Analiza Big Data informacji pochodzących z urządzeń końcowych STB oraz wykorzystanie metod uczenia maszynowego w celu usprawnieniu działania usługi, analizie doświadczeń klienta oraz zaspokojeniu jego oczekiwań

Big data analysis of information from STB end devices and the use of machine learning methods to improve the operation of the service with the customer experience analysis

Paweł Andruloniw

Wydział Informatyki i Telekomunikacji
Politechnika Poznańska

Rozprawa doktorska

Promotorzy:
dr hab. inż. Piotr Zwierzykowski, prof. PUT
dr inż. Karol Kowalik

Poznań 2025

Podziękowania

Pragnę złożyć serdeczne podziękowania Promotorowi niniejszej rozprawy, Profesorowi Piotrowi Zwierzykowskiemu, oraz Opiekunowi pomocniczemu Karolowi Kowalikowi za nieocenione wsparcie merytoryczne, życzliwość, praktyczne wskazówki oraz inspirację do badań naukowych. Ich zaangażowanie i otwartość towarzyszyły mi na każdym etapie realizacji pracy.

Wyrazy wdzięczności kieruję także do firmy INEA za stworzenie warunków umożliwiających realizację doktoratu wdrożeniowego. W szczególności dziękuję kolegom z Działu Usług Telewizyjnych, których pomoc, życzliwość i otwartość miały istotne znaczenie dla powstania niniejszej rozprawy.

Na końcu pragnę szczególnie podziękować mojej Żonie Darii, za cierpliwość, wyrozumiałość i nieustanne wsparcie, bez którego realizacja tej pracy nie byłaby możliwa.

Niniejsza praca została zrealizowana przy wsparciu finansowym Ministerstwa Nauki i Szkolnictwa Wyższego w ramach programu Doktorat Wdrożeniowy (nr DWD/4/24/2020).

Z wdzięcznością i miłością dedykuję tę pracę mojej Żonie Darii oraz Córce Helenie.

Abstract

HTTP adaptive streaming has become a key technology in modern video delivery systems, ranging from traditional web-based applications, through dedicated OTT platforms, to production-grade television systems. Research related to this service delivery technology remains an active area of investigation in many fields. One of the crucial topics is the assessment of service quality, which covers both the technical aspects and the user-centric dimension associated with the subjective perception of end users. The development and optimization of quality-related components — including video encoding, transmission, and other technical elements — should go hand in hand with the advancement of methods for assessing and estimating user satisfaction.

The subjective perception of video quality, referred to as *Quality of Experience* (QoE), describes how users perceive the service. QoE can be estimated objectively, based on measurable signal parameters such as playback delay, rebuffering events, and bitrate switches, or subjectively, by directly collecting user feedback. The relevant ITU recommendations provide a basis for designing methodologies to obtain such feedback. In subjective tests, users express their opinions on a predefined scale, evaluating quality based on a carefully prepared set of stimuli. This approach, however, yields a limited number of ratings due to the restricted size of the participant group and the time-consuming nature of the process. Based on such a dataset, supervised machine learning techniques can be applied to estimate QoE ratings for the remaining population of users.

The use of supervised machine learning makes it possible to capture nonlinear dependencies present in subjective rating data. The performance of such algorithms is closely related to the system architecture and the availability and quality of input data. In this dissertation, a strategy combining subjective QoE assessment with supervised machine learning has been proposed to estimate the perceived quality of services for users of the Hiway television system. A method for acquiring subjective video quality ratings has been developed, user surveys have been conducted, and a mechanism for automated data collection has been implemented. The collected data were used to design and optimize a multilayer perceptron configured as a soft classifier. The resulting rating distribution reflects subjective user evaluations with an accuracy exceeding 90%. The continuous evolution of television systems encourages further research in new directions, addressing both technical aspects and user experience.

5:5151497299

Streszczenie

Strumieniowanie adaptacyjne HTTP to technologia, która na stałe zagościła w szeroko rozumianych systemach wideo — począwszy od tradycyjnych aplikacji dostępnych z poziomu przeglądarek internetowych, przez dedykowane aplikacje OTT, aż po produkcyjne systemy telewizyjne. Tematyka badań związanych z tą technologią świadczenia usług jest wciąż aktywnym obszarem eksperymentów prowadzonych w wielu dziedzinach. Jednym z kluczowych zagadnień jest badanie jakości usług, obejmujące zarówno aspekt techniczny, jak i użytkowy, powiązany z subiektywnymi odczuciami odbiorców. Rozwój i optymalizacja elementów związanych z jakością — kodowania obrazu, transmisji oraz innych komponentów o charakterze technicznym — powinny iść w parze z udoskonalaniem metod oceny i estymacji poziomu satysfakcji użytkowników korzystających z tej technologii.

Subiektywna percepcja jakości wideo, określana jako *Quality of Experience* (QoE), opisuje sposób, w jaki użytkownicy postrzegają usługę. Jej oszacowania można dokonać w sposób obiektywny, bazując na mierzalnych parametrach sygnału, takich jak opóźnienia odtwarzania, zacięcia czy zmiany przepływności strumieni. W pełni subiektywne podejście polega natomiast na pozyskiwaniu ocen bezpośrednio od użytkowników. Odpowiednie rekomendacje ITU pozwalają na opracowanie metody pozyskiwania tego typu ocen. W badaniach ankietowych użytkownicy usługi wideo wyrażają swoją opinię w zdefiniowanej skali, oceniając jakość na podstawie starannie przygotowanego zestawu bodźców. W ten sposób uzyskuje się jednak ograniczoną liczbę ocen, wynikającą z ograniczonej liczby respondentów i czasochłonności procesu badawczego. Na podstawie takiego zestawu ocen możliwe jest zastosowanie metod nadzorowanego uczenia maszynowego w celu estymacji ocen dla pozostałych użytkowników usługi.

Wykorzystanie nadzorowanych metod uczenia maszynowego umożliwia uchwycenie nieliniowych zależności obecnych w danych z badań. Rozwój tego typu algorytmów jest ściśle powiązany z architekturą systemów oraz zakresem i jakością dostępnych danych pomiarowych. W prezentowanej rozprawie zaproponowano strategię łączącą badania subiektywnego postrzegania jakości z nadzorowanym uczeniem maszynowym w celu estymacji jakości usług QoE dla użytkowników systemu telewizyjnego Hiway. Opracowano metodę pozyskiwania subiektywnych ocen jakości wideo, przeprowadzono badania ankietowe oraz wdrożono mechanizm gromadzenia danych. Zebrane informacje wykorzystano do zaprojektowania i optymalizacji perceptronu wielowarstwowego w konfiguracji miękkiego klasyfikatora. Uzyskany rozkład ocen spełnia założenie wierności odwzorowania subiektywnych ocen użytkowników z dokładnością przekraczającą 90%. Dynamiczny rozwój systemów telewizyjnych stanowi zachętę do dalszej rozbudowy badań w nowych kierunkach, uwzględniających zarówno aspekt techniczny, jak i doświadczenia użytkowników.

6:4820680425

Spis treści

Podziękowania **iii**

Spis rysunków **iv**

Spis tabel **viii**

Lista algorytmów **xi**

1 Wstęp **1**
1.1 Cel rozprawy . . . 1
1.1.1 Hipoteza badawcza . . . 1
1.1.2 Zadania badawcze . . . 1
1.2 Zakres rozprawy . . . 2

2 Systemy telewizji bazujące na strumieniowaniu adaptacyjnym **4**
2.1 Kierunki badawcze dla systemów wideo . . . 4
2.2 Architektura systemu . . . 5
2.3 Ocena Quality of Experience . . . 6
2.4 Podsumowanie . . . 10

3 Subiektywne postrzeganie jakości wideo dla telewizji opartej na strumieniowaniu adaptacyjnym **12**
3.1 Skale ocen . . . 12
3.2 Skala DCR . . . 13
3.3 Skala CCR . . . 13
3.4 Skala CQS . . . 14
3.5 Dostępne metody . . . 14
3.5.1 Metoda pojedynczej stymulacji . . . 14
3.5.2 Wybrane metody podwójnej stymulacji . . . 15
3.6 Powiązane badania . . . 16
3.7 Założenia . . . 18
3.8 Realizacja badania . . . 19
3.8.1 Materiały wideo wykorzystane w sekwencjach . . . 19
3.8.2 Transkodowanie materiałów wideo . . . 19

7:1420884375

3.8.3 Organizacja badania . . . 20
3.8.4 Implementacja na dekoderze . . . 22
3.8.5 Agregacja ocen . . . 23
3.8.6 Dobór próby . . . 24
3.9 Analiza wyników . . . 25
3.9.1 Wyniki dla par kodowanych zgodnie ze standardem H.264 . . . 26
3.9.2 Wyniki dla par kodowanych zgodnie ze standardem H.265 . . . 37
3.9.3 Podsumowanie i wnioski . . . 49

4 Wykorzystanie metod uczenia maszynowego do analizy jakości wideo 52
4.1 Powiązane badania . . . 52
4.2 Nadzorowane uczenie maszynowe . . . 56
4.2.1 Sztuczna sieć neuronowa . . . 56
4.2.2 Funkcja aktywacji . . . 57
4.2.3 Funkcja softmax . . . 58
4.3 Zbiór cech wejściowych . . . 58
4.3.1 Elementy odstające . . . 58
4.3.2 Nowe cechy . . . 61
4.3.3 Macierz korelacji . . . 63
4.3.4 Analiza głównych składowych . . . 65
4.3.5 Liniowa analiza dyskryminacyjna . . . 66
4.4 Bazowy model sieci neuronowej . . . 69
4.5 Wyniki predykcji dla modeli bazowych . . . 74
4.6 Podsumowanie . . . 88

5 Optymalizacja modeli uczenia maszynowego 90
5.1 Optymalizacja dla modeli z jedną warstwą ukrytą . . . 90
5.2 Wyniki dla modeli z jedną warstwą ukrytą . . . 96
5.3 Optymalizacja dla modeli z wieloma warstwami ukrytymi . . . 110
5.4 Wyniki dla modeli z wieloma warstwami ukrytymi . . . 116
5.5 Podsumowanie . . . 130

6 Implementacja narzędzia do analizy QoE 132

7 Podsumowanie 135

Dodatek A Treść wideo wykorzystana w sekwencji nr 1 138

Dodatek B Treść wideo wykorzystana w sekwencji nr 2 141

Dodatek C Plansze informacyjne wykorzystane w sekwencjach badawczych 143
C.1 Grafiki informacyjne dla sekwencji badawczych . . . 143
C.2 Plansze zamienne dla sekwencji badawczej nr 2 . . . 148

8:7922563096

C.3 Implementacja badania na dekoderze 149

Dodatek D Szczegółowy zbiór ocen dla badań subiektywnej jakości wideo 151

Dodatek E Dorobek publikacyjny 154

Bibliografia 156

9:1076951622

Spis rysunków

2.1 Przykładowa architektura systemu telewizji bazującej na strumieniowaniu adaptacyjnym HTTP 6
2.2 Wynik obiektywnej oceny QoE dla referencyjnego STB dla 24–godzinnego interwału czasowego [1] 8
2.3 Wynik obiektywnej oceny QoE dla jednego z STB dla 24–godzinnego interwału czasowego [1] 8
2.4 Mapa cieplna dekoderów przypisanych ocenie 5 skali ACR. Oś x – router agregacyjny, oś y – zakres czasowy od 10:00 do 21:00, oś z – liczba STB (odcień koloru) [2] 9
2.5 Mapa cieplna dekoderów przypisanych ocenie 1 w skali ACR. Oś x – router agregacyjny, oś y – zakres czasowy od 10:00 do 21:00, oś z – liczba STB (odcień koloru) [2] 9

3.1 Organizacja logiczna sekwencji: plansze informacyjne oraz materiały testowe 21
3.2 Organizacja materiałów testowych w pary dla sekwencji 1. Każda kolumna zawiera wartość przepływności bitowej. Kolorem szarym zaznaczono materiały referencyjne. . 22
3.3 Organizacja materiałów testowych w pary dla sekwencji 2. Każda kolumna zawiera wartość przepływności bitowej. Kolorem szarym zaznaczono materiały referencyjne. . 22
3.4 Przykładowe widoki z aplikacji mobilnej do oceny jakości doświadczeń 24
3.5 Formularz ocen w formie papierowej 25
3.6 Rozkład ocen dla pary 1 sekwencji kodowanej zgodnie ze standardem H.264 27
3.7 Rozkład ocen dla pary 2 sekwencji kodowanej zgodnie ze standardem H.264 28
3.8 Rozkład ocen dla pary 3 sekwencji kodowanej zgodnie ze standardem H.264 28
3.9 Rozkład ocen dla pary 4 sekwencji kodowanej zgodnie ze standardem H.264 29
3.10 Rozkład ocen dla pary 5 sekwencji kodowanej zgodnie ze standardem H.264 30
3.11 Rozkład ocen dla pary 6 sekwencji kodowanej zgodnie ze standardem H.264 31
3.12 Rozkład ocen dla pary 7 sekwencji kodowanej zgodnie ze standardem H.264 31
3.13 Rozkład ocen dla pary 8 sekwencji kodowanej zgodnie ze standardem H.264 32
3.14 Rozkład ocen dla pary 9 sekwencji kodowanej zgodnie ze standardem H.264 33
3.15 Rozkład ocen dla pary 10 sekwencji kodowanej zgodnie ze standardem H.264 34
3.16 Rozkład ocen dla pary 11 sekwencji kodowanej zgodnie ze standardem H.264 34
3.17 Rozkład ocen dla pary 12 sekwencji kodowanej zgodnie ze standardem H.264 35
3.18 Rozkład ocen dla pary 13 sekwencji kodowanej zgodnie ze standardem H.264 36
3.19 Rozkład ocen dla pary 14 sekwencji kodowanej zgodnie ze standardem H.264 36

10:4372559642

3.20 Rozkład ocen dla pary 15 sekwencji kodowanej zgodnie ze standardem H.264 37
3.21 Rozkład ocen dla pary 1 sekwencji kodowanej zgodnie ze standardem H.265 38
3.22 Rozkład ocen dla pary 2 sekwencji kodowanej zgodnie ze standardem H.265 39
3.23 Rozkład ocen dla pary 3 sekwencji kodowanej zgodnie ze standardem H.265 39
3.24 Rozkład ocen dla pary 4 sekwencji kodowanej zgodnie ze standardem H.265 40
3.25 Rozkład ocen dla pary 5 sekwencji kodowanej zgodnie ze standardem H.265 41
3.26 Rozkład ocen dla pary 6 sekwencji kodowanej zgodnie ze standardem H.265 42
3.27 Rozkład ocen dla pary 7 sekwencji kodowanej zgodnie ze standardem H.265 42
3.28 Rozkład ocen dla pary 8 sekwencji kodowanej zgodnie ze standardem H.265 43
3.29 Rozkład ocen dla pary 9 sekwencji kodowanej zgodnie ze standardem H.265 44
3.30 Rozkład ocen dla pary 10 sekwencji kodowanej zgodnie ze standardem H.265 45
3.31 Rozkład ocen dla pary 11 sekwencji kodowanej zgodnie ze standardem H.265 46
3.32 Rozkład ocen dla pary 12 sekwencji kodowanej zgodnie ze standardem H.265 46
3.33 Rozkład ocen dla pary 13 sekwencji kodowanej zgodnie ze standardem H.265 47
3.34 Rozkład ocen dla pary 14 sekwencji kodowanej zgodnie ze standardem H.265 48
3.35 Rozkład ocen dla pary 15 sekwencji kodowanej zgodnie ze standardem H.265 49

4.1 Rozkład ocen dla strumieni kodowanych zgodnie ze standardem H.264 60
4.2 Rozkład ocen dla strumieni kodowanych zgodnie ze standardem H.265 61
4.3 Macierz korelacji Pearsona dla analizowanego zestawu danych 64
4.4 Macierz korelacji rang Spearmana dla analizowanego zestawu danych 65
4.5 Separowalność klas dla kombinacji liniowych LD1 i LD2 oraz LD3 i LD4 67
4.6 Separowalność klas dla kombinacji liniowych LD1, LD2 oraz LD3 68
4.7 Separowalność klas dla czterech kombinacji liniowych 69
4.8 Bazowy model sieci neuronowej wykorzystany do dalszych prac 70
4.9 Wartości SROCC dla modelu nr 1, dla którego liczba wejść N = 16 74
4.10 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Model $N = 16$ 75
4.11 Wartości SROCC dla modelu nr 2, dla którego liczba wejść N = 64 76
4.12 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Model $N = 64$ 76
4.13 Wartości SROCC dla modelu nr 3, dla którego liczba wejść N = 256 77
4.14 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Model $N = 256$ 78
4.15 Gęstość rozkładu wartości oczekiwanych 79
4.16 Rozkład uśrednionego prawdopodobieństwa każdej klasy 80
4.17 Wyniki predykcji dla próbki walidacyjnej i = 6 81
4.18 Wyniki predykcji dla próbki walidacyjnej i = 23 82
4.19 Wyniki predykcji dla próbki walidacyjnej i = 89 82
4.20 Wyniki predykcji dla próbki walidacyjnej i = 222 83
4.21 Wyniki predykcji dla próbki walidacyjnej i = 345 84
4.22 Wyniki predykcji dla próbki walidacyjnej i = 587 84

11:1001397255

4.23 Macierz pomyłek na modelu $N = 16$ 86
4.24 Macierz pomyłek na modelu $N = 64$ 87
4.25 Macierz pomyłek dla modelu $N = 256$ 88

5.1 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Zoptymalizowany model z pojedynczą warstwą ukrytą z próby 123 96
5.2 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Zoptymalizowany model z pojedynczą warstwą ukrytą z próby 37 97
5.3 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Zoptymalizowany model z pojedynczą warstwą ukrytą z próby 134 98
5.4 Gęstość rozkładu wartości oczekiwanych dla zoptymalizowanych modeli z pojedynczą warstwą ukrytą 99
5.5 Rozkład uśrednionego prawdopodobieństwa każdej klasy dla zoptymalizowanych modeli z pojedynczą warstwą ukrytą 100
5.6 Wyniki predykcji dla próbki walidacyjnej i = 6 dla zoptymalizowanych modeli z pojedynczą warstwą ukrytą 101
5.7 Wyniki predykcji dla próbki walidacyjnej i = 23 dla zoptymalizowanych modeli z pojedynczą warstwą ukrytą 102
5.8 Wyniki predykcji dla próbki walidacyjnej i = 89 dla zoptymalizowanych modeli z pojedynczą warstwą ukrytą 103
5.9 Wyniki predykcji dla próbki walidacyjnej i = 222 dla zoptymalizowanych modeli z pojedynczą warstwą ukrytą 104
5.10 Wyniki predykcji dla próbki walidacyjnej i = 345 dla zoptymalizowanych modeli z pojedynczą warstwą ukrytą 105
5.11 Wyniki predykcji dla próbki walidacyjnej i = 587 dla zoptymalizowanych modeli z pojedynczą warstwą ukrytą 106
5.12 Macierz pomyłek dla zoptymalizowanego modelu z pojedynczą warstwą ukrytą z próby 123 107
5.13 Macierz pomyłek dla zoptymalizowanego modelu z pojedynczą warstwą ukrytą z próby 37 108
5.14 Macierz pomyłek dla zoptymalizowanego modelu z pojedynczą warstwą ukrytą z próby 134 109
5.15 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Zoptymalizowany wielowarstwowy model z próby 236 . . . 116
5.16 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Zoptymalizowany wielowarstwowy model z próby 226 . . . 117
5.17 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Zoptymalizowany wielowarstwowy model z próby 170 . . . 118

12:9946055434

5.18 Gęstość rozkładu wartości oczekiwanych dla zoptymalizowanych modeli wielowarstwowych . . . 119
5.19 Rozkład uśrednionego prawdopodobieństwa każej klasy dla zoptymalizowanych modeli wielowarstwowych . . . 120
5.20 Wyniki predykcji dla próbki walidacyjnej i = 6 dla zoptymalizowanych modeli wielowarstwowych . . . 121
5.21 Wyniki predykcji dla próbki walidacyjnej i = 23 dla zoptymalizowanych modeli wielowarstwowych . . . 122
5.22 Wyniki predykcji dla próbki walidacyjnej i = 89 dla zoptymalizowanych modeli wielowarstwowych . . . 123
5.23 Wyniki predykcji dla próbki walidacyjnej i = 222 dla zoptymalizowanych modeli wielowarstwowych . . . 124
5.24 Wyniki predykcji dla próbki walidacyjnej i = 345 dla zoptymalizowanych modeli wielowarstwowych . . . 125
5.25 Wyniki predykcji dla próbki walidacyjnej i = 587 dla zoptymalizowanych modeli wielowarstwowych . . . 126
5.26 Macierz pomyłek dla zoptymalizowanego modelu wielowarstwowego z próby 236 . . . 127
5.27 Macierz pomyłek dla zoptymalizowanego modelu wielowarstwowego z próby 226 . . . 128
5.28 Macierz pomyłek dla zoptymalizowanego modelu wielowarstwowego z próby 170 . . . 129

6.1 Schemat narzędzia do analizy Quality of Experience dla treści prezentowanych przez dekodery . . . 132
6.2 Narzędzie do analizy QoE — widok sesji oglądania i średnich ocen dla kanału . . . 133
6.3 Narzędzie do analizy QoE — widok listy audycji . . . 134
6.4 Narzędzie do analizy QoE — widok szczegółowy audycji . . . 134

C.1 Informacja powitalna dla sekwencji numer 1 wyświetlana 17 sekund. . . . 143
C.2 Plansza ze szczegółową instrukcją wyświetlaą przez 40 sekund. W instrukcji przebiegu badania opisano procedurę oceniania treści oraz dostępną skalę ocen. . . . 144
C.3 Powtórzenie kodu QR na planszy trwającej 10 sekund. Sekwencje mają osobne formularze ze względów organizacyjnych związanych z gromadzeniem ocen. . . . 144
C.4 Informacja o rozpoczęciu sekwencji demonstracyjnej składajcej się z 2 par filmów. Plansza wyświetlana przez 8 sekund. . . . 145
C.5 Informacja o rozpoczęciu prezentacji pierwszego filmu testowego w ramach sesji demonstracyjnej. Każdy z 4 filmów posiada osobną planszę wyświetlaną przez 4 sekundy. 145
C.6 Informacja o rozpoczęciu właściwej części badania, czyli prezentacji 15 par materiałów testowych. Plansza wyświetlana przez 10 sekund. . . . 146
C.7 Informacja o rozpoczęciu prezentacji filmu 1A. Każdy z 30 filmów posiada własną planszę wyświetlaną przez 4 sekundy. . . . 146
C.8 Prośba o dokonanie oceny za pomocą formularza. Plansza wyświetlana przez 7 sekund po każej parze filmów. . . . 147

13:2516197006

C.9 Podziękowanie za wzięcie udziału w badaniu. Plansza wyświetlana przez 15 sekund na zakończenie sekwencji. 147
C.10 Informacja powitalna dla sekwencji numer 2 wyświetlana 17 sekund. 148
C.11 Ponowienie kodu QR na planszy trwającej 10 sekund. 148
C.12 Okładki VOD dla sekwencji badawczych 149
C.13 Sekwencje badawcze dostępne w bibliotece VOD usługi Hiway. 150
C.14 Film VOD widoczny na dekoderze. Przycisk **Oglądaj** kieruje użytkownika do badania. 150

Spis tabel

2.1 Parametry zdefiniowane do odtworzenia subiektywnej oceny QoE [1, 2] 7
2.2 Główne bloki architektury systemu telewizji bazującej na strumieniowaniu adaptacyjnym HTTP 11

3.1 Przymiotnikowa skala *ACR* (ang. *Absolute Category Rating*) [3] 13
3.2 Przymiotnikowa skala *DCR* (ang. *Degradation Category Rating*) [4] 13
3.3 Przymiotnikowa skala porównawcza *CCR* (ang. *Comparison Category Rating*) [4] . . 14
3.4 Parametry transkodowania dla sekwencji badawczych 20
3.5 Oceny uzyskane dla pary nr 1 kodowanej zgodnie ze standardem H.264 27
3.6 Oceny uzyskane dla pary nr 2 kodowanej zgodnie ze standardem H.264 28
3.7 Oceny uzyskane dla pary nr 3 kodowanej zgodnie ze standardem H.264 28
3.8 Oceny uzyskane dla pary nr 4 kodowanej zgodnie ze standardem H.264 29
3.9 Oceny uzyskane dla pary nr 5 kodowanej zgodnie ze standardem H.264 30
3.10 Oceny uzyskane dla pary nr 6 kodowanej zgodnie ze standardem H.264 30
3.11 Oceny uzyskane dla pary nr 7 kodowanej zgodnie ze standardem H.264 31
3.12 Oceny uzyskane dla pary nr 8 kodowanej zgodnie ze standardem H.264 32
3.13 Oceny uzyskane dla pary nr 9 kodowanej zgodnie ze standardem H.264 33
3.14 Oceny uzyskane dla pary nr 10 kodowanej zgodnie ze standardem H.264 33
3.15 Oceny uzyskane dla pary nr 11 kodowanej zgodnie ze standardem H.264 34
3.16 Oceny uzyskane dla pary nr 12 kodowanej zgodnie ze standardem H.264 35
3.17 Oceny uzyskane dla pary nr 13 kodowanej zgodnie ze standardem H.264 35
3.18 Oceny uzyskane dla pary nr 14 kodowanej zgodnie ze standardem H.264 36
3.19 Oceny uzyskane dla pary nr 15 kodowanej zgodnie ze standardem H.264 37
3.20 Oceny uzyskane dla pary nr 1 kodowanej zgodnie ze standardem H.265 38
3.21 Oceny uzyskane dla pary nr 2 kodowanej zgodnie ze standardem H.265 38
3.22 Oceny uzyskane dla pary nr 3 kodowanej zgodnie ze standardem H.265 39

14:1807900307

3.23 Oceny uzyskane dla pary nr 4 kodowanej zgodnie ze standardem H.265 40
3.24 Oceny uzyskane dla pary nr 5 kodowanej zgodnie ze standardem H.265 41
3.25 Oceny uzyskane dla pary nr 6 kodowanej zgodnie ze standardem H.265 41
3.26 Oceny uzyskane dla pary nr 7 kodowanej zgodnie ze standardem H.265 42
3.27 Oceny uzyskane dla pary nr 8 kodowanej zgodnie ze standardem H.265 43
3.28 Oceny uzyskane dla pary nr 9 kodowanej zgodnie ze standardem H.265 44
3.29 Oceny uzyskane dla pary nr 10 kodowanej zgodnie ze standardem H.265 45
3.30 Oceny uzyskane dla pary nr 11 kodowanej zgodnie ze standardem H.265 45
3.31 Oceny uzyskane dla pary nr 12 kodowanej zgodnie ze standardem H.265 46
3.32 Oceny uzyskane dla pary nr 13 kodowanej zgodnie ze standardem H.265 47
3.33 Oceny uzyskane dla pary nr 14 kodowanej zgodnie ze standardem H.265 48
3.34 Oceny uzyskane dla pary nr 15 kodowanej zgodnie ze standardem H.265 48
3.35 Oceny uczestników badania pogrupowane dla zastosowanych poziomów jakości . . . 50

4.1 Zestaw cech wykorzystany do predykcji 63
4.2 Wynik analizy PCA 66
4.3 Wynik analizy LDA 67
4.4 Parametry dla próbki $i = 6$ 81
4.5 Parametry dla próbki $i = 23$ 82
4.6 Parametry dla próbki $i = 89$ 83
4.7 Parametry dla próbki $i = 222$ 83
4.8 Parametry dla próbki $i = 345$ 84
4.9 Parametry dla próbki $i = 587$ 85
4.10 Metryki dla modelu $N = 16$ 86
4.11 Metryki dla modelu $N = 64$ 87
4.12 Metryki dla modelu $N = 256$ 88

5.1 Hiperparametry optymalizacyjne do poszukiwania optymalnej sieci neuronowej . . . 90
5.2 50 najlepszych modeli spośród 192 prób optymalizacji. Modele zostały posortowane malejąco według osiągniętego SROCC w pierwszej fazie 94
5.3 Parametry dla próbki $i = 6$ 101
5.4 Parametry dla próbki $i = 23$ 102
5.5 Parametry dla próbki $i = 89$ 103
5.6 Parametry dla próbki $i = 222$ 104
5.7 Parametry dla próbki $i = 345$ 105
5.8 Parametry dla próbki $i = 587$ 106
5.9 Metryki dla zoptymalizowanego modelu z pojedynczą warstwą ukrytą z próby 123 . 107
5.10 Metryki dla zoptymalizowanego modelu z pojedynczą warstwą ukrytą z próby 37 . . 108
5.11 Metryki dla zoptymalizowanego modelu z pojedynczą warstwą ukrytą z próby 134 . 110
5.12 Typy modeli testowanych w drugim podejściu optymalizacyjnym 111
5.13 50 najlepszych modeli spośród 200 prób optymalizacji dla architektury wielowarstwowej. Modele zostały posortowane malejąco według uśrednionych metryk 112

15:6395869471

5.14 Wartości metryk zoptymalizowanych wielowarstwowych modeli po 10-krotnej walidacji krzyżowej. Wartości obliczone dla wartości oczekiwanej wartości prawdziwych i predykcji 113
5.15 Parametry dla próbki $i = 6$ . . . 121
5.16 Parametry dla próbki $i = 23$ . . . 122
5.17 Parametry dla próbki $i = 89$ . . . 123
5.18 Parametry dla próbki $i = 222$ . . . 124
5.19 Parametry dla próbki $i = 345$ . . . 125
5.20 Parametry dla próbki $i = 587$ . . . 126
5.21 Metryki dla zoptymalizowanego modelu wielowarstwowego z próby 236 . . . 127
5.22 Metryki dla zoptymalizowanego modelu wielowarstwowego z próby 226 . . . 128
5.23 Metryki dla zoptymalizowanego modelu wielowarstwowego z próby 170 . . . 129
5.24 Porównanie wybranych modeli przedstawionych w rozdziale . . . 130

6.1 Parametry wdrożonego modelu z próby 236 . . . 132

A.1 Materiał audio wykorzystany w trakcie sekwencji badawczej nr 1 . . . 138
A.2 Treści wideo wykorzystane w sekwencji badawczej nr 1 dla kodowania zgodnie ze standardem H.264 . . . 138

B.1 Materiał audio wykorzystany w trakcie sekwencji badawczej nr 2 . . . 141
B.2 Treści wideo wykorzystane w sekwencji badawczej nr 2 dla kodowania zgodnie ze standardem H.265 . . . 141

D.1 Oceny uzyskane od uczestników badania dla sekwencji z materiałami wideo kodowanymi H.264. Skrajna lewa kolumna określa numer pary, natomiast nagłówek A i B określa materiał w ramach danej pary. . . . 152
D.2 Oceny uzyskane od uczestników badania dla sekwencji z materiałami wideo kodowanymi H.265. Skrajna lewa kolumna określa numer pary, natomiast nagłówek A i B określa materiał w ramach danej pary. . . . 152
D.3 Różnice ocen pomiędzy materiałami A i B dla danej pary określonej przez skrajną lewą kolumnę. Wartość ujemna oznacza spadek oceny dla materiału B w stosunku do A. Wartość dodatnia to wzrost oceny. Wartości dla sekwencji kodowanej H.264. . . . 153
D.4 Różnice ocen pomiędzy materiałami A i B dla danej pary określonej przez skrajną lewą kolumnę. Wartość ujemna oznacza spadek oceny dla materiału B w stosunku do A. Wartość dodatnia to wzrost oceny. Wartości dla sekwencji kodowanej H.265. . . . 153

E.1 Podsumowanie punktacji publikacji autora: punkty MEiN oraz punkty po uwzględnieniu wyższej wartości w przypadku podwójnej punktacji. . . . 155

16:8557691735

Lista algorytmów

1 Uczenie bazowej sieci neuronowej do klasyfikacji z miękkimi etykietami 71

2 Optymalizacja sieci neuronowej dla modeli z jedną warstwą ukrytą 91
3 Uczenie sieci neuronowej do klasyfikacji z miękkimi etykietami. Podejście z podziałem na zbiory: treningowy, walidacyjny i testowy w proporcjach 3:1:1 92
4 Uczenie sieci neuronowej do klasyfikacji z miękkimi etykietami. Podejście z 10-cio krotną walidacją krzyżową 93
5 Optymalizacja sieci neuronowej dla modeli z różną liczbą warstw ukrytych 111

17:5383284601

Rozdział 1

Wstęp

1.1 Cel rozprawy

Celem niniejszej rozprawy doktorskiej, mającej charakter wdrożeniowy, jest zbadanie możliwości zastosowania metod uczenia maszynowego do estymacji jakości usług telewizyjnych opartych na strumieniowaniu adaptacyjnym HTTP. Jako miarę jakości przyjęto *QoE* (ang. *Quality of Experience*), rozumianą jako subiektywną ocenę jakości odbioru dokonywaną przez użytkowników produkcyjnego systemu Hiway firmy INEA[1].

1.1.1 Hipoteza badawcza

Na podstawie powyższego celu sformułowano następującą hipotezę badawczą:

> **Hipoteza badawcza:** Możliwe jest dokonanie estymacji jakości usług telewizyjnych na podstawie analizy szerokiego spektrum parametrów dostarczanych przez system *OTT* (ang. *over-the-top*), w tym danych pochodzących z urządzeń końcowych. Zakłada się, że uzyskana w ten sposób estymata będzie odzwierciedlać subiektywną ocenę użytkowników z dokładnością przekraczającą 90%. Tak wyznaczona predykcja jakości, realizowana z wykorzystaniem metod uczenia maszynowego, może stanowić podstawę działań ukierunkowanych na poprawę jakości usług telewizyjnych.

1.1.2 Zadania badawcze

W celu weryfikacji hipotezy badawczej przyjęto następujące etapy badań:

- Analiza potencjału zastosowania nienadzorowanych metod uczenia maszynowego do detekcji degradacji jakości w sieci telewizyjnej.
- Opracowanie metod gromadzenia oraz przetwarzania danych pochodzących z dekoderów STB.
- Opracowanie metodologii pozyskiwania subiektywnych ocen jakości wideo.

[1] Ocena QoE została oparta na danych generowanych przez urządzenia klienckie typu *STB* (ang. *set-top box*).

18:1854843581

- Przeprowadzenie badań ankietowych oraz wdrożenie mechanizmu zbierania ocen użytkowników systemu Hiway.
- Projektowanie i optymalizacja metod nadzorowanego uczenia maszynowego w celu estymacji QoE na podstawie zgromadzonych danych.
- Opracowanie oraz przygotowanie do wdrożenia narzędzi służących do estymacji QoE w produkcyjnym systemie Hiway.

1.2 Zakres rozprawy

Rozprawa składa się z siedmiu rozdziałów oraz dodatków:

- Rozdział 2 opisuje architekturę systemu, dostępne dane oraz podejście do uczenia nienadzorowanego.
  - Podrozdział 2.2 przedstawia zarys systemu telewizyjnego opartego na strumieniowaniu adaptacyjnym HTTP.
  - Podrozdział 2.3 definiuje pojęcie *Quality of Experience* i jego znaczenie w kontekście badań.
  - Podrozdział 2.4 podsumowuje architekturę, stanowiącą podstawę dalszych analiz.
- Rozdział 3 prezentuje realizację badania subiektywnej oceny jakości w systemie Hiway.
  - Podrozdział 3.1 omawia dostępne skale ocen zgodne z rekomendacjami ITU.
  - Podrozdział 3.5 przedstawia metody badania percepcji jakości treści wideo.
  - Podrozdział 3.6 stanowi przegląd literatury dotyczącej badań subiektywnej percepcji jakości.
  - Podrozdział 3.7 formułuje założenia dotyczące implementacji badania jakości.
  - Podrozdział 3.8 opisuje podejście do wdrożenia badania w systemie Hiway.
  - Podrozdział 3.9 podsumowuje uzyskane wyniki badań.
- Rozdział 4 omawia zastosowanie metod uczenia nadzorowanego w celu estymacji subiektywnej jakości usług.
  - Podrozdział 4.1 przedstawia przegląd badań związanych z estymacją jakości usług wideo.
  - Podrozdział 4.2 opisuje podejście do wykorzystania uczenia nadzorowanego.
  - Podrozdział 4.3 omawia przygotowanie danych niezbędnych do trenowania modelu bazowego.
  - Podrozdział 4.4 przedstawia proces uczenia bazowej sieci neuronowej.
  - Podrozdział 4.5 zawiera analizę wyników predykcji.
  - Podrozdział 4.6 podsumowuje uzyskane rezultaty i wskazuje kierunki dalszych usprawnień.

19:8507610852

- Rozdział 5 poświęcony jest optymalizacji sieci neuronowych w celu uzyskania modeli gotowych do wdrożenia i spełniających założenia hipotezy.
  - Podrozdział 5.1 prezentuje optymalizację modeli o uproszczonej architekturze.
  - Podrozdział 5.2 przedstawia wyniki modeli z pojedynczą warstwą ukrytą.
  - Podrozdział 5.3 omawia optymalizację modeli wielowarstwowych.
  - Podrozdział 5.4 prezentuje wyniki predykcji dla modeli zoptymalizowanych.
  - Podrozdział 5.5 zawiera porównanie modeli i wskazuje jeden rekomendowany do wdrożenia.
- Rozdział 6 przedstawia propozycję wdrożenia systemu estymacji subiektywnej jakości usług na poziomie urządzeń STB.
- Rozdział 7 podsumowuje całość pracy, omawia uzyskane wyniki i wskazuje kierunki dalszych badań.
- Dodatki:
  - Dodatek A zawiera zestaw materiałów wykorzystanych w pierwszej sekwencji badawczej.
  - Dodatek B prezentuje materiały z drugiej sekwencji badawczej.
  - Dodatek C obejmuje przygotowane plansze instruktażowe wykorzystywane podczas badań.
  - Dodatek D zawiera szczegółowy zbiór ocen uzyskanych w trakcie badań subiektywnej percepcji jakości wideo.
  - Dodatek E przedstawia spis publikacji autora.

20:2718609168

Rozdział 2

Systemy telewizji bazujące na strumieniowaniu adaptacyjnym

Technologia strumieniowania adaptacyjnego HTTP jest stosowana nie tylko w aplikacjach *OTT* (ang. *over-the-top*), takich jak Netflix [5], Disney Plus [6], HBO Max [7] czy Amazon Prime [8], lecz także w nowoczesnych systemach telewizyjnych oferujących funkcje przesunięcia w czasie (ang. *time-shift*).

2.1 Kierunki badawcze dla systemów wideo

Systemy telewizyjne są jedną z kluczowych gałęzi nowoczesnych rozwiązań do świadczenia usług wideo. Po pandemii COVID-19 znaczący rozwój zanotowały platformy strumieniowania w czasie rzeczywistym *LVS* (ang. *Live Video Streaming*). Integracja systemów LVS z siecią dostępową operatora może pozwolić na ominięcie sieci szkieletowej i zmniejszenie opóźnień [9]. Aktywnymi obszarami badań w tym zakresie są: architektura sieciowa, algorytmy adaptacji strumieni, mechanizmy dostępu do usługi, zarządzanie QoS/QoE oraz utrzymanie użytkowników [10].

Zagadnienia architektury end-to-end obejmują także przetwarzanie wstępne (ang. *ingest*), transkodowanie, rozwój algorytmów adaptacji oraz systemów *CDN* (ang. *Content Delivery Network*) [11]. Wielodostępowe przetwarzanie brzegowe *MEC* (ang. *Multi-access Edge Computing*) [12] jest technologią, która może poprawić poziom QoE w usługach strumieniowania adaptacyjnego – zarówno poprzez pełnienie roli pamięci cache, jak i redukcję opóźnień. MEC może być również wykorzystywane do transkodowania segmentów wideo w sytuacji, gdy żądana jakość nie jest dostępna [13].

Rosnącym obszarem badań są także systemy wideo 360°, odtwarzane w okularach *HMD* (ang. *Head Mounted Display*). Badania koncentrują się na zapewnieniu odpowiedniego QoS i QoE, metodach oceny obiektywnej *VQA* (ang. *Video Quality Assessment*) oraz subiektywnych badaniach percepcji jakości [14–16]. Wzrost znaczenia odnotowują również systemy strumieniowania gier wideo, gdzie kluczowe są niski czas reakcji i wysoka jakość renderowanego obrazu [17–19].

Uczenie maszynowe znajduje zastosowanie zarówno w kompresji wideo [20], ocenie jakości QoE [21], optymalizacji sieci SDN [22], jak i w silnikach rekomendacji treści [23]. Kierunki rozwoju usług

21:1073211543

telewizyjnych i wideo można śledzić m.in. podczas corocznych targów *IBC* (ang. *International Broadcasting Convention*) w Amsterdamie [24, 25].

2.2 Architektura systemu

Na rysunku 2.1 przedstawiono przykładową architekturę systemu telewizji bazującej na strumieniowaniu adaptacyjnym. Architekturę można podzielić na dwie główne części: system telewizyjny oraz sieć *CDN* (ang. *Content Delivery Network*).

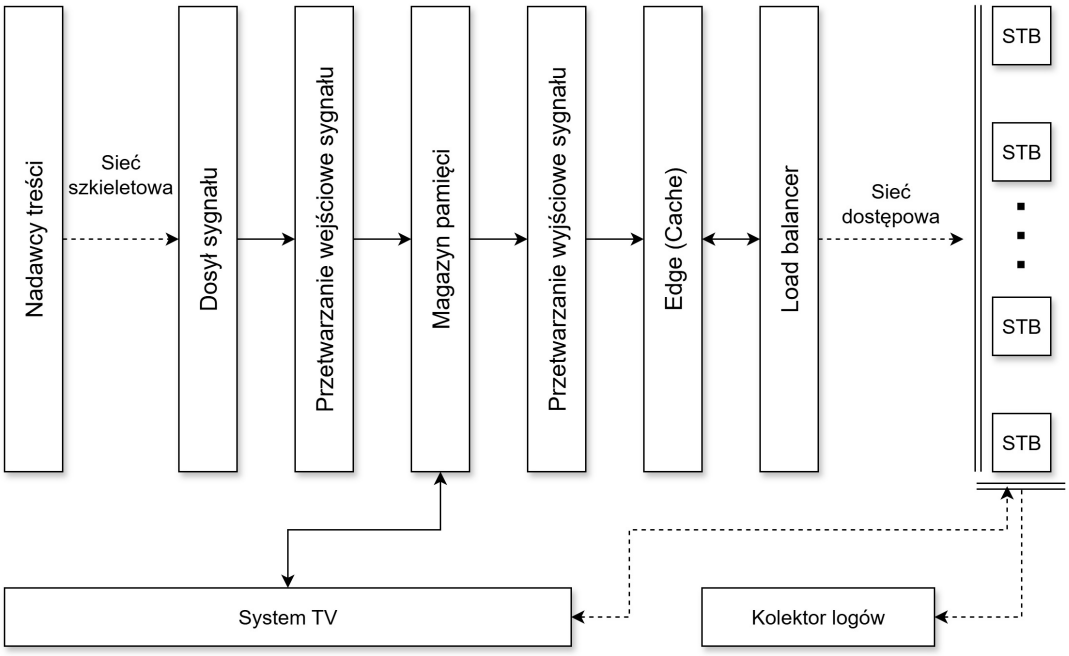
System składa się z następujących bloków:

- **Nadawcy treści** – tworzą i przesyłają sygnał kanałów liniowych do operatorów, najczęściej jako transmisję multicast. Medium transmisyjnym może być światłowód, transmisja satelitarna DVB [26] lub inne technologie. Treść wideo jest zazwyczaj kodowana w standardach H.264 [27] lub H.265 [28], a audio w MP2 [29], *AAC* (ang. *Advanced Audio Coding*) [30, 31] lub AC-3 [32].
- **Dosył sygnału** – odpowiada za niezawodny odbiór sygnału po stronie operatora i wykorzystuje mechanizmy redundancji (np. *VRRP* (ang. *Virtual Router Redundancy Protocol*) [33], *LACP* (ang. *Link Aggregation Control Protocol*) [34], *MPLS TE* (ang. *Traffic Engineering*), *FRR* (ang. *Fast Reroute*) [35]).
- **Przetwarzanie wejściowe** – obejmuje transkodowanie sygnału, zmianę standardów kodowania i przygotowanie do segmentacji wideo.
- **Magazyn pamięci** – przechowuje segmenty treści oraz nagrania zlecone przez użytkowników (PVR w chmurze) oraz umożliwia funkcję cofania programów (ang. *time-shift*) i dostęp do treści VOD.
- **Przetwarzanie wyjściowe** – odpowiada za wysyłkę segmentów do serwerów brzegowych i zarządzanie licencjami DRM (np. Google Widevine [36], Microsoft PlayReady [37], Apple FairPlay [38]).
- **Serwery brzegowe (Edge)** – pełnią funkcję pamięci podręcznej (cache) i dostarczają treści do dekoderów w standardach *MPEG-DASH* (ang. *Dynamic Adaptive Streaming over HTTP*), *Apple HLS* (ang. *HTTP Live Streaming*) i innych.
- **Load balancer** – równoważy obciążenie pomiędzy serwerami brzegowymi, zapewniając niskie opóźnienia i stabilny dostęp do treści [39].
- **System TV** – zespół serwerów i usług backendowych odpowiedzialnych za autoryzację i obsługę żądań dekoderów.
- **Kolektor logów** – zbiera dane telemetryczne z całego systemu, w tym z dekoderów, serwerów CDN i systemu TV.

