Imię i nazwisko autora rozprawy: mgr inż. **Dawid Weber**
Dyscyplina naukowa: Informatyka techniczna i telekomunikacja

ROZPRAWA DOKTORSKA

Tytuł rozprawy w języku polskim: Opracowanie metodologii rozpoznawania i klasyfikowania emocji w filmach przy użyciu sztucznych sieci neuronowych

Tytuł rozprawy w języku angielskim: Development of a methodology for recognizing and classifying emotions in videos using artificial neural networks

Promotor *podpis*	Drugi promotor *podpis*
prof. dr hab. inż. Bożena Kostek	<Tytuł, stopień, imię i nazwisko>
Promotor pomocniczy *podpis*	Kopromotor *podpis*
dr inż. Karolina Marciniuk	<Tytuł, stopień, imię i nazwisko>

Gdańsk, 2024

OŚWIADCZENIE

Autor rozprawy doktorskiej: **mgr inż. Dawid Weber**

Ja, niżej podpisany(a), oświadczam, iż jestem świadomy(a), że zgodnie z przepisem art. 27 ust. 1 i 2 ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (t.j. Dz.U. z 2021 poz. 1062), uczelnia może korzystać z mojej rozprawy doktorskiej zatytułowanej:
„Opracowanie metodologii rozpoznawania i klasyfikowania emocji w filmach przy użyciu sztucznych sieci neuronowych" do prowadzenia badań naukowych lub w celach dydaktycznych.*

Świadomy(a) odpowiedzialności karnej z tytułu naruszenia przepisów ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych i konsekwencji dyscyplinarnych określonych w ustawie Prawo o szkolnictwie wyższym i nauce (Dz.U.2021.478 t.j.), a także odpowiedzialności cywilno-prawnej oświadczam, że przedkładana rozprawa doktorska została napisana przeze mnie samodzielnie.
Oświadczam, że treść rozprawy opracowana została na podstawie wyników badań prowadzonych pod kierunkiem i w ścisłej współpracy z promotorem prof. dr hab. inż. Bożeną Kostek oraz promotorem pomocniczym – dr inż. Karoliną Marciniuk.
Niniejsza rozprawa doktorska nie była wcześniej podstawą żadnej innej urzędowej procedury związanej z nadaniem stopnia doktora.
Wszystkie informacje umieszczone w ww. rozprawie uzyskane ze źródeł pisanych i elektronicznych, zostały udokumentowane w wykazie literatury odpowiednimi odnośnikami, zgodnie z przepisem art. 34 ustawy o prawie autorskim i prawach pokrewnych.
Potwierdzam zgodność niniejszej wersji pracy doktorskiej z załączoną wersją elektroniczną.

Gdańsk, dnia ....
podpis doktoranta

Ja, niżej podpisany(a), wyrażam zgodę/~~nie wyrażam zgody~~* na umieszczenie ww. rozprawy doktorskiej w wersji elektronicznej w otwartym, cyfrowym repozytorium instytucjonalnym Politechniki Gdańskiej.

Gdańsk, dnia ....
podpis doktoranta

**niepotrzebne usunąć*

* Art. 27. 1. Instytucje oświatowe oraz podmioty, o których mowa w art. 7 ust. 1 pkt 1, 2 i 4–8 ustawy z dnia 20 lipca 2018 r. – Prawo o szkolnictwie wyższym i nauce, mogą na potrzeby zilustrowania treści przekazywanych w celach dydaktycznych lub w celu prowadzenia działalności naukowej korzystać z rozpowszechnionych utworów w oryginale i w tłumaczeniu oraz zwielokrotniać w tym celu rozpowszechnione drobne utwory lub fragmenty większych utworów.
2. W przypadku publicznego udostępniania utworów w taki sposób, aby każdy mógł mieć do nich dostęp w miejscu i czasie przez siebie wybranym korzystanie, o którym mowa w ust. 1, jest dozwolone wyłącznie dla ograniczonego kręgu osób uczących się, nauczających lub prowadzących badania naukowe, zidentyfikowanych przez podmioty wymienione w ust. 1.

POLITECHNIKA GDAŃSKA

OPIS ROZPRAWY DOKTORSKIEJ

Autor rozprawy doktorskiej: Dawid Weber

Tytuł rozprawy doktorskiej w języku polskim: Opracowanie metodologii rozpoznawania i klasyfikowania emocji w filmach przy użyciu sztucznych sieci neuronowych"

Tytuł rozprawy w języku angielskim:
Development of a methodology for recognizing and classifying emotions in videos using artificial neural networks

Język rozprawy doktorskiej: polski

Promotor rozprawy doktorskiej: prof. dr hab. inż. Bożena Kostek

Promotor pomocniczy rozprawy doktorskiej*: dr inż. Karolina Marciniuk

Data obrony:

Słowa kluczowe rozprawy doktorskiej w języku polskim: emocje, film, kolorystyka filmu, muzyka filmowa, testy subiektywne, sieci neuronowe, sieci splotowe, sieci rekurencyjne

Słowa kluczowe rozprawy doktorskiej w języku angielskim: emotions, film, film colors, film music, subjective tests, neural networks, convolutional networks, recurrent networks

Streszczenie rozprawy w języku polskim: Celem rozprawy doktorskiej jest opracowanie metodologii pozwalającej na rozpoznawanie i klasyfikację emocji w filmie za pomocą sztucznych sieci neuronowych. W pracy przedstawiono tematykę związaną z kolorowaniem sceny filmowej w kontekście oddziaływania koloru na emocje widza. W celu analizy wpływu filmów na emocje widza dokonano wyboru tytułów filmowych, następnie przeprowadzono szereg wstępnych testów subiektywnych pozwalających na wybór i potwierdzenie sześciokolorowego modelu emocji oraz przypisanie do danego fragmentu filmowego odpowiedniej etykiety emocji. Wyniki testów subiektywnych pozwoliły na przygotowanie bazy danych fragmentów filmów, którą następnie wykorzystano do treningu i testów modeli uczenia głębokiego. W drugiej części pracy przygotowano analizę sygnałów audio i wideo poprzez różne sposoby parametryzacji tych sygnałów, a następnie dokonano klasyfikacji klas emocji na podstawie sygnału audio oraz wideo. Modele o najwyższej dokładności dla zbioru testowego zostały wybrane do stworzenia modelu multimodalnego. W trzeciej części pracy przygotowano model bimodalny wykorzystujący dwa wybrane wcześniej modele klasyfikacji sygnałów fonicznych oraz wideofonicznych. Model bimodalny wykazał się wyższą dokładnością podczas testów niż pojedynczy model klasyfikacji wideo, przy niewielkim koszcie wzrostu liczby parametrów modelu i stopnia skomplikowania.

Streszczenie rozprawy w języku angielskim: The goal of the dissertation is to develop a methodology that allows recognizing and classifying emotions in a video using artificial neural networks. This Ph.D. dissertation presents the subject of coloring a film scene in the context of the effect of color on the viewer's emotions. To analyze the impact of films on viewers' emotions, a selection of film titles was made, followed by a series of preliminary subjective tests to select and confirm the six-color model of emotions and assign the appropriate emotion label to the film excerpt. The results of the subjective tests made it possible to prepare a database of movie excerpts, which was then used to train and test deep learning models. In the second part of the work, the analysis of audio and video signals was prepared by different ways of parameterizing these signals, and then the classification of emotion classes based on audio and video signals was performed. The models with the highest accuracy for the test set were selected to create a multimodal model. In the third part of the work, a bimodal model was prepared using two previously selected models for classifying audio and video signals. The bimodal model showed higher accuracy during testing than a single video classification model at a small cost of increasing the number of model parameters and complexity.

POLITECHNIKA GDAŃSKA

PODZIĘKOWANIA

Chciałbym podziękować mojej promotorce Pani Profesor Bożenie Kostek za wskazanie kierunku naukowej drogi.

Chciałbym podziękować moim rodzicom za trud włożony w moje wychowanie i wykształcenie, wiem, że była to niezwykle ciężka praca.

Dziękuję również żonie za ogrom cierpliwości i wsparcia, jakim się wykazała w stosunku do mnie, niewątpliwie bez niej ta rozprawa by nie powstała.

STRESZCZENIE

Celem rozprawy doktorskiej jest opracowanie metodologii pozwalającej na rozpoznawanie i klasyfikację emocji w filmie za pomocą sztucznych sieci neuronowych. W pracy przedstawiono tematykę związaną z kolorowaniem sceny filmowej w kontekście oddziaływania koloru na emocje widza. W celu analizy wpływu filmów na emocje widza dokonano wyboru tytułów filmowych, następnie przeprowadzono szereg wstępnych testów subiektywnych pozwalających na wybór i potwierdzenie sześciokolorowego modelu emocji oraz przypisanie do danego fragmentu filmowego odpowiedniej etykiety emocji. Wyniki testów subiektywnych pozwoliły na przygotowanie bazy danych fragmentów filmów, którą następnie wykorzystano do treningu i testów modeli uczenia głębokiego. W drugiej części pracy przygotowano analizę sygnałów audio i wideo poprzez różne sposoby parametryzacji tych sygnałów, a następnie dokonano klasyfikacji klas emocji na podstawie sygnału audio oraz wideo. Modele o najwyższej dokładności dla zbioru testowego zostały wybrane do stworzenia modelu multimodalnego. W trzeciej części pracy przygotowano model bimodalny wykorzystujący dwa wybrane wcześniej modele klasyfikacji sygnałów fonicznych oraz wideofonicznych. Model bimodalny wykazał się wyższą dokładnością podczas testów niż pojedynczy model klasyfikacji wideo, przy niewielkim koszcie wzrostu liczby parametrów modelu i stopnia skomplikowania.

Słowa kluczowe: emocje, film, kolorystyka filmu, muzyka filmowa, testy subiektywne, sieci neuronowe, sieci splotowe, sieci rekurencyjne

Dziedzina nauki i techniki, zgodnie z wymogami OECD: informatyka techniczna i telekomunikacja

ABSTRACT

The goal of the dissertation is to develop a methodology that allows recognizing and classifying emotions in a video using artificial neural networks. This Ph.D. dissertation presents the subject of coloring a film scene in the context of the effect of color on the viewer's emotions. To analyze the impact of films on viewers' emotions, a selection of film titles was made, followed by a series of preliminary subjective tests to select and confirm the six-color model of emotions and assign the appropriate emotion label to the film excerpt. The results of the subjective tests made it possible to prepare a database of movie excerpts, which was then used to train and test deep learning models. In the second part of the work, the analysis of audio and video signals was prepared by different ways of parameterizing these signals, and then the classification of emotion classes based on audio and video signals was performed. The models with the highest accuracy for the test set were selected to create a multimodal model. In the third part of the work, a bimodal model was prepared using two previously selected models for classifying audio and video signals. The bimodal model showed higher accuracy during testing than a single video classification model at a small cost of increasing the number of model parameters and complexity.

Keywords: emotions, film, film colors, film music, subjective tests, neural networks, convolutional networks, recurrent networks

WYKAZ WAŻNIEJSZYCH SKRÓTÓW

CIE – Comission Internationale de l'Eclairage (Międzynarodowa Komisja Oświetleniowa)
CNN – Convolutional Neural Network (neuronowa sieć splotowa)
EHD – Edge Histogram Descriptor
FFT – Fast Fourier Transform
GRU – Gated Recurrent Unit (bramkowana jednostka rekurencyjna)
HTD – Homogeneous Texture Descriptor
LPC – Linear Predictive Coding (liniowe kodowanie predykcyjne)
LSTM – Long short-term memory (rodzaj sieci rekurencyjnej; sieć z długoterminową pamięcią)
LUT – Look Up Table
MAE – Mean absolute error (średni błąd bezwzględny)
MER – Music Information Retrieval
MFCC – Mel-frequency cepstral coefficients (współczynniki mel-cepstralne)
MLP – Multilayer Perceptron (wielowarstwowy perceptron)
RAW – Cyfrowy negatyw (format RAW, 'surowy' – bez żadnej obróbki)
RGB – Red Green Blue
R^2 – R-Squared or the coefficient of determination (współczynnik determinacji, dopasowania modelu od danych uczących)
RNN – Recurrent Neural Network (neuronowa sieć rekurencyjna)
STFT – Short Time Fourier Transform
TBD – Texture Browsing Descriptor
ZCR – Zero-crossing rate

SPIS TREŚCI

1. Wstęp i cel pracy

Jednym z głównych aspektów muzyki i filmu jest ich oddziaływanie emocjonalne na słuchacza. Celem kompozytora jest zawarcie emocji w tworzonej muzyce, które następnie przez wykonanie dzieła trafiają do słuchacza. Zagadnienia te były studiowane na przestrzeni wieków [41, 85], zaś obecnie odnoszą się do badań znanych pod nazwą przetwarzania afektywnego (ang. *affective computing*) [50, 62, 83, 86]. W kontekście filmu, nadrzędnym czynnikiem jest niewątpliwie obraz i kompozycja koloru, które w równym lub nawet większym stopniu, co muzyka oddziałują na emocje widza − odbiorcy tego przekazu.

Istnieje wiele definicji emocji, przykładowo można się posłużyć opisem: „*intensywny stan psychiczny pobudzający układ nerwowy i wywołujący reakcje fizjologiczne*" [18]. Powołując się jednak na słowa Russella: „*Wszyscy wiedzą czym są emocje, dopóki nie zostaną poproszeni o ich definicję*" [24], należy zauważyć, że jednoznaczne określenie emocji jest problematyczne, gdyż obejmuje aktywność psychofizyczną danej osoby pobudzonej bodźcem zewnętrznym (bądź wewnętrznym) [76].

Zrozumienie, w jaki sposób muzyka przekazuje emocje nie jest łatwe. Kivy podaje dwie hipotezy: *„pierwsza może dotyczyć słuchowego podobieństwa między muzyką a naturalnym wyrażeniem emocji*" [48]. Niektóre muzyczne formy mogą powodować wywoływanie emocji, ponieważ zbliżone są one do mowy i jej ekspresji. Przykładem może być „złość" – gdzie głośność oraz dysonans widmowy [56] – to dwie charakterystyczne cechy dla tej emocji, zarówno w mowie, jak i muzyce. Druga hipoteza, jaką przywołuje Kivy, odnosi się do „*kulturowych zjawisk. które mają wspólne konotacje wraz z muzyką*". Grewe opisuje, że intensywność emocji może bazować na personalnym odczuciu dla danego utworu bądź też uzależniona jest od tła muzycznego [26]. Podaje przykład muzyka, który uczy się danego utworu muzycznego, a potem musi odegrać go na żywo − w ten sposób muzyk odczuwa emocje w sposób wzmożony. Z drugiej strony słuchanie danego utworu kilkukrotnie może spowodować zmiany w nastawieniu emocjonalnym dla tego dzieła. Warto jednak zaznaczyć, że odczuwanie emocji z muzyki nie jest uwarunkowane jedynie wykształceniem muzycznym czy wiekiem danej osoby − kształtowanie rozumienia muzyki dotyczy wczesnych lat (a nawet miesięcy) dziecka. Dowiedziono, iż nawet czteromiesięczne dzieci pokazują, który utwór jest przyjemniejszy w odbiorze, a który wprowadza je w przykry nastrój [117]. Są też osoby, które nie rozwinęły odczuwania emocji i nie są zdolne do zrozumienia muzyki [97].

Stworzenie filmu wymaga odpowiedniego scenariusza, doboru aktorów, wyznaczeniu lokalizacji kręcenia ujęć, a także późniejszego etapu postprodukcji, w skład, którego wchodzi między innymi wybór odpowiednich ujęć, tempo cięć, przejścia pomiędzy ujęciami, edycja kolorów, ale również edycja ścieżki dźwiękowej: dialogów czy też muzyki. Zabiegi postprodukcji mają za zadanie odpowiednio przyciągnąć uwagę widza, dodatkowo wywołując w nich konkretne reakcje na przekaz emocji. Użycie określonych kolorów w filmie definiuje klimat, nastrój czy też emocje, które mogą być wzmacniane u widzów. Reżyserzy często wykorzystują podstawowe kolory ciepłe bądź zimne do kreowania odpowiednich wrażeń. Kolor czerwony może być wykorzystany do podkreślenia pasji, miłości, romantyzmu, ale również niebezpieczeństwa czy agresji. Natomiast kolory chłodne, takie jak niebieski czy też zielony kojarzone są ze spokojem, melancholią bądź strachem. Psychologia kolorów jest szerokim obszarem badań w psychologii oraz w sztuce.

Prace nad kolorem w filmie rozpoczynają już na pierwszych etapach filmu, w trakcie przygotowań do tworzenia planu filmowego, kreując odpowiednie otoczenie, dobierając temperaturę barwową światła. Reżyserzy tacy, jak Wes Anderson, wykorzystują kolory nie tylko w postprodukcji filmu, ale również na początkowym etapie nagrywania ujęć, w kostiumach bohaterów czy w scenografii. Pozwala to zawrzeć dodatkową warstwę znaczeniową w filmie.

W dobie błyskawicznego rozwoju zasobów Internetu uzyskano możliwość wyboru interesujących filmów również serwisach *streamingowych* takich jak Netflix, Amazon Primie, itd. Do dyspozycji użytkowników udostępniane są ogromne bazy filmów. Kluczowa pozostaje jedynie

kwestia wyboru interesującego widza dzieła. Dlatego też, jeżeli widz nie ma wybranego tytułu filmowego, to może skorzystać z wyszukiwania treści i określenia preferencji pod kątem gatunku filmowego, ulubionych aktorów czy też wrażeń emocjonalnych, jakie za sobą niosą dane pozycje. Na tej podstawie również algorytmy rekomendacji dobierają widzom kolejne tytuły warte obejrzenia.

Analizując aktualny stan wiedzy, można zauważyć głównie trendy w odniesieniu do emocji zawartych w muzyce, tj. automatyczne przypisanie emocji do muzyki czy gatunku muzycznego, ale klasyfikacja emocji na podstawie filmu, bimodalnego źródła (AV), składającego się zarówno ze ścieżki dźwiękowej (A), jak i ruchomego obrazu (V) w dalszym ciągu pozostawia pole do dalszych badań. Większość badań związanych z analizą audio bądź wideo opiera się na predykcji stanów emocjonalnych słuchaczy czy widzów w kontekście wyznaczania wartości na płaszczyźnie wartościowości (walencji) oraz pobudzenia (ang. *valence-arousal*; VA). Brakuje jednak prac, które podchodzą do wyznaczania emocji na podstawie filmu jako spójnej całości w sposobie klasyfikacji konkretnych etykiet emocji. Może być to związane głównie z brakiem dostępności zbiorów danych, które zawierałyby źródła filmowe wraz z towarzyszącą ścieżką dźwięków odniesienia do wywołanych emocji.

Większość metod klasyfikacji klas emocji bądź predykcji wartości AV oparta jest na sieciach neuronowych. Architektury sieci przybierają różne formy w zależności od zastosowania oraz opracowanych metod przez twórców modeli. W najprostszej postaci algorytmy składające się z sieci neuronowych dzielą się na dwie części, w pierwszej części mechanizm przetwarzania danych dokonuje ekstrakcji cech obrazu bądź dźwięku, które następnie tworzą formę tensorów, druga część odpowiedzialna jest za klasyfikację danych [94].

W zależności od problemu klasyfikacji wybierane są różne architektury sieci neuronowych. W dziedzinie audio i filmu (jako danych czasowo-przestrzennych) wykorzystuje się najczęściej sieci splotowe (ang. *convolutional neural networks*, CNNs) [16, 30, 119] bądź rekurencyjne (ang. Recurrent neural networks, RNNs) [7, 25, 127]. Jednak intensywny rozwój sztucznych sieci neuronowych przynosi kolejne typy sieci neuronowych, które znajdują zastosowanie w klasyfikacji sygnałów fonicznych i wizyjnych [20, 112, 127].

W celu klasyfikacji sygnałów audio, częstym sposobem parametryzacji jest ekstrakcja współczynników mel-cepstralnych (ang. *Mel frequency cepstral coefficients*, MFCC) [68] z sygnału fonicznego bądź stworzenie na podstawie widma mocy sygnału spektrogramów w skali melowej. Parametry MFCC mogą być wykorzystywane w sieciach rekurencyjnych jako dane wejściowe, gdzie dla każdej ramki sygnału fonicznego obliczana jest wartość parametrów MFCC. Spektrogramy w skali melowej wykorzystywane są natomiast do zadań klasyfikacji dźwięku przy użyciu sieci splotowych, które doskonale sobie radzą z klasyfikacją na podstawie obrazu, czyli reprezentacji sygnału fonicznego w postaci dwuwymiarowej (2D).

Film jako reprezentacja w postaci danych przestrzenno-czasowych również może być klasyfikowany za pomocą sieci rekurencyjnych bądź splotowych. Popularnym podejściem do klasyfikacji filmu jest parametryzacja poszczególnych klatek filmowych, przy czym zazwyczaj ogranicza się liczbę klatek filmowych do kilkudziesięciu pierwszych bądź kilkudziesięciu klatek z całości fragmentu wybranych w równych odstępach. Ekstrakcji parametrów obrazu dokonuje się poprzez użycie sieci splotowych jako ekstraktora cech i następnie poddaje się je klasyfikacji za pomocą sieci rekurencyjnych [84]. Możliwe jest też wykorzystanie bardziej skomplikowanej operacji, jaką jest klasyfikacja obrazu przy użyciu trójwymiarowych sieci splotowych, które otrzymują tensor parametrów o długości liczby klatek filmu. W każdej warstwie tensora, odpowiadającej za poszczególną klatkę obrazu, znajduje się trójwymiarowy wektor parametrów zawierający wartości RGB dla poszczególnych pikseli [112].

Rozwój metod klasyfikacji emocji zawartych w muzyce

Klasyfikacja emocji na podstawie sygnału audio jest obecnie bardzo popularnym tematem prac i rozważań naukowych [16, 20, 25,, 94]. Często stosowanym rozwiązaniem jest użycie spektrogramów w skali melowej (zwykle zwanych mel-spektrogramami lub spektrogramami

melowymi) oraz sieci splotowych do tego typu zadań. Można również spotkać się z innymi podejściami takimi jak: połączenie sieci splotowych z rekurencyjnymi, wykorzystanie parametryzacji sygnału audio z użyciem parametrów MFCC oraz sieci rekurencyjnych do klasyfikacji. Wykorzystuje się również tradycyjne metody uczenia maszynowego, jakimi są algorytm *k*-najbliższych sąsiadów (ang. *k-Nearest Neighbors*, *k*-NN), maszyna wektorów nośnych (ang. *Support Vector Machine*, SVM), drzewa decyzyjne czy inne algorytmy [7, 20, 94] w połączeniu z wektorem parametrów opartym na standardzie MPEG 7 [25, 127].

Rozwój metod klasyfikacji emocji w filmie

Obecnie wykorzystywanym rozwiązaniem w klasyfikacji wideo czy filmu jest użycie sieci splotowych zarówno dwuwymiarowych, jak i trójwymiarowych, ale też są widoczne prace wykorzystujące podejście tradycyjne.

W tabeli 1.1 zestawiono przykłady wybranych badań z ostatnich lat, realizujące zadania związane z klasyfikacją/predykcją emocji zawartych w sygnałach fonicznych, wideo oraz wideo-fonicznych (filmach). Zestawienie to pozwala zauważyć, że wykorzystywane są w tej dziedzinie różne metody mechanizmów uczenia maszynowego, tj. klasyczne, jak np. SVM czy sztuczne sieci neuronowe (ang. *artificial neural network*, ANN), jak również modele głębokie, np. CNN, bramkowana jednostka rekurencyjna (ang. *gated reccurent unit*, GRU), regresja logistyczna, sieć pamięci długoterminowej (ang. *long short-term memory*, LSTM), dwukierunkowa sieć pamięci długoterminowej – (ang. *bidirectional long short-term memory*, BiLSTM). Dotyczy to również stosowania różnych baz danych sygnałów oraz miar oceny jakości w ocenie uzyskanych wyników. Wśród tych ostatnich można zauważyć, że obok typowego podejścia wykorzystującego miarę dokładności, stosuje się miarę AUC (ang. Area Under the Curve – pole powierzchni pod krzywą), R^2 (ang. Coefficient of Determination – współczynnik determinacji), mAP (ang. Mean Average Precision – uśredniona precyzja średnia), MSE (ang. Mean Squared Error – błąd średniokwadratowy), itd.

Tab. 1.1. Przegląd metod realizujących zadania związane z klasyfikacją sygnałów audio, wideo oraz audio-wideo w kontekście emocji zawartych w sygnałach

Autorzy	Rok	Zadanie	Typ danych wejściowych	Algorytm	Miary i wyniki
Behrouzi T., Toosi R., Akhaee M. A. [7]	2023	Klasyfikacja gatunków filmowych	Klatki filmowe, sygnał foniczny	CNN, SVM, GRU	F1-score: 0,66
Revathy V. R., Pillai A. S. [94]	2022	Klasyfikacja emocji	Wybrane parametry sygnału audio	SVM, Regresja logistyczna, ANN	Dokładność dla sieci neuronowej: 82%
Vrskova R., Hudec R., Kamencay P., Sykora P. [119]	2022	Klasyfikacja czynności ludzkich	Klatki filmowe	CNN	Dokładność dla zbioru UCF YouTube Action – 85,2% Dokładność dla zbioru UCF YouTube Action – 84,4%
Ciborowski T., Reginis S., Kurowski A., Weber D., Kostek B. [16]	2021	Klasyfikacja emocji	Mel-spektrogram	CNN	Dokładność dla klasyfikacji jednej klasy – 61,66% Dokładność dla klasyfikacji trzech klas – 78,71%
Hayat H., Ventura C., Lapedriza A. [30]	2021	Predykcja wartości Valence-Arousal	Klatki filmowe	CNN	Dokładność – 73,6%
Grekow J. [25]	2021	Predykcja wartości Valence-Arousal	Wybrane parametry sygnału audio	LSTM	R^2 Arousal – 0,73 R^2 Valence – 0,46 MAE Arousal – 0,12 MAE Valence – 0,12
Yu Y., Lu Z., Li Y., Liu D. [127]	2021	Klasyfikacja gatunków filmowych	Histogram kolorów RGB	CNN, LSTM	mAP – 0,609 HR@1 – 0,726 HR@3 – 0,834
Du P., Li X., Gao Y. [20]	2020	Predykcja wartości Valence-Arousal	Mel-spektrogram, cochleogram	CNN, BiLSTM	RMSprop Arousal – 0,07±0,05 RMSprop Valence – 0,06±0,04
Aslan F., Ekenel H. K. [3]	2019	Predykcja wartości Valence-Arousal	Klatki filmowe	CNN	MSE Arousal – 0,06 MSE Valence – 0,085
Sarkar R., Choudhury S., Dutta S., Roy A. [98]	2019	Klasyfikacja emocji	Mel-spektrogram	CNN	Dokładność – 67,7±3,6
Khanh-An Q., Vinh-Tiep N., Minh-Triet T. [47]	2018	Predykcja wartości Valence-Arousal	Klatki filmowe	CNN	MSE Arousal – 0,17 MSE Valence – 0,12

1.1 Cel pracy

Przegląd literatury obejmujący tematykę klasyfikacji sygnału wideo oraz audio wskazuje na znaczny niedosyt w kontekście klasyfikacji emocji na podstawie filmu, który jest złożonym medium audiowizualnym. Pierwszym z problemów jest brak ujednoliconych dostępnych baz

danych, co powoduje problem w porównaniu jakości różnych rozwiązań uczenia maszynowego. Zauważyć również należy stosowanie różnych miar oceny algorytmów oraz różnego podejścia, tj. klasyfikacji bądź predykcji emocji.

Celem niniejszej rozprawy doktorskiej jest opracowanie metodologii, która pozwoli na klasyfikację i rozpoznawanie emocji we fragmencie filmu z uwzględnieniem muzyki oraz koloru. Nowatorskim podejściem w stosunku do aktualnej literatury tematu jest podzielenie algorytmu klasyfikującego na dwa moduły, jednego odpowiedzialnego za klasyfikację sygnału wideo, drugiego - za sygnał foniczny. Podejście bimodalne powinno zapewnić większą skuteczność klasyfikacji emocji zawartych w treści multimedialnej. Jako narzędzie do przeprowadzenia tych badań wybrano sztuczne sieci neuronowe, które dominują w zadaniach klasyfikacji oraz predykcji sygnałów audio oraz wideo.

Ważnym aspektem pracy jest opracowanie własnego zbioru danych na podstawie wyników przeprowadzonych ankiet subiektywnych. Zbiór danych zawiera fragmenty filmu wraz z towarzyszącą muzyką, która została użyta oryginalnie w danym fragmencie filmu przez twórców.

1.2 Tezy rozprawy doktorskiej

W pracy zaproponowano następujące tezy pracy:

1. **Możliwe jest przygotowanie i wytrenowanie modeli sieci neuronowych, które osiągają dokładność klasyfikacji emocji zawartych w muzyce filmowej wyższą, tj. >90% oraz we fragmencie wideo pochodzącego z filmu, tj. >85%, w stosunku do wyników uzyskanych w literaturze (odpowiednio: 82% – Revathy, Pillai [94] oraz 73,6% – Hayat i in. [30]), co przewyższa aktualny stan wiedzy.**

2. **Wykorzystanie bimodalnego podejścia w uczeniu maszynowym, tj. jednoczesnej analizy sygnałów audio i wideo, pozwala zwiększyć dokładność klasyfikacji emocji zawartych w filmie do wartości 89% w stosunku do analizy jedynie w oparciu o sygnał wideo (najlepsza uzyskana dokładność dla modelu wideo na zbiorze testowym wyniosła 87%).**

1.3 Zawartość rozprawy

Praca została podzielona na 8 rozdziałów, których struktura została przedstawiona na rys. 1.1. W pierwszej kolejności podano genezę pracy z przeglądem rozwiązań w kontekście analizy sygnałów audio, sygnału wideo oraz sygnałów audiowizualnych, przedstawiono również cel oraz tezy rozprawy. W przeglądzie literatury odniesiono się w głównej mierze do tematyki zbliżonej do rozprawy, czyli klasyfikacji emocji, predykcji wartości walencji-pobudzenia, analizy sygnałów audio i wideo wraz z opisem miar i wyników uzyskanych przez przywołanych autorów.

Drugi rozdział obejmuje podstawy teorii kolorów w filmie oraz teorii muzyki. Stanowi on podstawę do analizy oraz parametryzacji sygnałów audio i wideo. W rozdziale tym przedstawiono również pokrótce metody analizy wideo, tj. wykorzystanie deskryptorów koloru, tekstury czy też wykorzystanie sieci splotowych do parametryzacji sygnałów wideo. Dalsza część tego rozdziału poświęcona jest analizie oraz parametryzacji sygnałów audio, tj. przywołane zostały deskryptory MPEG-7, parametryzacja za pomocą współczynników MFCC, jak również reprezentacje dwuwymiarowe w kontekście analizy sygnału audio.

Rozdział 3 obejmuje odniesienie do modeli emocji stosowanych w literaturze.

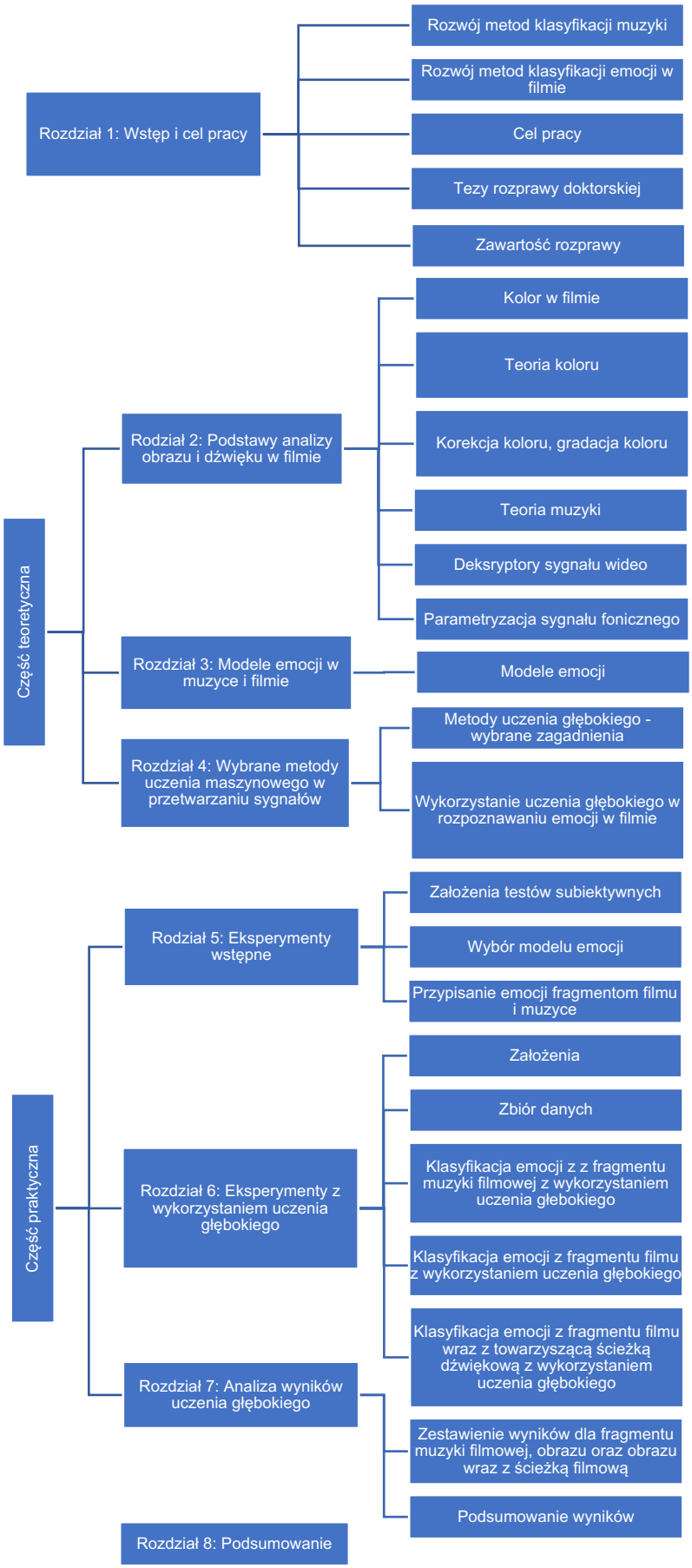
W czwartym rozdziale opisano wybrane zagadnienia związane z uczeniem maszynowym, tj. budowę sztucznych sieci neuronowych, w tym sieci splotowych oraz rekurencyjnych, opisano metody optymalizacji modeli oraz ich ocenę. Również w rozdziale czwartym zawarto opis eksperymentów na podstawie źródeł literatury, które wymieniono w tabeli 1.1.

Piąty rozdział opisuje wyniki badań wstępnych dotyczących zależności pomiędzy filmem a emocją. Na podstawie wyników ankiet określono model emocji składający się z sześciu kolorów oraz sześciu emocji, następnie potwierdzono korelację pomiędzy kolorami występującymi w filmie a emocjami, które towarzyszą tym fragmentom filmów. Pozwoliło to na stworzenie zbioru danych do treningu oraz ewaluacji zaproponowanych w kolejnym rozdziale metod uczenia głębokiego.

Metody parametryzacji sygnału audio oraz wideo, omówione w rozdziale 2, posłużyły do stworzenia klasyfikatora obejmującego film jako zbiór powiązanych, zmiennych w czasie danych audio i obrazu. W tym celu należy stworzyć architekturę pozwalającą na analizę tych dwóch sygnałów równolegle, tj. bimodalnie. Stworzenie takiej architektury opiera się o wyborze modeli dla poszczególnych modalności, które w warstwie wyjściowej przekażą do konkatenacji wstępne wyniki klasyfikacji dla poszczególnych modalności bądź parametry obliczone na podstawie danych wejściowych dla każdego z modeli z osobna. Następnie za podjęcie decyzji o ostatecznej klasyfikacji odpowiada mechanizm fuzji modeli [49, 126, 130] Rozdział szósty opisuje szereg eksperymentów mających na celu sprawdzenie najbardziej skutecznych metod parametryzacji oraz architektur sieci splotowych oraz rekurencyjnych, opis treningu sieci, wyników oraz przedstawienie modelu sieci bimodalnej wraz z uzyskanymi wynikami.

Siódmy rozdział jest podsumowaniem uzyskanych wyników dla przygotowanych w poprzednim rozdziale architektur. W ostatnim rozdziale na podstawie uzyskanych wyników wykazano osiągnięte cele badań oraz zaproponowano dalsze kierunki rozwoju.

W Dodatkach do pracy zamieszczono formularze opracowanych ankiet oraz szczegółowe wyniki testów subiektywnych.



Rys. 1.1. Struktura rozprawy doktorskiej

2. Podstawy analizy obrazu i dźwięku w filmie

2.1. Kolor w filmie

Kolor jest jednym z ważniejszych środków przekazu emocji, obok technik montażu i rodzaju doboru ujęć. Kolorystyka filmu wpływa na ustalenie klimatu sceny lub też zapewnia oddzielenie wizualne i emocjonalne danych scen, różnicuje gatunki jest znakiem rozpoznawczym reżysera (ang. *Director Trademarks*). Głównym zadaniem osób pracujących w tym kontekście przy produkcji filmu, czyli tzw. kolorystów (ang. *colorist)*, jest praca nad kolorystyczną obróbką materiału już nagranego. Najlepsi koloryści są również zapraszani na plan filmowy, aby wraz z reżyserem dyskutować o uzyskaniu odpowiednich barw ujęć jeszcze na etapie realizacji danych scen [13]. Teoria kolorów filmu bazuje na psychologii kolorów. Dla przykładu, kolor zielony ma znaczenie uspokajające, nie oznacza to jednak, że jeżeli scena w danym filmie ma wywołać stan spokoju u widza, to należy ją pokolorować, używając właśnie tej barwy, ważny jest bowiem ogólny kontekst.

Kolejnym istotnym elementem wpływającym na odbiór filmu przez widza jest muzyka filmowa, którą kompozytor tworzy na potrzeby danego filmu, skrupulatnie śledząc historię w nim zawartą. Muzyka jest nieodłączną częścią wpływania na emocje − reżyserzy działają wraz z kompozytorami i montażystami filmu, aby stworzyć muzykę pod fragment filmu bądź ułożyć montaż scen pod konkretny utwór [87].

Przyjmuje się, że pierwsza na świecie kamera została opatentowana w Anglii przez Francuza Louisa Le Prince'a w 1888 roku, a krótka scena nakręcona przy jej użyciu traktowana jest jako pierwszy na świecie zapis filmowy. W 28 grudnia 1895 roku miała miejsce pierwsza publiczna projekcja filmów braci Lumière w paryskiej kawiarni Grand Cafe, to właśnie tę datę uważa się za moment narodzin kina. W miarę rozwoju kinematografii zaczęły powstawać stałe kina, projekcje nie tylko odbywały się w parkach, na festynach czy jarmarkach. Pierwsze kino w Polsce powstało w Łodzi w 1899 roku przy ulicy Piotrowskiej 120 [74, 123, 125].

Filmy powstające na przełomie XIX i XX wieku były filmami niemymi – bez zapisu towarzyszącej ścieżki dźwiękowej, tj. dialogów, muzyki czy efektów dźwiękowych. Część przekazu emocjonalnego sceny wyrażała się w grze aktorów oraz w oprawie muzycznej, tj. muzykach, którzy synchronicznie wykonywali utwory muzyki filmowej do prezentowanego obrazu [123, 125]. Przełom nastąpił w połowie lat 20. XX wieku, kiedy wynaleziono możliwość rejestracji dźwięku na taśmie filmowej. Pełne udźwiękowienie filmu było zadaniem trudnym, nie można było nagrywać scen w plenerach, stosować dynamicznych zmian ujęć, wydłużano sceny, przez co stawały się bardziej statyczne. Z biegiem czasu technologia ewoluowała i jakość nagrań oraz odtwarzania dźwięku poprawiła się, zaczęto stosować też dubbing oraz postsynchronizację, czyli dogrywanie ścieżki dźwiękowej do nagranego wcześniej obrazu [13].

Kolejnym ważnym aspektem rozwoju kinematografii było opracowanie filmu kolorowego. W 1908 roku powstał pierwszy kolorowy film w technologii *kinemacolor*. Jednak przełom w tworzeniu kolorowych filmów powstał w latach 20. XX wieku, kiedy pojawiła się nowa technologia technicolor. System ten pozwalał na wykorzystanie dwóch, a w trakcie rozwoju koncepcji, trzech taśm światłoczułych, każda z taśm była odrębnym kolorem – czerwonym, zielonym i niebieskim, następnie w postprodukcji należało te negatywy ze sobą połączyć. Współcześnie rzadko powstają filmy w pełni czarno-białe, co nie oznacza, że zarówno reżyserzy, jak i koloryści nie wykorzystują tego typu kolorystyki przy tworzeniu obrazów. Jest to celowy krok, aby podkreślić artystyczny zamiar (rys. 2.1) [13, 123].

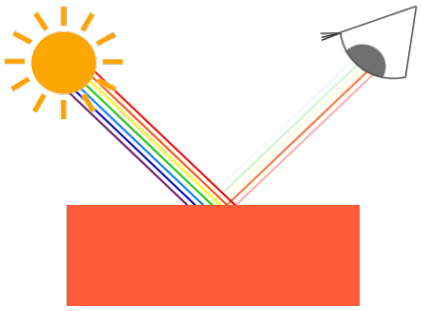
Rys. 2.1. Zwiększenie dramaturgii poprzez czarno-biały film

W latach 70. XX wieku nastąpił kolejny przełom w kinematografii; twórcami drogich i efektownych widowisk zostali George Lucas oraz Steven Spielberg. Premiera „Gwiezdnych Wojen" oraz „Bliskich Spotkań Trzeciego Stopnia" „przywróciła" rynek filmów przygodowych (przykład ujęcia z rys. 2.2). Dynamiczne cięcia, zmienna akcja oraz efekty specjalne zaczęły ponowie przyciągać uwagę widza. Wprowadzono termin „kino nowej przygody", ich twórcy nawiązywali do klasycznego hollywoodzkiego kina, ale również czerpali z różnorodności gatunków filmowych i starali się je mieszać, aby dostosowywać filmy do odbiorców w różnym wieku. Patrząc na historię kina, można zauważyć, że najważniejszym punktem w rozwoju było przejście z czarno-białego na obraz kolorowy oraz możliwość postprodukcji dźwięku. Możliwość rejestracji obrazu kolorowego oraz ścieżki dźwiękowej w filmie wzniosła odbiór dzieł filmowych na nowy poziom percepcji. Jednocześnie rozwój technologii dotyczący zarówno wyświetlania obrazu, jak i reprodukcji dźwięku zapewnił – nawet w warunkach domowych – obraz o dużej rozdzielczości i kontraście, a także systemy dźwięku dookólnego [13].

Rys. 2.2. Star Wars IV: Nowa Nadzieja – fragment filmu w kolorze

2.2. Teoria koloru

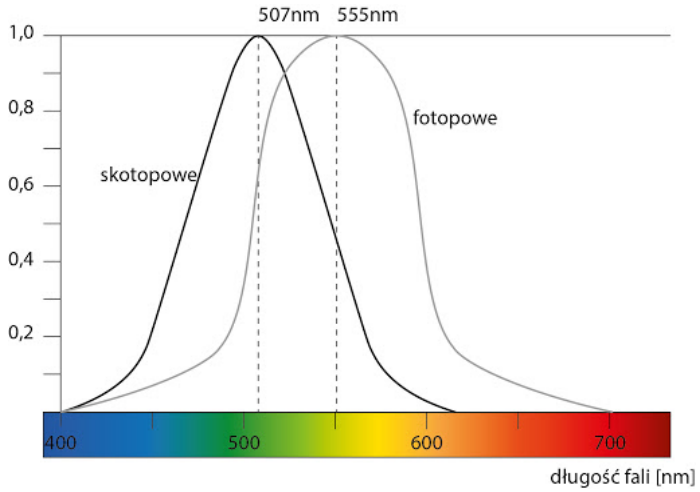
Cząsteczki zwane fotonami posiadają własności zarówno materii, jak i światła. W 1926 roku Lewis przedstawił hipotezę, według której fotony definiuje się, jako elementy świata wchodzące w interakcję z materią fizyczną, lecz niezawierające koloru i różniące się między sobą właściwościami energii [59]. Materia fizyczna odbija światło, a dzięki mechanizmom percepcji wzrok jest w stanie odpowiednio zinterpretować informację w postaci koloru (rys. 2.3). Ludzkie oko nie jest w stanie percypować całego widma ciągłego promieniowania elektromagnetycznego. w niższych częstotliwościach widma funkcjonowania widma istnieje pasmo między innymi transmisji radiowej, telewizyjnej czy też promieniowania podczerwonego. Spektrum widzialne, które ludzki narząd wzroku jest w stanie odbierać [51], znajduje się pomiędzy ultrafioletem a podczerwienią.



Rys. 2.3. Postrzeganie kolorów przez ludzki narząd wzroku

Widzenie barwne człowieka obejmuje przedział długości fali od 380 (czerwień) do 780 nm (fiolet); w skali kolorów tęczy – opisano ją cząstkowo kolorami: czerwonym, pomarańczowym, żółtym, zielonym, niebieskim, indygo oraz fioletem. Powyżej światła widzialnego znajdują się fale o wysokich częstotliwościach, między innymi ultrafiolet, promieniowanie „x" oraz promieniowanie gamma. Pomimo szerokiego spektrum kolorów tęczy światła widzialnego, podstawą teorii widzenia są trzy barwy: czerwona, niebieska oraz zielona (trójskładnikowa teoria widzenia barwnego). Przyczyną takiego widzenia jest budowa ludzkiego oka, czyli światłoczułe receptory znajdujące się w siatkówce oka – czopki oraz pręciki. Czopki odpowiadają za rozróżnianie wartości barwowych, natomiast pręciki odpowiadają za percepcję intensywności światła i stosunku światła do cieni. W siatkówce oka człowieka rozróżniane są trzy rodzaje czopków, które reagują na bodźcie w postaci długości fali światła: 400 nm dla barwy niebieskiej, 545 nm dla barwy zielonej oraz 580 nm dla barwy czerwonej (rys. 2.4) [51].

Pręciki jako drugi element receptorów światłoczułych, nie biorą udziału w widzeniu kolorów – służą uformowaniu całościowego obrazu pola widzenia poprzez reakcje na ilość światła wpadającego do oka. Dzięki pręcikom, które są czułe na niższe poziomy natężenia światła, człowiek jest w stanie dostrzec obiekty po zmroku – takie widzenie nazywane jest widzeniem nocnym bądź skotopowym [12, 39].



Rys. 2.4. Długość fali świetlnej dla widzenia nocnego i dziennego

2.2.1. Podstawowe cechy koloru

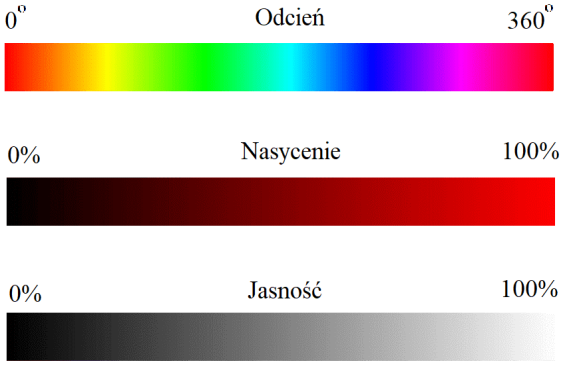
Kolor można scharakteryzować za pomocą czterech podstawowych cech: odcień (ang. *hue*), jasność (ang. *value*), nasycenie (ang. *chroma*) oraz temperatura barwowa (ang. *temperature*) [12, 118]:

- Odcień – określenie używane do konkretnej długości fali świetlnej (rys. 2.5). Cecha ta pozwala przypisać nazwy danym barwom. Człowiek jest w stanie rozróżnić około 160 odcieni barw. Newton zaproponował koło barw, na którym można zaobserwować ich ułożenie od czerwieni, idąc zgodnie z ruchem wskazówek zegara aż do najdłuższych fal, czyli do kolorów niebieskich i zielonych. W branży filmowej odcień nazywany jest również fazą (ang. *phase*) [12, 118].

Rys. 2.5. Koło odcieni kolorów [108]

- Jasność – opisuje poziom światła lub cienia na powierzchni o danym kolorze. Na jasność bardzo duży wpływ ma światło padające na przedmiot oraz materiał, z którego przedmiot jest wykonany – mianowicie jego współczynnik odbicia i pochłaniania światła [12, 118].
- Nasycenie – równoważne jest intensywności bądź wyrazistości danego koloru lub też jego przyćmienia. Największą wartość nasycenia posiadają barwy w swojej najczystszej postaci, dodawanie barw przeciwległych (według koła barw), czyli barw pochodnych, obniża intensywność. W nowoczesnych technologicznie kamerach cyfrowych, które posiadają możliwość korekcji nasycenia, podczas zmniejszania tego parametru, możliwy jest do uzyskania obraz czarno-biały [12, 118].
- Temperatura barwowa – parametr pozwalający na sklasyfikowanie koloru jako ciepły bądź zimny. Podstawą tego określenia jest psychologiczna reakcja na barwę – czerwień lub barwa podobna jest uznawana za najcieplejszą natomiast barwy niebieskie i jej odmiany za barwy najchłodniejsze. Jest to subiektywne określenie charakteru koloru. Temperatura barwowa wyrażana jest w stopniach Kelwina [K], powyżej 5000 K określone są barwy niebiesko-białe, zaś przedział 2700-3200 K zawiera barwy ciepłe (biel żółtawa po czerwień). Skala temperatury barwowej odnosi się do teoretycznego ciała doskonale czarnego – metalu pozbawionego właściwego koloru własnego nazywanego promiennikiem Plancka [12, 118].

Na rys. 2.6 pokazano skalę odcieni, nasycenia oraz jasności koloru stosowanych w edycji materiału filmowego.



Rys. 2.6. Skala odcieni, nasycenia oraz jasności koloru jako parametry w edycji materiału filmowego

Na kole barw zaproponowanym przez Isaaca Newtona można zauważyć trzy podstawowe kolory oraz trzy kolory, które są wynikiem łączenia kolorów podstawowych [12]:

- CZERWONY + NIEBIESKI = Magenta
- NIEBIESKI + ZIELONY = Cyjan
- CZERWONY + ZIELONY = Żółty

2.2.2. Modele barw

Wyznaczenie formy graficznej dla koloru próbowano opracować przez wiele lat. Większość opracowanych modeli barw stosuje się w celu klasyfikacji koloru w odniesieniu do nasycenia, dominanty, barwy czy też jasności. W środowisku filmowym – w celu określenia przestrzeni barwnych współczesnych kamer filmowych – stosuje się modele przestrzeni barw RGB (ang. *Red, Green, Blue*) / CIE (fr. *Comission Internationale de l'Eclairage*) [2, 108, 118].

System barw CIE został opracowany przez komisję CIE w roku 1931 i do tej pory stosowany jest jako standard pomiaru, wyznaczania i dopasowania koloru, lecz w zaktualizowanych wersjach. System CIE występuje w dwóch wariantach CIE RGB oraz CIE XYZ, który najczęściej stosowany jest w inżynierii wideo. Wszystkie możliwe kolory wyznaczane są za pomocą wykresu chrominancji, niezależnie od tego czy dane kolory pochodzą z emisji światła przez dane źródło, czy też są kolorami odbitymi od powierzchni, na którą pada światło.

Wśród systemów kolorów wyróżnia się też system addytywny oraz subtraktywny. System addytywny stosuje się szczególnie w wyświetlaczach sygnału telewizyjnego, monitorach komputerowych, gdzie generowane są kolorowe piksele poprzez użycie kolorów czerwonego, zielonego i niebieskiego, wszędzie tam, gdzie stosowane są w wyświetlaczach luminofory w ekranach odbiorników. Kolory addytywne powstają poprzez połączenie dwóch wiązek światła lub poprzez naprzemienne wyświetlanie z dużą częstotliwością dwóch kolorów. Subtraktywny system kolorów opiera się na barwach podstawowych RGB, zgodnie z teorią odbicia i pochłaniania kolorów, kolory na przedmiotach tworzone są na przedmiotach pochłaniających fale światła określonej długości, następnie pozostała część spektrum odbijana jest od powierzchni przedmiotu tworząc jego kolor [12, 118].

