks for music emotion recognition*, w: *14th Sound and Music Computing Conference*, Espoo, Finlandia, 2017, s. 208–213

[7] X. Yu, J. Zhang, J. Liu, W. Wan, W. Yang, *An audio retrieval method based on chromagram and distance metrics*, w: *2010 International Conference on Audio Language and Image Processing*, 2010, s. 425–428

[8] D. Grzywczak, G. Gwardys, *Audio features in music information retrieval*, w: Editors: Dominik Ślęzak, Gerald Schaefer, Son T. Vuong, Yoo-Sung Kim *Active Media Technology*, Lecture Notes in Computer Science, vol. 8610, Springer, Cham, 2014, s. 187–199

[9] G. Gwardys, D. Grzywczak, *Deep image features in music information retrieval*, International Journal of Electronics and Telecomunications 2014, vol. 60, s. 321–326

[10] J. Novet, *Google, Spotify & Pandora bet a computer could generate a better playlist than you*, w: *VenturaBeat*, 2014, https://venturebeat.com/business/deep-learning-music-streaming/, dostęp 12.10.2021

[11] C. Payne, *MuseNet. Open AI*, 2019, https://openai.com/blog/musenet/, dostęp 12.10.2021

[12] W. McCoulloch, W. Pitts, *A logical calculus of the ideas immanent in nervous activity*, Bulletin of Mathematical Biophysics 1943, vol. 5, s. 15–133

[13] J. Robinson, *Deeper than Reason: Emotion and its role in literature, music and art*, Oxford University Press, Oxford, 2005

[14] K. Sherer, M. Zentner, *Emotional effects of music: production rules*, P.N. Juslin, J.A. Sloboda (eds.), *Music and Emotion: Theory and Research*, Oxford University Press, Oxford, New York, 1989, s. 115–133

[15] Spotify, *Just the way you are: music listening and personality*, 2020, https://research.atspotify.com/just-the-way-you-are-music-listening-and-personality/, dostęp: 14.09.2021

[16] R. Orjesek, R. Jarina, M. Chmulik, M. Kuba, *DNN Based Music Emotion Recognition from Raw Audio Signal*, w: *29th International Conference Radioelektronika (RADIOELEKTRONIKA),* IEEE 2019, s. 1–4

[17] M. Bilal Er, I. Berkan Aydliek, *Music Emotion Recognition by Using Chroma Spectrogram and Deep Visual Features*, International Journal of Computional Intelligence Systems 2019, vol. 12(2), s. 1622–1634.

[18] Y.H. Yang, Y.C. Lin, Y.F. Su, H.H. Chen, *A regression approach to music emotion recognition*, w: *IEEE International Conference on Multimedia and Expo*, 2018, s. 448–457

[19] Epidemic Sound, *Royalty free music for your videos*, www.epidemicsound.com, dostęp 14.09.2021

[20] K. Hevner: *Experimental Studies of the Elements of Expression in Music*, The American Journal of Psychology 1936, vol. 48, s. 246–268

[21] J.A. Russel, *A circumplex model of affect*, Journal of Personality and Social Psychology 1980, vol. 39, s. 1161–1178

[22] D. Olson, C.S Russel, D.H. Sprenke (eds.), *Circumplex Model: Systemic Assessment and Treatment of Families*, Routledge, New York, 2014

[23] R.E. Thayer, *The Biopsychology of Mood and Arousal*, Oxford University Press, Oxford, 1989

[24] R.E. Thayer, R.J. McNally, *The biopsychology of mood and arousal*, Neuropsychiatry, Neuropsychology and Behavioral Neurology 1992, vol. 5, s. 65–74

[25] D. Watson, A. Tellegen, *Toward a consensual structure of mood*, Psychological Bulletin Journal 1985, vol. 98, s. 219–235

[26] A. Tellegen, D. Watson, L.A. Clark, *On the dimensional and hierarchical structure of affect*, Psychological Science 1999, vol. 10, s. 297–303
[27] M. Plewa, B. Kostek, *Music Mood Visualization Using Self-Organizing Maps*, Archives of Acoustic 2015, vol. 40, s. 513–525
[28] M. Plewa, *Automatic Mood Indexing of Music Excerpts based on Correlation Between Subjective Evaluation and Feature Vector*, rozprawa doktorska, Politechnika Gdańska, 2015
[29] B. Kostek, M. Plewa, *Rough Sets Applied to Mood of Music Recognition*, w: M. Ganhza, L. Maciaszek, M. Paprzycki (eds.), *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, 2016, s. 73–80, https://doi.org/10.15439/2016F548
[30] C. Lin, M. Liu, W. Hsiung, J. Jhang, *Music emotion recognition based on two-level support vector classification*, w: *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, IEEE 2016, s. 375–389
[31] S. Amiriparian, M. Gerczuk, E. Coutinho, A. Baird, S. Ottl, M. Milling, B. Schuller, *Emotion and Themes Recognition in Music Utilising Convolutional and Recurrent Neural Networks*, w: *MediaEval'19*, 27–29 October 2019, Sophia Antipolis, France, 2019.
[32] M. Bargaje, *Emotion recognition and emotion based classification of audio using genetic algorithm – an optimized approach*, w: *2015 International Conference on Industrial Instrumentation and Control (ICIC)*, IEEE 2015, s. 562–567
[33] R. Sarkar, S. Choudhury, S. Dutta, A. Roy, S.K. Saha, *Recognition of emotion in music based on deep convolutional neural network*, Mulitmedia Tools and Applications 2020, vol. 79, s. 765–783
[34] Y.S. Seo, J.H. Huh, *Automatic Emotion-Based Music Classification for Supporting Intelligent IoT Applications*, Electronics 2019, vol. 8, https://doi.org/10.3390/electronics8020164
[35] Keras, *Keras Applications*, https://keras.io/api/applications, dostęp 14.09.2021