Architektura może być realizowana zarówno na dedykowanych serwerach sprzętowych, jak i w środowiskach wirtualnych lub chmurowych. Urządzenia końcowe nie muszą być fizycznymi

22:1130617145

dekoderami – coraz częściej są to aplikacje działające na telewizorach typu Smart TV lub urządzeniach mobilnych.



Rysunek 2.1 Przykładowa architektura systemu telewizji bazującej na strumieniowaniu adaptacyjnym HTTP

2.3 Ocena Quality of Experience

Szacowanie jakości w systemach telewizyjnych może obejmować zarówno parametry jakości usług *QoS* (ang. *Quality of Service*) – takie jak opóźnienie, zmienność opóźnień czy utrata pakietów [39] – jak i parametry jakości doświadczeń *QoE* (ang. *Quality of Experience*), które opisują subiektywne postrzeganie jakości. Do najważniejszych czynników wpływających na QoE należą: zacięcia w odtwarzaniu, zmiany jakości strumienia oraz opóźnienia rozpoczęcia odtwarzania [40]. Badania wskazują, że zacięcia wpływają na spadek QoE nawet sześciokrotnie silniej niż wydłużenie czasu rozpoczęcia odtwarzania [41].

Ocena QoE może mieć charakter subiektywny (na podstawie opinii użytkowników) lub obiektywny (na podstawie mierzonych parametrów sygnału i sieci) [39]. Celem jest jak najlepsze odwzorowanie doświadczeń użytkowników przy pomocy danych pomiarowych.

Dane pochodzące z dekoderów związane ze zmianą odgrywanych zdarzeń, wykorzystaniem funkcji nieliniowych, czasem odtwarzania określonych profili wideo oraz wystąpieniem zacięć (czarny ekran z błędem odtwarzania) zostały wykorzystane do zdefiniowana czterech parametrów (2.1) [1, 2].

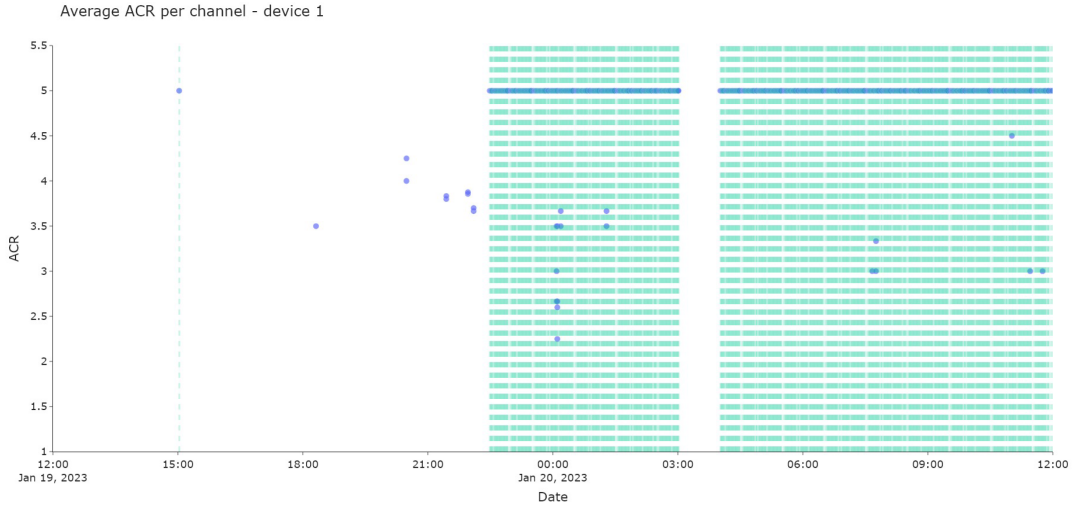
23:3408486473

Tabela 2.1 Parametry zdefiniowane do odtworzenia subiektywnej oceny QoE [1, 2]

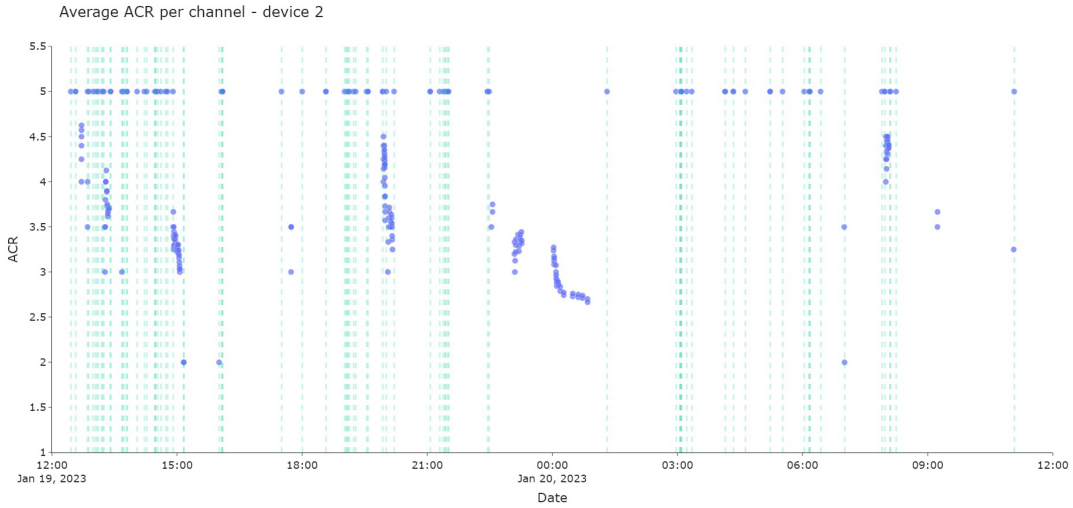
Parametr	Opis
SCI	Określa trend zmienności strumienia wideo w trakcie odtwarzania treści. Dodatnie wartości wskazuję na poprawę jakości, ujemne na degradację. Wartości skrajne oznaczają najlepszą jakość (1) lub zanik sygnału (-1).
$SCTI$	Współczynnik uwzględnia popularność danego kanału i czas odtwarzania strumienia określonego poziomu. W czasie odtwarzania profili o niskiej jakości, współczynnik ten przyjmuje wartości ujemne. Brak uzyskania strumienia o najwyższej jakości powoduje wzmocnienie współczynnika, przez co osiąga niższe wartości. Oglądalnośc kanału jest brana pod uwagę, gdyż często większa liczba użytkowników na danym kanale oznacza ważną audycje. Degradacje w takim momencie jest szczególnie niepożądana.
$VSBCT$	Współczynnik biorący pod uwagę częstotliwość zmian profili oraz występowania zacięć. Zwiększona liczba zdarzeń w krótkim okresie może mieć większy wpływ na postrzeganą jakość niż niewielka liczba pojedynczych zdarzeń w trakcie całej sesji oglądania. Znaczna liczba zdarzeń w krótkim okresie spowoduje znaczny spadek wartości wskaźnika. Gdy jakość na STB osiągnie maksymalny poziom, wartości charakterystyk zaczynają się poprawiać i po pewnym czasie osiągają wartość maksymalną.
$avErr_t$	Licznik zacięć (przerwa w sygnale) w trakcie czasu t, określającego czas trwania pojedynczego eventu, programu oraz sesji oglądania – od włączenia do wyłączenia dekodera.

Metryki te zostały wykorzystane w nienadzorowanym uczeniu maszynowym. Uzyskiwany w czasie rzeczywistym strumień danych z dekodera grupowano w sposób ciągły dwustopniowym algorytmem bazującym na DStream [42] oraz algorytmie hierarchicznym. W pierwszej fazie obliczono podsumowanie dla strumienia, aktualizowane przy każdej napływającej próbce danych. W kolejnej fazie agregowano podsumowanie strumienia, w celu uzyskania docelowych powiązań. Metoda aglomeracyjna i DStream została wybrana ze względu na szybkość i elastyczność [2]. Grupowanie wykonano w skali ACR.

24:8331545522



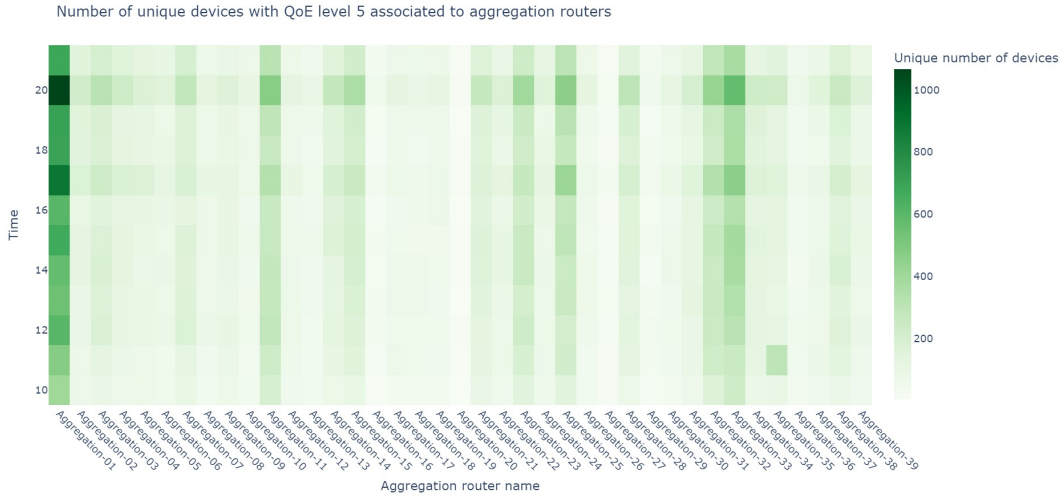
Rysunek 2.2 Wynik obiektywnej oceny QoE dla referencyjnego STB dla 24–godzinnego interwału czasowego [1]



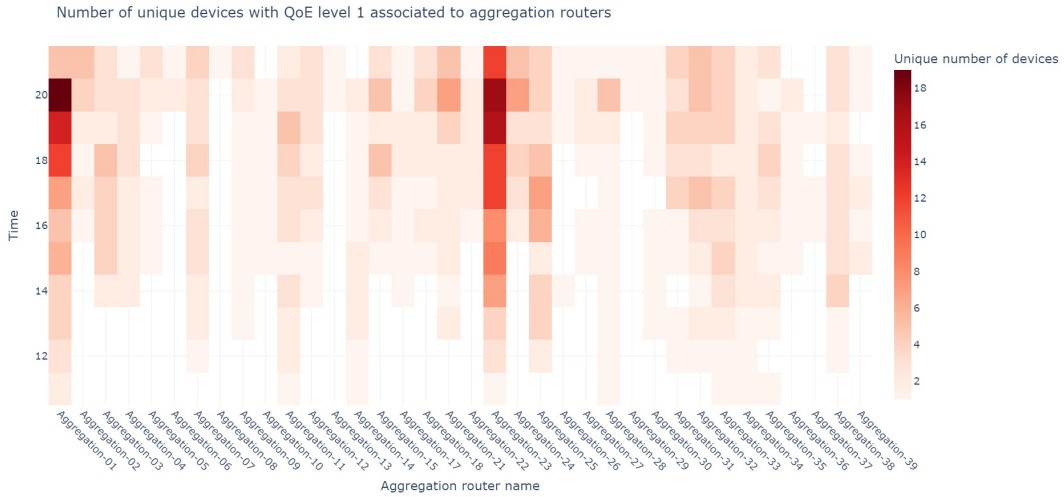
Rysunek 2.3 Wynik obiektywnej oceny QoE dla jednego z STB dla 24–godzinnego interwału czasowego [1]

Przykładowe grupowanie jakości obiektywnej oceny QoE zaprezentowano na rysunku 2.3. Pionowe linie oznaczają zmianę kanału. Dla 24–godzinnego interwału czasowego, oceniana jest jakość dla całego kanału, oglądanego przez użytkownika. Na urządzeniu obserwowano znaczną degradację jakości w przedziałach od 12:00 do 15:00, następnie od 18:00 do 21:00 oraz w okolicach północy. W tym czasie widać postępujący spadek oceny z najwyższej wartości do oceny niższej niż 3. Na urządzeniu referencyjnym (2.2) znajdującym się w firmowej serwerowni nie odnotowano problemów dla dwóch pierwszych przedziałów czasowych. Okolice północy to czas, w trakcie którego często następuje rozpoczęcie prac serwisowych, wtedy możliwe są wystąpienia zaplanowanych degradacji na poziomie globalnym.

25:1031580779



Rysunek 2.4 Mapa cieplna dekoderów przypisanych ocenie 5 skali ACR. Oś x – router agregacyjny, oś y – zakres czasowy od 10:00 do 21:00, oś z – liczba STB (odcień koloru) [2]



Rysunek 2.5 Mapa cieplna dekoderów przypisanych ocenie 1 w skali ACR. Oś x – router agregacyjny, oś y – zakres czasowy od 10:00 do 21:00, oś z – liczba STB (odcień koloru) [2]

Dekodery podłączone są do routerów agregacyjnych poprzez *ONT* (ang. *Optical Network Terminal*). Mapa cieplna na rysunku 2.4 przedstawia dekodery z przypisaną grupą odpowiadającą ocenie najwyższej (kolor zielony) w skali ACR, które przypisane są do określonych routerów agregacyjnych. Na rysunku 2.5 zaprezentowano sytuację odwrotną, w której dekodery uzyskują ocenę najniższą (kolor zielony). STB z najwyższa oceną są widoczne na każdym routerze agregacyjnym. Wyróżnia się jeden z nich (*Aggregation-01*) i znajduje się w centrum Poznania, do którego podłączona jest największa liczba dekoderów. Dwa routery agregacyjne zostały zidentyfikowane ze zwiększoną liczbą negatywnych ocen dla podłączonych STB. Dla urządzenia pierwszego (*Aggregation-01*) wydaje się, że jest to efekt liczby urządzeń, podobnie jak w pozytywnym przypadku (oceny najwyższe). Taka zależność nie występuje dla drugiego urządzenia (*Aggregation-23*), dla którego zaobserwowano tylko

26:5745719300

zwiększoną liczbę negatywnych ocen. Występowanie problemów powinno zostać zbadane przez zespół sieciowy opiekujący się danym routerem agregacyjnym.

2.4 Podsumowanie

Przedstawiona architektura systemu telewizji bazującej na strumieniowaniu adaptacyjnym HTTP pokazuje wielowarstwową strukturę, w której kluczowe znaczenie mają serwery brzegowe, pamięć masowa oraz systemy zapewniające QoE i bezpieczeństwo treści. Odpowiednia konfiguracja i monitorowanie tych elementów są kluczowe dla zapewnienia wysokiej jakości doświadczeń użytkowników. Estymacja obiektywnego wskaźnika QoE poprzez metody uczenia nienadzorowanego może być skuteczną metodą do wyszukiwania problemów w infrastrukturze dostawcy usług. Zastosowanie może być szczególnie przydatne wtedy, gdy nie ma dostępnych etykiet w zbiorze danych, umożliwiających zastosowanie metod nadzorowanych. Zaletą algorytmów grupowania ciągłego strumienia danych (ang. *stream clustering*) jest możliwość działania w systemach czasu rzeczywistego.

27:4028274924

Tabela 2.2 Główne bloki architektury systemu telewizji bazującej na strumieniowaniu adaptacyjnym HTTP

Blok systemu	**Opis**
Nadawcy treści	Tworzą kanały liniowe i przesyłają sygnał (np. w MPEG2-TS, SRT, RIST) do operatora za pomocą transmisji światłowodowej, satelitarnej DVB lub innych mediów. Kodowanie wideo najczęściej w H.264/H.265, audio w MP2, AAC lub AC-3.
Dosył sygnału	Zapewnia niezawodny odbiór sygnału po stronie operatora. Wykorzystuje mechanizmy redundancji i protokoły takie jak VRRP, LACP, MPLS TE/VPN, FRR.
Przetwarzanie wejściowe	Dostosowuje odebrany sygnał do platformy TV: transkodowanie, readresacja, zmiana tablic w strumieniu transportowym. Przygotowuje segmentację wideo dla CDN.
Magazyn pamięci	Macierze dyskowe przechowujące segmenty treści, nagrania (PVR w chmurze) oraz bufor dla funkcji cofania programów (*time-shift*) i usług VOD.
Przetwarzanie wyjściowe	Wysyła segmenty do serwerów brzegowych. W tym etapie stosowane są mechanizmy DRM (Google Widevine, Microsoft PlayReady, Apple FairPlay) w celu ochrony treści.
Serwery brzegowe (Edge)	Strumieniują treści do dekoderów. Mogą tworzyć architekturę wielowarstwową (cache), zmniejszając obciążenie macierzy dyskowych. Obsługują standardy MPEG-DASH, HLS, Smooth Streaming, CMAF.
Load balancer	Rozdziela ruch pomiędzy serwery brzegowe, zapewniając niskie opóźnienia i równomierne obciążenie systemu. Kluczowy element dostępności usług.
System TV	Backend systemu odpowiedzialny za autoryzację, obsługę żądań dekoderów i zarządzanie usługami (np. EPG, nagrywanie).
Kolektor logów	Zbiera dane telemetryczne z całego systemu: dekoderów, serwerów CDN i backendu. Służy do monitorowania QoE i diagnostyki.

28:2316720532

Rozdział 3

Subiektywne postrzeganie jakości wideo dla telewizji opartej na strumieniowaniu adaptacyjnym

Implementacja badania jakości wideo w produkcyjnym systemie telewizyjnym Hiway firmy INEA miała na celu analizę subiektywnego postrzegania jakości. Badanie umożliwiło określenie, w jaki sposób użytkownicy usługi telewizyjnej oceniają poziom jakości strumieni wideo w systemie wykorzystującym mechanizmy adaptacyjnego strumieniowania HTTP. W niniejszym rozdziale przedstawiono procedurę przeprowadzenia badań, opis implementacji oraz analizę uzyskanych wyników.

3.1 Skale ocen

Skale ocen należy dobrać w zależności od sposobu przeprowadzenia badania. Ocena bezwzględna pozwala uzyskać opinię na temat poziomu jakości danej treści lub stopnia jej degradacji, bez odniesienia do materiału referencyjnego. Wykorzystanie odpowiedniej skali umożliwia uporządkowanie ocen według ich istotności.

Dla określenia subiektywnego poziomu jakości fonii i wizji opracowano szereg skal zaproponowanych przez Międzynarodowy Związek Telekomunikacyjny (*ITU – International Telecommunication Union*) [3]. Skale ocen wykorzystywane są w wielu rekomendacjach dotyczących badania jakości transmisji, percepcji dźwięku w systemach telefonii, a także postrzegania jakości pojedynczych obrazów oraz ich sekwencji. Proponowane zakresy ocen dzielą się w zależności od ocenianego aspektu obrazu, którym może być zarówno poziom jakości, jak i poziom zniekształceń. Ocena może być dokonywana w odniesieniu do materiału referencyjnego lub bez takiego odniesienia, gdy brak jest obrazu wzorcowego.

Skala *ACR* (ang. *Absolute Category Rating*) jest pięciostopniową skalą numeryczną, uzupełnioną przymiotnikowymi odpowiednikami dla każdej wartości [3]. Oceny odnoszą się do poziomu jakości: najniższą wartością jest **1**, oznaczająca poziom *zły* (ang. *Bad*), natomiast najwyższą wartością jest **5**, określająca poziom *doskonały* (ang. *Excellent*). Skala ACR została przedstawiona w tabeli 3.1. Występuje także wariant skali, w którym wartości mieszczą się w zakresie od 0 do 4, gdzie 0 oznacza

29:2826807690

Tabela 3.1 Przymiotnikowa skala *ACR* (ang. *Absolute Category Rating*) [3]

Ocena	**Poziom postrzeganej jakości**
5	Doskonała
4	Dobra
3	Średnia
2	Słaba
1	Zła

Tabela 3.2 Przymiotnikowa skala *DCR* (ang. *Degradation Category Rating*) [4]

Ocena	**Poziom postrzeganych zniekształceń**
5	Brak zauważalnych zniekształceń
4	Zauważalne, ale nie przeszkadzające
3	Zauważalne i trochę przeszkadzające
2	Zauważalne i przeszkadzające
1	Zauważalne i bardzo przeszkadzające

najniższą ocenę. Średnia uzyskanych ocen w tej skali określana jest jako *MOS* (ang. *Mean Opinion Score*) [3].

3.2 Skala DCR

Odpowiednikiem skali ACR, służącym do oceny poziomu zniekształceń, jest skala *DCR* (ang. *Degradation Category Rating*) [3]. Podobnie jak w przypadku ACR, jest to skala niewymagająca materiału referencyjnego. Oceny przyjmują wartości całkowite od 1 do 5. Najniższa wartość oznacza zniekształcenia znacząco przeszkadzające w percepcji danego bodźca (ang. *Very annoying*), natomiast najwyższa wartość odpowiada sytuacji, w której zniekształcenia są niezauważalne (ang. *Imperceptible*). W tabeli 3.2 przedstawiono opis poszczególnych ocen. Średnia arytmetyczna ocen w tej skali określana jest jako *DMOS* (ang. *Degradation Mean Opinion Score*).

3.3 Skala CCR

Skala porównawcza *CCR* (ang. *Comparison Category Rating*) jest referencyjną metodą oceny jakości, obejmującą siedem wartości [3]. Oceniający ma do dyspozycji dwie próbki i musi określić, o ile jakość jednej z nich jest gorsza lub lepsza od drugiej. Środkowa ocena ma wartość **0** i oznacza, że obie próbki mają zbliżony poziom jakości. Najniższa wartość, **-3**, wskazuje na znacząco gorszą jakość w porównaniu z próbką referencyjną, natomiast maksymalna wartość, **+3**, oznacza znacząco lepszą jakość względem próbki referencyjnej. Opis skali CCR został przedstawiony w tabeli 3.3. Zastosowanie tej metody umożliwia nie tylko określenie, która próbka ma lepszą jakość, ale także w jakim stopniu różni się ona od drugiej. Średnia arytmetyczna uzyskanych ocen w tej skali nazywana jest *CMOS* (ang. *Comparison Mean Opinion Score*) [4].

30:1067505180

Tabela 3.3 Przymiotnikowa skala porównawcza *CCR* (ang. *Comparison Category Rating*) [4]

Ocena	**Poziom postrzeganej jakości**
-3	Znacznie gorszy
-2	Gorszy
-1	Nieznacznie gorszy
0	Taki sam
+1	Nieznacznie lepszy
+2	Lepszy
+3	Znacznie lepszy

3.4 Skala CQS

Pięciopunktowe skale *ACR* (ang. *Absolute Category Rating*) oraz *DCR* (ang. *Degradation Category Rating*) mogą zostać przekształcone w skalę ciągłą *CQS* (ang. *Continuous Quality Scale*) [3]. W tym przypadku przyjmuje się zakres ocen od **1** do **100**. Skala jest reprezentowana w postaci linii, której początek i koniec odpowiadają wartościom 1 i 100. Linia jest podzielona na pięć równych przedziałów, odpowiadających poszczególnym kategoriom ocen w skali pięciostopniowej. Zaznaczony przez oceniającego punkt można następnie przeskalować, dzieląc wartość przez 20, co umożliwia uzyskanie zmiennoprzecinkowego wyniku w zakresie od 1 do 5. Zaletą skali CQS jest dokładniejsze odwzorowanie subiektywnego odczucia jakości, zwłaszcza przy niewielkich zniekształceniach, oraz zmniejszenie ryzyka, że duże degradacje zostaną nieproporcjonalnie mocno zaakcentowane przez oceniających.

3.5 Dostępne metody

Grupa Radiokomunikacji ITU w rekomendacji BT.500-15 zaproponowała co najmniej osiem sposobów subiektywnej oceny jakości obrazu [3]. W niniejszym rozdziale opisano wybrane metody, które mogą być zastosowane w badaniach realizowanych w systemie telewizji Hiway [43].

3.5.1 Metoda pojedynczej stymulacji

W metodach pojedynczej stymulacji uczestnikowi prezentowany jest w danym momencie tylko jeden bodziec, bez możliwości porównania go z materiałem referencyjnym. W przypadku badań obrazu prezentowany jest pojedynczy kadr lub sekwencja obrazów.

Single-stimulus method (SS)

Metoda pojedynczej stymulacji *SS* (ang. *Single-stimulus*) należy do najprostszych metod badania jakości. Podczas eksperymentu wyświetlany jest pojedynczy, zazwyczaj zdegradowany obraz, który podlega ocenie. Aby uzyskać ocenę obrazu referencyjnego, należy go zaprezentować również jako pojedynczy bodziec. Typowe badanie obejmuje serię materiałów testowych, które powinny być prezentowane w kolejności losowej, a najlepiej tak, aby każdy uczestnik badania oglądał je w innej kolejności. Każdy materiał powinien być zaprezentowany jednokrotnie, a w wariancie z powtórzeniami

31:4603571160

– trzykrotnie. Procedura obejmuje planszę adaptacyjną, materiał do oceny oraz planszę umożliwiającą udzielenie oceny. Plansze powinny być neutralne (np. jednokolorowe), aby uniknąć dodatkowej stymulacji. W wariancie z powtórzeniami pierwszy pokaz obrazu traktowany jest jako faza adaptacji, a ostateczna ocena stanowi średnią z drugiego i trzeciego wyświetlenia. Zakładając identyczny czas prezentacji materiału, wariant bez powtórzeń jest o około 51% krótszy. Ocena w metodzie pojedynczej stymulacji dokonywana jest przy użyciu skali przymiotnikowej (tabela 3.1 lub 3.2) bądź skali ciągłej w zakresie od 1 do 100.

3.5.2 Wybrane metody podwójnej stymulacji

Prezentacja dwóch próbek umożliwia lepsze dostrzeżenie zniekształceń, które w metodzie pojedynczej stymulacji mogłyby pozostać niezauważone. W metodach podwójnej stymulacji uczestnik otrzymuje dwa bodźce tego samego typu, które są wyświetlane sekwencyjnie lub równocześnie. Celem badania jest ocena jakości obrazu zniekształconego w porównaniu do obrazu referencyjnego, choć możliwe jest również porównywanie par obrazów referencyjnych lub par obrazów zniekształconych.

Stimulus-comparison method (SC)

W metodzie porównawczej *SC* (ang. *Stimulus-comparison*) para obrazów wyświetlana jest naprzemiennie na jednym ekranie lub jednocześnie na dwóch skalibrowanych ekranach. Celem badania jest uzyskanie oceny relatywnej jakości wyrażonej za pomocą skali przymiotnikowej przedstawionej w tabeli 3.3. Możliwe jest również zastosowanie oceny liczbowej, poprzez zaznaczenie punktu na skali liniowej, której krańce oznaczone są jako *RÓWNY* (ang. *SAME*) oraz *RÓŻNY* (ang. *DIFFERENT*). Odległość pomiędzy zaznaczeniem a początkiem skali stanowi wartość oceny dla danej pary.

Double-stimulus impairment scale (DSIS)

Metoda *DSIS* (ang. *Double-stimulus impairment scale*) jest stosowana do badania nowych systemów oraz wpływu toru transmisyjnego na jakość obrazu. Uczestnikom prezentowane są pary obrazów: najpierw obraz referencyjny, następnie obraz zniekształcony. Oceniający proszeni są o wskazanie poziomu zniekształceń drugiego obrazu, pamiętając jego wersję wzorcową. Do zestawu należy dołączyć również obrazy referencyjne, aby weryfikować uwagę uczestników. Cała sesja nie powinna trwać dłużej niż 30 minut, ze względu na ryzyko zmęczenia oceniających. Metoda DSIS występuje w dwóch wariantach: w pierwszym obrazy prezentowane są tylko raz, natomiast w drugim – dwukrotnie. Drugi wariant pozwala lepiej oszacować niewielki poziom zniekształceń. Oceny dokonuje się za pomocą skali DCR (tabela 3.2). Pary obrazów należy prezentować w pseudolosowej kolejności, unikając sytuacji, w której dwie pary z rzędu są identyczne, nawet jeśli różnią się poziomem degradacji.

Double-stimulus continuous quality scale (DSCQS)

Metoda *DSCQS* (ang. *Double-stimulus continuous quality scale*) jest zbliżona do DSIS, jednak w tym przypadku oceniane są oba obrazy z pary. Obrazy wyświetlane są w kolejności losowej, a sesja badania nie powinna przekraczać 30 minut. Metoda występuje w wariancie indywidualnym, gdzie uczestnik może dowolnie przełączać się między obrazami A i B, oraz grupowym, gdzie obrazy

32:4842595029

wyświetlane są sekwencyjnie z ustalonym czasem prezentacji. Dla pojedynczych obrazów rekomenduje się czas wyświetlania 3–4 sekund i pięć powtórzeń, a dla sekwencji obrazów – 10 sekund i dwa powtórzenia [3]. W każdej parze jeden obraz stanowi referencję, ale nie musi być wyświetlany jako pierwszy. Kolejność należy zmieniać, aby uniknąć domyślnych założeń uczestników. Oceny dokonuje się na formularzu z dwoma pionowymi odcinkami, na których odwzorowana jest pięciopunktowa skala ACR. Oceniający zaznacza punkt w dowolnym miejscu odcinka, a wynik jest przeskalowywany do zakresu 0–100. Dopuszcza się naniesienie siatki pomocniczej, aby zmniejszyć błąd odczytu oceny. Metoda DSCQS została wykorzystana w badaniach postrzeganej jakości wideo użytkowników telewizji Hiway. Szczegóły implementacji opisano w kolejnych punktach tego rozdziału.

3.6 Powiązane badania

Subiektywne postrzeganie jakości w różnych technologiach wideo stanowi aktywny obszar badań. W pracy [44] zastosowano skalę ocen *DSIS* (ang. *Double-stimulus impairment scale*), aby sprawdzić, w jaki sposób oceny ekspertów w dziedzinie przetwarzania obrazu różnią się od ocen osób bez doświadczenia. Wyniki wskazują, że przy materiałach o wysokiej jakości różnice są niewielkie, natomiast eksperci skuteczniej wykrywają pogorszenie jakości przy niższych przepływnościach. Pozwala to wykorzystać oceny ekspertów do wyznaczenia progów widoczności zniekształceń.
W badaniu [45] wykazano, że metody takie jak *SSCQE* (ang. *Single Stimulus Continuous Quality Evaluation*), *DSCQS* (ang. *Double-stimulus continuous quality scale*) oraz *DSCS* (ang. *Double Stimulus Continuous Scoring*) mogą dawać porównywalne wyniki przy spełnieniu odpowiednich warunków. Ponadto stwierdzono, że w przypadku materiałów, w których jakość zmienia się w czasie, uczestnicy zapamiętują średnio około 15 sekund treści.
Autorzy [46] przeprowadzili badanie DSCQS dla materiałów wideo w jakości UHD, analizując wpływ kodowania w standardach H.264, H.265 i VP9 z różnymi poziomami kwantyzacji i liczbą klatek na sekundę. Stwierdzono, że kodowanie w H.265 zapewnia wyższe oceny jakości dla każdej kombinacji parametrów.
W badaniu [47] porównano osiem metod oceny jakości wideo, takich jak DSCQS [3], DSIS [3], ACR5 [3], ACR5-HR [48], ACR11 [48], ACR11-HR [48], SAMVIQ [3] oraz SAMVIQ-HR [3]. Badanie przeprowadzono na urządzeniach mobilnych. Najlepsze wyniki statystyczne uzyskano dla metod DSIS, ACR5 i SAMVIQ, a z uwagi na prostotę i krótki czas trwania badania, za optymalną uznano metodę ACR.
W artykule [49] zbadano jakość wideo dla treści HD, UHD i UHD z HDR. Wykorzystano metodę DCR wraz ze skalą EVP [3], stosowaną w badaniach eksperckich, oraz metodę ACR. Ostatecznie stwierdzono, że wyniki obu metod są porównywalne, a ich zastosowanie może znacząco ograniczyć koszty organizacji badań.
W badaniu [50] przedstawiono ocenę jakości QoE dla treści immersyjnych (360°) odtwarzanych na wyświetlaczach HMD (*head-mounted display*), wykorzystując metody DSIS i pojedynczej stymulacji.
W publikacji [51] zaprezentowano algorytm *HA-SVQA* (ang. *Human-Aware Subjective Video Quality Assessment*), który optymalizuje dobór materiałów testowych w celu redukcji redundancji i skrócenia czasu badań nawet o jedną trzecią.

33:8481073025

W pracy [52] zaproponowano wykorzystanie estymacji największej wiarygodności w celu eliminacji niedokładności i błędów w subiektywnych ocenach uczestników.
Autorzy [53] przeprowadzili badanie pojedynczej stymulacji z użyciem skali ACR dla treści UHD kodowanych w H.264, H.265 i VP9, porównując subiektywne i obiektywne metody oceny jakości.
W publikacji [54] przedstawiono badanie metodą SS na zbiorze materiałów UHD, który został udostępniony publicznie, oraz porównano wybrane algorytmy VQA (*Video Quality Assessment*), uzyskując współczynnik korelacji rang Spearmana (SROCC) na poziomie około 0,8.
W badaniu [55] analizowano jakość materiałów UGC (*user-generated content*) tworzonych na potrzeby mediów społecznościowych, z wykorzystaniem metody pojedynczej stymulacji i skali ACR.
Autorzy [56] zaproponowali narzędzie do przeprowadzania badań subiektywnej oceny jakości poza środowiskiem laboratoryjnym. Narzędzie wspiera zalecenia rekomendacji ITU-T P.910 [48] i umożliwia oceny w skalach ACR, DCR i CCR.
W publikacji [57] przedstawiono bazę MIND-VQ zawierającą materiały filmowe nagrane nocą oraz wyniki badania metodą SS. Stwierdzono, że największy wpływ na ocenę jakości mają cechy takie jak widoczność przestrzenna i czasowa treści, natomiast jasność, kolory czy stabilność obrazu mają mniejsze znaczenie.
W pracy [58] zaproponowano półautomatyczną metodę etykietowania treści wideo w celu odtworzenia ocen MOS (*Mean Opinion Score*) na podstawie poziomu kwantyzacji transkodowanych filmów. Zastosowanie tej metody umożliwiło stworzenie syntetycznych ocen dla 1170 filmów na podstawie 390 prawdziwych ocen subiektywnych.
W badaniu [18] oceniano jakość wideo w systemach *MCG* (ang. *Mobile Cloud Gaming*), gdzie strumień wideo przesyłany jest z chmury do urządzenia użytkownika.
W pracy [59] zastosowano metodę DSIS do oceny jakości transmisji SoftCast [60] i porównano wyniki z algorytmami obiektywnej oceny jakości. Najlepsze wyniki uzyskał algorytm SSIM [61] z korelacją rang Pearsona powyżej 0,96, natomiast VMAF [62] i MS-SSIM [63] uzyskały zbliżone, nieco niższe wartości.
W pracy [64] przeprowadzono badanie DSIS w celu oceny wpływu parametrów sieci, takich jak BER, PLR, jitter i przepływność kanału na jakość transmisji wideo.
Autorzy [65] przebadali materiały Full HD i UHD kodowane w H.264 i H.265, stwierdzając, że dla treści Full HD różnice w ocenach subiektywnych są na tyle niewielkie, że stosowanie H.265 może być nieuzasadnione.
W badaniu [66] sprawdzono wpływ liczby klatek na sekundę na ocenę jakości przy użyciu metody SAMVIQ. Dla materiałów nagranych w 120 kl./s ocena była o około 30 punktów (skala 1–100) wyższa względem materiałów nagranych w 24 kl./s, natomiast różnica między 60 i 120 kl./s wynosiła około 5 punktów.
W pracy [67] przeprowadzono badanie DSIS w celu wyznaczenia minimalnej przepływności dla materiałów UHD kodowanych w H.265 dla telewizji cyfrowej. Oszacowano, że przepływność 18 Mbit/s zapewnia wartości MOS powyżej 4,5, co uznano za wystarczające.
W badaniu [68] zastosowano metodę DSCQS w celu analizy wpływu warunków otoczenia i ustawień ekranu telewizora na postrzeganą jakość obrazu. Największy wpływ miały tryby wyświetlania, poziom oświetlenia oraz sama zawartość materiału.

34:3545678937

W pracy [69] zaproponowano system oceny QoE w czasie rzeczywistym, w którym ocena jakości udzielana jest dla całego filmu, a analizie poddano wpływ utraty pakietów i wstrzymania klatek.

3.7 Założenia

Badanie postrzeganej jakości wideo w produkcyjnym systemie telewizyjnym Hiway firmy INEA zostało zrealizowane metodą *DSCQS* (ang. *Double-stimulus continuous quality scale*), opisaną w rozdziale 3.5.2. Wybrano ją, ponieważ umożliwia prowadzenie zarówno badań indywidualnych, jak i grupowych, co odzwierciedla typowy sposób korzystania z usługi telewizyjnej. Celem badania było sprawdzenie, w jaki sposób użytkownicy telewizji Hiway oceniają poziom jakości treści wideo dostarczanych jako strumieniowanie adaptacyjne, z wykorzystaniem dekoderów *OTT* (ang. *Over-the-top*). Szczegółową architekturę systemu oraz sposób dystrybucji treści wideo przedstawiono w poprzednim rozdziale.
W badaniach przyjęto następujące założenia:

- Badanie ma weryfikować jakość treści wideo dostarczanych przez dekodery; zatem sekwencje użyte w teście muszą być odtwarzane przez użytkowników końcowych właśnie za pośrednictwem dekodera.
- Standardowy sposób konsumpcji treści telewizyjnych obejmuje oglądanie zarówno w grupie, jak i indywidualnie; metoda badania musi wspierać oba tryby.
- Implementację należy przeprowadzić tak, aby każdy użytkownik mógł wykonać test w domu.
- Udzielanie ocen powinno być możliwie proste; należy także zapewnić efektywny sposób zebrania ocen od użytkowników.
- Ocena jakości ma objąć każdy dostępny poziom przepływności strumienia wideo dla kodeków H.264 oraz H.265.
- Przyjęto liczbę użytkowników systemu w przedziale [100,000, 200,000].
- Instrukcje dotyczące metodyki badania należy przekazać w sposób prosty i czytelny.
- Komunikacja z klientem ma odbywać się wyłącznie poprzez dekoder, zgodnie z polityką firmy.
- Pojedyncza sesja badania nie może trwać dłużej niż 30 minut, aby ograniczyć zmęczenie i spadek koncentracji uczestników, zgodnie z zaleceniami ITU.

Wdrożeniowy charakter pracy wymagał implementacji badania na dekoderze. Takie podejście pozwalało testować strumienie pochodzące bezpośrednio z sieci *CDN* (ang. *Content Delivery Network*) Hiway [43], bez dodatkowego narzutu związanego z transkodowaniem treści przez aplikacje trzecie (np. YouTube). Metoda DSCQS umożliwia prowadzenie badań z pojedynczym uczestnikiem oraz z grupą osób. Dodatkowo procedura pozwala przygotować sekwencje badawcze na dekoderach w sposób bezobsługowy z punktu widzenia uczestnika. Ocena treści oraz gromadzenie wyników zostały zrealizowane za pomocą dedykowanego formularza w postaci aplikacji mobilnej. Na początku badania

35:7106217942

wyświetlany jest kod *QR* (ang. *Quick Response*) kierujący do formularza on-line, co umożliwia dokonywanie ocen z wykorzystaniem urządzenia mobilnego w warunkach domowych. Telewizja Hiway wykorzystuje strumienie wideo kodowane algorytmami H.264 i H.265, w zależności od kanału. Oceniane materiały zostały transkodowane tak, aby pokryć każdy poziom przepływności dla obu kodeków, a następnie połączone w sekwencje, by użytkownik nie musiał przełączać się pilotem między fragmentami. Przygotowano dwie osobne sekwencje dla H.264 i H.265, uzupełnione o instrukcje dotyczące przebiegu badania i kolejnych kroków. Każdą sekwencję badawczą wraz z niezbędnymi instrukcjami ograniczono do 23 minut. Szczegóły implementacji przedstawiono w kolejnym podrozdziale.

3.8 Realizacja badania

W niniejszym rozdziale opisano wdrożeniową realizację badania w produkcyjnym systemie telewizji Hiway, należącym do operatora INEA.

3.8.1 Materiały wideo wykorzystane w sekwencjach

Szczegółowy spis materiałów wideo użytych w obu sekwencjach zawarto w dodatkach A oraz B, odpowiednio dla treści kodowanych kodekami H.264 oraz H.265. Pozycje *Test 1* oraz *Test 2* wykorzystano jako materiały adaptacyjne i są identyczne dla obu sekwencji. W trakcie ich odtwarzania, poprzedzonego planszami objaśniającymi, uczestnicy zapoznają się z przebiegiem badania. Materiały od 1 do 15 tworzą pary, dla których zbierane są właściwe oceny. Zestaw treści dobrano tak, aby zapewnić zróżnicowaną zawartość. Każdy materiał posiadał przepływność wideo większą niż 10 Mbit/s. Łącznie wybrano 32 filmy, które poddano transkodowaniu do poziomów przepływności wykorzystywanych w telewizji Hiway.

3.8.2 Transkodowanie materiałów wideo

Dostosowanie filmów do profili wideo wykonano poprzez transkodowanie z użyciem oprogramowania Rhozet Workflow Manager firmy Harmonic [70]. Jest to transkoder o architekturze rozproszonej, składający się z kontrolerów, węzłów transkodujących oraz bazy SQL. Licencjonowanie umożliwia m.in. kodowanie wideo H.264 na poziomach 1–5.1 oraz H.265 na poziomach 1–6.2. Sygnał audio może być kodowany algorytmami AAC-LC, HE-AAC, Dolby Digital lub Dolby Digital Plus. Następnie treść może być spakowana do strumienia transportowego TS (ang. *Transport Stream*) lub do kontenera MP4 (ang. *MPEG-4 Part 14*). Zaletą oprogramowania jest możliwość transkodowania pojedynczego pliku wejściowego do wielu profili wyjściowych o różnych przepływnościach, które mogą być następnie multipleksowane do pojedynczego strumienia MP4. Na wyjściu otrzymuje się pojedynczy plik, a odtwarzana jakość wideo jest wybierana z poziomu odtwarzacza.