2.3. Korekcja koloru, gradacja koloru

Obecnie kamery filmowe oferują bardzo duży gamut koloru, jest to określenie zakresu kolorów reprodukowanych przez dany system przetwarzania. Wysokobudżetowe produkcje wymagają kamer sięgających kilkuset tysięcy dolarów, aby zapewnić jak najlepszą dynamikę kolorów, co za tym idzie najlepszą ich reprodukcję. Pomiar intensywności koloru w systemach cyfrowych mierzony jest w zakresie od 0 do 255, a sam kolor określony jest poprzez intensywność

kanałów RGB. Dla przykładu kolor żółty określony jest poprzez komputerowe wartości 255 dla kanału czerwonego, 255 dla kanału zielonego oraz 0 da kanału niebieskiego. Precyzja określenia odpowiedniego koloru jest sprawą kluczową dla osób zajmującą się pracą z kamerą oraz postprodukcją kolorów w filmie. Technika cyfrowa pozwala również na balans temperaturowy do rzeczywistej bieli, manualne bądź automatyczne ustawienie kamer do warunków oświetlenia pozwala na dokładną korekcję przesunięć kolorów względem emiterów światła [2, 118].

Możliwości cyfrowej obróbki sprawiły, że wszystkie procesy postprodukcyjne stały się nie tylko łatwiejsze, ale znacznie tę pracę przyspieszyły. Programy Adobe Premiere, Adobe After Effects, FinalCut, Sony Vegas czy też Davinci Resolve są to zaawansowane programy umożliwiające nie tylko pracę nad zarejestrowanym materiałem, lecz również skomplikowane prace nad korekcją kolorów filmu. Cyfryzacja kamer pozwoliła również na stworzenie cyfrowego negatywu, tzw. formatu RAW (z ang. surowy). Format RAW posiada bardzo dużą rozpiętość względem dynamiki obrazu, nie przypominając końcowej wersji filmu w kontekście kolorystyki, umożliwiają zapis informacji w zakresie dwunasto- lub czternastobitowym dla każdego z pikseli. Proces konwersji plików RAW nazywany jest wywołaniem plików RAW i nie stanowi jeszcze obrazu wynikowego, który nadawałby się do publikacji, trzeba wcześniej go odpowiednio przygotować [12].

W produkcji filmowej wyróżnia się kilka osobnych procesów powiązanych z kształtowaniem barwnym filmu [13]:

- Utworzeniu profilu obrazu – stworzenie charakterystyki jak film powinien wizualnie wyglądać;
- Korekcja kolorów materiału filmowego / ujęć;
- Gradacja kolorów w całym filmie.

Utworzenie odpowiedniego profilu obrazu jest wyborem odpowiedniego charakteru kolorystycznego filmu i ustaleniem go pomiędzy kolorystą a reżyserem filmu. Podczas prac nad uzyskaniem satysfakcjonujących kolorów danego filmu można wyróżnić dwa istotne terminy, tj.: *color correction* oraz *color grading*. Oba te pojęcia można znaleźć pod wspólnym hasłem kolor-korekcja i są one nierozłączne [13, 17, 18]. Oba te procesy są żmudne oraz trudne ze względu na cechy koloru, na które trzeba zwracać uwagę [106].

- Korekcja koloru – dopasowanie kolorystyczne – w zakres wchodzą takie prace jak balans kolorów, dopasowanie temperatury barwowej, dopasowanie kontrastu, nasycenia kolorów. Praca nad kolorami w kolejnych ujęciach musi trwać, dopóki kolorysta nie uzyska jednolitego ciągu wizualnego historii. Podczas nagrań wiele czynników wprowadza niespójności kolorystyczne, które należy skorygować, są to między innymi: różne źródła światła, różne pory dnia, zmiana pogody, różnorodność wnętrz, a także realizacja nagrań różnymi kamerami. Korekcje wykonuje się pomiędzy poszczególnymi ujęciami, aby zachować balans kolorów i światła w całościowym materiale filmowym (rys. 2.7) [106, 118].

Rys. 2.7. Korekcja koloru (ang. *Color correction*)

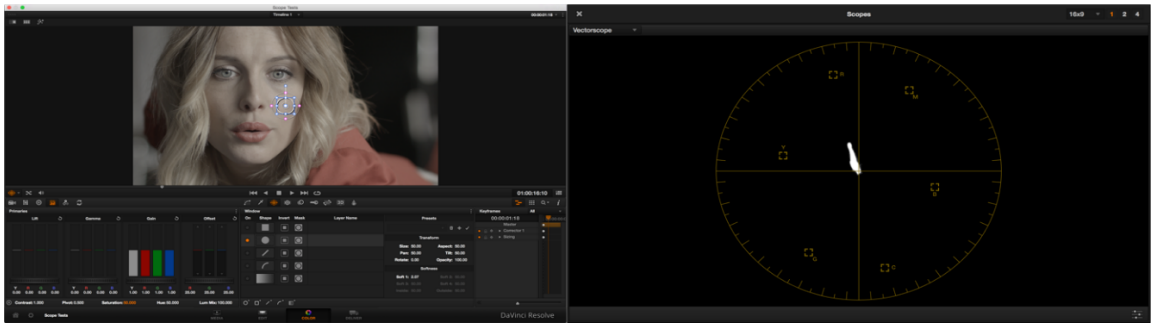
- Gradacja koloru – jest zarówno procesem kreatywnym, jak i pracą w aspekcie technicznym. Pozwala ukierunkować kolorystykę obrazu oraz wspomagać opowiadanie historii zawartej w filmie, co za tym idzie, pozwala wpływać na emocje widza podczas seansu. Prace nad *color gradingiem* przebiegają już na wcześniej przygotowanym materiale podczas kolor korekcji (poprzedni etap prac nad balansem kolorów), proces ten wykonuje się na obróbce całościowego materiału (rys. 2. 8) [12, 106].

Rys. 2.8. Gradacja koloru (ang. *color grading*)

W pracy nad zmianą kolorów w filmie, zarówno w procesie korekcji koloru czy na etapie gradacji koloru kolorysta ma do dyspozycji szereg narzędzi umożliwiających manipulację kolorami. Poza zmianą poszczególnych parametrów obrazu oraz zmianą wartości liczbowych wyróżnia się również główne narzędzia, które używane są podczas pracy nad kolorem w filmie, np.: oscyloskop, wektoroskop, krzywe koloru, histogram koloru [12, 13, 106 118].

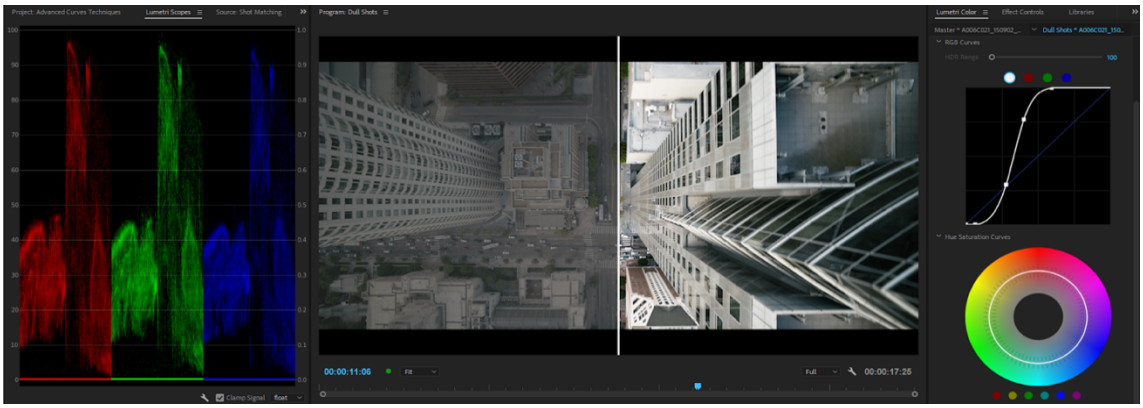
- Oscyloskop oraz wektoroskop są podstawowymi cyfrowymi narzędziami dostępnymi w każdym profesjonalnym oprogramowaniu służącym do edycji filmów. Oscyloskop wskazuje pomiar napięcia sygnału wideo w poszczególnych pasmach jasności, wskazuje na ewentualne przesterowanie sygnału. Natomiast wektoroskop służy ocenie balansu pomiędzy sześcioma barwami, trzema podstawowymi RGB oraz ich pochodnymi CMY, czyli do pomiaru chrominancji (koloru). Zmiany amplitudy i fazy sygnału zaznaczone na wektoroskopie wskazują na słabsze bądź mocniejsze odchylenia względem któregoś z

kolorów. Narzędzie to jest bardzo pomocne w celu określenia koloru skóry, zostało to zaprezentowane na rys. 2.9.



Rys. 2.9. Wektoroskop jako narzędzie do manipulowania kolorem podczas edycji filmu

- Krzywe koloru – jedno z ważnych narzędzi pozwalających na manipulację kanałami luminancji oraz chrominancji. Narzędzie to umożliwia punktowy zmiany krzywej, aby dopasować wysokość poziomów poszczególnych pasm wybranych kanałów (rys. 2.10).
- Histogram – histogram używany jest w celu sprawdzenia każdego z 255 pasm luminancji bądź kanałów RGB zarejestrowanego obrazu w celu odpowiedniego zbalansowania poszczególnych kanałów (rys. 2.10).



Rys. 2.10. Histogram oraz krzywe – narzędzia do edycji kolorów w filmie

Wstępnym ustaleniem interesujących kolorystę kolorów mogą być tablice LUT (ang. *Lookup Table*, LUT). Podczas pracy nad kolorystyką filmu używa się tablic LUT 3D, są to trójwymiarowe układy współrzędnych, każda z osi jest punktem wyjścia dla jednej z trzech barw składowych. Pozwalają na wstępną bądź końcową obróbkę kolorystyki obrazu z dostosowaniem nie tylko odpowiednich kolorów, ale również jasności, poprawy cieni, korygując ekspozycję i balans kolorów (temperaturę barwową). Używając do nagrań format RAW, otrzymuje się obraz bezpośrednio z matrycy światłoczułej, który należy poprzez wcześniej wspomniane działania przygotować do dalszej postprodukcji.

Autorka książki „Jeśli fiolet to ktoś umrze” [8] pokazuje na bazie dwudziestoletnich badań, że to kolor decyduje o ludzkich odczuciach. Kolor, mimo że często niezauważalny przez widzów może być narzędziem manipulacji, które potrafi odzwierciedlić nie tylko odczucia bohaterów, ale przedstawić daną scenę w bardziej dosłownym aspekcie. Niektórzy reżyserzy stosują zmianę kolorów w filmie, aby ukazać transformację bohaterów, inne mają wyjątkowe sceny ubrane w główne kolory, aby podkreślić rozwój fabuły. Kolory również służą w celu podkreślenia klimatu sceny lub jej oddzielenia wizualnego względem całego filmu [8, 70].

Poniżej wymieniono główne aspekty wykorzystania koloru w filmie:

- Rozwój akcji w filmie;
- Spotęgowanie emocji u widza;
- Umocnienie narracji wizualnej;
- Ukazanie rozwoju emocjonalnego bohaterów;
- Wskazanie ważnych detali w scenie;
- Ukazanie tonu filmu.

Używanie odpowiednich kolorów podczas opowiadania filmowej historii jest równie istotne jak dialogi, gra aktorska, ruch kamery czy też sceneria. Z psychologii kolorów korzystają nie tylko osoby zajmujące się produkcją filmu, lecz również graficy oraz kreatorzy znaków firmowych. Jednak można zauważyć, że artyści tacy, jak Tim Burton, Wes Anderson czy też Quentin Tarantino stworzyli na potrzeby wyróżnienia swoich produkcji własny system obrazowania kolorami, dzięki któremu można łatwo rozpoznać dzieło danego reżysera. Są to zmiany przemyślane i celowe, aby zwrócić uwagę widza na własne produkcje, które mogą pozostać w pamięci na długo [17]. Na rys. 2.11 przedstawiono koło barw, którym twórcy często się kierują w opowiadaniu filmowej historii.

Należy również wziąć pod uwagę aspekt zróżnicowania koła teorii kolorów względem różnych kultur i krajów. Przykładem jest chociażby kolor czarny, który w kulturach zachodnich może być utożsamiany ze śmiercią czy też żałobą, natomiast w kulturach wschodu w tych sytuacjach użytym kolorem będzie biały. Te same kolory mogą często odpowiadać za skrajne emocje, w zależności od kontekstu, stanu emocjonalnego bohaterów czy opisu sytuacji, w jakich dany kolor pojawia się w opowieści pokazywanej na ekranie [13, 70].

Rys. 2.11 Koło barw

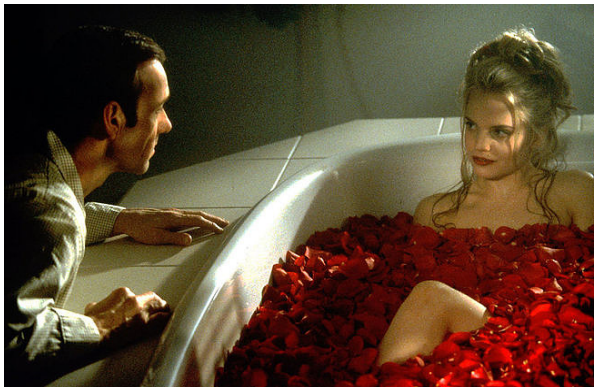
W kolejnych podrozdziałach podane zostanie odniesienie do tego, w jaki sposób twórcy filmowi wykorzystują wybrane kolory do przedstawienia fabuły, uwypuklenia danej sceny, nadania cech bohaterom, itd. W prezentacji tej tematyki posłużono się opisem filmów, których fragmenty zostały wykorzystane do stworzenia bazy danych oraz stały się podstawą testów subiektywnych.

2.3.1. Kolor czerwony

Kolor czerwony jest kolorem najbardziej wyrazistym. Określa się ten kolor jako żywą ilustrację uczuć człowieka, z jednej strony potrafi pobudzić widza do pozytywnych odczuć i emocji, a z drugiej strony może potęgować agresję czy złość.

Kolory w filmach pojawiają się najczęściej jako ich kombinacje. Aby jednak podkreślić istotę akcji można zauważyć, że niektóre kolory bądź jeden kolor jest kolorem dominującym. Przykładem dominanty koloru czerwonego jest film w reżyserii Sama Mendesa „American Beauty" (1999) (rys. 2.12). Opowieść o dysfunkcyjnej rodzinie zawarta jest w kolorach

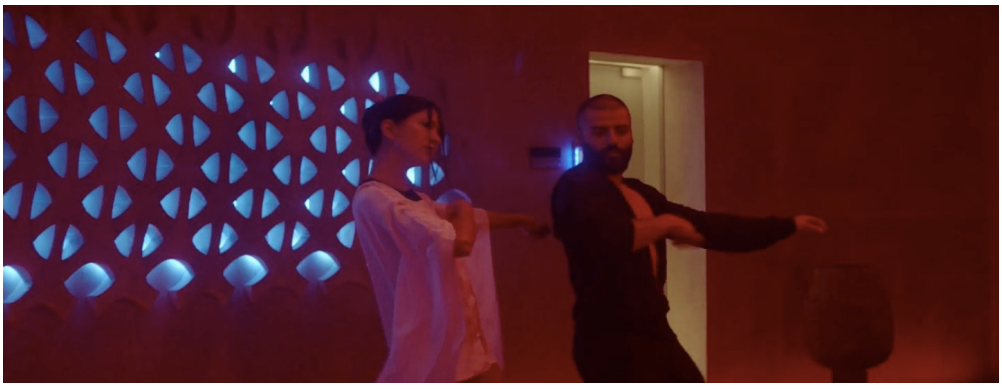
czerwonych, białych oraz błękitnych. Te trzy kolory mają dostarczyć widzowi odbiór − z jednej strony stereotypowej rodziny, a z drugiej uświadomić wewnętrzne problemy i rozterki wśród jej członków. Ostatecznie czerwień jest kolorem prowadzącym przez historię, z opowieści romantyczno-marzycielskiej przeistacza się w dramat, widz przyzwyczaja się do koloru czerwonego i kolor staje się nieodłączną częścią filmu [8, 77, 109].



Rys. 2.12. Ujęcie z filmu „American Beauty”

W roku 1999 na ekranach kin pojawił się film z Brucem Willisem pod tytułem „Szósty zmysł”. Już pierwsza scena atakuje widza kolorem czerwonym, pojawiają się czerwone drzwi kościoła, czerwony dywan w kościele i czerwona przystrojona statuetka Jezusa. Czerwień w tym filmie ma charakter przewodni oraz nieustannie niepokoi widza. Aspekt czerwieni nie tylko przejawia się w wystroju kościoła, ale również reżyser uchwycił w kolorze czerwieni siły dobra oraz zła. Ostatecznie czerwień − dominująca w całym filmie − jest kluczem do finałowej sceny i rozwiązania tajemnicy [70, 77, 109].

Przykładem nowszej produkcji jest film Alexa Garlanda z 2015 roku pod tytułem „Ex Machina” (rys. 2.13). Jest to film z gatunku thrillerów, którego głównym bohaterem jest Caleb, pracownik firmy informatycznej, w której wygrywa wewnętrzny konkurs – nagrodą w konkursie jest spędzenie tygodnia w posiadłości swojego szefa. Szef Caleba pracuje nad wizjonerską sztuczną inteligencją, która ma w przyszłości zastąpić gatunek ludzki. Garland wykorzystuje bogatą paletę kolorów w filmie, lecz jednym z najbardziej zauważalnych jest kolor czerwony, symbolizuje on emocje takie jak złość czy przerażenie, ale również miłosną ekstazę. Scena rozmowy ze sztuczną inteligencją oraz scena tańca, w której główny bohater tańczy z jedną z bohaterek, najbardziej zapadają widzowi w pamięć [70, 77, 109].



Rys. 2.13. Ujęcie z filmu „Ex Machina”

2.3.2. *Kolor żółty*

Kolor żółty jest kolorem pełen zaprzeczeń, z jednej strony kojarzony jest z kolorem słońca czy kwiatów, ale z drugiej strony można sobie wyobrazić jadowite owady, takie jak szerszenie czy też osy. Jego znaczenie można traktować jako przekaz ostrzeżeń, widocznych nie tylko w naturze, ale również we współczesnym świecie jako oznakowania na drodze czy budowie. Jednak przede wszystkim kolor ten kojarzony jest ze słońcem, życiodajną energią. Harry Hepner, profesor psychologii reklamy, twierdził, że kolor żółty jest kolorem, który ludzie zapamiętują najlepiej. Cecha ta sprawia, że kolor żółty staje się odniesieniem do obsesji, której przykładem są kadry filmu „Taksówkarz" z 1976 roku w roli głównej z Robertem De Niro (rys. 2.14). Już w pierwszych scenach Martin Scorsese sprawia, że żółty kolor prowadzi widza przez cały film [8, 77, 109].

Rys. 2.14. Ujęcie z filmu „Taksówkarz"

Jednym z najbardziej docenionych filmów kostiumowych ze względu na zdjęcia oraz muzykę jest niewątpliwie „Gladiator" w reżyserii Ridley'a Scotta z 2000 roku. Reżyser często pokazuje widzowi kadry, na których widoczne są łany zboża. Żółty kolor jest pokazany jako uczucie ukojenia i spokoju, których wizja przynosi głównemu bohaterowi na myśl dom rodzinny [8, 77, 109].

Kolor żółty posiada również funkcję relaksacyjną, widać to na przykładzie filmu „Hotel Chevalier" (2007) w reżyserii Wesa Andersona, który lubi intrygować widzów charakterystycznym i bardzo nasyconym w kolorystykę podejściem do kreowania filmów (rys. 2.15). „Hotel Chevalier" jest trzynastominutowym filmem traktowanym jako prolog do filmu „Pociąg do Dajeeling". Film kończy się sceną erotyczną i zostawia dużo swobody do dalszej interpretacji przez widza [8, 77, 109].

Rys. 2.15. Ujęcie z filmu „Hotel Chevalier"

2.3.3. *Kolor niebieski*

Psychologia przekonuje, że otoczenie, w którym się przebywa, oddziałuje na emocje i może powodować introspekcje. Pierwszym skojarzeniem, jakie może przyjść na myśl w odniesieniu do

koloru niebieskiego jest niebo bądź też woda, dlatego niebieski może być również odzwierciedleniem lojalności, czegoś co jest pewne, a w przypadku związków międzyludzkich − pożądane. Często używane są odmiany koloru niebieskiego − z badań wynika, że turkus jest kolorem inspirującym otwartość i interakcje, natomiast bledszy i chłodniejszy błękit powoduje wyciszenie się i daje poczucie spokoju.

Film „Skazani na Shawshank" (1994 r., reż. Frank Darabont) „atakuje" widza niebieskimi zimnymi kolorami. Zastosowany w opowieści szary błękit uwypukla wszechobecną w opowieści melancholię. W ciemnym i zimnym błękitno-niebieskim świetle przedstawiono naczelnika więzienia, który dopuszcza się na więźniach przemocy, przez co ten wątek staje się bardziej wyrazisty dla widza. Dodatkowo kolor niebieski jest odzwierciedleniem bezsilności, przedstawienie filmu właśnie w takiej kolorystyce pozwala widzom wczuć się w sytuację więźniów [70, 77, 109].

Dany film nie musi być w całości objęty jednym kolorem, poszczególne kolory mogą być użyte w kilku bądź w jednej scenie, aby zwrócić uwagę widza na konkretny aspekt filmu. Kolor niebieski w swoim znaczeniu ma również za zadanie dać widzowi odczuć chłód czy pasywność bohaterów względem jakichś wydarzeń. W filmie „Szósty zmysł" dominujący jest kolor czerwony, lecz w filmie pojawia się również kolor niebieski − został on użyty do przedstawienia jednej z postaci (rys. 2.16). Sceneria w kolorze błękitnym i niebieskim wzmaga emocje, szczególnie w momencie sceny brutalnego ataku [8, 77, 109].

Rys. 2.16. Ujęcie z filmu „Szósty zmysł"

2.3.4. Kolor pomarańczowy

Kolor pomarańczowy jest kolorem przyjaznym w odbiorze. Jest kojarzony z entuzjazmem i jest jednocześnie w najmniejszym stopniu powiązany z dramatyzmem. Kolor ten wspiera i oddaje ciepłe i przyjazne usposobienie akcji w filmie. Często jest spotykany w filmowych kadrach zachodzącego nieba, który uzmysławia widzowi przyjazną atmosferę i poszerza pole doświadczeń emocjonalnych. Jeśli jest zastosowany we wnętrzach, to można go odczytać jako kolor romantyczny. Pod kolor pomarańczowy zostały włączone również kolory ziemi, na który również człowiek reaguje pozytywnie.

Głównym motywem kolorystycznym w serii ujęć otwierającej dzieło Francisa Coppola „Ojciec Chrzestny" (1972) jest pomarańczowe, a wręcz bursztynowe światło (rys. 2.17). Realizacja z pomarańczowym światłem w zamyśle reżysera jest w jakimś stopniu przewrotna i jest odzwierciedleniem fascynacji przestępczym światem. W dalszej części filmu, w której odbywa się wesele, widać postaci noszące akcenty koloru pomarańczowego, a dokładnie pastelowo-pomarańczowego − oddaje to iluzję pozytywnych odczuć i przyjaznej atmosfery, nie budzi skojarzeń z mafią, lecz niewinnością [8, 77, 109].

Rys. 2.17. Ujęcie z filmu „Ojciec Chrzestny”

Harrison Ford znany jest głównie z ról w Gwiezdnych Wojnach oraz jako postać Indiana Jones. Aktor zagrał również „Łowcę Androidów” (1982) w reżyserii Ridleya Scotta. Kolor pomarańczowy jest widoczny w otoczeniu, rzuca poświaty poprzez gęste od spalin powietrze w mieście, można wyczuć przytłaczającą i toksyczną atmosferę miasta. Ridley Scott jest mistrzem w doborze odpowiednich kolorów do danych sytuacji, kolor pomarańczowy w jego produkcji z 1982 roku ma dwa aspekty, wcześniej wspomniany aspekt toksyczności miasta oraz otoczenia głównego bohatera, a z drugiej strony ukazanie atmosfery romantyzmu i spowodowanie przychylnego odbioru głównego bohatera, który zakochuje się w jednym z androidów, na które sam wcześniej polował.

W filmie Mad Max: Na drodze gniewu (2015) kolor pomarańczowy został bardzo mocno wyeksponowany w celu przybliżenia lokalizacji, w której dzieje się akcja, ogromnej pustyni, oraz aby pokazać agresywne emocje (rys. 2.18). Ognista wręcz paleta kolorów wskazuje na bardzo mocny konflikt głównego bohatera z antagonistą filmu. Dodatkowo dialogi pokazane są na mocniejszych zbliżeniach, kolorystyka pomarańczowa połączona z mocnym czarnym kontrastem sugeruje niebezpieczeństwo [8, 77, 109].

Rys. 2.18. Ujęcie z filmu „Mad Max: Na drodze gniewu”

2.3.5. Kolor zielony

Kolor zielony z jednej strony oznacza życie, dzięki pozytywnym skojarzeniom z naturą. Jednocześnie może się kojarzyć z niebezpieczeństwem, żeglarze mówią: „Strzeż się zielonej wody”. W „Cienkiej czerwonej linii” kiedy młody człowiek wpada do zielonej wody i znika w morzu wśród zielonych i falujących podwodnych traw, kolor zielony sprawia, że historia nabiera większego emocjonalnego znaczenia. W filmach Walta Disneya czy też w „Przekleństwie nienawiści” bądź „Tajemniczej zbrodni” można zauważyć, że kolor zielony został przypisany do trucizny i zła. W tym kontekście pojawił się już w roku 1937 w „Królewnie Śnieżce i siedmiu krasnoludkach” w reżyserii Davida Handa. Dlatego kolor zielony jest silnie zakorzeniony w ludzkich umysłach w odniesieniu do trucizny. Zieleń stała się metaforą strutego społeczeństwa. Jeżeli kolor zielony pojawia się w skojarzeniu z ludzkim ciałem i organizmem, to zieleń oznacza zazwyczaj chorobę lub zło [8, 77, 109].

Film „Szeregowiec Ryan" w reżyserii Stevena Spilberga z 1998 roku utrzymany jest w kolorystyce zimnego i mdłego koloru zielonego. Ujęcia, na których widoczne są śnieżnobiałe groby, które otacza późnowiosenna zieleń są alegorią do tego, że młodzi ludzie, polegli w trakcie II Wojny Światowej, nie zobaczą już kolejnej wiosny. Jest to zasadniczy kontrastowy element tej sceny. Znaczeniowo − podobne motywy koloru zielonego − reżyserzy wykorzystali między innymi w filmie „The Machanist", w którym widz również jest traktowany zimnym i mdłym kolorem zielonym, który wskazuje na monotonię oraz niechęć do życia głównego bohatera. Również w filmie „The Matrix" kolor zielony na początku filmu odwołuje się do świata wirtualnego, nawiązuje do skojarzeń programistycznych, widz może odczuć, że ten świat jest złowieszczy, mroczny oraz niebezpieczny (rys. 2.19) [8, 77, 109].

Rys. 2.19. Ujęcie z filmu „Matrix"

W filmie „Wszystko za życie" z 2007 roku w reżyserii Seana Penna widz szybko przywiązuje się do rozterek życiowych bohatera. Kiedy główny bohater wyrusza autostopem na Alaskę, kolor zielony odnosi się do sił natury.

Warto również zwrócić uwagę na animowane bajki Walta Disneya, gdzie kolor zielony jest przypisany złym postaciom. W każdej ze negatywnych postaci Disneya, gdy wśród atrybutów przypisanych do tych bohaterów pojawia się kolor zielony, można wyczuć ich charakterystyczne cechy: chciwość, podłość oraz chorobliwa chęć władzy. Postać czarownicy ze „Śpiącej Królewny" pojawiła się już w 1959 roku (rys. 2.20). Natomiast niedawno w kinach (tj. 2015 oraz 2019) pojawiła się ekranizacja animowanej bajki z Angeliną Jolie w roli głównej. Zarówno w filmie, jak i w animacji z lat 50. XX wieku można zauważyć, że czarownica stale jest przedstawiana w zielonym świetle: zielony blask wokół jej zamku, podczas rzucania klątwy towarzyszy jej również zielona poświata wraz z zielonym niebem, kiedy przybiera smoczą postać również można zauważyć kwaśno-zielony płomień. Disney ukazuje złe postaci, wiążąc je z kolorem zielonym. W przypadku czarownicy jej historia odzwierciedla negatywne cechy, które można z tym kolorem połączyć: kiedy zostaje odrzucona przez rodzinę królewską jest przepełniona urazą (zazdrość) wobec królowej, planuje i dokonuje zemsty (chciwość), rzucając na młodą księżniczkę klątwę (choroba) [8, 77, 109].

Rys. 2.20. Ujęcie z filmu animowanego „Śpiąca Królewna”

2.3.6. Kolor fioletowy

W poezji fiolet kojarzy się ze zmysłowością, ale też ze światem paranormalnym bądź mistycznym. Przykładem mistycznej transformacji jest film „Gladiator”, w którym już w pierwszych scenach pojawia się Marek Aureliusz rzymski cesarz, okryty fioletowym kapturem, niedługo później umiera. Podobną wymowę ma kolor fioletowy w filmie „Szósty zmysł”, w którym bohaterka w opalizującej purpurze, zastaje martwego męża. Wizualnie widz ma odczucie, że ta sceneria prowadzi w kierunku śmierci. Jak w każdym przypadku, kolor fioletowy ma również swoje pozytywne odzwierciedlenie, które uwypukla bohatera jako szlachetnego w swoich poczynaniach. Często też ten kolor nazywany jest kolorem królewskim, ale może być również łączony z aspektem seksualności [8].

W filmie „Lost River” w reżyserii Ryana Goslinga z 2014 widz spotyka się ze scenami, które przedstawiają jedną z najbardziej uwodzicielskich postaci w filmach z ostatnich dwudziestu lat (rys. 2.21). Bohaterka odsłania zmysłową sylwetkę w blasku ciemnej purpury, lecz jej twarz pozostaje w cieniu, co powoduje przekaz o tajemniczej naturze tej postaci. Motywy niejednoznaczności pojawiają się w tym filmie na wiele sposobów [8, 77, 109].

Rys. 2.21. Ujęcie z filmu „Lost River”

Fiolet czy też purpura wykorzystywane są często w filmach z gatunku *fantasy*. Jest to widoczne w filmie „Strażnicy Galaktyki” z 2014 roku, gdzie pokazany jest mistycyzm chwili, a też do ukazania transformacji głównego bohatera [8, 77, 109].

Obecnie coraz częściej stosuje się intonacje różnych kolorów w kontekście poszczególnych ujęć w filmach, w ramach podkreślenia emocji związanych z bohaterem i przybliżenia widza do tych wydarzeń. W filmie Joker (2019), w reżyserii Todda Philipsa gra kolorami opiera się głównie na trzech kolorach, których przykład odzwierciedla psychologię kolorów (rys. 2.22). Są to: kolor niebieski, żółty oraz czerwony. Użycie koloru niebieskiego podkreśla nieprzyjazne środowisko, w którym żyje bohater, całe miasto jest w tonacji blado-zimnego niebieskiego koloru, przez co pokazuje, że protagonista jest wyobcowany. Kolor niebieski jest dominującym kolorem

pierwszego aktu. Dobitnie pokazują to sceny, w której można wyczuć atmosferę chłodu oraz melancholii, sytuacja, w której bohater nie może poradzić nic na chorobę matki i jego pasywność w działaniach. Dodatkowo bierna postawa terapeutki w stosunku do tytułowego „Jokera" również w scenerii koloru niebieskiego, daje odczucie izolacji bohatera. Ilekroć widz widzi kolor niebieski, to może odnieść wrażenie, że ten kolor działa przeciwko bohaterowi [8, 77, 109].

Rys. 2.22. Ujęcie z filmu „Joker"

Podsumowując, kolor jest ściśle powiązany z emocjami i pośrednio określa gatunek filmowy.

2.4. Teoria muzyki

"Muzyka" należy do jednych z trudniejszych do zdefiniowania terminów. Można znaleźć filozoficzną myśl Jacquesa Attaliego: "muzyka jest dźwiękowym wydarzeniem pomiędzy hałasem a ciszą" [5] czy filozofa Heideggera: "muzyka jest czymś w czym działa prawda" [32], po bardziej konserwatywne i techniczne określenia terminu. Muzykolog Charles Seeger zaznacza, że: "muzyka jest to system komunikacji zawierający dźwięki stworzone przez członków społeczności, aby porozumiewać się z innymi członkami tejże społeczności". Według Basongye'a: "muzyka jest czystym produktem ludzkim, kiedy jesteś zadowolony, to śpiewasz, kiedy jest zły, to śpiewasz" [32]. Podsumowując, niezależnie od różnorodności muzyki w różnych częściach świata i różnych kulturach, można wskazać, że muzyka [71]:

- składa się z dźwięku;
- składa się również z ciszy;
- jest świadomie stworzoną sztuką;
- jest zorganizowanym po ludzku dźwiękiem.

2.4.1. Składowe utworu muzycznego

Na postawie powyższych przesłanek, można stworzyć roboczą definicję terminu „muzyka". Muzyka jest celowo zorganizowaną formą sztuki, której medium jest dźwięk i cisza, z podstawowymi elementami wysokości (melodia i harmonia), rytmu (metrum, tempo i artykulacja), dynamiki, oraz walory barwy i faktury (rys. 2.23) [15, 61]. Zatem muzyka jest zjawiskiem rozchodzącym się w dźwięku i czasie, jest komunikacją, która wymaga słuchania, przetwarzania i odpowiadania. Jones [42] zdefiniował cztery elementy: harmonię, melodię, rytm i tempo, natomiast w pracy Peretza [87] pojawiają się takie opisy, jak: struktura, melodia, rytm i artykulacja. Poniżej, na rys. 2.23 przedstawiono zestawienie wyżej wspomnianych elementów. Zostały one pogrupowane w cztery różne klasy w zależności od cech wspólnych, związanych z czasem, wysokością tonu, dynamiką oraz interpretacją. Każdy element z jednej grupy wpływa na element z innej grupy, pomimo ich odrębnego pogrupowania [72].

Na podstawie czasu	Wysokość tonu	Dynamika	Interpretacja
• Rytm • Tempo • Metrum	• Wysokość tonu • Melodia • Harmonia	• Głośność • Dynamika • Agogika	• Artykulacja • Barwa • Fraza

Rys. 2.23. Składowe utworu muzycznego

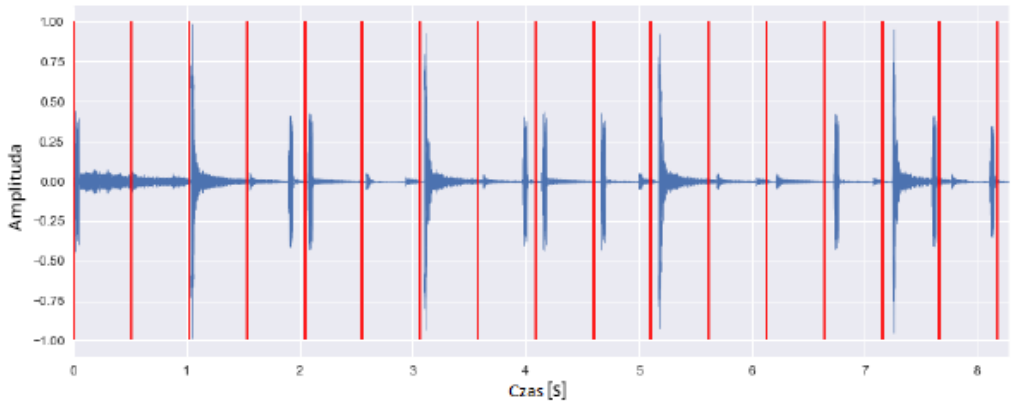
2.4.2. Elementy określane na podstawie czasu

Warto zauważyć, że składowe opisujące muzykę są w ostatnich dekadach wykorzystywane w obszarze, jakim jest automatyczne wyszukiwanie informacji muzycznej MIR (ang. *Music Information Retrieval*). Parametry związane z cechami muzyki wykorzystywane są często do automatycznej klasyfikacji utworów pod kątem gatunków czy też automatycznego tworzenia list muzycznych i rekomendacji [22, 69, 102, 124]. Jedną z metod wykrywania tempa jest kwantyzacja próbek sygnału audio, w tym przypadku utworu muzycznego, wykrywanie rytmu jest nieco prostsze. Skupia się ono na wykrywaniu czasu trwania poszczególnych dźwięków w dziele muzycznym. W systemach MIR badacze skupiają się na wykrywaniu niskopoziomowych cech rytmicznych, w szczególności wykrywanie tonów basowych, które wykorzystywane są do organizowania i przewidywania sygnału muzycznego.

Rytm bazuje na charakterystycznych akcentach, pauzach oraz dźwiękach w utworze muzycznym [55], widoczny jest w równomiernie rozłożonych w czasie wzrostach energii, zatem jest to miara dla podstawowej okresowości muzyki [56]. Związane z rytmem jest tempo **(**agogika**)**, które oznacza, jak szybko dany utwór ma być wykonany. Rytm może również wskazywać na podobieństwo cech gatunkowych w muzyce [99].

Metrum opisywane jest jako regularne akcenty, które porządkują rytm utworu muzycznego. Wyróżnia się metrum parzyste (4/4, 6/8, 2/4) oraz nieparzyste (3/4, 3/8. 9/8). Górna liczba stanowi wartość nut w takcie. Dolna określa, jakimi nutami metrum jest liczone (np. ćwierćnuty, ósemki) [99].

Tempo określa liczbę uderzeń na minutę (ang. *beats per minute* – BPM). Określa, ile ćwierćnut znajduje się w minucie i wyznacza czas trwania nut (rys. 2.24) [99].

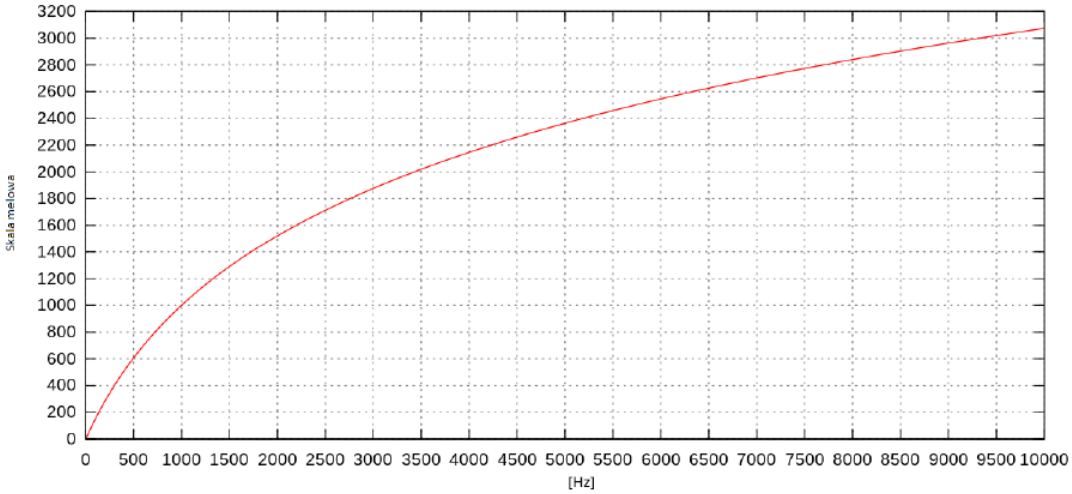


Rys. 2.24. Przykładowy wykres ilustrujący tempo utworu

2.4.3. Elementy określane na podstawie wysokości tonu

Wyznaczenie wysokości tonu umożliwia określenie położenia tonu na skali częstotliwości dla danego dźwięku (rys. 2.25). Wysokość tonu związana jest z częstotliwością w hercach (*Hz*), ale

też wykorzystuje się jednostkę *mel*, która odnosi się do percepcji słyszenia. Skalę melową wyznaczono zgodnie z definicją wysokości tonu, czyli na podstawie tonów prostych. W wyniku doświadczeń wrażenie wysokości dźwięku zależy w dużej mierze również od poziomu głośności dźwięku, dlatego też w definicji punkt referencyjny 1000 meli przyjęto w zależności od częstotliwości 1000 Hz dla tonu [66].



Rys. 2.25. Wykres obrazująca zależność wysokości tonu w [Hz] i melach

Melodia – definicja podaje, że jest to zorganizowana grupa dźwięków, mających różną wysokość. Może mieć również różny zakres częstotliwościowy. Melodię można podzielić na monofonię oraz polifonię. Monofonia oznacza jedną linię melodyczną, przykładem monofonii jest chorał gregoriański, natomiast polifonia to jednoczesne współdzielenie dwóch lub wielu linii melodycznych. Relacja pomiędzy kolejnymi dźwiękami (nutami) zależy od przerw pomiędzy nimi, których liczba zależy od liczby półtonów. Obecnie rzadko spotykane są utwory monofoniczne, które były domeną wczesnej muzyki średniowiecznej. Współczesna muzyka jest muzyką polifoniczną, w której dominuje zjawisko zachowania balansu pomiędzy różnymi tonami, czyli zachowana jest harmonia. Poszczególne dźwięki czy też głosy łączą się w jedną spójną całość, taką całością nazywa się akord. Kompozytorzy, tworząc partytury, mają na uwadze całe grupy oraz pojedyncze instrumenty. Jeśli wszyscy muzycy zagrają dźwięki wpasowujące się w ten sam akord, nazywa się to harmonią zgodną, jeżeli jeden dźwięk nie będzie pasował do akordu nazywany jest dysonansem. W jazzie, muzycy używają tak zwanej implikowanej harmonii, to znaczy, że nie zagrają wszystkich nut w akordzie, dzięki czemu odbiorca sam jest w stanie dopowiedzieć sobie niezagraną nutę [6, 65].

2.4.4. Elementy dynamiki w utworze muzycznym

Głośność jest zjawiskiem subiektywnym i rośnie wraz ze wzrostem ciśnienia akustycznego (ang. Sound Pressure Level, SPL), co za tym idzie rośnie wraz ze wzrostem poziomu dźwięku. Danemu poziomowi natężenia dźwięku w [dB] odpowiada poziom głośności liczony w fonach, który jest wyznaczony przez krzywe głośności (tzw. krzywe izofoniczne). Wprowadzono też skalę głośności [son]. Oparta jest ona na porównaniu głośności danego dźwięku do dźwięku wzorcowego o częstotliwości 1 kHz i poziomie 40 dB. Skalę głośności ustalono na podstawie wyników oceny słuchaczy [43, 90].

Obecnie w obszarze inżynierii dźwięku stosuje się skale, które mają odniesienie do percepcji: LU – jednostka głośności (ang. *Loudness Units*), LUFS – jednostka głośności w odniesieniu do pełnej skali (cyfrowej) (ang. *Loudness Unit Referenced to Full Scale*). Są to jednostki pozwalające ujednolicić poziomy głośności − w głównej mierze w telewizji, radiu, ale również w produkcjach wydawnictw cyfrowych (Netflix, Amazon Prime, YouTube, Spotify) oraz kinie. Zostały

opracowane, aby zapobiegać różnicom w odczuwalnej głośności dźwięku pomiędzy poziomem reklam a właściwym materiałem czy też głośnością kolejnych po sobie odtwarzanych filmów i utworów muzycznych. Utwory muzyczne cechują się również dynamiką. Dźwięku o różnej głośności wpływają na dynamikę utworu muzycznego, w notacji muzycznej głośność określona jest za pomocą znaków i transkrypcji w zapisie nutowym, są to m. in. oznaczenia forte (głośno), piano (cicho), crescendo (stopniowy wzrost dynamiki dźwięków). Zmiany głośności mają bardzo duży wpływ na odbiór przez słuchaczy dzieła muzycznego. Wcześniej wspomniany element opisu muzyki – agogika – odnosi się do regulowania tempa utworu muzycznego. Tempo w utworze muzycznym może być stałe (tempo wolne – adagio, umiarkowane – moderato, żywe – vivo) oraz tempo zmienne [43, 90].

2.5. Deskryptory sygnału wideo

Standard MPEG-7 [103] formułuje zapisy dotyczące metadanych sygnałów wizyjno-fonicznych, opisuje cechy sygnałów cyfrowych i zawartych w nich treści. Opis sygnałów wideo wykorzystywany jest wyszukiwania określonych treści czy charakterystyki materiałów multimedialnych.

Deskryptory, zawarte w standardzie MPEG-7, odnoszące się do cech wideo, można podzielić na kilka głównych grup [103]:

- Struktury podstawowe:
  - rozmieszczenie siatki (ang. *Grid Layout*) – kompozycja deskryptorów wizualnych obliczanych w siatce danego obrazu,
  - widok złożony w 2D-3D (2D-3D *Multiple View*) – deskryptory obliczane dla obrazu 2D i 3D tej samej sceny.
- Deskryptory koloru:
  - dominujące koloru (ang. *Dominant Color*) – informacja o dominujących kolorach ekstrahowanych z obiektu wizualnego (ramka, region ramki, sekwencja ramek, obszar czasowy),
  - rozmieszczenie koloru (ang. *Color Layout*) – informacja o współczynnikach DCT dla składowych luminancji i chrominancji,
  - struktura koloru – informacja o strukturalnym histogramie kolorów HMMD (ang. *Hue, Min, Max, Difference*).
- Deskryptory tekstury:
  - ruch kamery (ang. *Camera Motion*) – informacja o parametrach ruchomej kamery,
  - tor ruchu (ang. *Motion Trajectory*) – informacja o trajektorii ruchomego obiektu
  - aktywność ruchu (ang. *Motion Activity*) – informacja o aktywności ruchu obiektu mierzona w blokach o wymiarach 16x16.

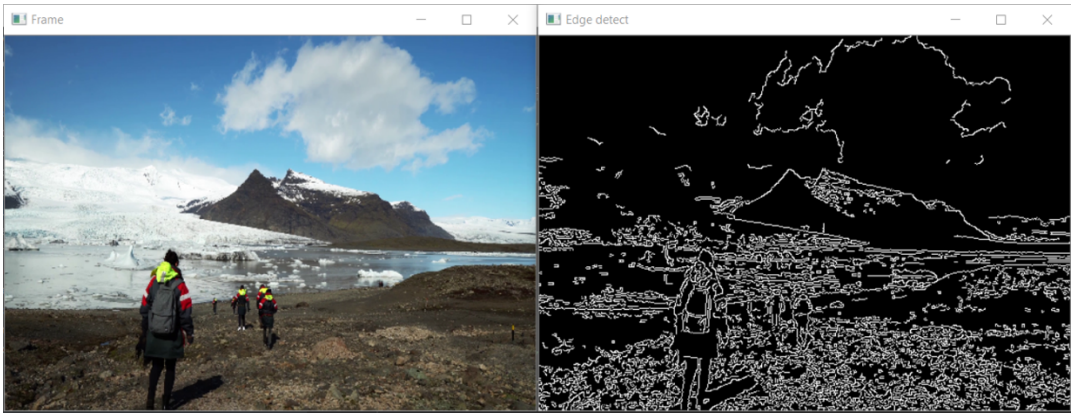
W skład opisu deskryptorów opisanych w MPEG-7 wchodzą również deskryptory tekstury [103]:

- HTD (ang. *Homogeneus Texture Descriptor*) – opisuje jednorodne tekstury wizyjne,
- TBD (ang. *Texture Browsing Descriptor*) – opisuje percepcyjne właściwości tekstur,
- EHD (ang. *Edge Histogram Descriptor*) – opisuje rozkład oraz kierunkowość tekstur pomiędzy regionami o różnicy tekstur w obrazie.

Istnieje również szereg narzędzi do ekstrakcji cech obrazu wideo, jak np. biblioteka OpenCV [40]. Biblioteka open CV jest biblioteką w pełni otwartą i multiplatformową, co oznacza, że może być użyta na systemach operacyjnych Windows, Linux czy MacOS, jak również w systemach działających na urządzeniach mobilnych: Andorid, iOS. Oferuje ona szereg funkcji do

przetwarzania i modyfikowania zarówno obrazów statycznych jak i ruchomych. Pozwala na analizę wielu formatów obrazu ruchomego i statycznego, w tym: JPEG, PNG, MP4, AVI.

Dodatkowymi funkcjami biblioteki OpenCV jest detekcja oraz śledzenie obiektów. Detekcja obiektów opiera się na implementacji wytrenowanych głębokich sieci neuronowych bądź klasyfikatora kaskadowego Haara (ang. *Haar Cascade*) [78], które zawierają szablony dla różnych obiektów (takich jak np.: samochody, twarze, oczy itp.). Biblioteka pozwala również na śledzenie obiektów w materiałach wideo za pomocą filtrów korelacyjnych bądź algorytmu śledzenia obiektów w czasie rzeczywistym (tj. *MedianFlow Tracker)*. Dodatkowo posiada implementację metod pozwalających na usuwanie tła oraz wykrywania kształtu i konturów obiektów. Przykład detekcji krawędzi został zaprezentowany na rysunku 2.26 [40].



Rys. 2.26. Przykład detekcji krawędzi za pomocą biblioteki OpenCV [40]

Analiza sygnału wideo w odniesieniu do klasyfikacji w sieciach neuronowych obejmuje:

- parametryzację sygnału wideo przy pomocy cech i deskryptorów MPEG-7 (ang. Moving Pictures Expert Group),
- bibliotek przeznaczonych do analizy sygnału wideo (np. OpenCV),
- użycie sieci CNN do ekstrakcji cech obrazu (pojedynczych ramek fragmentu filmu),
- klasyfikację wideo przy pomocy trójwymiarowych sieci splotowych (CNN 3D),
- wygenerowanie cech sekwencyjnych przy pomocy sieci rekurencyjnych.

Oprócz wymienionych technik stosuje się również inne metody wydobywania cech z obrazu czy wideo. Zostaną one przybliżone w rozdziale 4.

2.6. Parametryzacja sygnału fonicznego

Parametryzacja sygnałów fonicznych jest intensywnie rozwijającym obszarem w kontekście automatycznego przetwarzania sygnałów, spowodowane to zastało ogromnym przyspieszeniem rozwoju sieci neuronowych i ich wykorzystania do inteligentnego przetwarzania muzyki. W przetwarzaniu sygnałów audio obecnie stosuje się dwa podejścia [68]:

- Parametryzacja, mająca za zadanie ekstrakcję najbardziej istotnych cech sygnału;
- reprezentacja 2D (spektrogramy, chromagramy, mel-cepstrogramy, itp.).

2.6.1. Reprezentacja 2D sygnału muzycznego

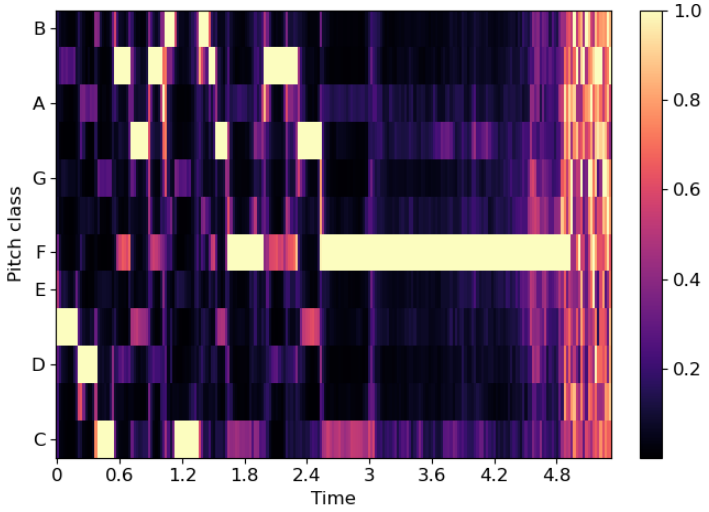
Przekształcenie sygnału dźwiękowego z funkcji czasu do funkcji częstotliwości pozwala na analizę audio w dziedzinie częstotliwości, reprezentacja ta nazwana jest widmem (ang. spectrum). W tym celu stosowana jest transformacja Fouriera, w przypadku sygnałów cyfrowych transformacja ta nazwana jest dyskretnym przekształceniem Fouriera i służy do tego algorytm

FFT – szybkiego przekształcenia Fouriera (ang. *Fast Fourier Transform*). Do analizy sygnałów muzycznych, które posiadają różne składowe częstotliwościowe używana jest transformata STFT – krótkookresowe przekształcenie Fouriera (ang. *Short-Time Fourier Transform*). STFT polega na podziale sygnału na okna czasowe, a następnie na obliczeniu dla każdego okna transformaty Fouriera. Wynikowo otrzymywana jest dwuwymiarowa macierz, gdzie oś pozioma reprezentuje czas, a oś pionowa częstotliwości. Funkcje okna stosuje się przed obliczeniem transformacji Fouriera. W praktyce stosuje się różne okna, zazwyczaj do analizy sygnałów muzycznych wykorzystywane są okna Hanninga bądź Hamminga. Wynik STFT można przedstawić w formie dwuwymiarowej bądź trójwymiarowej, częstym rozwiązaniem w sieciach neuronowych jest wykorzystanie reprezentacji 2D sygnałów audio w formie spektogramu [19, 68].