Materiały wideo transkodowano z parametrami zestawionymi w tabeli 3.4 dla treści kodowanych H.264 oraz H.265. Ustawienia te odpowiadają konfiguracji kanałów telewizyjnych w usłudze komercyjnej. Dla treści H.264 uzyskano przepływności: 800, 1,500, 3,000, 6,000, 8,000 oraz 10,000 Kbit/s. Wymiary pojedynczej klatki to 800×640 lub 854×480 dla 800 Kbit/s, 1280×720 dla 1,500 Kbit/s

36:6942256946

Tabela 3.4 Parametry transkodowania dla sekwencji badawczych

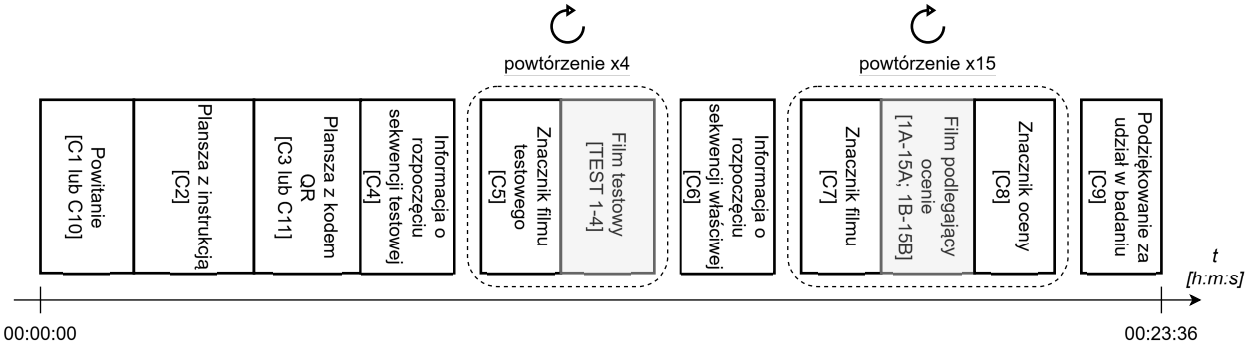
Parametr w Rhozet Workflow Manager	**Sekwencja 1**	**Sekwencja 2**
Koder wideo (ang. *Video Codec*)	H.264	H.265
Profil wideo (ang. *Video Profile*)	H.264 High	H.265 Main
Poziom wideo (ang. *Video Level*)	H.264 Level 4.2	H.265 Level 5.1
Usuwanie przeplotu (ang. *De-interlace Input*)	Dynamic Deinterlace	Dynamic Deinterlace
Opóźnienie VBV (ang. *VBV Delay*)	Default	Default
Interwał IDR (ang. *IDR Period*) [s]	2	2
Opcje GOP (ang. *GOP Option*)	Auto	Auto
Płynność (ang. *Smoothness*)	Balanced	Balanced
Przepływność wideo (ang. *Video Bitrate*) [Kbit/s]	800, 1,500, 3,000, 6,000, 8,000, 10,000	1,255, 2,276, 3,298, 5,342, 6,364
Szerokość i wysokość (ang. *Width and Height*)	800×640, 854×480, 1280×720, 1920×1080	1280×720, 1920×1080
Koder audio (ang. *Audio Codec*)	AAC-LC	AAC-LC
Częstotliwość próbkowania (ang. *Sample Rate*) [kHz]	48	48
Przepływność audio (ang. *Audio Bitrate*) [Kbit/s]	128	128

oraz 1920×1080 dla 3,000–10,000 Kbit/s. Najniższe przepływności rzędu 800 i 1,500 Kbit/s odpowiadają jakości *SD* (ang. *Standard Definition*), natomiast od 3,000 Kbit/s w górę jakości *HD* (ang. *High Definition*). Treści wideo kodowane H.265 posiadają przepływności rzędu 1,255 Kbit/s dla 1280×720 oraz 2,276, 3,298, 5,342 i 6,364 Kbit/s dla 1920×1080. Głębię bitową ustawiono na 8 bitów. Dla każdego materiału H.264 przygotowano 7 profili wideo (łącznie 119 filmów) do wykorzystania w sekwencji badawczej nr 1. Dla H.265 każdy materiał posiadał 5 profili (85 pozycji) dla sekwencji nr 2. Łącznie przygotowano 204 materiały, z których wybrano po 34 pozycje do każdego badania.

3.8.3 Organizacja badania

Utworzenie sekwencji badawczych zgodnych z protokołem DSCQS w warunkach ocen zbiorowych wymagało konkatenacji materiałów testowych oraz podlegających ocenie w jedną długą sekwencję. Udostępnienie badania użytkownikom telewizji wymagało przygotowania graficznego instruktażu osadzonego w sekwencji. W tym celu przygotowano zestaw grafik informacyjnych, które przekształcono w materiały wideo o tak dobranym czasie trwania, aby umożliwić spokojne zapoznanie się z treścią plansz. Sekwencje badawcze miały identyczny zestaw plansz, z wyjątkiem dwóch grafik zawierających kody QR: C.1 oraz C.3 w pierwszej sekwencji zastąpiono odpowiednio C.10 i C.11 w drugiej. Zestaw wykorzystanych grafik oraz ich czas trwania podano w dodatku C.

37:2636404242



Rysunek 3.1 Organizacja logiczna sekwencji: plansze informacyjne oraz materiały testowe

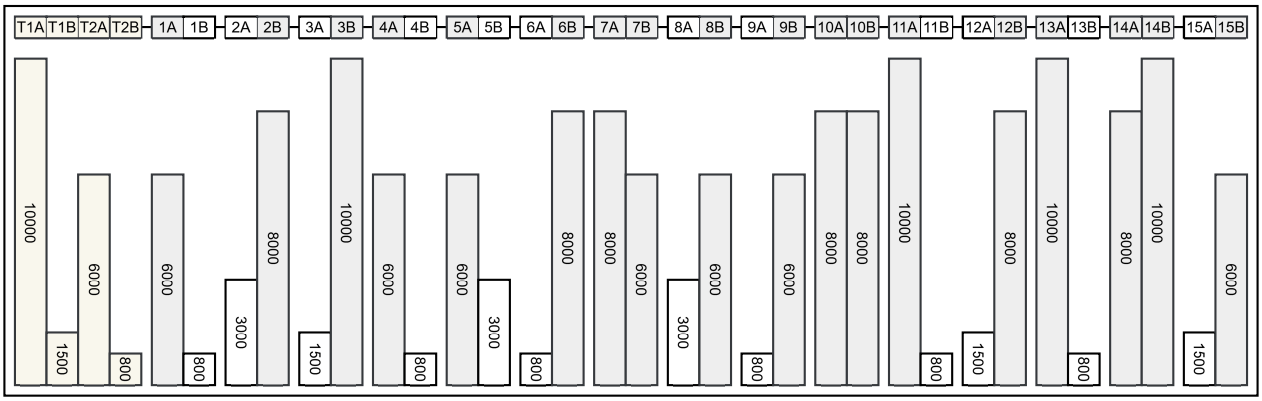
Rysunek 3.1 ilustruje układ materiałów w obu sekwencjach badawczych. Każda sekwencja trwa 23 minuty i 36 sekund, a jej organizacja jest następująca:

- Plansza powitalna z przewidywanym czasem trwania i kodem QR kierującym do formularza ocen w wersji mobilnej. Sekwencje H.264 i H.265 mają osobne formularze, stąd dwie różne grafiki (C.1, C.10). Czas: 17 s.
- Grafika C.2 z opisem i instrukcją przebiegu badania: prośba o ocenę po obejrzeniu pary materiałów, prezentacja skali ocen i jej zakresu, informacja o dodatkowej ankiecie (4 pytania). Czas: 40 s.
- Trzecia plansza to prosty ekran z kodem QR (C.3 dla H.264 oraz C.11 dla H.265), ułatwiający otwarcie formularza przed rozpoczęciem części zasadniczej. Czas: 10 s.
- Plansza C.4 zapowiada sekwencję adaptującą (dwie pary filmów), które nie podlegają ocenie. Czas: 10 s.
- Sekwencja adaptująca: czterokrotnie wyświetlany blok składający się z grafiki sygnalizacyjnej C.5 (5 s) oraz materiału (30 s).
- Plansza C.6 informująca o rozpoczęciu właściwej części badania i przypominająca o dokonaniu oceny. Czas: 10 s.
- Część właściwa: piętnastokrotnie powtarzany blok: plansza dla filmu testowego A (3 s), materiał A (30 s), plansza dla filmu testowego B (3 s), materiał B (30 s), oraz grafika zachęcająca do oceny (8 s). Pojedynczy blok trwa 74 s.
- Plansza końcowa C.9 z podziękowaniem za udział.

Ze względu na wdrożeniowy charakter badania, przepływności bitowe materiałów dobrano tak, aby odpowiadały profilom strumieniowania adaptacyjnego stosowanym w sieci operatora INEA. Różnorodność przepływności wynika z faktu świadczenia kanałów w różnych jakościach. Na rysunku 3.2 przedstawiono przepływności dla treści kodowanych H.264. Każda kolumna reprezentuje materiał filmowy; w kolumnie umieszczono wartość przepływności, a nad nią identyfikator materiału. Materiały TXA i TXB należą do sekwencji adaptującej, natomiast XA i XB to pary przeznaczone do oceny

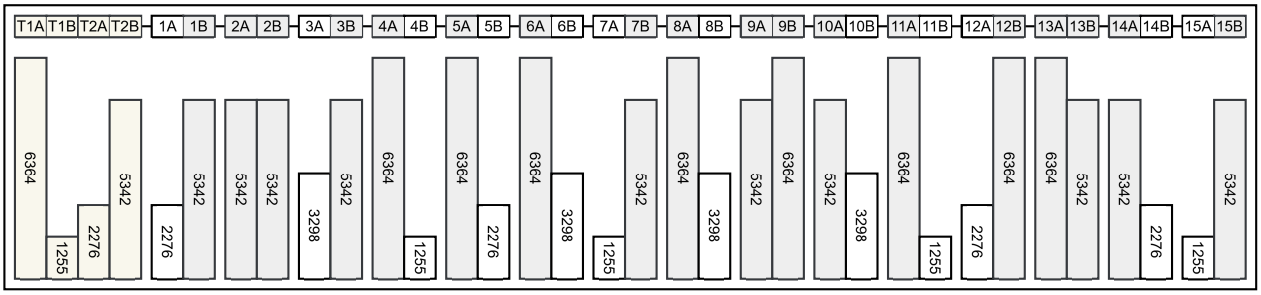
38:7087545348

(X oznacza numer pary). Co najmniej jeden z materiałów w parze jest treścią o podwyższonej przepływności (referencją). Dla H.264 przyjęto 6,000 Kbit/s jako minimum dla treści referencyjnej (jest to jednocześnie najwyższa przepływność dla standardowych kanałów 1920×1080), natomiast 8,000 i 10,000 Kbit/s odpowiadają kanałom o podwyższonej jakości (premium). Pozostałe materiały mają 800, 1,500 i 3,000 Kbit/s. Każdy materiał podlega ocenie, łącznie z referencjami. Pary 7, 10 i 14 zawierają referencje zarówno dla obrazów A, jak i B, co służy utrzymaniu uwagi uczestników oraz uniknięciu przekonania, że zawsze jeden z materiałów ma gorszą jakość. Pozycja referencji (A/B) jest zmieniana między parami, aby uniknąć efektu uczenia się.



Rysunek 3.2 Organizacja materiałów testowych w pary dla sekwencji 1. Każda kolumna zawiera wartość przepływności bitowej. Kolorem szarym zaznaczono materiały referencyjne.

Na rysunku 3.3 przedstawiono przepływności dla kodowania H.265, dobrane na bazie tych samych kryteriów. Obrazami referencyjnymi są treści o przepływnościach 5,342 i 6,364 Kbit/s. Pozostałe materiały mają 1,255, 2,276 oraz 3,298 Kbit/s. Podobnie jak wcześniej, pozycje referencji w parach są zmienne. Pary 2, 9 i 13 zawierają wyłącznie referencje.



Rysunek 3.3 Organizacja materiałów testowych w pary dla sekwencji 2. Każda kolumna zawiera wartość przepływności bitowej. Kolorem szarym zaznaczono materiały referencyjne.

3.8.4 Implementacja na dekoderze

Badanie udostępniono na dekoderach usługi TV Hiway (INEA). Implementacja w formule wideo na życzenie *VOD* (ang. *Video on Demand*) wymagała spełnienia założeń technicznych i marketingowych; wykorzystane grafiki i metadane zatwierdził dział marketingu. Założono użycie biblioteki VOD

39:9050763581

do prezentacji sekwencji badawczych poprzez dekoder, co wymagało przygotowania sekwencji jako pojedynczych materiałów wideo.

Pierwszym krokiem było przekształcenie statycznych plansz informacyjnych (dodatek C) w materiały wideo o odpowiedniej długości. Wykorzystano aplikację Movie Maker – Video Editor [71], która po dodaniu ścieżki dźwiękowej wygenerowała plik o wysokiej przepływności bitowej. Następnie pliki te transkodowano z parametrami z tabeli 3.4, analogicznie do materiałów ocenianych. Przepływność wyjściową ustawiono na 6,000 Kbit/s (H.264) oraz 6,364 Kbit/s (H.265), co jest wystarczające dla statycznej grafiki z tekstem. Transkodowanie grafik do tego samego formatu co materiały oceniane umożliwiło konkatenację całości w jedną sekwencję wideo dla każdego kodeka. Materiały uporządkowano zgodnie z opisem organizacji badania, a następnie połączono oprogramowaniem FFmpeg [72] w dwie osobne sekwencje (H.264 oraz H.265).

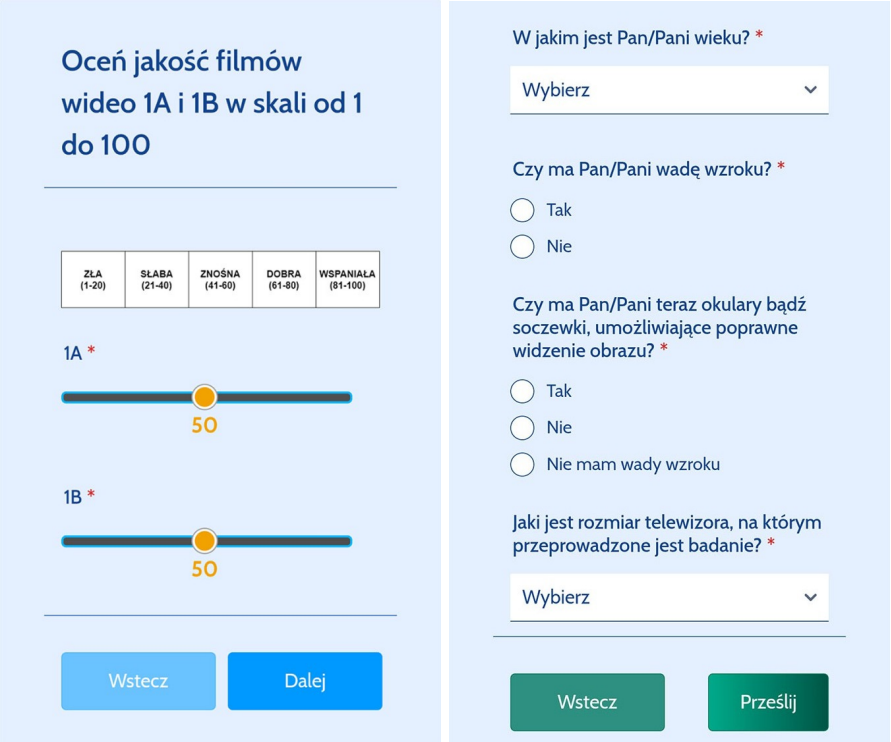
Dla obu sekwencji przygotowano plakaty graficzne (C.12a, C.12b) na wzór pozycji VOD dostępnych z poziomu dekodera. Grafiki zawierają kod QR prowadzący do mobilnego formularza ocen. Ostatnim krokiem było przygotowanie metadanych opisujących zasób VOD. Utworzono plik XML zgodny ze standardem CableLabs Content 3.0 [73]. Tak przygotowany zestaw plików pozwolił utworzyć dwa materiały dostępne w bibliotece VOD usługi Hiway. Treści umieszczono w katalogu *Badanie Jakości* (Rys. C.13). Na karcie filmu VOD (Rys. C.14) znajduje się opis zachęcający do udziału. Po wybraniu przycisku *Oglądaj* użytkownik uruchamia sekwencję, a dalszy przebieg badania jest prowadzony przez plansze informacyjne.

3.8.5 Agregacja ocen

Gromadzenie ocen odbywa się głównie poprzez formularz on-line; alternatywnie udostępniono kartę ocen w formie papierowej. Formularz on-line utworzono w narzędziu jotform.com [74], umożliwiającym tworzenie zaawansowanych formularzy dostosowanych zarówno do urządzeń mobilnych, jak i ekranów większych. Dostęp do formularza uzyskuje się przez zeskanowanie kodu QR wyświetlanego w trakcie badania. Każda sekwencja ma osobny kwestionariusz, a zebrane oceny agregowane są w odrębnych arkuszach. Pierwsza strona formularza zawiera instrukcję przebiegu badania. Kolejne 15 stron zawiera po dwa suwaki (ang. *sliders*) w skali 1–100. Dwie ostatnie strony to krótka ankieta (4 pytania) oraz podziękowanie. W momencie projektowania jedynie jotform.com zapewniał suwak z dedykowaną skalą, co pozwoliło odwzorować skalę ciągłą w protokole DSCQS (opis w rozdziale 3.5.2). Nad suwakami (Rys. 3.4a) umieszczono graficzną reprezentację skali ACR z przedziałami i przymiotnikowymi etykietami; wartość domyślna każdego suwaka to 50. Przedostatnia strona (Rys. 3.4b) zawiera pytania pomocnicze: zakres wieku, występowanie i korekcję wady wzroku, oraz rozmiar ekranu urządzenia. Badanie zaprojektowano tak, aby mogło być realizowane również grupowo.

Dla osób bez urządzeń mobilnych przygotowano alternatywny formularz papierowy (Rys. 3.5). Arkusz składa się z dwóch stron. Pierwsza zawiera 15 par pionowych odcinków podzielonych na pięć równych części; każda para służy do oceny subiektywnej jakości dla odpowiadającej pary materiałów. Po lewej stronie umieszczono legendę z przymiotnikowymi ocenami ACR, dopasowaną rozmiarem do odcinków, aby ułatwić mapowanie oceny. Każdy odcinek podzielono dodatkowo na cztery równe fragmenty, by zminimalizować błąd odczytu. Druga strona zawiera te same cztery pytania, co formularz mobilny. Ocena elektroniczna jest wartością dokładną, natomiast wartości odczytane

40:4886876990



(a) Sposób oceniania w formularzu mobilnym

(b) Dodatkowe pytania na zakończenie badania

Rysunek 3.4 Przykładowe widoki z aplikacji mobilnej do oceny jakości doświadczeń

z arkusza papierowego mogą być obarczone błędem; dlatego formularz papierowy stosowano jedynie pomocniczo.

3.8.6 Dobór próby

W celu oszacowania minimalnej liczby osób niezbędnej do uzyskania maksymalnego błędu na poziomie 10% wykorzystano wzór 3.1 na liczebność próby [75]:

$$n_{\min} = \frac{z^2\, p\,(1-p)}{e^2} \tag{3.1}$$

gdzie:

$n_{\min}$ liczebność próby,

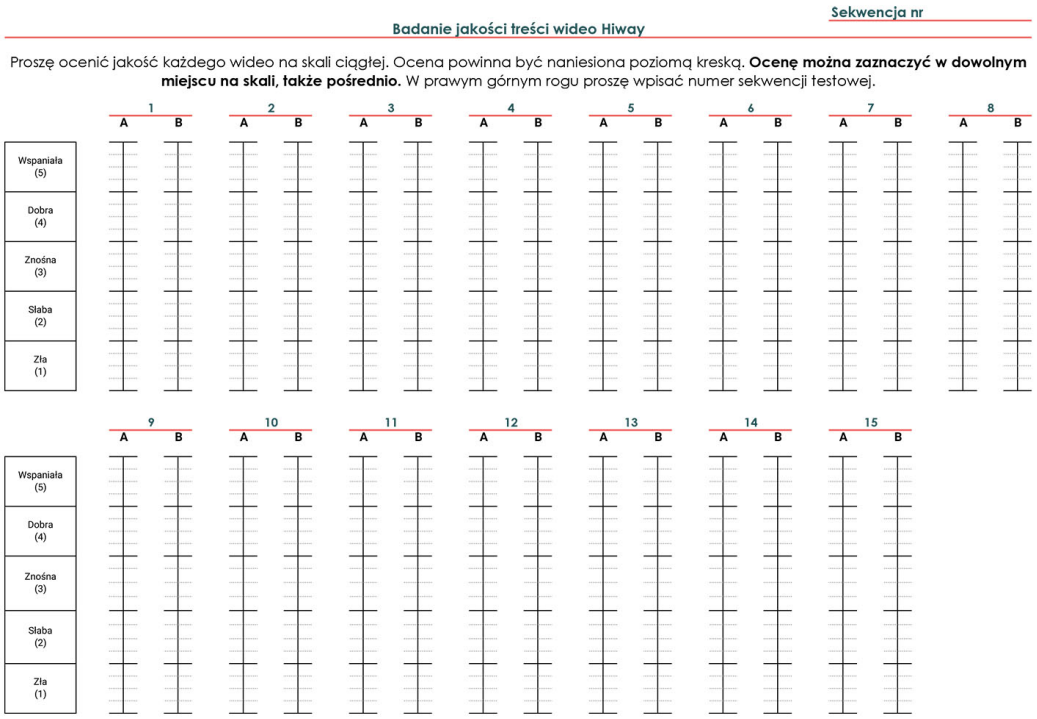
z wartość krytyczna z rozkładu normalnego dla zadanego poziomu ufności,

p szacowana wielkość frakcji,

e maksymalny dopuszczalny błąd.

Przyjęto poziom ufności 95%, dla którego wartość krytyczna wynosi $z = 1{,}96$ [76]. Wielkość frakcji przyjęto jako $p = 0{,}5$ (brak wiedzy o preferencjach badanych; wartość ta maksymalizuje

41:7777046905



Rysunek 3.5 Formularz ocen w formie papierowej

wariancję i tym samym wymaganą liczebność). Maksymalny dopuszczalny błąd ustalono na 10%. Otrzymano:

$$n_{\min} = \frac{1{,}96^2 \cdot 0{,}5 \cdot (1 - 0{,}5)}{0{,}1^2} \tag{3.2}$$

co daje $n_{\min} \approx 96{,}04$, czyli 96 osób.

Sprawdzono również wpływ skończonej wielkości populacji ($N = 100{,}000$ oraz $N = 200{,}000$) za pomocą korekty 3.3 [76]:

$$n_{\min}^{(\text{finite})} = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \tag{3.3}$$

gdzie:

$n_{\min}^{(\text{finite})}$ liczebność próby po korekcie,

n_0 liczebność próby bez korekty,

N liczebność populacji.

Po korekcie otrzymano również 96 osób (w zaokrągleniu). Zatem należało przebadać co najmniej tylu uczestników, aby uzyskać błąd nie większy niż 10%.

3.9 Analiza wyników

Badanie postrzegania subiektywnej jakości wideo dla telewizji Hiway było wykonywane zarówno w wariancie indywidualnym jak i grupowym. Sumarycznie uzyskano wyniki od 108 osób. 64 osoby

42:9700543195

wykonały badanie dla sekwencji kodowanej H.264, a 44 osoby dla H.265. 6 uczestników badania skorzystało z papierowej karty ocen, reszta została zebrana za pomocą formularza mobilnego. Wstępna analiza ocen wykazała, że dwa wyniki z sekwencji H264 oraz jeden wynik dla H265 znacząco różnią się od pozostałych. Pierwszy wynik dla H264 zawierał wartości ocen wynoszące 50 lub 100 punktów. Drugi wynik to oceny, o wartościach bliskich pojedynczych punktów bez względu na jakość wideo danego materiału. Dla H265 uzyskano wartości ocen równe 50 punktów dla każdego materiału. Wynik to oceny zebrane dla pojedynczej ankiety zawierającej 30 ocen. Otrzymywanie znacząco różniących się ocen jest jednym z ryzyk badania, które może być wykonane w pełni zdalne. Trzy wyniki zostały uznane za odstające z dwóch powodów. Po pierwsze, ocena wynosząca 50 punktów to domyślna ocena w formularzu mobilnym, natomiast 100 to ocena maksymalna. Wątpliwe jest, aby użytkownik wykonał badanie z należytą starannością, jeżeli 30 ocen ma takie same wartości lub przyjmuje tylko wartości domyślne lub maksymalne. Po drugie, brak staranności potwierdza czas trwania wypełnienienia formularza ocen. Dla tych przypadków czas ten był nieco powyżej jednej minuty, czyli statystycznie znacząco odbiegał od średniego czasu potrzebnego na wykonanie badania. Czas wypełnienia ankiety powinien być zbliżony do czasu trwania sekwencji badawczej, czyli ponad 20 minut. Można więc założyć, że takie wyniki mogą być następstwem ciekawości użytkowników telewizji. Na ekranie dekodera pojawia się nowa zawartość VOD, do której dostęp jest bez ograniczeń. Użytkownik dekodera może chcieć sprawdzić co nowego udostępnił operator w swojej usłudze, zarówno pod kątem treści wideo, jak i treści dodatkowej kryjącej się za kodem QR.

W tabelach D.1 oraz D.2 dostępnych w dodatku D przedstawiono szczegółowe podsumowanie zebranych ocen dla sekwencji materiałów kodowanych H264 oraz H265. Dla każdego materiału określono wartość minimalną (MIN), pierwszy kwartyl (Q1), medianę (Q2), trzeci kwartyl (Q3), 95 percentyl (95%), wartość średnią (AVG) oraz odchylenie standardowe (σ). Ponadto, dla każdej pary wykonano analizę porównawczą, dla której sprawdzono w jaki sposób odczuwane jest polepszenie lub pogorszenie jakości obrazu. Zebrane w całość wyniki zostały przedstawione w tabelach D.3 oraz D.4 w dodatku D. Ujemna wartość oznacza pogorszenie odczuwanej jakości wideo drugiego wyświetlenia i została obliczona jako różnica ocen wyświetlenia B i A. Dodatnia to polepszenie jakości obrazu względem pierwszego materiału.

3.9.1 Wyniki dla par kodowanych zgodnie ze standardem H.264

Dla każdej pary kodowanej kodekiem H.264 wyniki zostały wyodrębione i opisane poniżej. W kolumnie Kbit/s podano przepływność danego strumienia. Dla wiersza zawierającego różnicę ocen zawarto wartość wzrostu lub spadku przepływności pomiędzy dwoma wyświetleniami. Dla każdego materiału z pary wygenerowano histogram ocen wraz z estymatorem gęstości, aby dokładniej porównać rozkłady ocen.

Para nr 1

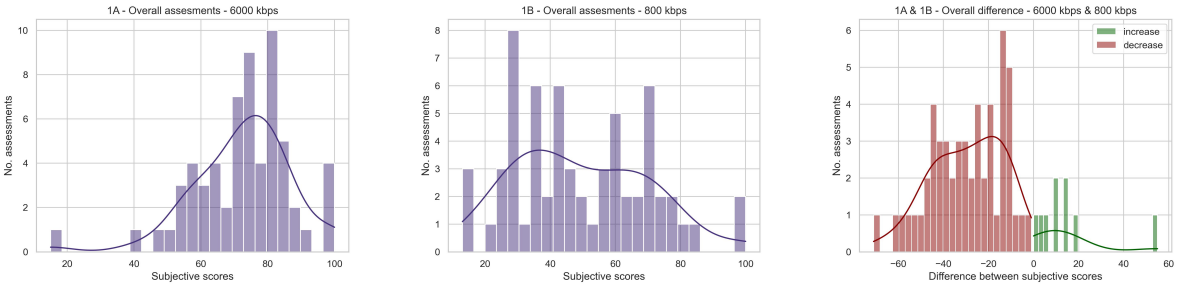
Treści o przepływnościach odpowiednio 6000 Kbit/s (A) oraz 800 Kbit/s (B). Wartości minimalne są zbliżone. Mediana ocen jest wyższa o 28,5 punktu, średnia arytmetyczna natomiast o 23,2 punktu na korzyść obrazu o wyższej przepływności. 75% populacji oceniło, że obraz o przepływności 800 Kbit/s zasługuje na ocenę 65,5 punktu lub niższą. Według przymiotnikowej

43:8546070122

skali ACR obrazy te zasługją średnio na ocenę odpowiednio 3,625 (jakość średnia, dążąca do dobrej) oraz 2,465 (jakość słaba, dążąca do złej). Na wykresie 3.6 (a) i (b) widać, że dla pierwszego materiału rozkład ocen jest skupiony w sąsiedztwie 70-80 punktów, natomiast dla obrazu drugiego oceny są nieco bardziej rozproszone na skali. Na wykresie 3.6 (c) widać, że większość oceniających prawidłowo oznaczyła strumień o pogorszonej jakości. 75% uczestników badania określiło, że spadek jakości wynosił co najmniej 12 punktów. Połowa oceniających uznała, że jakość spadła o minimum 24 punkty, podczas gdy jedna czwarta oceniających zaznaczyła ocenę niższą o 40 punktów. Pojedyncze osoby zaznaczyły, że to drugi obraz ma lepszą jakość. Niemniej jednak określono, że średni spadek jakości wyniósł 23,2 punktu.

Tabela 3.5 Oceny uzyskane dla pary nr 1 kodowanej zgodnie ze standardem H.264

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
1A	6000	15,0	64,3	74,0	81,0	99,7	72,5	14,9
1B	800	13,0	32,0	45,5	65,5	80,9	49,3	20,7
1A $\neq$ 1B	spadek o 5200	-71	$-40,0$	$-24,0$	$-12,0$	13,8	$-23,2$	22,4



(a) 1A (b) 1B (c) 1A $\neq$ 1B

Rysunek 3.6 Rozkład ocen dla pary 1 sekwencji kodowanej zgodnie ze standardem H.264

Para nr 2

Treści o przepływnościach odpowiednio 3000 Kbit/s (A) oraz 8000 Kbit/s (B). Dla tej pary zarówno średnia arytmetyczna jak i mediana jest wyższa dla materiału A, odpowiednio o 5,2 punktu oraz 5 punktów. Na podstawie wartości odchylenia standardowego można stwierdzić, że oceny dla materiału B są bardziej rozproszone. Potwierdza to wartość minimalna wynosząca 1 punkt. Na wykresach 3.7 (a) i (b) można dostrzec, że rozkłady są do siebie zbliżone. Niemniej jednak, kilka ocen znajduje się w okolicy najniższej punktacji. Przeciętnie oceniający uznali, że drugi obraz jest słabszy o 5,2 punktu. 25% uczestników badania przyznało materiałowi B ocenę lepszą o co najmniej 6 punktów. Na wykresie 3.7 (c) widać jednak, że to obraz pierwszy, o niższej przepływności, częściej uzyskiwał oceny wyższe od materiału B.

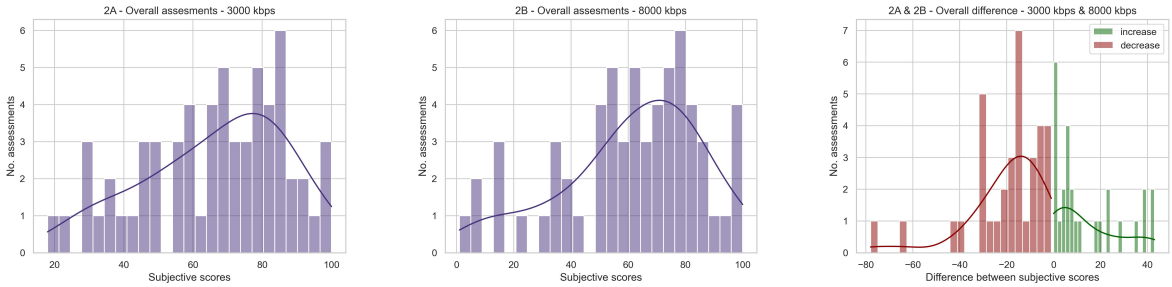
Para nr 3

Treści o przepływnościach odpowiednio 1500 Kbit/s (A) oraz 10000 Kbit/s (B). Mediana jest wyższa o 10 punktów, a średnia arytmetyczna o 4,9 punktu dla materiału B. W 5-cio stopniowej skali ACR mediana osiągnęła ocenę dobrą, czyli wartość 4. Histogramy 3.8 (a) i (b)

44:1153180512

Tabela 3.6 Oceny uzyskane dla pary nr 2 kodowanej zgodnie ze standardem H.264

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
2A	3000	18,0	51,3	70,0	81,0	95,8	66,3	20,7
2B	8000	1,0	50,0	65,0	78,8	97,9	61,1	24,5
2A $\neq$ 2B	Wzrost o 5000	−78,0	−16,8	−4,5	6,0	38,8	−5,2	23,1



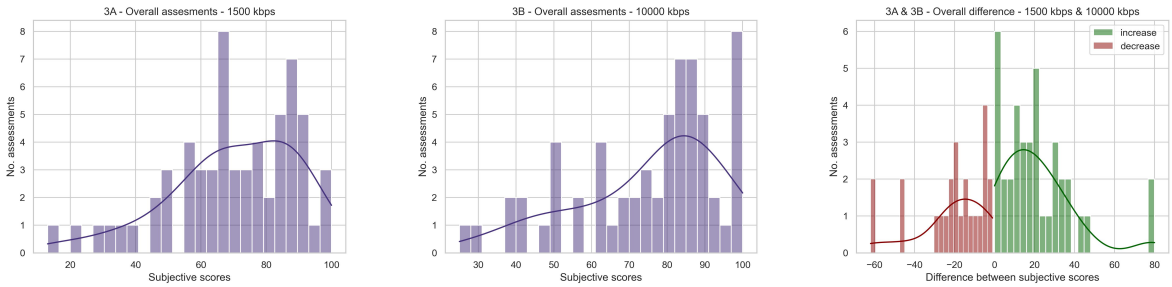
(a) 2A (b) 2B (c) 2A $\neq$ 2B

Rysunek 3.7 Rozkład ocen dla pary 2 sekwencji kodowanej zgodnie ze standardem H.264

wskazują, że największa liczba ocen dla pierwszego materiału miała wartości w okolicach 55 punktów. Natomiast dla drugiego była to wartość maksymalna wynosząca 100 punktów. Obraz ten uzyskał znaczną liczbę ocen doskonałych według 5-cio stopniowej skali ACR. Ze względu na wartości odstające, mediana oraz średnia arytmetyczna różnicy ocen osiągnęła pojedyncze wartości punktowe. Na histogramie 3.8 (c) widać jednak przewagę zielonego koloru, mówiącego o tym, że obraz B posiada subiektywnie lepszą jakość wideo.

Tabela 3.7 Oceny uzyskane dla pary nr 3 kodowanej zgodnie ze standardem H.264

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
3A	1500	13,0	59,0	70,0	86,8	95,8	70,0	19,8
3B	10000	25,0	63,0	80,0	87,8	100,0	74,9	19,3
3A $\neq$ 3B	wzrost o 8500	−62,0	−12,3	4,0	21,8	43,7	4,9	27,3



(a) 3A (b) 3B (c) 3A $\neq$ 3B

Rysunek 3.8 Rozkład ocen dla pary 3 sekwencji kodowanej zgodnie ze standardem H.264

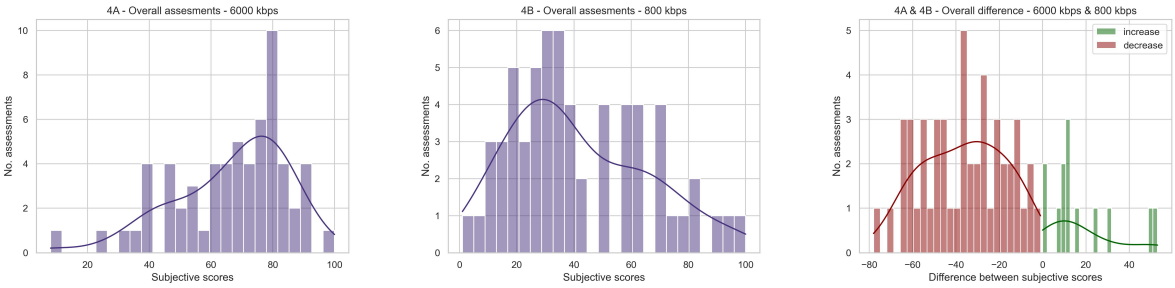
45:2386260771

Para nr 4

Treści o przepływnościach odpowiednio 6000 Kbit/s (A) oraz 800 Kbit/s (B). Dla tej pary odnotowano znaczącą różnicę ocen na niekorzyść materiału B. Mediana jest niższa o 35,5 punktu, natomiast średnia arytmetyczna o 25,1 punktu. Mediana dla materiału A wynosi 70,5 punktu i jest zbliżona do mediany pozostałych materiałów z grupy A wyświetlonych do tej pory. Na wykresach 3.9 (a) i (b) widać, że rozkład ocen dla materiału A jest lewoskośny. Uczestnicy badania byli w miarę zgodni co do poziomu jakości tego materiału. Dla obrazu B rozkład jest prawoskośny. Najczęściej występująca ocena jest mniejsza niż mediana i średnia, ale pewna liczba ocen jest w okolicach wartości domyślnej, czyli 50 punktów. Co najmniej 75% uczestników badania oceniło, że obraz B ma gorszą jakość niż obraz A. Mediana różnicy ocen wynosi 28 punktów na korzyść obrazu A. Najczęściej oceniający obniżali ocenę dla gorszego obrazu o 38 punktów. Na wykresie 3.9 (c) widać jednak, że część uczestników badania przyznała lepszą ocenę dla filmu o słabszej jakości.

Tabela 3.8 Oceny uzyskane dla pary nr 4 kodowanej zgodnie ze standardem H.264

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
4A	6000	8,0	55,0	70,5	80,0	90,0	66,6	18,5
4B	800	1,0	24,3	35,0	60,0	83,9	41,5	24,0
4A $\neq$ 4B	spadek o 5200	−78,0	−47,5	−28,0	−7,8	23,6	−25,1	29,1



(a) 4A (b) 4B (c) 4A $\neq$ 4B

Rysunek 3.9 Rozkład ocen dla pary 4 sekwencji kodowanej zgodnie ze standardem H.264

Para nr 5

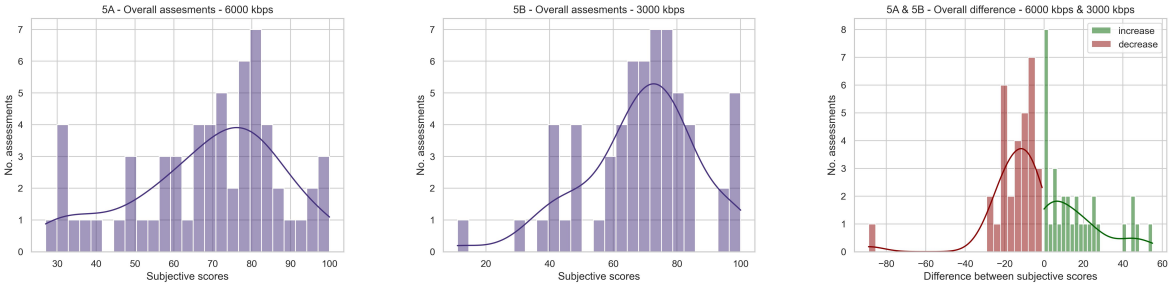
Treści o przepływnościach odpowiednio 6000 Kbit/s (A) oraz 3000 Kbit/s (B). Mediana i średnia arytmetyczna ocen jest zbliżona i wynosi odpowiednio 72,5 oraz 70 punktów, a także 68,5 i 68,7 punktu. Oceny różnią się skrajną wartością minimalną, która dla obrazu B o dwukrotnie niższej przepływności jest niższa o 16 punktów. Rozkład ocen dla materiału A widoczny na wykresie 3.10 (a) obejmuje zakres od 30 do 100 punktów, z tendencją zwiększonego występowania wartości w sąsiedztwie 70-80 punktów. Rozkład 3.10 (b) dla materiału B wykazuje zwiększone występowanie ocen dla wartości z zakresu od 60 do 80 punktów. Obecna jest także skrajnie słaba ocena o wartości poniżej 20 punktów. Na histogramie 3.10 (c) widać, że rozkład różnic ocen jest symetryczny za wyjątkiem pojedynczego elementu odstającego po ujemniej stronie wykresu. Mediana i średnia różnic ocen to odpowiednio 0,5 punktu na korzyść obrazu A oraz

46:4990908837

0,2 punktu na korzyść obrazu B. Uczestnicy badania uznali więc, że przeciętnie jakość obu wyświetleń jest taka sama.

Tabela 3.9 Oceny uzyskane dla pary nr 5 kodowanej zgodnie ze standardem H.264

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
5A	6000	27,0	58,3	72,5	81,0	95,0	68,5	18,9
5B	3000	11,0	60,3	70,0	78,8	98,0	68,7	18,1
5A $\neq$ 5B	spadek o 3000	−89,0	−12,5	−0,5	11,5	43,8	0,2	22,3



(a) 5A (b) 5B (c) 5A $\neq$ 5B

Rysunek 3.10 Rozkład ocen dla pary 5 sekwencji kodowanej zgodnie ze standardem H.264

Para nr 6

Treści o przepływnościach odpowiednio 800 Kbit/s (A) oraz 8000 Kbit/s (B). Pierwszy materiał uzyskał medianę ocen rzędu 63 punktów, natomiast drugi 78 punktów. Różnica między średnią ocen jest nieco mniejsza i wynosi 11,5 punktu. Wartości minimalne są zbliżone, jednak oceny maksymalne są zdecydowanie na korzyść drugiego materiału. Na wykresie 3.11 (a) widać, że rozkład ocen dla pierwszego materiału przypomina rozkład normalny z delikatną asymetrią lewostronną. Rozkład ocen dla drugiego materiału na rysunku 3.11 (b) pokazuje, że zdecydowana większość wartości znajduje się w górnej części skali ocen. Na histogramie 3.11 (c) widać, że co najmniej 75% oceniających prawidłowo określiło wyświetlenie B jako obraz o lepszej jakości wideo. Mediana i średnia arytmetyczna dla różnicy ocen obu wyświetleń wyniosła odpowiednio 13,5 oraz 11,6 punktu.

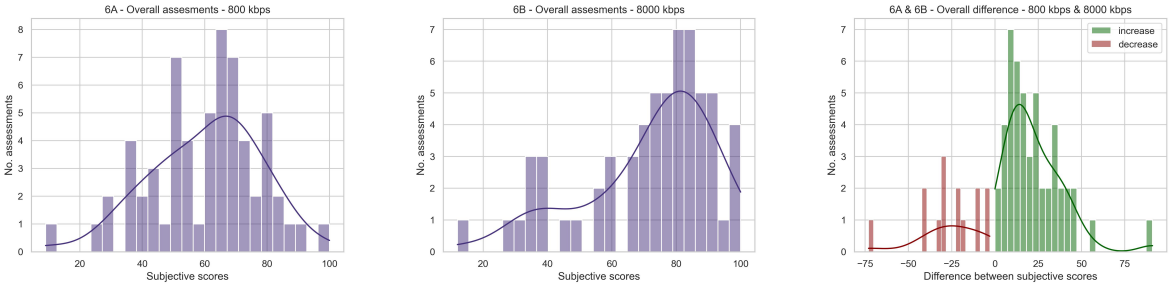
Tabela 3.10 Oceny uzyskane dla pary nr 6 kodowanej zgodnie ze standardem H.264

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
6A	800	9,0	50,0	63,0	70,8	84,9	60,3	17,6
6B	8000	12,0	60,5	78,0	85,0	99,7	71,8	20,2
6A $\neq$ 6B	wzrost 7200	−73,0	1,0	13,5	27,3	44,9	11,6	26,5

Para nr 7

Treści o przepływnościach odpowiednio 8000 Kbit/s (A) oraz 6000 Kbit/s (B). W parze numer 7 oba materiały mają zbliżoną przepływność bitową i mogą być traktowane jako para treści referencyjnych mająca na celu weryfikację czujności oceniajacych. Mediana ocen różni

47:2089932523



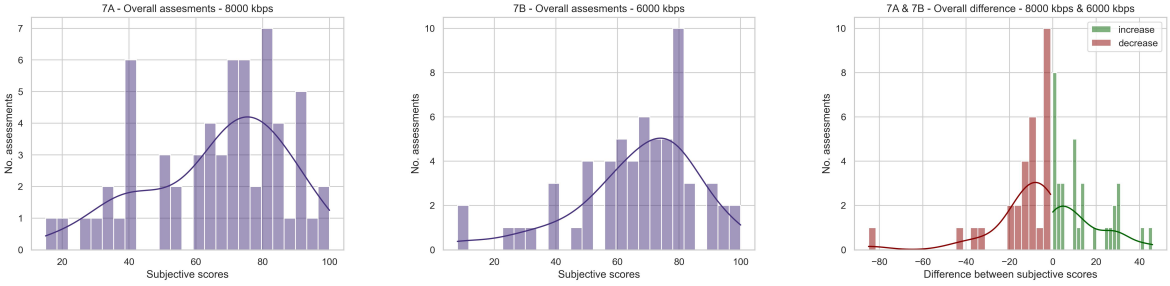
(a) 6A (b) 6B (c) 6A ≠ 6B

Rysunek 3.11 Rozkład ocen dla pary 6 sekwencji kodowanej zgodnie ze standardem H.264

się tylko o 1 punkt i wynosi 71 dla pierwszego obrazu oraz 70 dla drugiego obrazu. Średnia arytmetyczna ocen różni się zaledwie o 0,4 punktu i stanowi 66,3 oraz 66,7 punktów. Odchylenie standardowe jest zbliżone i osiąga wartość zbliżoną do 20 punktów. Na wykresie 3.12 (a) i (b) widać, że rozkłady ocen mają podobną tendencję do gromadzenia się w trzecim kwartale skali ocen. Ponadto, materiał A uzyskał pewną liczbę ocen dla wartości rzędu 40 punktów. Część uczestników badania była w stanie wskazać, że obraz drugi ma niższą przepływność i ocenić obraz B niżej niż obraz A. Na wykresie 3.12 (c) widać wartość dominującą w okolicy zera. Wyjątkiem jest nieznaczna liczba ocen, dla których różnica pomiędzy wyświetleniem A i B jest elementem odstającym.