Reprezentacje te przedstawiają amplitudę widmową w postaci barw w przestrzeni czasowo-częstotliwościowej. Istnieją różne skale dla osi częstotliwości w których spektrogramy mogą być wykreślane [105]:

- Liniowa – jest najprostszą i najbardziej intuicyjną skalą, w tym przypadku wartości częstotliwości są wykreślone równomiernie na osi. Posiada najdokładniejsze odwzorowanie wartości częstotliwości.
- Logarytmiczna – skala ta używa logartym z wartości częstotliwości, umożliwia to odzwierciedlenie ludzkiej percepcji dźwięków.
- Melowa – używana jest w dziedzinie percepcji dźwięków, jest często używana w analizie sygnałów muzycznych. Skala melowa reprezentuje równomierne odległości pomiędzy dźwiękami w odniesieniu do perceptualnej skali wysokości.
- Oktawowa – w skali tej częstotliwości rozmieszczone są w równych odstępach proporcjonalnych do ich potęgi dwójki. W muzyce oddalone od siebie dźwięki o jedną oktawę traktowane są jako wariacje tego samego dźwięku, ale o innej częstotliwości.

W reprezentacji graficznej sygnałów muzycznych często stosowane są również chomagramy. Chomagram jest przedstawieniem akordów w muzyce na podstawie analizy widma częstotliwościowego sygnału audio. Do stworzenia chromagramu używa się skali równomiernie temperowanej, która pozwali podzielić oktawę na 12 półtonów (klasy chromatyczne częstotliwości). Każda komórka w chromagramie reprezentuje intensywność danego dźwięku w danej klasie częstotliwości. Chromagramy pozwalają na określenie, jakie dźwięki lub akordy obecne są w danym utworze, przykład chromagramu został zaprezentowany na rys. 2.27 [115].

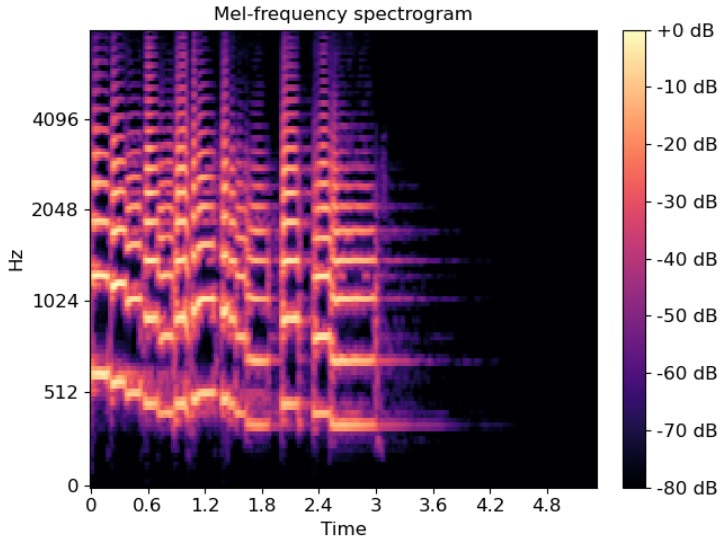


Rys. 2.27. Przykład chromagramu analizy sygnału audio [115]

Najczęściej jednak analiza dźwięków muzycznych wykorzystuje reprezentację 2d w postaci mel-spektogramów. Reprezentacja ta łączy ze sobą dwie analizy: obliczenie energii widma oraz skalę melową. Proces obliczania widma w skali melowej jest następujący [116]:

1) Obliczenie transformacji Fouriera (FFT) z sygnału audio.
2) Przekształcenie widma częstotliwościowego za pomocą filtrów melowych. Filtry projektowane są tak, aby miały kształt trójkątny, a ich wierzchołki były równomiernie rozłożone na skali melowej.
3) Obliczenie energii w każdym paśmie melowym poprzez zsumowanie amplitud w każdym paśmie melowym.

Wynikiem mel-spektogramu jest dwuwymiarowa macierz, gdzie na osi poziomej jest reprezentowany czas, a na osi pionowej różne pasma melowe. Intensywność koloru w każdej komórce macierzy odpowiada za ilość obliczonej energii w danym paśmie skali melowej i w danym czasie, przykład spektrogramu w skali melowej został zaprezentowany na rys. 2.28 [116].



Rys. 2.28. Przykład spektrogramu melowego analizy sygnału audio [63]

2.6.2. Reprezentacja parametryczna sygnału muzycznego

Jedną z głównych reprezentacji sygnałów audio jest wektor parametrów mel-cepstralnych (ang. *Mel-Frequency Cepstral Coefficient*, MFCC). Proces ekstrakcji parametrów MFCC obejmuje kilka kroków:

1) Podział sygnału na ramki. Sygnał dzielony jest na krótkookresowe nachodzące na siebie ramki czasowe.
2) Dla każdej ramki obliczana jest transformacja Fouriera (FFT).
3) Przekształca się widmo częstotliwościowe na skalę melową, co pozwala na uzyskanie mel-spektogramu.
4) Oblicza się logarytm z energii mel-spektogramu.
5) Oblicza się ponownie transformację Fouriera, wracając w dziedzinę czasu, uzyskując w ten sposób w cepstrum, a następnie z logarytmicznego spektogramu uzyskuje się współczynniki cepstralne.
6) Wybiera się najlepsze współczynników, aby uzyskać optymalny wektor cech MFCC.

Parametry MFCC mają tę zaletę, że wydobywają istotne elementy dźwięku, np.: pozwalają różnicować cechy tonalne, różnice wysokości dźwięków, struktury harmoniczne, itp. Dodatkowo

pozwalają zredukować wymiar danych, dzięki temu można zachować zestaw istotnych wartości służących do analizy, jednocześnie zmniejszając ich liczbę [52, 54].

Jednym z głównych standardów opisów parametrycznych plików audio jest czwarta część wspomnianego wcześniej standardu MPEG-7 ISO/IEC 15938-4:2002 [33]. Standard ten zawiera 17 deskryptorów niskiego poziomu (ang. *Low Level Descriptor*, LLD), które zostały podzielone na 6 grup, przedstawiono je w tab. 2.1 [33]:

- Basic – wykorzystywane do wizualizacji sygnału, opisują przebieg sygnału,
- BasicSpectral – opisują podstawowe własności widma,
- SpectralBasis – zawierają informację o dynamice widmowej sygnału,
- SignalParameters – zawierają istotne informację dla sygnałów prawie-okresowych lub okresowych,
- TimbralTemporal – grupa parametrów opisujących barwę dźwięku w sygnale,
- TimbralSpectral – grupa parametrów opisujących barwę dźwięku w dziedzinie częstotliwości

Tab. 2.1. Deskryptory standardu MPEG-7 dla sygnału audio [33]

Grupa deskryptorów	Deskryptory niskiego poziomu
Basic	• AudioWaveform (AW) – określa minimalne i maksymalne wartości próbek w przebiegu czasowym • AudioPower (AP) – uśredniona moc sygnału w czasie
BasicSpectral	• AudioSpectrumEnvelope (ASE) – krótkookresowy opis energii widma w skali logarytmicznej. Graficzna reprezentacja tego parametru to spektogram • AudioSpectrumCentroid (ASC) – określa środek ciężkości widma. Opisuje widmo sygnału w funkcji częstotliwości. • AudioSpectrumSpread (ASS) – wariancja energii widma sygnału względem jego środka ciężkości wyznaczonego przez ASC • AudioSpectrumFlatness (ASF) – określa różnice pomiędzy kształtem sygnału widma w danym paśmie od idealnej charakterystyki
SpectralBasis	• AudioSpectrumBasis (ASB) – opisuje funkcje bazowe widma • AudioSpectrumProjection (ASP) – opisuje funkcje przekształcające widma
SignalParameters	• AudioHarmonicity (AH) – określa stopień harmoniczności sygnału • AudioFundamentalFrequency (AFF) – zawiera informację o częstotliwości podstawowej sygnału
TimbralTemporal	• LogAttackTime (LAT) – opisuje czas narastania obwiedni sygnału, podawany w skali logarytmicznej • TemporalCentroid (TC) – zawiera informacje o środku ciężkości badanego sygnału
TimbralSpectral	• SpectralCentroid (SC) – określa środek ciężkości widma mocy sygnału poprzez wartość średniej częstotliwości • HarmonicSpectralCentroid (HSC) – określa środek ciężkości składowych harmonicznych • HarmonicSpectral-Spread (HSS) – określa odchylenie harmonicznych względem środka ciężkości • HarmonicSpectralVariation (HSV) – określa współczynnik korelacji harmonicznych z pary ramek sygnału • HarmonicSpectralDeviation (HSD) – określa odchylenie wartości amplitud harmonicznych od ich wartości średniej

W obliczeniach parametrów wykorzystuje się bibliotekę OpenSMILE, która obecnie jest powszechnie używana do ekstrakcji cech parametrycznych dźwięku, w kontekście analizy emocji i przetwarzania mowy. Biblioteka OpenSMILE jest narzędziem dostępnym na wielu platformach, w celach badawczych może być wykorzystywana z darmową licencją. Biblioteka posiada wiele mechanizmów ekstrakcji parametrów, pozwala między innymi na: analizę emocji, rozpoznawanie mówcy, wykrywanie mowy, analizę muzyczną czy też tworzenie aplikacji interaktywnych. Główne deskryptory, które biblioteka udostępnia, są to [80]:

- Parametry MFCC,
- ZeroCrossing Rate (ZRC) – określa liczbę przejść sygnału przez zero w ciągu jednej sekundy,
- Chroma – opisuje parametry tonalności i struktury harmonicznej dźwięku,

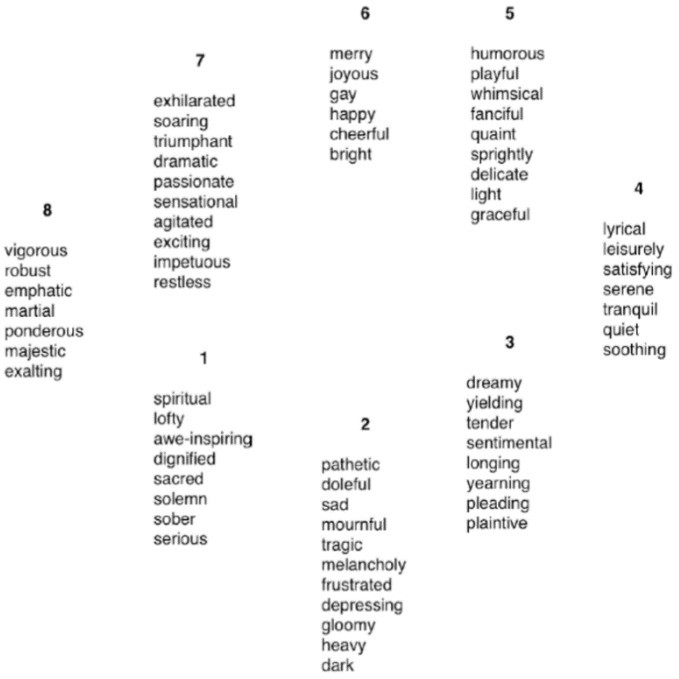
- Pitch – określa wysokość dźwięku,
- Energia – określa amplitudę sygnału,
- Intensywność formantów – format to pasmo częstotliwości, w granicach którego wszystkie tony ulegają wzmocnieniu, odpowiada za barwę dźwięku. Może dostarczyć informacji o mówcy i jego charakterystyce głosu,
- Współczynniki LPC (ang. *Linear Predicitve Coding*) – współczynniki LPC są wagami modelu liniowego, który odwzorowuje rzeczywisty sygnał.

3. Modele emocji

W literaturze obecnie można zauważyć dwa trendy skupiające się na klasyfikacji emocji:

- podejście kategoryczne – emocje opisywane są za pomocą przymiotników,
- podejście tzw. wymiarowe – emocje klasyfikowane są na podstawie ich położenia w przestrzeni dwuwymiarowej,

Istnieje wiele modeli emocji dostępnych w psychologii, ale też powiązanych z dźwiękiem czy obrazem [88, 89, 111]. Jednym z przykładów kategorycznych, często przytaczanych w literaturze, jest model Kate Hevner, który zawiera w sumie 67 przymiotników podzielonych na osiem różnych grup. Model ten został zaprezentowany na rys. 3.1 [34].



Rys. 3.1. Model emocji Hevner [34]

Przykładami wykorzystywanych modeli emocji jest model Thayera [113] oraz model Schuberta, który − na podstawie dwuwymiarowego modelu Thayera − został przemapowany na 46 przymiotników podzielonych na 9 grup [100]. Zostały one przedstawione w tabeli 3.1 (pozostawiono nazwy w języku angielskim ze względu na możliwe wystąpienie niejednoznaczności w tłumaczeniu na j. polski) [100].

Tab. 3.1. Model emocji zaproponowany przez Schuberta [100]

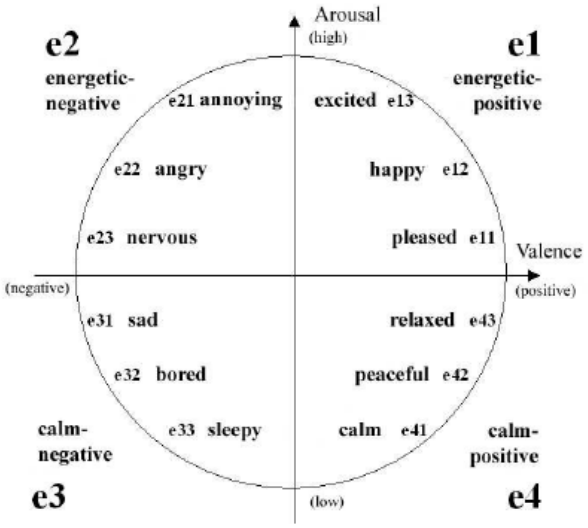
Numer grupy	Emocje zawarte w grupie
1	*Bright, cheerful, happy, joyous*
2	*Humorous, light, lyrical, merry, playful*
3	*Calm, delicate, graceful, quiet, relaxed, serene, soothing, tender, tranquil*
4	*Dreamy, sentimental*
5	*Dark, depressing, gloomy, melanchoky, mournful, sad, solemn*
6	*Heavy, majestic, scared, serious, spiritual, vigorous*
7	*Tragic, yearning*
8	*Agitated, angry, restless, tense*
9	*Dramatic, exciting, exhilarated, passionate, sensational, soaring, triumphant*

Kolejnym przykładem jest podział na grupy, zaproponowany podczas międzynarodowego konkursu *Music Information Retrieval Evaluation eXchange*. Model ten składa się z 29 przymiotników podzielonych na pięć grup. Zostały one przedstawione w tabeli 3.2 [36].

Tab. 3.2. Model emocji przedstawiony podczas konkursu MIR *Evaluation eXchange* [36]

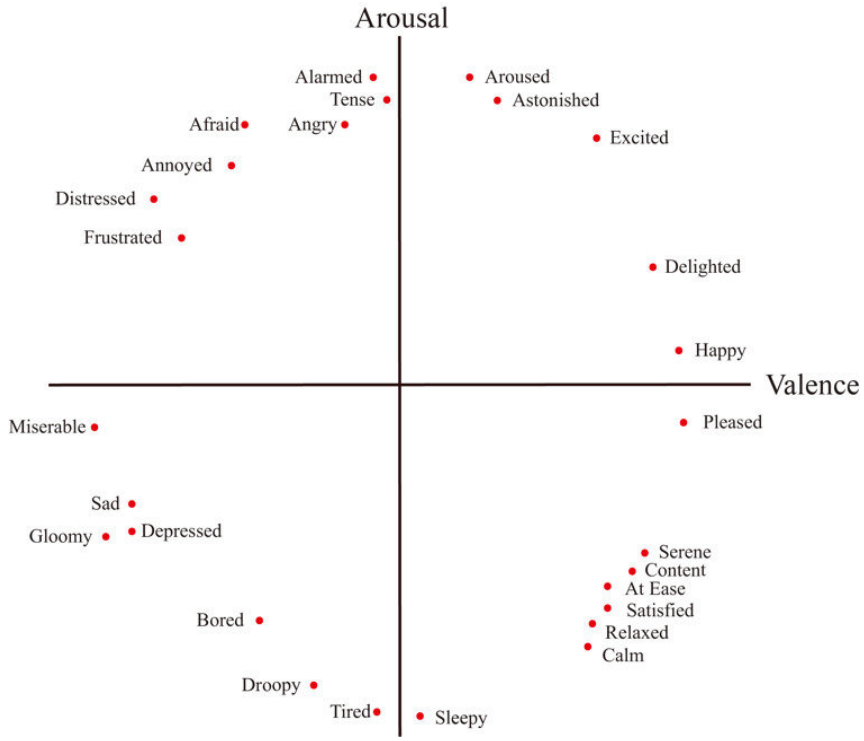
Numer grupy	Emocje zawarte w grupie
1	*Passionate, rousing, boisterous, rowdy*
2	*Rollcking, cheerful, fun, sweet, amible*
3	*Literate, poignant, wistful, bittersweet, autumnal, brooding*
4	*Humorous, silly, campy, quirky, whimsical, witty, wry*
5	*Aggressive, fiery, tense, intense, volatile, visceral*

W odniesieniu do modeli wymiarowych w literaturze można wymienić kilka podstawowych modeli. Jednym z nich jest model bazujący na uproszczonej wersji modelu dwuwymiarowego, w której główną rolę odgrywa napięcie i energia. Rozłożone są one na dwóch prostopadłych do siebie osiach. Napięcie jest zmienną względną – od emocji pozytywnych do negatywnych, natomiast energia opisuje emocje spokojne i bardziej wzniosłe. Obecnie stosuje się bardziej rozbudowaną wersję modelu Thayera, która w czterech obszarach powstałych przez podział osi zawiera trzy emocje na każdym obszarze. Zostało to przedstawione na rys. 3.2 [113].



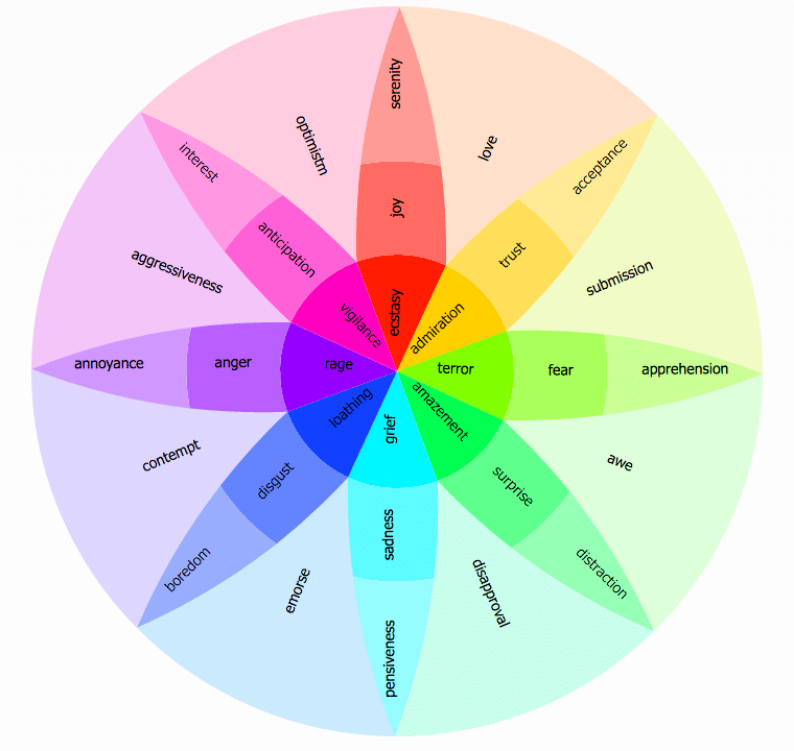
Rys. 3.2. Rozbudowany model emocji Thayera [113]

Kolejną odmianą podstawowego modelu emocji Thayera jest model Russela, który zachował pierwotny podział na emocje. Russel swój model przedstawił jako odwzorowanie 28 emocji zaznaczonych na wykresie walencji (ang. *valence*; wartościowość, znak emocji) i pobudzenia (ang. *arousal*), przedstawione na rys. 3.3 [95].



Rys. 3.3. Model emocji Russela [95]

Jednym z najbardziej rozbudowanych modeli wymiarowych jest model Plutchika, model ten zawiera podział na osiem głównych emocji, które są podstawą dla pozostałych. Są one rozłożone przeciwstawnie, tj.: radość i smutek, akceptacja i wstręt, strach i złość, zaskoczenie i przewidywanie. Przedstawione zostało to na rys. 3.4 [91].



Rys. 3.4. Model emocji Plutchika [91]

W pracach badawczych z zakresu rozpoznawania i klasyfikacji emocji w treściach audiowizualnych wykorzystywane są powyżej wspomniane modele emocji. Modele te dostarczają informacji z dziedziny psychologii do budowy systemów opartych o uczenie maszynowe i lepsze zrozumienie aspektów emocjonalnych w analizowanych treściach, oferują one również różne perspektywy podziału na kategorię emocji. Najczęstszym rozwiązaniem stosowanym w literaturze odnoszącej się do badań nad klasyfikacją emocji z danych audio [20, 25, 94, 98] jest wykorzystanie modeli emocji Thayera [113] bądź Russella [95], gdzie rozwiązania klasyfikacji emocji oparto na predykcji wartości *valence-arousal*. Spotykanym rozwiązaniem jest również wykorzystanie klasyfikacji emocji na podstawie etykiet, które zostały przygotowane w oparciu o kolorowy model Plutchika [91], rozwiązanie takie zostało zaproponowane w pracy Ciborowskiego i in. [16] oraz w pracy Plewy [89]. W analizie sygnałów wideo w celach klasyfikacji emocji wykorzystuje się predykcję wartości VA. W pracach Aslana i in. [3] oraz Hayata i in. [30] powołano się między innymi na model Russela [95].

W niniejszej rozprawie doktorskiej wybrano podejście kategoryczne. Na podstawie propozycji zawartych w literaturze, dotyczącej psychologii kolorów w filmie w połączeniu z kolorowym modelem Plutchika [91], zaproponowano nowy model emocji, który następnie zastosowano do przygotowania etykiet w zbiorze danych użytym do uczenia algorytmów głębokich sieci neuronowych.

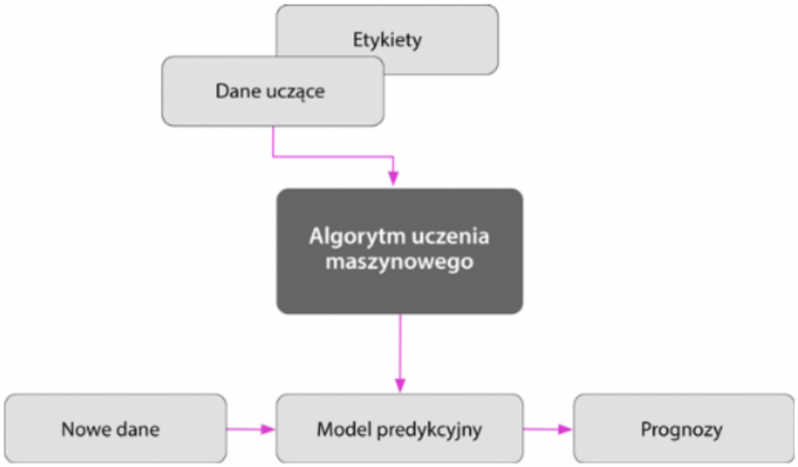
4. Wybrane metody uczenia maszynowego stosowane w przetwarzaniu sygnałów

4.1. Metody uczenia maszynowego – wybrane zagadnienia

Artur Samuel w 1959 roku zdefiniował uczenie maszynowe w następujący sposób: „Uczenie maszynowe jest to dziedzina nauki dająca komputerom możliwość uczenia się bez konieczności ich jawnego programowania". Natomiast bardziej dokładną definicję podał w 1997 r. Tom Mittchel: „Mówimy, że program komputerowy uczy się na podstawie doświadczenia E, na podstawie zadania T i pewnej miary wydajności P, jeśli miara wydajności P wobec zadania T rośnie wraz z nabywanym doświadczeniem E". Wszystko jednak można uprościć do programowania komputerów do wykonywania zadania uczenia się danych [27].

W literaturze rozróżnia się podział uczenia maszynowego na: uczenie nadzorowane, uczenie nienadzorowane oraz uczenie ze wzmocnieniem [93].

Uczenie nadzorowane, w którym skupiono się w niniejszej pracy, jest uczeniem, gdzie dane treningowe przekazywane algorytmowi posiadają tak zwane etykiety (ang. *labels*). Na rys. 4.1 widać cykl pracy uczenia nadzorowanego, w którym specjalnie oznakowane dane przekazywane są do modelu predykcji, który przewiduje wyniki na nowych, nieoznakowanych danych wejściowych [93].

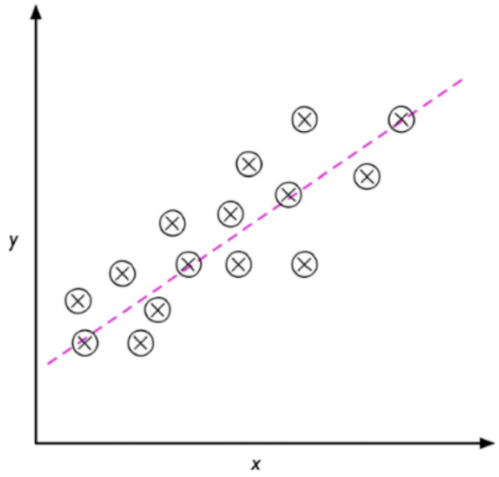


Rys. 4.1. Cykl uczenia nadzorowanego [93]

Uczenie nadzorowane można podzielić na zadania klasyfikacji i regresji. W rozprawie doktorskiej do rozwiązania problemu wybrano algorytmy klasyfikacji. Klasyfikacja pozwala na przewidywanie etykiet klas nowych danych na podstawie wcześniejszych obserwacji. Etykiety klas są wartościami dyskretnymi, które określają przynależność danych do wyznaczonych grup. Klasyfikację można podzielić na klasyfikację binarną oraz wieloklasową. Binarna występuje, gdy algorytm uczy się reguł pozwalających na rozróżnienie dwóch możliwych klas. Przykładem prostej klasyfikacji binarnej może być filtr spamu, gdzie algorytm klasyfikuje wiadomości e-mail na wartościowe wiadomości (klasa pozytywna) oraz na spam (klasa negatywna). Model uczenia maszynowego może również przyjmować dowolną liczbę etykiet dla nowych nieoznakowanych przykładów. Jednym z klasycznych przykładów klasyfikacji wieloklasowej jest rozróżnianie cyfr ręcznie pisanych [93].

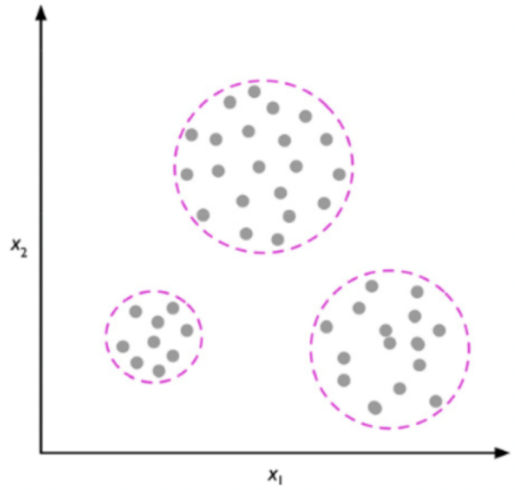
Drugim rodzajem uczenia nadzorowanego jest regresja, czyli przewidywanie wyników ciągłych (rys. 4.2). Model bazuje na cechach, czyli zmiennych objaśniających (prognozujących) i jego zadaniem jest predykcja wartości zmiennej celu, czyli zmiennej objaśnianej (prognozowanej). Na rys. 4.2 zaprezentowano koncepcję regresji liniowej, gdzie dla przykładowych danych wyznaczana jest prosta, w jak najmniejszej odległości od punktów danych.

Uzyskiwana jest ona zazwyczaj metodą najmniejszych kwadratów. W celu predykcji zmiennych celu na podstawie zadanych cech wykorzystuje się w tym przypadku punkt przecięcia prostej z osią współrzędnych oraz jej nachylenie [93].



Rys. 4.2. Przykład regresji liniowej [93]

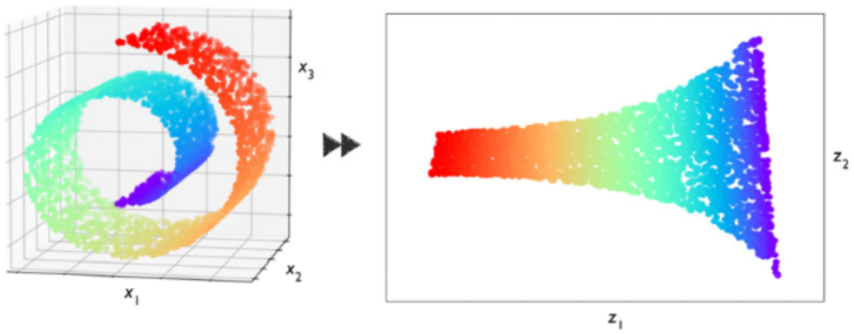
W uczeniu nienadzorowanym (ang. *unsupervised learning*) dane wynikowe mogą być nieoznaczone bądź o nieznanej strukturze. Jedną z metod uczenia nienadzorowanego jest grupowanie (klasteryzacja), pozwala ona na organizację danych na podzbiory (klastry) bez wcześniejszej wiedzy dotyczącej podziału grupowego danych. Klastry grupowane są na zasadzie wspólnych podobieństw wynikających z analizy danych, pozwala to na strukturyzację danych oraz wyznaczanie powiązań między nimi (rys. 4.3) [93].



Rys. 4.3. Grupowanie danych na podstawie cech x1 i x2 [93]

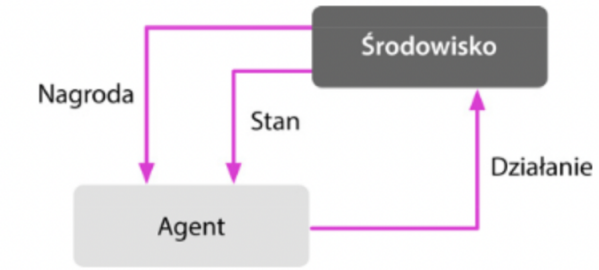
Poza klasteryzacją uczenie nienadzorowane stosowane jest jako narzędzie pozwalające na redukcję wymiarowości. Przykład redukcji wymiarów został zaprezentowany na rys. 4.4. Dane wielowymiarowe wymagają dużej ilości zasobów zarówno w kontekście pojemności pamięci, jak i mocy obliczeniowej. Redukcja wymiarów stosowana jest często przy wstępnym przetwarzaniu danych w celu wyeliminowania szumów danych i pozwala zwiększyć skuteczność algorytmów uczenia maszynowego. W efekcie otrzymywana jest podprzestrzeń o mniejszej liczbie wymiarów bez utraty istotnej części informacji. Drugim przykładem wykorzystania redukcji wymiarów jest wizualizacja danych, gdzie zbiory cech wielowymiarowych można rzutować na dogodne formy

wykresów punktowych bądź histogramów, przykład redukcji wymiarów przedstawiono na rys. 4.4. [93].



Rys. 4.4. Redukcja wymiarów na dogodne formy [93]

Trzecim rodzajem uczenia maszynowego jest uczenie przez wzmacnianie (ang. *reinforcement learning*). System (agent) poprawia swoją własność predykcji na podstawie interakcji ze środowiskiem, poprawę skuteczności działania poprzez funkcję nagrody. Rys. 4.5 przedstawia metodologię działania algorytmu w uczeniu nienadzorowanym [93]



Rys. 4.5. Uczenie nienadzorowane [93]

Agent zawsze próbuje zmaksymalizować nagrodę poprzez szereg działań związanych ze środowiskiem, każdy obecny stan agenta można powiązać z ujemną bądź dodatnią nagrodą [93].

4.1.1. Metody uczenia głębokiego

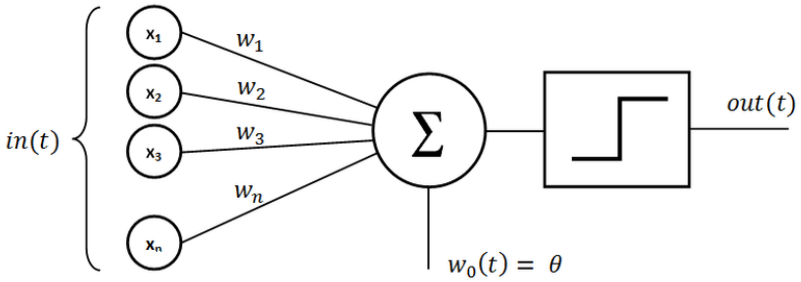
Sieci neuronowe zyskały obecnie dużą popularność w rozwiązywaniu problemów nieliniowych, których cechą jest duża ilość danych. Podstawowa koncepcja sieci neuronowej bazuje na modelach, które odnoszą się do funkcjonowania ludzkiego mózgu. W rozprawie skupiono się na głębokich sieciach neuronowych ze względu na ich szerokie wykorzystanie w dziedzinach przetwarzania obrazu oraz dźwięku.

Model uczenia maszynowego definiowany jest jako funkcja matematyczna, która ma celu ma odwzorowanie danych wejściowych ze zbioru danych o wspólnych właściwościach na dane wyjściowe ze zbioru, który jest z nim powiązany.

W 1957 roku Frank Rosenblatt zaproponował modyfikacje sztucznego neuronu binarnego i przekształcił go w perceptron, co stało się podstawą głębokich sieci neuronowych. Była to kluczowa zmiana w koncepcji sztucznych sieci neuronowych. Zmiany zaproponowane przez Rosenblatta dotyczyły między innymi (rys. 4.6) [93, 122]:

a) Na wejściu i wyjściu pojawiły się wartości liczbowe zamiast wartości binarnych – węzły.
b) Połączenia węzłów otrzymały odpowiednią wagę.

c) Wartość wyjściowa jest sumą wartości z poprzednich warstw pomnożonych przez wagi oraz przekształcenie tej wartości w węźle przez funkcje aktywacji, której zadaniem jest wprowadzenie nieliniowości do modelu.



Rys. 4.6. Perceptron jako przykład sztucznej sieci neuronowej

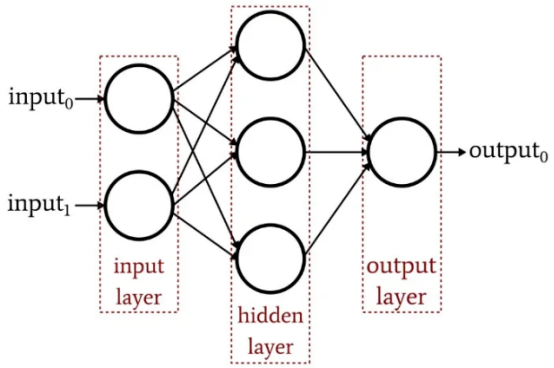
Gdzie:

- x_n – wejścia sieci neuronowej,
- w_n – wagi sieci neuronowej,
- Σ - suma ważona,
- Po sumie znajduje się funkcja aktywacji, a następnie wyjście sieci neuronowej.

W sieci neuronowej liczba neuronów może się różnic między warstwami, stała zostaje jednak funkcja aktywacji dla wszystkich neuronów w danej warstwie. Sieć neuronowa składa się z trzech warstw:

- Warstwa wejściowa (ang. *Input Layer*) – ma za zadanie zbierania danych i przekazywania ich dalej,
- Warstwa ukryta (ang. *Hidden Layer*) – są to stany pośrednie, odpowiadające za znalezienie nieliniowych zależności pomiędzy danymi,
- Warstwa wyjściowa (ang. *Output Layer*) – warstwa odpowiadająca za zwrócenie wyniku.

Połączenia występują asymetrycznie jedynie z warstwami sąsiadującymi według zasady „każdy z każdym”, przykład połączeń został zaprezentowany na rys. 4.7 [93, 122]. Nie istnieje jednak liczba warstw głębokiej sieci neuronowej, którą można przyjąć za ostateczną [4, 93, 122].



Rys. 4.7. Przykład sieci neuronowej składającej się z trzech warstw

Wśród głębokich sieci neuronowych można wyróżnić w szczególności trzy typy, których wybór zależy od rozpatrywanego problemu oraz posiadania danych [4]:

- Perceptrony wielowarstwowe (ang. *Multi-Layer Perceptron,* MLP),
- Splotowe sieci neuronowe (ang. *Convolutional Neural Network*, CNN),
- Rekurencyjne sieci neuronowe (ang. *Recurrent Neural Network*, RNN).

Model MLP jest klasycznym przykładem jednokierunkowej sztucznej sieci neuronowej oraz siecią o topologii pełnych połączeń (ang. *fully connected*). MLP są wykorzystywane zazwyczaj w problemach logistycznych oraz predykcji wartości liczbowych za pomocą regresji liniowej. Konstrukcja sieci MLP nie pozwala jednak na przetwarzanie znacznej ilości danych wymiarowych oraz danych sekwencyjnych [4, 93].

Sieci splotowe jest najczęściej wykorzystany w klasyfikacji obrazów. Inspiracją do stworzenia splotowych sieci neuronowych był mechanizm działania kory wzrokowej ludzkiego mózgu podczas zadania rozpoznawania przedmiotów [60]. Zaletą sieci CNN jest skuteczne wydobywanie najbardziej istotnych cech z przekazanych danych wejściowych. Sieci splotowe tworzą tak zwane hierarchię cech poprzez łączenie cech podstawowych czy też ogólnych w warstwy, dzięki temu uzyskane zostają cechy wysokopoziomowe. W zadaniu analizy obrazu sieci CNN wydobywają podstawowe cechy, którymi mogą być krawędzie przedmiotów, a następnie składane są one w bardziej szczegółowe cechy, np. kształty przedmiotów [4]. Są to tak zwane mapy cech, gdzie każdy element mapy cech pochodzi z lokalnego pola recepcyjnego (ang. *local receptive field*). Sieci splotowe bardzo dobrze radzącą sobie z problemami dotyczącymi przetwarzania obrazów ze względu na:

- rzadkość połączeń – pojedynczy element w mapie cech jest połączony z niewielkim obszarem pikseli,
- współdzielenie parametrów – te same wagi używane są do różnych obszarów obrazu.

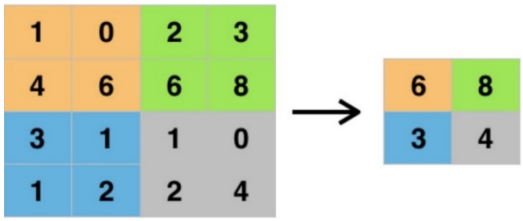
Pozwala to na redukowanie parametrów modelu, co jest istotne w procesie ekstrakcji cech i parametrów danych wejściowych.

Sieci splotowe składają się zazwyczaj z kilku warstw splotowych i warstw podpróbkowania (ang. *subsampling layers*), po których występuje minimum jedna w pełni połączenia warstwa (ang. *fully-connected layer*). Warstwy podpróbkowania nie zawierają modyfikowalnych parametrów, w przeciwieństwie do dwóch wcześniej wymienionych warstw [93].

Jedną z podstawowych operacji w sieciach CNN jest splot dyskretny. Dodatkowo stosuje się proces zwany uzupełnianie zerami (ang. *zero-padding* lub ang. *padding*), aby uniknąć problemu indeksowania w zakresie nieskończoności i otrzymywania wektora wyjściowego o nieskończonym rozmiarze. Istnieją trzy tryby uzupełniania zerami, które przydatne są w praktyce: pełny, krawędziowy oraz zerowy. Najczęściej używanym jednak trybem jest tryb krawędziowy, *padding* pozwala na uzupełnienie próbki i otrzymanie wyjścia takiego samego rozmiarowo jak wejście, generuje to co prawda dodanie sztucznych wag na krawędziach, ale ponieważ uzupełnienie następuje zerami, pozwala to zapewnić prostotę obliczeń i wydajność algorytmu [4, 93].

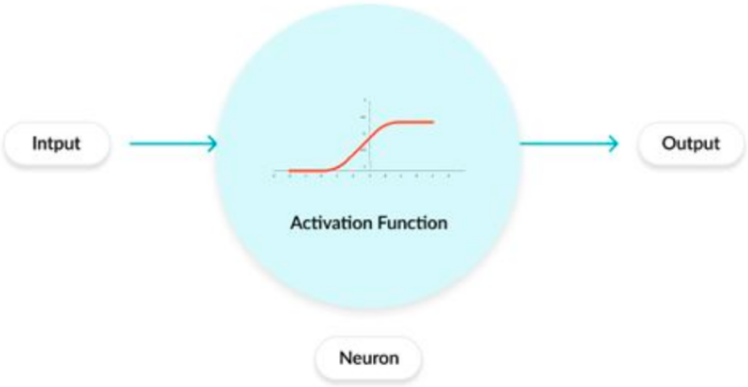
Jednym z hiperparametrów splotu jest przesunięcie, czyli tak zwany krok (ang. *stride*), definiuje on, o ile pikseli powinno zostać przesunięte jądro (inaczej okno filtra) [4].

Jedną z istotnych warstw sieci splotowej poza warstwą splotową jest warstwa łączenia, w tym celu wykorzystuje się warstwy MaxPooling2D (w przypadku sieci splotowych dwuwymiarowych, przykład techniki *MaxPooling* został zaprezentowany na rys. 4.8) lub też *AveragePooling*, jako techniki kompresji i łączenia [4].



Rys. 4.8. MaxPooling jako technika kompresji i łączenia warstw sieci splotowej [4]

Każdy neuron głębokiej sieci neuronowej posiada funkcję aktywacji, jest ona istotna zarówno dla sieci MLP, jak i CNN. Określa ona wyjście neuronu i to czy powinien on zostać aktywowany, czy też nie. Funkcje aktywacji pozwalają również znormalizować wyjście każdego neuronu do zakresu 0 do 1 lub -1 do 1. Jednym z głównych zadań funkcji aktywacji jest to, że muszą one być wydajne obliczeniowo, ponieważ są stosowane na tysiącach, a nawet milionach neuronów dla każdej próbki danych [4].



Rys. 4.9. Przykład funkcji aktywacji – funkcja sigmoidalna [4]

Współczesne modele sieci neuronowych wykorzystują nieliniowe funkcje aktywacji, aby umożliwić tworzenie złożonych mapowań pomiędzy warstwami, dzięki temu można przetwarzać w sieciach neuronowych dane audio lub wideo. Pozwalają one na transformację sumy danych wejściowych na dane wyjściowe, wykorzystywane są do generowania predykcji. Wybierając funkcję aktywacji, należy brać pod uwagę zarówno wydajność, złożoność sieci neuronowych, jak również wynik klasyfikacji. Wśród funkcji aktywacji wyróżnia się (tab. 4.1) [93]:

Tab. 4 1. Funkcje aktywacji w sieci neuronowej [93]

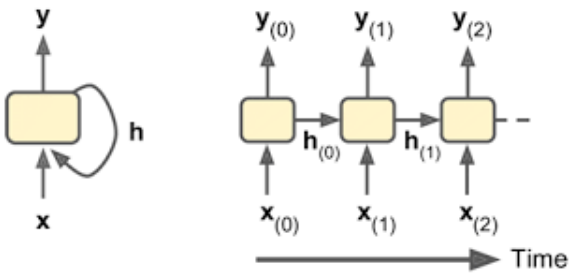
Funkcja	Wzór	Właściwości
Sigmoidalna	$\frac{1}{1+\ e^{-x}}$	• Gładka • Różniczkowalna • Wartości wyjściowe ograniczone są od 0 do 1 • Wykorzystywana głównie do klasyfikacji binarnej
ReLU	$\max(0, x)$	• Nieliniowa • Umożliwia szybką zbieżność sieci
Softmax	$\frac{e^{x}}{\sum_{i=0}^{n} e^{x_i}}$	• Wykorzystywana do obsługi wielu klas • Normalizuje wynik dla każdej klasy od 0 do 1

Dobór funkcji aktywacji warstwy wyjściowej uzależniony jest od typu rozwiązywanego problemu. W klasyfikacji binarnej będzie to funkcja sigmoidalna, natomiast w klasyfikacji

wieloklasowej będzie to softmax. Funkcja softmax powoduje znormalizowanie wyniku, prognozuje przynależności do określonych klas w środowiskach wieloklasowych. Dzięki tej funkcji aktywacji nie otrzymuje się pojedynczych indeksów klas, lecz prawdopodobieństwo przynależności do każdej z nich [93].

W przypadku funkcji prostowania jednostkowego (ang. *Rectified Linear Unit*, ReLU) pozwala uniknąć problem zanikającego gradientu. ReLU jest funkcją nieliniową. Problem zanikającego gradientu jest rozwiązany poprzez pochodną funkcji ReLU, która dla wartości dodatnich zawsze wynosi 1 [93].

Trzecim rodzajem sieci, które zostaną wykorzystane w niniejszej rozprawie, są sieci RNN, czyli rekurencyjne sieci neuronowe. Sieci RNN w przeciwieństwie do wcześniej omówionych sieci MLP i CNN posiadają funkcję pamięci sekwencyjnej. W standardowych sieciach jednokierunkowych dane przekazywane są z warstwy wejściowej do warstwy ukrytej, a następnie do warstwy wyjściowej. W sieciach rekurencyjnych warstwa ukryta otrzymuje dane zarówno z warstwy wejściowej, jak i warstwy ukrytej z poprzedniej iteracji. Schemat sieci rekurencyjnej został zaprezentowany na rys. 4.10 [4, 93].



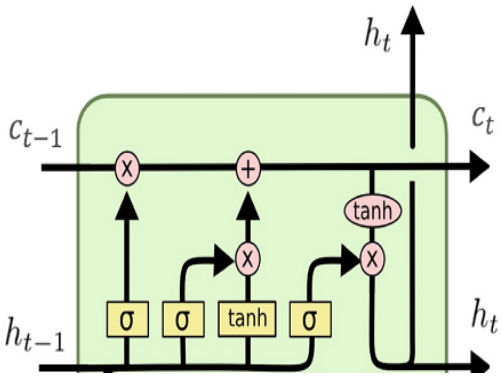
Rys. 4.10. Schemat sieci rekurencyjnej [4]

Gdzie:

- x – wektor wejściowy,
- y – wektor wyjściowy,
- h – ukryty stan wejściowy.

W zastosowaniach przetwarzania sygnałów audiowizualnych często używa się dwóch rodzajów sieci rekurencyjnych, tj. sieci LSTM (ang. *Long Short-Term Memory*), czyli tak zwanych długich pamięci krótkotrwałych oraz sieci GRU (ang. *Gated Recurrent Unit*), czyli rekurencyjnych jednostek bramkujących. Sieci LSTM rozwiązują problem znikających/eksplodujących gradientów, dzięki zastosowaniu bramek. W komórkach LSTM występują trzy rodzaje bramek, zaprezentowane na rys. 4.11 [4]:

- Bramka zapominająca (ang. *forget gate*) – umożliwia zerowanie komórki pamięci;
- Bramka wejściowa (ang. *input gate*) – odpowiedzialna jest ona za zaktualizowanie stanu komórki;
- Bramka wyjściowa (ang. *output gate*) – określa sposób aktualizacji wag.



Rys. 4.11. Schemat sieci LSTM [4]

Gdzie:

- x – wektor wejściowy,
- h – stan ukryty krótkotrwały,
- c – stan ukryty długotrwały,
- σ - kontrolery bramkowe: bramka „zapomnij”, bramka wejściowa, bramka wyjściowa.

Sieć GRU jest uproszczonym modelem architektury LSTM, który uzyskuje się poprzez ograniczenie mechanizmu bramkowania. Posiada jedynie dwie bramki, aktualizacji i resetowania, nie posiada zapamiętywania stanu komórki. Model bardzo dobrze radzi sobie z problemami modelowania sekwencyjnego. Bramka aktualizacji decyduje o ilości informacji, z aktualnego stanu komórki, które powinny zostać uwzględnione w następnym stanie komórki sieci. Wartość wyznaczane są z przedziału od 0 do 1, gdzie wartość 0 oznacza, że cała informacja zostanie zapomniana, zaś wartość 1 informuje o zapamiętaniu całej informacji komórki. Bramka aktualizacji pozwala na zapamiętanie, ile informacji jest przechowywanych z poprzedniego stanu komórki, dzięki temu sieci GRU są w stanie przechowywać informacje dotyczące dłuższych okresów czasowych dotyczących konkretnych danych. Bramka resetowania odpowiada za informacje, które powinny być zapomniane bądź zachowane z poprzedniego stanu komórki. Bramka ta również zawiera wartości z przedziału od 0 do 1. Zarówno wartości bramki aktualizacji, jak i resetowania obliczana jest na podstawie danych wejściowych, stanu ukrytego komórki z poprzedniego kroku czasowego oraz wag i funkcji aktywacji warstwy [4].

W modelach klasyfikacji wieloklasowej istotnym elementem jest również funkcja straty, która służy do oceny błędu, czyli rozbieżności modelu pomiędzy wynikami prawdziwymi, a przewidywanymi. Używana jest ona do minimalizacji tej różnicy w procesie uczenia się modelu. Najpopularniejszą funkcją straty używaną podczas klasyfikacji wieloklasowej jest funkcja kategorycznej entropii krzyżowej (ang. *categorical cross-entropy*). Zdefiniowana jest ona jako średnia suma policzonych strat dla każdej klasy. Dla każdej klasy porównuje ona prawdziwe etykiety klas z etykietami przewidzianymi, wyrażonymi jako rozkład prawdopodobieństwa. Mniejsza wartość funkcji straty definiuje większą zgodność przewidywania z rzeczywistą etykietą. Proces uczenia modelu bazuje na minimalizowaniu funkcji straty z wykorzystaniem metod optymalizacji [4].

4.1.2. Optymalizacja modelu

Najczęściej stosowanymi metodami optymalizacji modeli sieci neuronowych są [4, 93]:

- Algorytm spadku gradientu (ang. *Gradient Descent*) – polega na iteracyjnym aktualizowaniu wag sieci w przeciwnym kierunku do gradientu funkcji straty. Często

stosowanym wariantem jest spadek gradientu ze stałym krokiem (ang. *Stochastic Gradient Descent,* SGD).

- RMSprop (ang. *Root Mean Square Propagation)* – algorytm adaptujący, dostosowuje kroki uczenia dla każdego parametru sieci neuronowej na podstawie estymacji średniej kumulowanej wartości kwadratów gradientu. Pozwala to zmniejszyć krok uczenia dla parametrów, które mają duże gradienty, jednocześnie zwiększając krok uczenia dla parametrów z małym gradientem, dzięki temu przyspiesza to uczenie.
- Adam (ang. *Adaptive Moment Estimation)* – jest algorytmem adaptacyjnym, który łączy zalety algorytmów RMSprop oraz SGD. Pozwala na adaptację kroków uczenia dla każdego parametru indywidualnie oraz jest odporny na skalowanie gradientów. Ogromną zaletą tego algorytmu jest osiąganie dobrych wyników optymalizacji algorytmów i szybkie zbieżność.

Podczas uczenia sieci neuronowych mogą wystąpić dwie skrajne i niepożądane zdarzenia: niedouczenie modelu (ang. *underfitting*) oraz przetrenowanie modelu (ang. *overfitting*). Niedouczenie występuje, gdy model nie jest w stanie się dopasować zarówno do danych treningowych, jak i do nowych danych. Problem z generalizowaniem nowych danych może być spowodowany małą liczbą warstw bądź parametrów sieci neuronowej. Problem może również leżeć po stronie dostarczonych do modelu danych, które mogą być zbyt podobne, przez co model nie jest w stanie modelować nieliniowych związków pomiędzy danymi bądź ignorować ważniejszych cech danych. W kontekście niedouczania stosuje się zwiększenie architektury modelu o kolejne warstwy bądź zwiększenie ich złożoności, odpowiednie dostosowanie hiperparametrów, zwiększenie liczby epok bądź ilości danych do treningu [4].

O przeuczeniu modelu można mówić, gdy model dobrze dostosowuje się do danych treningowych, natomiast słabo generalizuje nowe dane. Powodem przetrenowania może być również zbytnie dopasowanie do szumów i odstających danych ze zbioru. Aby zapobiec nadmiernemu podopasowywaniu się modelu do danych treningowych, zalecane jest użycie technik regularyzacji modelu [4, 93]:

- Regularyzacja L1 – regularyzacja Lasso (ang. *Least Absolute Shrinkage and Selection Operator*) wprowadza karę proporcjonalną do wartości bezwzględnej wag. Stosowanie regularyzacji L1 powoduje selekcję istotnych cech, dzięki doborowi odpowiednich wag oraz selekcji nieprzydatnych cech i zerowaniu ich wag.
- Regularyzacja L2 – wprowadza karę proporcjonalną do kwadratu wartości wag. Nie zmienia wartości wag na zero, tylko zmniejsza różnice pomiędzy wagami, sprzyja tworzeniu mniejszych wag.

Regularyzacje L1 I L2 stosuje się w zależności od problemu, L1 stosowana jest w przypadku znalezienia istotnych cech, natomiast L2 w momencie zmniejszenia złożoności modelu [93].

- Warstwa pomijania (ang. *Dropout Layer*) – określając w warstwie regularyzacji współczynnik pomijania (ang. *dropout*) losowo zerowane są współczynniki danej warstwy, co powoduje, że w treningu następnej zostaną one wyłączone i zmniejszona zostaje liczba parametrów. Współczynnik dropout równy 0,4 powoduje wyzerowanie 40% wag warstwy, do której został zaadaptowany.
- W ramach niwelacji przeuczenia się modelu zostaje również możliwość zwiększenia liczby danych w zbiorze uczącym.
- Można również zadbać o odpowiednie hiperparametry modelu przy zastosowaniu różnych technik. Dwoma podstawowymi algorytmami tego typu są: przeszukiwanie siatki (ang. *Grid Search*) bądź użycie walidacji krzyżowej. Dzięki tym technikom można porównać optymalne wartości hiperparametrów, które dadzą najlepsze wyniki dla przygotowanych zbiorów danych.

- Uogólnioną metryką oceny wydajności modelu jest walidacja krzyżowa, która jest bardziej stabilna i dokładna niż użycie standardowego podziału danych na zbiór uczący oraz zbiór testowy. W przypadku walidacji krzyżowej trenowanych jest wiele modeli ze względu na wielokrotny podział danych, najczęściej stosuje się *k*-krotną walidację krzyżową, gdzie *k*-jest parametrem określanym przez użytkownika (zazwyczaj jest to liczba 5 lub 10). Dla każdego z podzbiorów trenowany jest model oraz liczona jest jego dokładność. Pierwszy model trenowany jest z podzbioru pierwszego jako zestawu testowego, a pozostałe służą jako podzbiory służą jako zestawy uczące. Model budowany jest na podstawie podzbiorów uczących, następnie jego dokładność liczona jest na zestawie pierwszym. Kolejny model budowany jest na podstawie podzbioru drugiego (jako podzbioru testowego), pozostałe zestawy wraz z zestawem pierwszym z poprzedniej iteracji służą jako zestawy uczące. Ostatecznie otrzymywane są wyniki dokładności w takiej samej liczbie, ile było *k*-iteracji walidacji. Algorytm przy użyciu walidacji krzyżowej zapewnia uogólnienie wszystkich próbek w zestawie danych. Główną wadą walidacji krzyżowej (ang. *cross-validation*) jest duży koszt obliczeniowy, ponieważ trenowane jest *k*-modeli zamiast jednego. Założeniem walidacji krzyżowej jest wskazanie, na ile dobrze model będzie uogólniał na podstawie dostarczonego mu zestawu danych.
- Zwykła walidacja krzyżowa nie sprawdza się na danych, które są posortowane w swoim zestawie, dlatego też − w takim przypadku − najczęściej stosuje się stratyfikowaną *k*-krotnej walidacji. W tego rodzaju walidacji proporcje pomiędzy klasami zostają zachowane zarówno w każdym podzbiorze, jak i całej klasie. W zależności od danych, które posłużą do uczenia algorytmu można posiłkować się również walidacją krzyżową z pominięciem (ang. *leave-one out*), gdzie dla każdego podzbioru wybierany jest jeden punkt danych jako zestaw testowy, jest to czasochłonne, ale skutecznie szacuje dokładności modeli, które uczone są na małym zestawie danych. Warto również wspomnieć o walidacji krzyżowej z podziałem losowym (ang. *shuffle-split*), pozwala ona na kontrolę nad liczbą iteracji niezależnie od wielkości zestawu uczącego i treningowego. W szczególności przydatna jest w przypadku dużych zestawów danych.
- Walidacja krzyżowa pozwoli ocenić, na ile dobrze model potrafi uogólniać. Kolejnym krokiem jest poprawienie wydajności modelu przez dostrajanie hiperparametrów.

Zarówno tradycyjne algorytmy uczenia maszynowego, jak i sieci neuronowe posiadają modyfikowalny zestaw parametrów, nazywane hiperparametrami. Natomiast parametry, które są obliczane poprzez algorytm nazywane są wagami.

Innymi metodami doskonalenia modelu są techniki przeszukiwania siatki oraz losowego przeszukiwania zbioru [4, 9].

Przeszukiwanie siatki (ang. *Grid Search*). W zależności od liczby hiperparametrów danego algorytmu poprzez przeszukiwanie siatki powstaje tyle różnych modeli, ile wynosi iloczyn wszystkich hiperparametrów zdefiniowanych do sprawdzenia. Jednak przeszukiwanie siatki jest czasochłonne i w przypadku dużej liczby hiperparametrów nie jest skuteczną metodą;

Random Search jest drugą techniką przeszukiwania hiperparametrów modelu. Celem algorytmu Random Search jest losowe przeszukiwanie zbioru zdeklarowanych hiperparametrów w celu znalezienia ich najlepszych kombinacji. Na modele uczenia głębokiego składa się wiele hiperparametrów: współczynnik uczenia, liczba warstw, ilość połączeń głębokich, funkcja aktywacji. Znalezienie optymalnej kombinacji pozwala na uzyskanie najlepszych wyników. Algorytm losowego przeszukiwania działa w następujący sposób [9]:

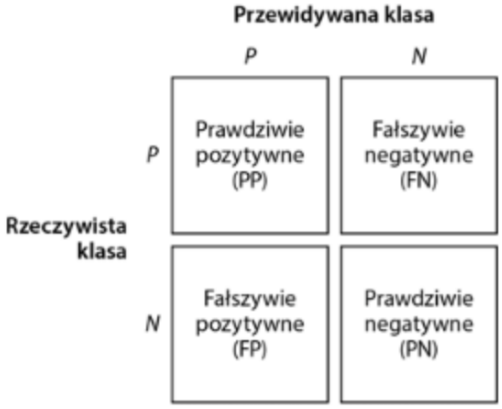
- Określenie zakresu hiperparametrów do sprawdzenia, np.: zakres współczynnika uczenia od 0,0001 do 0,001,
- Losowe próbkowanie – algorytm przeszukuje najlepsze zestawy hiperparametrów,
- Trenowanie i ocena – następuje trening modelu na wybranym zestawie hiperparametrów i jego ocena na zbiorze walidacyjnym pod kątem wybranych metryk,

- Zapisanie wyników i powtórzenie próby dla innych zestawów parametrów
- Wybór najlepszego zestawu hiperparametrów spośród przeprowadzonych prób.

4.1.3. Ocena modelu

Zarówno w klasie binarnej, jak i rozszerzeniu algorytmu na klasyfikację wieloklasową mówi się o klasie pozytywnej oraz o klasie negatywnej.

Macierz pomyłek pozwala na określenie skuteczności algorytmu i zarazem wstępem do innych bardziej złożonych metryk od dokładności precyzji. Jest to macierz macierzą kwadratowa, w której występują zliczenia klas przewidywań: prawdziwie pozytywna (ang. *true positive*), fałszywie pozytywna (ang. *false positive*), prawdziwie negatywna (ang. *true negative*), fałszywie negatywna (ang. *false negative*). Prognoza fałszywie pozytywna nazywana jest również błędem typu I, natomiast prognoza fałszywie negatywna jest nazywana w statystyce błędem typu II. Macierz pomyłek dla klasyfikacji binarnej została zaprezentowana na rys. 4.12 [12, 93, 122].



Rys. 4.12. Macierz pomyłek dla klasyfikacji binarnej [12, 93, 122]

Na podstawie macierzy pomyłek można wnioskować kilka podstawowych miar opisujących skuteczność skonstruowanego modelu. Jedną z takich miar jest związek z dokładnością, gdzie dokładność to liczba poprawnych przewidywań (TP i TN) względem wszystkich próbek (suma wszystkich wpisów macierzy błędu) (4.1) [9].

$$Dokładność = \frac{TP+TN}{TP+FP+TN+FN} \tag{4.1}$$

W przypadku niezrównoważenia klas można mówić o dwóch wskaźnikach skuteczności: odsetek prawdziwie pozytywnych (ang. *true positive rate*) (4.2) oraz odsetek fałszywie pozytywnych (ang. *false positive rate*) (4.3) [9].

$$OFP = \frac{FP}{N} = \frac{FP}{FP+PN} \tag{4.2}$$

$$OPP = \frac{FP}{N} = \frac{FP}{FP+PN} \tag{4.3}$$

Na podstawie przedstawionych powyżej dwóch metryk stworzone zostały wskaźniki wydajności: precyzja (ang. *Precision*) (4.4) oraz pełność lub inaczej czułość (ang. *Recall*), gdzie pełność jest synonimem OPP, natomiast czułość opisana jest wzorem (4.5) [9]:

$$Precyzja = \frac{TP}{TP+FP} \tag{4.4}$$

$$Czułość = \frac{TP}{TP+FN} \quad (4.5)$$

Precyzja jest miarą używaną, gdy istnieje potrzeba ograniczenia liczby fałszywie pozytywnych. Natomiast czułość jest miarą używaną do określenia liczby wszystkich próbek pozytywnych [9].

Często jednak wykorzystuje się kombinację dwóch powyższych wskaźników wydajności, jest to miara F1-score. Miara ta pozwala ona zróżnicować wady i zalety związane z optymalizacją powyższych metryk. Jest ona kompromisem pomiędzy czułością a optymalizacją precyzji. Kwalifikacja wszystkich próbek jako pozytywne spowoduje dużo fałszywych wyników. Model, zakwalifikuje w takim przypadku tylko jedną próbkę z dużą pewnością jako pozytywną, a pozostałe punkty jako ujemne. To oznacza, że będzie duża precyzja, ale niska czułość. Dlatego też stosuje się często miarę F1, która uwzględnia zarówno czułość, jak i precyzję (4.6) [9].

$$F1 = \frac{Precyzja\ x\ Czułość}{Pprecyzja + Czułość} \quad (4.6)$$

Metodą analizy doboru modeli klasyfikujących jest krzywa charakterystyki roboczej odbiornika (ang. *Receiver operating characteristic*), częściej używa się określenia krzywa ROC. Krzywe ROC bazują na skuteczności policzonych na odsetku prawdziwie pozytywnych oraz odsetku negatywnie pozytywnych. Im bliżej prawego górnego rogu znajduje się krzywa ROC, tym lepszy jest klasyfikator, punkt ten oznacza wysoką precyzję oraz czułość. Na podstawie krzywej ROC można obliczyć pole pod wykresem (ang. *Area under curve*) AUC, które opisuje skuteczność modelu, Krzywa AUC podobnie, jak w przypadku krzywej ROC, uwzględnia wszystkie możliwe zakresy dla klasyfikatora, lecz zamiast wskazywać wartości dokładności i czułości, pokazuje współczynnik wyników fałszywie pozytywnych FPR w porównaniu z wynikiem fałszywie pozytywnych. Wartość AUC można interpretować jako wartość, która wskazuje, że wszystkie losowo wybrane punkty pozytywnej będą miały wyższy wyniki prawdopodobieństwa niż klasy negatywne [9].

Powyższe wskaźniki modeli dotyczą zarówno predykcji binarnej, jak i wieloklasowej. W przypadku użycia powyższych miar dla klasyfikacji wieloklasowej występują metody makro- i mikro uśredniania, pozwalają one na stosowanie powyższych wskaźników poprzez klasyfikację jeden przeciwko wszystkim. Mikro-uśrednianie używane jest podczas analizy każdej prognozy (prawdziwie pozytywna, fałszywie pozytywna, prawdziwie negatywna, fałszywie negatywna) w taki sam sposób, natomiast makro-uśrednianie równorzędnie waży wszystkie klasy w celu ogólnej skuteczności klasyfikatora, średnia ta przydaje się szczególnie w przypadku niezrównoważenia klas [9, 93].

Najczęściej używaną metryką dla zestawów niezrównoważonych wieloklasowych jest wieloklasowa odmiana metryki F1. Jak wspomniano wcześniej, ideą wieloklasowych miar jest obliczanie binarnej miary dla danej klasę, przy czym ta klasa jest klasą pozytywną, pozostałe stanowią klasy negatywne. W kolejnym kroku dokonuje się uśredniania przy użyciu uśredniania makro, uśredniania ważonego bądź uśredniania mikro.

Uśrednianie makro oblicza nieważone miary F1 dla każdej z klas, zachowując w ten sposób równowagę pomiędzy klasami niezależnie od ich wielkości. Uśrednianie ważone oblicza średnią F1 dla każdej z klas, natomiast uśrednianie mikro oblicza całkowitą liczbę wyników fałszywie pozytywnych, fałszywie negatywnych i prawdziwie pozytywnych we wszystkich klasach, gdzie następnie liczona jest precyzja, czułości i miara F1 [9].

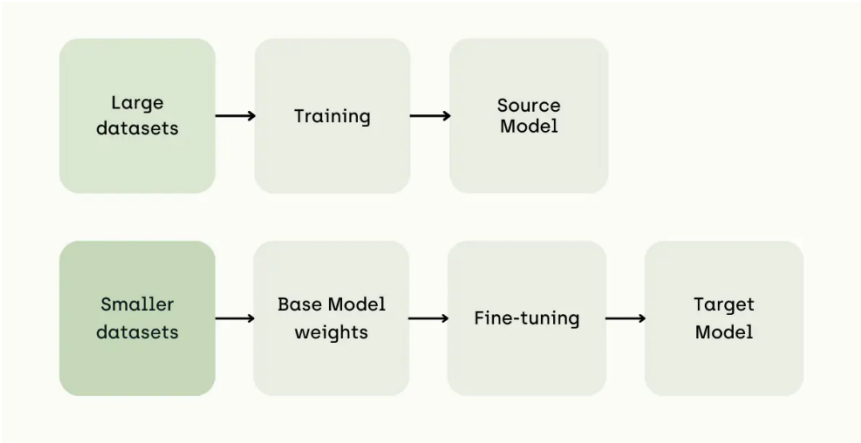
Do ekstrakcji cech wideo w dużym stopniu wykorzystuje się wstępnie wytrenowane modele głębokie. W szczególności dotyczy to modeli CNN oraz RNN. Splotowe sieci posiadają zdolność hierarchicznej ekstrakcji cech o różnym poziomie abstrakcji. Proces ten polega na przygotowaniu wielu różnych warstw sieci splotowej i warstw gęstych, które przetwarzają wzorce danych

reprezentatywnych. Hierarchiczna ekstrakcja cech i parametrów danych przebiega następująco [45]:

1) Filtry warstw splotowych uczą się podstawowych wzorców danych obrazu czy wideo, takich jak krawędzie, narożniki i wzorce kolorów co prowadzi do wyodrębnienia podstawowych cech przestrzennych.
2) Warstwy splotowe głębsze następnie analizują bardziej skomplikowane wzorce, takie jak: tekstury obiektów, kształty obiektów czy części jednostkowe obiektów. Warstwy te ekstrahują bardziej złożone parametry danych wejściowych, które następnie mogą być kombinacją cech wyodrębnionych w poprzednim etapie.
3) Warstwy połączeń gęstych są ostatnimi warstwami sieci splotowej, które agregują parametry z różnych cech obiektów i uczą się globalnych cech rozpoznawania wzorców używanych do zadań takich jak klasyfikacja czy detekcja obiektów.

Hierarchiczna ekstrakcja parametrów pozwala na zmniejszenie informacji i wydobywanie jedynie istotnych cech obiektów. Pozwala to na bardziej efektywne wykorzystanie modeli CNN do analizy obrazów wideo.

Kolejnym wykorzystaniem sieci neuronowych do zadań ekstrakcji cech wideo jest skorzystanie z metody uczenia transferowego (ang. *Transfer Learning*). Jest to technika polegająca na wykorzystaniu wcześniej wytrenowanych sieci neuronowych przygotowanych do rozwiązania innego, ale podobnego problemu. W kontekście ekstrakcji cech z wideo korzysta się z wstępnie wytrenowanych modeli CNN, które zostały uczone na dużych zbiorach obrazów. Zaletą tego rozwiązania jest brak potrzeby przygotowania większych zbiorów danych, przygotowanie dużych zbiorów danych jest pracochłonne i istnieje potrzeba posiadania dużych zasobów zarówno pamięciowych i obliczeniowych komputerów. Wykorzystanie metody *Transfer Learning* pozwala również na uniknięcie przetrenowania modelu, modele wcześniej przygotowane są regularyzowane, dalsze przygotowanie modelu (ang. *fine-tuning*) ogranicza się głównie do dodania kilku warstw sieci głębokich. Proces *fine-tuningu* polega na dalszym uczeniu wcześniej przygotowanego modelu na nowych danych treningowych, wstępnie wybiera się model, który został przygotowany do podobnego zadania, następnie należy zamrozić warstwy modelu, pozwala to na zapobiegnięcie zmianom w nauczonych już cechach modelu, które są użyteczne do nowego zadania. Po dodaniu kilku nowych warstw następuje dostosowanie ich wag poprzez przekazanie do modelu nowych danych uczących. Schemat różnic pomiędzy uczeniem modelu od podstaw oraz strojenia gotowego modelu zaprezentowano na rys. 4.13 [128].



Rys. 4.13. Różnice uczenia modelu od podstaw oraz strojenia gotowego modelu [128]

Sieci splotowe pozwalają również na ekstrakcje cech przestrzennych i temporalnych, które pozwalają na analizę kolejnych klatek obrazu wideo i wykrywanie zmian pomiędzy nimi.

4.2. Wykorzystanie uczenia głębokiego w rozpoznawaniu emocji w filmie

Netflix obecnie pracuje nad kolejnym rozszerzeniem rekomendacji produkcji filmowych swoim subskrybentom. W tym celu tworzone jest rozwiązane, które na celu ma skrócenie oryginalnych zwiastunów (*trailerów*) filmów, które proponowane są użytkownikom, tak aby bardziej emocjonalnie wpływały na widza i tym samym zachęcały go do obejrzenia filmu.

Początkowo rekomendacja opierała się na dostarczaniu, dostarczane były opisów filmu w postaci słownej, aby użytkownik mógł zapoznać się z jego treścią oraz zaproponowano system rekomendacyjny oparty na metadanych filmu i porównań go z innymi filmami wcześniej oglądanymi i ocenianymi przez widza [28]. Następnie wdrożono dodatkowo − przy wybieranych przez widza pozycjach − prezentowanie zwiastunów kinowych oraz ich ewentualnego skracania tylko do kilkunastu początkowych sekund, aby widz mógł zadecydować czy dana pozycja go interesuje. Obecnie trwające prace skupiają się na rozpoznawaniu emocji widza w oparciu o analizę wyrazu twarzy lub innych reakcji mimowolnych. Ma to na celu wychwycenie najważniejszych elementów filmu i prezentowanie ich widzowi w formie *trailera*, aby ten podjął decyzję o obejrzeniu filmu w całości. Model emocji zbudowany został na podstawie prac Ekmana, który w 1972 zaproponował model składający się z sześciu emocji: radość, smutek, strach, wstręt, złość, zaskoczenie [21]. Badania wykazały, że w wyniku analizy bardzo łatwo znaleźć odwzorowanie radości na ludzkiej twarzy, jednak pozostałe emocje są trudniejsze do odczytania [57].

Kolejnymi serwisami, które budują swoje metody filtracji produkcji filmowych i muzycznych są serwisy udostępniające twórcom muzycznym i filmowym inne utwory, m.in.: muzykę, pojedyncze ujęcia czy całe fragmenty filmu. Mogą zostać one użyte w innych produkcjach za uiszczeniem opłaty i poszanowaniem licencji twórcy. Przykładowo serwis Artlist pozwala na wyszukiwanie ujęć czy fragmentów filmowych na podstawie kilku zapytań związanych z kategorią ujęć, długości czy jakości ujęcia, a nawet liczby osób znajdujących się na ujęciu [2].

Z kolei wyszukiwanie muzyki np. w serwisach Epidemic Sound czy Artlist opiera się na informacjach zawartych w metadanych utworów, lecz również umożliwione jest przeszukiwanie bazy pod kątem przynależności utworów do grup tematycznych bądź etykiet nastrojów, które mogą towarzyszyć muzyce podczas odsłuchiwania [2, 23]. Epidemic Sound stworzył algorytm klasyfikujący utwory różnych gatunków muzycznych do ponad 30 nastrojów, które mogą towarzyszyć słuchaczom. Użył do tego głębokich sieci splotowych, które jako dane wejściowe wykorzystywały spektrogramy danych utworów muzycznych, udało im się osiągnąć na tak dużej liczbie klas 91% dokładności [35].

Spotify czy Apple również zaczęły łączyć domeny wykrywania emocji z użyciem uczenia maszynowego oraz rozwijania systemu rekomendacji użytkownikom [1, 75]. Kolejne badania firmy Spotify pokazują silną korelacje pomiędzy nastrojem i gatunkiem muzycznym pod kątem badania osobowości użytkowników [1]. Anderson wraz z zespołem postanowili udowodnić, że rodzaj słuchanej muzyki jest zależny od osobowości danej osoby. W tym celu grupa ankietowanych miała za zadanie przez trzy miesiące korzystać z serwisu Spotify, a następnie odpowiedzieć na kwestionariusz składający się z 44 pytań, po analizie którego można określić typ osobowości danej osoby [114] – otwartość, sumienność, ekstrawersja, ugodowość i stabilność emocjonalna. Następnie wytrenowano model uczenia maszynowego w celu przewidzenia danego typu osobowości ze względu na preferencje muzyczne i informację demograficzną o użytkownikach [1].

Firma Apple skupiła się na stworzeniu algorytmu umożliwiającego automatyczne wykrycie nastroju pochodzącego z danego utworu muzycznego. Zbudowano model uczenia głębokiego w oparciu o parametry utworów muzycznych, począwszy od metadanych (tytuł, autor, gatunek itp.) utworów poprzez parametry audio utworów muzycznych (długość utworu, skala, dynamika, energia, taneczność itp.). Następnie podjęto próbę zebrania etykiet opisowych tych utworów muzycznych, które opisywały emocje użytkowników podczas odsłuchu danego utworu muzycznego i predykcji utworów pod wybrany przez użytkowników nastrój bądź samopoczucie [67,75].

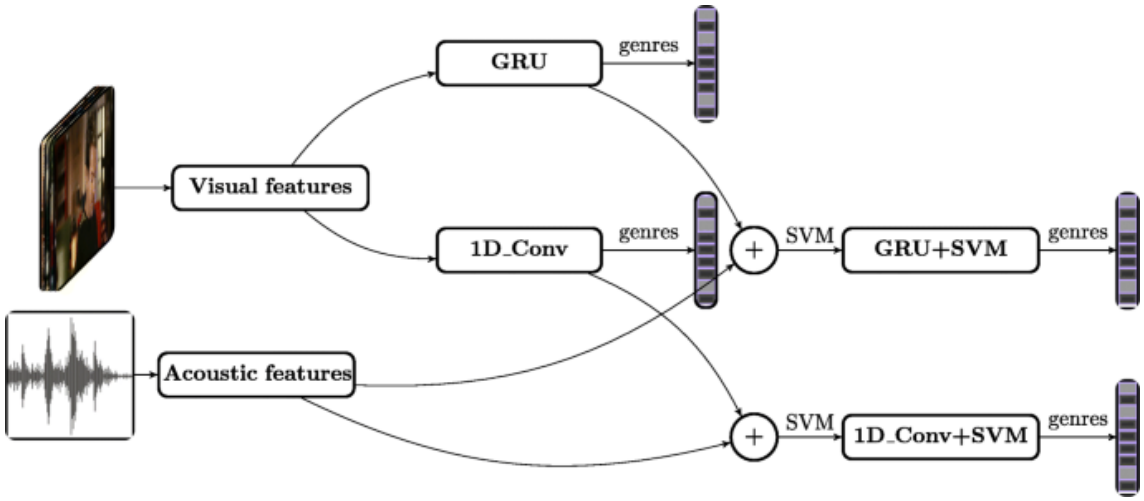
4.2.1. *Wykorzystanie sieci splotowych i rekurencyjnych do klasyfikacji z wykorzystaniem danych audio oraz wideo*

W pracy Behrouzi i in. przedstawiono problem klasyfikacji trailerów filmowych przy użyciu rekurencyjnych sieci GRU (ang. Gated Recurrent Unit) [7].

Autorzy zaproponowali odseparowanie danych audio od danych wideo z pliku. Do badań użyto zbioru LMTD, który zawiera wieloetykietowy opis około 4 tysiąca trailerów filmowych. Łączna liczba klas to 9, wśród których znajdują się: Action, Adventure, Comedy, Crime, Drama, Horror, Romance, Sci-Fi oraz Thriller. Zbiór danych LMTD nie jest zbiorem zbalansowanym, do badań zbiór ten został podzielony na zbiór treningowy, walidacyjny oraz testowy w proporcji 7:1:2.

Zaproponowane metody sieci neuronowych wykorzystują zarówno sieci rekurencyjne, jak i jednowymiarowe sieci splotowe, o rozmiarze paczki równym 32 oraz współczynniku uczenia 0,0001. Długość treningu modeli wyniosła 100 epok.

Do ekstrakcji cech filmowych użyto sieci VGG16 oraz Resnet_152, ekstrakcji cech dokonano na 240 klatkach fragmentu trailera [7]. Z fragmentu audio wydobyto współczynniki MFCC oraz współczynniki LPC. Następnie przeprowadzono szereg badań nad dokładnością klasyfikacji. Dokonano klasyfikacji gatunków filmu jedynie na podstawie cech wizualnych filmu za pomocą splotowej sieci jednowymiarowej. W dalszym kroku użyto dodatkowo algorytm SVM, gdzie wektor cech powstał na podstawie łączenia parametrów wizualnych przygotowanych przez sieć GRU oraz parametrów audio. Ostatnią porównywaną architekturą była architektura algorytmów łączonych, tj. ekstrakcji cech wizyjnych poprzez sieć splotową jednowymiarową oraz cech akustycznych. Architekturę zaproponowaną przez Behrouzi i in. przedstawiono na rys. 4.14 [7].



Rys. 4.14. Zaproponowana architektura w pracy Behrouzi i in. [7]

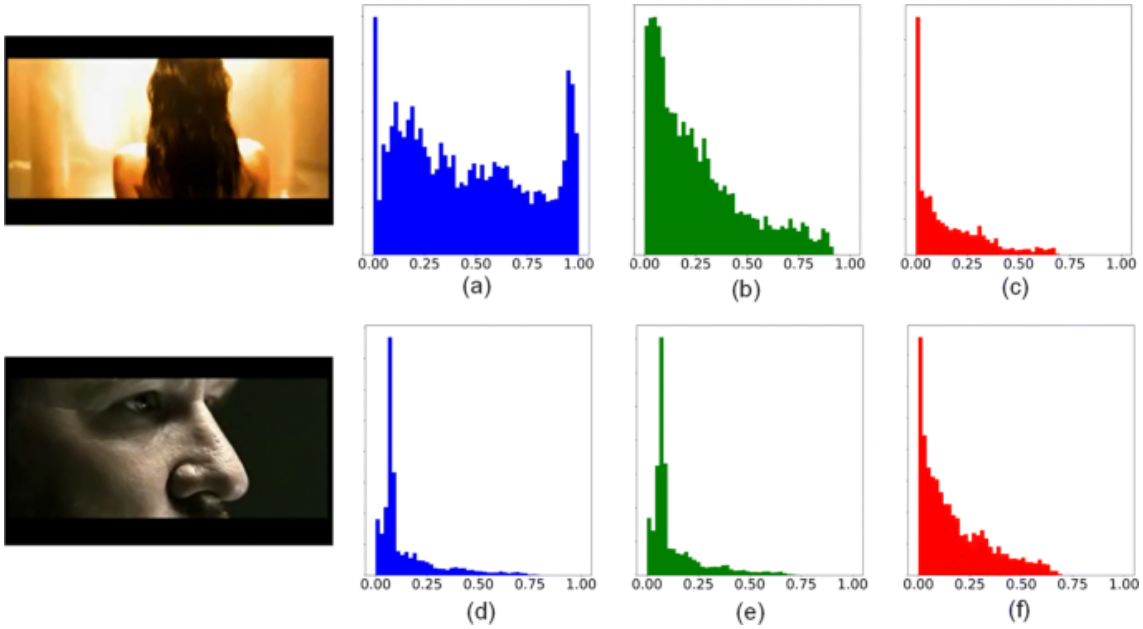
Najlepszym uzyskanym wynikiem był wynik średniej ważonej dla miary AUC na poziomie 0,864 oraz F1-score 0,66 dla architektury GRU + SVM. Wykorzystanie tylko cech wizualnych okazało się niewystarczające do uzyskania wysokich wyników. Poniżej przedstawiono wyniki dla metryki F1-score (tab. 4.2) [7].

Tab. 4.2. Wyniki F1-score dla architektur zaproponowanych w pracy Behrouzi i in. [7]

Model	**F1-score**
1D_Conv_V	0,61
1D_Conv+SVM_M	0,63
GRU_V	0,65
GRU_SVM_M	0,66

Przykładem wykorzystania sieci rekurencyjnych do klasyfikacji gatunków filmowych opartym wyłącznie na obrazie wideo jest praca Yu i in. [127]. W pracy tej zaproponowano klasyfikację gatunków filmowych w oparciu o przestrzenno-czasowe cechy sygnału wideo. Zastosowane podejście ma na celu wyodrębnienie globalnych, semantycznych i sekwencyjnych cech fragmentu wideo.

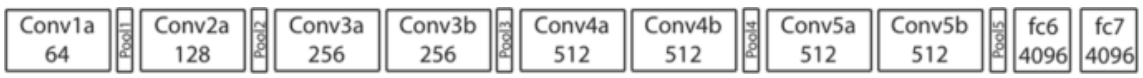
Do badań nad klasyfikacją gatunku filmowego skorzystano z ogólnie dostępnego zbioru danych, który zawiera 14 tys. trailerów filmowych pobranych z serwisu YouTube, o średniej długości 126 sekund, gdzie jednej sekundzie materiału wideo odpowiada 14 klatek filmu. Każdy trailer opisany jest odpowiednią etykietą gatunku filmowego. Z ponad 100 dostępnych etykiet wybrano 19 najczęściej powtarzających się. Do badań nad zadaniem klasyfikacji gatunku filmowego wybrano sieci rekurencyjne BiLSTM W pierwszej fazie wyekstrahowano klatki filmowe z trailerów filmowych na podstawie podobieństwa występowania między sobą, w oparciu o histogram kolorów RGB (rys. 4.15). Algorytm składa się z trzech modułów: wybór sąsiadujących ze sobą ramek z uwzględnieniem podobieństwa w histogramach koloru, moduł deskrypcji cech czasowo-przestrzennych oraz moduł sekwencyjny, który ponownie tworzy nową sekwencje będącą danymi wejściowymi dalszego poziomu algorytmu.



Rys. 4.15. Ekstrakcja cech wykorzystująca histogram kolorów RGB [127]

Wejście algorytmu jest czterowymiarowe, macierz cech zawiera następujące dane: liczbę klatek, wysokość klatki, szerokość klatki oraz dane kolorystyczne dla każdego z kanałów RGB. Następnie każda klatka jest pomniejszana do wielkości wejścia algorytmu, to jest 112x112.

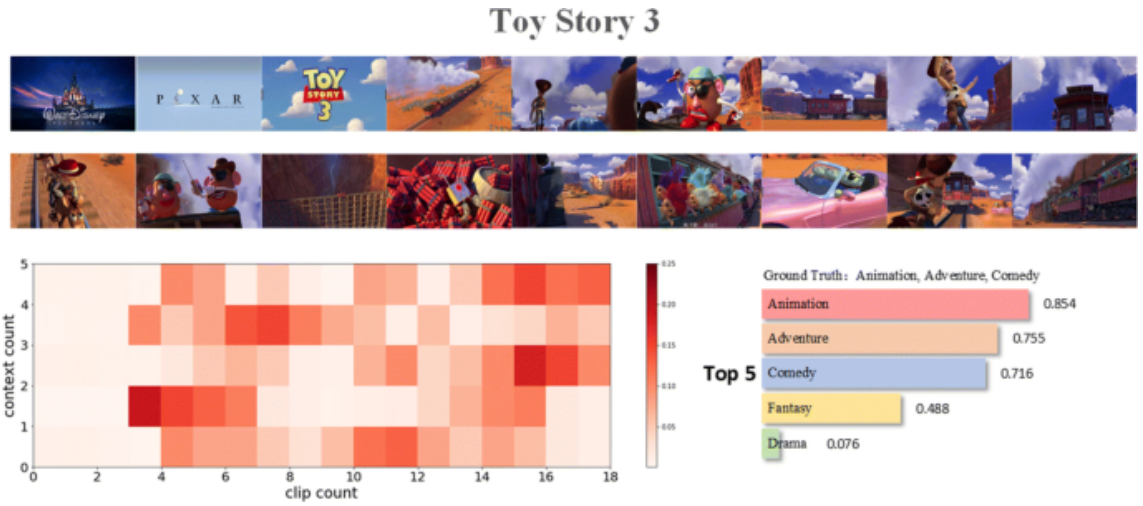
Tak przygotowane dane poddano wydobyciu cech za pomocą trójwymiarowej sieci splotowej, aby uzyskać cechy przestrzenno-czasowe obrazu. Architektura sieci splotowej zawierała 8 warstw splotowych, 5 warstw max-pooling oraz 2 warstwy w pełni połączone. Architekturę zaproponowaną przez Yu i in. przedstawiono na rysunku 4.16 [127].



Rys. 4.16. Architektura sieci splotowej zaproponowana przez Yu i in. [127]

Cechy przestrzenno-czasowe podano na wejście rekurencyjnej sieci LSTM zakończonej funkcją sigmoidalną pozwalającą na predykcję modelu z odniesieniem do każdej klasy. Do

porównania z innymi rozwiązaniami opartymi na sieciach splotowych oraz sieciach rekurencyjnych użyto metryki Hit Rate oraz mAP (ang. *Mean Average Precision*). Metryka Hit Rate jest miarą pozwalającą na ocenę przewidywań systemu na podstawie współczynnika liczby poprawnych przewidywań do liczby całkowitych prób. Współczynnik Hit Rate wykorzystuje trzy rodzaje przewidywań oparte na prawidłowej predykcji jednej klasy (HR@1), trzech (HR@3) oraz pięciu klas (HR@5). Na rys. 4.17 przedstawiono interfejs wizualny pracy algorytmu.



Rys. 4.17. Klasyfikator gatunków filmowych w pracy Yu i in. [127]

Poniżej (tab. 4.3) przedstawiono zestawienie proponowanej architektury na tle innych badań nad klasyfikacją gatunku filmowego przy użyciu danego zbioru danych [127].

Tab. 4.3. Wyniki badań nad algorytmem klasyfikacji w pracy Yu Y. [127]

Model	**mAP**	**HR@1**	**HR@3**
Resnet-50 [He K.] + Self-LSTM	0,609	0,726	0,834
TSN [Wang]	0,642	0,768	0,918
VLF [Rasheed]	0,586	0,624	0,702
CNN-MoTion [Simoes	0,677	0,816	0,855
CTT-MMC [Wehrman]	0,669	0,801	0,905
Self-C3DConvLSTM (BiLSTM)	0,685	0,883	0,921

4.2.2. Wykorzystanie metod uczenia głębokiego do klasyfikacji emocji z muzyki

Obecnie bardzo duże zainteresowanie zarówno w aspekcie badawczym, jak i komercyjnym wzbudza klasyfikacja emocji na podstawie muzyki bądź filmu. Istotnym aspektem jest również dobór modelu emocji.

Du i in. zaproponowali predykcję wartości VA w muzyce na podstawie wartości *arousal-valence* poprzez użycie kombinacji mel-spektrogramów oraz cochleogramów [20]. Pierwszym poziomem wielowarstwowej architektury jest sieć CNN, która posiada dwa wejście dla wcześniej wspomnianych danych. Mel-spektrogramy podawane są na wejście sieci neuronowej o wielkości 96x44 natomiast cochleogramy posiadają wielkość 12x44. Następie dane wejściowe separowalnie poddawane są operacji splotu w warstwie splotowej, *poolingu* oraz spłaszczania,

gdzie w warstwie w pełni połączonej są ze sobą łączone i przekazywane przez jedno wejście na sieć rekurencyjną BiLSTM, która dodatkowo odpowiada za czasowe cechy sygnału.

Zbiorem danych był zbiór liczący tysiąc 45-sekundowych przykładów z etykietami muzyki zachodniej Amazon Mechanical Turk. Dodatkowo zbiór danych został poddany augmentacji danych o zmianę częstotliwości oraz zmianę tonu utworu.

Dynamiczna predykcja obywała się na postawie analizy 0,5 sekundy utworu, oceny predykcji dokonano na podstawie metryki RMS. W tab. 4.4 przedstawiono wyniki RMSprop dla zaproponowanego algorytmu oraz innych rozwiązań literaturowych.

Tab. 4.4. Wyniki architektury BiLSTM w pracy Du i in. [20]

RUN	Arousal	Valence
RND	0,25±0,13	0,23±0,11
BSL	0,25±0,11	0,23±0,10
TUM	0,08±0,05	0,08±0,04
Du i in.	0,07±0,05	0,06±0,04

Revathy i Pillai [94] zaproponowali podejście klasyfikacji wieloetykietowej na podstawie 13 parametrów, które między innymi zawierały parametry dotyczące energii oraz walencyjności utworu, następnie wartości z danych kwadrantów wykresu VA (*Valence*/*Arousal*) przemapowano na 4 klasy emocji, przedstawiono to w tabeli 4.5 [94].

Tab. 4.5 Podział kwadrantów wykresu VA na cztery klasy emocji w pracy Revathy i Pillaia [94]

Range	Quadrant	Class	Emotional quadrant
Energy >= 0.5 and valence >=0.5	I	0	Happy
Energy >= 0.5 and valence < 0.5	II	1	Angry
Energy < 0.5 and valence <0.5	III	2	Sad
Energy < 0.5 and valence >= 0.5	IV	3	Relaxed

Spośród dostępnych 13 parametrów wybrano cztery parametry o największym współczynniku korelacji *r*-Pearsona, tj. *popularity*, *danceability*, *valence* oraz *arousal*.
Architektura modelu przygotowana została na bazie czterech warstw MLP: warstwa ukryta, warstwa dropout, warstwa ukryta, warstwa ukryta. W pracy osiągnięto dokładność 82% podczas ewaluacji modelu [94].

W pracy Grekowa i współpracowników przedstawiono predykcję wartości V/A na modelu emocji Russela. Zbiór danych obejmował 324 sześciosekundowe fragmenty adnotowanej muzyki poprzez ekspertów z wykształceniem muzycznym [25].

Następnie użyto narzędzia Marsyas, które pozwala na ekstrakcję cech audio, w sumie wyekstrahowano 31 cech, między innymi:

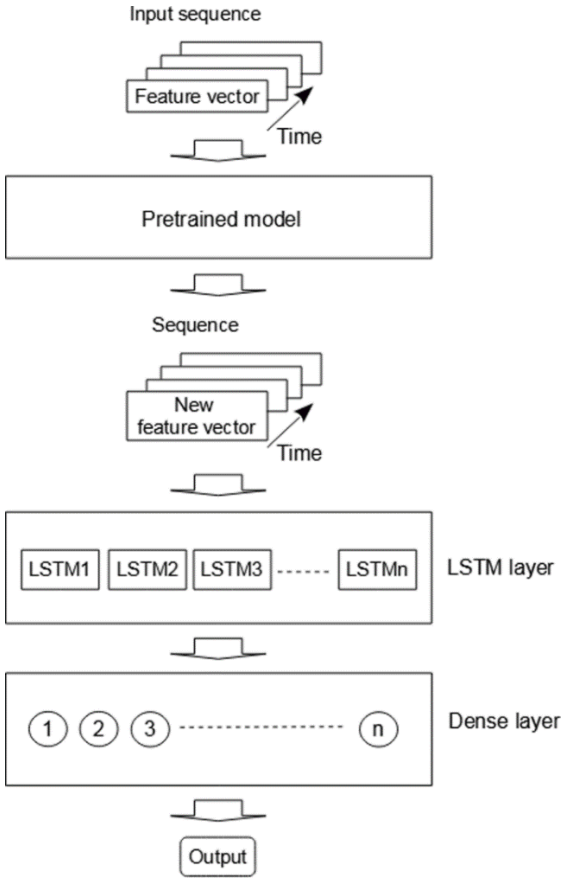
- *Zero Crossing;*
- *Spectral Centroid;*

- *Spectral Flux;*
- *Spectral Rolloff;*
- *Mel-Frequency Cepstral.*

Architektura sieci neuronowej wykorzystuje model sieci LSTM (patrz rys. 4.18), druga warstwa architektury to warstwa głęboka z funkcją aktywacji *tanh*. Trening odbył się na 100 epokach i został przerwany po 10 ze względu na brak poprawy nauki sieci.

Eksperymenty zostały przeprowadzone pięciokrotnie:

- Wykorzystano wszystkie cechy wyeksportowane poprzez rozwiązanie Marsyas;
- Wykorzystano parametry MFCC;
- Wykorzystano cechy chromatyczne;
- Wykorzystano cechy z narzędzia Essentia, które pokrywają cechy z narzędzia Marsyas i dodatkowo posiadają cechy wysokopoziomowe sygnału;
- Wykorzystano wstępnie wytrenowane sieci neuronowej do ekstrakcji cech sygnału.



Rys. 4.18. Architektura sieci klasyfikatora emocji zaproponowana w pracy Grekowa i in. [25]

Użycie wysokopoziomowych cech sygnału audio oraz przetrenowanych sieci neuronowych pozwoliło na osiągnięcie w badaniu najlepszych wyników. Zostały one zawarte w tab. 4.6. Autorzy odnieśli się miar: R^2 (współczynnik dopasowania) oraz MAE (Mean absolute error; średni błąd bezwzględny).

Tab. 4.6. Wyniki architektury klasyfikatora emocji w pracy Grekowa i in. [25]

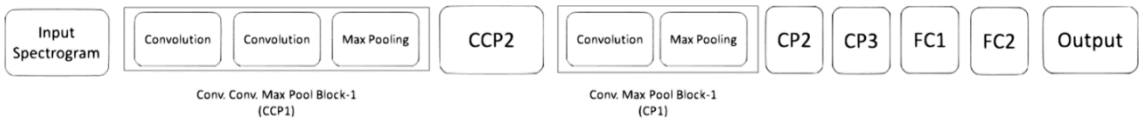
	Arousal		**Valence**	
Model	R^2	MAE	R^2	MAE
RNN1(124, LSTM)	0,63	0,13	0,35	0,14
RNN2(124, 124, LSTM)	0,73	0,11	0,42	0,13
RNN3(248, LSTM)	0,68	0,12	0,36	0,14
RNN4(248, 248, LSTM)	0,73	0,11	0,46	0,12
RNN5(529, LSTM)	0,69	0,12	0,38	0,13
RNN6(592, 529, LSTM)	0,71	0,12	0,46	0,12

Jak można zauważyć, najlepsze rezultaty osiągnięto dla sieci RNN z dwiema warstwami zawierającymi 248 jednostek. W przypadku *arousal* współczynnik dopasowania R^2 wyniósł 0,73, a dla *valence* – współczynnik wyniósł 0,46 [25].

W pracy Sarkara i in. [98] przedstawiono również metodykę klasyfikacji emocji. Klasyfikację przeprowadzono na postawie modelu emocji Russela, każdy z kwadrantów modelu opisano jedną emocją: szczęście, złość, smutek oraz emocja neutralna. W pracy użyto dwóch zbiorów danych: Soundtrack, Bi-Modal [98].

Każdy z plików audio w zbiorze danych znormalizowano do wartości od -1 do 1. Następnie pliki audio podzielono na 5-sekundowe fragmenty oraz na ich podstawie wygenerowano spektrogramy w skali melowej.

W ten sposób przygotowane dane zostały przekazane na wejście dwuwymiarowej splotowej sieci neuronowej. Dodatkowo zastosowano regularyzację L2 na warstwach w pełni połączonych (ang. *fully connected*). Architektura sieci została oparta na popularnej architekturze VGGNet. Sieć obejmowała ponad 1,2 miliona parametrów, wielkość wkładu ustalono na 64. Wykorzystano optymalizację Adam, zaś współczynnik uczenia wynosił 0,001. Na rys. 4.19 przedstawiono architekturę sieci zaproponowaną w pracy Sarkara i in. [98]



Rys. 4.19. Architektura sieci w pracy Sarkara i in. [98]

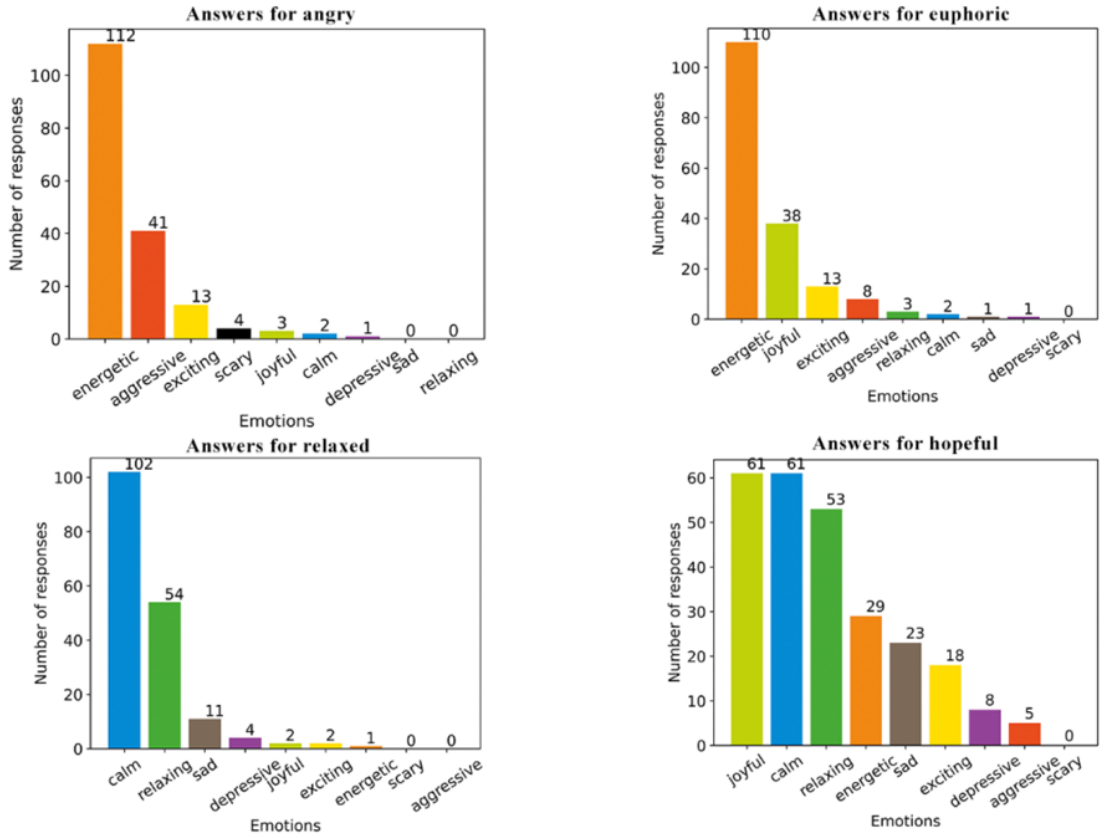
Zaproponowane rozwiązanie porównano z innymi rozwiązaniami opartymi na tych samych zbiorach danych. Wyliczono współczynniki dokładności, precyzji, czułości oraz wartości F1-score. Poniżej, w tab. 4.7, przedstawiono wyniki zaproponowanej architektury [98].

Tab. 4.7. Wyniki badań w pracy Sarkara i in. [98]

Metodologia	Dokładność (%)
K_NN – BE Saari i in. [96]	56,5±2,8
SVM – BE Saari i in. [96]	54,3±1,9
Model uczenia głębokiego Sarkara i in. [98]	67,7±3,6

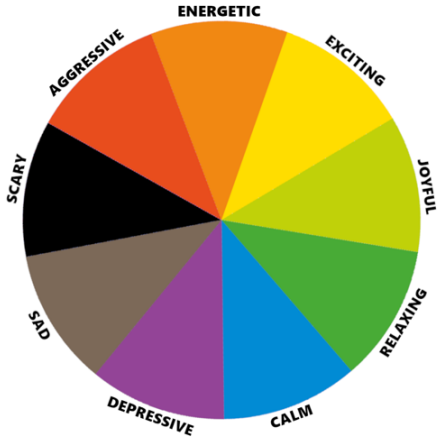
W pracy Ciborowskiego i in. wykorzystano muzykę z bazy Epidemic Sound, która wstępnie została zakwalifikowana do 19 różnych nastrojów [16]. W pracy zaproponowano model emocji wykorzystujący dziewięć elementów, do których zostały przypisane kolory. Osoby biorące udział

w ankiecie miały za zadanie określić, jakie dwie emocje odczuwają podczas słuchania danego utworu oraz dodatkowo zaznaczyć kolor, który kojarzy im się z wybraną emocją. Na podstawie wyników z ankiety (rys. 4.20) zaproponowano 10-elementowy model emocji wpisany w okrąg (rys. 4.21). Następnie przemapowane utwory na model emocji przygotowano zbiór danych do uczenia splotowych sieci neuronowych. Przemapowane emocje poddano analizie statystycznej, wyznaczono statystycznie znaczące różnice w parach emocji za pomocą testu chi-kwadrat, dodatkowo obliczono współczynnik r-Pearsona do sprawdzenia korelacji pomiędzy dopasowaniem koloru a wybraną przez uczestników emocją zawartą w opracowanej ankiecie [16].



Rys. 4.20. Histogram odpowiedzi dla mapowanych emocji [16]

Na podstawie wyników z ankiety zaproponowano model emocji:



Rys. 4.21. Zaproponowany model emocji w pracy Ciborowskiego i in. [16]

Spośród dużej bazy utworów Epidemic Sound wybrano 420 utworów, na podstawie których wygenerowano mel-spektrogramy z różną długością analizy:

- Utwory podzielone na fragmenty o długości 30 sekund analizowano oknem długości kolejno: 2, 4, 6, 8 oraz 10 sekund, dla których wyznaczono mel-spektrogramy.
- Utwory podzielone na fragmenty o długości 15 sekund przeanalizowano oknem o długości 10 sekund.

Następnie przygotowano trzy różne warianty wielkościowe 224x224, 299x299 mel-spektrogramów jako wejścia do splotowych sieci neuronowych.

Przetestowano łącznie pięć różnych architektur sieci splotowych:

- Xception – z wejściem o wielkości 299x299,
- VGG19 – z wejściem o wielkości 224x224,
- InceptionResNet50 – z wejściem o wielkości 224x224,
- Inception v3 – z wejściem o wielkości 299x299.

W procesie uczenia sieci neuronowych zaproponowano własny model liczenia dokładności sieci splotowych oparty na sprawdzeniu trzech najbardziej znaczących predykcji emocji, gdyż dokładność klasyfikacji binarnej okazała się niewystarczająca. Zaproponowane obliczenie dokładności modelu składa się na wyborze trzech emocji z największym procentem predykcji oraz porównaniu jej z listą etykiet treningowych. Co ważne, jeżeli lista testowa zawiera etykiety [a, b, c], a lista z nową predykcją zawiera etykiety [a, c, b], to dokładność wynosi 100%, ponieważ te same etykiety pojawiły się w predykcjach. Jeśli jednak lista testowa zawiera etykiety [a, b, c], lecz model przewidział etykiety [c, d, e], to dokładność wynosi 33%. Współczynnik uczenia ustalono na 0,001 oraz 0,0001.

Duża liczba zmiennych parametrów spowodowała cztery różne etapy uczenia modeli ze zmianą wielkość wsadu, współczynnika uczenia, itd. Poniżej, w tab. 4.8, przedstawiono wyniki dla czwartego etapu uczenia.