Tabela 3.11 Oceny uzyskane dla pary nr 7 kodowanej zgodnie ze standardem H.264

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
7A	8000	15,0	51,0	71,0	81,5	93,0	66,3	20,7
7B	6000	8,0	59,0	70,0	80,0	93,8	66,7	19,9
7A ≠ 7B	spadek o 2000	−85,0	−10,0	0,0	10,0	30,0	0,4	20,4



(a) 7A (b) 7B (c) 7A ≠ 7B

Rysunek 3.12 Rozkład ocen dla pary 7 sekwencji kodowanej zgodnie ze standardem H.264

Para nr 8

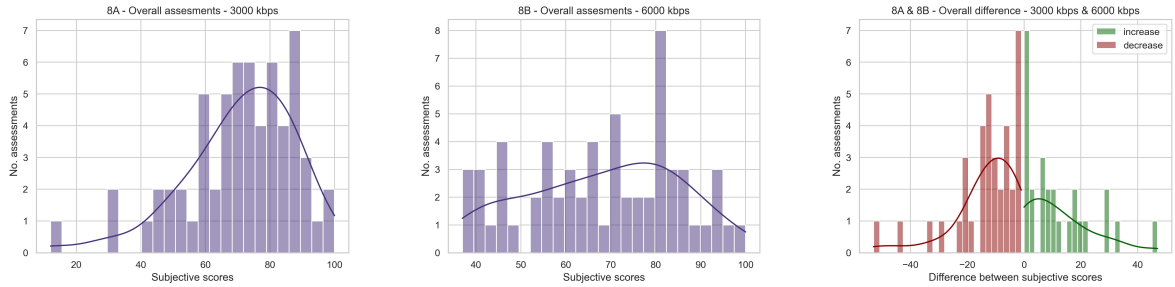
Treści o przepływnościach odpowiednio 3000 Kbit/s (A) oraz 6000 Kbit/s (B). W parze 8 treści mają taki sam poziom bitowy strumieni jak w parze numer 5, zmieniono jednak kolejność obrazów. Materiał A uzyskał medianę na poziomie 74 punktów i średnią arytmetyczną

48:4456578512

wynoszącą 70,7 punktu. Wartość środkowa ocen dla drugiego materiału to 70 punktów, a średnia arytmetyczna to 67,9. Na wykresach 3.13 (a) i (b) widać, że dla pierwszego materiału oceny są skupione wokół wyższych wartośći, jednak pojawiają się oceny skrajne w dolnej części skali. Dla drugiego materiału oceny mają tendencję do rozkładu jednostajnego z mniejszą liczbą ocen o najwyższych wartościach. Ta cecha jest niewelowana przez wartość minimalną, która wynosi 37 punktów i jest zauważalnie wyższa w stosunku do innych materiałów. Wykres różnic ocen 3.13 (c) ma podobną charakterystykę jak w parze poprzedzającej. Wykres jest symetryczny, wartość dominująca jest w okolicy zera. Mediana oraz średnia różnic ocen są nieznacznie ujemne, co wskazuje na znikome oszacowanie obrazu o wyższej przepływności jako ten o niższym poziomie jakości.

Tabela 3.12 Oceny uzyskane dla pary nr 8 kodowanej zgodnie ze standardem H.264

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
8A	3000	12,0	61,5	74,0	83,8	91,0	70,7	17,2
8B	6000	37,0	55,3	70,0	80,0	93,0	67,9	16,9
8A $\neq$ 8B	wzrost o 3000	−53,0	−13,0	−3,0	6,8	29,6	−2,8	17,2



(a) 8A (b) 8B (c) 8A $\neq$ 8B

Rysunek 3.13 Rozkład ocen dla pary 8 sekwencji kodowanej zgodnie ze standardem H.264

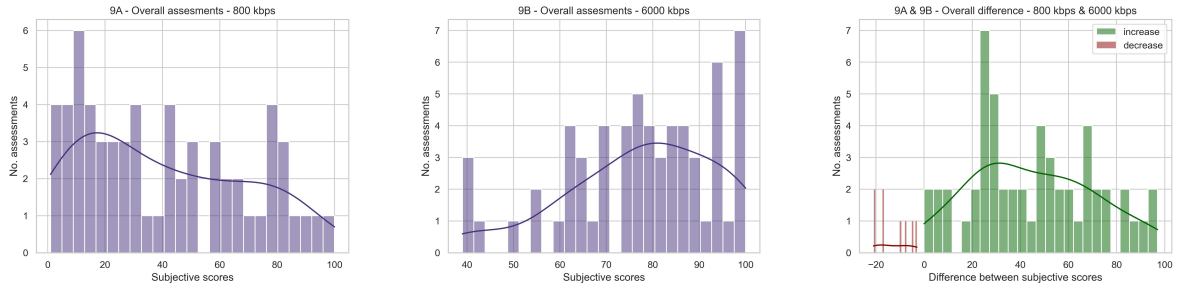
Para nr 9

Treści o przepływnościach odpowiednio 800 Kbit/s (A) oraz 6000 Kbit/s (B). Dobór przepływności strumieni jest taki sam jak dla pierwszej i czwartej pary. Tym razem materiał o mniejszej przepływności wideo jest prezentowany jako pierwszy. Mediana dla treści A to zaledwie 33,5 punktu. Średnia arytmetyczna ma wyższą wartość i wynosi 40 punktów. Materiał B osiągnął wartość mediany równą 79,5 punktu oraz 77,6 punktów dla średniej arytmetycznej. Wartość minimalna dla tego materiału wynosi 39 punktów i podobnie jak dla treści B w parze ósmej jest nieco wyższa od pozostałych materiałów. Rozkłady ocen na wykresach 3.14 (a) i (b) dla tych materiałów są naprzemiennie asymetryczne. Oceny dla materiału A obejmują całą skalę, jednak większość jest skumulowana w dolnej zakresu. Istotna część ocen dla materiału B to wartości powyżej 80 punktów. Wykres różnic ocen 3.14 (c) dla tej pary pokazuje, że większość oceniających określiło obraz B jako ten o lepszej jakości. Przeciętnie ocena jakości wideo zwiększyła się o ponad 37 punktów względem obrazu A.

49:3980108837

Tabela 3.13 Oceny uzyskane dla pary nr 9 kodowanej zgodnie ze standardem H.264

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
9A	800	1,0	14,3	33,5	63,5	85,0	40,0	28,6
9B	6000	39,0	66,8	79,5	90,0	100,0	77,6	16,2
9A $\neq$ 9B	wzrost o 5200	−21,0	20,8	36,0	59,8	88,8	37,7	31,0



(a) 9A (b) 9B (c) 9A $\neq$ 9B

Rysunek 3.14 Rozkład ocen dla pary 9 sekwencji kodowanej zgodnie ze standardem H.264

Para nr 10

Treści o przepływnościach odpowiednio 8000 Kbit/s (A) oraz 8000 Kbit/s (B). W tej parze materiały A i B są zdublowane. Jest to materiał referencyjny wyświetlony dwukrotnie. Mediana ocen dla wyświetlenia A i B to odpowiednio 72 i 73 punkty. Średnia arytmetyczna ocen wyniosła 67,8 i 69,1 punktu. Rozkłady ocen 3.15 (a) i (b) są do siebie zbliżone. Materiał A uzyskał więcej ocen w zakresie od 40 do 60 punktów i zakresie od 80 do 100 punktów. Treść B uzyskała mniej więcej równą liczbę ocen dla przedziału od 60 do 100 punktów. Dla obu materiałów największa liczba ocen to okolice wartości 80 punktów. Finalnie, obie treści uzyskały zbliżoną średnią ocenę. Uczestnicy badania najczęściej oceniali w podobny sposób materiały A i B. Świadczy o tym wartość dominująca w okolicy zera widoczna na rozkładzie różnic ocen 3.15 (c).

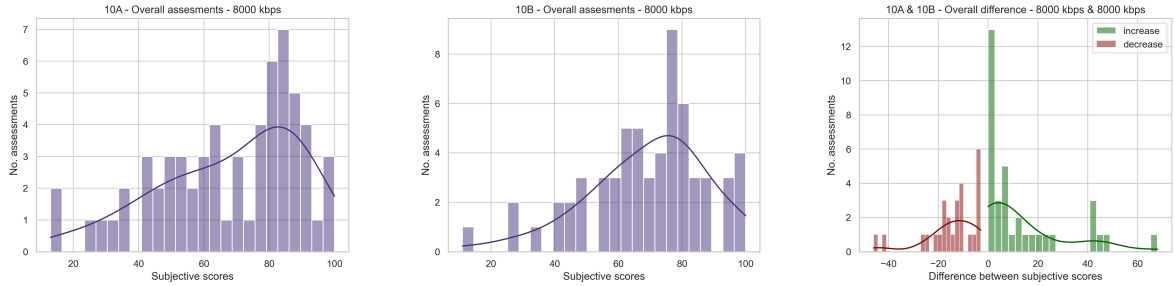
Tabela 3.14 Oceny uzyskane dla pary nr 10 kodowanej zgodnie ze standardem H.264

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
10A	8000	13,0	52,3	72,0	85,0	94,0	67,8	21,9
10B	8000	11,0	59,3	73,0	80,8	99,7	69,1	19,0
10A $\neq$ 10B	bez zmiany	−46,0	−11,0	0,0	7,0	42,9	1,3	20,5

Para nr 11

Treści o przepływnościach odpowiednio 10000 Kbit/s (A) oraz 800 Kbit/s (B). Materiały te posiadają skrajne wartości przepływności bitowej, odpowiednio maksymalną i minimalną dostępną w tej sekwencji. Treść A uzyskała medianę ocen wynoszącą 78 punktów, natomiast film B uzyskał 56 punktów. Średnia arytmetyczna ocen wynosi 72,9 oraz 53,8 punktu. Anomalią może być wartość minimalna dla materiału o najwyższej przepływności, która jest niższa niż dla drugiego materiału. Ogólnie, na rozkładach ocen 3.16 (a) oraz (b) widać, że wartości dla

50:3921850892



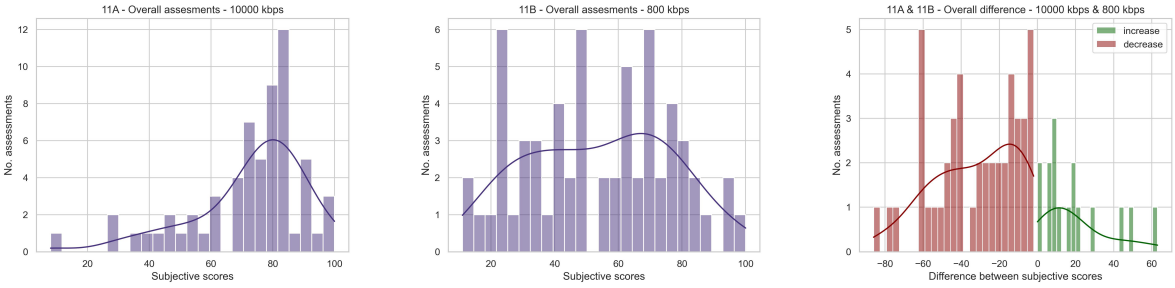
(a) 10A (b) 10B (c) 10A ≠ 10B

Rysunek 3.15 Rozkład ocen dla pary 10 sekwencji kodowanej zgodnie ze standardem H.264

materiału A są skupione w górnej części skali, natomiast dla B są one obecne w prawie całym dostępnym zakresie. 75% uczestników prawidłowo oceniło spadek jakości wideo materiału B względem materiału A. Przeciętnie spadek oceny wynosił 19 punktów. Na histogramie 3.16 (c) widać wartości dodatnie w zakresie od 0 do 60. Wartości powyżej 20 to pojedyncze wystąpienia. Oznacza to, że znikoma liczba uczestników określiła obraz B jako znacznie lepszy. Pozostali uczestnicy byli bardziej jednomyślni więc może się okazać, że dodatnie różnice będą zawierać się w elementach odstających dla tej pary.

Tabela 3.15 Oceny uzyskane dla pary nr 11 kodowanej zgodnie ze standardem H.264

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
11A	10000	8,0	68,5	78,0	85,0	94,9	72,9	18,5
11B	800	11,0	34,3	56,0	70,8	86,9	53,8	22,7
11A ≠ 11B	spadek o 9200	−86,0	−42,8	−16,0	−0,5	28,7	−19,0	31,4



(a) 11A (b) 11B (c) 11A ≠ 11B

Rysunek 3.16 Rozkład ocen dla pary 11 sekwencji kodowanej zgodnie ze standardem H.264

Para nr 12

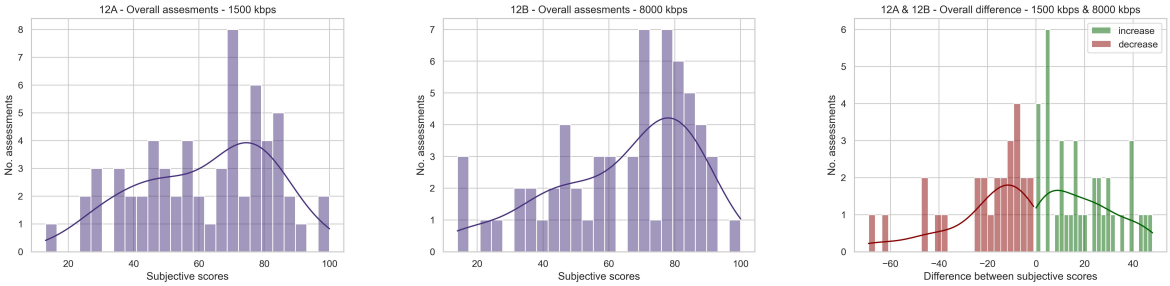
Treści o przepływnościach odpowiednio 1500 Kbit/s (A) oraz 8000 Kbit/s (B). Mediana ocen i średnia arytmetyczna dla materiału A to odpowiednio 68 punktów i 62 punkty. Dla treści B jest to 70 punktów i 64,4 punktu. Rozkłady ocen 3.17 (a) i (b) posiadają ujemną asymetryczność. Materiał B uzyskał nieco więcej ocen o wartościach w zakresie od 80 do 90 punktów, pojawiły

51:6263344811

się także oceny poniżej 20 punktów. Histogram różnic ocen 3.17 (c) wskazuje na nieco większą liczbę ocen wyższych dla materiału B. Wartość dominująca jest w pierwszym decylu wartości.

Tabela 3.16 Oceny uzyskane dla pary nr 12 kodowanej zgodnie ze standardem H.264

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
12A	1500	13,0	45,0	68,0	78,0	87,0	62,0	20,6
12B	8000	14,0	47,3	70,0	80,8	90,0	64,4	21,6
12A $\neq$ 12B	wzrost o 6500	−69,0	−10,0	4,0	19,8	40,0	2,4	25,7



(a) 12A (b) 12B (c) 12A $\neq$ 12B

Rysunek 3.17 Rozkład ocen dla pary 12 sekwencji kodowanej zgodnie ze standardem H.264

Para nr 13

Treści o przepływnościach odpowiednio 10000 Kbit/s (A) oraz 800 Kbit/s (B). Układ przepływności jest taki sam jak w parze nr 11. Mediana ocen dla najlepszego pod względem przepływu bitów materiału to 83 punkty, natomiast średnia wynosi 78,9 punktu. Są to najwyższe wartości dla zbioru materiałów sekwencji kodowanej H264. 25% uczestników badania przyznało ocenę wynoszącą co najmniej 90 punktów. Materiał o najsłabszej jakości wideo uzyskał odpowiednio wartości 64 i 61,5 dla mediany i średniej arytmetycznej. Rozkłady ocen 3.18 (a) i (b) pokazują, że znaczna liczba ocen dla materiału A jest powyżej 70 punktów, odchylenie standardowe jest jednym z niższych i wynosi 16 punktów. Rozkład dla materiału B obejmuje prawie cały zakres, a najwięcej ocen uzyskano w okolicach środkowego przedziału. Obraz B został przez większość uczestników określony jako ten o gorszej jakości wideo co wskazuje histogram 3.18 (c). Połowa oceniających uznała, że spadek jakości wynosi co najmniej 15,5 punktu.

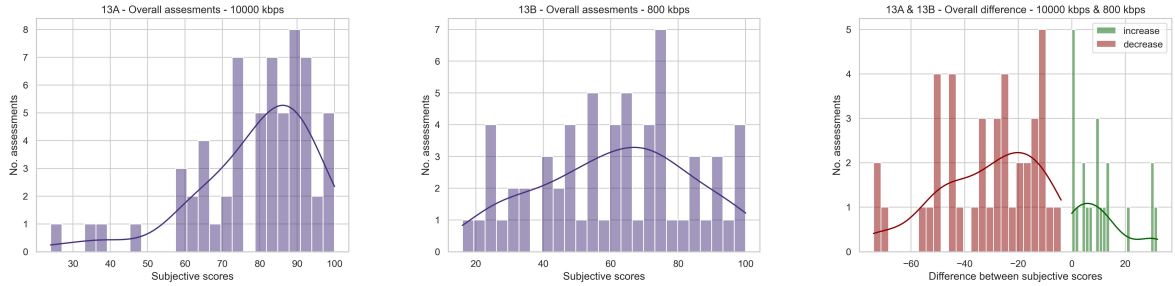
Tabela 3.17 Oceny uzyskane dla pary nr 13 kodowanej zgodnie ze standardem H.264

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
13A	10000	24,0	72,3	83,0	90,0	100,0	78,9	16,0
13B	800	16,0	45,5	64,0	75,8	99,7	61,5	22,5
13A $\neq$ 13B	spadek o 9200	−74,0	−34,0	−15,5	1,8	20,7	−17,4	25,8

Para nr 14

Treści o przepływnościach odpowiednio 8000 Kbit/s (A) oraz 10000 Kbit/s (B). W tej parze zasotosowano przepływności referencyjne. Mediana ocen dla treści A i B to odpowiednio 63,5

52:4165452794



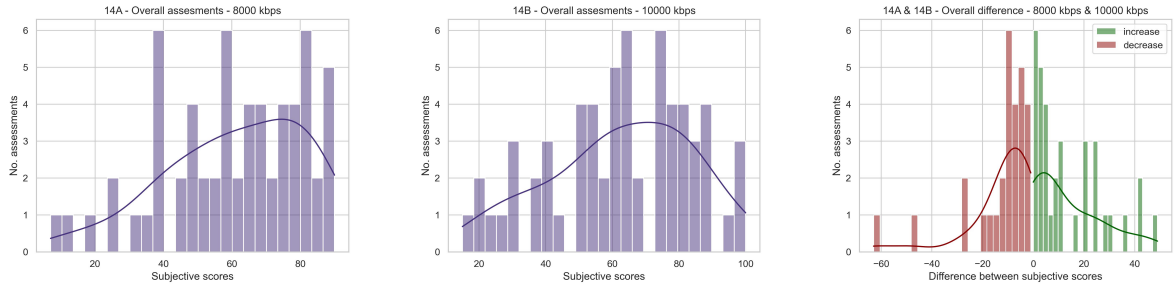
(a) 13A (b) 13B (c) 13A ≠ 13B

Rysunek 3.18 Rozkład ocen dla pary 13 sekwencji kodowanej zgodnie ze standardem H.264

punktu oraz 65 punktów. Średnia arytmetyczna to 60,8 i 62,6 punktu. Materiały zostały ocenione na podobnym poziomie, jednak wartości ocen wydają się dość niskie w stosunku do przepływności materiałów. Rozkłady ocen 3.19 (a) i (b) pokazują, że materiał A uzyskał dużą liczbę ocen na skraju przedziału, czyli 40, 60 i 80 punktów. Oceny dla materiału B mają tendencję do skupiania się w czwartym przedziale, czyli zakresie od 60 do 80 punktów. Mediana oraz średnia wartość różnicy ocen wyniosła odpowiednio 0 oraz 1,8 punktu. Uczestnicy podobnie ocenili wyświetlenia A i B. Na wykresie 3.19 (c) widać jednak, że istnieje niewielka liczba par ocen, dla których różnice zostały wskazane jako znaczne.

Tabela 3.18 Oceny uzyskane dla pary nr 14 kodowanej zgodnie ze standardem H.264

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
14A	8000	7,0	48,0	63,5	79,5	89,0	60,8	20,4
14B	10000	15,0	50,0	65,0	77,8	93,8	62,6	21,5
14A ≠ 14B	wzrost o 2000	−63,0	−7,5	0,0	9,5	36,7	1,8	19,2



(a) 14A (b) 14B (c) 14A ≠ 14B

Rysunek 3.19 Rozkład ocen dla pary 14 sekwencji kodowanej zgodnie ze standardem H.264

Para nr 15

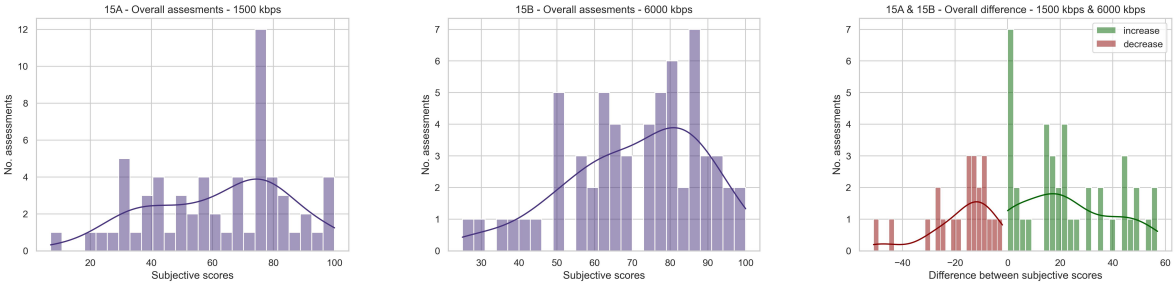
Treści o przepływnościach odpowiednio 1500 Kbit/s (A) oraz 6000 Kbit/s (B). Ostatnia para filmów uzyskała medianę ocen wynoszącą 68 punktów oraz 74,5 punktów, odpowiednio dla materiałów A i B. Średnia arytmetyczna ocen jest nieco niższa i wynosi 62,2 i 70,8 punktu. Rozkład ocen 3.20 (a) dla materiału A pokazuje, że dominanta jest w okolicach 80 punktów

53:9114563088

i znacznie się wyróżnia na tle częstotliwości występowania pozostałych ocen. Rozkład 3.20 (b) dla materiału B jest lewoskośny, najczęściej występujące oceny mają wartości z zakresu od 50 do 90 punktów. Histogram 3.20 wskazuje, że znacznie częściej drugie wyświetlenie zostało ocenione jako lepsze. Wartość dominująca wynosi 0, co może sugerować, że część uczestników nie była zdecydowana przy próbie określenia referencji.

Tabela 3.19 Oceny uzyskane dla pary nr 15 kodowanej zgodnie ze standardem H.264

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
15A	1500	7,0	41,5	68,0	76,8	99,7	62,2	22,6
15B	6000	25,0	60,3	74,5	85,0	95,9	70,8	17,9
15A $\neq$ 15B	wzrost o 4500	−51,0	−9,8	4,5	23,5	49,9	8,6	25,2



(a) 15A (b) 15B (c) 15A $\neq$ 15B

Rysunek 3.20 Rozkład ocen dla pary 15 sekwencji kodowanej zgodnie ze standardem H.264

3.9.2 Wyniki dla par kodowanych zgodnie ze standardem H.265

Podobnie jak w poprzednim podrozdziale, wyniki dla par kodowanych zgodnie ze standardem H.265 zostały wyodrębione i opisane poniżej. W kolumnie Kbit/s podano przepływność danego strumienia. Dla wiersza zawierającego różnicę ocen zawarto wartość wzrostu lub spadku przepływności pomiędzy dwoma wyświetleniami. Dla każdego materiału z pary wygenerowano histogram ocen wraz z estymatorem gęstości, aby dokładniej porównać rozkłady ocen.

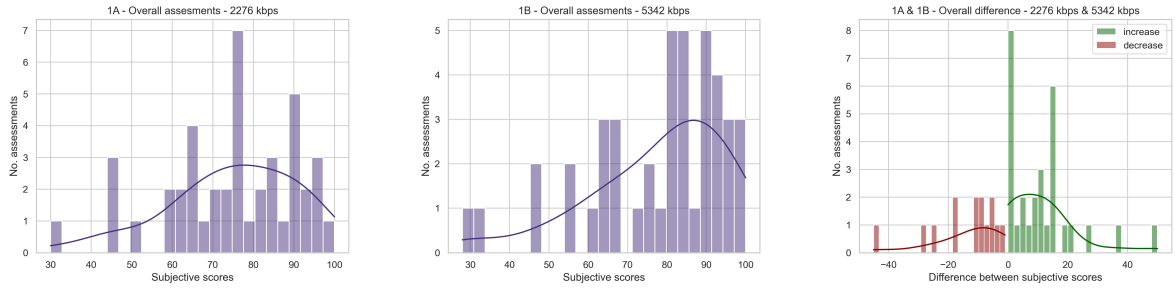
Para nr 1

Treści o przepływnościach odpowiednio 2276 Kbit/s (A) oraz 5342 Kbit/s (B). Mediana i średnia ocen dla materiału A to odpowiednio 76 i 74,7 punktów. Dla drugiego obrazu wartości wynoszą 82 i 77,6 punktów. Wartości odchylenie standardowego dla obu treści wydają się być nieco mniejsze niż dla materiałów z poprzedniego badania i wynoszą odpowiednio 16,1 i 17,7 punktów. Referencyjny materiał B uzyskał ocenę o wartości co najmniej 90 punktów od 25% uczestników badania. Rozkłady ocen przedstawiono na wykresach 3.21 (a) i (b). Dla obu treści uzyskano nieznaczną liczbę ocen poniżej 50 punktów. Rozkład dla materiału A wydaje się być spłaszczony w porównaniu do rozkładu ocen drugiego materiału. Histogram 3.21 (c) wskazuje, że obraz B uzyskał znacznie częściej oceny wyższe od wyświetlenia A. Wartością dominującą jest jednak wartość zerowa. Dodatkowo wartości różnicy ocen skupione są w sąsiedztwie dominanty. Może to sugerować brak jednoznaczności przy określaniu materiału o lepszej jakości wideo.

54:7586338027

Tabela 3.20 Oceny uzyskane dla pary nr 1 kodowanej zgodnie ze standardem H.265

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
1A	2276	30,0	65,5	76,0	87,5	95,9	74,7	16,1
1B	5342	28,0	67,0	82,0	90,0	99,5	77,6	17,7
1A $\neq$ 1B	wzrost o 3066	−45,0	−5,0	0,0	13,0	25,4	2,9	16,5



(a) 1A (b) 1B (c) 1A $\neq$ 1B

Rysunek 3.21 Rozkład ocen dla pary 1 sekwencji kodowanej zgodnie ze standardem H.265

Para nr 2

Treści o przepływnościach odpowiednio 5342 Kbit/s (A) oraz 5342 Kbit/s (B). W tej parze oba materiały to taka sama treść referencyjna. Mediana dla wyświetlenia A i B jest taka sama i wyniosła 70 punktów. Średnia różni się o nieco ponad 2 punkty i wynosi odpowiednio 67,8 i 65,6 punktów. Wartości minimalne dla obu wyświetleń to 33 punkty dla obrazu A i 20 punktów dla obrazu B. Dla pierwszego kwartyla różnica pomiędzy dokonanymi ocenami zmniejsza się i wynosi 5,5 punktu. Dla 25% badanych obrazy różniły się o 1,5 punktu. Rozkłady ocen dla obu wyświetleń przedstawiono na wykresach 3.22 (a) i (b). Rozkład ocen materiału A ma charakter symetryczny dla zakresu wartości od 30 do 100 punktów. Obraz B uzyskał niższą ocenę minimalną, więc rozkład jest asymetryczny. Wartość dominująca dla obu rozkładów jest różna o około 10 punktów. Wykres 3.22 (c) pokazuje, że dla tej pary uczestnicy badania głosowali w zróżnicowany sposób dla obu wyświetleń. Część uczestników oszacowała, że obraz B ma gorszą jakość i przyznali oceny niższe od kilku do prawie 50 punktów. Pozostali uczestnicy określili, że obraz B miał lepszą jakość i to on zasługuje na punktację wyższą, podczas gdy obie treści to był ten sam materiał wyświetlony dwukrotnie.

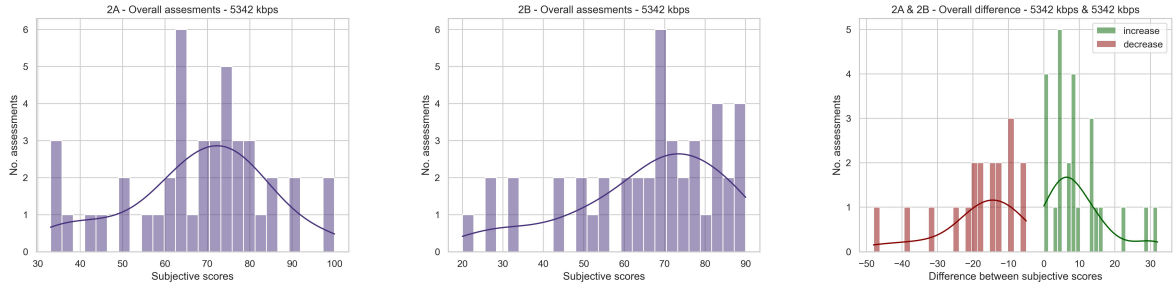
Tabela 3.21 Oceny uzyskane dla pary nr 2 kodowanej zgodnie ze standardem H.265

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
2A	5342	33,0	60,5	70,0	77,5	90,0	67,8	16,6
2B	5342	20,0	55,0	70,0	79,0	89,9	65,6	18,6
2A $\neq$ 2B	bez zmiany	48,0	−13,5	0,0	8,0	21,4	−2,2	17,0

Para nr 3

Treści o przepływnościach odpowiednio 3298 Kbit/s (A) oraz 5342 Kbit/s (B). Mediana ocen dla treści A wynosi 71 punktów, a średnia 70,1 punktu. Obraz B, o wyższej przepływności

55:9791809723



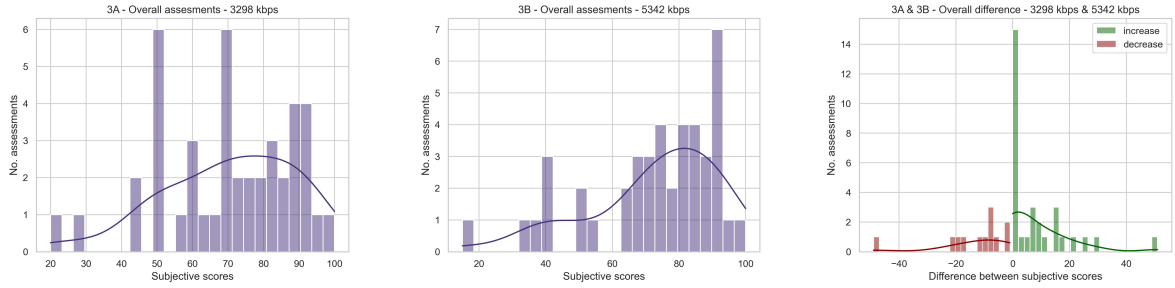
(a) 2A (b) 2B (c) 2A ≠ 2B

Rysunek 3.22 Rozkład ocen dla pary 2 sekwencji kodowanej zgodnie ze standardem H.265

uzyskał medianę wyższą o 6 punktów, natomiast średnią wyższą o 2,3 punktu. Minimalna ocena jest jednak niższa dla obrazu B o 5 punktów. Na rozkładzie ocen 3.23 (a) i (b) widać, że ocena minimalna dla drugiego obrazu może mocno wpływać na jego średnią, gdyż pozostałe oceny są skupione w zakresie od 65 do 100 punktów. Obraz A uzyskał kilka ocen o niższych wartościach, wartością dominującą jest ocena domyślna, czyli wartość 50 oraz okolice wartości 70 punktów. Wyświetlenia A i B w tej parze najczęściej uzyskiwały takie same oceny. Histogram 3.8 (c), że wskazana różnica jakości między tymi materiałami wynosiła 0.

Tabela 3.22 Oceny uzyskane dla pary nr 3 kodowanej zgodnie ze standardem H.265

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
3A	3298	20,0	58,5	71,0	84,5	92,9	70,1	18,6
3B	5342	15,0	66,0	77,0	87,0	93,0	72,4	19,6
3A ≠ 3B	wzrost o 2044	−49,0	−1,5	0,0	9,0	24,7	2,3	15,4



(a) 3A (b) 3B (c) 3A ≠ 3B

Rysunek 3.23 Rozkład ocen dla pary 3 sekwencji kodowanej zgodnie ze standardem H.265

Para nr 4

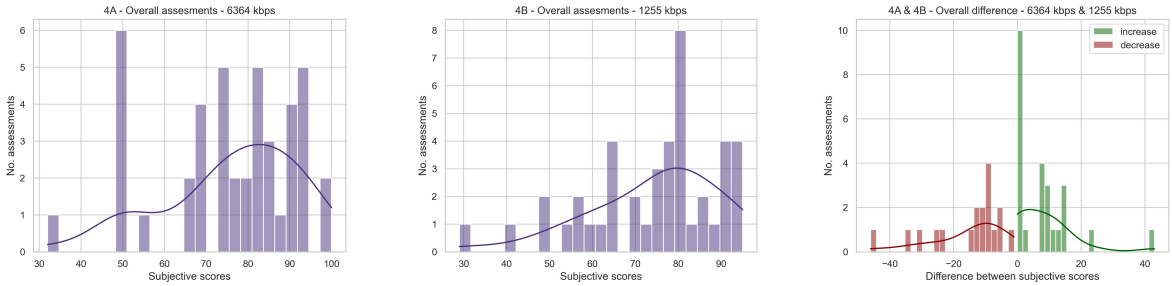
Treści o przepływnościach odpowiednio 6364 Kbit/s (A) oraz 1255 Kbit/s (B). W tej parze wyświetlane są materiały o najwyższej i najniższej dostępnej przepływności w ramach sekwencji kodowanej H265. Materiał A uzyskał medianę na poziomie 80 punktów, średnia wynosi 76 punktów. Materiał B to mediana równa 77 punktów, natomiast średnia wyniosła 74,2 punktu. W ogólnym podsumowaniu widać, że ponad pięciokrotnie niższy strumień uzyskał zbliżone

56:1618794683

oceny. Wartości minimalne ocen były podobne i wynosiły odpowiednio 32 i 29 punktów. 25% badanych oceniłoby obrazy na co najmniej 89,5 oraz 84 punktów. Rozkłady ocen 3.24 (a) i (b) pokazują, że dla materiału A wartością dominującą jest 50 i jest to kolejny obraz, którego najczęstsza wartość oceny to ocena domyślna. Może to sugerować, że uczestnicy badania nie są zdecydowani jaką ocenę mają przyznać i przechodzą dalej, do kolejnej strony udzielając niejawnie oceny domyślnej. Uczestnicy, którzy jawnie wybrali ocenę, zaznaczyli przedział od 70 do 90 punktów. Ocena domyślna może więc zaniżać średnią ocenę dla treści A. Materiał B uzyskał pojedyncze oceny poniżej 40 punktów, natomiast większość ocen zawierało się w przedziale od 60 do 90 punktów. Podobnie jak dla poprzedniej pary, wskazana różnica jakości dla obu materiałów najczęściej wynosiła 0. Wykres 3.24 (c) pokazuje, że rozkład różnic ocen pomiędzy wyświetleniami jest symetryczny. Rozkład ten zawiera się głownie w zakresie od 0 do 20 punktów na korzyść zarówno materiału A oraz B z pojedynczymi elementami odstającymi.

Tabela 3.23 Oceny uzyskane dla pary nr 4 kodowanej zgodnie ze standardem H.265

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
4A	6364	32,0	69,5	80,0	89,5	94,0	76,0	16,2
4B	1255	29,0	65,0	77,0	84,0	94,0	74,2	15,2
4A $\neq$ 4B	spadek o 5109	−46,0	−10,0	0,0	8,0	15,0	−1,7	15,8



(a) 4A (b) 4B (c) 4A $\neq$ 4B

Rysunek 3.24 Rozkład ocen dla pary 4 sekwencji kodowanej zgodnie ze standardem H.265

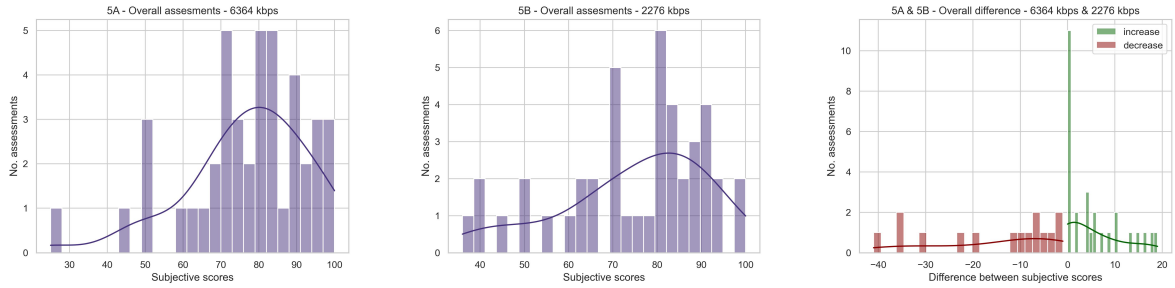
Para nr 5

Treści o przepływnościach odpowiednio 6364 Kbit/s (A) oraz 2276 Kbit/s (B). Mediana ocen dla materiału o najwżyszej przepływności wynosi 80 punktów, natomiast średnia to 76,7 punktów. Obraz B uzyskał taką samą medianę ocen, natomiast średnia wyniosła 74,7. Rozkład ocen 3.25 (a) oraz (b) ma taką samą asymetrię, zakres ocen jest zbliżony z wyłączeniem oceny minimalnej dla obrazu A, która wydaje się odstawać od reszty wartości. Dla obrazu B rozkład jest spłaszczony względem drugiego materiału. Jest to związane z większą liczbą ocen o wartości poniżej 60 punktów. Materiał A uzyskał zdecydowaną większość ocen w zakresie wartości od 70 do 100. Jest to kolejna para obrazów, w której najczęściej udzielono takiej samej oceny dla obu wyświetleń. Na wykresie 3.25 (c) widać, że wartość dominująca ma znaczną przewagę nad pozostałymi różnicami ocen. Podobnie jak poprzednio, uczestnicy podzielili się na dwie, podobnie licznie grupy, doceniając zarówno obraz A jak i B.

57:1112243322

Tabela 3.24 Oceny uzyskane dla pary nr 5 kodowanej zgodnie ze standardem H.265

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
5A	6364	25,0	70,0	80,0	88,0	98,7	76,7	16,1
5B	2276	36,0	66,0	80,0	87,5	94,0	74,7	16,6
5A $\neq$ 5B	spadek o 4088	−41,0	−6,0	0,0	5,5	15,9	−2,0	13,8



(a) 5A (b) 5B (c) 5A $\neq$ 5B

Rysunek 3.25 Rozkład ocen dla pary 5 sekwencji kodowanej zgodnie ze standardem H.265

Para nr 6

Treści o przepływnościach odpowiednio 6364 Kbit/s (A) oraz 3298 Kbit/s (B). Materiał A uzyskał medianę ocen o wartości 75 punktów, średnia wyniosła 75,6. Dla materiału B mediana i średnia to odpowiednio 71 oraz 68,7 punktów. Minimalna ocena dla obrazu A wynosi 40 punktów i jest to najwyższa ocena minimalna spośród wszystkich obrazów tej sekwencji. Rozkłady ocen 3.26 (a) oraz (b) pokazują, że oceny dla materiału A są rozrzucone w górnej połowie skali i generalnie nie mają tendencji do skupiania w jednym ograniczonym zakresie. Wartości dominujące to oceny powyżej 90 punktów. Rozkład ocen dla materiału B ma podobne cechy, jednak wartością dominującą jest ponownie wartość domyślna, czyli 50 punktów. Podobniej jak dla poprzednich par, oceniający często udzielili takiej samej oceny dla obu materiałów. Niemniej jednak, na wykresie 3.11 (c) przeważa kolor czerwony mówiący o tym, że dla obrazu B uczestnicy obniżali ocenę względem wyświetlenia A. Pojedyncze oceny zostały obniżone aż o 50 punktów, jednak w głównej mierze wartości zostały obniżone o maksymalnie 30 punktów. Podobnie jak poprzednio, część uczestników udzieliło wyższej punktacji materiałowi B, który miał niższą jakość.

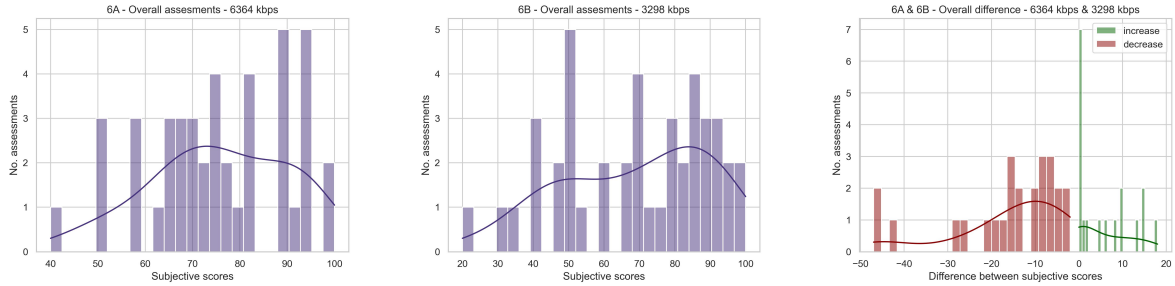
Tabela 3.25 Oceny uzyskane dla pary nr 6 kodowanej zgodnie ze standardem H.265

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
6A	6364	40,0	66,5	75,0	90,0	94,0	75,6	14,8
6B	3298	20,0	50,0	71,0	86,0	94,9	68,7	21,0
6A $\neq$ 6B	spadek o 3066	−47,0	−14,5	−4,0	0,5	14,8	−6,9	15,2

Para nr 7

Treści o przepływnościach odpowiednio 1255 Kbit/s (A) oraz 5342 Kbit/s (B). Materiał A posiada najniższą dostępną przepływność bitową w tej sekwencji. Pomimo tego uzyskał taką

58:5899446035



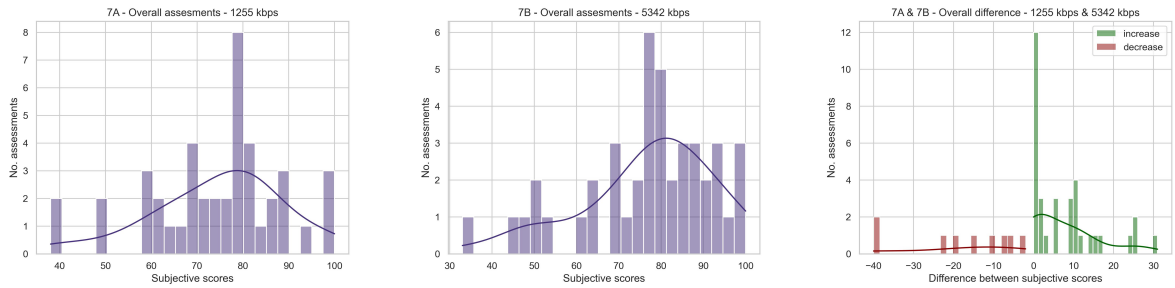
(a) 6A (b) 6B (c) 6A ≠ 6B

Rysunek 3.26 Rozkład ocen dla pary 6 sekwencji kodowanej zgodnie ze standardem H.265

samą medianę ocen jak materiał B. Wartość ta wyniosła 79 punktów. Średnia ocen dla tych materiałów to odpowiednio 74,4 oraz 76,7 punktów. Wartość odchylenia standardowego dla materiału A wynosi 14,4 punktu i jest najniższa spośród wszystkich treści w obu badaniach. Rozkłady ocen dla tych materiałów zostały przedstawione na wykresach 3.27 (a) i (b). Zakres uzyskanych ocen jest zbliżony. Materiał B uzyskał niższą wartość minimalną tj. 33 punkty wobec 38 dla materiału A. Wartości dominujące znajdują się na tym samym poziomie i obejmują sąsiedztwo 80 punktów. Wykres 3.27 (c) ponownie prezentuje dominującą wartość zerową nad pozostałymi różnicami ocen. Spadek oceny dla strumienia o przepływności 5342 Kbit/s aż o 40 punktów względem strumienia o przepływności 1255 Kbit/s wydaję się być elementem odstającym. Wartości są subiektywnym odczuciem, więc poddając je analizie na tym etapie, można stwierdzić, że dla tej pary 25% uczestników błędnie wskazało strumień o wyższej przepływności bitowej.

Tabela 3.26 Oceny uzyskane dla pary nr 7 kodowanej zgodnie ze standardem H.265

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
7A	1255	38,0	67,5	79,0	81,5	97,6	74,4	14,4
7B	5342	33,0	68,5	79,0	87,0	99,6	76,7	15,7
7A ≠ 7B	wzrost o 4087	−40,0	0,0	2,0	10,0	24,9	2,3	14,4



(a) 7A (b) 7B (c) 7A ≠ 7B

Rysunek 3.27 Rozkład ocen dla pary 7 sekwencji kodowanej zgodnie ze standardem H.265

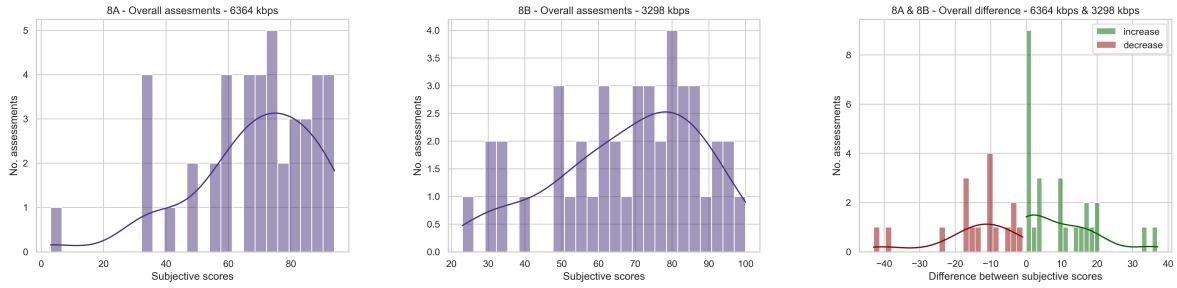
59:4394929349

Para nr 8

Treści o przepływnościach odpowiednio 6364 Kbit/s (A) oraz 3298 Kbit/s (B). Mediana ocen dla obu treści wynosi odpowiednio 72 i 71 punktów, natomiast średnia to 68,2 i 68 punktów. Średnia ocen dla materiału o wyższej przepływności może być nieco zaniżona ze względu na bardzo niską ocenę minimalną. Jeden z uczestników badania przyznał ocenę o wartośći 3. Jest to jedyna tak niska ocena. Rozkłady ocen dla tych materiałów są widoczne na wykresach 3.28 (a) i (b). Wartości ocen dla materiału A są skupione w zakresie od 65 do 95 punktów. Oceny dla materiału B znajdują się na nieco szerszym zakresie skali i obejmują na równym poziomie punktację od 50 do 95. Zarówno w pierwszym jak i drugim wyświetleniu treści uzyskały nieznaczną liczbę ocen poniżej 40 punktów. Histogram 3.28 (c) prezentuje symetryczny rozkład z wartością dominującą w centrum. Pomimo niższej o połowę przepływności dla obrazu B, wskazane różnice ocen rozkładają się po równo. Zarówno mediana jak i średnia przyjmuje wartości zerowe lub bliskie zeru. Pierwszy i trzeci kwartyl przyjmuje takie same wartości absolutne. Rozkład ten świadczy o prawie idealnym podziale uczestników badania i problemie związanym z trudnością w oszacowaniu obrazu o lepszej i gorszej jakości wideo dla zastosowanego kodowania.

Tabela 3.27 Oceny uzyskane dla pary nr 8 kodowanej zgodnie ze standardem H.265

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
8A	6364	3,0	59,0	72,0	84,0	92,8	68,2	20,1
8B	3298	23,0	56,0	71,0	84,0	93,9	68,0	19,7
8A $\neq$ 8B	spadek o 3066	−43,0	−10,5	0,0	10,0	20,0	−0,1	16,1



(a) 8A (b) 8B (c) 8A $\neq$ 8B

Rysunek 3.28 Rozkład ocen dla pary 8 sekwencji kodowanej zgodnie ze standardem H.265

Para nr 9

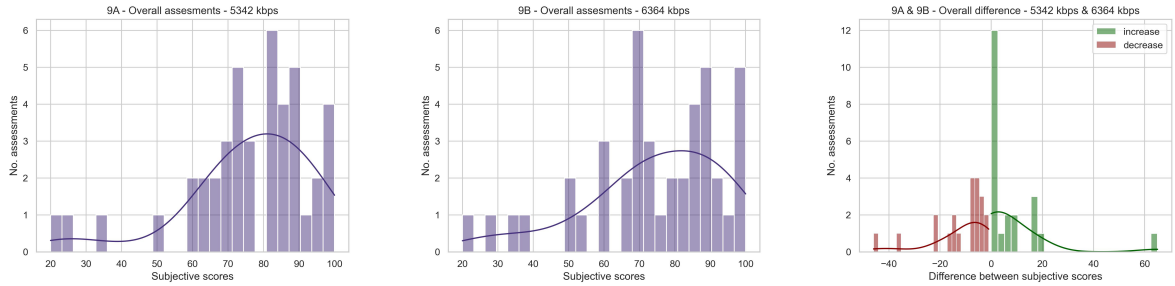
Treści o przepływnościach odpowiednio 5342 Kbit/s (A) oraz 6364 Kbit/s (B). W tej parze oba wyświetlenia mogą być traktowane jako zestaw treści referencyjnych. Mediana ocen wynosi odpowiednio 81 oraz 77 punktów dla wyświetlenia A i B oraz 75,9 i 74,2 dla średniej ocen. Wartości są zbliżone do siebie. Ocena minimalna jest taka sama i wynosi 20 punktów. Na rozkładach ocen 3.29 (a) i (b) widać, że udzielone oceny obejmują taki sam zakres, czyli od 20 do 100 punktów. Poniżej wartości 50 znajduje sie mała liczba ocen dla obu treści. Dla pierwszego wyświetlenia szczyt popularności ocen znajduje się w sąsiedztwie 80 punktów, ze zwiększoną liczbą ocen o wartościach około 70 oraz 85 punktów. Dla materiału B dominanta

60:8918709519

to wartość około 70 punktów, natomiast dla zakresu od 85 do 100 także obserwowana jest zwiększona liczba ocen. Rozkład różnic ocen 3.29 (c) zawiera wartości odstające w sąsiedztwie oceny o 60 punktów wyższej na korzyść obrazu B oraz około 40 punktów wyższej na korzyść obrazu A. Za wyjątkiem tych wartości, rozkład jest zbliżony do rozkładu dla pary poprzedniej. Dla tej pary symetryczny rozkład różnic ocen jest oczekiwany, gdyż para składa się z dwóch obrazów referencyjnych.

Tabela 3.28 Oceny uzyskane dla pary nr 9 kodowanej zgodnie ze standardem H.265

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
9A	5342	20,0	69,0	81,0	88,0	99,7	75,9	18,3
9B	6364	20,0	65,0	77,0	88,5	100,0	74,2	19,8
9A $\neq$ 9B	spadek o 1022	−46,0	−7,0	0,0	1,5	16,0	−1,7	16,3



(a) 9A (b) 9B (c) 9A $\neq$ 9B

Rysunek 3.29 Rozkład ocen dla pary 9 sekwencji kodowanej zgodnie ze standardem H.265

Para nr 10

Treści o przepływnościach odpowiednio 5342 Kbit/s (A) oraz 3298 Kbit/s (B). Mediana dla obrazu A to 85 punktów, natomiast średnia to 82,5 punktu. Drugie wyświetlenie uzyskało oceny na poziomie 80 punktów dla mediany i 78,5 punktu dla średniej. Zaróno mediana jak i średnia ocen dla treści A jest najwyższa z całego zbioru ocen sekwencji badawczej. Z drugiej strony materiał ten uzyskał niższą ocenę minimalną niż drugie wyświetlenie posiadające niższą przepływność bitową. Różnica między tymi wartościami to 10 punktów. Oba wyświetlenia posiadają niską wartość odchylenia standardowego, różniącą się tylko jednym punktem. Potwierdzają to rozkłady danych 3.30 (a) i (b). Dla obu materiałów największa liczba ocen to przedział od 70 do 100 punktów. Nieznaczna liczba ocen posiada wartości poniżej 50 punktów. Dla wyświetlenia A znacząco dominującym zakresem jest punktacja od 90 do 100. Ponownie rozkład różnic ocen 3.30 (c) prezentuje znaczącą dominację równej oceny dla obu wyświetleń. Pozostałe, niezerowe wartości, znajdują się po ujemnej stronie tego wykresu. Wskazuje to, że część uczestników prawidłowo rozpoznała obraz o niższej przepływności i obniżyła swoją ocenę dla drugiego wyświetlenia.

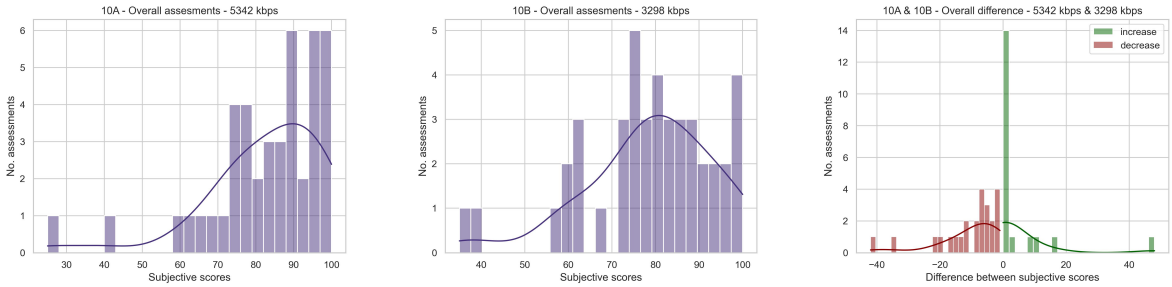
Para nr 11

Treści o przepływnościach odpowiednio 6364 Kbit/s (A) oraz 1255 Kbit/s (B). W tej parze

61:4042976463

Tabela 3.29 Oceny uzyskane dla pary nr 10 kodowanej zgodnie ze standardem H.265

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
10A	5342	25,0	75,5	85,0	94,5	100,0	82,5	15,8
10B	3298	35,0	73,0	80,0	88,0	99,9	78,5	14,8
10A $\neq$ 10B	spadek o 2044	−42,0	−7,5	−2,0	0,0	9,8	−4,0	13,2



(a) 10A (b) 10B (c) 10A $\neq$ 10B

Rysunek 3.30 Rozkład ocen dla pary 10 sekwencji kodowanej zgodnie ze standardem H.265

ponownie wyświetlane są materiały o najwyższej i najniższej przepływności bitowej. Materiał A uzyskał medianę ocen równą 68 punktów oraz średnią wynoszącą 65,4. Drugie wyświetlenie posiada medianę o wartości zaledwie 56 punktów i średnią wynoszącą 55,4. Różnica w punktacji pomiędzy wyświetleniami jest zauważalna, jednak oceny dla treści A nieco odbiegają od podobnych poziomów przepływności bitowej w innych parach. Wyświetlenia uzyskały taką samą ocenę minimalną równą 18 punktów. Rozkłady ocen 3.31 (a) oraz (b) posiadają ten sam zakres ocen za wyjątkiem jednej maksymalnej oceny dla materiału A. Rozkład dla treści A wykazuje asymetrię lewostronną. Dla treści B rozkład ocen jest symetryczny i spłaszczony. Wykres różnic ocen 3.31 (c) wskazuje na częstsze obniżenie oceny dla obrazu o niższej przepływności. Dominuje spadek oceny o co najmiej 15 punktów. Część oceniających przyznało taką samą ocenę dla obu wyświetleń podobnie jak dla wcześniejszych par.

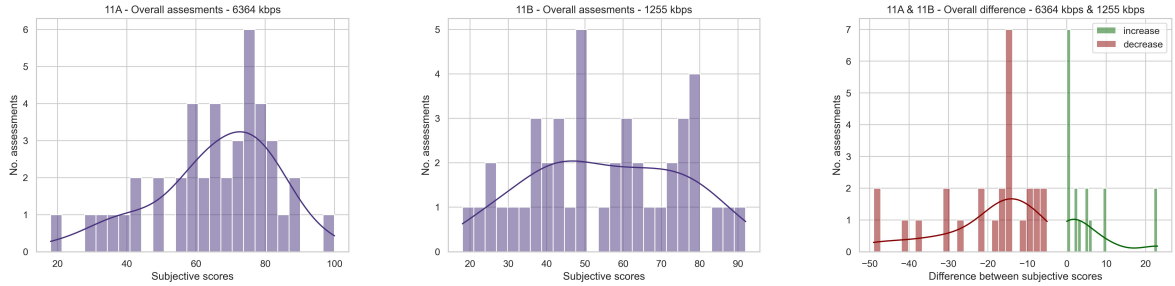
Tabela 3.30 Oceny uzyskane dla pary nr 11 kodowanej zgodnie ze standardem H.265

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
11A	6364	18,0	58,0	68,0	77,0	87,8	65,4	17,8
11B	1255	18,0	40,5	56,0	72,5	85,4	55,4	19,6
11A $\neq$ 11B	spadek o 5109	−49,0	−16,0	−10,0	0,0	10,0	−10,0	16,5

Para nr 12

Treści o przepływnościach odpowiednio 2276 Kbit/s (A) oraz 6364 Kbit/s (B). Mediana ocen dla treści A wynosi 73, a średnia to 68,7 punktów. Dla materiału B są to wartości wynoszące 80 oraz 75,9 punktów. Materiał A uzyskał ocenę minimalną o najmniejszej dostępnej wartości czyli 1. Rozkłady ocen 3.32 (a) oraz (b) pokazują, że materiał A uzyskał więcej ocen poniżej 50 punktów. Materiał B uzyskał więcej ocen w zakresie od 90 do 100 punktów. Ocena dominująca jest przesunięta o około 10 punktów na korzyść treści o wyższej przepływności. Histogram

62:7597280599



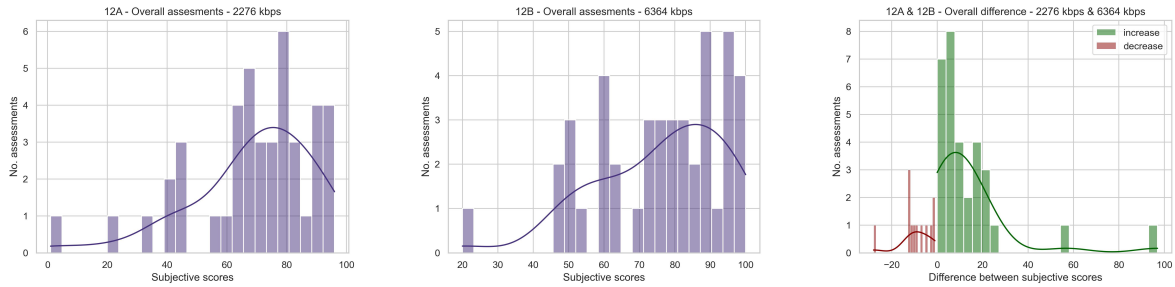
(a) 11A (b) 11B (c) 11A ≠ 11B

Rysunek 3.31 Rozkład ocen dla pary 11 sekwencji kodowanej zgodnie ze standardem H.265

3.32 (c) wskazuje, że znaczna większość oceniających dostrzegła poprawę jakości podczas wyświetlenia B. Dominuje udzielenie oceny do 20 punktów wyższej. Skrajny wzrost oceny jest widoczny dla obu wyświetleń, przy czym dla obrazu A jest niższy.

Tabela 3.31 Oceny uzyskane dla pary nr 12 kodowanej zgodnie ze standardem H.265

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
12A	2276	1,0	62,0	73,0	81,5	93,0	68,7	20,6
12B	6364	20,0	61,5	80,0	89,5	99,8	75,9	18,3
12A ≠ 12B	wzrost o 4088	−28,0	−1,5	5,0	12,0	25,4	7,2	19,6



(a) 12A (b) 12B (c) 12A ≠ 12B

Rysunek 3.32 Rozkład ocen dla pary 12 sekwencji kodowanej zgodnie ze standardem H.265

Para nr 13

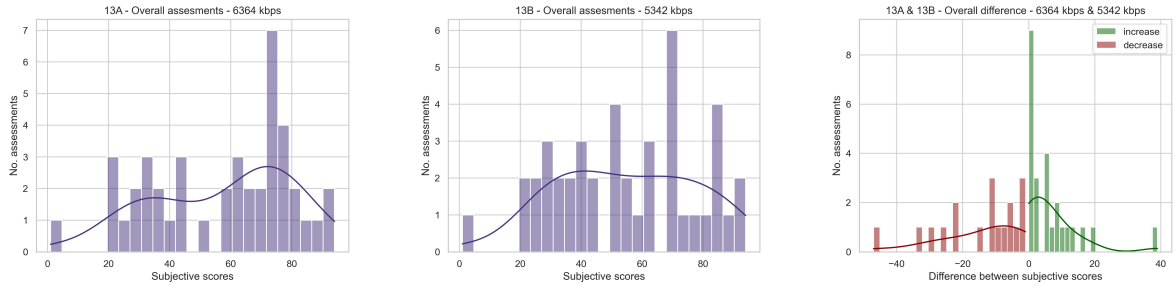
Treści o przepływnościach odpowiednio 6364 Kbit/s (A) oraz 5342 Kbit/s (B). Wyświetlenia A i B w tej parze ponownie mogą być traktowane jako wyświetlenia referencyjne, gdyż przepływności bitowe obu materiałów są jednymi z najwyższych w sekwencji badawczej. Mediana ocen dla obu materiałów to zaledwie 61 i 54 punkty, średnia osiągnęła wartości wynoszące odpowiednio 57 oraz 54,6. Oba materiały posiadają wartość minimalną na poziomie 1 punktu. Odchylenie standardowe jest wyższe niż obserwowane dla pozostałych materiałów i wynosi 23,1 oraz 22,8 punktu. Rozkłady ocen 3.33 (a) i (b) są zbliżone do siebie i obejmują ten sam zakres ocen. Znaczna liczba ocen zawiera się w zakresie od 20 do 45 punktów. Drugim takim zakresem są wartości od 60 do 90 punktów. Pomimo wysokich przepływności bitowych, które

63:3391053670

powinny gwarantować dobrą jakość uzyskano oceny niższe niż oceny pozostałych materiałów z takim samym strumieniem bitowym. Porównanie ocen widoczne na wykresie 3.33 ponownie pokazuje, że wartość dominująca wynosi zero. Wzrost oceny względem drugiego wyświetlenia jest widoczny dla obu obrazów. Przepływność bitowa materiałów jest zbliżona, więc różnice między ocenami w sąsiedztwie zera są oczekiwane. Niemniej jednak pojawiły się pojedyncze elementy odstające.

Tabela 3.32 Oceny uzyskane dla pary nr 13 kodowanej zgodnie ze standardem H.265

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
13A	6364	1,0	36,5	61,0	75,0	88,6	57,0	23,1
13B	5342	1,0	37,5	54,0	70,0	86,8	54,6	22,8
13A $\neq$ 13B	spadek o 1022	−47,0	−8,5	0,0	5,0	16,6	−2,3	14,9



(a) 13A (b) 13B (c) 13A $\neq$ 13B

Rysunek 3.33 Rozkład ocen dla pary 13 sekwencji kodowanej zgodnie ze standardem H.265

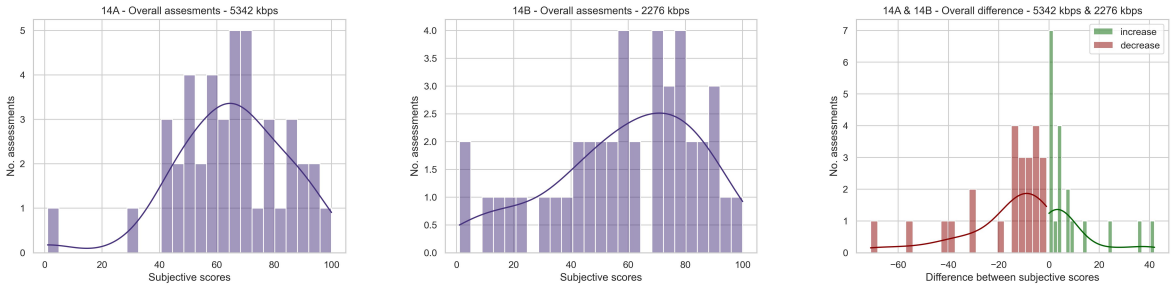
Para nr 14

Treści o przepływnościach odpowiednio 5342 Kbit/s (A) oraz 2276 Kbit/s (B). Mediana ocen dla materiału A to 66 punktów, średnia to 65,1 punktu. Treść B uzyskała medianę ocen równą 62 punkty i średnią na poziomie 59,1 punktu. Podobnie jak poprzednio oceny są nieco niższe. Oba materiały uzyskały ocenę minimalną o wartości 1. Odchylenie standardowe dla materiału B posiada najwyższą wartość ze wszystkich treści i wynosi 25,5 punktu. Rozkłady ocen dostępne na wykresie 3.34 (a) i (b) potwierdzają poziom odchylenia standardowego dla materiału B, gdyż oceny są rozrzucone na całej skali. Dla tego materiału uzyskano znaczną liczbę ocen poniżej 50 punktów, jednak dominującym przedziałem jest zakres od 70 do 80 punktów. Dla materiału A wartości są nieco bardziej skupione w zakresie od 60 do 70 punktów. Widoczny jest duży udział ocen z zakresu wartości od 50 do 60. Poniżej 40 punktów są tylko dwie oceny. Jedna z nich to wartość 1, natomiast druga to 30 punktów. Warto zauważyć, że jest to 3 para filmów z rzędu, w której pojawiła się wartość 1 jako ocena minimalna. Na wykresie 3.34 (c) wyróżnia się spadek oceny w zakresie do 20 punktów dla obrazu B. Mediana i średnia to odpowiednio spadek o 2 i 6 punktów. Wartości te mogą nie oddać w pełni różnicy pomiędzy strumieniami, ze względu na wartość dominującą wynoszącą zero. Pomimo ponad 2-krotnej różnicy w poziomie przepływności, część użytkowników udzieliła oceny na tym samym poziomie dla obu obrazów.

64:1147415111

Tabela 3.33 Oceny uzyskane dla pary nr 14 kodowanej zgodnie ze standardem H.265

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
14A	5342	1,0	52,5	66,0	79,0	92,9	65,1	19,3
14B	2276	1,0	43,0	62,0	78,5	90,9	59,1	25,5
14A $\neq$ 14B	spadek o 3066	−71,0	−12,0	−2,0	3,5	23,9	−6,0	20,5



(a) 14A (b) 14B (c) 14A $\neq$ 14B

Rysunek 3.34 Rozkład ocen dla pary 14 sekwencji kodowanej zgodnie ze standardem H.265

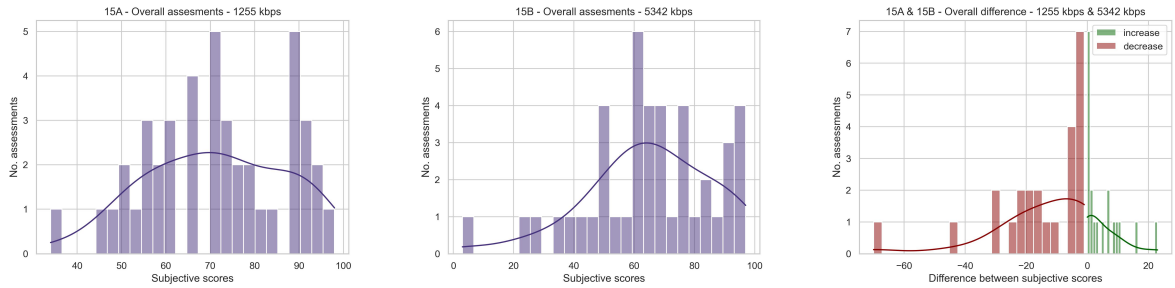
Para nr 15

Treści o przepływnościach odpowiednio 1255 Kbit/s (A) oraz 5342 Kbit/s (B). Dla ostatniej pary z tej sekwencji uzyskano medianę ocen wynoszącą 71 dla filmu A i 66 dla filmu B. Średnia ocen to odpowiednio 71,2 oraz 64,9. Jest to pierwsza para, w której film o przepływności niższej uzyskał lepszą ocenę. Przy czym porównane zostały treści o najniższym strumieniu oraz materiał o prawie najwyższej jakości bitowej. Materiał B uzyskał ocenę minimalną na poziomie zaledwie 3 punktów. Rozkłady ocen widoczne są na wykresach 3.35 (a) i (b). Dla obrazu A oceny w zakresie od 45 do 85 punktów wykazują rozkład normalny. Pojawia się jednak drugi dominujący obszar w zakresie od 90 do 100 punktów. Może to sugerować, że część uczestników badania założyło, że ten obraz jest referencyjny i przyznali wyższą ocenę. Dla obrazu B widać szeroki przedział unikalnych wartości. Wiele ocen zostało przydzielonych tylko raz. Dominującym zakresem wydaje się być przedział od 60 do 70, jednak wartości w sąsiedztwie 50, 75 i 100 punktów zostały kilka razy powtórzone. Na wykresie 3.35 (c) prezentującym rozkład różnic ocen dla tej pary widać, że zmienność ocen dominuje w sąsiedztwie zera. Uczestnicy udzieli niższej oceny dla obrazu o wyższej jakości, jednak duża część została obniżona o wartości bliskie zeru. Średni spadek oceny wynosił 6,3 punktu.

Tabela 3.34 Oceny uzyskane dla pary nr 15 kodowanej zgodnie ze standardem H.265

#	Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ
15A	1255	34,0	59,0	71,0	85,5	93,8	71,2	15,8
15B	5342	3,0	53,0	66,0	78,5	94,9	64,9	20,9
15A $\neq$ 15B	wzrost o 4087	−70,0	−15,0	−1,0	1,0	10,9	−6,3	16,5

65:6648300899



(a) 15A (b) 15B (c) 15A ≠ 15B

Rysunek 3.35 Rozkład ocen dla pary 15 sekwencji kodowanej zgodnie ze standardem H.265

3.9.3 Podsumowanie i wnioski

Dla sekwencji kodowanej H.264 zazwyczaj strumień o wyższej przepływności uzyskiwał lepszą ocenę. Uczestnicy badania byli w stanie wykryć kiedy są wyświetlane dwa obrazy referencyjne. Wiele razy uzyskiwano oceny z całego zakresu skali. Oceny minimalne o pojedynczych wartościach nie są rzadkością, podobnie jak oceny bardzo wysokie lub maksymalne. Ze statystycznego punktu widzenia część ocen to wartości odstające, które dla dalszego przetwarzania wymagają szczegółowej analizy. Na tym etapie badań podsumowano wszystkie oceny, ze względu na ich subiektywność. Materiały dla pierwszej sekwencji wydają się znacznie łatwiejsze w ocenie w porównaniu do drugiej sekwencji kodowanej algorytmem H.265. W tej sekwencji różnice pomiędzy materiałami wydają się być znacznie trudniej dostrzegalne. Wykrycie obrazu o niższej przepływności i słabszej jakości sprawiło znaczną trudność oceniającym. Nie byli w stanie jednoznacznie wskazać, który obraz ma niższą jakość lub czy dana para materiałów składa się z dwóch referencji. W rezultacie oceny dla strumieni drugiej sekwencji są bardzo do siebie zbliżone. Często są one wyższe dla obrazów o niższej przepływności. Może to wskazywać na próbę przypuszczania, że dany obraz ma niższą jakość i zakładania z góry, że jeden z obrazów pary jest zdegradowany. W sekwencji drugiej widoczny jest także brak zdecydowania, gdyż wiele ocen uzyskało wartość domyślną. Być może uczestnicy badania oczekiwali, że kolejne wyświetlenia będą miały jeszcze lepszą jakość i szacowali materiały filmowe zachowując rezerwę na skali ocen. Oceny domyślne mogą także sugerować zmęczenie danym badaniem. W tej sekwencji degradacja może nie być widoczna na pierwszy rzut oka. Zwiększona uwaga w trakcie badania mogła spowodować zmęczenie pod koniec badania objawiające się udzieleniem oceny domyślnej. Dla pary nr 13 w sekwencji kodowanej H.265 uzyskano niskie wartości ocen, odbiegające od innych materiałów. Być może źródłem tak niskiego osądu jest jakość materiału źródłowego, która okazała się niewystarczająca. Ostatecznie, dla sekwencji kodowanej H.264 uzyskano oceny o średniej arytmetycznej wynoszącej od 40 do 78,9 punktu. W pięciostopniowej skali ACR wspomniane wartości odpowiadały by punktacji wynoszącej od 2 punktów do 3.945 punktu. W skali przymiotnikowej można by określić poziom jakości tych obrazów jako odpowiednio słaba oraz dobra, przy założeniu, że zaokrąglimy punktację do liczb całkowitych. Dla sekwencji kodowanej H.265 uzyskano oceny o średniej arytmetycznej wynoszącej od 54,6 punktu do 82,5 punktu. Po zaokrągleniu, w skali ACR oceny te wyniosły by odpowiednio 3, czyli jakość średnia oraz 4, czyli jakość dobra.

66:8246485826

Tabela 3.35 Oceny uczestników badania pogrupowane dla zastosowanych poziomów jakości

Przepływność [Kbit/s]	MIN	Q1	Q2	Q3	95%	AVG	σ	ALG
800	1,0	30,0	50,0	70,0	90,0	50,8	24,3	H.264
1500	7,0	50,0	69,0	80,5	95,8	64,8	21,2	H.264
3000	11,0	58,0	70,0	81,0	95,8	68,3	18,7	H.264
6000	8,0	60,0	74,0	82,0	95,7	70,1	18,0	H.264
8000	1,0	50,0	71,0	81,0	94,0	65,8	21,5	H.264
10000	8,0	61,0	77,5	85,0	100,0	72,3	19,8	H.264
1255	18,0	58,0	72,0	81,0	94,0	68,8	18,0	H.265
2276	1,0	60,0	73,5	85,0	95,0	69,3	20,9	H.265
3298	20,0	59,8	74,5	85,3	95,5	71,3	19,0	H.265
5342	1,0	60,0	74,0	85,0	96,0	69,9	19,9	H.265
6364	1,0	60,0	74,0	85,0	96,0	71,1	19,4	H.265

Trzydzieści par materiałów wideo o różnej treści pozwoliło na zebranie sumarycznych wyników dla wszystkich przepływności bitowych stosowanych w strumieniach dla telewizji Hiway. Zebrane wyniki przedstawiono w tabeli 3.35. W tabeli zawarto zarówno wyniki dla kodowania treści algorytmem H.264 oraz H.265. Na tym etapie oceny kwalifikujące się jako elementy odstające nie zostały odfiltrowane, za wyjątkiem przypadków, o których wspomniano na początku tego podrozdziału. Dla strumienia o przepływności 800 Kbit/s uzyskano średnią ocenę na poziomie 50,8 punktu. 25% uczestników badania przyznało ocenę nie wyższą niż 30 punktów, podczas gdy 75% oceniających przyznało maksymalnie 70 punktów. Strumień o nieco wyższej przepływności wynoszące 1500 Kbit/s uzyskał średnio 64,8 punktu. Mediana wyniosła 69 punktów. 25% uczestników przyznało ocenę nie wyższą niż połowa skali, natomiast 75% oceniających udzieliło oceny maksymalnie wynoszącej 80,5 punktu. Dla kolejnego strumienia wynoszącego 3000 Kbit/s uzyskano średnią ocenę na poziomie 68,3 punktu. Mediana wyniosła 70 punktów. Dla I kwartyla uzyskano wartość oceny wynoszącą 58 punktów, natomiast dla III kwartyla wartość wyniosła 81 punktów. Dla strumienia 6000 Kbit/s średnia arytmetyczna ocen wyniosła 70,1 punktów, natomiast mediana 74 punkty. 25% najwyższych ocen to wartości wynoszące powyżej 82 punktów. Z drugiej strony dla I kwartyla wartości nie przekroczyły 60 punktów. Nieznaczny spadek wartości ocen nastąpił dla strumienia 8000 Kbit/s. Mediana wyniosła 71 punktów, natomiast średnia osiągnęła poziom zaledwie 65,8 punktu. Dla I kwartyla uzyskano maksymalną punktację niższą o 10 punktów w stosunku do poprzedniego strumienia. 25% najwyższych ocen jest nie mniejsza niż 81 punktów, więc poziom jest zbliżony do poprzednich strumieni. Strumień o najwyższej przepływności uzyskał medianę ocen wynoszącą 77,5 punktu. Średnia arytmetyczna wyniosła 72,3 punktu. Obie wartości są najwyższe w tym zestawieniu, włączając w to obrazy kodowane algorytmem H.265. Dla I kwartyla uzyskano ocenę wynoszącą 61 punktów, natomiast dla III kwartyla jest to 85 punktów. Strumień ten uzyskał oceny o wartości maksymalnej wynoszącej 100 punktów. Strumień o przepływności 1255 Kbit/s to najniższy z dostępnych dla treści kodowanych H.265. Średnia ocen to 68,8 punktu, mediana to 72 punkty. Dla I i III kwartyla uzyskano odpowiednio 58 oraz 81 punktów. Druga dostępna przepływność, czyli 2276 Kbit/s uzyskała nieco wyższe wartości ocen. Mediana ocen to 73,5 punktu, natomiast średnia

67:5955103384

to 69,3 punktu. Strumień 3298 Kbit/s uzyskał 74,5 punktów dla mediany oraz 71,3 dla średniej. Dla I kwartyla oceny są zbliżone dla wszystkich strumieni kodowanych H.265. Dla tego strumienia ocena wyniosła 59,8 punktu. Ocena dla III kwartyla także jest zbliżona dla strumieni H.265, za wyjątkiem tego o najniższej przepływności, dla którego ocena odbiega o 4 punkty. Ostatecznie, ocena dla strumienia 3298 Kbit/s w III kwartylu wyniosła 85,3 punktu. Dla strumienia 5342 Kbit/s średnia ocena wyniosła 69,9 punktu. Jest to nieznaczny spadek w stosunku do poprzedniego strumienia. Strumień o najwyższej przepływności uzyskał zbliżoną ocenę do strumienia 5342 Kbit/s. Mediana wyniosła 74 punkty, natomiast średnia 71,1 punktu. Dla pierwszego kwartyla ocena wyniosła 60 punktów, dla trzeciego było to 96 punktów. Wartości są takie same dla obu najwyższych strumieni.

Zebrane oceny zostały wykorzystane do wytrenowania modeli uczenia maszynowego, które są w stanie odtworzyć postrzeganie jakości przez uczestników badania. W ocenach widać nieznaczny wzrost wartości w stosunku do przepływności. Z jednej strony subiektywny charakter wartości ocen może być jedną z trudności związanych z predykcją poziomu QoE przez modele uczenia maszynowego. Praktycznie dla każdej przepływności występują oceny skrajne, a średnia i mediana są do siebie zbliżone. Z drugiej strony oceny mogą świadczyć o poziomie jakości doświadczeń użytkowników dekoderów i mogą nieść nieco bardziej rozbudowaną informację dla operatora niż surowe, obiektywne dane z czujników.

68:1819773715

Rozdział 4

Wykorzystanie metod uczenia maszynowego do analizy jakości wideo

Dane zebrane z badania oceniającego subiektywny poziom postrzegania jakości wideo zostały wykorzystane do modelowania oceny Quality of Experience dla telewizji Hiway. W niniejszym rozdziale opisano podejście do predykcji oceny QoE bazujące na modelach sztucznej sieci neuronowej.

4.1 Powiązane badania

Badanie i szacowanie poziomu jakości wideo jest często poruszaną tematyką badawczą. Jednym z ważniejszych algorytmów, które powstały do tego celu jest *VMAF* (ang. *Video Multimethod Assessment Fusion*) zaprezentowany przez firmę Netflix [62]. Algorytm ten wykorzystuje architekturę maszyny wektorów nośnych *SVM* (ang. *Support Vector Machine*) [77] do predykcji ocen w skali DMOS. Model wykonuje predykcje bazując między innymi na metrykach *VIF* (ang. *Visual Information Fidelity*) [78] określającej utratę wierności obrazu oraz *DLM* (ang. *Detail Loss Metric*) [79] wskazującą ilość utraconych szczegółów obrazu. Do treningu algorytmu wykorzystano oceny w skali DCR zebrane od uczestników badania przeprowadzonego w protokole DSIS. Badanie to pokazało, że wykorzystanie metod uczenia maszynowego może pozwolić na uzyskanie bardziej celnych predykcji subiektywnych odczuć użytkownika niż zastosowanie klasycznych metod oceny jakości wideo takich jak PSNR, czy też *SSIM* (ang. *Structural Similarity Index Measure*) [61]. *VQM_VFD* (ang. *Video Quality Model for Variable Frame Delay*) [80] opracowany przez amerykańską agencję *NTIA* (ang. *National Telecommunications and Information Administration*) [81] jest innym ważnym modelem oceny jakości wideo. Model ten wykorzystuje sieć neuronową do predykcji subiektywnej jakości wideo bazując na ośmiu obiektywnych i mierzalnych parametrach sygnału wideo. Metoda wykorzystuje odwołanie do obrazu referencyjnego (ang. *Full Reference*) i porównuje parametry klatek tego sygnału z przesuniętymi i zniekształconymi klatkami testowanego materiału. Do trenowania modelu wykorzystano oceny DMOS dla ponad 11000 materiałów wideo uzyskując dla zbioru testowego współczynnik korelacji Pearsona na poziomie 0,9. Model oceniający w skali ACR bazujący na maszynie wektorów nośnych SVM został zaprezentowany w badaniu [82]. Algorytm ten szacuje jakość materiałów znajdujących się na platformie YouTube [83]. Cechy wejściowe dla modelu to głównie

69:8659837041

liczba zacięć (ang. *stalls*), długość trwania materiału, a także całkowity czas zacięć. Autorzy pokazali w jaki sposób korzystając z platform umożliwiających crowdsourcing można zebrać subiektywne oceny od użytkowników Internetu. W pracy [84] autorzy zaproponowali metrykę *VsQM* (ang. *Video streaming Quality Metrics*), która uwzględnia wpływ przerw w dostępie do treści wideo oraz czasu ich trwania w stosunku do czasu trwania segmentów czasowych materiałów wideo. Założeniem tego badania było precyzyjniejsze oszacowanie oceny MOS dla materiałów wideo, których odtwarzanie zostało przerwane jako skutek problemów transmisyjnych. W badaniu [85] przedstawiono predykcyjny model *SQI* (ang. *Streaming QoE Index*). Ocenia on jakość materiału wideo uwzględniając obiektywną jakość obrazu, występowanie zacieć oraz efektu pamięci dla kolejnych zacięć występujących w krótkim czasie. Ocena estymowana przez SQI znacznie spada dla segmentu wideo, w którym pojawiła się pewna liczba zacięć w krótkim odstępie czasu. Wraz z poprawą transmisji i ustąpieniem negatywnego efektu przerwy w odtwarzaniu obrazu, indeks ten wraca do początkowej wyższej oceny. *Video ATLAS* (ang. *Video Assessment of TemporaL Artifacts and Stalls*) [86] to model bazujący na algorytmach *SVR* (ang. *Support Vector Regression*), regresji grzbietowej (ang. *Ridge regression*) i Lasso. Jest on w stanie osiągać lepsze predykcje niż wspomniany wcześniej indeks SQI. Model ten uwzględnia obiektywne cechy jakościowe segmentu wideo, występowanie zacięć tj. efekt buforowania, efektu pamięci związanego z postępującą degradacji oraz parametrów czasowych określających czas występowania degradacji oraz poprawnie odtwarzanego materiału wideo. Predykcyjna analiza parametrów jakościowych transmisji może być też wykorzystywana do poprawy algorytmów adaptacji strumienia wideo HTTP. W ramach badania [87] autorzy przedstawili algorytm kontroli adaptacji MPC (ang. model predictive control) sterujący stanem bufora odtwarzacza w taki sposób, aby maksymalizować poziom jakości QoE. Założono, że na postrzeganą jakość najbardziej wpływa średni poziom przepływności wideo oraz jej wariancja, a także zacięcia i opóźnienie w rozpoczęciu odtwarzania. Algorytm *QDASH* (ang. *QoE-aware DASH system*) zaproponowany w [88] także dostosowuje poziom przepływności wideo do panujących warunków transmisyjnych. W tym badaniu autorzy zwrócili uwagę, że w celu zwiększenia poziomu QoE należy zapewnić dostęp do pośrednich strumieni wideo. Brak dostępu do takich strumieni HTTP powoduje przerwę w transmisji, która jest subiektywnie mniej akceptowalna niż przejście do odtwarzania pośredniego profilu wideo. W badaniu [89] przedstawiono model regresji liniowej, który na podstawie metryk jakości obrazu obliczonych przez inne algorytmy był w stanie estymować subiektywną ocenę zebraną od uczestników badania w formule DSIS. Sprawdzono także skuteczność regresji SVR, drzew decyzyjnych, lasów losowych, wzmacniania gradientowego *GBM* (ang. *Gradient Boost Machine*) oraz metody najbliższych sąsiadów *kNN* (ang. *k-Nearest Neighbours*). Badanie [90] wykorzystuje metryki takie jak opóźnienie odtwarzania, liczba zacięć oraz średnia jakość wideo wraz z liczbą zmian profili wideo do stworzenia architektury sieci definiowanej programowo *SDN* (ang. *Software Defined Network*). Architektura SDNDASH zarządza alokacją zasobów oraz optymalizuje sieć pod kątem jakości QoE uwzględniając parametry takie jak zajętość bufora, dostępna przepustowość, ograniczenie typu treści czy też możliwości urządzeń odtwarzających. W pracy [91] autorzy wykorzystali zespół modeli Hammerstein-Wienera w celu modelowania percepcji jakości z efektem pamięci. Badanie to uwzględniało spadek postrzeganej jakości po wystąpieniu wielu zakłóceń w krótkim odstępie czasu w trakcie odtwarzania treści wideo. Reprezentacja pamięci oraz parametry materiałów wideo zostały wykorzystane do wytrenowania modelu regresji SVR szacują-

70:3589734066

cego ocenę QoE. W publikacji [92] autorzy zaprezentowali parametryczny model obiektywnej oceny QoE bazujący na opisie modelu ludzkiego układu wzrokowego *HVS* (ang. *Human Visual System*). Model wykorzystujący kombinacje konwolucyjnej sieci neuronowej *CNN* (ang. *Convolutional Neural Network*) i *GRU* (ang. *Gate Recurrent Unit*) zaprezentowano w badaniu [93]. Model konfirmacyjny *ECT* (ang. *Expectation confirmation theory*) [94] [95] został wykorzystany w badaniu [96]. Autorzy zauważyli, że osoba oglądająca materiał wideo może postrzegać poziom jakości w zależności od założonych oczekiwań. Wraz z upływem czasu oczekiwany poziom jakości może spadać, przez co subiektywna satysfakcja z jakości oglądanej treści stabilizuje się lub nawet rośnie. Rekurencyjna sieć neuronowa *LSTM* (ang. *Long Short-Term Memory*) została wykorzystana w modelu LSTM-QoE do predykcji jakości QoE w sposób ciągły [97]. Autorzy wykorzystali parametr *STSQ* (ang. *Short Time Subjective Quality*) mierzący jakość wideo jako podobieństwo obrazu zniekształconego do oryginalnego. Przykładowym algorytmem wykorzystanym do pomiaru metryki STSQ jest *MS-SSIM* (ang. *Multi-Scale Structural Similarity Index*) [63]. Dodatkowo wykorzystano współczynnik określający stan odtwarzacza (zacięcie lub odtwarzanie) oraz czas od wystąpienia ostatniego zacięcia. Metrykę STSQ wykorzystano również w badaniach [98], [99], [100]. Głębokie uczenie zostało także wykorzystane w badaniu [101], w którym autorzy wykorzystali osadzanie słów (ang. *word embedding*) oraz trójwymiarową sieć konwolucyjną C3D do ekstrakcji cech materiałów wideo. Następnie cechy te wykorzystano w sieci neuronowej do klasyfikacji i regresji poziomu jakości. Jako cechy wejściowe wykorzystano czynniki oddziałujące *IFs* (ang. *Influence factors*). Czynniki te można sklasyfikować jako systemowe (np. PSNR), kontekstowe (np. miejsce oglądania treści) oraz ludzkie (np. zainteresowanie daną treścią). W badaniu [102] autorzy przygotowali wielopoziomowy system łączący uczenie nienadzorowane wykorzystujące model ograniczonej Maszyny Boltzmana RBM (ang. Restricted Boltzmann Machine) oraz klasyfikację. Celem było efektywniejsze zarządzanie dostępem do treści wideo we współdzielonym paśmie i zapewnienie odpowiedniego poziomu jakości QoE. Analiza QoE dla adaptacyjnych treści wideo jest także przedmiotem badań w sieciach mobilnych. W badaniu [103] autorzy wykorzystali parametry QoS zebrane z urządzeń serwerowych, aby szacować jakość QoE dla transmisji w sieciach mobilnych w czasie rzeczywistym. W podejściu [104] autorzy wykorzystali uogólniony model liniowy *GLM* (ang. *Generalized Linear Model*) oraz algorytm maszyny wektorów nośnych SVM. Celem było zbadanie jakości QoE w kontekście zapewnienia odpowiedniej przepustowości i zapewnienia dostępu do zasobów w trakcie strumieniowania treści wideo. W badaniu [105] wykorzystano dane od tureckiego operatora sieci mobilnej aby zmapować parametry sieci bezprzewodowej na parametry QoE dla treści wideo. Celem było wykorzystanie modeli głębokiego uczenia do predykcji strat propagacji i szacowania oceny MOS. Szacowanie jakości QoE jest także przedmiotem badań w sieciach 5G. W badaniu [106] autorzy zasymulowali analizę danych pochodzących z komponentu Funkcji Analityki Danych Sieciowych *NWDAF* (ang. *Network Data Analytics Function*). Celem badania było sprawdzenie przydatności parametrów dostępnych w NWDAF w celu szacowania QoE poprzez model regresji liniowej i SVR. W pracy [107] wykorzystano czas podróży w obie strony *RTT* (ang. *Round-Trip Time*), liczbę pakietów oraz przepływność w celu predykcji QoE w sieciach 5G. Sprawdzono w jaki sposób algorytmy regresji drzewa decyzyjnego *DTR* (ang. *Decision Tree Regression*), regresji wielokrotnej *MLR* (ang. *Multi-Linear Regression*) oraz regresji lasu losowego *RFR* (ang. *Random Forest Regression*) radzą sobie w przypadku odtwarzania

71:6183246306

treści w miejscu oraz w trakcie poruszania się. Algorytm DeSVQ wykorzystujący sieć LSTM został zaprezentowany w badaniu [108]. W pierwszej fazie cechy materiału wideo są wyznaczane przez obiektywne algorytmy *VQA* (ang. *Video Quality Assessment*). W kolejnej fazie wartości te trafiają na wejście sieci LSTM, która wykonuje predykcję QoE. Na wyjściu wykorzystano algorytm *XGBoost* (ang. *eXtreme Gradient Boosting*)[109], którego celem jest wzmocnienie skuteczności predykcji. W badaniu [110] autorzy analizowali postrzeganą jakość dla filmów odtwarzanych w aplikacji YouTube [83] na urządzeniach mobilnych. Do analizy wykorzystano 275 cech związanych m.in. z odbieranym poziomem sygnału, liczbą przepięć pomiędzy stacjami bazowymi, statystykami związanymi z ruchem sieciowym, liczbą zacięć w trakcie oglądania treści wideo, czasie opóźnienia w rozpoczęciu oglądania oraz poziomem strumienia i liczbą zmian strumienia. Do predykcji QoE wykorzystano algorytm lasu losowego z dziesięcioma drzewami decyzyjnymi. Architektura bazująca na uczeniu przez wzmacnianie oraz sieci LSTM została przedstawiona w badaniu [111]. Autorzy zaprezentowali podejście serwer - klient, w którym serwer steruje przydziałem odpowiedniego pasma bazując na algorytmie predykcji jakości QoE. Serwer dynamicznie modyfikuje pliki manifestów *MPD* (ang. *Media Presentation Description*) [112] ograniczając dostępne profile dla odtwarzacza. W rezultacie dostępne pasmo jest sprawiedliwie dzielone wśród wszystkich klientów. W badaniu [113] autorzy przedstawili metodę wykrywania zacięć, średniej przepływności oraz zmian w przepływności dla zaszyfrowanego ruchu HTTP w sieci komórkowej. Porównując rozkłady ruchu szyfrowanego z nieszyfrowanym stworzono narzędzie do oceny jakości QoE. Algorytm lasu losowego oraz cechy takie jak m.in. RTT, *BDP* (ang. *Bandwidth-delay product*), średnia i maksymalna przepływność, procentowa ilość utraconych i retransmitowanych pakietów oraz rozmiar i czas dla fragmentu strumieniowania adaptacyjnego zostały wykorzystane do wytrenowania modelu oceniającego QoE. Autorzy badania [114] zamiast oceniać poziom jakości QoE zaproponowali algorytm, który oszacuje metryki wpływające na QoE. Wspomniane parametry, podobnie jak dla poprzedniego badania są estymowane dla szyfrowanego ruchu pochodzącego z aplikacji Youtube[83]. Predykcja stanu zapełnienia bufora odtwarzacza, jego tendencji oraz rozdzielczości obrazu została przeprowadzona przez algorytm bazujący na lasach losowych. Do treningu wykorzystano dane związane z przepływem pakietów w trakcie odtwarzania zaszyfrowanego materiału wideo.