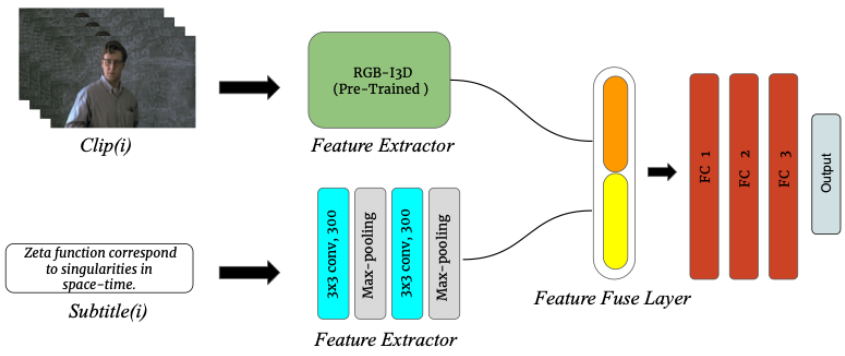
Tab. 4.8. Wyniki zaproponowanych modeli uczenia głębokiego dla klasyfikacji emocji w pracy Ciborowskiego i in. [16]

Nazwa modelu	Xception	ResNet50V2	InceptionResNetV2	InceptionV3
Rozmiar wsadu	64	16	64	16
Współczynnik uczenia	10^{-4}			
Zbiór danych	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 walidacyjny [%]	90,54	87,63	89,17	93,16
Dokładność [%]	60,86	59,9	61,11	61,66
Test własny [%]	76,58	75,57	77,9	78,71
Liczba epok	150	150	141	68

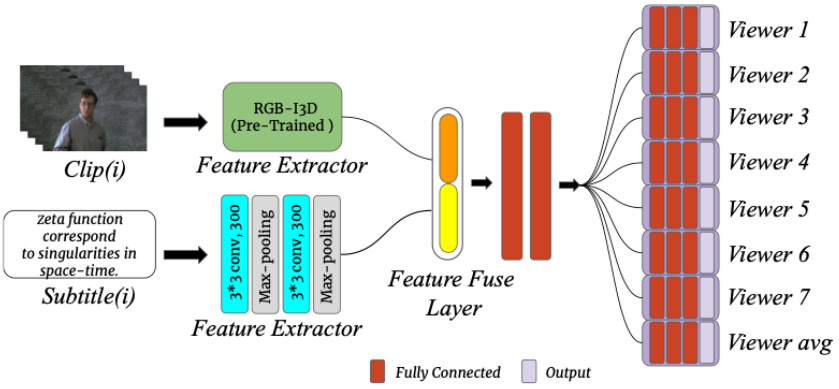
Najbardziej efektywnym modelem okazała się architektura sieci InceptionV3, która następnie została zaimplementowana do aplikacji pozwalającej na predykcję emocji na podstawie fragmentu utworu, który użytkownik podaje na wejście aplikacji i zaznacza fragment utworu [16].

4.2.3. Wykorzystanie metod uczenia głębokiego do klasyfikacji emocji z fragmentów filmu

Hayat i in. proponują dwa sposoby użycia sieci splotowych do klasyfikacji emocji odczuwanych przez widza z fragmentów obrazu: modelu o pojedynczym zadaniu (ang. *single-task model*) (rys 4.22) oraz wielozadaniowego modelu (rys. 4.23.) (ang. *Multi-task model*) [30]. Analizie poddano zarówno obraz wideo, jak i również napisy tłumaczące oryginalne głosy lektorów. Każdy z modeli posiada osobne wejścia dla części analizy wideo oraz części analizy tekstu. Parametry dla tekstu zostały wyznaczone jako macierz dystrybucji słów (ang. *word-embeddings matrix*), dzięki temu sekwencje wyrazów mogą zostać przetransponowane na wektory liczbowe, każdy wektor posiada taką samą liczbę słów równą 18. Parametryzację wideo przeprowadzono na podstawie parametrów czasoprzestrzennych, na wejściu przygotowano wstępnie wytrenowany model RGB-I3D [58] na zbiorze Kinectic-400 [44].



Rys. 4.22. Zaproponowana architektura w pracy Hayata i in., tzw. model pojedynczo-zadaniowy [30]



Rys. 4.23. Zaproponowana architektura w pracy Hayata i in., tzw. model wielozadaniowy [30]

Oba modele zostały wytrenowane na zbiorze COGNIMUSE [83]. Fragmenty filmów ze zbioru zostały oznaczone na modelu wymiarowym V/A. Klatkaż filmu wynosi 25 klatek na sekundę, każda klatka opisana jest współrzędnymi znajdującymi się na modelu wymiarowym A/V. Natomiast do zbioru tekstowego wybrano ramki co 40 ms filmu. W pracy skupiono się głównie na wartościach walencyjności emocji. Do treningu użyto 6 różnych fragmentów filmów po 30 minut każdy oraz jednego fragmentu 30-minutowego do walidacji modelu. Model wielozadaniowy (ang. *multi-task,* MT) uzyskuje predykcje dla każdego z poszczególnych oglądających, natomiast model pojedynczo-zadaniowy (*ang. single-task,* ST) uzyskuje predykcje ogólną dla całości pojedynczego filmu. Model MT wypadł znacznie lepiej pod kątem dokładności przewidywania: 73,6%. Natomiast model ST uzyskał dokładność na poziomie 65,6% [30].

Aslan i in. zaproponowali użycie sieci splotowych do klasyfikacji emocji na podstawie filmu [3]. Dane treningowe, walidacyjne oraz testowe zostały przygotowane w oparciu o zbiór danych LIRIS-ACCEDE. Zbiór danych zawiera łącznie 53 różne produkcje filmowe, które zostały uszeregowane względem gatunku filmowego. Następnie z 53 fragmentów filmów przygotowano pojedyncze ramki o wielkości 256x256, jedna taka ramka zawiera etykietę *valence-arousal*.

Jako architektury sieci zaproponowano wstępnie wytrenowane modele sieci splotowych, między innymi: VGG19, ResNet50 oraz InceptionV3. Dla każdej z sieci funkcją kosztu była MSE oraz MAE, dodatkowo zastosowano metody optymalizacji SGD, Adam oraz RMSProp [3].

Podejściem predykcji wartości na wykresie V/A było przygotowanie zbioru danych podzielonego pod katem gatunków filmowych. Dane zostały opisane poprzez etykiety z wartościami *valence* oraz *arousal* z przedziału [-1, 1]. Zbiór danych zawiera opis poszczególnych klatek: deskryptor koloru oraz krawędzi, kolor warstwy, histogram krawędzi, opisy ruchu, jak również przypisanie do gatunku filmowego [3]. Na rys. 4.24 pokazano przykładowe wyniki predykcji dla omawianej architektury.

Rys. 4.24. Klasyfikacja wartości V/A na podstawie zaproponowanej architektury w pracy Aslana i in. [3]

W tab. 4.9 przedstawiono wyniki dla predykcji wartości *arousal* oraz *valence* dla poszczególnych wybranych architektur sieci splotowych. Skuteczność modeli oparto na wskaźnikach MSE oraz współczynniku korelacji Pearsona (PCC).

Tab. 4.9. Wyniki dla zaproponowanej architektury w pracy Aslana i in. [3]

Gatunek filmu	Aktualne badania [3]		Wcześniejsze badania [47]	
	MSE	PCC	MSE	PCC
Valence	0,085	0,243	0,115	0,14
Arousal	0,06	0,18	0,17	0,091

Wyniki badania przywołują wskaźniki związane ze skutecznością sieci ResNet50, gdyż zarówno błąd średniokwadratowy, jak i współczynnik korelacji Pearsona pozwala określić czy sieć najlepiej generalizuje wartości V/A dla danych testowych [3].

Inne podejście do predykcji wartości *valence* oraz *arousal* przy użyciu wcześniej wspomnianego zbioru danych jest przedstawione w pracy Khanh-Ana i in. [47]. Ze zbioru danych LIRIS-ACCEDE z każdego filmu wyekstrahowano jedną ramkę na jedna sekundę filmu. Następnie użyto wstępnie wytrenowanej sieci ResNet50 do ekstrakcji cech z każdej ramki obrazu, otrzymano w ten sposób 2048 wektor cech. Do zaimplementowanej sieci ResNet50 dołączono na końcu dwie warstwy w pełni połączonej. Jako funkcji kosztu użyto RMSProp oraz ustalono współczynnik uczenia na 0,0001. Predykcje policzono osobno dla wartości arousal oraz wartości valence. Dodatkowo zastosowano technikę średniego okna przesuwnego, aby pozbyć się losowego szumu z danych, przetestowano okna o wielkości: 3, 5 oraz 7.

Przetestowano pięć różnych konfiguracji architektur pod kątem złożoności warstw w pełni połączonych:

- Przebieg pierwszy oparty był o 128 połączeń w pierwszej warstwie oraz 512 w drugiej warstwie zarówno dla wartości *arousal* oraz *valence*. Trening został ustalony na 20 epok.
- Przebieg drugi powtórzono w oparciu o hiperparametry przebiegu pierwszego, lecz dodatkowo zastosowano średnie okno przesuwne o wielkości 7.
- Przebieg trzeci oparto na 256 połączeniach w pierwszej warstwie oraz 512 połączeniach w warstwie drugiej dla wyznaczenia wartości *valence*. Dla wartości *arousal* zarówno dla warstwy pierwszej, jak i drugiej liczbę połączeń ustalono na 512. Trening ustalono w oparciu o 15 epok.
- W przebiegu czwartym dla przewidywań wartości walencji pierwsza warstwa zawierała 256 połączeń natomiast druga 512. Dla przewidywań wartości pobudzenia pierwsza warstwa zawierała 128 połączeń, druga warstwa 512. Model został trenowany przez 10 epok.
- W ostatnim piątym przebiegu dla przewidywań wartości walencji warstwa pierwsza, jak i druga zawierały 512 połączeń oraz trenowane były przez 10 epok.

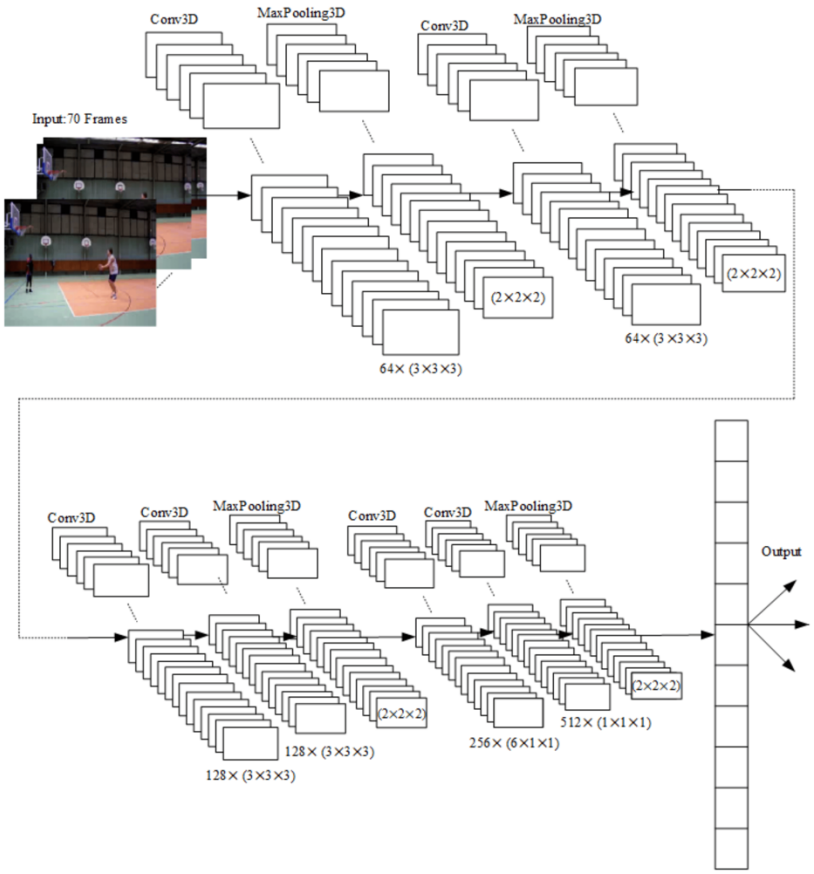
W tab. 4.10 przedstawiono wyniki eksperymentu dla wartości MSE oraz współczynnika korelacji *r*.

Tab. 4.10. Wyniki zaproponowanej architektury w pracy Khanh-Ana i in. [47]

Konfiguracja architektury	*Valence*		*Arousal*	
	MSE	r	MSE	r
1	0,12	0,11	0,17	0,05
2	0,11	0,15	0,17	0,07
3	0,12	0,14	0,17	0,07
4	0,12	0,14	0,18	0,02
5	0,11	0,14	0,17	0,09

Eksperymentalne wyniki pokazują, że można uzyskać satysfakcjonujące wyniki, korzystając z nieskomplikowanych architektur modeli głębokich, w tym sieci splotowych [47].

Dużą popularność zdobyły sieci splotowe trójwymiarowe, które pozwalają nie tylko na analizę przestrzenną obrazów, ale również na analizę czasową poszczególnych fragmentów czy ramek filmu. Klasyfikację aktywności przedstawiono za pomocą trójwymiarowych sieci splotowych w pracy „*Human Activity Classification Using the 3DCNN Architecture*" [119]. W pracy tej zaproponowano 6-warstwową architekturę trójwymiarowych sieci splotowych (rys. 4.25).



Rys. 4.25. Zaproponowana architektura sieci splotowej trójwymiarowej [119]

- Pierwsza warstwa jest warstwa CONV3D zawierająca 64 filtry oraz jądro o wielkości 3x3x3;
- Kolejną warstwą jest warstwa MaxPooling3D (2x2x2) oraz występująca po niej warstwa normalizująca paczkę wsadową BatchNormalization;
- Druga warstwa splotowa również zawiera 64 filtry oraz jądro o wielkości 3x3x3;
- Następnie zastosowano ponownie warstwę MaxPooling3D oraz BatchNormalization w takiej samej konfiguracji co poprzednio;

- Kolejnymi warstwami są dwie warstwy splotowe o liczbie filtrów równych 128 oraz rozmiarze jądra 3x3x3;
- Po dwóch warstwach splotowych powtórzono warstwę MaxPooling3D oraz warstwę normalizującą;
- Piąta i szósta warstwa splotowa zawiera liczbę filtrów 256 i 512 oraz rozmiary jądra 6x1x1 i 1x1x1;
- Przed warstwą gęstą Dense zastosowano w takiej samej konfiguracji warstwę MaxPooling3D oraz BatchNormalization.

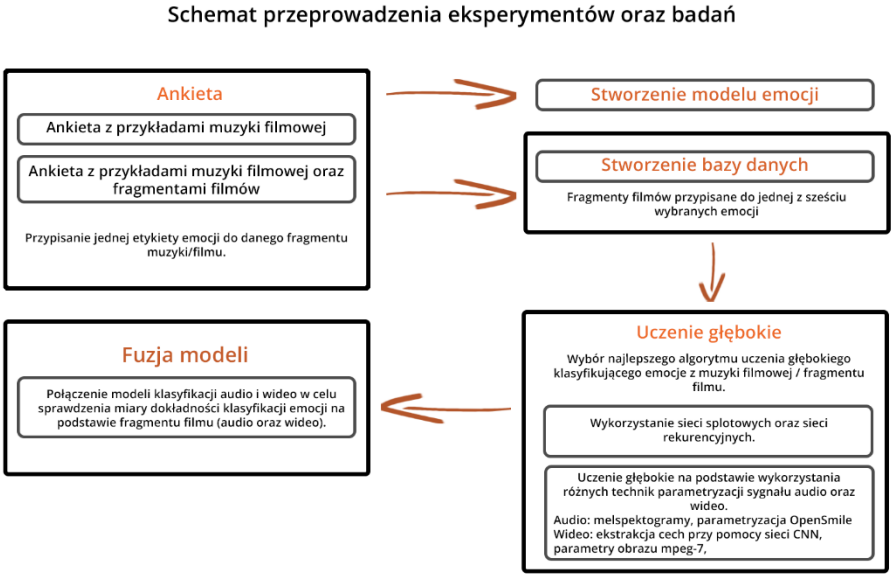
W pracy zdecydowano się na zmniejszenie rozmiaru obrazu wejściowego dla każdej ramki na rozmiar 32x32 piksele oraz ustalono stałą długość filmu na 70 pierwszych klatek wideo.

Użytym zbiorem danych do badań jest zbiór UCF YouTube Action, zawiera on 11 klas różnych aktywności: gra w koszykówkę, jazda na rowerze, pływanie, gra w golfa, żonglerka piłkarska, taniec, gra w tenisa, itp. W sumie zbiór zawiera 1160 różnych nagrań w formacie mp4 o rozdzielczości 320x240 z klatkażem 29,97. Drugim użytym zbiorem był zbiór UCF101 zawierający 101 klas różnych czynności wykonywanych przez ludzi, zbiór zawiera ponad 13 tysięcy filmów. Z tych dwóch zbiorów ostatecznie wybrano 1160 filmów, który to zbiór podzielono w proporcji 70:10:20 na zbiór treningowy, walidacyjny i testowy. Dokładność dla danych walidacyjnych podczas treningu sieci osiągnęła wartość 95%. Architekturę przetestowano na obu zbiorach, dla zbioru UCF YouTube Action osiągnięto dokładność 85,2% natomiast dla zbioru UCF101 osiągnięto dokładność równą 84,4% [119].

5. Eksperymenty wstępne

Pierwszym etapem było stworzenie odpowiedniej bazy fragmentów filmów i utworów muzycznych, które zostały zaproponowane w literaturze opisującej korelacje kolorów dominujących w filmie z emocjami widza, a następnie – na ich podstawie – przygotowanie dane niezbędne do przetwarzania przez modele uczenia głębokiego.

Na rys. 5.1 przedstawiono schemat prowadzenia eksperymentów oraz badań dotyczących zadania rozpoznawania i klasyfikacji emocji.



Rys. 5.1. Schemat prowadzenia eksperymentów oraz badań

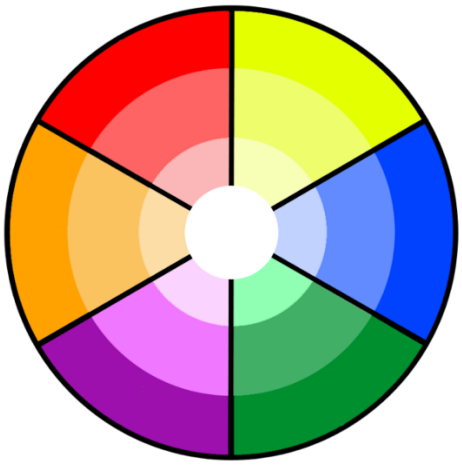
W badaniach prowadzonych przez autora rozprawy zdecydowano się na uproszczenie modelu emocji poprzez ograniczenie palety kolorów do sześciu wybranych. Wybrano trzy podstawowe kolory służące do określenia głównego modelu barwnego używanego w produkcji filmowej – modelu RGB. Model ten stosowany jest w każdym środowisku wyświetlającym kolor, zarówno podczas produkcji obrazu, jak i w grafice komputerowej. W skład modelu RGB wchodzą kolory takie jak: czerwony, zielony oraz niebieski. Dodatkowo do stworzenia nowego modelu emocji posłużono się literaturą z zakresu psychologii filmów „Jeśli to fiolet, ktoś umrze. Teoria koloru w filmie" autorstwa Bellantoni. W tej pozycji książkowej analizie emocjonalnej poddano sześć kolorów: czerwony, pomarańczowy, żółty, fioletowy, niebieski oraz zielony. Na podstawie tych dwóch uwarunkowań zaproponowano nowy wymiarowy model emocji, okrąg podzielono na sześć równych części, do której przypisano jeden kolor i jedną emocję. Analizę korelacyjną przedstawiono w rozdziale 6.

5.1. Założenia testów subiektywnych

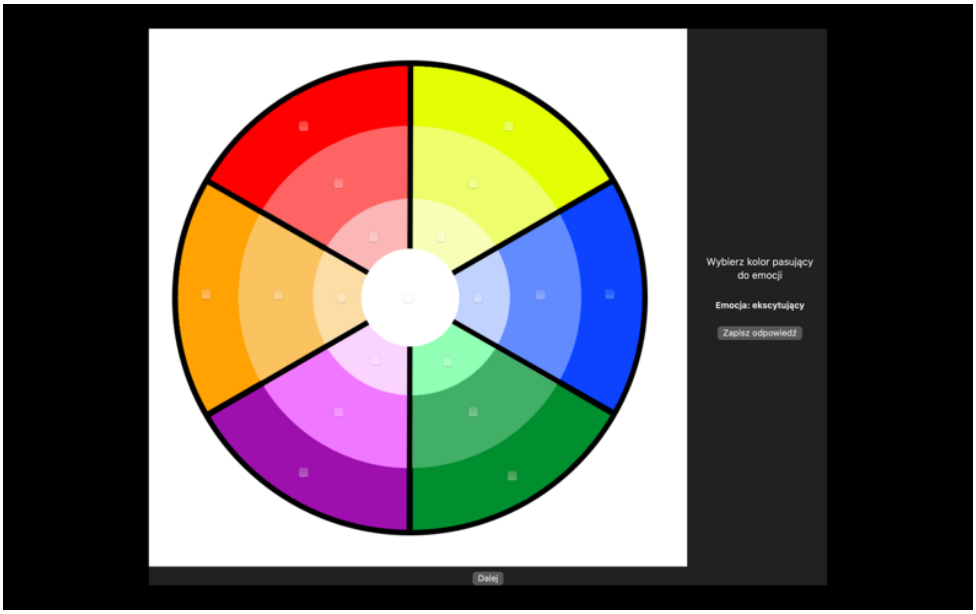
Celem testów subiektywnych było stworzenie uproszczonego modelu emocji, który jest bezpośrednio skorelowany z podstawowymi kolorami z palety barw używanej w filmie oraz odpowiednia etykietyzacja zbioru uczącego i testowego dla algorytmów uczenia głębokiego.

W ramach pracy przeprowadzone zostały trzy różne ankiety, pierwsza ankieta dotyczyła przypisania odpowiedniej emocji do koloru na kole, które zostało podzielone na sześć części (rys. 5.2). Ankietowani mieli za zadanie przypisać każdej z części koła jedną emocję. Każdy z kolorów koła posiada swój stopień gradacji w kierunku białego (neutralnego w wymowie). Na potrzeby testów subiektywnych została stworzona aplikacja w systemie operacyjnym MacOS z interfejsem

użytkownika pozwalającym na przypisanie konkretnej emocji do odpowiedniego punktu na kole emocji (rys. 5.3).



Rys. 5.2. Koło koloru z przyporządkowaniem emocji w aplikacji do testów subiektywnych



Rys. 5.3. Okno opracowanej aplikacji do testów subiektywnych

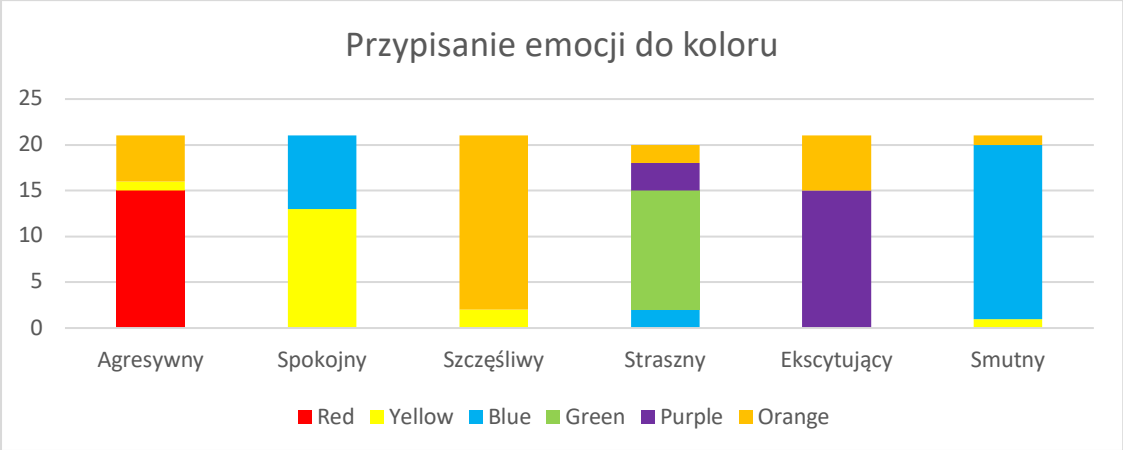
Ponadto uczestnicy ankiety zostali poproszeni o uzupełnienie krótkiego kwestionariusza osobowego dotyczącego zapytania o płeć, określenia przedziału wiekowego oraz znajomości tematyki filmowej. W trakcie badań wykorzystano również urządzenie do śledzenia ruchu gałek ocznych (ang. *Eye Tracker*) firmy Tobii, które pozwoliło na śledzenie wzroku ankietowanego i obliczenie czasu skupienia na ekranie. Czas skupienia wzroku użytkownika pozwala na analizę dodatkowych danych związanych z czytaniem [37] bądź wywoływaniem/przekazywaniem emocji [10]. Dzięki tym pomiarom sprawdzono czas skupienia ankietowanego na poszczególnych odpowiedziach oraz pomierzono całkowity czas, wymagany do uzyskania odpowiedzi z ankiety. Odpowiedzi z drugiej oraz trzeciej ankiety pozwoliły na przyporządkowanie emocję do fragmentu filmu oraz muzyki filmowej, pozwoliło to na stworzenie odpowiedniego zbioru danych do treningu sieci neuronowych.

5.2. Wybór modelu emocji

Pierwsza ankieta obejmowała zdanie przypisania kolorom odpowiednich emocji. W ankiecie udział wzięło 21 osób, którzy na specjalnie zaprojektowanej aplikacji, z pracującym w tle *Eye Trackerem*, mieli zadanie dopasowania etykiety emocji do koloru na zaproponowanym modelu. Poniżej przedstawione zostały wyniki pierwszej ankiety (tab. 5.1., rys. 5.4.).

Tab. 5.1. Wyniki zliczeń odpowiedzi w pierwszym teście subiektywnym

Aplikacja	**Agresywny**	**Spokojny**	**Szczęśliwy**	**Straszny**	**Ekscytujący**	**Smutny**
Red	15					
Yellow	1	13	2			1
Blue		8		2		19
Green				13		
Purple				3	15	
Orange	5		19	2	6	1



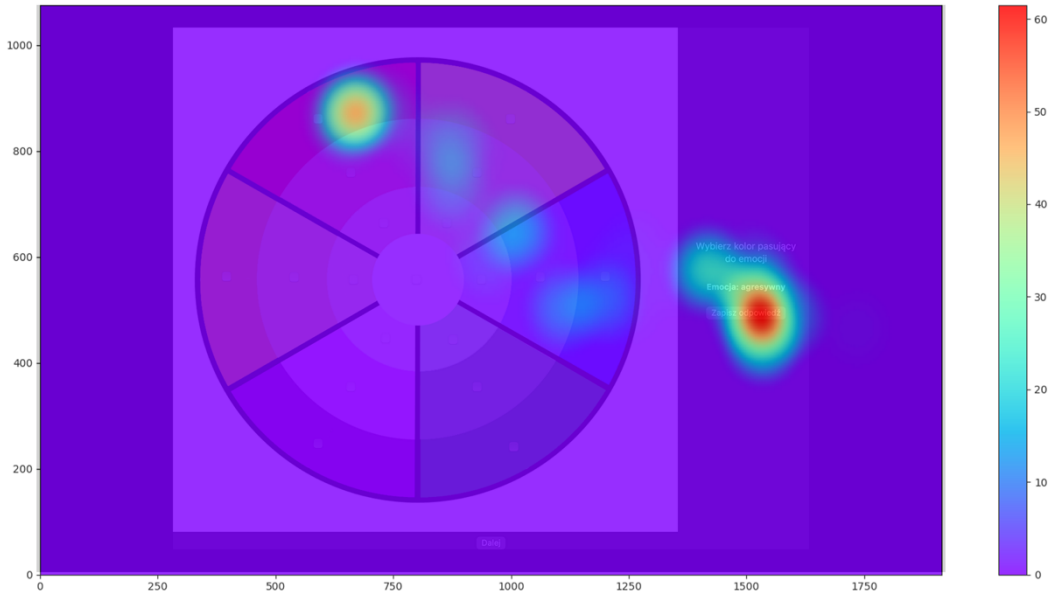
Rys. 5.4. Wykres kolumnowy odpowiedzi ankietowanych w pierwszym teście subiektywnym

Eye Tracker firmy Tobii pozwolił na policzenie czasu skupienia wzroku na ekranie dla każdej odpowiedzi. Przykładowe zliczenia przedstawiono w tab. 5.2.

Tab. 5.2. Fragment wyników czasu pomiaru skupienia wzorku z Eye Trackera.

Osoba	Odpowiedź	Mierzona wielkość	Sekundy (s)
1	1	Czas pomiaru skupienia wzroku	4,92
		Najdłuższy czas skupienia na odpowiedzi	2,65
		Drugi co do długości czas skupienia na odpowiedzi	1,37
	2	Czas pomiaru skupienia wzroku	12,87
		Najdłuższy czas skupienia na odpowiedzi	6,61
		Drugi co do długości czas skupienia na odpowiedzi	3,34

Mapa ciepła (ang. *heatmap*) przedstawia czas skupienia rozłożony na kolisty gradient w kolorze o zielonego do czerwonego. Im bardziej ciepły kolor, tym czas skupienia wzroku na danym punkcie ekranu był dłuższy. Badania jednoznacznie wskazują, że użytkownicy najczęściej skupiają wzrok na obszarach zainteresowań, co ostatecznie skutkuje szybszym zaznaczeniem danej odpowiedzi [10, 79, 82]. Na rys. 5.5 przedstawiono przykładową mapę skupienia wzroku na odpowiedzi.



Rys. 5.5. Przykładowa mapa skupienia wzroku na odpowiedzi

Czas skupienia wzroku został obliczony jedynie dla grupy pikseli znajdujących się w obszarze zainteresowania, czyli dla modelu emocji zaproponowanego w badaniach. W tab. 5.3 przedstawiono wyniki odpowiedzi dla najdłuższego czasu skupienia.

Tab. 5.3. Wyniki odpowiedzi dla najdłuższego czas skupienia sczytanych z Eye Trackera

Tobi	**Agresywny**	**Spokojny**	**Szczęśliwy**	**Straszny**	**Ekscytujący**	**Smutny**
Red	14	0	0	0	0	0
Yellow	1	13	2	0	0	1
Blue	0	7	0	2	0	16
Green		0	0	13	0	3
Purple	1	0	1	3	15	0
Orange	5	1	18	2	6	1

Poniżej przedstawiono wartość współczynnika korelacji zaznaczenia odpowiedzi w aplikacji i czasu skupienia wzroku na odpowiedzi. Test korelacji pozwala na zbadanie siły oraz kierunku korelacji pomiędzy danymi zmiennymi.

Korelację *r* Pearsona można obliczyć ze wzoru:

$$r = \frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{\sqrt{\sum_{i=0}^{n}(x_i - m_x)^2}\sqrt{\sum_{i=0}^{n}(y_i - m_y)^2}} \tag{5.1}$$

gdzie:
- *r* – współczynnik korelacji Pearsona,
- *x*, *y* – porównywane grupy,
- m_x, m_y – średnie wartości porównywanych grup,
- *n* – liczba elementów w porównywanych grupach.

Zwracaną wartością testu korelacji *r* Pearsona jest wartość *r* ($r \in [-1, 1]$), którą można interpretować w sposób następujący [110]:

Dla $|r|$:

- < 0.2 – brak korelacji pomiędzy zmiennymi,
- 0.2 – 0.4 – słaba korelacja,
- 0.4 – 0.7 – umiarkowana korelacja,
- 0.7 – 0.9 – dość silna korelacja,
- > 0.9 – bardzo silna korelacja.

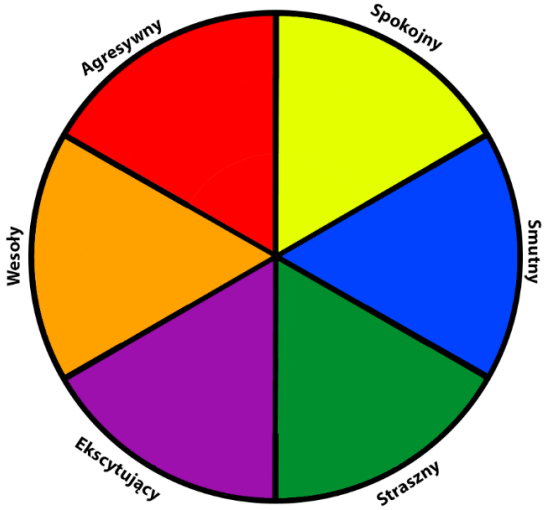
W tab. 5.4 przedstawiono współczynniki korelacji dla poszczególnych emocji zbadanych zarówno poprzez zliczenie odpowiedzi zaznaczone przez ankietowanego, jak i czas skupienia wzroku na danych odpowiedziach.

Tab. 5.4. Współczynnik korelacji dla poszczególnych emocji

Korelacja emocji	Wsp. korelacji
Agresywny	1,00
Spokojny	0,99
Szczęśliwy	1,00
Straszny	1,00
Ekscytujący	1,00
Smutny	0,98

Jak można zauważyć, dla każdego z przypadków analizy, współczynnik korelacji grup odpowiedzi z aplikacji i z systemu służącego do śledzenia wzroku, jest potwierdzony bardzo silną korelacją.

Na podstawie wyników z przeprowadzonych testów subiektywnych przedstawiono propozycję modelu emocji. W zaproponowanym modelu emocji nie uwzględniono gradacji kolorów, gdyż uczestnicy ankiety praktycznie w każdym przypadku odpowiedzi zaznaczyli najbardziej intensywny kolor (rys. 5.6).



Rys. 5.6. Zaproponowany model emocji

W modelu emocji wyróżniono sześć grup kolor-emocja:

- Czerwony – agresywny;
- Żółty – spokojny;
- Niebieski – smutny;
- Zielony – straszny;
- Fioletowy – ekscytujący;
- Pomarańczowy – szczęśliwy.

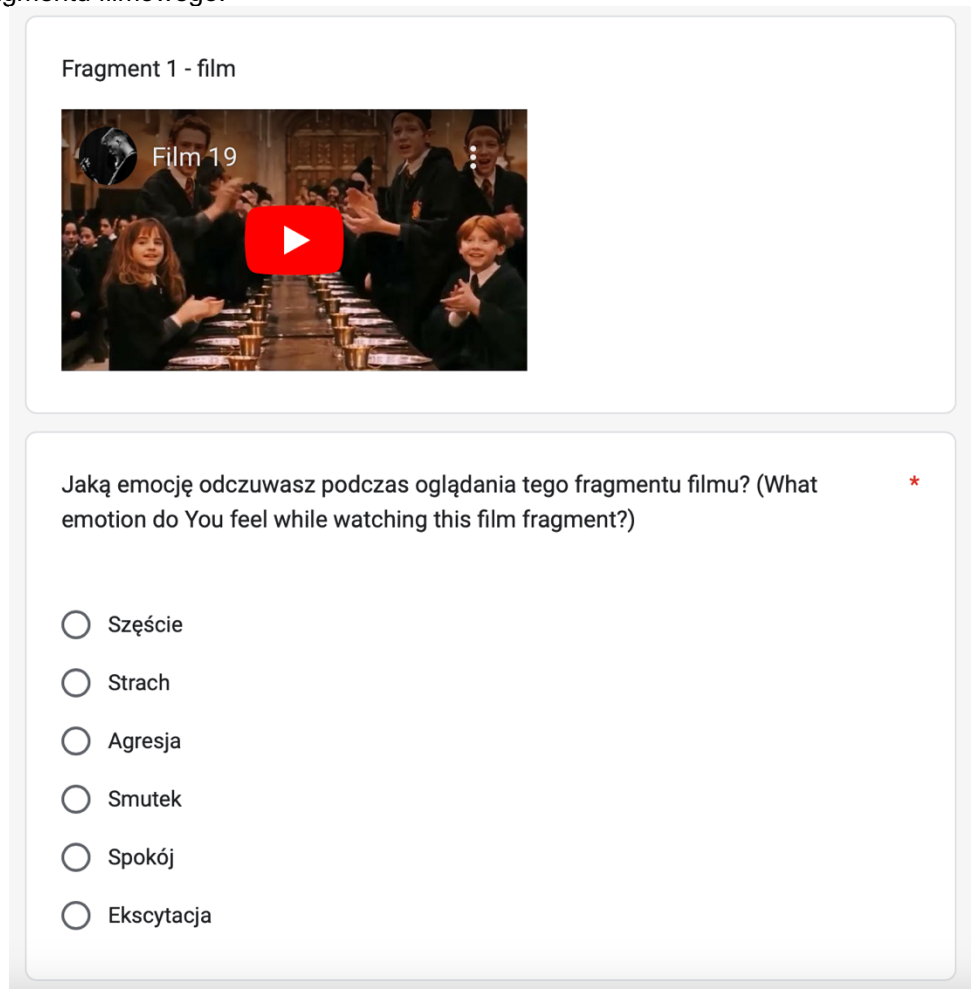
5.3. Przypisanie emocji fragmentom filmu oraz muzyce filmowej

W drugiej części testów subiektywnych przygotowano dwa warianty kwestionariusza, pierwszy wariant zakłada odsłuchanie fragmentu muzyki filmowej i przypisanie mu etykiety emocji, jaką widz odczuł podczas słuchania. Druga ankieta filmowa zawiera powtórzony fragment

muzyczny z poprzedniej ankiety, ale dodatkowo wzbogacony jest fragmentem filmowym. Zarówno fragment muzyki, jak i fragment filmowy pochodzą z tej samej części linii czasowej wybranego filmu. Każda z ankiet składa się z dwóch kwestionariuszy, które naprzemiennie zostały pokazywane ankietowanym, w sumie zaprezentowano 36 przykładów fragmentów filmów z 36 różnych tytułów filmowych, które cechują się charakterystyczną kolorystyką oraz pochodzą z różnych gatunków filmowych. Tytuły filmowe zostały wybrane z książki „Jeśli to fiolet, ktoś umrze" [8] oraz na podstawie źródeł internetowych traktujących o psychologii kolorów w filmie [17, 77 109]. W ankiecie wzięło udział w sumie 60 osób różnych płci i z różnej kategorii wiekowej. W ankietach zamieszczono 15-sekundowe fragmenty filmów oraz fragmenty muzyki filmowej. Były to fragmenty między innymi z następujących filmów:

- „Gladiator" w reżyserii Ridley'a Scotta;
- „Kill Bill vol. 2" w reżyserii QuentinaTarantino;
- "Titanic" w reżyserii Jamesa Camerona;
- „The Shawshank Redepmtion" w reżyserii Franka Darabonta.

Na rys. 5.7 przedstawiono fragment ankiety dotyczącej przypisania emocji do danego fragmentu filmowego.



Rys. 5.7. Fragment ankiety dotyczącej przypisaniu odpowiedniej emocji fragmentom filmu

Fragmenty filmów wybrane zostały na podstawie dominanty sześciu kolorów użytych do stworzenia propozycji filmowego modelu emocji. Ankiety zostały przygotowane z wykorzystaniem formularzy Google'a i udostępnione uczestnikom w formie on-line.

Pierwsza ankieta polegała na wypełnieniu kwestionariusza osobowego w pierwszej części oraz w drugiej części dopasowaniu emocji, którą słuchający odczuwa podczas odtwarzania fragmentu utworu muzyki filmowej. Druga ankieta polegała na przypisaniu emocji, do fragmentu filmu wraz z towarzyszącą mu muzyką. W tab. 5.5 przedstawiono wyniki obu ankiet dla pierwszych dziesięciu pytań. Wartości do porównania zostały przyjęte ze stanu wiedzy (ang. *state-of-the-art*) [8 17, 77, 109].

Tab. 5.5. Porównanie odpowiedzi dla poszczególnych testów subiektywnych wraz z odniesieniem do stanu wiedzy

		Muzyka vs emocja		**Film vs emocja**		**Odniesienie do stanu wiedzy**	
Lp.	**Tytuł filmu**	**Odp.**	**Emocja**	**Odp.**	**Emocja**	**Kolor dominujący w ujęciu**	**Emocja**
1	A Very Long Engagement	26	Smutek	17	Szczęśliwy	Pomarańczowy	Szczęśliwy
2	Amelie	29	Szczęście	30	Szczęśliwy	Pomarańczowy	Szczęśliwy
3	Titanic	26	Szczęście	27	Szczęśliwy	Pomarańczowy	Szczęśliwy
4	The Thin Red Line	19	Ekscytacja	27	Strach	Zielony	Strach
5	Into the Wild	17	Strach	30	Strach	Zielony	Strach
6	Kill Bill vol. 2	23	Ekscytacja	28	Strach	Zielony	Strach
7	Rebel without a Cause	24	Agresja	29	Agresja	Czerwony	Agresja
8	Inglorious Bastards	23	Agresja	24	Agresja	Czerwony	Agresja
9	Ex Machina	16	Agresja	29	Agresja	Czerwony	Agresja
10	Spirited Away	26	Smutek	28	Smutek	Niebieski	Smutek

Dla każdej z 36 pozycji filmu zebrano 30 odpowiedzi. Następnie porównano je z wynikami zawartymi w aktualnej literaturze. Obliczono poziom istotności za pomocą testu chi-kwadrat. Test chi-kwadrat pozwala na określenie czy występuje pomiędzy zmiennymi zależność statystyczna. Przyjmuje się w tym celu hipotezę zerową oraz hipotezę alternatywną, które można zweryfikować w testach:

- Hipoteza zerowa – zmienne są od siebie zależne;
- Hipoteza alternatywna – zmienne nie są od siebie zależne.

W przypadku badania zależności zmiennych wykorzystuje się wzór:

$$X^2 = \sum \frac{(O-E)^2}{E} \tag{5.2}$$

gdzie:
- X^2 - wartość statystyki testowej,
- *O* - wartość zaobserwowana,
- *E* - wartość oczekiwana,
- *n* - liczba porównań.

Za hipotezę zerową przyjęto:

*Emocja wybrana większością głosów dla danego fragmentu filmu **jest** tożsama z przyjętymi wartościami z odniesienia do stanu wiedzy.*

Natomiast hipoteza alternatywna brzmi:

*Emocja wybrana większością głosów dla danego fragmentu filmu **nie jest** tożsama z przyjętymi wartościami z odniesienia do stanu wiedzy.*

Za wartość oczekiwaną przyjęto emocję *z odniesienia do stanu wiedzy,* natomiast za wartość obserwowaną przyjęto odpowiedzi ankietowanych ankiety Film vs Emocja, ponieważ przykłady znajdujące się w literaturze opisują emocje wywoływane u widza poprzez obraz i nie dotyczą muzyki filmowej.

Następnie obliczono poziom istotności odpowiedzi ankietowanych oraz wartość rozkładu testu chi-kwadrat. Założono poziom istotności równy 0,05. W celu obliczenia statystyki i poziomu istotności wykorzystano język programowa Python oraz bibliotekę scipy.stats [101]. Statystyka wyniosła: 12,56, natomiast poziom istotności: 0,99. Liczba stopni swobody w teście wyniosła 35, z tablicy rozkładu chi-kwadrat wartość teoretyczną dla p = 0,05 oraz 35 stopni swobody można odczytać jako: 49,81. W przeprowadzonym teście zarówno poziom istotności jest większy od zakładanego (0,99 - obliczony > 0,05 - zakładany) oraz obliczona wartość statystyki jest mniejsza niż teoretyczna (12,56 - obliczona < 49,81- teoretyczna). Na podstawie powyższej analizy można przyjąć odrzucenie hipotezy alternatywnej i przyjęcie hipotezy zerowej [107].

Obliczono również korelacje pomiędzy odpowiedziami z ankiety Muzyka vs. Emocja oraz Film vs. Emocja, aby określić relację pomiędzy muzyką a filmem. Jako główną zmienną wybrano odpowiedzi z ankiety Film vs. Emocja i porównano je z liczbą odpowiedzi tej samej odpowiedzi z ankiety „Muzyka vs. Emocja". W tab. 5.6 przedstawiono liczbę odpowiedzi ankietowanych względem najczęściej wybieranej emocji z ankiety „Film vs Emocja".

Tab. 5.6. Liczba odpowiedzi dla najczęściej zaznaczanej emocji

Lp.	Tytuł filmu	Odp. film vs emocja	Odp. muzyka vs emocja	Emocja
1	A Very Long Engagement	17	0	Szczęśliwy
2	Amelie	30	29	Szczęśliwy
3	Titanic	27	26	Szczęśliwy
4	The Thin Red Line	27	7	Strach
5	Into the Wild	30	30	Strach
6	Kill Bill vol. 2	28	2	Strach
7	Rebel without a Cause	29	24	Agresja
8	Inglorious Bastards	24	23	Agresja
9	Ex Machina	29	16	Agresja
10	Spirited Away	28	26	Smutek

Obliczony współczynnik korelacji równy 0,48 sugeruje umiarkowaną korelację pomiędzy emocjami wywołanymi przez muzykę towarzyszącą filmowi a emocjami wynikającymi z treści audio-wizualnej.

Ze względu na obliczoną powyższej statystykę i uwzględnienie zależności w odpowiedziach etykiety zbioru przyporządkowano względem odpowiedzi z testu subiektywnego sprawdzającego

dopasowanie emocji do fragmentu filmu, gdzie występuje zarówno obraz, jak i ścieżka dźwiękowa.

6. Eksperymenty z wykorzystaniem uczenia głębokiego

6.1. Założenia

Celem tej części rozprawy doktorskiej było zaproponowanie optymalnego w odniesieniu do skuteczności oraz wydajności algorytmu do klasyfikacji sześciu emocji na podstawie fragmentów filmów. Badania podzielono na trzy etapy:

- Pierwszy etap obejmuje przeprowadzenie testów z wykorzystaniem uczenia głębokiego do zadania klasyfikacji sześciu etykiet emocji na podstawie wyłącznie ścieżki audio z fragmentu filmu. Przeprowadzone testy obejmują zarówno wykorzystanie parametrów sygnału audio, jak i reprezentacji graficznej sygnału w postaci mel-spektrogramów.
- Drugi etap obejmuje przeprowadzenie testów z wykorzystaniem algorytmów uczenia głębokiego do klasyfikacji sześciu etykiet emocji na podstawie wyłącznie obrazu wideo pochodzącego z fragmentu filmu. Analiza sygnału wideo opiera się na analizie parametrycznej sygnału przy wykorzystaniu m.in. narzędzi OPENSmile oraz hierarchicznej ekstrakcji cech przy pomocy sieci splotowych.
- Trzeci etap testów obejmuje klasyfikację sześciu klas emocji za pomocą metod uczenia głębokiego z wykorzystaniem zarówno sygnału audio, jak i sygnału wideo. Testy obejmują architektury łączone, które pozwalają na analizę sygnałów pochodzących z dwóch źródeł. W testach zaproponowano rozwiązania oparte zarówno na sieciach splotowych, jak i rekurencyjnych.

Implementacji modeli dokonano przy pomocy modułów Keras oraz Tensorflow, korzystając z narzędzia Google Colaboratory Pro. Trening modeli prowadzony był z wykorzystaniem GPU Nvidia Tesla T4, zawierającej 16 GB dostępnej pamięci RAM [29]. W ramach eksperymentów sprawdzono dziewięć różnych architektur sieci neuronowych, z których wybrano dwa z najwyższą dokładnością, które posłużyły do stworzenia modelu bimodalnego, dlatego kolejne rozdziały stanowią rodzaj raportu dotyczącego efektywności treningu, walidacji oraz testów sieci, w tym zaadaptowanych z literatury.

6.2. Zbiór danych

W zbiorze danych znalazło się w sumie 36 tytułów filmowych oznaczonych zarówno kolorem (dominującym kolorem w fragmencie filmu) oraz etykietą emocji, które wcześniej zostały przyporządkowane podczas testów subiektywnych. W tab. 6.1 przedstawiono przypisanie etykiet emocji dla poszczególnych tytułów filmowych ze zbioru danych.

Tab. 6.1. Przypisanie etykiet emocji dla poszczególnych tytułów filmowych wykorzystanych w zbiorze danych

Tytuł filmu	**Kolor**	**Emocja**
A Very Long Engagement	Pomarańczowy	Szczęśliwy
Amelie	Pomarańczowy	Szczęśliwy
Titanic	Pomarańczowy	Szczęśliwy
The Thin Red Line	Zielony	Straszny
Into the Wild	Zielony	Straszny
Kill Bill vol. 2	Zielony	Straszny
Rebel without a Cause	Czerwony	Agresywny
Inglorious Bastards	Czerwony	Agresywny
Ex Machina	Czerwony	Agresywny
Spirited Away	Niebieski	Smutny
The Revenant	Niebieski	Smutny
The Grand Budapest Hotel	Niebieski	Smutny
Chicago	Fioletowy	Ekscytujący
Lost River	Fioletowy	Ekscytujący
La La Land	Fioletowy	Ekscytujący
Hotel Chevalier	Żółty	Spokojny
The Age of Innocence	Żółty	Spokojny
Gladiator - opening scene	Żółty	Spokojny
Harry Potter	Pomarańczowy	Szczęśliwy
Don Juan de Marco	Pomarańczowy	Szczęśliwy
The Martian	Pomarańczowy	Szczęśliwy
Maleficient	Zielony	Straszny
Gattaca	Zielony	Straszny
Taxi Driver	Zielony	Straszny
Deadpool	Czerwony	Agresywny
Malcom X	Czerwony	Agresywny
Dick and Tracy	Czerwony	Agresywny
The Shawshank Redemption	Niebieski	Smutny
The Corpse Bride	Niebieski	Smutny
Working Girl	Niebieski	Smutny
Gladiator	Fioletowy	Ekscytujący
Cabaret	Fioletowy	Ekscytujący
Rushmore	Fioletowy	Ekscytujący
The Lion King	Żółty	Spokojny
Elisabeth Golden Age	Żółty	Spokojny
English Patient	Żółty	Spokojny

Następnie z powyżej przedstawionych tytułów filmowych wybrano od dwóch do czterech 15-sekundowych fragmentów które dodatkowo podzielono na 5-sekundowe fragmenty z sekundowym oknem przesuwnym. Podział na fragmenty oraz przydzielenie fragmentów do odpowiednich klas został przeprowadzony poprzez skrypt automatyzujący, przygotowany w języku Python z wykorzystaniem bibliotek moviepy oraz ffmpeg. Pozwoliło to na wygenerowanie 154 fragmentów filmowych dla każdej z klas.

W zbiorze danych znalazło się dzięki temu podziałowi znalazło się 924 fragmentów filmów o łącznej długości 77 minut.

Zbiór ten został podzielony w proporcji: 70% (zbiór treningowy), 10% (zbiór walidacyjny) oraz 20% (zbiór testowy):

- Treningowy zawierający 644 5 sekundowych fragmentów filmów;
- Walidacyjny zawierający 90 5 sekundowych fragmentów filmów;
- Testowy zawierający 190 5 sekundowe fragmenty filmów.

6.3. Klasyfikacja emocji z fragmentu muzyki filmowej z wykorzystaniem uczenia głębokiego

6.3.1. Klasyfikacja emocji na podstawie spektrogramów melowych ścieżki audio fragmentów filmu

Do klasyfikacji sześciu emocji z przedstawionego wcześniej modelu emocji wybrano trzy popularne rozwiązania sieci InceptionV3, InceptionResNetV2 oraz Xception. Modele uczenia głębokiego zostały wybrane na podstawie wyników treningu i walidacji sieci na zbiorze ImageNet [Image-net] oraz na podstawie pracy Ciborowskiego i in. [16], gdzie wykorzystano te modele do klasyfikacji emocji na podstawie muzyki [16].

W implementacji wykorzystano wagi architektur sieci, które zostały wytrenowane wstępnie na zbiorze obrazów ImageNet, który zawiera 1000 klas. Do przeprowadzenia testów na wyżej wymienionych architekturach użyto techniki przenoszenia wiedzy (ang. *transfer learning*) ze zmianą ostatniej warstwy sieci głębokiej [128].

Dodatkowo wykorzystano metodę przeszukiwania losowego hiperparametrów (ang. *Random Search*), gdzie wskazano zakres jednostek przedostatniej warstwy głębokiej od 256 do 1024 z krokiem co 256 oraz zmiennym współczynnikiem uczenia od 0,001 do 0,0001. Dzięki metodzie przeszukiwania losowego sprawdzono 10 prób (10 losowych zestawów hiperparametrów), spośród których otrzymano najlepsze hiperparametry dla każdej z sieci pod kątem dokładności sieci oraz minimalizacji funkcji straty. Dodatkowo zastosowano algorytm wczesnego zatrzymania (ang. *Early Stopping*) z czułością równą 3 epoki. W przypadku braku postępów w nauce modelu od trzeciej epoki przerywana jest nauka i model zachowuje najlepsze wagi w momencie przerwania nauki. Parametrem mierzalnym była wartość funkcji straty na zbiorze walidacyjnym.