Ze względu na szeroki obszar zastosowań analiza jakości QoE dla treści wideo jest aktywną tematyką badań. Powiązane badania przedstawiają różne podejścia do analizy jakości doświadczeń. Proponowane algorytmy przedstawiają często wysoki stopień zaawansowania i skomplikowania. Może to powodować trudność w zastosowaniu w środowisku operatorskim, w którym dostępność różnorodnych danych często przekracza środowiska laboratoryjne. W takim przypadku niejednokrotnie należy się liczyć z ograniczeniami technicznymi nałożonymi przez postanowienia komercyjne. W tym rozdziale przedstawiono model sieci neuronowej, którego celem jest wskazanie szacunkowego poziomu jakości doświadczeń QoE. Model ten bazuje na danych zebranych w trakcie badania w formule DSCQS opisanego w poprzednim rozdziale.

72:2244796405

4.2 Nadzorowane uczenie maszynowe

Przewidywanie subiektywnego poziomu jakości wideo prezentowanego za pośrednictwem dekodera podłączonego do systemu telewizji Hiway zostało sprowadzone do problemu, który rozwiązano z wykorzystaniem metod uczenia maszynowego. Dane zebrane podczas badań, opisanych w poprzednim rozdziale, posłużyły do trenowania sztucznej sieci neuronowej. Subiektywny charakter ocen sprawia, że zależności pomiędzy niektórymi zmiennymi nie zawsze mogą być jednoznacznie określone. Przykładowo, dwóch obserwatorów o podobnym wieku i zainteresowaniach może odmiennie ocenić tę samą treść. Materiały wyświetlane na ekranach telewizorów różnych producentów, choć o identycznych rozmiarach, ale odmiennych ustawieniach wyświetlania, również mogą być różnie postrzegane. Można zatem stwierdzić, że każdy użytkownik usługi funkcjonuje w indywidualnym środowisku, w którym konsumuje treści wideo. Wykorzystanie sieci neuronowej do predykcji postrzeganej jakości umożliwiło uśrednienie ocen użytkowników, co pozwoliło na oszacowanie poziomu jakości odtwarzanego wideo na danym dekoderze.

4.2.1 Sztuczna sieć neuronowa

Szacowanie poziomu jakości treści odtwarzanej za pośrednictwiem dekodera zostało ujęte jako problem klasyfikacji wieloklasowej bazującej na architekturze perceptronu wielowarstwowego *MLP* (ang. *Multi-layer Perceptron*). Wyjście pojedynczego neuronu można opisać jako 4.1 [115]:

$$y = \psi\left(\sum_{i=1}^{n} \theta_{[i]} x_{[i]} + \theta_{[n+1]}\right), \tag{4.1}$$

gdzie:

y – wyjście neuronu

$\psi(s)$ – funkcja aktywacji (np. ReLU, sigmoid, tanh), $\psi : \mathbb{R} \to \mathbb{R}$,

$x_{[i]}$ – wektor wejść i-tego neuronu (ang. *neuron inputs*), $x \in \mathbb{R}^n$

$\theta_{[i]}$ – wektor wag i-tego neuronu (ang. *neural weights*), $\theta \in \mathbb{R}^{n+1}$, waga $\theta_{[n+1]}$ określona jest jako obciążenie (ang. *bias*) neuronu.

Neuron jest liniowy jeżeli $\psi(s) = s$; sigmoidalny jeżeli ψ jest funkcją sigmoidalną; lub typu *ReLU* (ang. *Rectified Linear Unit*), gdy $\psi(s) = \max\{0, s\}$. Sieć neuronowa składa się z warstw neuronów, w których wyjścia jednych neuronów mogą stanowić wejście dla kolejnych. Wejścia neuronów pochodzą z zewnątrz sieci, albo jako wyjścia innych neuronów. Wyjście dla ostatniej warstwy może stanowić wynik całej sieci [115]. Perceptron wielowarstwowy to struktura sieci neuronowej, składająca się z warstwy wejściowej (ang. *input layer*), jednej lub więcej warstw ukrytych (ang. *hidden layers*) oraz warstwy wyjściowej (ang. *output layer*). Neurony w perceptronie wielowarstwowym w warstwach innych niż ostatnia zwracają wartość 4.2 [115]:

$$y^k_{[j]} = \psi\left(\theta^k_{[j]}\, y^{k-1}\right) \tag{4.2}$$

gdzie:

73:2807523410

$y^k_{[j]}$ – wyjście j-tego neuronu w k-tej warstwie,

ψ – funkcja aktywacji (np. ReLU, sigmoid, tanh),

$\theta^k_{[j]}$ – wektor wag przypisanych do j-tego neuronu w k-tej warstwie,

y^{k-1} – wektor wyjść z poprzedniej warstwy.

Neurony w ostatniej warstwie perceptronu wielowarstwowego zwracają wartość 4.3 [115]:

$$y^K_{[j]} = \theta^K_{[j]}\, y^{K-1} \tag{4.3}$$

gdzie:

$y^K_{[j]}$ – wyjście j-tego neuronu w K-tej warstwie wyjściowej,

$\theta^K_{[j]}$ – wektor wag j-tego neuronu w K-tej warstwie wyjściowej,

y^{K-1} – wektor wyjść z poprzedniej warstwy.

4.2.2 Funkcja aktywacji

Funkcja aktywacji określa wartość wyjścia pojedynczego neuronu. Do najczęściej stosowanych funkcji aktywacji należą: funkcja liniowa, logistyczna, sigmoidalna, oraz tangens hiperboliczny.

Funkcja ReLU

ReLU (ang. *Rectified Linear Unit*) to funkcja zwracająca wartość 0 dla ujemnych argumentów oraz wartość argumentu dla argumentów dodatnich. Neurony aktywowane przez ReLU pozwalają na przyśpieszenie obliczeń oraz szybsze uzyskanie zbieżności w trakcie uczenia sieci. Wynika to z faktu, że wartość pochodnej przyjmuje wartość 1 dla argumentów dodatnich. W porównaniu do funkcji logistycznej lub sigmoidalnej nie ma potrzeby obliczania wartości funkcji eksponencjalnej. Wadą funkcji ReLU może być tzw. zjawisko wymierania neuronów (ang. *dead ReLU*) ze względu na zerową wartość pochodnej dla argumentów ujemnych. Neurony objętę tym problemem zwracają wartość wyjściową równą zero, niezależnie od wartości na wejściu. Może to prowadzić do zaniku ich aktywności w trakcie uczenia. W przypadku wystąpienia takiego problemu rozwiązaniem może być zastosowanie funkcji Leaky ReLU (LReLU) lub Parametric ReLU (PReLU), które modyfikują zachowanie funkcji w zakresie ujemnych argumentów [115].

Dywergencja Kullbacka-Leiblera

W trakcie uczenia modeli miękkiej klasyfikacji (ang. *soft classification*) przyjęto Dywergencję Kullbacka-Leiblera jako funkcję straty. Opisana wzorem 4.4 funkcja mierzy rozbieżność pomiędzy przewidzianym przez model rozkładem prawdopodobieństwa Q, a prawdziwym rozkładem P [116]. Dla klasyfikacji ocen w skali ACR rozkłady prawdopodobieństwa to pięcioelementowe wektory dla każdej próbki.

74:9017450144

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \tag{4.4}$$

gdzie:

D_{KL} – Dywergencja Kullbacka-Leiblera pomiędzy rozkładami Q i P,

x – zdarzenia od 1 do 5 reprezentujące klasy ocen w skali ACR,

$P(x)$ – prawdopodobieństwo oceny x w rzeczywistym rozkładzie P,

$Q(x)$ – prawdopodobieństwo oceny x w przewidywanym rozkładzie Q.

4.2.3 Funkcja softmax

Funkcja softmax 4.5 skaluje wektor wyjściowy danej warstwy do zakresu [0, 1], w taki sposób aby suma składowych równała sie wartości 1. Skalowanie w ostatniej warstwie może być interpretowane jako rozkład prawdopodobieństwa dla problemu klasyfikacji wieloklasowej. W takim przypadku wektor wyjściowy określa szansę przynależności do danej klasy [115].

$$\mathrm{softmax}(z)_{[j]} = \frac{e^{z_{[j]}}}{\sum_{k=1}^{n} e^{z_{[k]}}} \tag{4.5}$$

gdzie:

$\mathrm{softmax}(z)_{[j]}$ – wartość funkcji softmax dla klasy j, interpretowana jako prawdopodobieństwo przynależności do tej klasy,

$z_{[j]}$ – j-ty element nieznormalizowanego wyjściowego wektora z,

n – liczba elementów wektora z, odpowiadająca liczbie klas.

4.3 Zbiór cech wejściowych

Dane zebrane w trakcie badania dotyczącego postrzegania poziomu jakości wideo odtwarzanego za pośrednictwem dekodera telewizji Hiway zostały wykorzystane do uczenia sieci neuronowej bazującej na architekturze perceptronu wielowarstwowego. Przygotowanie zbioru danych (ang. dataset) obejmowało między innymi filtrowanie elementów odstających, tworzenie nowych cech, kodowanie cech kategorycznych oraz standaryzację.

4.3.1 Elementy odstające

Elementy odstające dla ocen w ramach danej przepływności zostały zidentyfikowane z wykorzystaniem rozstępu międzykwartylowego IQR (ang. *Interquartile Range*) określonej jako 4.6 [117].

$$IQR = Q_3 - Q_1 \tag{4.6}$$

gdzie:

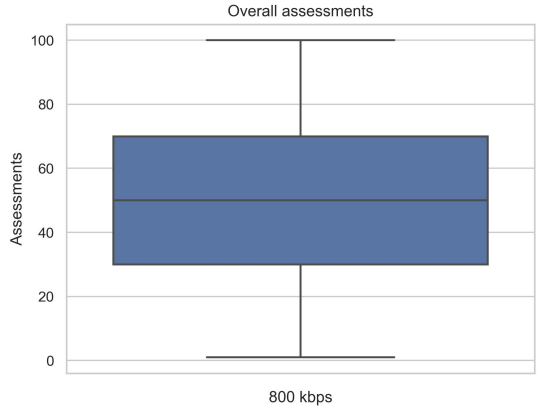
75:4595713989

Q_3 – trzeci kwartyl,

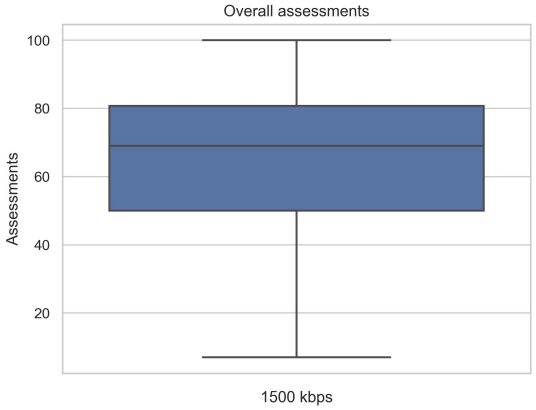
Q_1 – pierwszy kwartyl.

Wartość jest określana jako odstająca jeżeli przekracza dolny przedział równy $Q_1 - 1,5 \times IQR$ lub górny wynoszący $Q_3 + 1.5 \times IQR$. Wartości, które przekraczały dolny lub górny przedział zostały zastąpione średnią oceną dla każdej przepływności. Na wykresach pudełkowych (ang. *boxplot*) przedstawiono rozkład ocen dla każdej przepływności. Dane pochodzą z badania opisanego we wcześniejszym rozdziale. Dolną i górną granicę dla elementów odstających zaznaczono wąsami (ang. *whiskers*). Na rysunku 4.1 przedstawiono rozkłady dla ocen materiałów kodowanych H.264. Poniżej dolnej oraz powyżej górnej granicy nie znajduje się żaden element. Podobnie jest dla rozkładów ocen treści kodowanych H.265 widocznych na rysunku 4.2. Dane dla każdego strumienia nie zawierają wartości odstających.

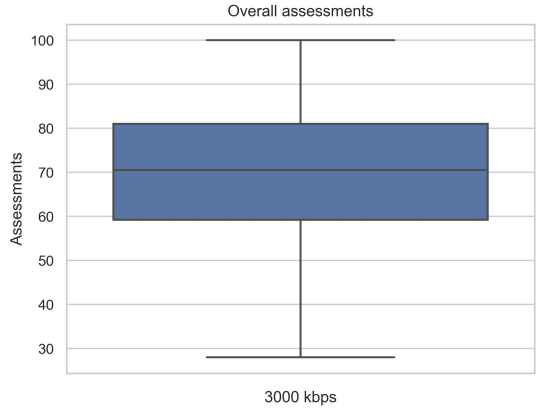
76:6044749039



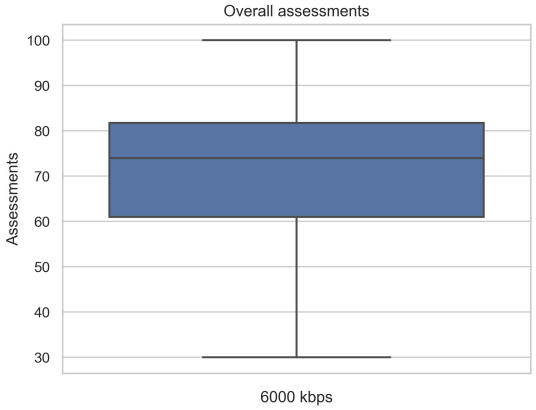
(a) 800 Kbps



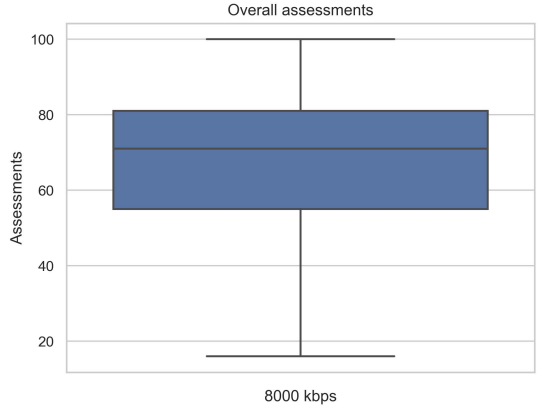
(b) 1500 Kbps



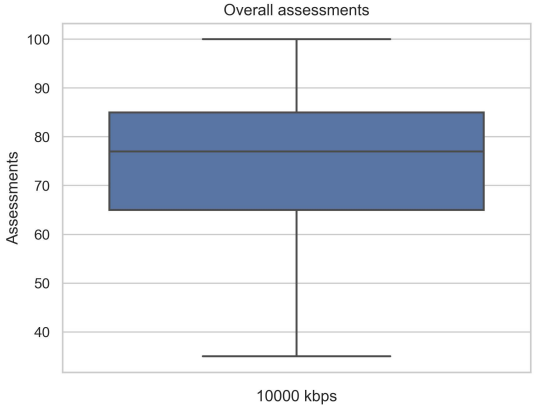
(c) 3000 Kbps



(d) 6000 Kbps



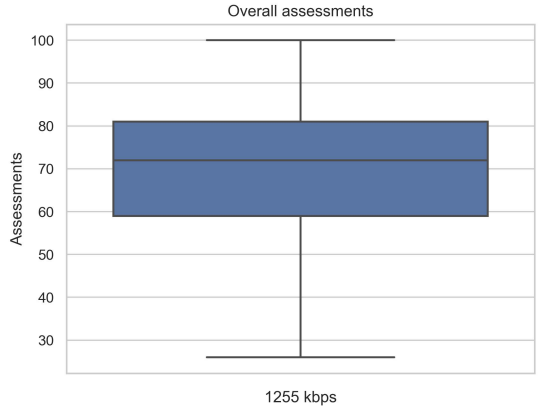
(e) 8000 Kbps



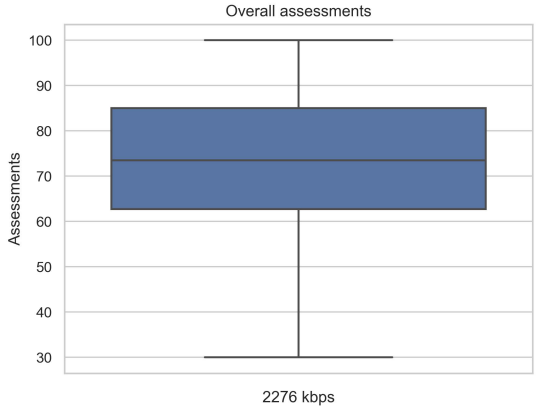
(f) 10000 Kbps

Rysunek 4.1 Rozkład ocen dla strumieni kodowanych zgodnie ze standardem H.264

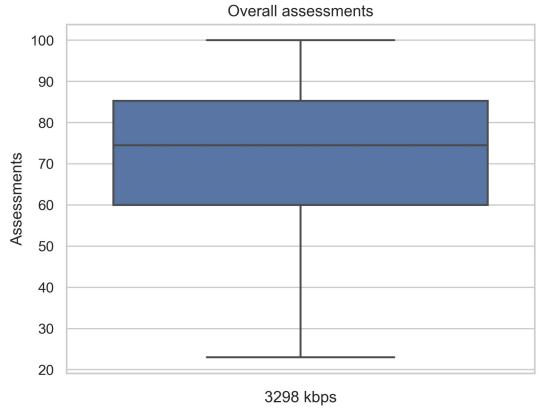
77:2455368416



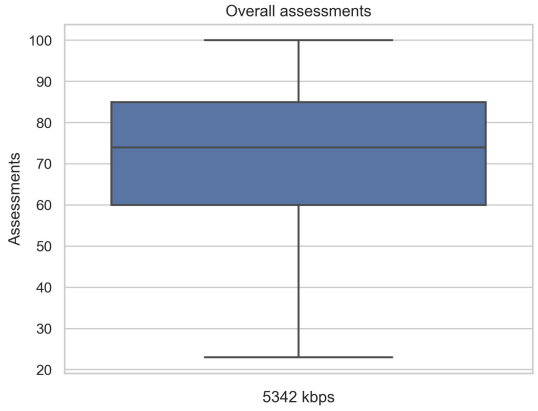
(a) 1255 Kbps



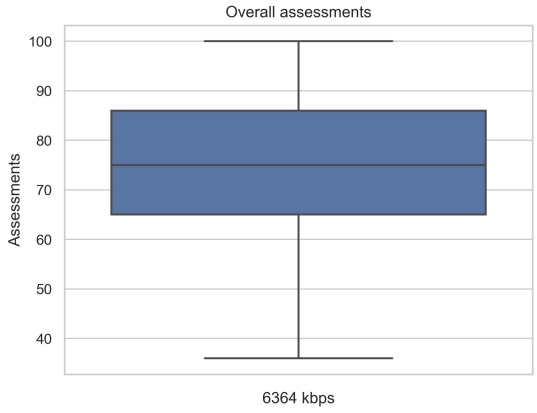
(b) 2276 Kbps



(c) 3298 Kbps



(d) 5342 Kbps



(e) 6364 Kbps

Rysunek 4.2 Rozkład ocen dla strumieni kodowanych zgodnie ze standardem H.265

4.3.2 Nowe cechy

Do zbioru danych dodano nowe cechy 4.7 oraz 4.8 bazujące na aktualnie istniejących:

78:2407458412

$$current_assessment_class = \min\left(\left\lfloor \frac{current_assessment}{20} \right\rfloor,\ 4\right) \tag{4.7}$$

gdzie:

current_assessment_class – obliczona ocena w pięciostopniowej skali ACR dla aktualnie odgrywanego fragmentu wideo,

current_assessment – aktualna ocena w skali od 1 do 100.

$$previous_assessment_class = \min\left(\left\lfloor \frac{previous_assessment}{20} \right\rfloor,\ 4\right) \tag{4.8}$$

gdzie:

previous_assessment_class – obliczona ocena w pieciostopniowej skali ACR dla poprzednio odgrywanego fragmentu wideo,

current_assessment – ocena dla poprzedniego fragmentu wideo w skali od 1 do 100

Oceny obliczone w 4.7 oraz 4.8 to oceny w skali ACR, jednak dla ułatwienia dalszych obliczeń pozostawiono zakres od 0 do 4, gdzie 0 jest odpowiednikiem najniższej oceny.

Dla uzyskanego podzbioru danych wejściowych 4.9:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \tag{4.9}$$

gdzie:

$x_i \in \mathbb{R}^n$ – wektor cech i-tej próbki. Wektor ten zawiera cechy: *current_bitrate*, *previous_bitrate*, *tv_screen_size* oraz *age_range*,

$y_i \in \{0, 1, 2, 3, 4\}$ – etykieta klasy przypisana próbce,

N – liczba próbek w zbiorze.

wyznaczono rozkład prawdopodobieństwa $\mathbf{p}$ dla etykiet y_i w zbiorze $\mathcal{D}$ 4.10:

$$\mathbf{p} = [p_0,\ p_1,\ \ldots,\ p_{K-1}] \tag{4.10}$$

gdzie prawdopodobieństwo określonej oceny w skali ACR to 4.11:

$$p_k = \frac{1}{N}\sum_{i=1}^{N} 1(y_i = k) \tag{4.11}$$

liczba wszystkich próbek w zbiorze jest określona jako 4.12:

$$N = \sum_{k=0}^{K-1} n_k \tag{4.12}$$

Ostatecznie uzyskano etykiety jako wektor prawdopodobieństw 4.13:

79:9358160529

$$\mathbf{y}_i = [p_0,\ p_1,\ p_2,\ p_3,\ p_4] \tag{4.13}$$

Dla każdego wektora x_i określone zostało prawdopodobieństwo uzyskania wskazanej oceny w skali ACR. Pojedyncza próbka w zbiorze reprezentuje teraz 60-sekundową sekwencję, która składa się z dwóch materiałów wideo, czyli treści A i B. Finalnie do dalszego przetwarzania wykorzystano zestaw cech:

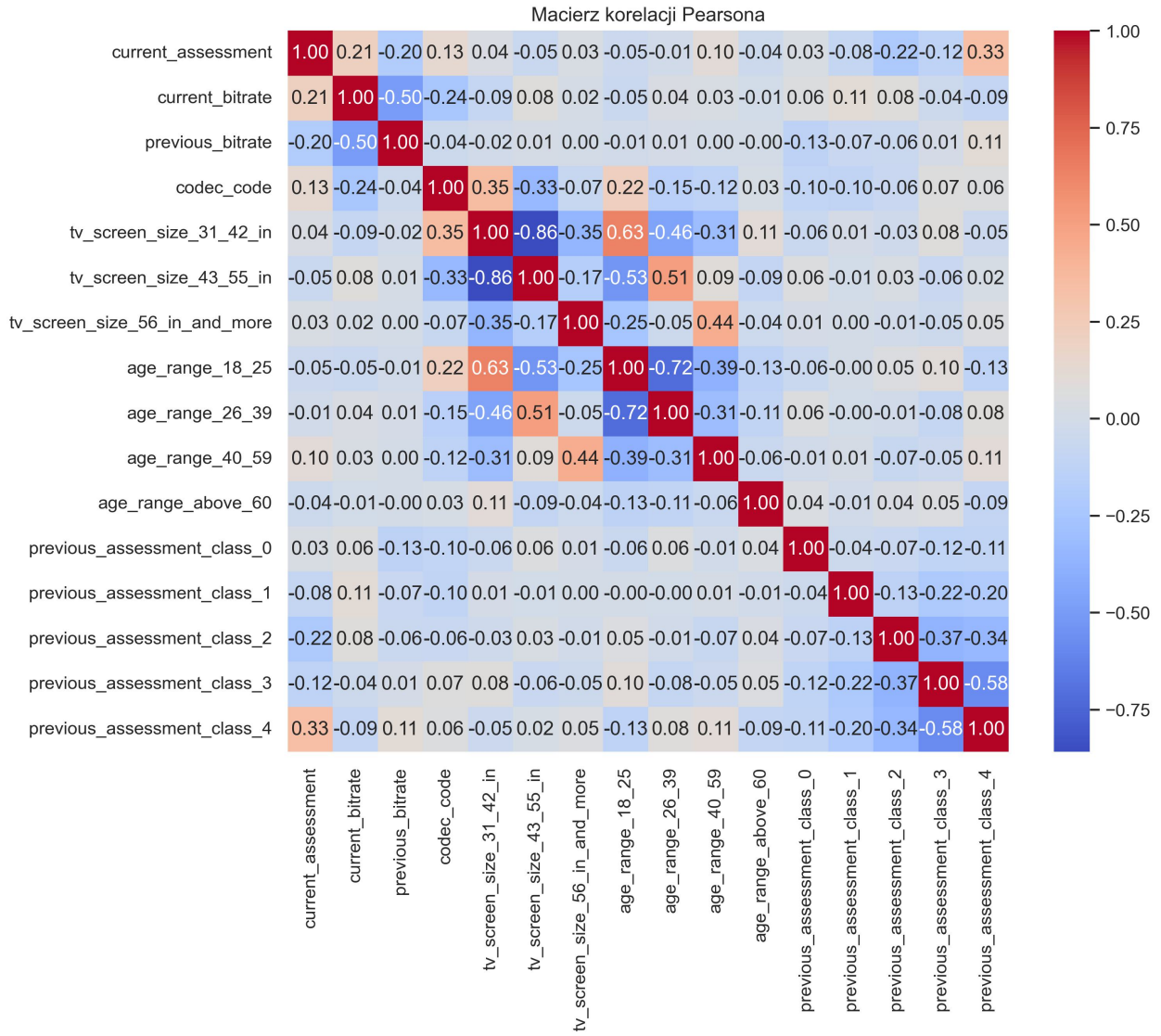
Tabela 4.1 Zestaw cech wykorzystany do predykcji

#	Nazwa cechy	Opis
1	current_bitrate	poziom przepływności w Kbit/s dla aktualnej próbki
2	previous_bitrate	poziom przepływności w Kbit/s dla poprzedniej próbki
3	codec_code	znacznik zastosowanego algorytmu kodowania wideo
4	tv_screen_size_31_41_in	zakres rozmiaru ekranu TV od 31 do 42 cali
5	tv_screen_size_43_55_in	zakres rozmiaru ekranu TV od 43 do 55 cali
6	tv_screen_size_56_in_and_more	zakres rozmiaru ekranu TV powyżej 56 cali
7	age_range_18_25	zakres wieku od 18 do 25 lat
8	age_range_26_39	zakres wieku od 26 do 39 lat
9	age_range_40_59	zakres wieku od 40 do 59 lat
10	age_range_above_60	zakres wieku powyżej 60 lat
11	previous_assessment_class_0	ocena równa 0 dla poprzedniej próbki
12	previous_assessment_class_1	ocena równa 1 dla poprzedniej próbki
13	previous_assessment_class_2	ocena równa 2 dla poprzedniej próbki
14	previous_assessment_class_3	ocena równa 3 dla poprzedniej próbki
15	previous_assessment_class_4	ocena równa 4 dla poprzedniej próbki

4.3.3 Macierz korelacji

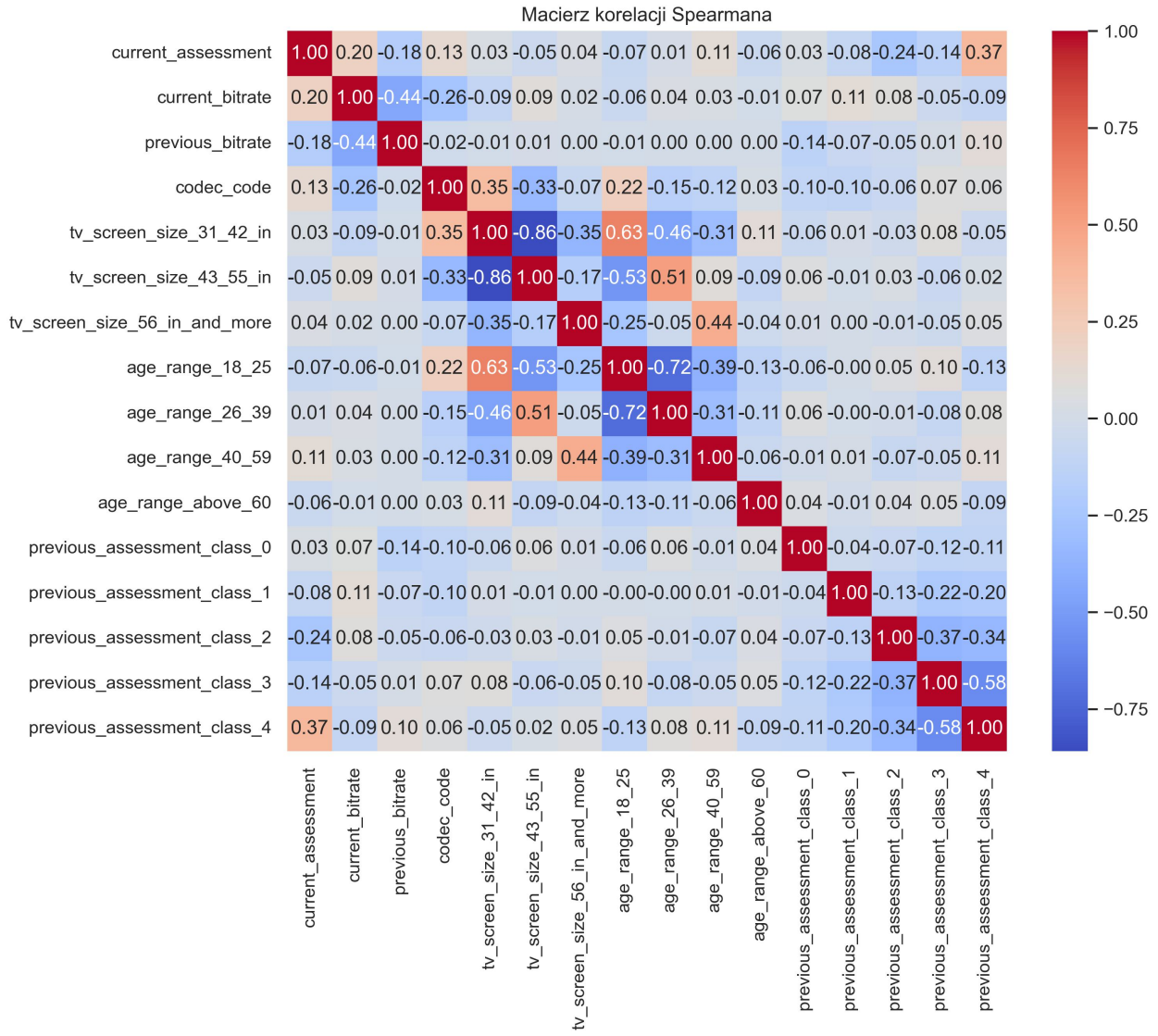
Współczynniki korelacji liniowej Pearsona [117] oraz współczynniki korelacji rang Spearmana [117] zostały obliczone dla analizowanych cech w celu sprawdzenia siły liniowej zależności między zmiennymi oraz ich monotoniczności. Ze względu na subiektywną naturę ocen sprawdzono zależność pomiędzy cechami wejściowymi oraz udzieloną oceną *current_assessment*. Największa dodania zależność występuje dla cech *current_bitrate* oraz *previous_assessment_class_*4, czyli zmiennych odpowiednio określającej aktualną przepływność oraz poprzednią ocenę w przedziale wartości wynoszącym 81 – 100. Wyliczone współczynniki zostały umieszczone na macierzach korelacji na rysunkach 4.3 oraz 4.4 odpowiednio dla korelacji Pearsona oraz korelacji rang Spearmana.

80:9381574816



Rysunek 4.3 Macierz korelacji Pearsona dla analizowanego zestawu danych

81:7650351441



Rysunek 4.4 Macierz korelacji rang Spearmana dla analizowanego zestawu danych

4.3.4 Analiza głównych składowych

Metoda analizy głównych składowych *PCA* (ang. *principal component analysis*) umożliwia redukcję liczby wymiarów zmiennych wejściowych poprzez zastąpienie ich nowymi zmiennymi. Nowe zmienne, zwane głównymi składowymi są liniowymi kombinacjami oryginalnych zmiennych utworzonych w taki sposób, aby utracić jak najmniej informacji [118]. Analiza PCA obejmuje etapy takie jak: wyliczenie średniej i macierzy kowariancji zmiennych, znalezienie wektorów własnych (ang. *eigenvectors*) i ich wartości własnych (ang. *eigenvalues*), wybór wektorów o największych wartościach własnych oraz projekcję oryginalnych punktów do przestrzeni wyznaczonej przez wybrane wektory własne. W analizie PCA zawartość informacji w składowych głównych można mierzyć wariancją [118]. Wykorzystując bibliotekę scikit-learn [119] wykonano analizę PCA dla przygotowanych danych, wyniki przedstawiono w tabeli 4.2. Obliczono, że aby zawartość informacji była na poziomie co najmniej 90% potrzeba 8 zmiennych głównych, które sumarycznie osiągają wariancję na poziomie

82:3803159460

92,88%. Wyjaśnienie przynajmniej 95% wariancji wymaga już 9 zmiennych głównych, którch sumaryczna wariancja wynosi 96,82%. Dla 10 składowych głównych wartość wariancji sumuje się do 100% co oznacza, że całość informacji zawarta jest w tych 10 kombinacjach liniowych. Możliwa jest redukcja wymiarowości danych do 10 zmiennych. Należy jednak pamiętać, że korzystając z głównych składowych można utracić szansę na interpretowalność cech w stosunku do wyniku predykcji. W takim przypadku dane wejściowe obejmować będą cechy PC1-PC10. Przykładowo stwierdzenie, w jaki sposób zakres wieku lub rozmiar ekranu wpływa na wynik predykcji może nie być możliwe w bezpośredni sposób.

Tabela 4.2 Wynik analizy PCA

Składowa główna	Cząstkowa wariancja wyjaśniona	Skumulowany udział wariancji
PC1	0,238	0,238
PC2	0,157	0,395
PC3	0,127	0,522
PC4	0,124	0,646
PC5	0,083	0,729
PC6	0,079	0,808
PC7	0,064	0,872
PC8	0,056	0,928
PC9	0,039	0,967
PC10	0,033	1

4.3.5 Liniowa analiza dyskryminacyjna

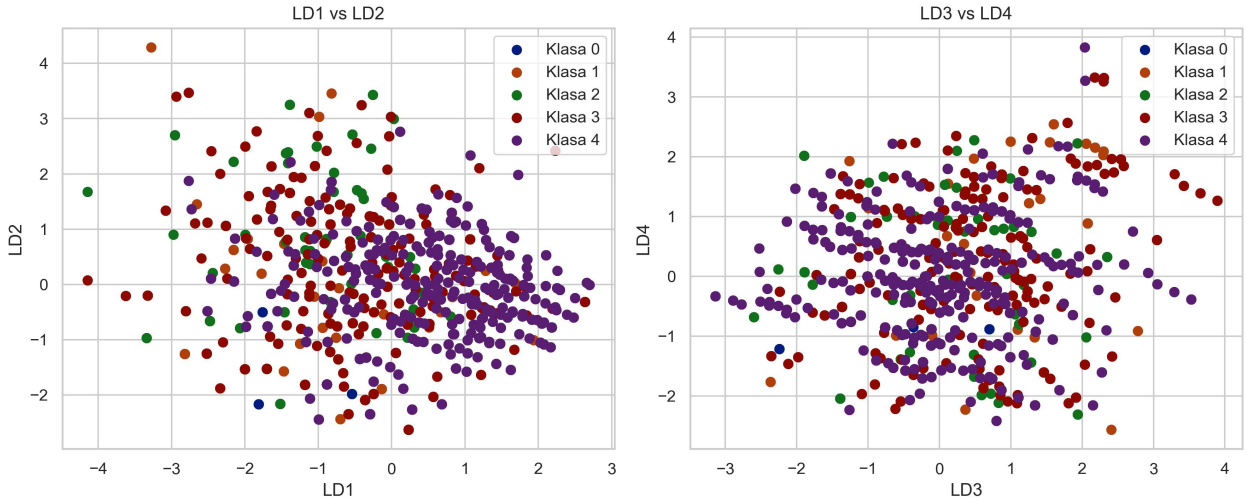
Liniowa analiza dyskryminacyjna LDA (ang. *linear discriminant analysis*) to metoda umożliwiająca wyznaczenie reguły decyzyjnej bazującej na funkcji liniowej dla problemów klasyfikacji. W LDA możliwa jest redukcja wymiarowości danych oraz maksymalizacja separacji pomiędzy klasami. Etapy LDA obejmują obliczenie średniej całkowitej oraz średnich dla klas, wyznaczenie macierzy rozproszenia wewnątrzgrupowego minimalizując wariancję wewnątrz każdej klasy, wyznaczenie zmienności międzygrupowej i maksymalizacja separacji między klasami. LDA prowadzi do wyznaczenia hiperpłaszczyzny dyskryminacyjnej, która dzieli przestrzeń w taki sposób, aby wyznaczyć podzbiory określające przynależność do grup. Maksymalna liczba składowych dyskryminacyjnych możliwych do uzyskania wynosi $K - 1$, gdzie K to liczba klas [120]. Analiza LDA została wykonana za pomocą biblioteki scikit-learn [119]. Analizą objęto pięć klas, co pozwoliło na uzyskanie maksymalnie czterech składowych. W tabeli 4.3 zawarto poziom wyjaśnionej wariancji dla każdej składowej. Dwuwymiarowa przestrzeń utworzona jako kombinacja liniowa LD1 oraz LD2 zachowuje 95% informacji rozróżniających klasy. Skumulowana suma dla trzech składowych wynosi 98,2%..

83:2941876289

Tabela 4.3 Wynik analizy LDA

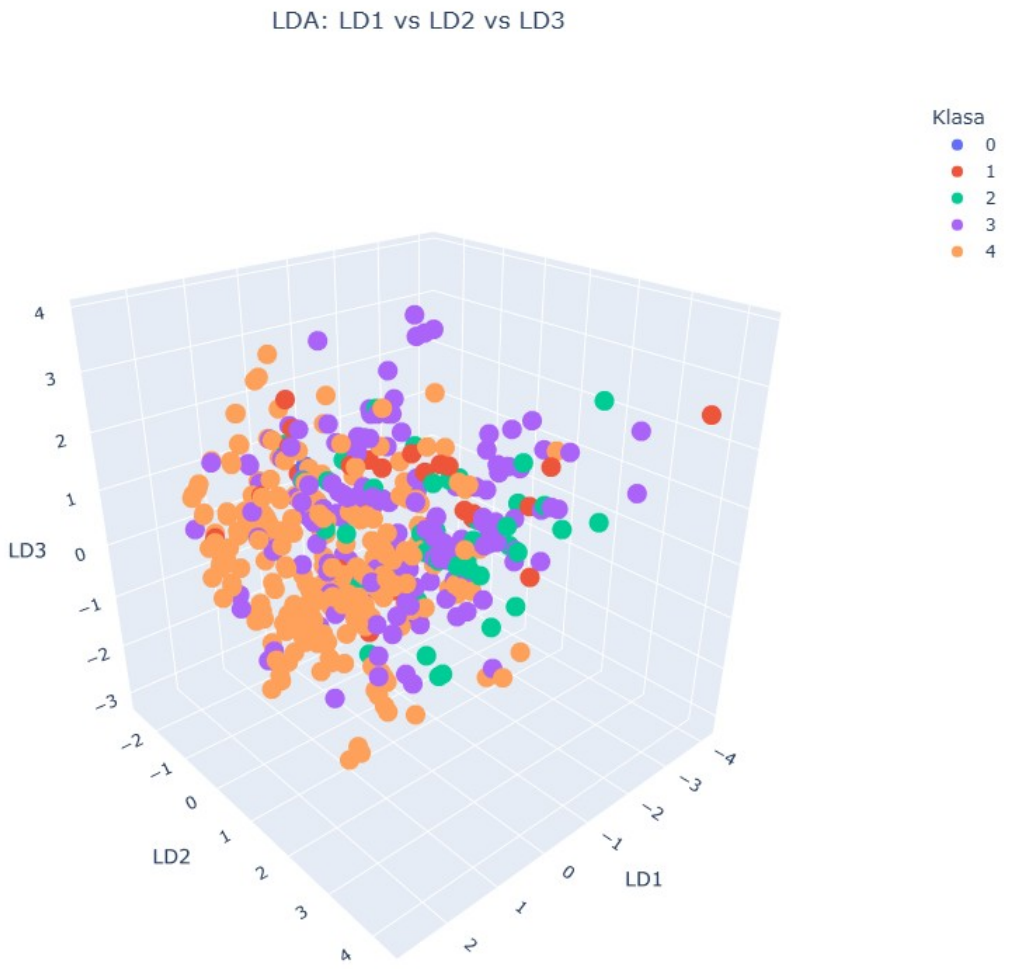
LD1	LD2	LD3	LD4
0,800	0,152	0,030	0,018

Finalny model uczenia maszynowego powinien operować na pięciu klasach, ze względu na łatwość interpretacji wyników. W celu sprawdzenia poziomu separowalności klas przygotowano wizualizację poszczególnych czynników dyskryminacyjncych. Na wykresie 4.5 przedstawiono separowalność klas w przestrzeni LD1 i LD2 oraz LD3 i LD4. Wykres LD1 vs LD2 prezentuje poziom separacji dla 5 klas z wykorzystaniem pierwszych dwóch zmiennych dyskryminacyjnych, które łącznie wyjaśniają co najmniej 95% wariancji. Wykres LD3 vs LD4 prezentuje separację dla mniej istotnych kombinacji, które sumarycznie wyjaśniają poniżej 5% wariancji. Klasy nie są jednoznacznie separowalne zarówno na pierwszym, jak i drugim wykresie. Pogłębiona analiza dla 3 składowych dyskryminacyjnych LD1, LD2 oraz LD3 zaprezentowana na wykresie 4.6 również nie wskazuje na jednoznaczną separację klas. Brakuje skupisk, wyraźnie oddzielonych od pozostałych punktów. Reprezentacja graficzna 4.7 prezentuje zestaw dwuwymiarowych wykresów porównujących zależności pomiędzy wszystkimi składowymi dyskryminacyjnymi. Wykresy wskazują na niejednoznaczne separację klas, co może sugerować nieliniowość związaną z subiektywnym charakterem uzyskanych ocen.



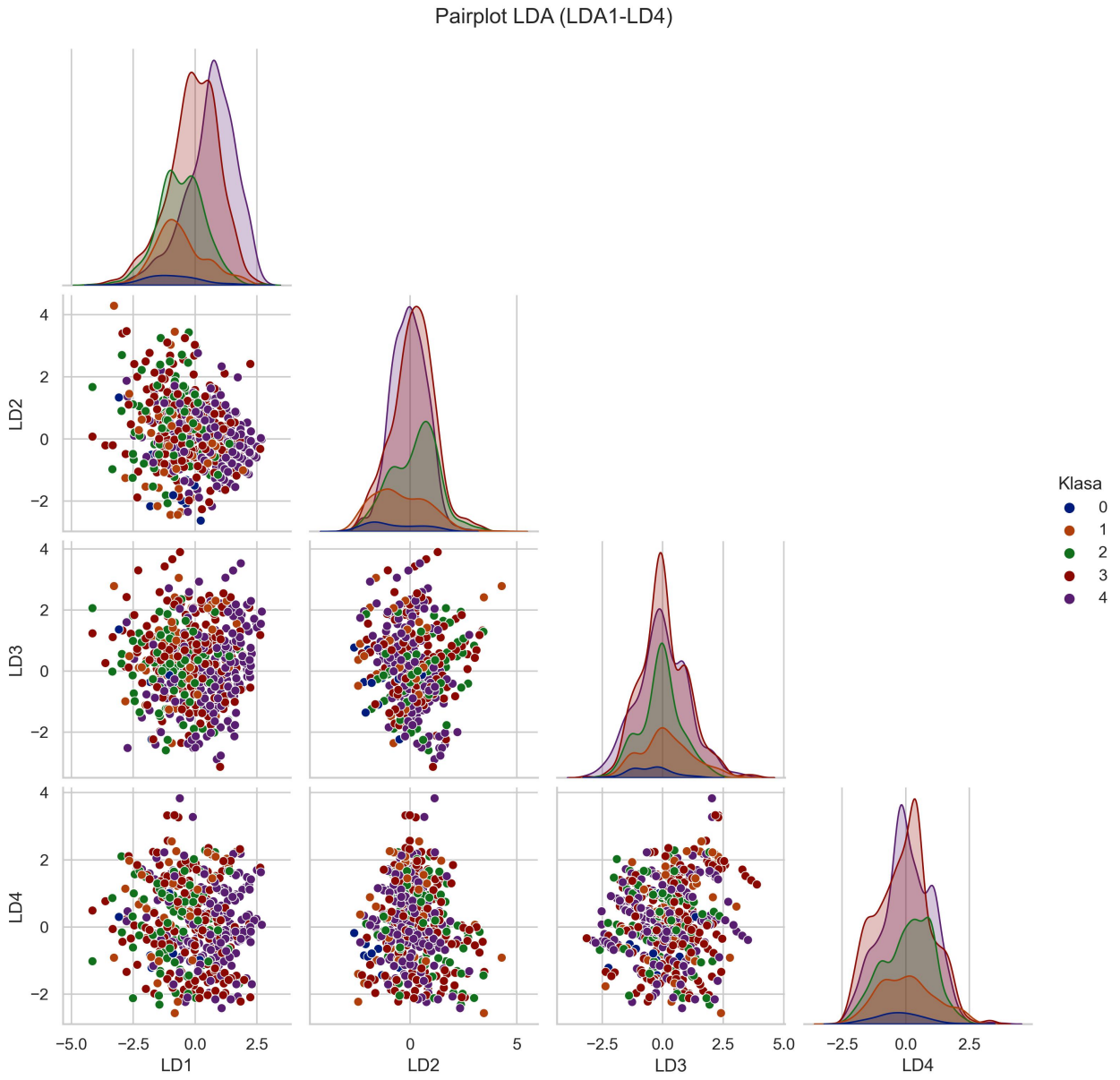
Rysunek 4.5 Separowalność klas dla kombinacji liniowych LD1 i LD2 oraz LD3 i LD4

84:9906881439



Rysunek 4.6 Separowalność klas dla kombinacji liniowych LD1, LD2 oraz LD3

85:5935098413



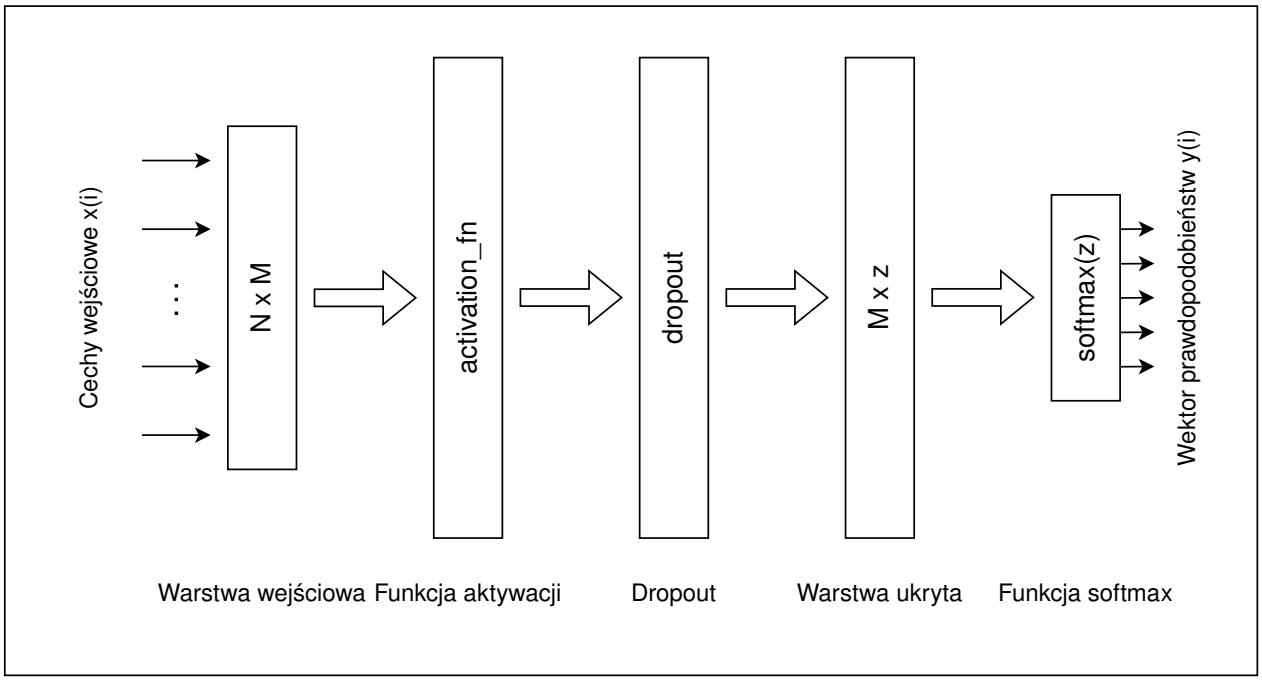
Rysunek 4.7 Separowalność klas dla czterech kombinacji liniowych

4.4 Bazowy model sieci neuronowej

Na schemacie 4.8 przedstawiono ogólną architekturę sieci do predykcji ocen dla użytkowników telewizji Hiway. Bazuje ona na sieci neuronowej posiadającej liniową warstwę wejściową z liczbą N wejść, gdzie $N = 15$ oraz M wyjść. Warstwa ukryta to kombinacja warstwy liniowej, aktywacji oraz warstwy regularyzacyjnej. Aktywacja wyjść opisana jest parametrem *activation_fn*, który wskazuje stosowaną funkcję aktywacji. Technika regularyzacji opuszczeń (ang. *dropout*) umożliwia dezaktywację wybranych neuronów wraz z ich wejściami i wyjściem. Pozwala to na lepszą generalizację oraz zapobiega przeuczeniu (ang. *overfitting*) sieci neuronowej [121]. Dropout w bibliotece PyTorch jest określony poprzez parametr p będący prawdopodobieństwem wyzerowania (wyłączenia) danej jednostki sieci.

86:1164802698

Liniowa warstwa wyjściowa posiada M wejść oraz Z wyjść, gdzie $Z = 5$. Na wyjściu tej warstwy zwracany jest pięcioelementowy wektor nieznormalizowanych wartości liczbowych (ang. *logits*). Funkcja aktywacji *softmax* przekształca nieznormalizowany wektor wejściowy w pięcioelementowy wektor prawdopodobieństw. Suma wartości wynosi 1, a każda wartość to prawdopodobieństwo danej oceny w skali ACR dla zadanych cech wejściowych.



Rysunek 4.8 Bazowy model sieci neuronowej wykorzystany do dalszych prac

Środowiskiem uruchomieniowym dla testów był komputer z procesorem Intel Core i5-1145G7 z zainstalowaną pamięcią RAM wynoszącą 16 GB pracujący pod kontrolą systemu operacyjnego Windows 11. Kod do przygotowania danych i uczenia modeli sieci neuronowych napisano w języku Python w wersji 3.9.16. Wykorzystano w głównej mierze narzędzie i biblioteki takie jak: JupyterLab [122], PyTorch [123], NumPy [124], scikit-learn [119], SciPy [125], Pandas [126] oraz Optuna [127].

Przygotowano algorytm 1 w celu sprawdzenia wstępnych założeń. Wytrenowano 3 podstawowe modele dla 16, 64 oraz 256 wejść. Funkcja ReLU została wykorzystana do aktywacji wyjść, natomiast jako funkcję straty przyjęto dywergencję Kullbacka-Leiblera (KLDivLoss). Dane wejściowe zostały podzielone na trzy zbiory: treningowy, walidacyjny oraz testowy przyjmując, że 60% zbioru wejściowego to dane treningowe, 20% to dane walidacyjne i 20% to dane testowe. Przyjęto po 2 próbki jako rozmiar wsadu do uczenia (ang. *batch size*). Do optymalizacji sieci wykorzystano optymalizator Adam z współczynnikiem uczenia (ang. *learning rate*) wynoszącym 0.001. Metoda wczesnego zatrzymania (ang. *early stopping*) została wykorzystana, aby uniknąć przeuczenia modelu. Jeżeli przez 10 kolejnych epok nie będzie poprawy wyniku funkcji straty dla zbioru walidacyjnego to uczenie zostanie zatrzymane. Jako punkt końcowy przyjęto maksymalnie 200 epok uczenia. Dla warstwy regularyzacji opuszczeń (ang. *dropout*) ustalono stałą wartość p wynoszącą 0,3.

87:6375371739

Algorytm 1: Uczenie bazowej sieci neuronowej do klasyfikacji z miękkimi etykietami

Dane wejściowe: Iterator zbioru treningowego: *train_loader*, Iterator zbioru walidacyjnego: *validation_loader*, zbiór hiperparametrów modelu: H

Dane wyjściowe: Najlepszy model: *best_model*

```
improve_limit ← 10
without_improve_ct ← 0
model ← utwórz instancję modelu bazując na hiperparametrach H
Dla epoch ← 1 do 200 wykonaj
    Dla każdego train_batch w train_loader wykonaj
        y_train_predicted ← model(x_train_batch)
        loss ← oblicz stratę KLDivLoss(y_predicted, y_train_batch)
        model ← wykonaj propagację wsteczną i zaktualizuj wagi modelu
    Koniec
    Dla każdego validation_batch w validation_loader wykonaj
        y_predicted ← model(x_validation_batch)
        y_predicted_probs ← softmax(y_predicted)
        predictions_list ← dodaj do listy y_predicted_probs
        truth_list ← dodaj do listy y_validation_batch
    Koniec
    Jeśli loss < best_loss to
        best_loss ← loss
        without_improve_ct ← 0
        best_model ← model
    w przeciwnym razie
        without_improve_ct + +
        Jeśli without_improve_ct >= improve_limit to
            Przerwij ze względu na brak poprawy
Koniec
Zwróć best_model
```

Metryki oceniające skuteczność modeli

Dla każdego modelu obliczono metryki typowe dla algorytmów klasyfikacji takie jak: trafność, dokładność, czułość, średnia harmoniczna F1.
Trafność klasyfikacji (ang. *Accuracy*), opisana równaniem 4.14 to stosunek liczby prawidłowo sklasyfikowanych próbek do liczby wszystkich próbek w zbiorze.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.14}$$

gdzie:

Accuracy – dokładność

TP – próbki prawdziwie dodatnie (ang. *true positive*)

TN – próbki prawdziwie ujemne (ang. *true negative*)

88:7944271832

FP – próbki fałszywie dodatnie (ang. *false positive*)

FN – próbki fałszywie ujemne (ang. *false negative*)

Dokładność (ang. *Precision*) jest opisana równaniem 4.15. Określa odsetek przewidzianych pozytywnych próbek, które jest faktycznie pozytywne dla danej klasy. Jest to stosunek prawdziwie dodatnich próbek do sumy próbek prawdziwie dodatnich i fałszywie dodatnich.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.15}$$

gdzie:

Precision – dokładność

TP – próbki prawdziwie dodatnie (ang. *true positive*)

FP – próbki fałszywie dodatnie (ang. *false positive*)

Czułość (ang. *Recall*) została opisana równaniem 4.16. Określa odsetek przewidzianych pozytywnych próbek danej klasy, które zostały prawidłowo sklasyfikowane jako pozytywne. Jest to stosunek elementów prawdziwie dodatnich do sumy wartości prawdziwie dodatnich i fałszywie ujemnych.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.16}$$

gdzie:

Recall – czułość

TP – próbki prawdziwie dodatnie (ang. *true positive*)

FN – próbki fałszywie ujemne (ang. *false negative*)

Miara F1 opisana równaniem 4.17. Mierzy skuteczność modelu biorąc pod uwagę dokładność i czułość.

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.17}$$

Współczynnik korelacji rang Spearmana

Współczynnik korelacji rang Spearmana SROCC (ang. *Spearman's Rank-Order Correlation Coefficient*) zdefiniowany jako 4.18 to nieparametryczna miara monotonicznej zależności dla dwóch uporządkowanych zbiorów cech (rang) [128].

$$SROCC = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{4.18}$$

gdzie:

$SROCC$ – współczynnik korelacji rang Spearmana ,

d_i – różnica pomiędzy i-tą rangą w danych wektorach,

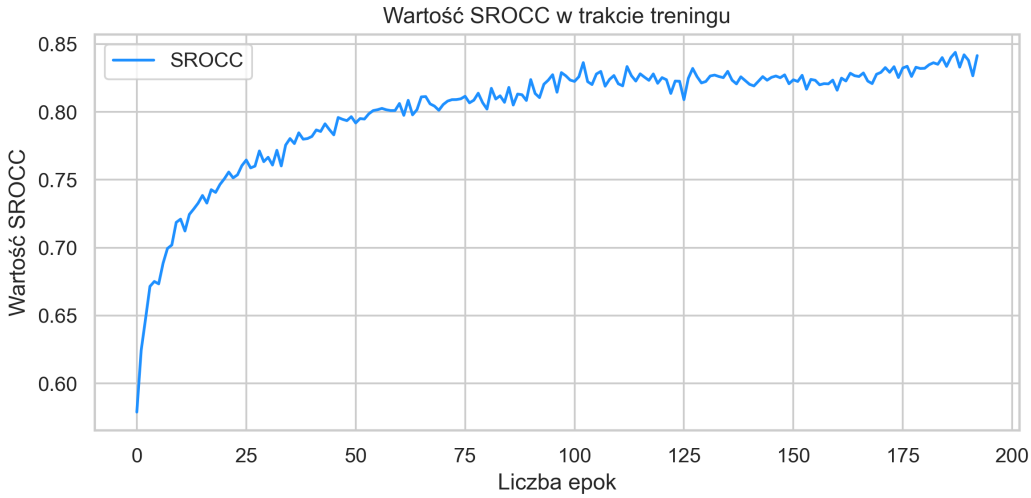
89:4386583414

n – liczba rang w wektorze.

Współczynnik SROCC przyjmuje wartości z zakresu od -1 do 1 i umożliwia określenie poziomu korelacji rangowej dla uporządkowanych zbiorów. Są to przykładowo wektor wartości prawdopodobieństw dla subiektywnych ocen w skali ACR, będących źródłowymi etykietami dla modeli predykcyjnych i wektor przewidywanych wartości prawdopodobieństw będących wyjściem tych modeli. Współczynnik SROCC pozwala na ocenę skuteczności modeli klasyfikacji wieloklasowej bazującej na miękkich etykietach, mierząc zgodność kolejności wartości prawdopodobieństw w obu wektorach. Współczynnik SROCC jest regularnie wykorzystywany w publikacjach naukowych poświęconych ocenie jakości wideo. Dla każdego wyuczonego modelu wykorzystano SROCC jako podstawową metrykę porównawczą.

90:1496964758

4.5 Wyniki predykcji dla modeli bazowych

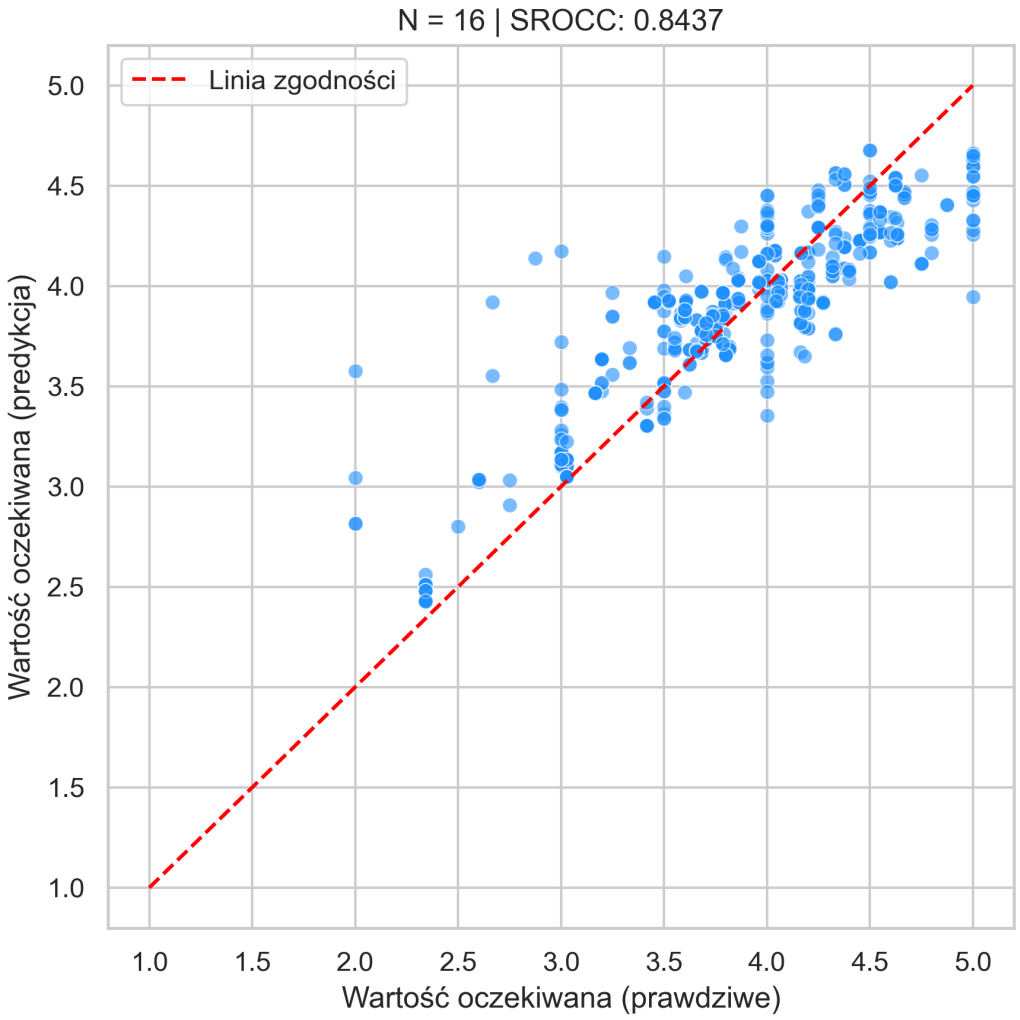


Rysunek 4.9 Wartości SROCC dla modelu nr 1, dla którego liczba wejść N = 16

Na wykresie 4.9 przedstawiono zależność wartości SROCC do liczby epok uczenia modelu $N = 16$. Model uzyskał SROCC wynoszące 0,8415 na poziomie zbioru walidacyjnego po ponad 190 epokach uczenia. Dla wektorów rzeczywistych prawdopodobieństw oraz wektorów przewidzianych prawdopodobieństw obliczono wartości oczekiwane z rozkładu klas w skali ACR.

Na rysunku 4.10 przedstawiono porównanie wartości oczekiwanych dla prawdziwych prawdopodobieństw (oś X) oraz wartości prawdopodobieństw dla predykcji (oś Y). Czerwona linia wyznacza zgodność pomiędzy predykcją, a wartością prawdziwą. Rozproszone niebieskie punkty określają błędy predykcji. Dla modelu $N = 16$ wykres pokazuje, że dla ocen poniżej 4 model przeszacowuje predykcje (punkty powyżej linii zgodności), natomiast powyżej oceny 4 ocena jest niedoszacowana (punkty poniżej linii zgodności). Duże rozproszenie punktów sugeruje znaczącą liczbę błędów.

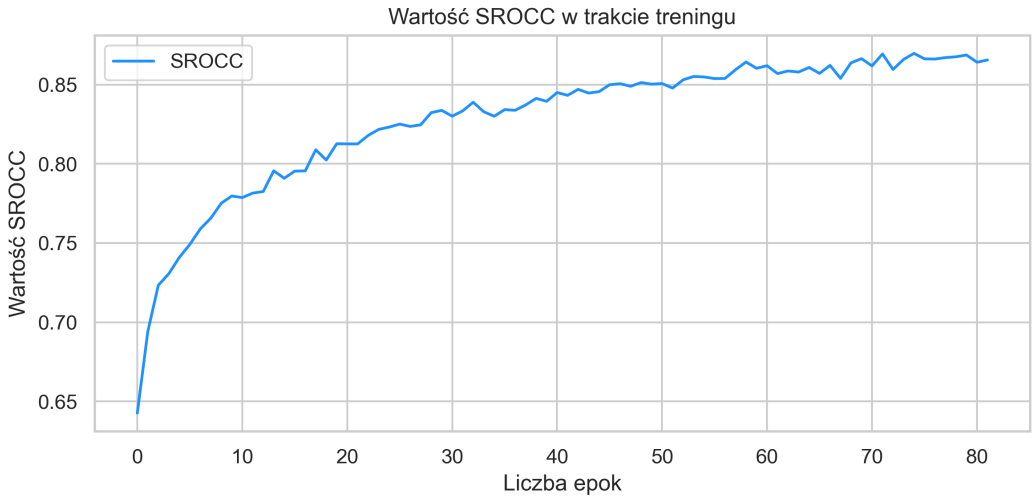
91:1080824105



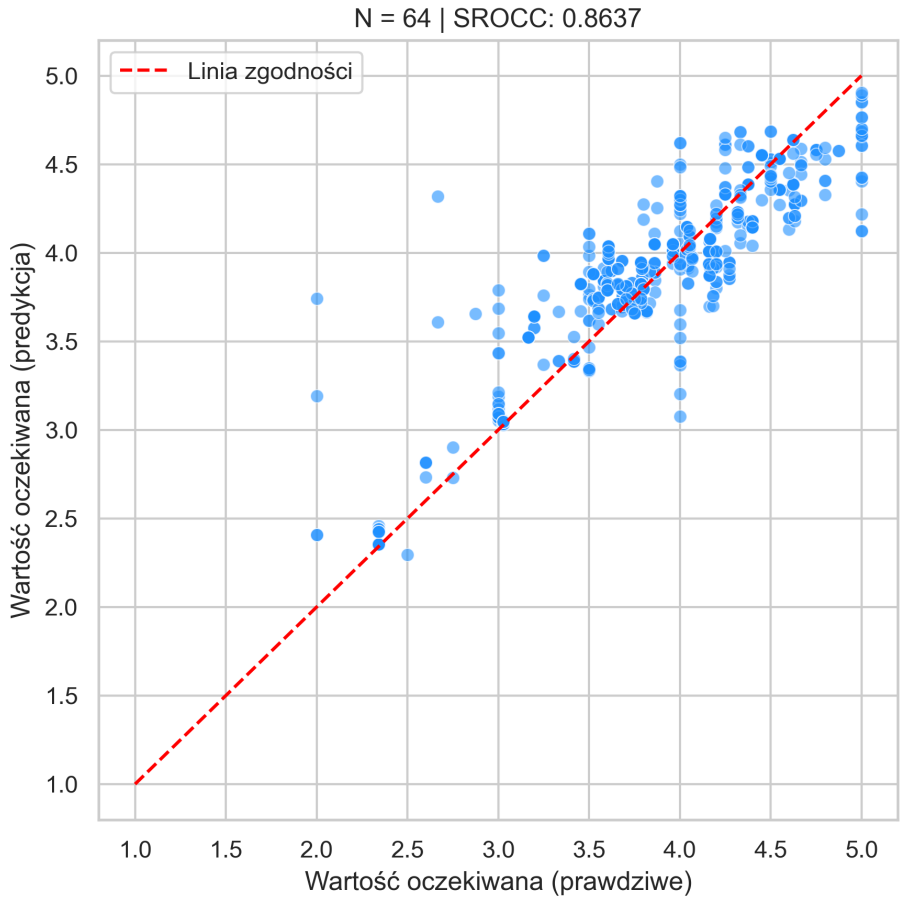
Rysunek 4.10 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Model $N = 16$

Na wykresie 4.11 widac, że model z liczbą wejść wynoszącą $N = 64$ uzyskał SROCC na poziomie 0,864. Jest to wartość zbliżona do modelu wcześniejszego. Na rysunku 4.12 widać, że rozproszenie wartości oczekiwanych jest zbliżone do poprzedniego modelu.

92:5201842398



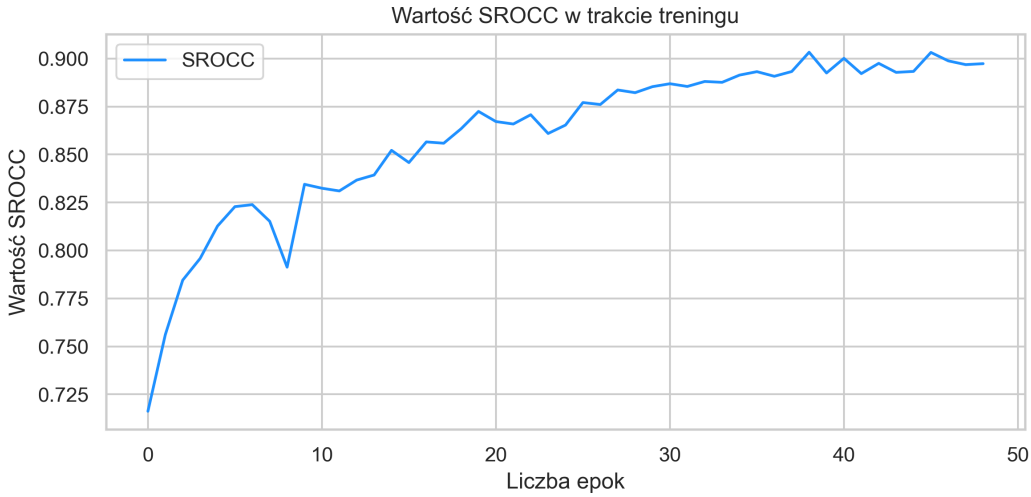
Rysunek 4.11 Wartości SROCC dla modelu nr 2, dla którego liczba wejść N = 64



Rysunek 4.12 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Model $N = 64$

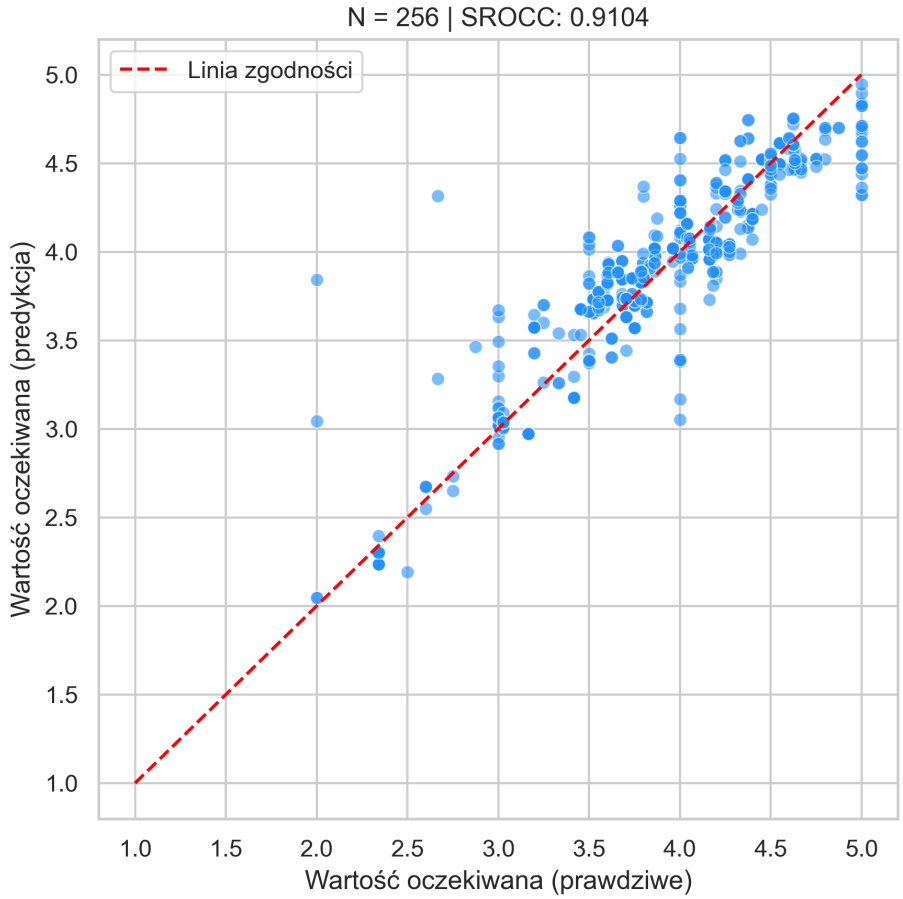
93:9608611390

Trzeci model o liczbie wejść $N = 256$ uzyskał SROCC o wartości bliskiej 0,91, co widać na wykresie 4.13. Zgodność szeregowania prawdopodobieństw w wektorach jest więc na poziomie powyżej 90%. Niemniej jednak, jak widać na wykresie 4.14 wartości oczekiwane szacowanych klas są nadal dość mocno rozbieżne w stosunku do oczekiwanych wartości prawdziwych. Trzeci model potrzebował ponad 140 epok uczenia mniej od modelu pierwszego oraz około 30 iteracji mniej od modelu drugiego aby osiągnąć maksymalny poziom SROCC dla danej architektury.



Rysunek 4.13 Wartości SROCC dla modelu nr 3, dla którego liczba wejść N = 256

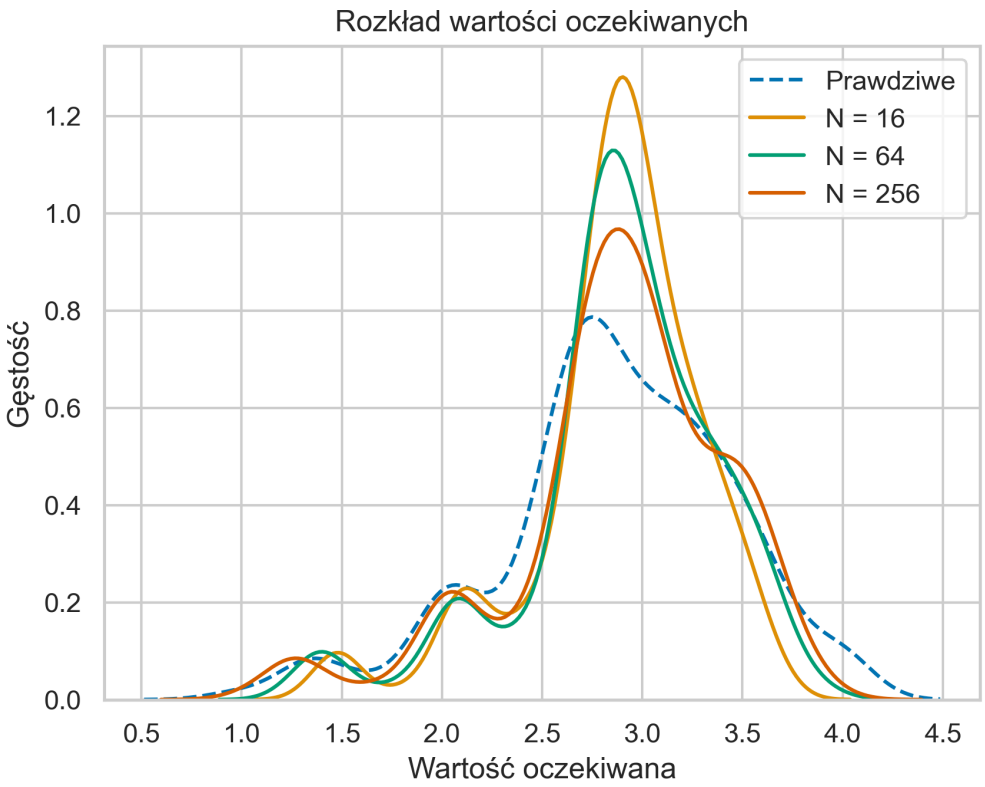
94:6883276450



Rysunek 4.14 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Model $N = 256$

Rozłożenie wartości oczekiwanych dla rozkładów prawdopodobieństw ocen rzeczywistych oraz szacowanych zostało przedstawione na wykresie gęstości 4.15. Linią przerywaną zaznaczono wartości prawdziwe, natomiast linią ciągłą, w zmiennym kolorze zaznaczono predykcję danego modelu. Wykresy mają podobna charakterystykę. Dla wartości prawdziwych najwięcej ocen skupionych jest pomiędzy klasą 2 oraz 3. Dla wszystkich modeli widoczne jest nieznaczne przesunięcie. Modele niedoszacowują najwyższej oceny, gdyż charakterystyka predykcji zanika po ocenie o wartości 4. W zbiorze walidacyjnym jest nieznaczna liczba ocen o najniższej wartości. Z tego względu prawidłowość predykcji dla najniższej oceny będzie widoczna na macierzy pomyłek (ang. *confusion matrix*) modelu. Zbliżona charakterystyka gęstości dla wartości prawdziwych i predykcji może oznaczać, że kierunek poszukiwań optymalnego modelu może być prawidłowy.

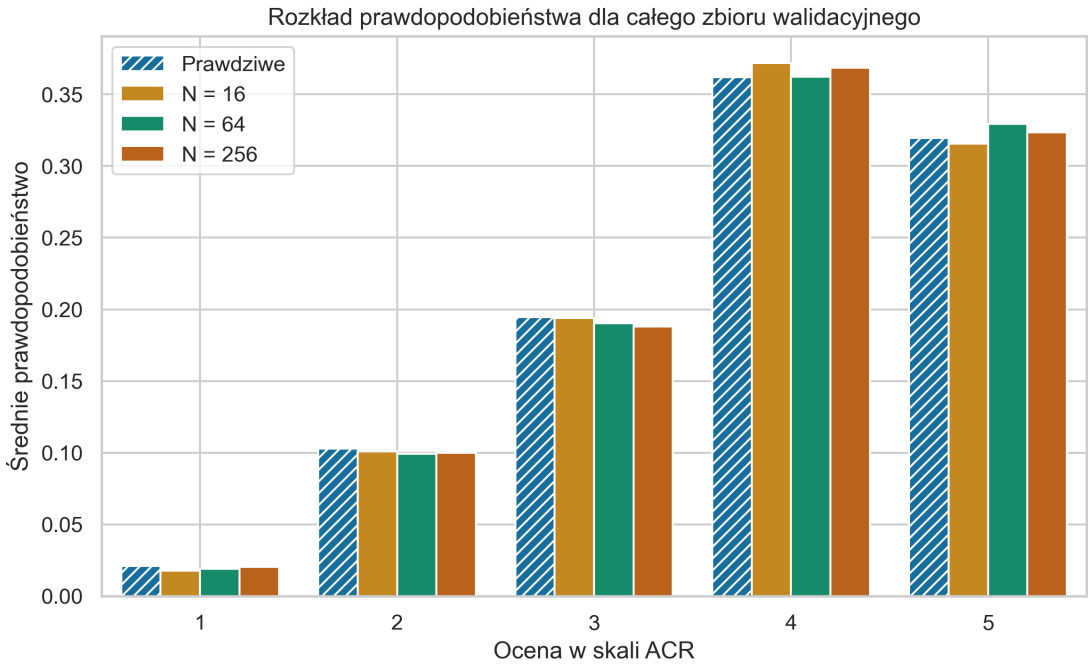
95:1111352784



Rysunek 4.15 Gęstość rozkładu wartości oczekiwanych

Rozkład uśrednionego prawdopodobieństwa każdej klasy dla wartości prawdziwych i predykcji przedstawiono na wykresie 4.16. Na wykresie widać, w jaki sposób modele są pewne danej klasy w sposób globalny. Oceny 4 i 5 przeciętnie często uzyskują najwyższe prawdopodobieństwo. Predykcje modeli są dość dobrze odwzorowane w stosunku do rzeczywistych etykiet. Taka charakterystyka rozkładów oraz wyniki poszczególnych metryk sugerują, że modele potrafią odwzorować zależności zbioru, jednak mogą popełniać błędy predykcji na korzyść sąsiednich klas.

96:5797177131



Rysunek 4.16 Rozkład uśrednionego prawdopodobieństwa każdej klasy

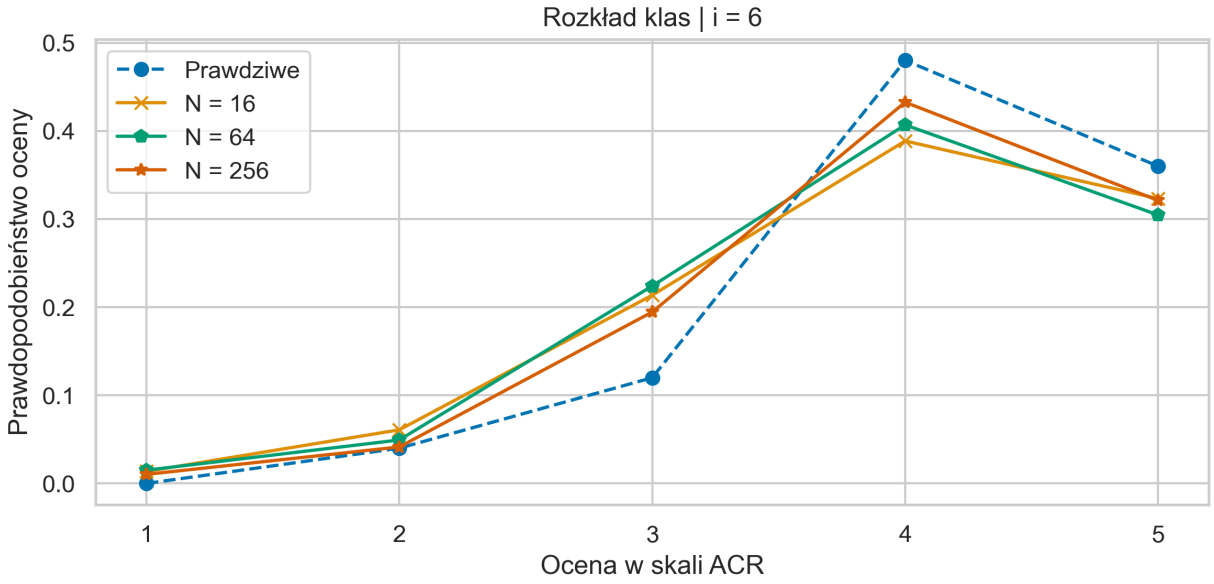
Wykresy: 4.17, 4.18, 4.19, 4.20, 4.21 oraz 4.22 prezentują predykcję prawdopodobieństw dla każdej oceny w skali ACR oraz obliczone prawdziwe prawdopodobieństwa (linia przerywana) dla wybranej próbki ze zbioru walidacyjnego. Wykorzystano wykresy liniowe, aby w lepszy sposób zobrazować skuteczność predykcji. W tabelach 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, zawartych pod każdym z wykresów umieszczono parametry danej próbki. Każda próbka odpowiada 60-cio sekundowej sekwencji wideo. Każda próbka złożona jest z dwóch 30-sekundowych materiałów wideo (para filmów). Nagłówki tabel przedstawione jako akronimy objaśniono poniżej:

- i – numer próbki,
- c_a – current_assessment - subiektywna ocena w skali od 1 do 100 aktualnie odtwarzanego materiału wideo (materiał B z pary A i B),
- p_a – previous_assessment - subiektywna ocena w skali od 1 do 100 poprzednio odtwarzanego materiału wideo (materiał A z pary A i B),
- c_b – current_bitrate - przepływność aktualnie odtwarzanego materiału wideo,
- p_b – previous_bitrate - przepływność poprzednio odtwarzanego materiału wideo,
- cdc – codec - typ kodeka,
- t_v_s – zakres rozmiaru ekranu TV w calach,
- a_r – zakres wieku oglądającego,

97:7832536571

- p_1, p_2, p_3, p_4, p_5 – wektor rzeczywistych prawdopodobieństw przydzielenia danej oceny w skali ACR.