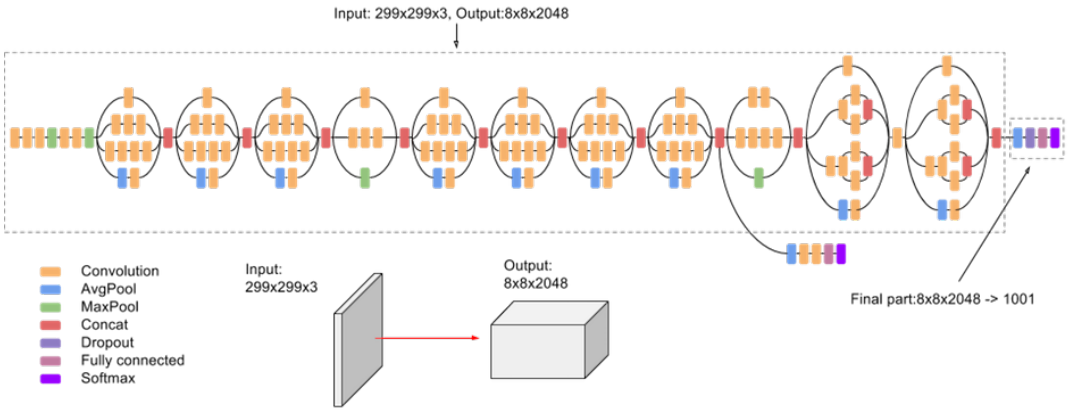
Z każdego z 924 fragmentów filmów przygotowano dane treningowe w postaci mel-spektrogramów przy pomocy biblioteki Librosa w skrypcie napisanym w języku Python.

Dla każdego z modeli obliczono dokładność na zbiorze treningowym, walidacyjnym oraz testowym. Zostały również obliczone współczynniki precyzji, czułości oraz współczynnika F1-score dla każdej z klas oraz jako średnia dla wszystkich klas.

W testach etapu pierwszego porównano ze sobą trzy modele sieci neuronowych splotowych:

Architektura InceptionV3 [46] – model ten trenowany był na obrazach o rozmiarze 299x299, stąd też takie samo wejście zastosowano podczas wykorzystania tego modelu do klasyfikacji 6 emocji. Architektura InceptionV3 posiada warstwy splotowe o rozmiarze jądra 3x3 oraz 5x5 oraz warstwy AvgPooling() i MaxPooling() służące do redukcji wielowymiarowości danych. Moduły Concat powodują konkatenację różnych wariantów splotu. Na ostatniej wprowadzono modyfikację warstwy gęstej z różnymi wartościami jednostek połączeń gęstych: 256, 512, 768 oraz 1024 w celu wyznaczenia najlepszej dokładności klasyfikacji modelu. Na rys. 6.1

przedstawiono architekturę sieci InceptionV3. Wielkość wytrenowanego najlepszego modelu pokazano w tab. 6.2.



Rys. 6.1. Architektura modelu InceptionV3 [46]

Tab. 6.2. Wielkość wytrenowanego najlepszego modelu InceptionV3

Całkowita liczba parametrów	22 380 870
Liczba parametrów trenowanych	526 000
Liczba parametrów nietrenowanych	21 802 784

W tab. 6.3 przedstawiono wyniki dla najlepszej próby.

Tab. 6.3. Wyniki treningu i testu modelu InceptionV3

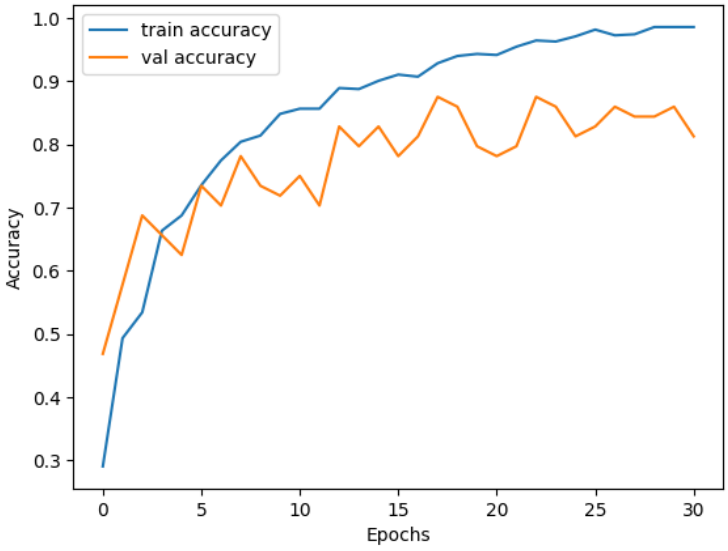
Numer próby	**Wartości hiperparametrów**	**Dokładność modelu dla danych walidacyjnych**	**Dokładność modelu dla danych testowych**
5	Jednostki: 768 Współczynnik uczenia: 0,001	0,93	0,89

Dodatkowo obliczono wartości precyzji, czułości oraz współczynnika F1-score (tab. 6.4).

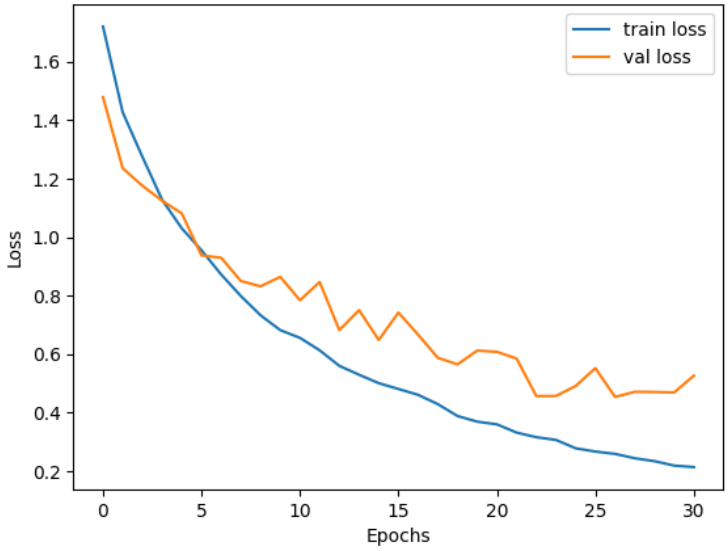
Tab. 6.4. Raport klasyfikacji dla modelu InceptionV3

Klasa	**Precyzja**	**Czułość**	**F1-Score**
Agresywny	1,00	0,91	0,95
Ekscytujący	0,85	0,88	0,86
Smutny	0,89	0,78	0,83
Spokojny	0,93	0,88	0,90
Straszny	0,91	0,91	0,91
Szczęśliwy	0,80	1,00	0,89
Średnia – macro avg	0,90	0,89	0,89

Na rys. 6.2. I 6.3 przedstawiono wykresy odpowiednio dla dokładności oraz wartości funkcji straty względem kolejnych epok uczenia. Zastosowany algorytm wczesnego zatrzymania zakończył proces uczenia w 27 epoce.

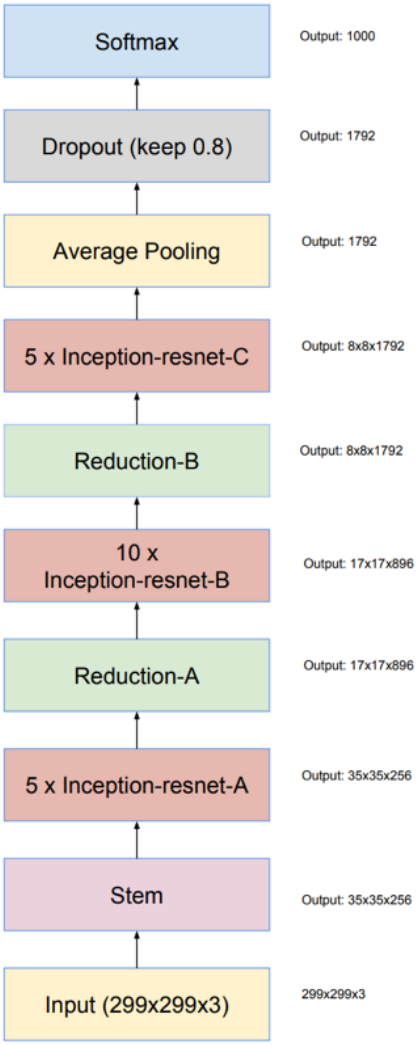


Rys. 6.2. Dokładność modelu InceptionV3 wraz z kolejnymi epokami nauki



Rys. 6.3. Wartość funkcji straty modelu InceptionV3 wraz z kolejnymi epokami nauki

InceptionResNetV2 [46] – architektura tego modelu przygotowana jest również dla danych wejściowych o rozmiarze 299x299. Architektura zawiera bloki rezydualne powielone z architektury ResNet oraz dodatkowo moduły warstw splotowych z architektury Inception. Natomiast zamieszczona warstwa Global Average Pooling dokonuje uśredniania cech przestrzennych danych. Tak, jak w poprzednim eksperymencie, również w tym przypadku zastosowano przeszukiwanie losowe hiperparametrów i dodano jako ostatnią warstwę gęstą z różnymi wartościami połączeń gęstych, od 256 do 1024 z krokiem co 256. Na rys. 6.4 przedstawiono architekturę sieci InceptionResNetV2. Szczegółowe informacje dotyczące wytrenowanego najlepszego modelu zawarto w tab. 6.5. W tab. 6.6 przedstawiono wyniki dla najlepszej próby.



Rys. 6.4. Architektura modelu InceptionResNetV2 [46]

Tab. 6.5. Wielkość wytrenowanego najlepszego modelu InceptionResNetV2

Całkowita liczba parametrów	55 126 758
Liczba parametrów trenowanych	790 022
Liczba parametrów nietrenowanych	54 336 736

Tab. 6.6. Wyniki treningu i testu modelu InceptionResNetV2

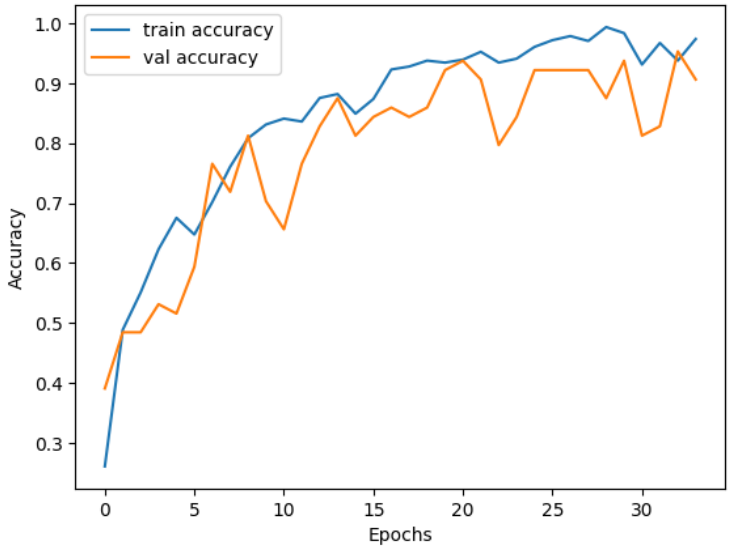
Numer próby	**Wartości hiperparametrów**	**Dokładność modelu dla danych walidacyjnych**	**Dokładność modelu dla danych testowych**
6	Jednostki: 768 Współczynnik uczenia: 0,001	0,91	0,85

Również dla modelu opartym na InceptionResNetV2 policzono wartości dla precyzji, czułości oraz współczynnika F1-score (tab. 6.7).

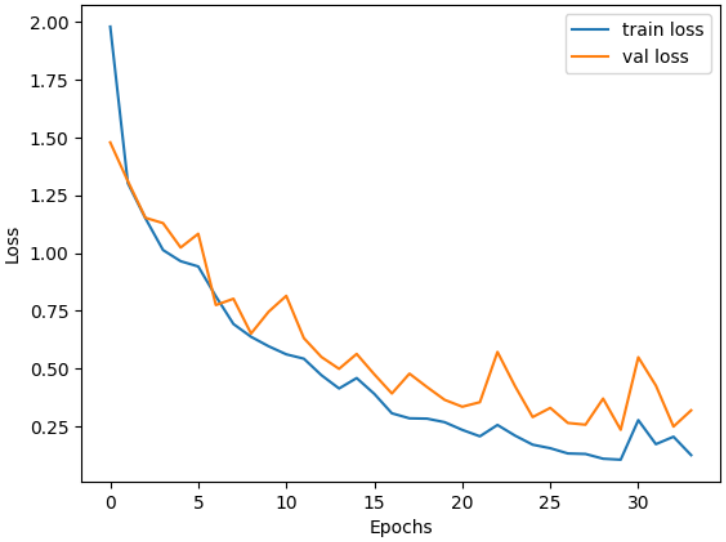
Tab. 6.7. Raport klasyfikacji dla modelu InceptionResNetV2

Klasa	**Precyzja**	**Czułość**	**F1-score**
Agresywny	1,00	0,69	0,81
Ekscytujący	0,79	0,94	0,86
Smutny	0,82	0,84	0,83
Spokojny	1,00	0,72	0,84
Straszny	0,74	1,00	0,85
Szczęśliwy	0,88	0,91	0,89
Średnia – macro avg	0,87	0,85	0,85

Na rys. 6.5 i 6.6 przedstawiono wykresy odpowiednio dla dokładności oraz wartości funkcji straty względem kolejnych epok uczenia, zastosowany algorytm wczesnego zatrzymania zakończył proces uczenia w 30 epoce.

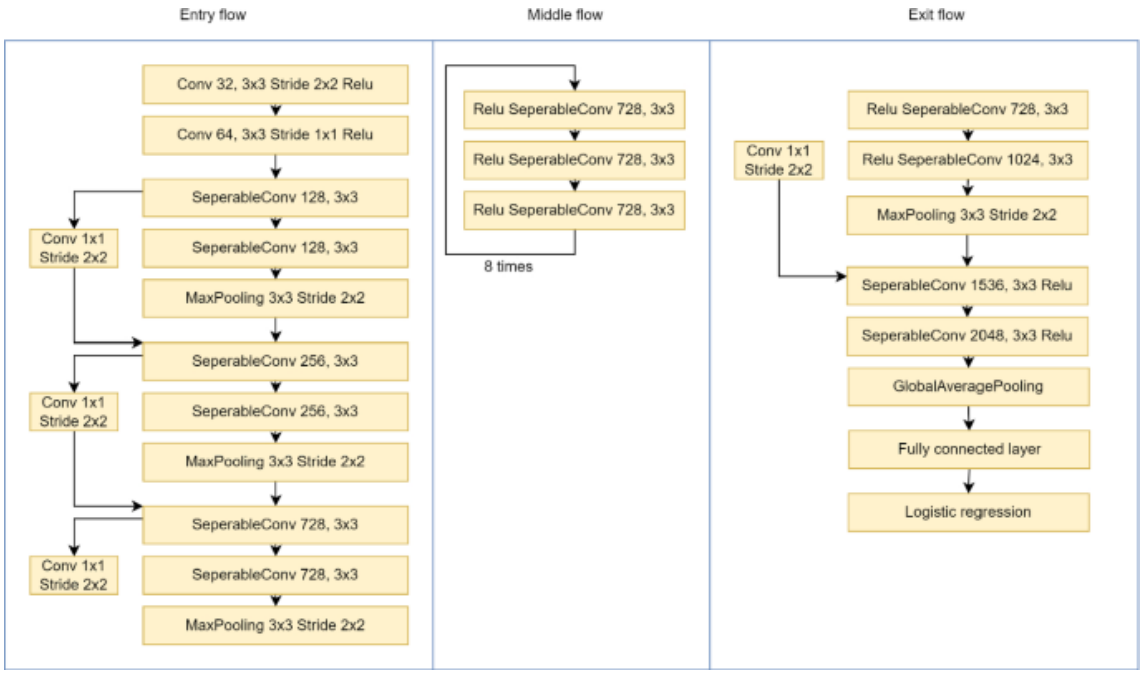


Rys. 6.5. Dokładność modelu InceptionResNetV2 wraz z kolejnymi epokami nauki



Rys. 6.6. Wartość funkcji straty modelu InceptionResNetV2 wraz z kolejnymi epokami nauki

Xception [46] – warstwa wejściowa również jest ustalona tak jak w poprzednich modeli na rozmiar obrazka 299x299. Extreme Inception, została zaproponowana jako skomplikowana architektura o dużych możliwościach klasyfikacji w zadaniach widzenia komputerowego. Koncepcja modelu wykorzystuje technikę splotów rozdzielnych. Technika ta pozwala na osobną naukę funkcji przestrzennych, jak i kanałowych, co skutkuje wydajniejszym procesem uczenia się modelu. Sploty głęboko separowalne zmniejszają liczbę parametrów modelu oraz poprawiają jego możliwości pod kątem wychwycenia bardziej złożonych wzorców danych. Poza warstwami splotów rozdzielnych architektura posiada bloki resztkowe oraz bloki modelu Inception, które łączą operacje splotu dla różnych wielkości jąder. Na rys. 6.7 zaprezentowano architekturę modelu Xception.



Rys. 6.7. Architektura modelu Xception [46]

Wielkość wytrenowanego najlepszego modelu zawarto w tab. 6.8. Tab. 6.9 przedstawia wyniki treningu i testu modelu Xception, zaś w tab. 6.10 zamieszczono raport klasyfikacji dla modelu Xception.

Tab. 6.8. Wielkość wytrenowanego najlepszego modelu Xception

Całkowita liczba parametrów	21 387 566
Liczba parametrów trenowanych	526 086
Liczba parametrów nietrenowanych	20 861 480

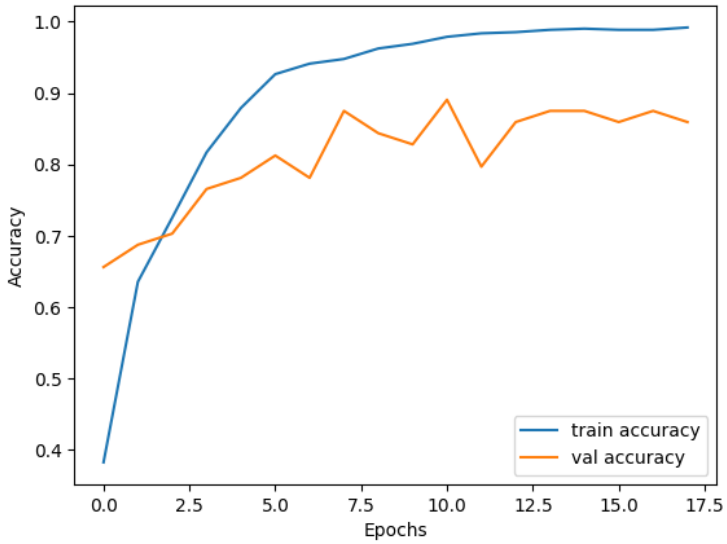
Tab. 6.9. Wyniki treningu i testu modelu Xception

Numer próby	**Wartości hiperparametrów**	**Dokładność modelu dla danych walidacyjnych**	**Dokładność modelu dla danych testowych**
1	Jednostki: 512 Współczynnik uczenia: 0,001	0,91	0,86

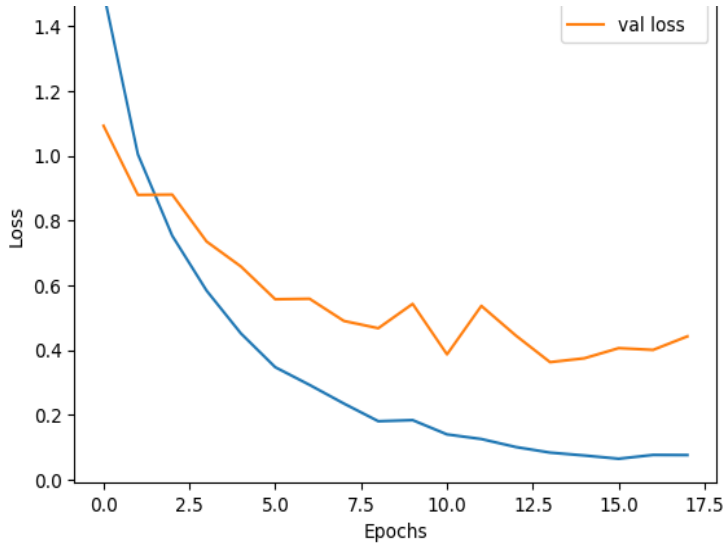
Tab. 6.10. Raport klasyfikacji dla modelu Xception

Klasa	**Precyzja**	**Czułość**	**F1-score**
Agresywny	0,93	0,88	0,90
Ekscytujący	0,84	0,84	0,84
Smutny	0,90	0,84	0,87
Spokojny	0,82	0,84	0,83
Straszny	0,78	0,91	0,84
Szczęśliwy	0,93	0,88	0,90
Średnia – macro avg	0,87	0,86	0,87

Na rys. 6.8 i 6.9 przedstawiono wykresy odpowiednio dla dokładności oraz wartości funkcji straty względem kolejnych epok uczenia, zastosowany algorytm wczesnego zatrzymania zakończył proces uczenia w 14 epoce.



Rys. 6.8. Dokładność modelu Xception wraz z kolejnymi epokami nauki



Rys. 6.9. Wartość funkcji straty modelu Xception wraz z kolejnymi epokami nauki

6.3.2. *Klasyfikacja emocji na podstawie parametrów MFCC ścieżki audio fragmentów filmu*

Do klasyfikacji sześciu modeli emocji na podstawie parametrów MFCC użyto sieci rekurencyjnej opartej o rodzaj sieci GRU. Naukę oparto na prostym modelu sieci, zawierającym jedynie 51 tysięcy parametrów, co w porównaniu do wcześniej wykorzystanych modeli jest znaczącym zmniejszeniem struktury i parametrów sieci neuronowej.

Dodatkowo użyto techniki redukcji liczby neuronów (ang. *dropout*) o 20%, model wykorzystuje trzy warstwy oraz zastosowano również algorytm przeszukiwania losowego hiperparametrów:

- Warstwa pierwsza: GRU – sprawdzono różną liczbę jednostek warstwy GRU, która odpowiedzialna jest za pojemność modelu, do każdej z prób losowo były wybierane wartości od 32 do 128 z krokiem co 32 jednostki;
- Warstwa druga: Dropout = 0,2;

- Warstwa trzecia: warstwa głęboka z funkcją aktywacji softmax w celu predykcji klasy emocji.

W tab. 6.11 przedstawiono informacje dotyczące wytrenowanego najlepszego modelu GRU dla nauki parametrami MFCC.

Tab. 6.11. Wielkość wytrenowanego najlepszego modelu GRU dla nauki parametrami MFCC

Całkowita liczba parametrów	51 078
Liczba parametrów trenowanych	51 078
Liczba parametrów nietrenowanych	0

Sprawdzono również różne wartości współczynnika uczenia począwszy od 0,00001 do 0,001. Aby zapobiec nadmiernej nauce sieci, ustanowiono algorytm wczesnego przerwania treningu (ang. *Early Stopping*). Tak, jak w poprzednim przypadku uczenia sieci splotowych, algorytm posiadał czułość na poziomie 3 epok, dzięki temu proces treningu zatrzymał się dla najlepszych hiperparametrów w 32 epoce. Dla modelu o najlepszych wartościach hiperparametrów obliczono zarówno dokładność dla zbioru testowego, jak i czułość, precyzję oraz współczynnik F1-score.

W tab. 6.12 zawarto wynik ewaluacji modelu oraz hiperparametry dla najlepszego modelu uzyskanego podczas 10 prób przeszukiwania losowego. Tabela 6.13 przedstawia raport klasyfikacji dla poszczególnych klas. Rysunek 6.11 i rys. 6.12 przedstawiają wartość funkcji straty oraz wartość dokładności dla kolejnych epok nauki.

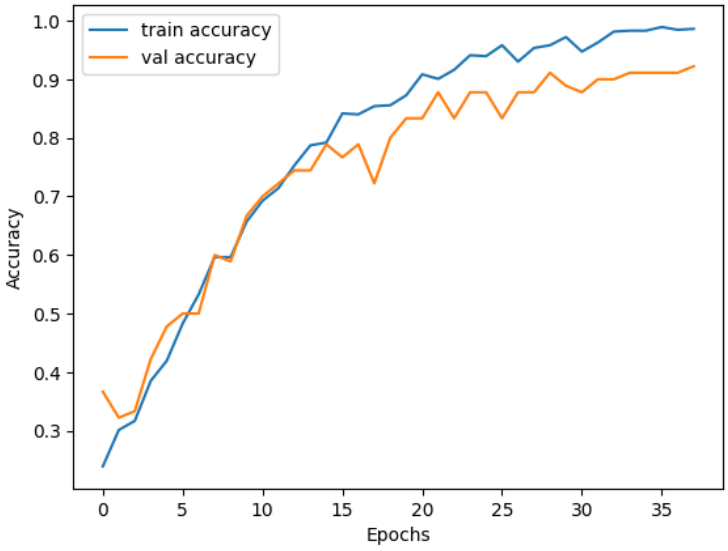
Tab. 6.12. Wyniki treningu i testu modelu GRU dla nauki parametrami MFCC

Numer próby	**Wartości hiperparametrów**	**Dokładność modelu dla danych walidacyjnych**	**Dokładność modelu dla danych testowych**
9	Jednostki: 128 Współczynnik uczenia: 0,001	0,92	0,90

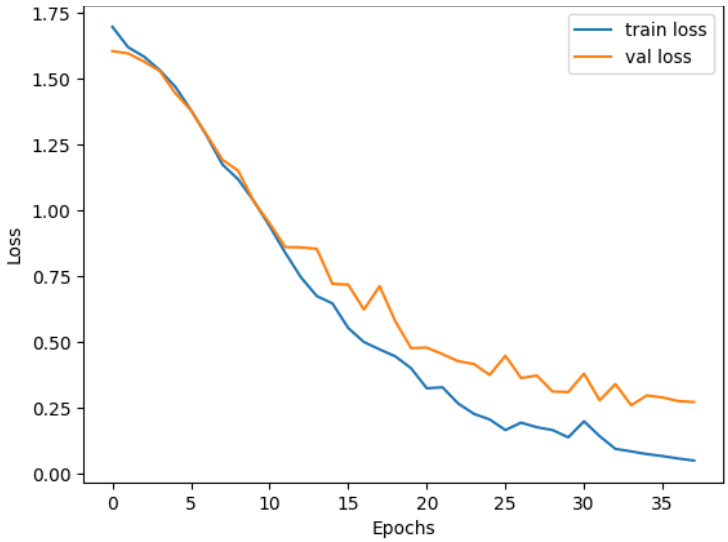
Tab. 6.13. Raport klasyfikacji dla modelu GRU dla nauki parametrami MFCC

Klasa	**Precyzja**	**Czułość**	**F1-score**
Agresywny	0,85	0,81	0,88
Ekscytujący	0,91	0,94	0,92
Smutny	0,94	0,91	0,92
Spokojny	0,85	0,88	0,86
Straszny	0,97	0,88	0,92
Szczęśliwy	1,00	1,00	1,00
Średnia – macro avg	0,92	0,92	0,92

Na rys. 6.10 i 6.11 przedstawiono wyniki dokładności modelu oraz wartość funkcji straty względem kolejnych epoko nauki sieci.



Rys. 6.10. Dokładność modelu GRU dla nauki parametrami MFCC wraz z kolejnymi epokami nauki



Rys. 6.11. Wartość funkcji straty modelu GRU dla nauki parametrami MFCC wraz z kolejnymi epokami nauki

6.3.3. Klasyfikacja emocji na podstawie parametrów OPENSmile ścieżki audio fragmentów filmu

W badanym przypadku na wejście sieci neuronowej przygotowano zestaw parametrów, które uzyskano poprzez metody wdrożone z biblioteki OPENSmile do analizy plików dźwiękowych, które podano na wejście identycznej sieci rekurencyjnej, przedstawionej w poprzednim teście. Każdy zestaw parametrów podzielony został na parametry funkcjonalne oraz parametry niskopoziomowe, w zależności o wyboru zestawu zmienia się liczbę parametrów, które można uzyskać podczas analizy narzędziem OPENSmile. W tab. 6.14. przedstawiono zestawy parametrów dla narzędzia OPENSmile [80].

Tab. 6.14. Zestawy parametrów dla narzędzia parametryzacji OPENSmile [80]

Nazwa zestawu parametrów	Liczba parametrów niskopoziomowych / funkcjonalnych
ComParE_2016	65 / 6373
GeMAPSv01a	18 / 62
GeMAPSv01b	18 / 62
eGeMAPSv01a	23 / 88
eGeMAPSv01b	23 / 88
eGeMAPSv02	25 / 88

Do analizy plików audio posłużono się zestawem parametrów eGeMAPSv02. Dla parametrów niskopoziomowych wyznaczono między innymi (całkowita suma parametrów niskopoziomowych wynosi 25, które są rozłożone w czasie i dotyczą fragmentów):

- Ton podstawowy F0;
- Współczynniki mel-cepstralne;
- Zawartość energetyczna w różnych pasmach częstotliwości;
- Zawartość harmoniczną;
- Spektralne cechy dźwięku;
- Częstotliwości dźwięku o niskiej amplitudzie.

Dla parametrów funkcyjnych ze zbioru eGeMAPSv02 można zaliczyć te, które odnoszą się do parametrów niskopoziomowych (łączna ich suma wynosi 88):

- Średnia wartość poszczególnych parametrów niskopoziomowych;
- Odchylenie standardowe dla poszczególnych parametrów niskopoziomowych;
- Maksimum oraz minimum parametrów niskopoziomowych;
- Entropia;
- Zmiana parametrów niskopoziomowych w czasie;
- Inne.

Założenia nauki modelu zostały powielone z poprzedniego przykładu, natomiast dane audio zostały poddane dwukrotnej analizie. W tym przypadku zostały przygotowane dwa modele, jeden oparty o parametry niskopoziomowe, drugi o parametry funkcyjne.

W tab. 6.15 przedstawiono całkowitą liczbę parametrów modelu. Tabela 6.16 zawiera wyniki ewaluacji najlepszego modelu oraz jego hiperparametry, zaś tab. 6.17 przedstawia raport klasyfikacji dla poszczególnych klas emocji. Na rys. 6.13 oraz 6.14 zawarto wykresy wartości funkcji straty oraz wartości dokładności dla kolejnych epok nauki modelu.

Tab. 6.15. Wielkość wytrenowanego najlepszego modelu GRU dla nauki parametrami niskopoziomowymi eGeMAPSv02

Całkowita liczba parametrów	53 568
Liczba parametrów trenowanych	53 568
Liczba parametrów nietrenowanych	0

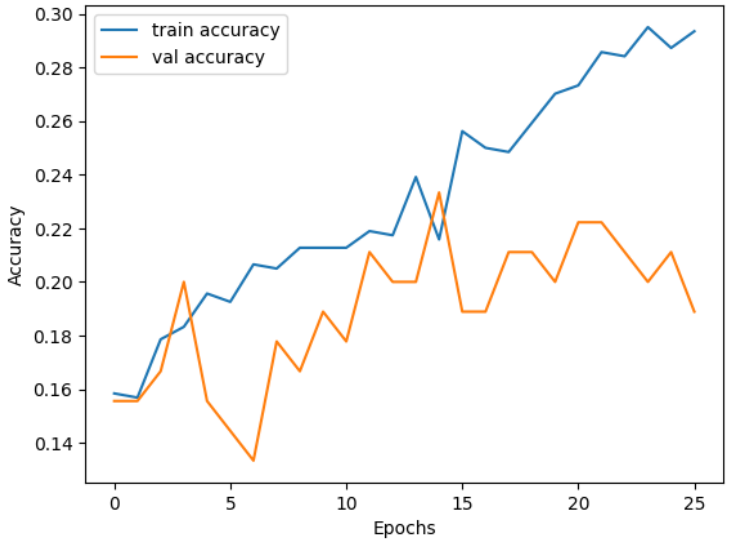
Tab. 6.16. Wyniki treningu i testu modelu GRU dla nauki parametrami niskopoziomowymi eGeMAPSv02

Numer próby	**Wartości hiperparametrów**	**Dokładność modelu dla danych walidacyjnych**	**Dokładność modelu dla danych testowych**
9	Jednostki: 128 Współczynnik uczenia: 0,0001	0,32	0,24

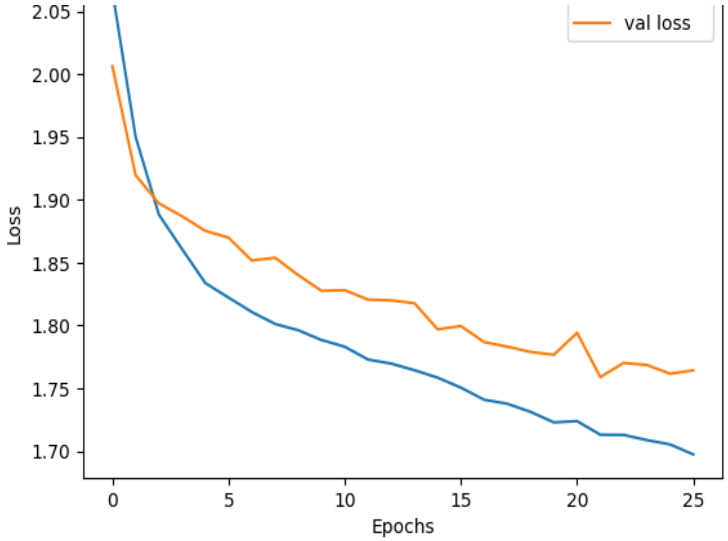
Tab. 6.17. Raport klasyfikacji dla modelu GRU dla nauki parametrami niskopoziomowymi eGeMAPSv02

Klasa	**Precyzja**	**Czułość**	**F1-score**
Agresywny	0,11	0,13	0,12
Ekscytujący	0,38	0,39	0,38
Smutny	0,24	0,28	0,26
Spokojny	0,35	0,38	0,36
Straszny	0,17	0,16	0,16
Szczęśliwy	0,23	0,16	0,19
Średnia – macro avg	0,25	0,25	0,25

Model − poprzez użycie techniki wczesnego zatrzymania − zakończył naukę w 25 epoce. Na rys. 6.12 i 6.13 przedstawiono dokładność oraz wartość funkcji straty modelu.



Rys. 6.12. Dokładność modelu GRU dla nauki parametrami niskopoziomowymi eGeMAPSv02 wraz z kolejnymi epokami nauki



Rys. 6.13. Wartość funkcji straty modelu GRU dla nauki parametrami niskopoziomowymi eGeMAPSv02 wraz z kolejnymi epokami nauki

Drugim podejściem w analizie cech wydobytych przy pomocy narzędzia OPENSmile był trening modelu GRU przy użyciu cech funkcjonalnych (88 cech). Poniżej w tab. 6.18 przedstawiono liczbę parametrów wytrenowanego modelu. Z kolei tab. 6.19 i 6.20 przedstawiają kolejno wyniki ewaluacji modelu oraz hiperparametry dla najlepszego modelu uzyskane poprzez mechanizm przeszukiwania losowego oraz raport klasyfikacji dla tego modelu.

Tab. 6.18. Wielkość wytrenowanego najlepszego modelu GRU dla nauki parametrami funkcjonalnymi eGeMAPSv02

Całkowita liczba parametrów	59 520
Liczba parametrów trenowanych	59 520
Liczba parametrów nietrenowanych	0

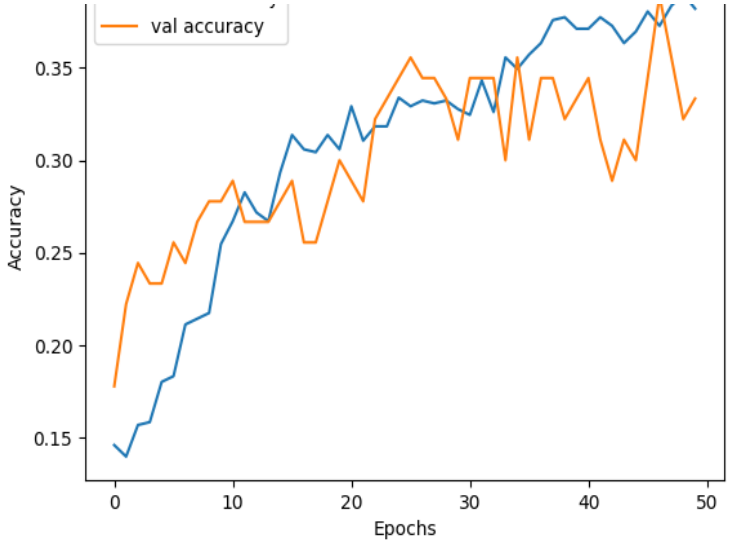
Tab. 6.19. Wyniki treningu i testu modelu GRU dla nauki parametrami funkcjonalnymi eGeMAPSv02

Numer próby	**Wartości hiperparametrów**	**Dokładność modelu dla danych walidacyjnych**	**Dokładność modelu dla danych testowych**
9	Jednostki: 128 Współczynnik uczenia: 0,001	0,38	0,37

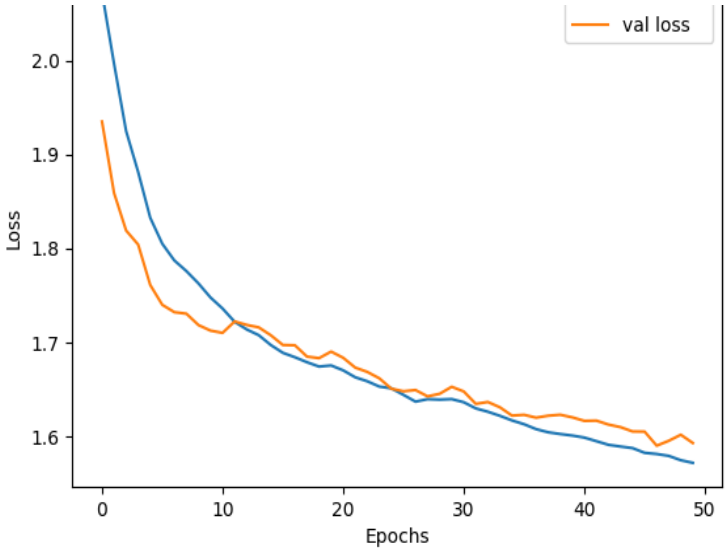
Tab. 6.20. Raport klasyfikacji dla modelu GRU dla nauki parametrami funkcjonalnymi eGeMAPSv02

Klasa	**Precyzja**	**Czułość**	**F1-score**
Agresywny	0,32	0,39	0,35
Ekscytujący	0,28	0,23	0,25
Smutny	0,47	0,62	0,53
Spokojny	0,47	0,66	0,58
Straszny	0,21	0,16	0,18
Szczęśliwy	0,40	0,19	0,26
Średnia – macro avg	0,36	0,37	0,35

Model − poprzez użycie techniki wczesnego zatrzymania – zakończył naukę w 47 epoce. Rysunki 6.14 oraz 6.15 przedstawiają kolejno wyniki wartości funkcji straty oraz dokładności dla wybranego modelu.



Rys. 6.14. Dokładność modelu GRU dla nauki parametrami funkcjonalnymi eGeMAPSv02 wraz z kolejnymi epokami nauki



Rys. 6.15. Wartość funkcji straty modelu GRU dla nauki parametrami funkcjonalnymi eGeMAPSv02 wraz z kolejnymi epokami nauki

6.4. Klasyfikacja emocji z fragmentu filmu z wykorzystaniem uczenia głębokiego

6.4.1. Klasyfikacja emocji na podstawie ekstrakcji parametrów poprzez sieć CNN z fragmentu filmu

W kolejnym teście modeli sieci głębokich do zadania klasyfikacji sześciu emocji użyto architektury sieci łączonych, sieć splotowa odpowiadała za ekstrakcje cech z wideo, które następnie zostały przekazane na wejście sieci GRU, gdzie dokonana została klasyfikacja emocji na podstawie fragmentów filmów.

Do ekstrakcji cech została użyta gotowa sieć splotowa InceptionV3, która pierwotnie trenowana była na dużym zbiorze danych w postaci obrazów ImageNet. Sieć splotowa InceptionV3 jest siecią dwuwymiarową, dlatego wcześniej przeprowadzono wstępne przetwarzanie danych filmowych, aby na wejście sieci podać ramki obrazu wideo. W tab. 6.21 podano liczbę parametrów dla sieci splotowej Inception V3. Dodatkowo, aby dane nie zajmowały zbyt dużo pamięci RAM, podczas ekstrakcji cech wideo z sieci splotowej, użyto jedynie pierwszych 40 ramek każdego z fragmentów filmu oraz ekstrakcji 256 cech dla każdej ramki obrazu. Ekstrakcję ramek z fragmentów filmów przygotowano za pomocą biblioteki OpenCV, następnie dla każdej ramki obrazu wykonywana była ich analiza. Dla każdego fragmentu filmu powstała macierz o wielkości 40x256 (liczba ramek x liczba wyekstrahowanych cech obrazu).

Następnie obliczone parametry zostały przekazane do rekurencyjnej jednowarstwowej sieci GRU, która odpowiada za klasyfikację fragmentów filmów do odpowiedniej emocji. W tym przypadku treningu sieci rekurencyjnej również skorzystano z metody dopasowania najlepszych hiperparametrów, jaką jest *Random Search*, trening sieci ograniczono do 10 prób, w których losowo dobierano parametry nauki sieci, jakimi była ilość jednostek sieci z zakresu od 32 do 128 z krokiem co 32 jednostki oraz współczynnik uczenia: 0,00001, 0,0001, 0,001. W tabeli 6.22 wskazano liczbę parametrów dla najlepszego modelu sieci GRU.

Tab. 6.21. Całkowita liczba parametrów siec splotowej InceptionV3 służącej do ekstrakcji cech obrazu wideo

Całkowita liczba parametrów	22 327 328
Liczba parametrów trenowanych	22 292 896
Liczba parametrów nietrenowanych	34 432

Tab. 6.22. Całkowita liczba parametrów siec rekurencyjnej GRU

Całkowita liczba parametrów	102 534
Liczba parametrów trenowanych	102 534
Liczba parametrów nietrenowanych	0

W tab. 6.23 przedstawiono wyniki dla ewaluacji modelu dla najlepszego modelu wybranego podczas przeszukiwania losowego. Z kolei, tab. 2.24 zawiera raport klasyfikacji dla modelu GRU,

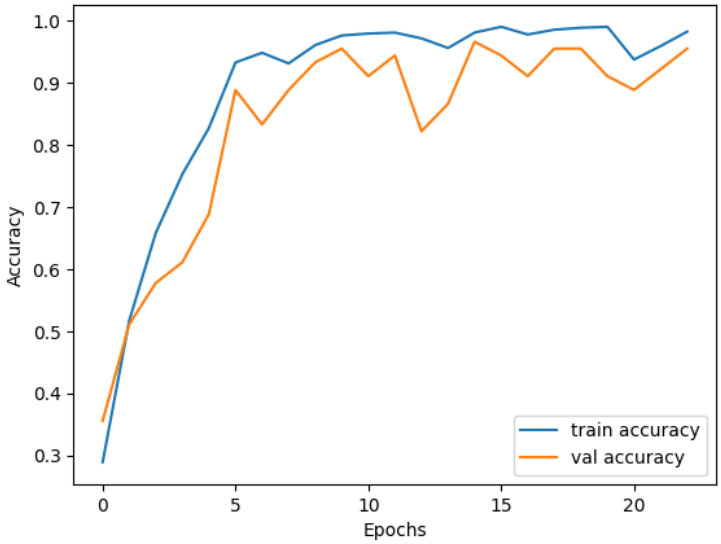
Tab. 6.23. Wyniki treningu i testu modelu GRU z parametrów wyekstrahowanych z sieci splotowej

Numer próby	**Wartości hiperparametrów**	**Dokładność modelu dla danych walidacyjnych**	**Dokładność modelu dla danych testowych**
1	Jednostki: 96 Współczynnik uczenia: 0,001	0,97	0,16

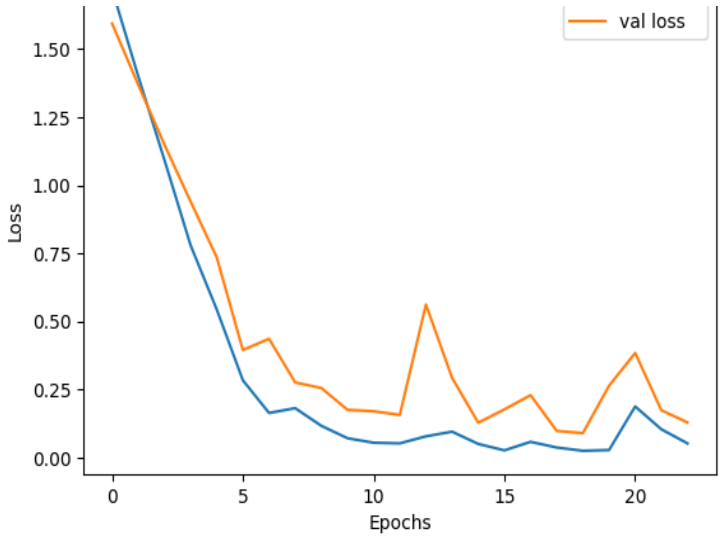
Tab. 6.24. Raport klasyfikacji dla modelu GRU dla nauki parametrami wyekstrahowanych z sieci splotowej

Klasa	**Precyzja**	**Czułość**	**F1-score**
Agresywny	0,00	0,00	0,00
Ekscytujący	0,17	0,03	0,05
Smutny	0,00	0,00	0,00
Spokojny	0,00	0,00	0,00
Straszny	0,17	0,91	0,28
Szczęśliwy	0,00	0,00	0,00
Średnia – macro avg	0,06	0,16	0,06

Model − poprzez użycie techniki wczesnego zatrzymania − zakończył naukę w 20 epoce. Na rys. 6.16 i rys. 6.17 przedstawiono wartości funkcji straty oraz dokładności modelu w trakcie nauki.



Rys. 6.16. Dokładność modelu GRU dla nauki parametrami obrazu wyekstrahowanych siecią splotową



Rys. 6.17. Wartość funkcji straty dla modelu GRU dla nauki parametrami obrazu wyekstrahowanych siecią splotową

6.4.2. Klasyfikacja emocji poprzez sieć GRU na podstawie wartości histogramów RGB z obrazu wideo

Histogramy zostały przygotowane za pomocą biblioteki OpenCV. Podobnie jak w poprzednim przypadku do realizacji zadania klasyfikacji wybrano 40 pierwszych klatek każdego filmu na podstawie, których został wygenerowane histogramy. Aby zachować niski poziom skomplikowania sieci GRU podczas przetworzenia danych, zredukowano rozmiar każdej klatki do wymiaru 256x256 pikseli, a następnie tak przygotowane klatki filmów poddano wygenerowaniu histogramów kolorów RGB. Sieci rekurencyjne przyjmują dane sekwencyjne, dlatego też nie zdecydowano się na obliczenie histogramu z uśrednionych wartości histogramów dla

poszczególnych klatek całego fragmentu filmu, a na wejście sieci neuronowej przekazano wektor wartości histogramu dla pojedynczych klatek w formie sekwencji.

Wartości histogramu przed wejściem na sieć neuronową zostały znormalizowane oraz spłaszczone do jednego wymiaru, gdyż pierwotnie histogram został przygotowany jako macierz 3D zawierająca każdy kanał RGB osobno.

Podobnie jak w poprzednich eksperymentach wybór odpowiednich hiperparametrów oparto na algorytmie Random Search dla 10 prób. W tab. 6.25 przedstawiono wyniki dla najlepszej próby, w której uzyskano najwyższą wartość dokładności modelu dla danych walidacyjnych.

Tab. 6.25. Całkowita liczba parametrów modelu GRU opartą o klasyfikację emocji na podstawie histogramów RGB

Całkowita liczba parametrów	87 558
Liczba parametrów trenowanych	87 558
Liczba parametrów nietrenowanych	0

Tabela 6.26 przedstawia wyniki ewaluacji modelu GRU dla nauki wartościami histogramów RGB.

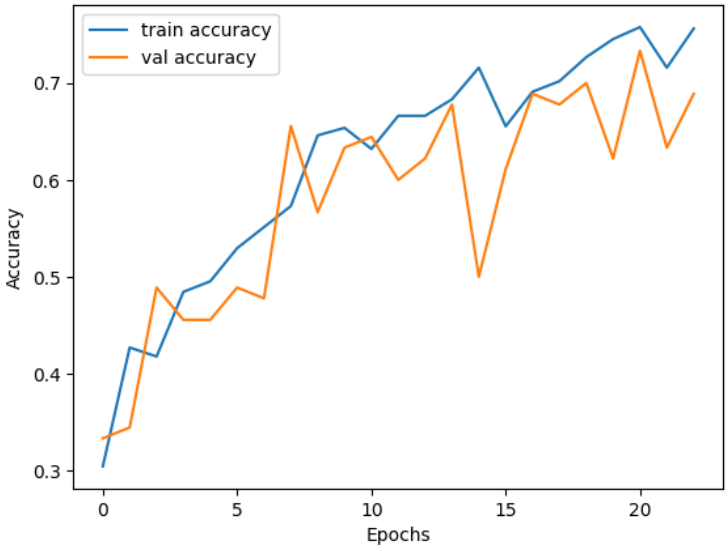
Tab. 6.26. Wyniki treningu i testu modelu GRU opartą o klasyfikację emocji na podstawie histogramów RGB

Numer próby	**Wartości hiperparametrów**	**Dokładność modelu dla danych walidacyjnych**	**Dokładność modelu dla danych testowych**
1	Jednostki: 96 Współczynnik uczenia: 0,001	0,68	0,70

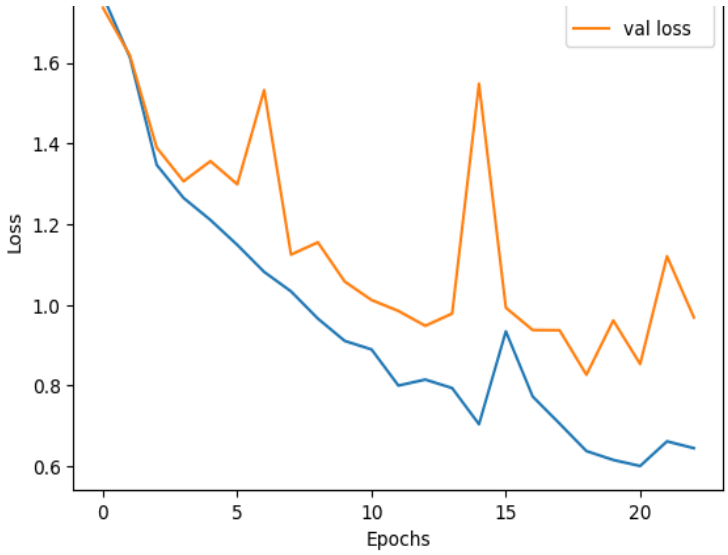
Tab. 6.27. Raport klasyfikacji dla modelu GRU dla nauki wartościami histogramów RGB

Klasa	**Precyzja**	**Czułość**	**F1-score**
Agresywny	0,95	0,58	0,72
Ekscytujący	0,58	0,71	0,64
Smutny	0,57	0,66	0,61
Spokojny	0,69	0,62	0,66
Straszny	0,88	0,88	0,88
Szczęśliwy	0,71	0,78	0,75
Średnia – macro avg	0,73	0,70	0,70

Model − poprzez użycie techniki wczesnego zatrzymania − zakończył naukę w 22 epoce, na rys. 6.18 oraz 6.19 przedstawiono wyniki wartości funkcji straty oraz dokładności modelu podczas treningu w kolejnych epokach.



Rys. 6.18. Dokładność modelu GRU dla nauki wartościami histogramów RGB



Rys. 6.19. Wartość funkcji straty dla modelu GRU dla nauki wartościami histogramów RGB

6.4.3. Klasyfikacja emocji za pomocą architektury 3D sieci splotowej

Ostatnią testowaną architekturą była architektura trójwymiarowej sieci splotowej. Sieci splotowe 3D (Conv3D) znajdują zastosowania w analizie obrazu wideo ze względu na możliwość analizy obiektów zarówno w wartościach przestrzennych, jak i czasowych. Conv3D przyjmują wartości trójwymiarowe, co jest typowe dla danych sekwencyjnych, w przypadku analizy wideo dane te posiadają wartości dotyczące czasu, wysokości, szerokości, liczby kanałów. Podobnie jak w przypadku zwykłych dwuwymiarowych sieci splotowych, także trójwymiarowe sieci splotowe posiadają jądra splotowe. Jednak w tym przypadku jądra te przesuwają się w trzech wymiarach (szerokość, wysokość oraz czas), aby wydobyć parametry przestrzenne i czasowe.

Dzięki tym zaletom przygotowano model sieci CNN 3D w oparciu o pracę zespołu Vrskrovej [119], który zaprezentował architekturę sieci splotowej trójwymiarowej (patrz rozdz. 4.4.3 − Wykorzystanie metod uczenia głębokiego do klasyfikacji emocji z fragmentów filmu).

Sieci splotowe 3D wymagają dużej ilości zasobów pamięciowych, dlatego też ograniczono rozmiar danych poprzez redukcję wymiaru filmu, każdą ramkę filmu zmniejszono do rozmiaru

32x32 piksele. Podział na zbiór treningowy, walidacyjny i testowy pozostał taki sam jak w poprzednich badaniach innych architektur i metod klasyfikacji.

W tab. 6.28 przedstawiono wyniki dla najlepszego modelu wybranego spośród 10 prób algorytmem przeszukiwania losowego hiperparametrów.

Tab. 6.28. Całkowita liczba parametrów modelu CNN3D

Całkowita liczba parametrów	3 612 358
Liczba parametrów trenowanych	3 611 206
Liczba parametrów nietrenowanych	1152

W tabeli 6.29 przedstawiono wyniki ewaluacji oraz hiperparametry dla najlepszego modelu wyznaczonego poprzez przeszukiwanie losowe.

Tab. 6.29. Wyniki treningu i testu modelu CNN3D

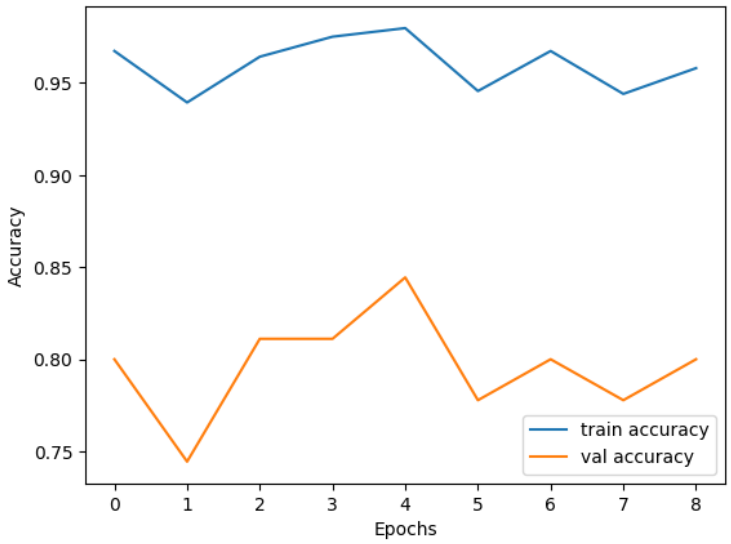
Numer próby	**Warstwa**	**Wartości hiperparametrów**	**Dokładność modelu dla danych walidacyjnych**	**Dokładność modelu dla danych testowych**
10	Cov3D_1	Liczba filtrów: 32 Rozmiar jądra: 5x5x5	0,83	0,87
	Cov3D_2	Liczba filtrów: 32 Rozmiar jądra: 5x5x5		
	Cov3D_3	Liczba filtrów: 96 Rozmiar jądra: 5x5x5		
	Cov3D_4	Liczba filtrów: 128 Rozmiar jądra: 5x5x5		
	Cov3D_5	Liczba filtrów: 192 Rozmiar jądra: 5x5x5		
	Cov3D_6	Liczba filtrów: 384 Rozmiar jądra: 5x5x5		
	Dense	Liczba jednostek: 128		
	Dropout	0,4		

Tabela 6.30 przedstawia raport klasyfikacji dla modelu trójwymiarowej sieci splotowej.

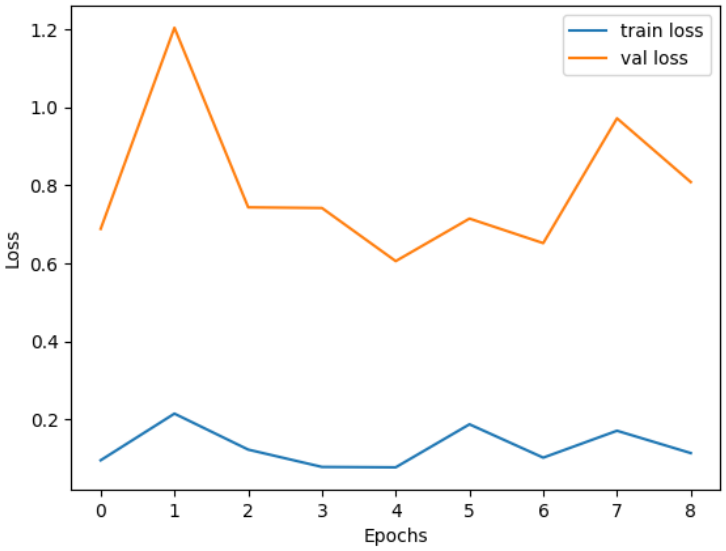
Tab. 6.30. Raport klasyfikacji dla modelu 3D CNN

Klasa	**Precyzja**	**Czułość**	**F1-score**
Agresywny	0,94	0,94	0,94
Ekscytujący	0,83	0,81	0,82
Smutny	0,87	0,84	0,86
Spokojny	0,89	0,78	0,83
Straszny	0,90	0,84	0,87
Szczęśliwy	0,80	1,00	0,89
Średnia – macro avg	0,87	0,87	0,87

Model − poprzez użycie techniki wczesnego zatrzymania treningu − zakończył naukę w 5 epoce. Na rys. 6.20 oraz 6.21 przedstawiono zmiany wartości funkcji straty oraz wartość dokładności modelu dla kolejnych epok uczenia.



Rys. 6.20. Dokładność modelu CNN3D



Rys. 6.21. Wartość funkcji straty dla modelu CNN3D

6.5. Klasyfikacja emocji z fragmentu filmu wraz z towarzyszącą ścieżką dźwiękową z wykorzystaniem uczenia głębokiego

Do klasyfikacji filmu na podstawie danych audiowizualnych użyto metodologii fuzji dwóch wcześniej przygotowanych oraz przetestowanych modeli sieci neuronowych. W tym przypadku wykorzystano dwa oddzielne modele dla sygnału audio oraz sygnału wideo. Następnie wykorzystano metodę konkatenacji wyników do jednego wspólnego wektora i poddano ponownej klasyfikacji, jest to jedna z metod łączenia wyników klasyfikacji dla modeli multimodalnych.

Metody łączenia w zespoły klasyfikatorów (ang. *Ensemble Learning*) wyróżniają różne techniki [14, 64, 73, 129]:

- *Bagging* – technika ta polega na trenowaniu kilku kopii modeli na różnych próbkach następnie łączenie wyników poprzez głosowanie lub uśrednianie wyników;
- *Boosting* – modele trenowane są sekwencyjnie, a każdy poprzedni model skupia się na poprawie wyników poprzedniego;
- *Stacking* – wyniki trenowanych wcześniej modeli przekazywane stają się danymi i przekazywane są jako dane wejściowe do nowego modelu, który dokonuje klasyfikacji;
- *Voting* – modele dokonują wstępnej predykcji, a ostateczna decyzja jest podejmowana na podstawie głosowania.

Ponadto ensemble learning dzieli się na techniki homogeniczne oraz heterogeniczne. Podczas korzystania z technik homogenicznych zespoły modeli stworzone są na bazie tych samych architektur. Techniki heterogeniczne, są technikami, gdzie w komitecie używa się różnych modeli, oznacza to, że modele różnią się między sobą architekturą, danymi wejściowymi oraz metodyką uczenia się. W pracy przedstawiono dla wcześniej przygotowanych modeli oraz różnych metod przetwarzania danych metodykę łączenia modeli (ang. *stacking*).

Pierwszym etapem procesu klasyfikacji dla modeli multimodalnych jest przetworzenie danych wejściowych dla każdego z początkowych modeli z osobna, gdzie każdy model przygotowuje dane wejściowe na swój unikalny sposób. Dla metody *stackingu* kolejnym krokiem jest fuzja modeli, gdzie wyniki z poprzednich modeli łączone są i przekazywane jako wejście do nowego modelu bądź warstwy decyzyjnej. Ostatecznie przetworzony i połączony wektor cech z etapu fuzji jest używany do klasyfikacji etykiet.

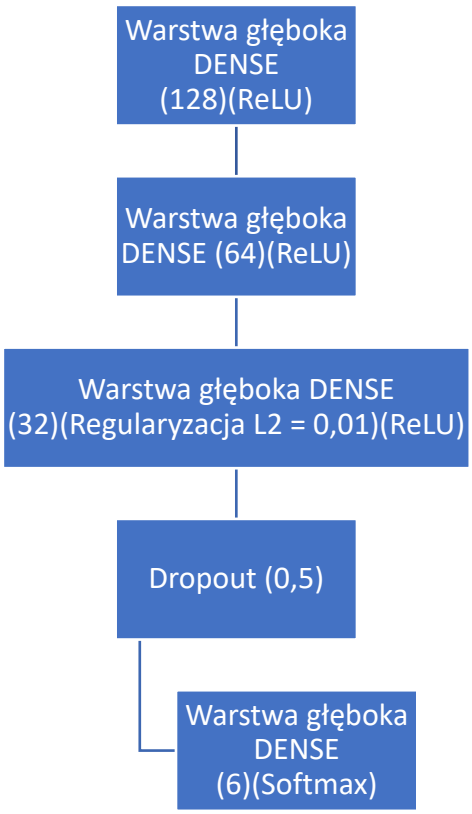
Na podstawie przeprowadzonych badań (rozdział 6.3 i 6.4) wybrano dwa najlepsze modele pod katem skuteczności dla danych testowych oraz wielkości modelu, aby docelowo zachować optymalny model pod kątem zarówno klasyfikacji, jak i potrzebnych zasobów do przetwarzania danych. W tab. 6.31 przedstawiono dwa wybrane modele dla modalności audio oraz wideo

Dla danych audio zdecydowano się użyć modelu wykorzystującego parametryzację sygnału audio, tj. współczynniki MFCC, a następnie wykorzystanie modelu GRU o wcześniej wskazanych hiperparametrach do utworzenia wektora cech sygnału audio (patrz tab. 6.12). Dla danych wideo wybrano model oparty na modelu trójwymiarowej sieci splotowej.

Następnie oba modele poddano metodzie konkatenacji, gdzie wektor parametrów przekazano na wejście sieci składającej się z trzech warstw gęstych. Dodatkowo sieć została zmodyfikowana o techniki regularyzacji dropout oraz L2 (patrz rys. 6. 22).

Tab. 6.31. Wyniki wybranych najlepszych modeli dla modelu multimodalnego

Model	Liczba parametrów	Dokładność dla zbioru walidacyjnego	Dokładność dla zbioru testowego
GRU MFCC – model audio	51 078	0,92	0,90
3D CNN – model wideo	3 612 358	0,83	0,87



Rys. 6.22. Schemat sieci głębokiej do klasyfikacji multimodalnej

W tab. 6.32 przedstawiono wielkość sieci multimodalnej oraz jej wyniki klasyfikacji sześciu klas emocji.

Tab. 6.32. Całkowita liczba parametrów modelu sieci multimodalnej w metodzie fuzji

Całkowita liczba parametrów	3 746 878
Liczba parametrów trenowanych	3 745 798
Liczba parametrów nietrenowanych	1080

Tabela 6.33 przedstawia wyniki dokładności siec bimodalnej w metodzie fuzji obu modalności: audio i wideo. Tab. 6.34 zawiera raport klasyfikacji dla modelu multimodalnego.

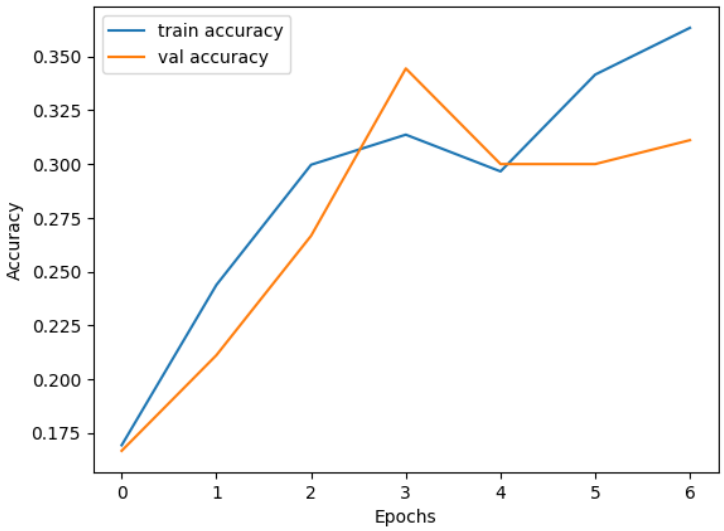
Tab. 6.33. Wyniki dokładności modelu sieci bimodalnej w metodzie fuzji

Dokładność modelu dla danych walidacyjnych	**Dokładność modelu dla danych testowych**
0,88	0,89

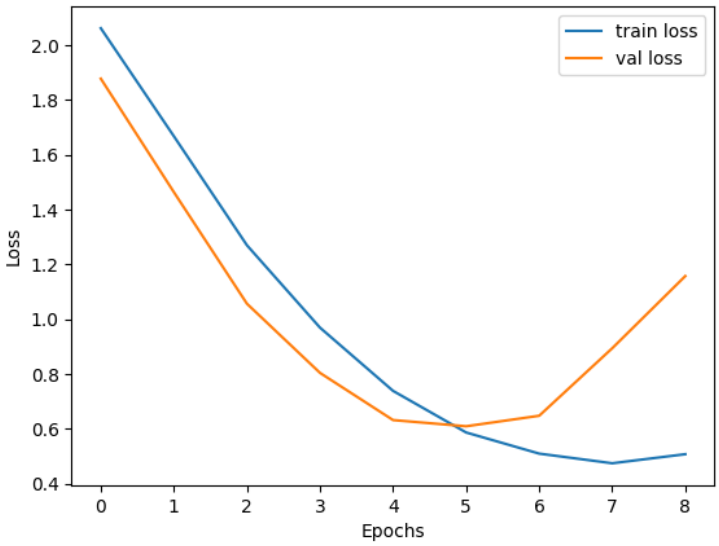
Tab. 6.34. Raport klasyfikacji dla modelu bimodalnego w metodzie fuzji

Klasa	**Precyzja**	**Czułość**	**F1-score**
Agresywny	0,92	0,69	0,79
Ekscytujący	0,88	0,88	0,88
Smutny	0,93	0,88	0,90
Spokojny	0,94	1,00	0,97
Straszny	0,84	0,97	0,90
Szczęśliwy	0,83	0,91	0,87
Średnia – macro avg	0,89	0,89	0,89

Model − poprzez użycie techniki wczesnego zatrzymania − uzyskał najlepsze wyniki dla wartości straty w 4 epoce, co można zauważyć na rys. 6.24. Na rys. 6.25 przedstawiono wartość dokładności i jej zmiany względem treningu w kolejnych epokach



Rys. 6.23. Dokładność modelu multimodalnego w metodzie fuzji



Rys. 6.24. Wartości funkcji straty dla modelu multimodalnego w metodzie fuzji

7. Analiza wyników uczenia głębokiego

7.1. Zestawienie wyników dla fragmentu muzyki filmowej, obrazu oraz obrazu wraz z ścieżką filmową

W tabeli 7.1 przedstawiono wyniki dokładności na zbiorze testowym ze wszystkich przeprowadzonych testów na różnych modelach, w których podjęto analizę wyłącznie wideo bądź analizę multimodalną oraz liczbę parametrów poszczególnych modeli użytych do stworzenia modelu bimodalnego. Modele oparte na parametrach analizy sygnału audio miały na celu wybór najlepszego modelu do stworzenia modelu multimodalnego.

Tab. 7.1. Podsumowanie modeli klasyfikacji emocji na podstawie fragmentu filmu

Model	**Liczba parametrów modelu**	**Dokładność na zbiorze walidacyjnym**	**Dokładność na zbiorze testowym**
Model bimodalny	3 746 878	0,88	0,89
Model 3D CNN	3 612 358	0,83	0,87
Model GRU (Histogram RGB)	87 558	0,68	0,70
Model GRU (parametry sieci Inception V3)	Inception V3: 22 327 328	0,97	0,16
	GRU: 102 534		
	W sumie: 22 429 862		

Dodatkowo dla modelu bimodalnego uzyskano wartość F1-score na poziomie 0,89, co przewyższa wynik udokumentowany w pracy Behrouzi i in., gdzie uzyskano wartość metryki F1-score na poziomie 0,66 [7]. Ponadto została obliczona metryka mAP dla modelu bimodalnego. Zaproponowany w pracy model osiągnął wartość metryki mAP na poziomie 0,89, co również przewyższa wynik udokumentowany w literaturze, tj. pracy Yu i in., gdzie uzyskano mAP na poziomie 0,61 [127].

Wyniki przedstawione w tab. 7.1 pozwoliły na udowodnienie tezy nr 1, tj.:

1. **Możliwe jest przygotowanie i wytrenowanie modeli sieci neuronowych, które osiągają dokładność klasyfikacji emocji zawartych w muzyce filmowej wyższą, tj. >90% oraz we fragmencie wideo pochodzącego z filmu, tj. >85%, w stosunku do wyników uzyskanych w literaturze (odpowiednio: 82% – Revathy, Pillai [94] oraz 73,6% – Hayat i in. [30]), co przewyższa aktualny stan wiedzy.**

oraz tezy nr 2 rozprawy, tj.:

2. **Wykorzystanie bimodalnego podejścia w uczeniu maszynowym, tj. jednoczesnej analizy sygnałów audio i wideo, pozwala zwiększyć dokładność klasyfikacji emocji zawartych w filmie do wartości 89% w stosunku do analizy jedynie w oparciu o sygnał wideo (najlepsza uzyskana dokładność dla modelu wideo na zbiorze testowym wyniosła 87%).**

7.2. Podsumowanie analizy uczenia głębokiego

Uzyskane wyniki potwierdzają słuszność podejścia analizy bimodalnej, gdzie pod uwagę wzięto zarówno parametry sygnału audio (tj. parametry MFCC), jak również parametry sygnału wideo, które zostały wyekstrahowane i sklasyfikowane za pomocą trójwymiarowej sieci splotowej.

Do analizy filmu w kontekście wartości czasowo-przestrzennych użyto metod uczenia głębokiego, takich jak sieci rekurencyjne oraz trójwymiarowe sieci splotowe. Filmy są sekwencją klatek, w których zachowana jest przestrzenna struktura, dlatego też wykorzystanie trójwymiarowych sieci splotowych pozwoliło na analizę przestrzenną informacji z uwzględnieniem zmian na linii czasu. Dodatkowo wykorzystanie sieci rekurencyjnej dało możliwość przeprowadzenia analizy w kontekście temporalnym przy uwzględnieniu poprzednich, jak i kolejnych zmian w sygnale audio.

W analizie sygnału audio poprzez parametryzację MFCC oraz wykorzystaniu sieci rekurencyjnej osiągnięto wysoką dokładność modelu na poziomie 90%. Połączenie modelu trójwymiarowej sieci splotowej oraz sieci rekurencyjnej GRU z analizą parametrów MFCC sygnału audio pozwoliło na uzyskanie wzrostu dokładności modelu multimodalnego o 2 punkty procentowe w stosunku do modelu trójwymiarowej sieci splotowej opartego na analizie samego sygnału wideo, kosztem niewielkiego wzrostu złożoności modelu sieci neuronowej.

Model sieci rekurencyjnej o małej złożoności oparty na analizie obrazu wideo jako sekwencji klatek na podstawie wartości histogramów RGB osiągnął niezadowalający wynik dokładności na poziomie 70%, dodatkowo analiza raportu klasyfikacji (tab. 6.27) pokazała problemy ze zbalansowaną wartością wyniku F1-score dla wszystkich klas. Niskie wartości parametru precyzji dla tego modelu wskazują, że model często fałszywie przypisuje właściwemu przypadkowi odpowiednią klasę.

Model sieci GRU wykorzystujący parametryzację sygnału wideo poprzez sieć splotową Inception V3, pomimo wysokiego wyniku dokładności na zbiorze walidacyjnym, uzyskał bardzo niski wynik na zbiorze testowym, tj. 16%. Spowodowane jest to prawdopodobnie przetrenowaniem nieskomplikowanej sieci rekurencyjnej GRU i zbytnim dopasowaniem do danych treningowych modelu bądź zbyt dużą złożonością danych wejściowych podawanych na model sieci rekurencyjnej.

8. Podsumowanie

W przedstawionej dysertacji przenalizowane zostały aspekty koloru poszczególnych filmów/ujęć w kontekście ich wpływu na emocje widza. Zarówno analiza statystyczna, jak i analiza za pomocą algorytmów uczenia głębokiego potwierdzają zależność odpowiednio dobranej kolorystyki filmu/ujęć do odczuwania emocji poprzez widza.

Przeprowadzono trzy ankiety subiektywne, które miały na celu wskazanie powiązania emocji z kolorem, dzięki czemu powstał sześcioelementowy kolorowy model emocji. Następnie na tej podstawie przygotowano bazę tytułów filmowych, których fragmenty zostały przedstawione w kolejnych dwóch ankietach. Pierwsza ankieta zawierała wyłącznie muzykę filmową, natomiast w drugiej ankiecie zawarto fragmenty filmów z towarzyszącą im muzyką filmową. Zadaniem ankietowanych było przypisanie do fragmentów muzyki, jak i filmu emocji, które respondenci odczuwają podczas odsłuchiwania/oglądania fragmentów muzyki/filmu (rozdział 5).

Testy subiektywne pozwoliły również na przypisanie odpowiednich etykiet emocji dla danych tytułów filmowych, z których następnie wydzielono pojedyncze fragmenty. W ten sposób zbudowano zbiór danych do uczenia głębokiego.

Podjęto również wszechstronną analizę sygnału audio i wideo, przeprowadzono parametryzację tych sygnałów, dzięki czemu możliwe był wybór sposobu parametryzacji oraz modelu sieci neuronowej do analizy sygnału audio oraz analizy sygnału wideo w celu zbudowania architektury bimodalnej.

W celu potwierdzenia pierwszej tezy przeprowadzono szereg badań z wykorzystaniem algorytmów uczenia głębokiego, w których sprawdzono różne metody parametryzacji sygnału audio, jak i wideo w połączeniu z działaniem rozwiązań sieci splotowych oraz sieci rekurencyjnych do celów klasyfikacji sygnałów przestrzenno-czasowych. W ramach eksperymentów sprawdzono dziewięć różnych architektur sieci neuronowych. Eksperyment ten potwierdza tezę nr. 1, tj.:

„Możliwe jest przygotowanie i wytrenowanie modeli sieci neuronowych, które osiągają dokładność klasyfikacji emocji zawartych w muzyce filmowej wyższą, tj. >90% oraz we fragmencie wideo pochodzącego z filmu, tj. >85%, w stosunku do wyników uzyskanych w literaturze (odpowiednio: 82% – Revathy, Pillai [94] oraz 73,6% – Hayat i in. [30]), co przewyższa aktualny stan wiedzy" (rozdział 6)**.**

Film jest sygnałem zarówno przestrzennym, jak i czasowym, dlatego też skupiono się na architekturze umożliwiającej analizę tych dwóch zależności jednocześnie. Konstrukcja architektury bimodalnej pozwoliła na zwiększenie dokładności sieci w zadaniu klasyfikacji emocji na podstawie fragmentu filmu.

Efektem analizy tych powiązań był wybór najlepiej sprawdzających się metod klasyfikacji dla sześciu emocji. Zostały one wykorzystane do stworzenia architektury bimodalnej, której dokładność klasyfikacji okazała się wyższa w porównaniu do klasyfikacji samego obrazu pochodzącego z fragmentu filmu. W ten sposób udowodniono tezę nr 2, tj.:

„Wykorzystanie bimodalnego podejścia w uczeniu maszynowym, tj. jednoczesnej analizy sygnałów audio i wideo, pozwala zwiększyć dokładność klasyfikacji emocji zawartych w filmie do wartości 89% w stosunku do analizy jedynie w oparciu o sygnał wideo (najlepsza uzyskana dokładność dla modelu wideo na zbiorze testowym wyniosła 87%)" (rozdział 6).

Użycie trójwymiarowej sieci splotowej pozwoliło na uwzględnienie trzech wymiarów danych, długości, szerokości oraz koloru, oznacza to, że bardziej złożona 3D CNN umożliwia analizę kolorystyczną w różnych wymiarach abstrakcji. Uzupełnieniem architektury bimodalnej była nieskomplikowana sieć GRU wykorzystująca analizę za pomocą parametrów MFCC. Pozwoliła ona na analizę w kontekście czasowym sygnału audio (w tym przypadku muzyki filmowej), sieć rekurencyjna z dużą skutecznością potrafiła generalizować parametry muzyczne.

W niniejszej rozprawie można wskazać kilka oryginalnych rozwiązań zaproponowanych przez autora:

- W testach subiektywnych wykorzystano technologię śledzenia wzroku (ang. *gaze-tracking; eye-tracking*), która pozwoliła na dodatkową ocenę odpowiedzi i jednocześnie zwiększenie wiarygodności tej oceny;
- Badania nad przypisaniem odpowiednich etykiet do kolorów filmu zostały potwierdzone trzema różnymi testami subiektywnymi, w których określono model emocji składający się z sześciu kolorów;
- Przygotowano zbiór danych wykorzystujący fragmenty filmów, do których etykiety zostały przygotowane na podstawie odpowiedzi w testach subiektywnych;
- Sprawdzono różne rodzaje parametryzacji sygnału audio oraz sygnału wideo w kontekście emocji zawartych w filmie;
- Sprawdzono dwa rodzaje sieci neuronowych: sieci splotowe oraz sieci rekurencyjne do rozpoznawania i klasyfikacji emocji zawartych w filmie;
- Zaproponowano architekturę sieci bimodalnej składającej się z sieci rekurencyjnej i trójwymiarowej sieci splotowej do klasyfikacji emocji na podstawie 4-sekundowych fragmentów filmów.

Propozycje dalszych prac

Obecnie rozwój platform strumieniowania (*streamingowych*) czy serwisów zawierających treści związane z produkcją filmową posiada wiele form profilowania swojej zawartości poprzez zastosowanie odpowiednich filtrów dla różnych użytkowników. Odpowiednie algorytmy analizy filmów pozwalają na dodatkowe opcje sortowania zawartości tych stron pod kątem wyszukiwania przez użytkowników interesujących dla nich pozycji. W ramach kontynuacji tego tematu, jedną z dalszych prac nad tymi zagadnieniami jest przygotowanie modelu w oparciu o przekazanie wiedzy (ang. *transfer learning*) i osiągnięcie lepszej generalizacji podczas klasyfikacji emocji dla danych fragmentów filmu.

W celu osiągnięcia lepszej generalizacji istotnym elementem byłoby poszerzenie zbioru o kolejne tytuły filmowe oraz wiarygodne przypisanie etykiet emocji dla danych fragmentów filmów. Można by również poszerzyć zbiór etykiet o dodatkowe emocje.

Dodatkowym aspektem klasyfikacji emocji mogłaby być klasyfikacja wieloetykietyowa danych, tzw. *multi-labeling*. Pozwoliłoby to na przypisanie nie jednej, a na przykład dwóch klas emocji dla danego fragmentu filmu. Jednak wiąże się to z przebudową zaproponowanej architektury w kontekście klasyfikacji emocji na ostatniej warstwie i zastosowanie innej funkcji aktywacji odpowiadającej za wyznaczenie prawdopodobieństwa wystąpienia danej klasy, czyli funkcji sigmoidalnej oraz zmiany mechanizmu wnioskowania. W takim przypadku można by też zastosować model pozwalający na predykcję, a nie jedynie klasyfikację etykiet emocji.

Poprzez uzyskanie większych zdolności generalizujących modelu możliwe byłoby wdrożenie dodatkowej filtracji na podmioty stron internetowych czy serwisów zajmujących się tematyką filmową, które udostępniają do filmy, czy też ujęcia filmowe (np. serwis Youtube).

Kolejną dodatkową funkcją zaproponowanego modelu mógłby być system wspomagający osoby pracujące nad filmem w trakcie ostatniej fazy postprodukcji, jaką jest gradacja kolorystyczna (ang. *color grading*). System tego typu – poprzez analizę ścieżki muzycznej, ujęcia filmowe oraz scenopisu – mógłby wskazać kierunek bądź zaproponować ustawienie odpowiednich parametrów obrazu, tj. dostosowanie kontrastu i kolorów obrazu w celu uzyskania ostatecznej kolorystyki filmu/ujęć. Należy jednak pamiętać, że ten etap ma również charakter kreatywny i stylistyczny, dlatego dalsze badania w kontekście stworzenia narzędzia automatyzującego proces gradacji kolorów wymagają nie tylko odpowiednio przygotowanej bazy danych, która zawierałaby zarówno ujęcia w formacie RAW oraz odpowiadające im fragmenty filmów ze ścieżką audio, ale również adnotację w postaci oceny takich zestawów ujęć. Pozwoliłoby to na wykorzystanie metod uczenia głębokiego do analizy zarówno danych audio,

jak i wideo w kontekście usprawnienia pracy twórcom filmów. Obecnie nie istnieje jednak zbiór danych umożliwiający przeprowadzenie badań w tym aspekcie oraz przygotowania mechanizmu automatyzacji gradacji kolorów w filmie.

WYKAZ LITERATURY

1. Anderson I., „Just The Way You Are': Music Listening and Personality", https://research.atspotify.com/2020/12/just-the-way-you-are-music-listening-and-personality/, data dostępu: 13.09.2023.
2. Artlist, https://artlist.io, data dostępu: 13.09.2023.
3. Aslan F., Ekenel H. K., „Emotion Prediction in Movies Using Visual Features & Genre Information.", 2019 4th International Conference on Computer Science and Engineering (UBMK), str. 1-5, Turcja, https://doi.org / 10.1109/UBMK.2019.8907100, 2019.
4. Atienza R., „Deep learning z TensorFlow 2 i Keras dla zaawansowanych", Helion, Warszawa, 2022.
5. Atalli J., „Noise: The Political Economy of Music", Manchester, 1977.
6. Bartsch M. A., Bartsch A., Wakefield G. H., „To catch a chorus: Using chroma-based representations for audio thumbnailing", IEEE Workshop Applications of Processing to Audio and Acoustic, https://doi.org/10.1109/ASPAA.2001.969531, 2001.
7. Behrouzi T., Toosi R., Akhaee M. A., „Multimodal movie genre classification using recurrent neural network.", Multimed Tools Appl, 82, str. 5763-5784, https://doi.org/10.1007/s11042-022-13418-6, 2023.
8. Bellantoni P., „Jeśli to fiolet, ktoś umrze. Teoria koloru w filmie.", Wydawnictwo Wojciech Marzech, Warszawa, 2009.
9. Bergstra J., Bengio Y., „Random Search for Hyper-Parameters Optimization", Journal of Machine Learning Research, str. 281-305, 2012.
10. Bindemann M., Burton A. M., Langton S. R. H., „How do eye gaze and facial expression interact?", Visual Cogn., 16, str. 708–733, https://doi.org/10.1080/13506280701269318, 2008.
11. Blog Statystyczny, „Macierz pomyłek i co z tego wynika?", https://www.statystyczny.pl/macierz-bledow-raport-dokladnosc-czulosc-precyzja/, data dostępu: 13.09.2023.
12. Brinkmann R., „The Art and Science of Digital Com", Burlington: Elsevier, 2008.
13. Brown B., „Cinematography: Sztuka Operatorska", Wydawnictwo Wojciech Marzec, Warszawa, 2018.
14. Brownlee J., "A Gentle Introduction to Ensemble Learning Algorithms", https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/, data dostępu: 11.11.2023.
15. Byrd G., „A similarity scale for content-based music IR.", http://homes.sice.indiana.edu/donbyrd/MusicSimilarityScale.HTML, data dostępu: 23.09.2023.
16. Ciborowski T., Reginis S., Kurowski A., Weber D., Kostek B., „Classifying Emotions in Film Music - A Deep Learning Approach.", Electronics, 10, https://doi.org/10.3390/electronics10232955, 2021.
17. Czekaj N., „Kolory w roli głównej. Filmy w których paleta barw bryluje na pierwszym planie.", https://papaya.rocks/pl/opinions/kolor-w-roli-glownej-filmy-w-ktorych-paleta-barw-bryluje-na, data dostępu: 11.09.2023.
18. Damasio A., Damasio H., Adolphs R., „Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala", Nature, 372, str. 669-672, 1994.
19. Deep Drive PL, „Audio-wizja w przebraniu?", https://deepdrive.pl/audio-wizja-w-przebraniu/, data uzyskania dostępu: 12.09.2023.
20. Du P., Li X., Gao Y., „Dynamic Music emotion recognition based on CNN-BiLSTM", IEEE 5th Information Technology and Mechatronics Engineering Conference, str. 1372-1376, https://doi.org/10.1109/ITOEC49072.2020.9141729, 2020.

21. Ekman P., Wallace V. F., „Hand Movements”, Journal of Communication, 22(4), str. 353-374, https://doi.org/10.1111/j.1460-2466.1972.tb00163.x, 1972.
22. Ellis D., Poliner G., „Identifying cover songs with chroma features and dynamic programming beat tracking”, 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, https://doi.org /10.1109/ICASSP.2007.367348, 2007.
23. Epidemic Sound, https://epidemicsound.com, data dostępu: 12.09.2023.
24. Fehr B., Russell J. A., „Concept of emotion viewed from a prototype perspective.”, Journal of Experimental Psychology: General, 113, str. 464-486, http://dx.doi.org/10.1037/0096-3445.113.3.464, 1984.
25. Grekow J., „Music emotion recognition using recurrent neural networks and pretrained models”, Intell Inf Syst, 57, str. 531-546, https://doi.org/10.1007/s10844-021-00658-5, 2021.
26. Gerawe O., Nagel F., Kopiez R., Altenmuller E., „Emotions over time: Synchronicity and development of subjective,physiological and facial affective reactions to music.”, Emotion, 7(4), str. 774-788, https://doi.org/10.1037/1528-3542.7.4.774, 2007.
27. Geron A., „Uczenie maszynowe z użyciem Sci-Kit Learn i Tensorflow”, Helion, Warszawa, 2018.
28. Gomez-Uribe C. A., Hunt N., "The Netflix Recommender System: Algorithms, Business Value, and Innovation", ACM Trans. Manage. Inf. Syst, 6(4), http://dx.doi.org/10.1145/2843948, 2015.
29. Google Colaboratory, https://colab.research.google.com, data dostępu 20.06.2023.
30. Hayat H., Ventura C., Lapedriza A., „Recognizing Emotions evoked by Movies using Multitask Learning”, 9th International Conference on Affective Computing and Intelligent Interaction (ACII), https://doi.org/10.48550/arXiv.2107.14529, 2021.
31. He K., Zhang X., Ren S., Sun J., “Deep residual learning for image recognition”, Proceedings of the IEEE conference on computer vision and pattern recognition, str. 770–778, https://doi.org/10.48550/arXiv.1512.03385, 2016.
32. Heideggera M., „The Will to Power as Art”, Routledge & Kegan, Londyn, 1981.
33. Hellmuth O., Allamanche E., Herre J., Kastner T., Cremer M., Hirsch W., „Advanced Audio Identification Using MPEG-7 Content Description”, Fraunhofer Institute for Integrated Circuits IIS-A, 2001.
34. Hevner K., „Experimental Studies of the Elements of Expression in Music”, The American Journal of Psychology, 49, str. 246-268, 1936.
35. Hilmkil A., Thome C., Arpteg A., „Perceiving Music Quality with GANs”, Northern Lights Deep Learning Workshop, https://doi.org/10.48550/arXiv.2006.06287, 2020.
36. Hu X., Downie S. J., Laurier C., Bay M., „MIREX audio mood classification task: Lesson learned”, 9th International Conference on Music Information Retrieval, 2008.
37. Hyrskykari A., „Utilizing eye movements: Overcoming inaccuracy while tracking the focus of attention during reading.”, Comput. Hum. Behav, 22, str. 657–671, https://doi.org/10.1016/j.chb.2005.12.013, 2006.
38. Image-net, https://www.image-net.org, data dostępu: 20.06.2023.
39. James J., „Digital Intermediates for Film and Video.”, Burlnigton: Focal Press, 2006.
40. Javaheri B., „Feature extraction and image classification using Deep Neural Networks and OpenCV”, https://domino.ai/blog/feature-extraction-and-image-classification-using-deep-neural-networks, data dostępu: 12.09.2023.
41. Juslin P. N., „Musical Emotions Explained", Oxford University Press, 2019.
42. Jones R., Fay R., Popper A., „Music Perception”, Springer-Verlag, Nowy Jork, 2010.
43. Kardas P., „Homerecording dla każdego”, Piotr Kardas Productions, Warszawa, 2015.
44. Kay W., Carreria J., Simonyan K., Zhang B., Hiller C., The Kinetics Human Action Video Dataset, Computer Vision and Pattern Recognition, https://doi.org/10.48550/arXiv.1705.06950, 2017.

45. Keras, „Video Classification with a CNN-RNN Architecture”, https://keras.io/examples/vision/video_classification/, data dostępu: 12.09.2023.
46. Keras, „Keras Applications", https://keras.io/api/applications/, data dostępu: 23.09.2023.
47. Khanh-An Q., Vinh-Tiep N., Minh-Triet T., „Frame-based Evaluation with Deep Features to Predict Emotional Impact of Movies”, CEUR Workshops Procedings, 2018.
48. Kivy P., „Sound Sentiment: An Essay on the Musical Emotions”, Including the Complete Text of The Corded Shell, Temple University Press, Filadelfia, 1989.
49. Kline A., „Multimodal Machine Learning: Data Fusion", https://pub.towardsai.net/multimodal-machine-learning-data-fusion-d1d8776e2cb0, data dostępu: 12.12.2023.
50. Kochanowska E., Majzner R., „Muzyka w dialogu z edukacją", Libron, Kraków, 2015,
51. Kostek B., „Anatomia i fizjologia narządu wzroku”, https://sound.eti.pg.gda.pl/student/pp/oko-budowa_i_wlasnosci.pdf, data dostępu: 2.10.2023.
52. Kostek B., Plewa M., „Parametrization and Correlation Analysis Applied to Music Mood Classification”, International Journal of Computional Inteligentce Studies, 2, https://doi.org/10.1504/IJCISTUDIES.2013.054734, 2013.
53. Kostek B., Plewa M., „Testing a variety of Features for Music Mood Recognition”, Journal of the Acoustical Society of America, 134, https://doi.org/10.1121/1.4830570, 2013.
54. Kostek B., Wójcik J., „Forming and Ranking Musical Rhythm Hypotheses”, Lecture Notes in Computer Science, http://dx.doi.org/10.1007/978-3-540-30132-5_102, 2004.
55. Lartillot O., Eerola T., Toiviainen P., Fornari J., „Multi-feature modeling of pulse clarity: Design, validation, and optimization.”, 9th International Conference on Music Information Retrieval, Filadelfia, 2018.
56. Laurie C., Herrera P., „Automatic Detection of Emotion in Music: Interaction with Emotionally Sensitive Machines”, https://doi.org/ 10.4018/978-1-60566-354-8.ch002, 2012.
57. Lawton G., „How Netflix uses emotional analytics to improve CX”, TechTarget, https://www.techtarget.com/searchcustomerexperience/feature/How-Netflix-uses-emotional-analytics-to-improve-CX, data dostępu: 12.09.2023.
58. Lee C., Su Y-R., Chen C-H., „End-to-end deep learning for recognition of ploidy status using time-lapse videos", Journal of Assisted Reproduction and Genetics, 38(6), https://doi.org/10.1007/s10815-021-02228-8, 2021.
59. Lewis G. N., „The Conversation of the Photons", Nature, 118, str. 874-875, 1926.
60. Le Cun Y., Boser B., Denker J. S., Henderson D., Howard R. E., Hubbard W., Jackel L. D., „Handwritten Digit Recognition with a Back-Propagation Network", IEEE Communications Magazine, 27(11), https://doi.org/10.1109/35.41400, 1989.
61. Levitin D. J., „This is Your Brain on Music: The Science of a Human Obsession.”, Grove/Atlantic, Londyn, 2008.
62. Livingstone S. R., Muhlenberg R., Brown A. R., Loch A., „Controlling musical emotionality: an affective computational architecture for influencing musical emotions", Digital Creativity, 18, https://doi.org/10.1080/14626260701253606, 2007.
63. Librosa, https://librosa.org/doc/latest/index.html, data dostępu: 12.09.2023.
64. Lutins E., "Ensemble Methods in Machine Learning: What are They and Why Use Them?", https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f, data dostępu: 11.11.2023.
65. Łopatka K., „Akustyka muzyczna. Podstawy notacji muzycznej i teorii muzyki”, https://sound.eti.pg.gda.pl/student/akmuz/01-PodstawyNotacji.pdf, data dostępu: 12.09.2023.
66. Marciniuk K., „Akustyka Muzyczna. Metody rekonstrukcji i archiwizacji nagrań muzycznych”, https://multimed.org/student/akmuz/07-Rekonstrukcja.pdf, data dostępu: 12.09.2023.

67. MacRumors, „Apple Acquires AI Music Startup That Can Generate Dynamic Soundtracks”, https://www.macrumors.com/2022/02/07/apple-acquires-ai-music/, data dostępu: 12.09.2023.
68. Medium, Handling audio data for machine learning, https://medium.com/mlearning-ai/handling-audio-data-for-machine-learning-7ba225f183cb, data dostępu 12.12.2023.
69. Mcauley J. D., „Tempo and Rhythm”, Music Perception, str. 165-199, https://psycnet.apa.org/doi/10.1007/978-1-4419-6114-3_6, 2010.
70. Mercado G., „Okiem filmowca: Nauka i łamanie zasad filmowej kompozycji”, Wydawnictwio Wojciech Marzec, Warszawa, 2011.
71. Merriam A., „The Anthropology of Music”, The Journal of Aesthetic Education, 1968.
72. Moore B. C., „An Introduction to the Psychology of Hearing”, Academic Press, Nowy Jork, 1997.
73. Mohammed A., Kora R., „A comprehensive review on ensemble deep learning: Opportunities and challenges", Journal of King Saud University - Computer and Information Sciences, 35(2), str. 757-774, https://doi.org/10.1016/j.jksuci.2023.01.014, 2023.
74. Mostowska J., „Elementarz młodego kinomana”, https://edukacjafilmowa.pl/elementarz-mlodego-kinomana/, data dostępu: 12.09.2023.
75. MusConv, „Does Apple Music Use AI?”, https://musconv.com/does-apple-music-use-ai/, data dostępu: 13.09.2023.
76. Norman G. J., Necka E., Bernston G. G., „The Psychophysiology of Emotions", Emotion Measurment, str. 83-89, https://doi.org/10.1016/B978-0-08-100508-8.00004-7, 2016.
77. No Film School, „The Color Psychology in Film”, https://nofilmschool.com/color-psychology-in-film, data dostępu: 11.09.2023.
78. OpenCV, Cascade Classifier, https://docs.opencv.org/4.x/db/d28/tutorial_cascade_classifier.html, data dostępu: 12.12.2023.
79. Olmsted-Hawala E., Hawala T., Nichols E., „Answers for Self and Proxy—Using Eye Tracking to Uncover Respondent Burden and Usability Issues in Online Questionnaires”, Lecture Notes in Computer Science, 8516, 2014.
80. OPENSmile, https://www.audeering.com/research/opensmile/, data dostępu: 12.09.2023.
81. Quan K., "Frame-based Evaluation with Deep Features to Predict Emotional Impact of Movies", MediaEval Workshop, 2018.
82. Palermo E., „Examples of Heat Map Survey Questions”, https://www.driveresearch.com/market-research-company-blog/examples-of-heat-map-survey-questions, data dostępu: 13.09.2023.
83. Pandeya Y. R., Bhattarai B., Lee J., „Deep-Learning-Based Multimodal Emotion Classification for Music Videos", Sensors, 2021, https://doi.org/10.3390/s21144927.
84. Paul S., „Video Classification with a CNN-RNN Architecture", Keras, https://keras.io/examples/vision/video_classification/, data dostępu: 12.12.2023.
85. Park J., Sloboda J. A., "Handbook of Music and Emotion: Theory, Research, Applications", Music Library Association, 67(4), str. 760-765, 2011.
86. Picard R., "Affective computing", http://affect.media.mit.edu/, data dostępu: 12.12.2023.
87. Peretz I., Zatorre R. J., „The Cognitive Neuroscience of Music”, Oxford University Press, Oksford, 2003.
88. Plewa M., Kostek B., „Music Mood Visualization Using Self-Organizing Maps.", Archives of Acoustic, 40, str. 513-525, http://dx.doi.org/10.1515/aoa-2015-0051, 2015.
89. Plewa M., „Automatic Mood Indexing of Music Excerpts Based on Correlation between Subjective Evaluation and Feature Vector", rozprawa doktorska, Politechnika Gdańska, promotor: Kostek B., 2015.
90. Pluta M., „Zasady muzyki i notacja muzyczna”, Wydawnictwo AGH, Kraków, 2012.

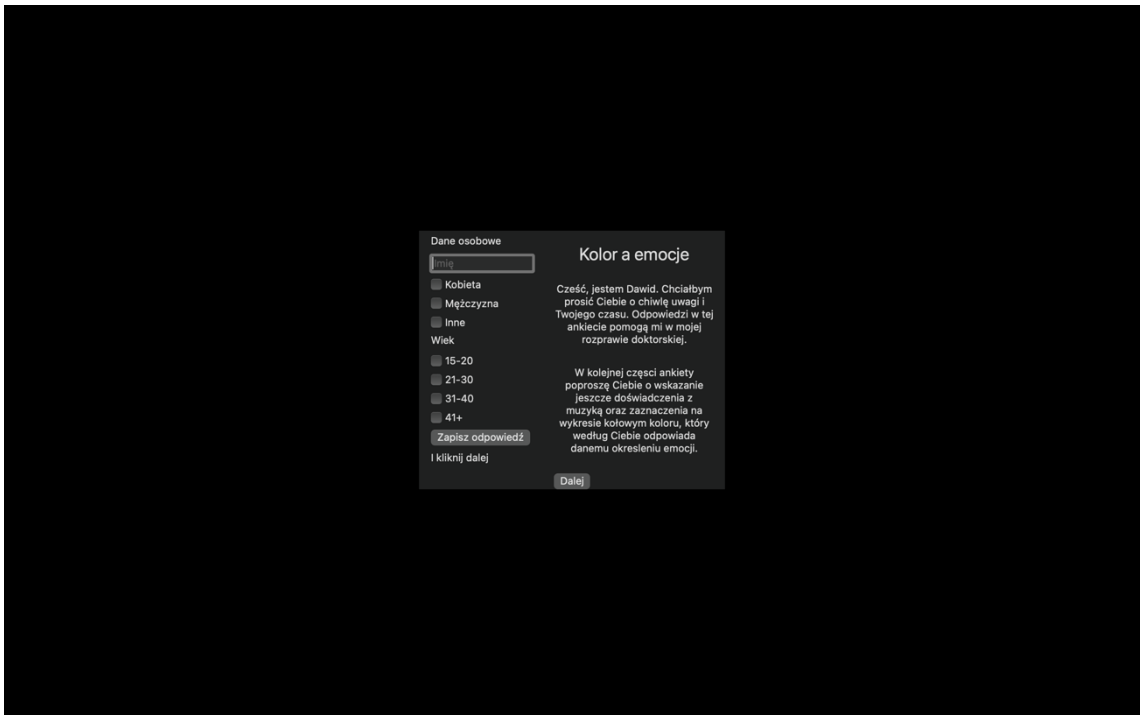
91. Plutchik's Color of Emotion, https://www.6seconds.org/2022/03/13/plutchik-wheel-emotions/, data dostępu: 12.09.2023.
92. Rasheed Z., Sheikh Y., Shah M., „On the use of computable features for film classification.”, IEEE Transactions on Circuits And Systems for Video Technology, str. 52–64, http://dx.doi.org/10.1109/TCSVT.2004.839993(410), 2016.
93. Raschka S., Mirjalili V., „Python. Machine learning i deep learning”, Helion, Warszawa, 2021.
94. Revathy V. R., Pillai A. S., „Multi-class classification of song emotions using Machine learning", 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering, str. 2317-2322, 2022, https://doi.org/10.1109/ICACITE53722.2022.9823535.
95. Russel J.A., „A cicumplex model of affect”, Journal of Personality and Social Psychology, 39, str. 1161-1178, 1980.
96. Saari P., Eerola T., Lartillot O., „Generalizability and simplicity as criteria in feature selection: Application to mood classification in music", IEEE Trans Audio Speech Lang Process 19(6), str.1802–1812, https://doi.org/10.1109/TASL.2010.2101596, 2011.
97. Sacks O., Freeman A., „An Anthropologist on Mars”, Journal of Consciousness Studies, str. 234-240, 1994.
98. Sarkar R., Choudhury S., Dutta S., Roy A., „Recognition of emotion in music based on deep convolutional neural network”, Multimedia Tools and Applications, 79, Springer, 2019.
99. Scholes P., „Metre and Rhythm”, The Oxford Companion to Music, 1977.
100. Schubert E., „Update of the Hevner adjective checklist”, Perceptual and Motor Skills, 96, str. 1117-1123, https://doi.org/10.2466/pms.2003.96.3c.1117, 2012.
101. SciPy, „scipy.stats.pearsonr”, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html, data dostępu: 13.09.2023.
102. Serra J., Gomez A., „A cover song identification system based on sequences of tonal descriptors.”, Internation Conference Music Information Retrieval, Wiedeń, 2007.
103. Sikora T., „The MPEG-7 Visual Standard for Content. Description—An Overview”, IEEE Transactions on Circuits and Systems for Video Technology, 11(6), https://doi.org/10.1109/76.927422, 2001.
104. Simões G., Wehrmann J., Barros R., Ruiz D., „Movie genre classification with convolutional neural networks”, pp. 259-266, https://doi.org/10.1109/IJCNN.2016.7727207.
105. Smith J., „Mathematics of the Discrite Fourier Transform with Audio Applications”, Standford University, Standford, 2007.
106. Splash Media, „Co to jest color grading?”, http://splashmedia.pl/blog/co-to-jest-color-grading/, data dostępu: 7.09.2023.
107. Statystyka-Pomoc, „Testy chi-kwadrat”, http://statystyka-pomoc.com/Chi-kwadrat.html, data dostępu: 13.09.2023.
108. Studiobinder, „Color grading vs color correction process”, https://www.studiobinder.com/blog/color-grading-vs-color-correction-process/, data dostępu: 13.09.2023.
109. Studiobinder, „How to use color in film”, https://www.studiobinder.com/blog/how-to-use-color-in-film-50-examples-of-movie-color-palettes/, data dostępu: 11.09.2023.
110. Szczygieł M., „Korelacja Pearsona”, Cyrkiel.Info, https://cyrkiel.info/statystyka/korelacja-pearsona/, data dostępu: 13.09.2023.
111. Tellagen A., Watsond D., Clark L.A., „On the dimensional and hierarchical structure of affect", Psychology Science, 10, str. 297-303, https://doi.org/10.1111/1467-9280.00157, 1999.
112. Tensorflow, „Video classification with a 3D convolutional neural network", https://www.tensorflow.org/tutorials/video/video_classification, data dostępu 12.12.2023.

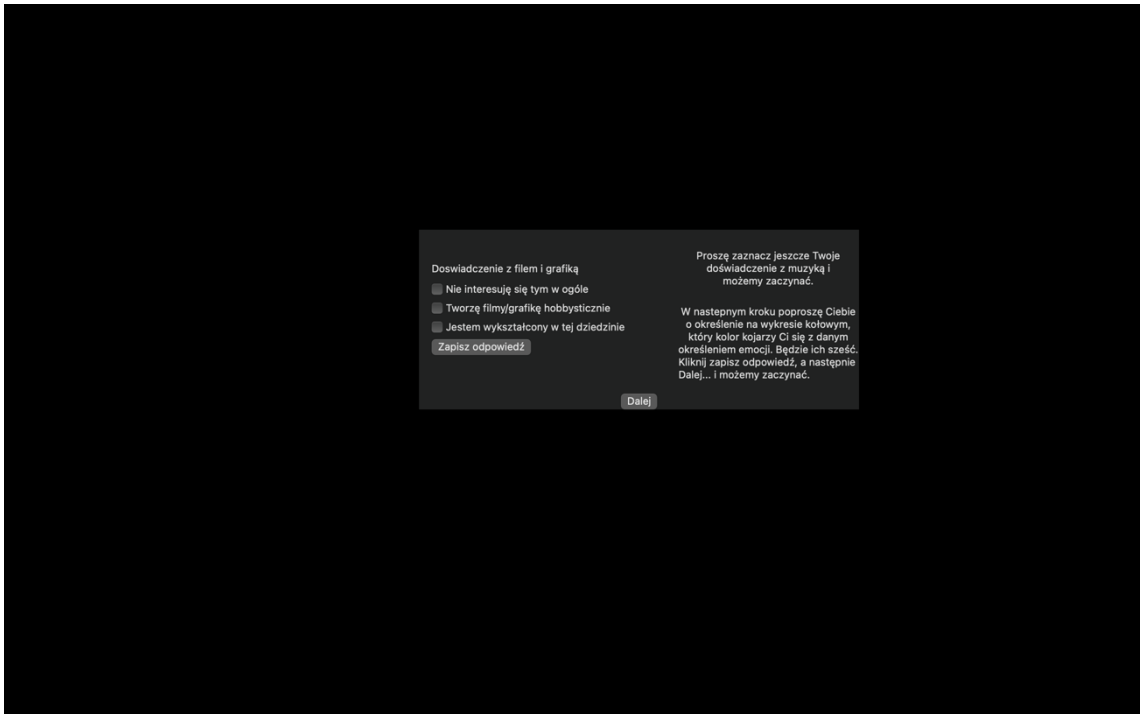
113. Thayer R. E., „The Biopsychology of Mood and Arousal”, Oxford University Press, Oksford, 1989.
114. The Big Five Project - Personality Test, https://www.outofservice.com/bigfive/, data dostępu: 8.11.2023
115. Towards Data Science, „Learning from Audio: Pitch and Chromagrams”, https://towardsdatascience.com/learning-from-audio-pitch-and-chromagrams-5158028a505, data dostępu: 12.09.2023.
116. Towards Data Science, „Why Mel Spectrograms perform better”, https://medium.com/towards-data-science/audio-deep-learning-made-simple-part-2-why-mel-spectrograms-perform-better-aad889a93505, data dostępu: 12.09.2023.
117. Trainor L. J., Tsang C. D., Cheung V. H., „Preference for sensory consonance in 2- and 4-month-old infants”, Music Perception, 20, str.187-194, http://dx.doi.org/10.1525/mp.2002.20.2.187, 2002.
118. Van Hurkman A., „Color Correction Handbook”, Pearson Education, USA, 2013.
119. Vrskova R., Hudec R., Kamencay P., Sykora P., „Human Activity Classification Using the 3DCNN Architecture", Applied Sciences, 12(9), https://doi.org/10.3390/app12020931, 2022.
120. Wang L., Xiong Y., Zhe W., Yu Q., Lin D., Tang X., „Gool L V (2016) Temporal segment networks: towards good practices for deep action recognition”, Eccv, https://doi.org/10.48550/arXiv.1608.00859, 2016.
121. Wehrmann J., Barros R. C., Simões G. S., Paula T. S., Ruiz DD., „Deep Learning from frames.”, Intelligent systems, 2017.
122. Weidman S., „Uczenie głębokie od zera”, Helion, Warszawa, 2020.
123. Wojnicka J., Katafiasz O., „Słownik wiedzy o filmie”, Park, Polska, 2005.
124. Wójcik J., Kostek B., „Intelligent Methods for Musical Rhythm Retrieval”, International Series on Advenced Intelligence, 2004.
125. Yewdall D. L., „Dźwiek w filmie”, Wydawnictwo Wojciech Marzec, 2011.
126. Yang F., Yang J., Zhong J., Wang H., „A Fusion Model for Road Detection based on Deep Learning and Fully Connected CRF", 13th Annual Conference on System of Systems Engineering, 2018, Francja, https://doi.org/10.1109/SYSOSE.2018.8428696.
127. Yu Y., Lu Z., Li Y., Liu D., „ASTS: attention based spatio-temporal sequential framework for movie trailer genre classification”, Multimed Tools Appl, https://doi.org/10.1007/s11042-020-10125-y, 2021.
128. Zero to Mastery TensorFlow Deep Learning, „Transfer Learning with Tensorflow”, https://dev.mrdbourke.com/tensorflow-deep-learning/04_transfer_learning_in_tensorflow_part_1_feature_extraction/, data dostępu: 12.09.2023.
129. Zhen M., Yi M., Luo T., Wang F., „Application of a Fusion Model Based on Machine Learning in Visibility Prediction", Remote Sensing, 15, 5, 2023, https://doi.org/10.3390/rs15051450.
130. Zlatintsi A., Koutras P., Evangelopoulos G., Malandrakis N., Efthymiou N., Pastra K., Potamianos A., Maragos P., „COGNIMUSE: A multimodal video database annotated with saliency, events, semantics and emotion with application to summarization”, EURASIP Journal on Image and Video Processing, 2017.

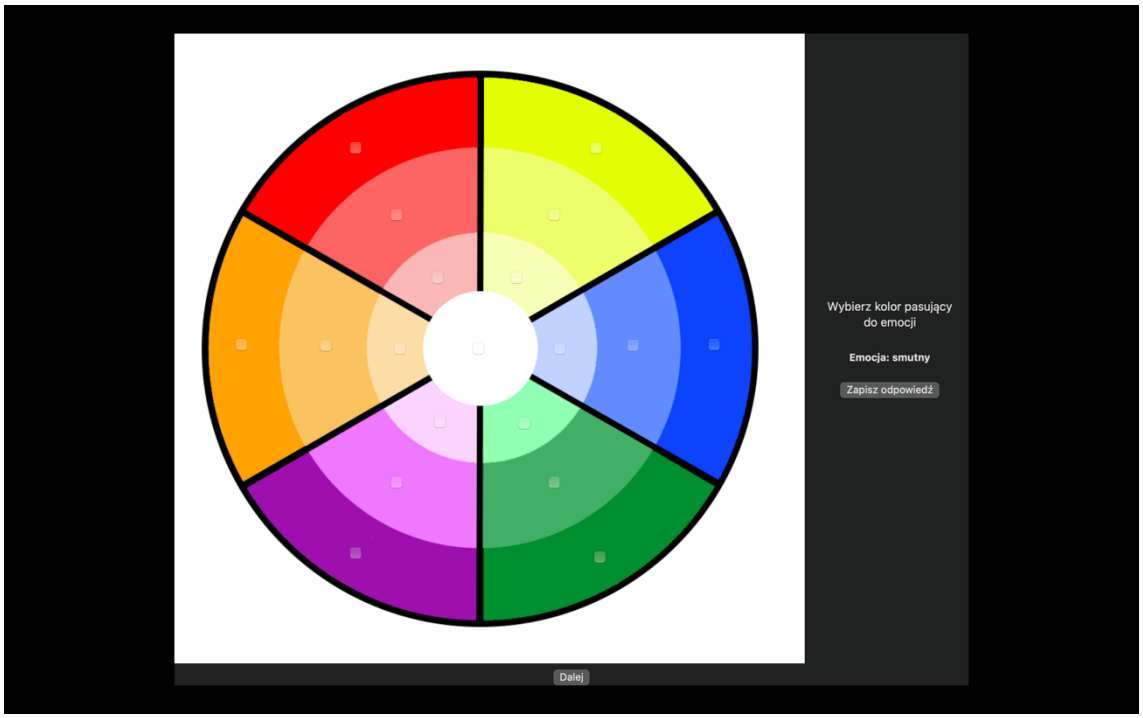
SPIS RYSUNKÓW

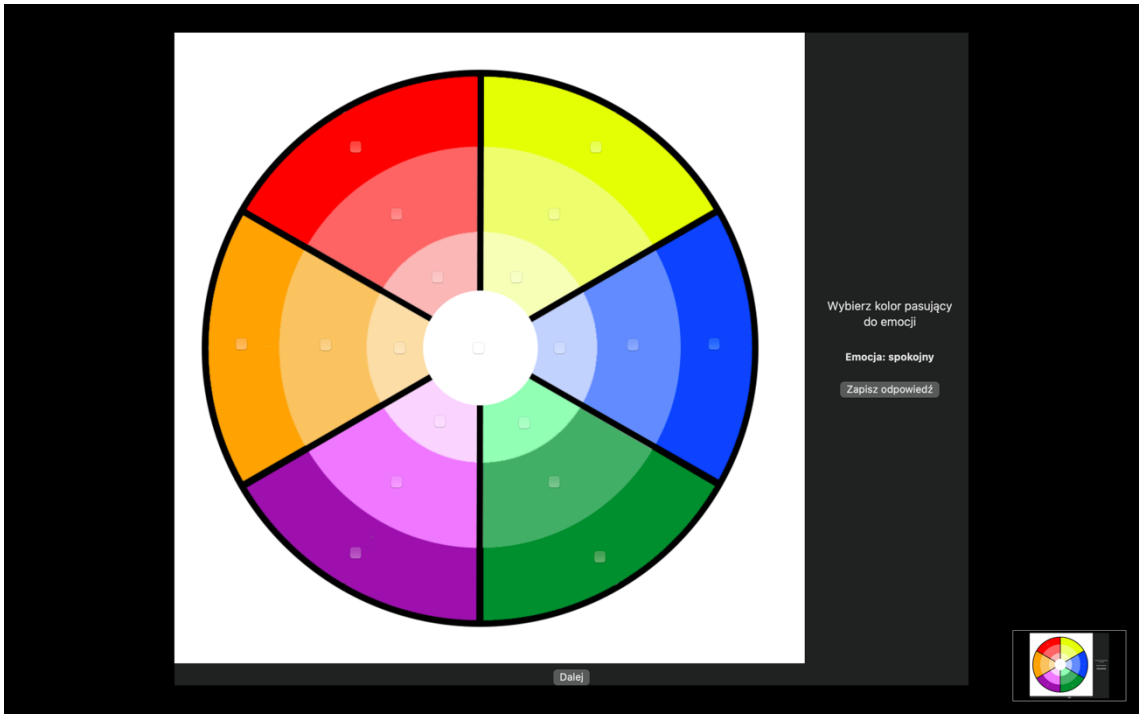
SPIS TABEL

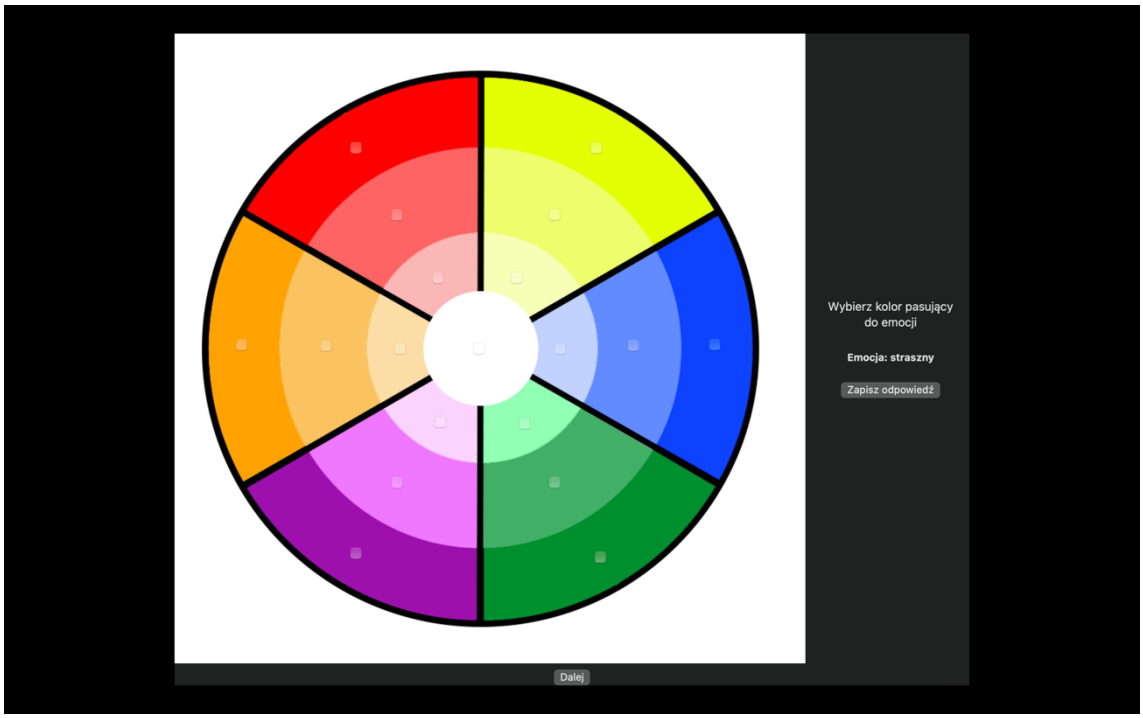
Załącznik A: Ankieta I – dopasowanie emocji do koloru

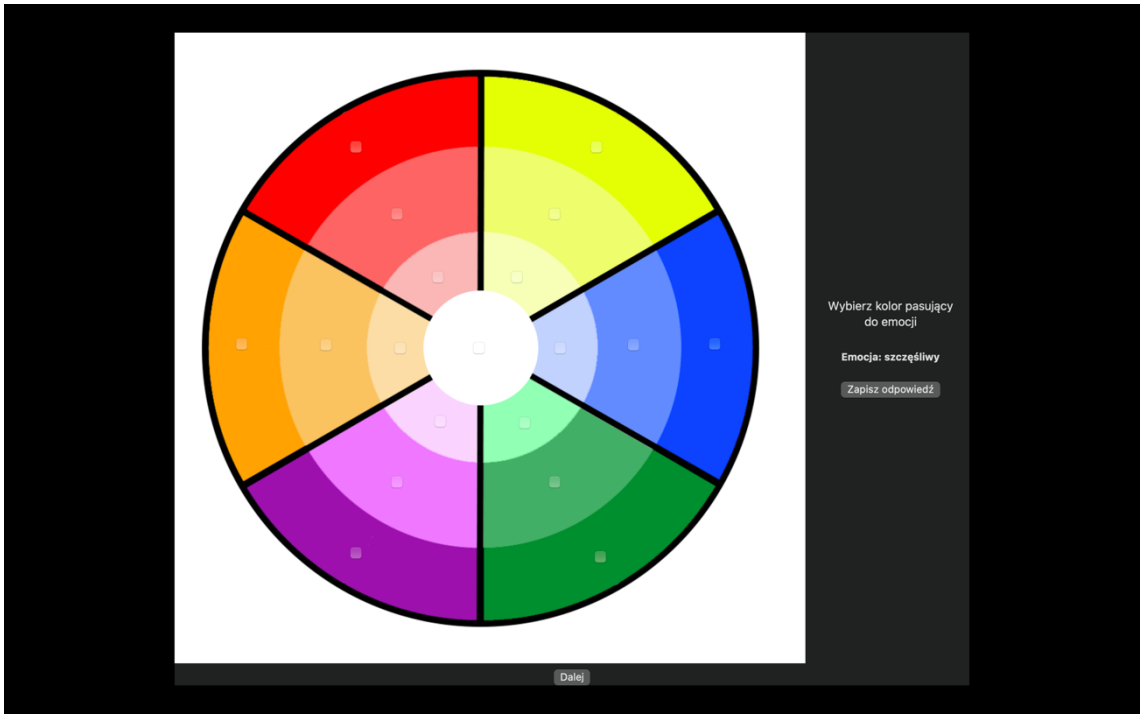


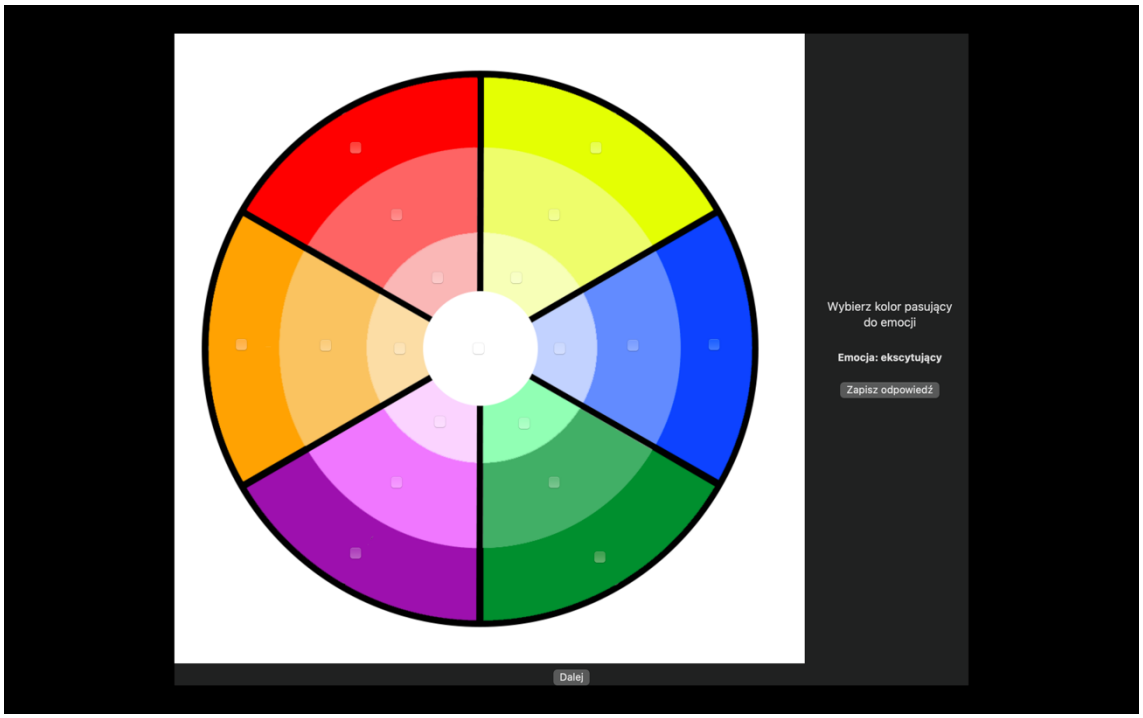


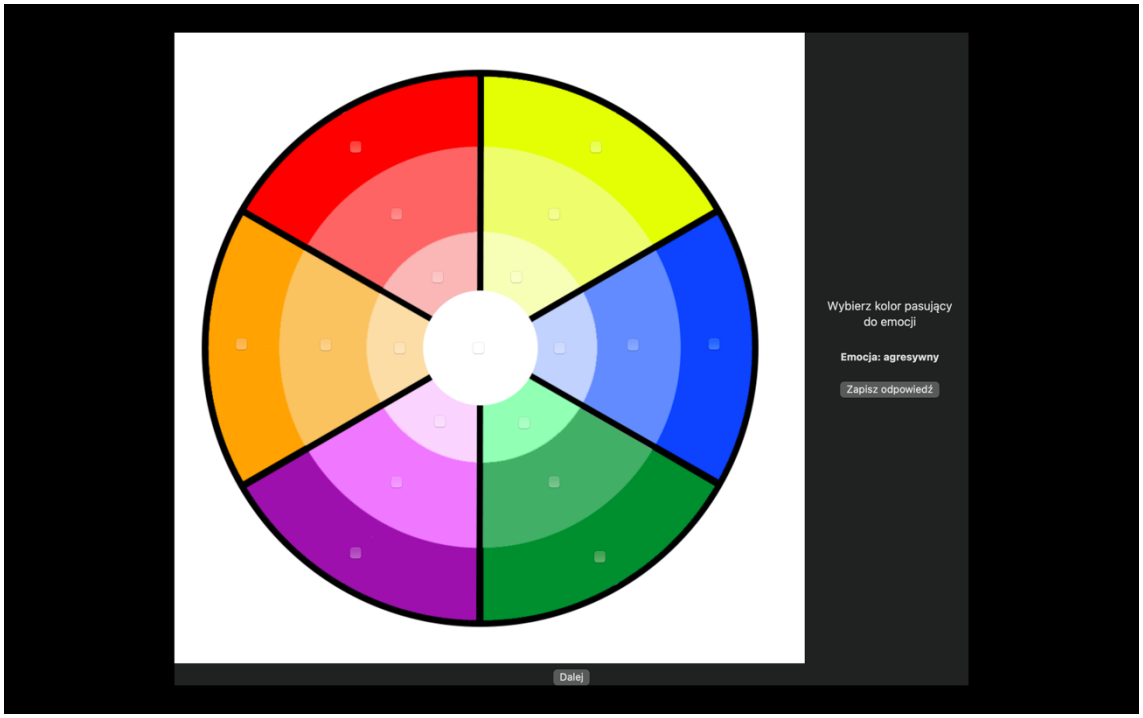












Załącznik B: Ankieta II A– dopasowanie emocji fragmentu muzyki

Emocje w muzyce filmowej - emotions in film music.

Cześć! Jestem Dawid i na potrzeby rozprawy doktorskiej prosiłbym Ciebie o uzupełnienie krótkiej ankiety dotyczącej przypisania odpowiedniej emocji (jednej z sześciu) do słuchanego fragmentu muzycznego. Ankieta składa się z 18 pytań. Ta ankieta pozwoli mi na kontynuowanie badań odnośnie porównań kolorystyki filmu, a muzyki filmowej pod kątem utwierdzania w nas pewnych emocji podczas seansu filmowej.

Poniżej jeszcze trzy krótkie pytania:

English version:

Hi! I am Dawid and for the purposes of my doctoral dissertation, I would ask you to complete a short questionnaire on assigning the right emotions (one of six) to music fragment. Survey have 18 questions. This survey will allow me to continue my research on the comparisons of film colors and film music in terms of confirming certain emotions in us while watching films.

Twój przedział wiekowy *

- ◯ 15-20
- ◯ 21-30
- ◯ 31-40
- ◯ 41+

Twoja płeć / Your gender *

- ◯ Kobieta / female
- ◯ Mężczyzna / male
- ◯ Inne: _____

Jak mocno jesteś związany z produkcją filmową? / How much You are involved in film production? *

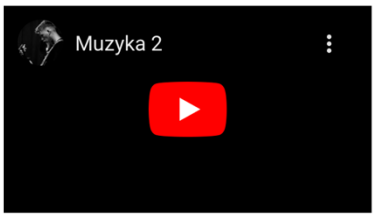
- ◯ Jestem professional reżyserem / filmowcem / aktorem / I am profesional director / filmmaker / actor
- ◯ Film to moje hobby (tworzę amatorskie dzieła) / Film is only my hobby (I create amateurs films)
- ◯ Tylko oglądam filmy / seriale / I only watch films / TV series
- ◯ Nie oglądam filmów / seriali / I don't watch films / TV series

Fragment 1 - muzyka

Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

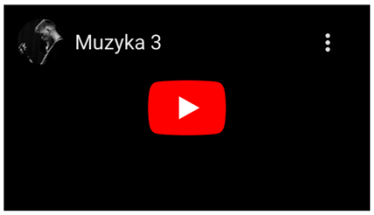
Fragment 2 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

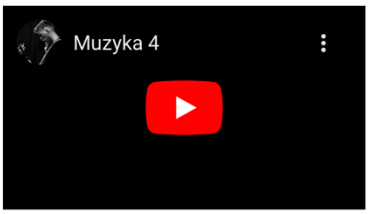
- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 3 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?)

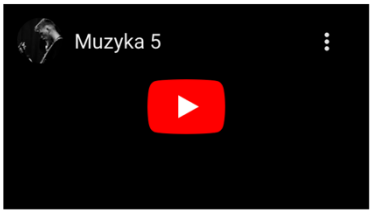
- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

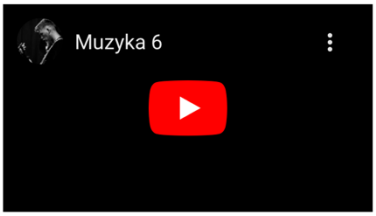
Fragment 5



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

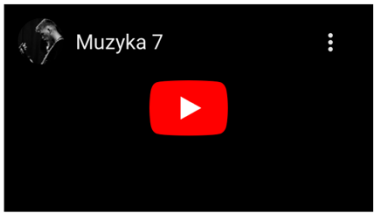
Fragment 6 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

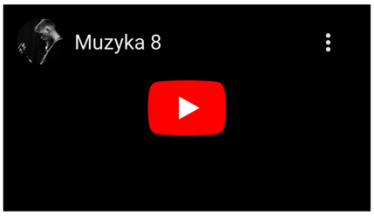
Fragment 7 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

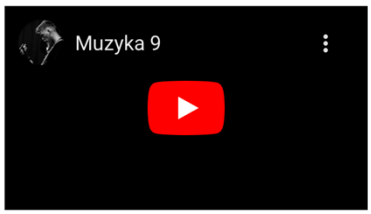
Fragment 8 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 9 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?)

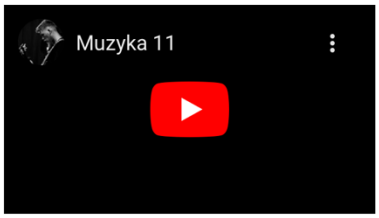
- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 10 - muzyka

Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

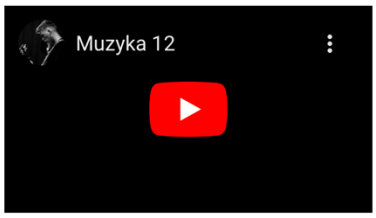
Fragment 11 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

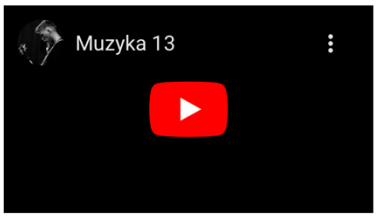
Fragment 12 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

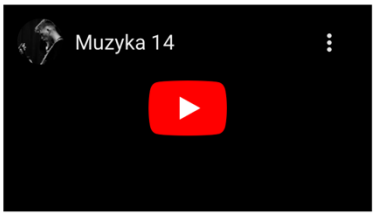
Fragment 13 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

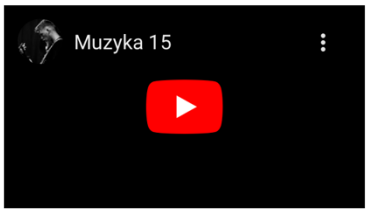
Fragment 14



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?)

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

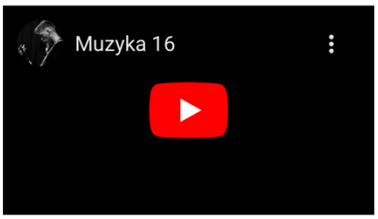
Fragment 15 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

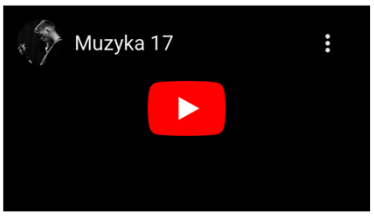
Framgent 16 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

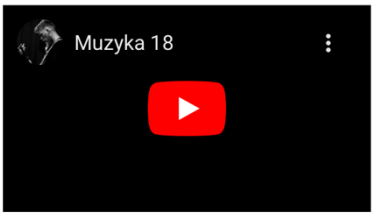
Fragment 17 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?)

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 18 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Załącznik B: Ankieta II B– dopasowanie emocji fragmentu muzyki

Emocje w muzyce filmowej - emotions in film music.

Cześć! Jestem Dawid i na potrzeby rozprawy doktorskiej prosiłbym Ciebie o uzupełnienie krótkiej ankiety dotyczącej przypisania odpowiedniej emocji (jednej z sześciu) do słuchanego fragmentu muzycznego. Ankieta składa się z 18 pytań. Ta ankieta pozwoli mi na kontynuowanie badań odnośnie porównań kolorystyki filmu, a muzyki filmowej pod kątem utwierdzania w nas pewnych emocji podczas seansu filmowej.

Poniżej jeszcze trzy krótkie pytania:

English version:

Hi! I am Dawid and for the purposes of my doctoral dissertation, I would ask you to complete a short questionnaire on assigning the right emotions (one of six) to music fragment. Survey have 18 questions. This survey will allow me to continue my research on the comparisons of film colors and film music in terms of confirming certain emotions in us while watching films.

Twój przedział wiekowy *

- ◯ 15-20
- ◯ 21-30
- ◯ 31-40
- ◯ 41+

Twoja płeć / Your gender *

- ◯ Kobieta / female
- ◯ Mężczyzna / male
- ◯ Inne: _____

Jak mocno jesteś związany z produkcją filmową? / How much You are involved in film production? *

- ◯ Jestem professional reżyserem / filmowcem / aktorem / I am profesional director / filmmaker / actor
- ◯ Film to moje hobby (tworzę amatorskie dzieła) / Film is only my hobby (I create amateurs films)
- ◯ Tylko oglądam filmy / seriale / I only watch films / TV series
- ◯ Nie oglądam filmów / seriali / I don't watch films / TV series

Fragment 1 - muzyka

Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

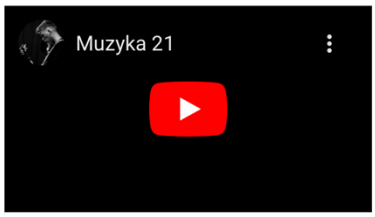
- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 2 - muzyka

Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

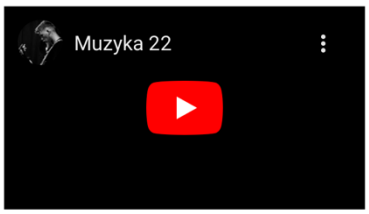
Fragment 3 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?)

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

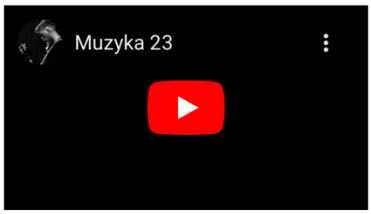
Fragment 4 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

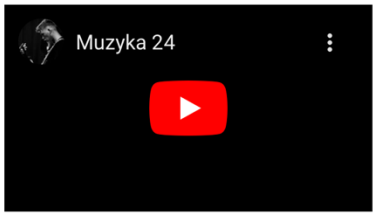
Fragment 5



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

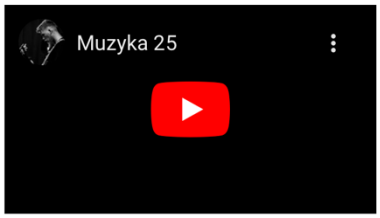
Fragment 6 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

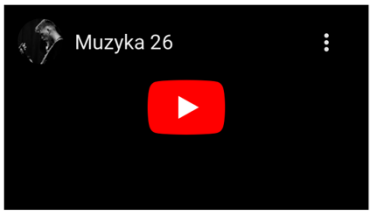
Fragment 7 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

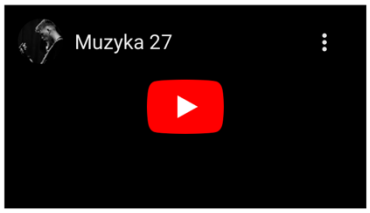
Fragment 8 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

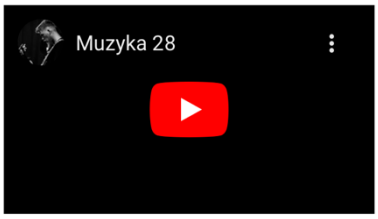
Fragment 9 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?)

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

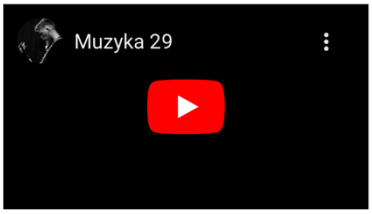
Fragment 10 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 11 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

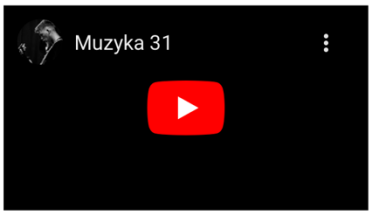
- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 12 - muzyka

Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

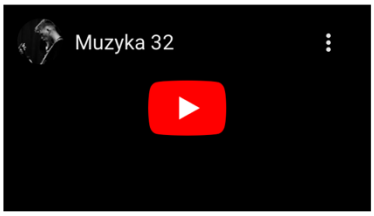
Fragment 13 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

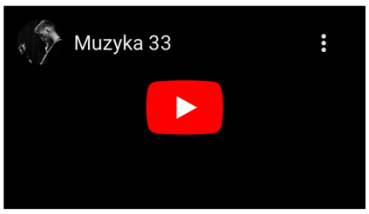
Fragment 14



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?)

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

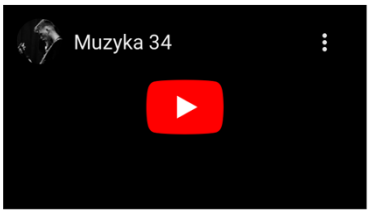
Fragment 15 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

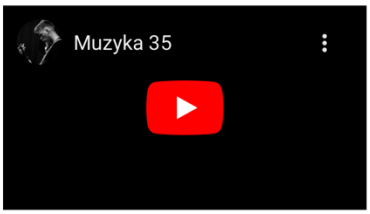
Framgent 16 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

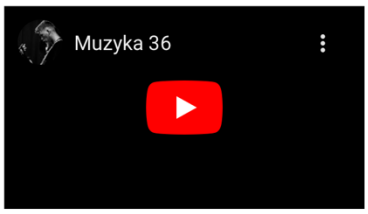
Fragment 17 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?)

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 18 - muzyka



Jaką emocję odczuwasz podczas słuchania tego fragmentu utworu muzycznego? (What emotion do You feel while listening this music fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Załącznik D: Ankieta III A – dopasowanie emocji fragmentu filmu

Emocje w filmie - emotions in film.

Cześć! Jestem Dawid i na potrzeby rozprawy doktorskiej prosiłbym Ciebie o uzupełnienie krótkiej ankiety dotyczącej przypisania odpowiedniej emocji (jednej z sześciu) do oglądanego fragmentu filmu wraz z jego oryginalnym podkładem muzycznym. Ankieta składa się z 18 pytań. Ta ankieta pozwoli mi na kontynuowanie badań odnośnie porównań kolorystyki filmu, a muzyki filmowej pod kątem utwierdzania w nas pewnych emocji podczas seansu filmowej.

Poniżej jeszcze trzy krótkie pytania:

English version:

Hi! I am Dawid and for the purposes of my doctoral dissertation, I would ask you to complete a short questionnaire on assigning the right emotions (one of six) to film fragment with his original soundtrack. Survey have 18 questions. This survey will allow me to continue my research on the comparisons of film colors and film music in terms of confirming certain emotions in us while watching films.

Twój przedział wiekowy *

- ○ 15-20
- ○ 21-30
- ○ 31-40
- ○ 41+

Twoja płeć / Your gender *

- ○ Kobieta / female
- ○ Mężczyzna / male
- ○ Inne: _____

Jak mocno jesteś związany z produkcją filmową? / How much You are involved in film production? *

- ◯ Jestem professional reżyserem / filmowcem / aktorem / I am profesional director / filmmaker / actor
- ◯ Film to moje hobby (tworzę amatorskie dzieła) / Film is only my hobby (I create amateurs films)
- ◯ Tylko oglądam filmy / seriale / I only watch films / TV series
- ◯ Nie oglądam filmów / seriali / I don't watch films / TV series

Fragment 1 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

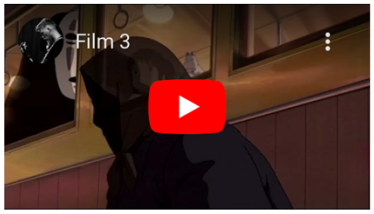
- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 2 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 3 - film



Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?)

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 4 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 5 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 6 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

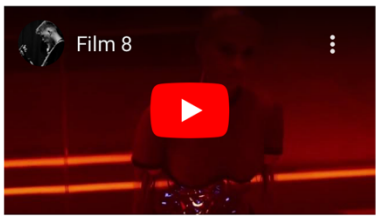
- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 7 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 8 - film



Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

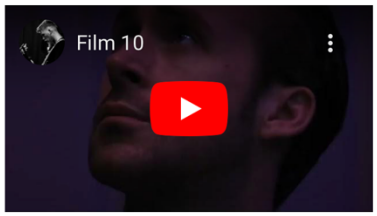
- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 9 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?)

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 10 - film



Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 11 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 12 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 13 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 14 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?)

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 15 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Framgent 16 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 17 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?)

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 18 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Załącznik E: Ankieta III B – dopasowanie emocji fragmentu filmu

Emocje w filmie - emotions in film.

Cześć! Jestem Dawid i na potrzeby rozprawy doktorskiej prosiłbym Ciebie o uzupełnienie krótkiej ankiety dotyczącej przypisania odpowiedniej emocji (jednej z sześciu) do oglądanego fragmentu filmu wraz z jego oryginalnym podkładem muzycznym. Ankieta składa się z 18 pytań. Ta ankieta pozwoli mi na kontynuowanie badań odnośnie porównań kolorystyki filmu, a muzyki filmowej pod kątem utwierdzania w nas pewnych emocji podczas seansu filmowej.

Poniżej jeszcze trzy krótkie pytania:

English version:

Hi! I am Dawid and for the purposes of my doctoral dissertation, I would ask you to complete a short questionnaire on assigning the right emotions (one of six) to film fragment with his original soundtrack. Survey have 18 questions. This survey will allow me to continue my research on the comparisons of film colors and film music in terms of confirming certain emotions in us while watching films.

Twój przedział wiekowy *

- ◯ 15-20
- ◯ 21-30
- ◯ 31-40
- ◯ 41+

Twoja płeć / Your gender *

- ◯ Kobieta / female
- ◯ Mężczyzna / male
- ◯ Inne: _____

Jak mocno jesteś związany z produkcją filmową? / How much You are involved in film production? *

- ◯ Jestem professional reżyserem / filmowcem / aktorem / I am profesional director / filmmaker / actor
- ◯ Film to moje hobby (tworzę amatorskie dzieła) / Film is only my hobby (I create amateurs films)
- ◯ Tylko oglądam filmy / seriale / I only watch films / TV series
- ◯ Nie oglądam filmów / seriali / I don't watch films / TV series

Fragment 1 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 2 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 3 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?)

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 4 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

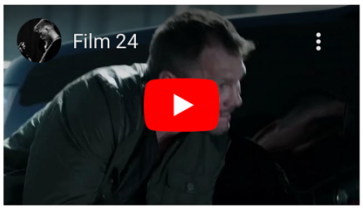
- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 5 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 6 - film



Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 7 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

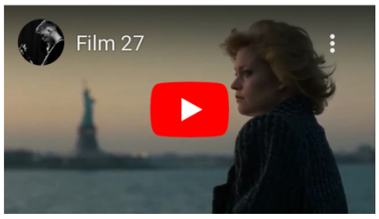
- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 8 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 9 - film



Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?)

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 10 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 11 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

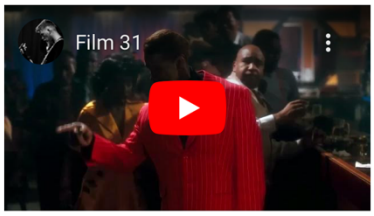
- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 12 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

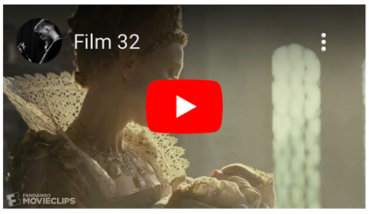
Fragment 13 - film



Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 14 - film



Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?)

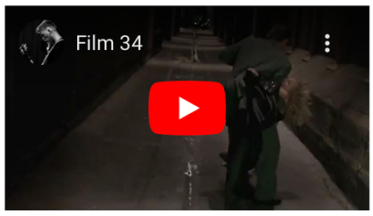
- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 15 - film

Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

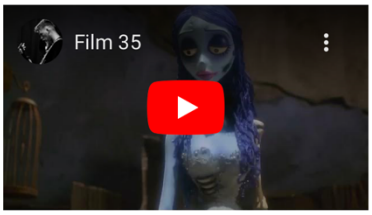
Framgent 16 - film



Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- Szęście
- Strach
- Agresja
- Smutek
- Spokój
- Ekscytacja

Fragment 17 - film



Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?)

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Fragment 18 - film

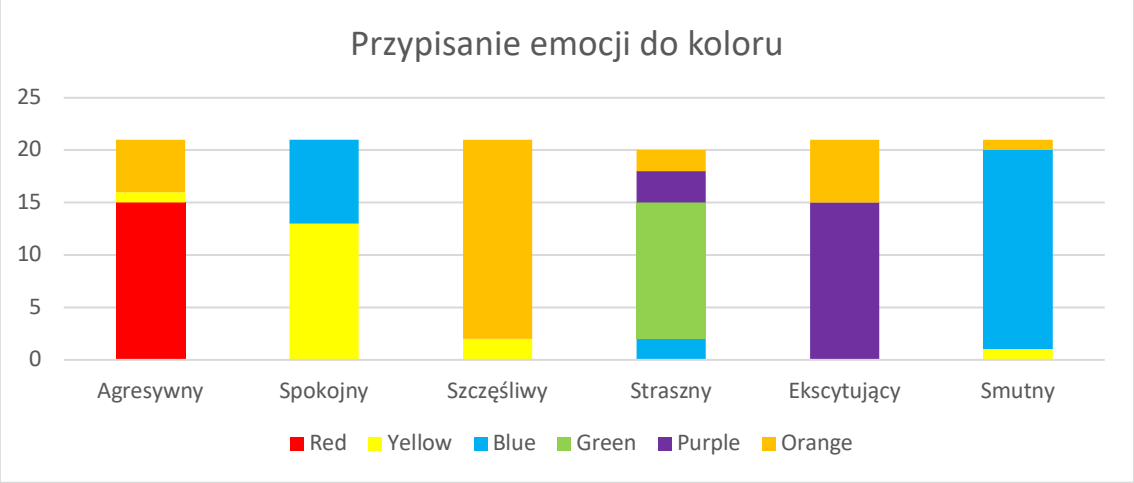
Jaką emocję odczuwasz podczas oglądania tego fragmentu filmu? (What emotion do You feel while watching this film fragment?) *

- ◯ Szęście
- ◯ Strach
- ◯ Agresja
- ◯ Smutek
- ◯ Spokój
- ◯ Ekscytacja

Załącznik F: Wyniki testów subiektywnych dla Ankiety I

Aplikacja	**Agresywny**	**Spokojny**	**Szczęśliwy**	**Straszny**	**Ekscytujący**	**Smutny**
Czerwony	15	0	0	0	0	0
Żółty	1	13	2	0	0	1
Niebieski	0	8	0	2	0	19
Zielony	0	0	0	13	0	0
Fioletowy	0	0	0	3	15	0
Pomarańczowy	5	0	19	2	6	1

Eye Tracker	**Agresywny**	**Spokojny**	**Szczęśliwy**	**Straszny**	**Ekscytujący**	**Smutny**
Czerwony	14	0	0	0	0	0
Żółty	1	13	2	0	0	1
Niebieski	0	7	0	2	0	16
Zielony	0	0	0	13	0	3
Fioletowy	1	0	1	3	15	0
Pomarańczowy	5	1	18	2	6	1



Załącznik G: Wyniki testów subiektywnych dla Ankiety II

LP	Tytuł filmu	Odp	Emocja
1	A Very Long Engagement	26	Smutek
2	Amelie	29	Szczęście
3	Titanic	26	Szczęście
4	The Thin Red Line	19	Ekscytacja
5	Into the Wild	17	Strach
6	Kill Bill vol. 2	23	Ekscytacja
7	Rebel without a Cause	24	Agresja
8	Inglorious Bastards	23	Agresja
9	Ex Machina	16	Agresja
10	Spirited Away	26	Smutek
11	The Revenant	20	Smutek
12	The Grand Budapest Hotel	16	Smutek
13	Chicago	25	Ekscytacja
14	Lost River	23	Ekscytacja
15	La La Land	23	Ekscytacja
16	Hotel Chevalier	30	Szczęście
17	The Age of Innocence	25	Spokoj
18	Gladiator - opening scene	17	Spokoj
19	Harry Potter	27	Szczęscie
20	Don Juan de Marco	21	Spokój
21	The Martian	22	Szczęscie
22	Maleficient	23	Strach
23	Gattaca	25	Strach
24	Taxi Driver	19	Strach
25	Deadpool	28	Agresja
26	Malcom X	18	Ekscytacja
27	Dick and Tracy	21	Agresja
28	The Shawshank Redemption	23	Smutek
29	The Corpse Bride	27	Smutek
30	Working Girl	22	Spokój
31	Gladiator	25	Smutek
32	Cabaret	18	Ekscytacja
33	Rushmore	23	Ekscytacja
34	The Lion King	23	Szczęscie
35	Elisabeth Golden Age	28	Spokój
36	English Patient	27	Spokój

Załącznik H: Wyniki testów subiektywnych dla Ankiety III

LP	Tytuł filmu	Odp	Emocja
1	A Very Long Engagement	17	Szczęscie
2	Amelie	30	Szczęscie
3	Titanic	27	Szczęscie
4	The Thin Red Line	27	Strach
5	Into the Wild	30	Strach
6	Kill Bill vol. 2	28	Strach
7	Rebel without a Cause	29	Agresja
8	Inglorious Bastards	24	Agresja
9	Ex Machina	29	Agresja
10	Spirited Away	28	Smutek
11	The Revenant	30	Smutek
12	The Grand Budapest Hotel	20	Smutek
13	Chicago	29	Ekscytacja
14	Lost River	27	Ekscytacja
15	La La Land	23	Ekscytacja
16	Hotel Chevalier	26	Spokój
17	The Age of Innocence	26	Spokój
18	Gladiator - opening scene	22	Spokój
19	Harry Potter	29	Szczęscie
20	Don Juan de Marco	16	Szczęscie
21	The Martian	25	Szczęscie
22	Maleficient	27	Strach
23	Gattaca	26	Strach
24	Taxi Driver	27	Strach
25	Deadpool	30	Agresja
26	Malcom X	27	Agresja
27	Dick and Tracy	23	Agresja
28	The Shawshank Redemption	30	Smutek
29	The Corpse Bride	30	Smutek
30	Working Girl	29	Smutek
31	Gladiator	23	Ekscytacja
32	Cabaret	19	Ekscytacja
33	Rushmore	21	Ekscytacja
34	The Lion King	19	Spokój
35	Elisabeth Golden Age	30	Spokój
36	English Patient	25	Spokój

Załącznik I: Wyniki testów subiektywnych dla Ankiety II i Ankiety III z porównaniem do wartości z literatury

		Muzyka vs emocja		Muzyka vs film		Wartości z literatury	
LP	**Tytuł filmu**	**Odp.**	**Emocja**	**Odp.**	**Emocja**	**Kolor**	**Emocja**
1	A Very Long Engagement	26	Smutek	17	Szczęście	Pomarańczowy	Szczęście
2	Amelie	29	Szczęście	30	Szczęście	Pomarańczowy	Szczęście
3	Titanic	26	Szczęście	27	Szczęście	Pomarańczowy	Szczęście
4	The Thin Red Line	19	Ekscytacja	27	Strach	Zielony	Straszny
5	Into the Wild	17	Strach	30	Strach	Zielony	Straszny
6	Kill Bill vol. 2	23	Ekscytacja	28	Strach	Zielony	Straszny
7	Rebel without a Cause	24	Agresja	29	Agresja	Czerwony	Agresywny
8	Inglorious Bastards	23	Agresja	24	Agresja	Czerwony	Agresywny
9	Ex Machina	16	Agresja	29	Agresja	Czerwony	Agresywny
10	Spirited Away	26	Smutek	28	Smutek	Niebieski	Smutny
11	The Revenant	20	Smutek	30	Smutek	Niebieski	Smutny
12	The Grand Budapest Hotel	16	Smutek	20	Smutek	Niebieski	Smutny
13	Chicago	25	Ekscytacja	29	Ekscytacja	Fioletowy	Ekscytujący
14	Lost River	23	Ekscytacja	27	Ekscytacja	Fioletowy	Ekscytujący
15	La La Land	23	Ekscytacja	23	Ekscytacja	Fioletowy	Ekscytujący
16	Hotel Chevalier	30	Szczęście	26	Spokój	Żółty	Spokojny
17	The Age of Innocence	25	Spokój	26	Spokój	Żółty	Spokojny
18	Gladiator	17	Spokój	22	Spokój	Żółty	Spokojny
19	Harry Potter	27	Szczęście	29	Szczęście	Pomarańczowy	Szczęście
20	Don Juan de Marco	21	Spokój	16	Szczęście	Pomarańczowy	Szczęście
21	The Martian	22	Szczęście	25	Szczęście	Pomarańczowy	Szczęście
22	Maleficient	23	Strach	27	Strach	Zielony	Straszny
23	Gattaca	25	Strach	26	Strach	Zielony	Straszny
24	Taxi Driver	19	Strach	27	Strach	Zielony	Straszny
25	Deadpool	28	Agresja	30	Agresja	Czerwony	Agresywny
26	Malcom X	18	Ekscytacja	27	Agresja	Czerwony	Agresywny

27	Dick and Tracy	21	Agresja	23	Agresja	Czerwony	Agresywny
28	The Shawshank Redemption	23	Smutek	30	Smutek	Niebieski	Smutny
29	The Corpse Bride	27	Smutek	30	Smutek	Niebieski	Smutny
30	Working Girl	22	Spokój	29	Smutek	Niebieski	Smutny
31	Gladiator	25	Smutek	23	Ekscytacja	Fioletowy	Ekscytujący
32	Cabaret	18	Ekscytacja	19	Ekscytacja	Fioletowy	Ekscytujący
33	Rushmore	23	Ekscytacja	21	Ekscytacja	Fioletowy	Ekscytujący
34	The Lion King	23	Szczęście	19	Spokój	Żółty	Spokojny
35	Elisabeth Golden Age	28	Spokój	30	Spokój	Żółty	Spokojny
36	English Patient	27	Spokój	25	Spokój	Żółty	Spokojny

Załącznik J: Lista publikacji

1. Wędołowska A., Weber D., Kostek B., „Predicting emotion from color present in images and video excerpts by machine learning”, IEEE Access, str. 66357-66473, DOI: 10.1109/ACCESS.2023.3289713, 2023.
2. Ciborowski T., Reginis Sz., Weber D., Kurowski A., Kostek B., „Classifying Emotions in Film Music - A Deep Learning Approach”, Electronics, DOI: 10.3390/electronics10232955, 2021, Szwajcaria.
3. Weber D., „Wpływ kolorystyki ujęć oraz ścieżki dźwiękowej na emocje widza - wstępne eksperymenty”, Akademicka Oficyna Wydawnicza EXIT, str. 99-123, 2019, Warszawa.
4. Weber D., Zaporowski Sz., Korzekwa D., „Constructing a Dataset of Speech Recordings with Lombard Effect”, SPA2020, Institute of Electrical and Electronics Engineers (IEEE), DOI: 10.23919/SPA50552.2020.9241266, 2020, Poznań.
5. Ciborowski T., Reginis Sz., Weber D., Kurowski A., Kostek B., „Klasyfikacja emocji w muzyce filmowej z wykorzystaniem uczenia głębokiego”, Postępy w inżynierii dźwięku i psychoakustyce, Wydawnictwo AGH w Krakowie, 2022, Kraków.
6. Weber D., Kostek B., „Subjective tests for gathering knowledge for applying color grading to video clips automatically”, Signal Processing: Algorithms, Architectures, Arrangements, and Applications, DOI: 10.23919/SPA.2019.8936722, 2019, Poznań.
7. Weber D., Kostek B., „Analiza kolorów scen filmowych w kontekście color gradingu”, Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej, DOI: 10.32016/18.12, 2019, Gdańsk.
8. Weber D., Kostek B., „A Concept of Automatic Film Color Grading Based on Music Recognition and Evoked Emotions”, The International Conference on Digital Image & Signal Processing DISP'19, 2019, Oxford.
9. Blaszke M., Weber D., Zaporowska M., Zaporowski Sz., „Influence of the Delay in Monitor System on the Motor Coordination of Musicians while Performing”, 146th Audio Engineering Society Convention, 2019, Dublin.
10. Weber D., Kostek B., „The influence of sound track on the viewer’s emotions and correction of the color in the film”, XVII Międzynarodowe Sympozjum Nowości w Technice Audio i Wideo, 2018, Poznań.
11. Blaszke M., Weber D., Zaporowski., „Measurement of Latency in the Android Audio Path”, 144th Audio Engineering Society Convention, 2018, Mediolan.
12. Blaszke M., Weber D., Zaporowski Sz., „Pomiary wartości opóźnień w torze audio urządzeń z systemem Android”, Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskie, DOI: 10.32016/1.60.01, 2018, Gdańsk.
13. Koszewski D., Weber D., „Przykład zastosowania przetworników piezoelektrycznych do stworzenia elektronicznych padów na platformę sprzętową Arduino”, 4. Ogólnopolska Studencka Konferencja Akustyków, 2017, Kraków.
14. Blaszke M., Zaporowski S., Weber D., Lech M., „Procesor efektów dźwiękowych do gitary na urządzenia mobilne”, XVI Międzynarodowe Sympozjum Nowości w Technice Audio i Wideo, Rzeszów, 2016.

oraz praca w recenzji:
Weber D., Kostek B., „Harnessing expertise and experience for automatic emotion assignment in films based on a machine-learning approach”, Information Sciences (in review, 2024).