Dla próbki $i = 6$ wszystkie modele wskazały ocenę 4 z największym prawdopodobieństwem. Próbka $i = 23$ to prawdziwie ocena 4, z mniejszym prawdopodobieństwem oceny 5. Dla pozostałych ocen prawdziwe prawdopodobieństwo wynosi zero. Modele przewidziały ocenę 4 i 5 z najwyższym prawdopodobieństwem. Wskazały jednak niezerowe predykcje dla klas niższych. W ten sam, jednoznaczny sposób przewidziano ocenę 5 dla próbki $i = 89$. Modele popełniły błąd dla próbki $i = 222$, dla której prawidłowa ocena to 3. Predykcja wskazała jednoznacznie najwyższe prawdopodobieństwo dla oceny 4. Ponadto, modele przypisały kolejne najwyższe prawdopodobieństwo ocenie najwyższej (5). Dopiero na trzecim miejscu znajduje się prawidłowa wartość. Próbka $i = 345$ została prawidłowo oszacowana na ocenę 5. Kolejna próbka $i = 587$ została oszacowana z najwyższymi prawdopodobieństwami jako ocena 4 lub 5. Prawdziwa ocena o wartości 5 dominuje jednak nad 4, więc modele nie są pewne oceny dla tej próbki. Monotoniczność wykresów predykcji względem wykresu wartości prawdziwych wskazuje, że wybrane podejście może być odpowiednie do problemu klasyfikacji jakości wideo bazującej na danych subiektywnych. Zróżnicowane prawdopodobieństwa dla różnych klas wskazują, że modele nie uogólniają swoich predykcji.

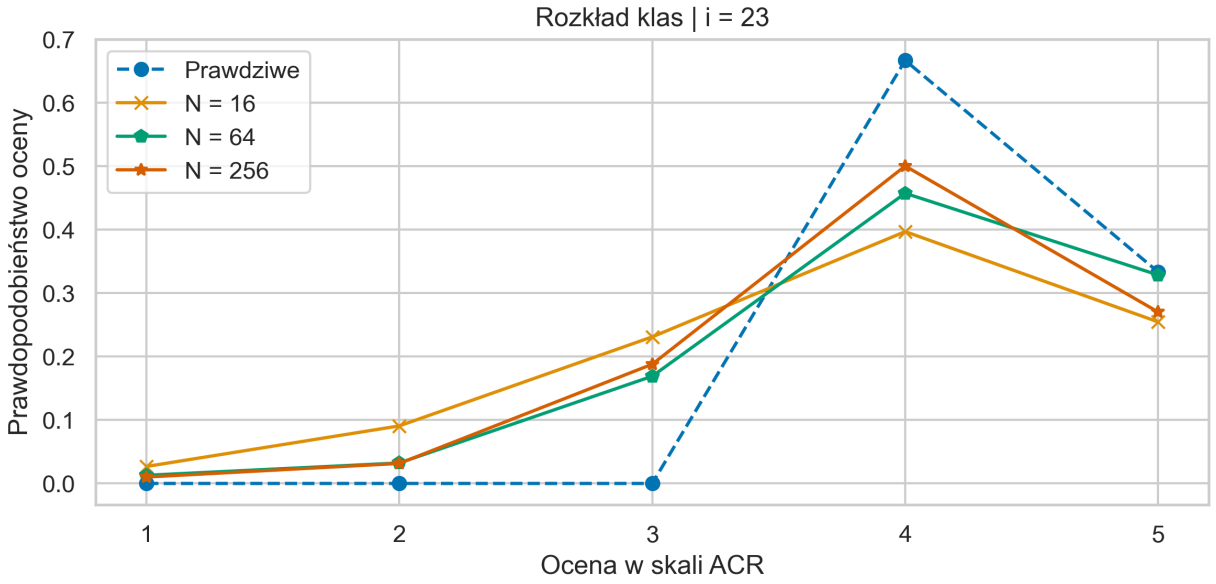


Rysunek 4.17 Wyniki predykcji dla próbki walidacyjnej i = 6

Tabela 4.4 Parametry dla próbki $i = 6$

i	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
6	79	79	3298	5342	H.265	31 - 42	18 - 25	0,000	0,040	0,120	0,480	0,360

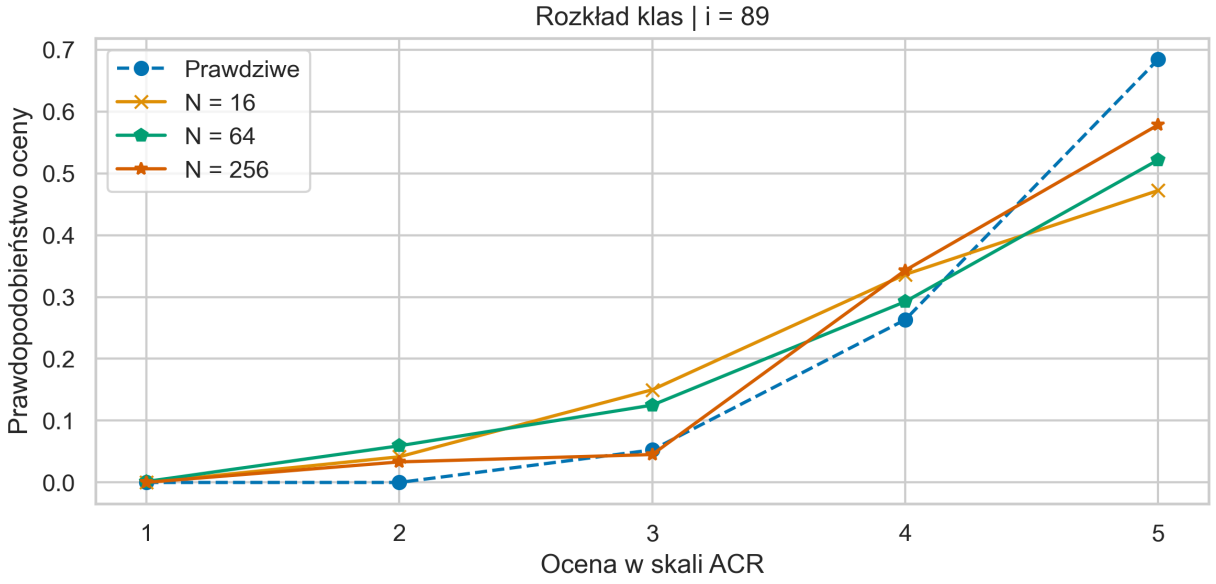
98:6170067045



Rysunek 4.18 Wyniki predykcji dla próbki walidacyjnej i = 23

Tabela 4.5 Parametry dla próbki $i = 23$

i	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
23	68	90	3000	6000	H.264	> 56	40 - 59	0,000	0,000	0,000	0,667	0,333

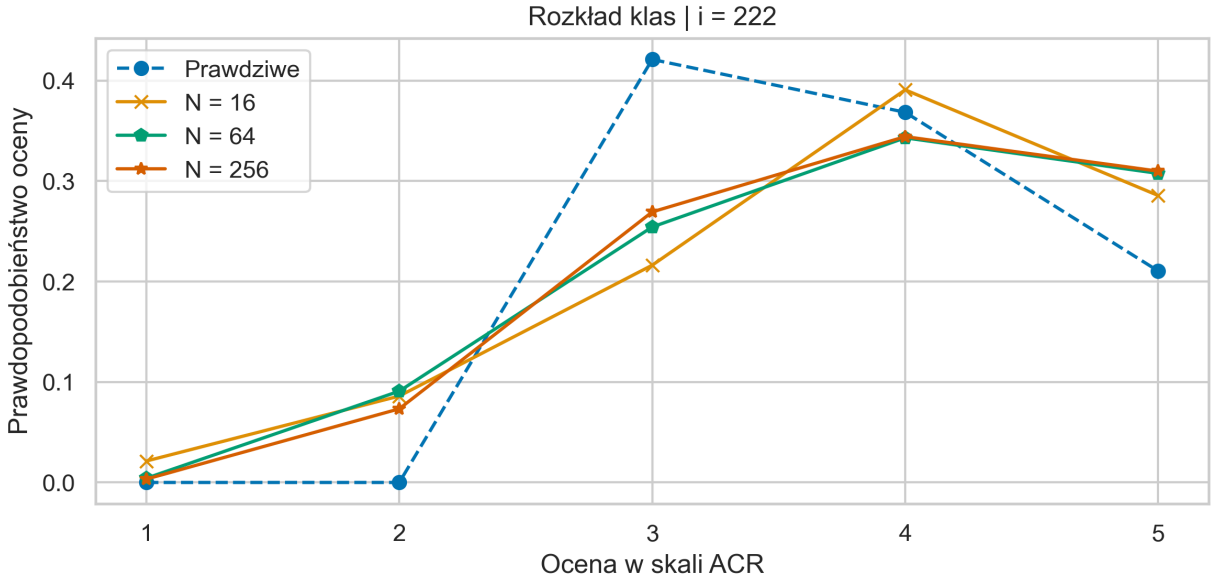


Rysunek 4.19 Wyniki predykcji dla próbki walidacyjnej i = 89

99:1660662787

Tabela 4.6 Parametry dla próbki $i = 89$

i	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
89	85	75	10000	1500	H.264	43 - 55	26 - 39	0,000	0,000	0,053	0,263	0,684

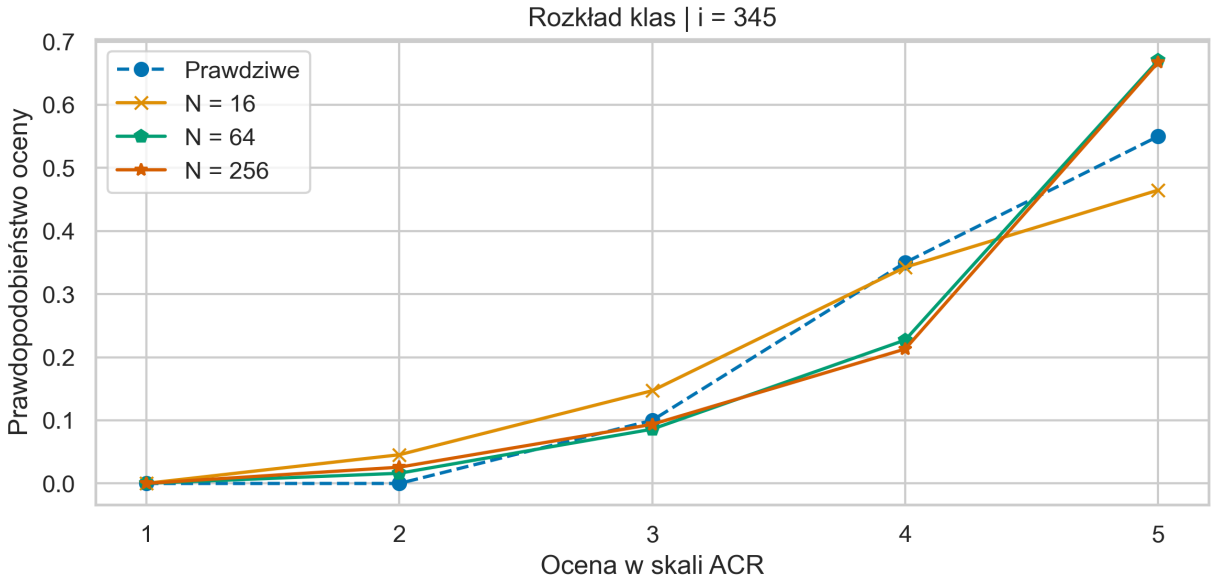


Rysunek 4.20 Wyniki predykcji dla próbki walidacyjnej i = 222

Tabela 4.7 Parametry dla próbki $i = 222$

i	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
222	63	60	6000	1500	H.264	43 - 55	26 - 39	0,000	0,000	0,421	0,368	0,211

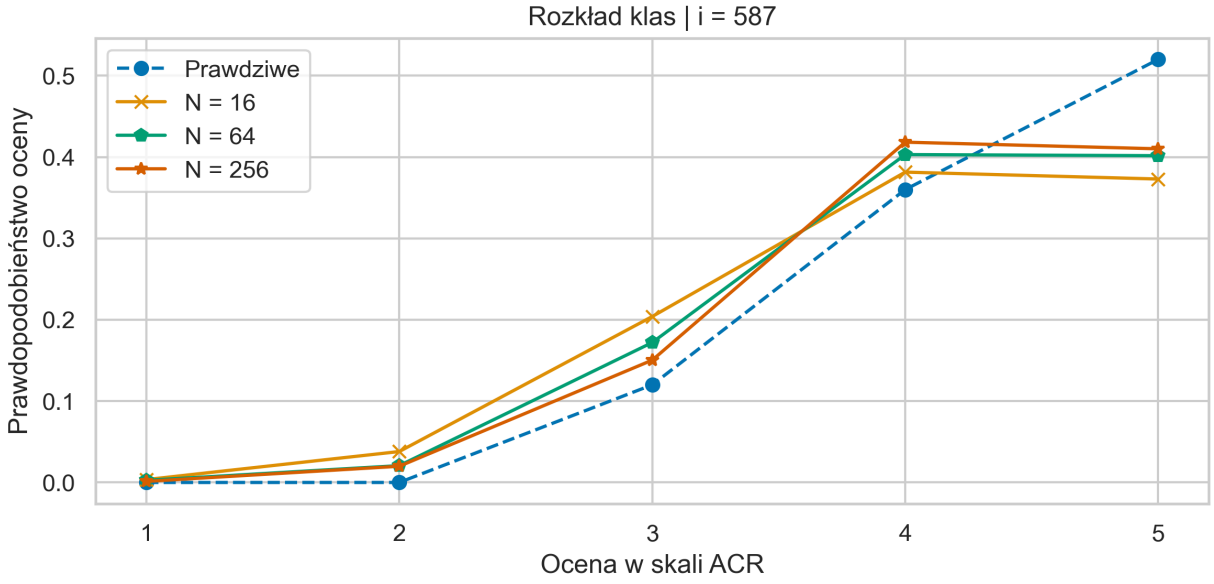
100:1088248782



Rysunek 4.21 Wyniki predykcji dla próbki walidacyjnej i = 345

Tabela 4.8 Parametry dla próbki $i = 345$

i	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
345	77	88	3298	6364	H.265	43 - 55	26 - 39	0,000	0,000	0,100	0,350	0,550



Rysunek 4.22 Wyniki predykcji dla próbki walidacyjnej i = 587

101:6761329401

Tabela 4.9 Parametry dla próbki $i = 587$

i	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
587	93	81	5342	2276	H.265	31 - 42	18 - 25	0,000	0,000	0,120	0,360	0,520

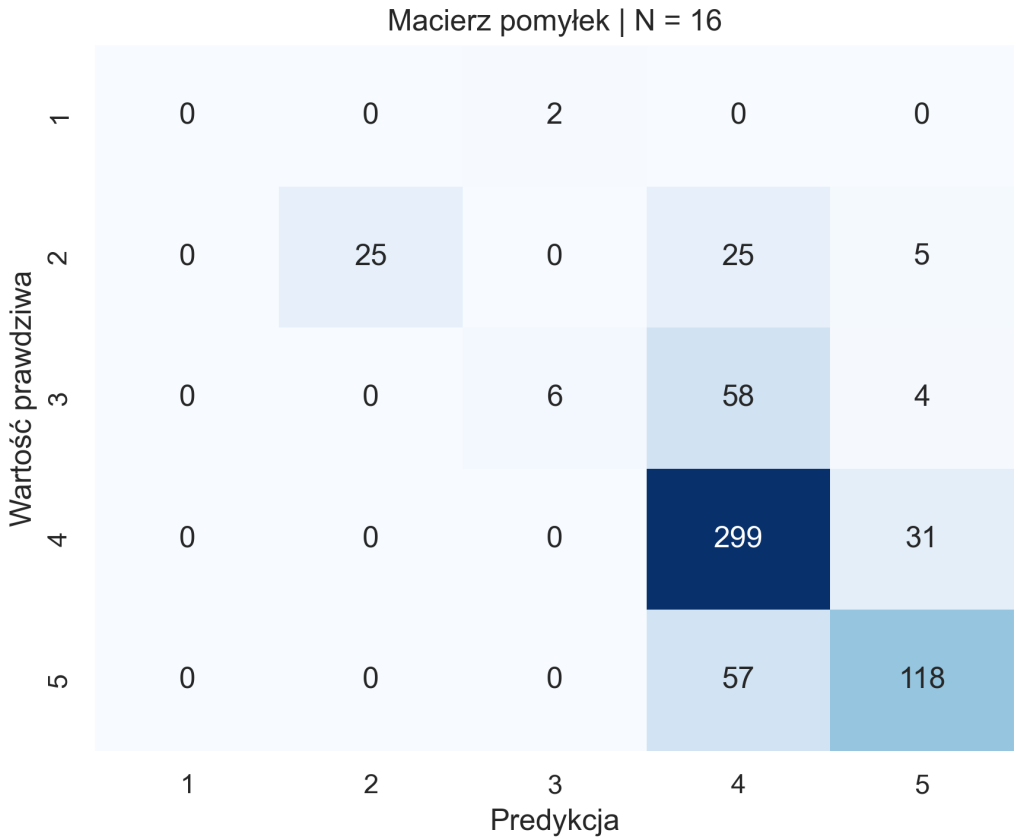
Dla każdego modelu wygenerowano macierz pomyłek (ang. *confusion matrix*). Na osi Y przedstawione są klasy prawdziwe, natomiast na osi X wartości przewidziane przez dany model.

W tabelach 4.10, 4.11 oraz 4.12 umieszczonych poniżej macierzy pomyłek zawarto metryki podsumowujące skuteczność bazowych modeli uczenia maszynowego do klasyfikacji oceny w skali ACR. Wyjściem modeli jest pięcioelementowy wektor prawdopodobieństw. Do obliczenia metryk przyjęto ocenę z najwyższym prawdopodobieństwem, czyli $argmax\, y(i)$.

Dokładność, czyli ułamek prawidłowych klasyfikacji dla trzech wyuczonych modeli wyniosła odpowiednio 0,711, 0,733 oraz 0,795. Dla precyzji, czyli stosunku wyników prawdziwie dodatnich do wszystkich wyników dodatnich osiągnięto wartości w okolicy 0,63 dla modeli $N = 16$ oraz $N = 64$. Model $N = 256$ uzyskał wynik precyzji bliski 0,85. Wartości czułości, czyli prawidłowo sklasyfikowanych rzeczywistych wyników pozytywnych dla modeli $N = 16$ oraz $N = 64$ oscylują w okolicy 0,63. Dla modelu $N = 256$ czułość osiągnęła wartość bliską 0,85. Dla czułości, czyli współczynnika określającego zdolność do identyfikacji pozytywnych przypadków w zbiorze uzyskano wartości poniżej 0,5 dla modeli $N = 16$ oraz $N = 256$ oraz bliskie 0,74 dla modelu $N = 256$. Dla parametru F1 uzyskano wartości w okolicy 0,5 dla pierwszych dwóch modeli oraz powyżej 0,77 dla modelu trzeciego.

Dla macierzy 4.23 wygenerowanej dla modelu $N = 16$ widać, że model nie był w stanie oszacować próbek o najniższej jakości i pomylił je z oceną 3. Dla oceny 2 w 50% procentach predykcji model wykonał błąd i przypisał próbki do oceny 4. Klasa 3 została w znacznej większości sklasyfikowana jako klasa 4. Tylko 6 próbek z 68 zostało prawidłowo oznaczonych. Dla klas 4 i 5 wystąpiła największa zgodność. Pomiędzy tymi klasami pojawiło się też najwięcej pomyłek. Udzielone oceny bazujące na subiektywnej opinii najczęściej zawierały się w zakresie ocen 4 i 5. Zbiór danych wykorzystany do uczenia, testowania i walidacji nie jest zbalansowany, gdyż uczestnicy badania rzadziej udzielali ocen o najniższych wartościach. Z tego względu liczba ocen danej klasy wzrasta wraz z wartością tej oceny.

102:1892290608



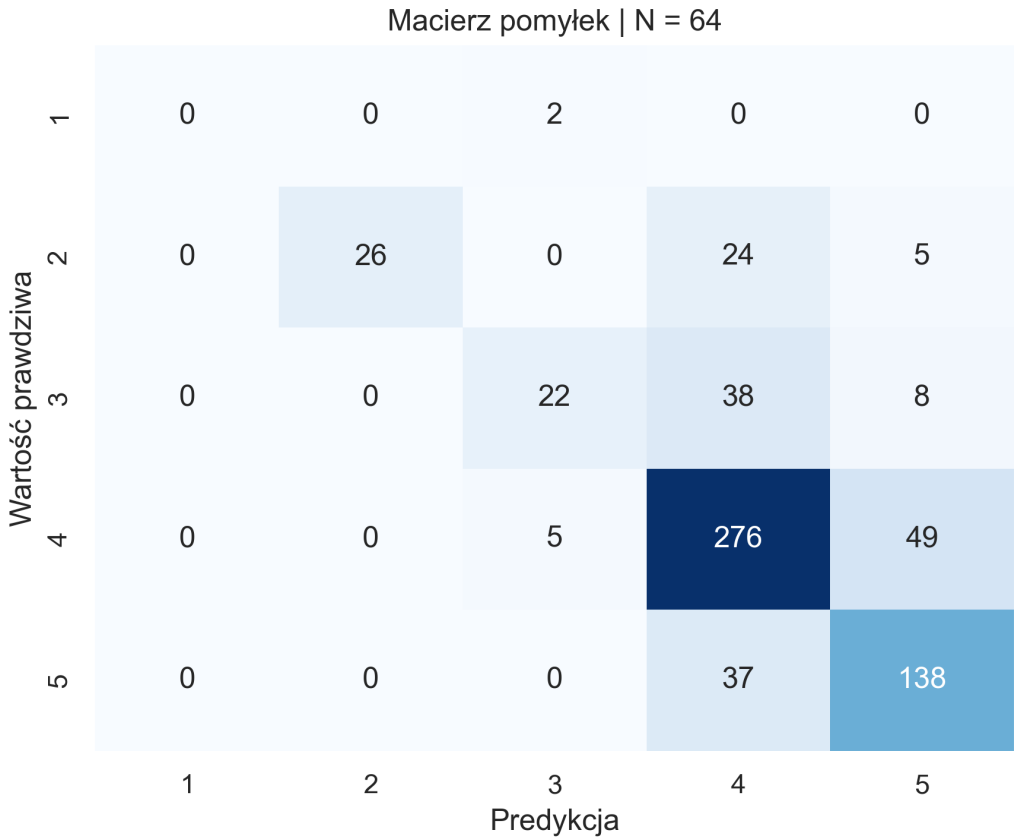
Rysunek 4.23 Macierz pomyłek na modelu $N = 16$

Tabela 4.10 Metryki dla modelu $N = 16$

Model	SROCC	Accuracy	Precision	Recall	F1
N=16	0,844	0,711	0,636	0,425	0,454

Macierz pomyłek 4.24 reprezentuje model $N = 64$. Dla oceny 1 i 2 występuje podobna do poprzedniego modelu skala błędów. Model ten nieco lepiej radzi sobie z klasą 3, dla której prawidłowo przewidział 22 próbki, 38 pomylił z klasą 38, a 8 próbek sklasyfikował jako ocena 5. Poziom błędów dla ocen 4 i 5 jest podobny do modelu $N = 16$. Występuje znaczoną liczba błędów pomiędzy obiema klasami.

103:3607317296



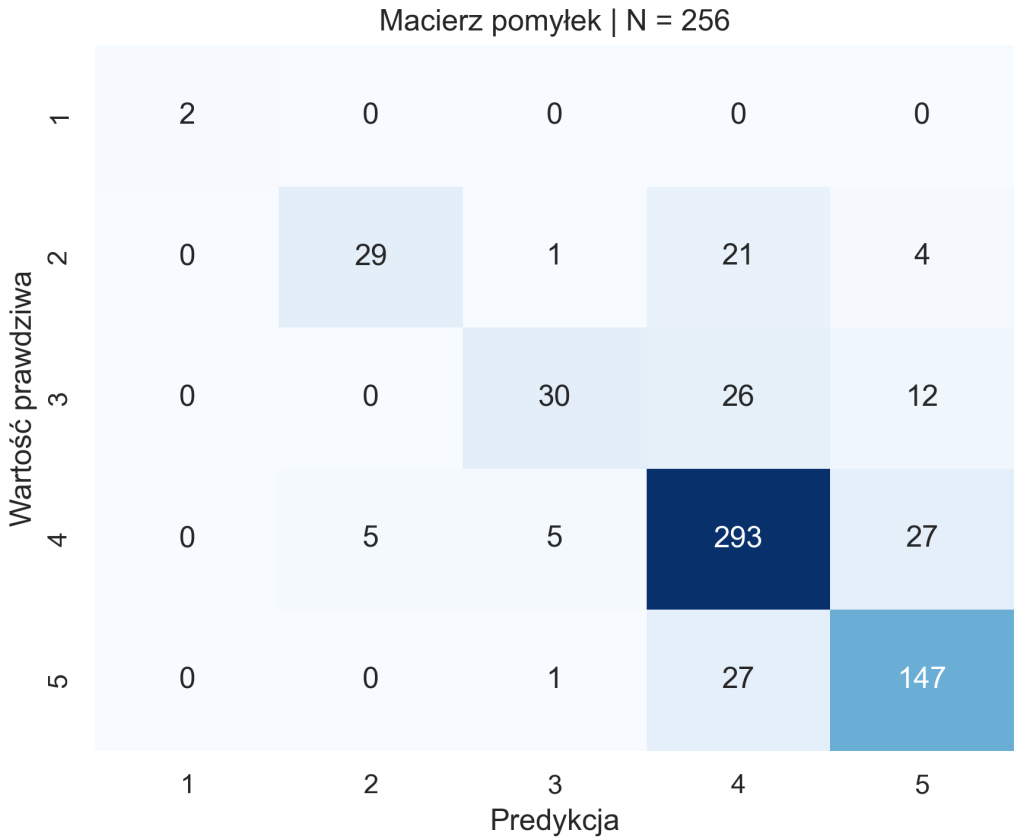
Rysunek 4.24 Macierz pomyłek na modelu $N = 64$

Tabela 4.11 Metryki dla modelu $N = 64$

Model	SROCC	Accuracy	Precision	Recall	F1
N=64	0,864	0,733	0,637	0,484	0,523

Na macierzy pomyłek 4.25 dla modelu $N = 256$ widać, że prawidłowo została rozpoznana najniższa klasa. Dla klas 2 i 3 poziom błędów jest podobny jak dla obu poprzednich modeli. Predykcja dla dwóch najwyższych ocen jest nieco skuteczniejsza, jednak nadal występuje znaczna liczba pomyłek.

104:7762924798



Rysunek 4.25 Macierz pomyłek dla modelu $N = 256$

Tabela 4.12 Metryki dla modelu $N = 256$

Model	SROCC	Accuracy	Precision	Recall	F1
N=256	0,910	0,795	0,847	0,739	0,774

4.6 Podsumowanie

Wytrenowane bazowe modele miękkiej klasyfikacji (ang. *soft classification*) stanowią trafne podejście do szacowania oceny QoE bazującej na subiektywnych danych. Skuteczność modeli opisanych w niniejszym rozdziale może być jeszcze daleka od zadowalającej, niemniej wyznacza kierunek dalszych badań. Współczynnik korelacji rang Spearmana SROCC dla modelu $N = 256$ osiągnął wartość powyżej 0,9, co oznacza, że przewidziane prawdopodobieństwa w ponad 90% są zgodne z wektorami rzeczywistych prawdopodobieństw. Biorąc pod uwagę pozostałe metryki, których wartości pozostają niezadowalające, można stwierdzić potrzebę zwiększenia złożoności i precyzji modelu.

Istotnym aspektem jest również charakter zbioru danych, którego główną cechą jest subiektywność ocen użytkowników. Zbiór ten jest niezbalansowany – rozkład ocen wskazuje, że uczestnicy badania bardzo rzadko udzielali skrajnie niskich ocen (1). Oceny 2 i 3 pojawiały się częściej, jednak nadal w mniejszym stopniu niż najwyższe klasy (4 i 5). Istnieją metody balansowania klas w zbiorze danych,

105:6389236245

takie jak nadpróbkowanie (ang. *oversampling*) czy *SMOTE* (ang. *Synthetic Minority Over-sampling Technique*) [129], jednak ze względu na charakter pracy techniki te nie zostały wykorzystane.

Strategia miękkiej klasyfikacji wydaje się właściwym podejściem do analizy zbioru subiektywnych danych, ponieważ umożliwia reprezentację niejednoznaczności i pozostawia przestrzeń dla próbek, które mogą nie być logiczne w kontekście innych obserwacji. W kolejnym rozdziale przedstawiono metody poszukiwania i optymalizacji algorytmów miękkiej klasyfikacji w celu znalezienia modeli, które będą charakteryzowały się zarówno wysoką wartością współczynnika korelacji SROCC, jak i zadowalającymi wynikami pozostałych metryk.

106:9860366730

Rozdział 5

Optymalizacja modeli uczenia maszynowego

5.1 Optymalizacja dla modeli z jedną warstwą ukrytą

Parametry zawarte w tabeli 5.1 zostały wykorzystane w celu znalezienia modelu sieci neuronowej zoptymalizowanego pod kątem metryki SROCC. W pierwszej fazie optymalizacji skupiono się na modelach o jednej warstwie ukrytej. Przeszukiwana przestrzeń jest rozszerzeniem bazowego modelu. Przyjęto liczbę wejść do warstwy ukrytej jako funkcję wykładniczą 2^n w zakresie n od 4 do 9, co daje od 16 do 512 wejść. Liczbę próbek (ang. *batch size*) przyjęto w zakresie od 2 do 64. Testowane funkcje aktywacji to: ReLU, LeakyReLU, Tanh oraz GELU, natomiast jako optymalizator przyjęto: Adam, AdamW, SGD i RMSProp. Ponadto obszar przeszukiwań obejmuje zmienne wartości dla hiperparametrów takich jak *dropout* w zakresie od 0,1 do 0,5, współczynnik uczenia (ang. *learning rate*) od 0,0001 do 0,001 oraz regularyzację L2 (ang. *weight decay*) w zakresie od 0,000001 do 0,001. Optymalizację modelu przeprowadzano za pomocą pakietu Optuna [127].

Tabela 5.1 Hiperparametry optymalizacyjne do poszukiwania optymalnej sieci neuronowej

Parametr	Wartości
Liczba wejść warstwy ukrytej	16, 32, 64, 128, 256, 512
Regularyzacja opuszczeń (ang. *dropout*)	0,1 - 0,5
Współczynnik uczenia (ang. *learning rate*)	0,0001 - 0,01
Regularyzacja L2 (ang. *weight decay*)	0,000001 - 0,001
Liczba próbek (ang. *batch size*)	2, 4, 8, 16, 32, 64
Funkcja aktywacji	ReLU, LeakyReLU, Tanh, GELU
Optymalizator	Adam, AdamW, SGD, RMSProp

Optymalizację podzielono na dwie główne fazy. W pierwszej fazie wykonano 200 iteracji (triali) przeszukiwania przestrzeni hiperparametrów zawartych w tabeli 5.1 zgodnie z algorytmem 2. Zbiór danych został podzielony na podzbiory: treningowy - 60% danych, walidacyjny - 20% danych oraz testowy - pozostałe 20% danych. Podział został wykonany w taki sposób, aby w każdym podzbiorze były zawarte próbki z prawdopodobieństwem przynależności do każdej klasy. Uczenie

107:8701762236

Algorytm 2: Optymalizacja sieci neuronowej dla modeli z jedną warstwą ukrytą

Dane wyjściowe: Lista triali: *trials_list*, Zbiór metryk triali: *trials_metrics*
trials_list ← []
trials_metrics ← []
Dla *trial* ← 1 *do* 200 **wykonaj**
    hidden_layer_n ← wybierz liczbę wejść warstwy ukrytej z: $[16, 32, 64, 128, 256, 512]$
    dropout ← wybierz wartość dropout z zakresu: $0,1 - 0,5$
    learning_rate ← wybierz wartość współczynnika uczenia z zakresu: $0,0001 - 0,01$
    weight_decay ← wybierz wartość współczynnika regularyzacji L2 z zakresu: $0,000001 - 0,001$
    batch_size ← wybierz wartość liczby próbek z: $[2, 4, 8, 16, 32, 64]$
    activation_func ← wybierz funkcję aktywacji z: [*ReLU*, *LeakyReLU*, *Tanh*, *GELU*]
    optimizer ← wybierz optymalizator z: [*Adam*, *AdamW*, *SGD*, *RMSProp*]
    H ← przypisz do zbioru wybrane hiperparametry:
      hidden_layer_n, *dropout*, *learning_rate*, *weight_decay*, *batch_size*, *activation_func*, *optimizer*
    train_loader ← utwórz iterator zbioru treningowego
    validation_loader ← utwórz iterator zbioru walidacyjnego
    model ← utwórz instancję modelu bazując na hiperparametrach H
    trial_model, *trial_metric* ← wywołaj *Algorytm 3* z argumentami:
      model, *train_loader*, *validation_loader*
    trials_list ← dodaj do listy *trial_model*
    trials_metrics ← dodaj do listy *trial_metric*
Koniec
Zwróć *trials_list*, *trials_metrics*

modeli odbyło się na zbiorze treningowym. Epoki uczenia były weryfikowane pod kątem metryki SROCC na zbiorze walidacyjnym. Uczenie modeli odbywało zgodnie z algorytmem 3 wywoływanym w ramach pętli testującej zbiór hiperparametrów. Mechanizm wczesnego zatrzymania (ang. *early stopping*), zapobiegł przetrenowaniu modeli (ang. *overfitting*) w trakcie uczenia. Model z najlepszym wynikiem dla danego podejścia został zapisany wraz z jego parametrami. Druga faza optymalizacji jest powiązana z subiektywnym charakterem zbioru danych. Podział zestawu danych na podzbiór treningowy, walidacyjny i testowy w stosunku 3:1:1 powoduje, że algorytmy nie mogą uczyć się na danych przewidzianych do walidacji i testów.

W drugiej fazie, ze względu na mały rozmiar zbioru danych zdecydowano się na wdrożenie strategii 10-krotnej walidacji krzyżowej (ang. *10-fold cross validation*) dla 50-ciu modeli o najwyższym SROCC uzyskanym w pierwszej fazie. W tym celu przygotowano algorytm 4. Jest on wywoływany w pętli, która przekazuje do algorytmu zainicjowany model z określonymi hiperparametrami. Zastosowana walidacja krzyżowa dzieli wskazany zbiór danych na 10 podzbiorów. Jeden z podzbiorów jest wykorzystany do walidacji, a pozostałe 9 do treningu. Operacja ta jest powtarzana 10 razy, a SROCC danego modelu to uśredniona wartość z 10 sprawdzianów krzyżowych. Walidacja krzyżowa została wdrożona zamiast podziału w stosunku 3:1:1, a jako zbiór do podziału wskazano 80% całości zbioru danych. Pozostałe 20% pozostawiono do finalnej weryfikacji modelu. W drugiej fazie uczenia modeli nadal obowiązywała strategia wczesnego zatrzymania dla 200 epok uczenia. Modele były inicjowane według zapisanych hiperparametrów dla 50 modeli o najwyższym SROCC w pierwszej fazie. Były jednak trenowane od zera, aby nie bazować na wcześniej ustalonych wagach.

108:1013554684

Algorytm 3: Uczenie sieci neuronowej do klasyfikacji z miękkimi etykietami. Podejście z podziałem na zbiory: treningowy, walidacyjny i testowy w proporcjach 3:1:1

Dane wejściowe: Iterator zbioru treningowego: *train_loader*, Iterator zbioru walidacyjnego: *validation_loader*, zainicjowany z określonymi hiperparametrami model: *model*
Dane wyjściowe: Najlepszy model: *best_model*, Zbiór metryk: *metrics*

```
improve_limit ← 10
without_improve_ct ← 0
Dla epoch ← 1 do 200 wykonaj
    Dla każdego train_batch w train_loader wykonaj
        y_train_predicted ← model(x_train_batch)
        loss ← oblicz stratę KLDivLoss(y_predicted, y_train_batch)
        model ← wykonaj propagację wsteczną i zaktualizuj wagi modelu
    Koniec
    Dla każdego validation_batch w validation_loader wykonaj
        y_predicted ← model(x_validation_batch)
        y_predicted_probs ← softmax(y_predicted)
        predictions_list ← dodaj do listy y_predicted_probs
        truth_list ← dodaj do listy y_validation_batch
    Koniec
    metrics ← oblicz metryki bazując na y_predicted_probs oraz y_validation_batch
    Jeśli loss < best_loss to
        best_loss ← loss
        without_improve_ct ← 0
        best_model ← model
    w przeciwnym razie
        without_improve_ct + +
        Jeśli without_improve_ct >= improve_limit to
            Przerwij ze względu na brak poprawy
Koniec
Zwróć best_model, metrics
```

W pierwszej pętli operującej na podziale zbioru na treningowy i testowy wytrenowano w sumie 192 modele. Bazując na wynikach SROCC wybrano do dalszej analizy 50 modeli o najwyższej wartości tego wskaźnika. W tabeli 5.2 umieszczono wyniki dla 50 prób o najwyższych wartościach SROCC dla pierwszej fazy uczenia (kolumna *SROCC 1*). W kolumnie *SROCC 2* umieszczono wartości współczynnika korelacji Spearmana dla drugiej fazy uczenia.
Poszczególne kolumny zawierają hiperparametry dla modeli:

- *N* – liczba wejść warstwy ukrytej z zakresu 16 - 512,
- *DO* – wartość współczynnika dropout z zakresu 0,1 - 0,5,
- *LR* – wartość współczynnika uczenia z zakresu 0,0001 - 0,01,
- *WD* – wartość współczynnika regularyzacji L2 z zakresu 0,000001 - 0,001,
- *BS* – liczba próbek w jednym kroku propagacji z zakresu 2 - 64,
- *Act* – funkcja aktywacji,
- *Opt* – zastosowany optymalizator.

109:2739090122

Algorytm 4: Uczenie sieci neuronowej do klasyfikacji z miękkimi etykietami. Podejście z 10-cio krotną walidacją krzyżową

Dane wejściowe: Zbiór treningowy: *train_d*, Zbiór walidacyjny: *validation_d*, zainicjowany z określonymi hiperparametrami model: *model*
Dane wyjściowe: Najlepszy model: *best_model*, Zbiór średnich metryk: *avg_metrics*

improve_limit ← 10
without_improve_ct ← 0
cross_val_metrics ← []
Dla *train_indexes*, *validation_indexes* ← *cross_validation*(*k do* 10 *z train_d*, *validation_d)* **wykonaj**
    train_loader ← zainicuj iterator zbioru treningowego bazując na *train_indexes*
    validation_loader ← zainicuj iterator zbioru walidacyjnego bazując na *validation_indexes*
    Dla *epoch* ← 1 *do* 200 **wykonaj**
        Dla każdego *train_batch w train_loader* **wykonaj**
            y_train_predicted ← *model*(*x_train_batch*)
            loss ← oblicz stratę KLDivLoss(y_predicted, y_train_batch)
            model ← wykonaj propagację wsteczną i zaktualizuj wagi modelu
        Koniec
        Dla każdego *validation_batch w validation_loader* **wykonaj**
            y_predicted ← *model*(*x_validation_batch*)
            y_predicted_probs ← $softmax$(*y_predicted*)
            predictions_list ← dodaj do listy *y_predicted_probs*
            truth_list ← dodaj do listy *y_validation_batch*
        Koniec
        metrics ← oblicz metryki bazując na *y_predicted_probs* oraz *y_validation_batch*
        Jeśli *loss* < *best_loss* **to**
            best_loss ← *loss*
            without_improve_ct ← 0
            best_model ← *model*
        w przeciwnym razie
            without_improve_ct + +
            Jeśli *without_improve_ct* >= *improve_limit* **to**
                Przerwij ze względu na brak poprawy
    Koniec
    cross_val_metrics ← dodaj do listy metryki *metrics* dla k-tego sprawdzianu krzyżowego
Koniec
avg_metrics ← oblicz średnie metryki z wszystkich walidacji z *cross_val_metrics*
Zwróć *best_model*, *avg_metrics*

W pierwszej fazie uczenia najlepsze wyniki zostały osiągnięte przez próby o numerach porządkowych 132, 153 oraz 177. Modele te osiągneły SROCC o wartości powyżej 0,9455. Pierwsze dwa modele charakteryzują się liczbą wejść $N = 128$, czterema próbkami w jednym kroku propagacji oraz funkcją aktywacji *GELU*. Trzeci model, o nieco niższej wartości SROCC posiadał liczbę wejść wartswy ukrytej wynoszącą $N = 512$. Liczba próbek w jednym kroku propagacji to dwie zamiast czterech. Aktualizacja wag sieci została wykonana przez optymalizator *AdamW*. Ogólnie, wyższe wartości SROCC w pierwszej fazie uczenia były osiągane przez algorytmy z małą liczbą próbek w jednym kroku propagacji. Tylko 3 modele z 50 posiadały liczbę próbek z zakresu 16 - 32. Ponadto, najwyższe wyniki były osiągane zawsze dla funkcji aktywacji *GELU* i optymalizatora *Adam* lub *AdamW*.

110:1091215146

Tabela 5.2 50 najlepszych modeli spośród 192 prób optymalizacji. Modele zostały posortowane malejąco według osiągniętego SROCC w pierwszej fazie

Trial	SROCC 1	SROCC 2	N	DO	LR	WD	BS	Act	Opt
132	0,9477	0,9249	128	0,319	0,0049	4,796E-06	4	GELU	Adam
153	0,9475	0,9304	128	0,293	0,0052	4,475E-06	4	GELU	Adam
177	0,9455	0,9253	512	0,321	0,0045	2,082E-06	2	GELU	AdamW
182	0,9438	0,9215	512	0,332	0,0051	1,774E-06	2	GELU	AdamW
122	0,9432	0,9272	128	0,316	0,0054	3,011E-06	4	GELU	Adam
108	0,9427	0,9033	128	0,331	0,0046	2,527E-06	4	GELU	Adam
97	0,9416	0,9126	128	0,301	0,0043	4,864E-06	4	GELU	Adam
92	0,9414	0,9168	128	0,289	0,0059	4,016E-06	4	GELU	Adam
185	0,9408	0,9104	512	0,333	0,0011	5,920E-06	2	GELU	AdamW
107	0,9397	0,9236	128	0,330	0,0047	4,811E-06	4	GELU	Adam
124	0,9396	0,9266	128	0,309	0,0072	2,643E-06	4	GELU	Adam
183	0,9392	0,9309	512	0,330	0,0050	6,108E-06	2	GELU	AdamW
123	0,9390	0,9378	128	0,312	0,0054	3,105E-06	4	GELU	Adam
180	0,9385	0,9147	512	0,317	0,0059	5,706E-06	2	GELU	AdamW
81	0,9385	0,9258	128	0,322	0,0047	5,449E-06	4	GELU	AdamW
45	0,9384	0,9324	128	0,355	0,0040	1,591E-06	4	GELU	Adam
188	0,9383	0,9069	512	0,347	0,0011	5,707E-06	2	GELU	AdamW
169	0,9379	0,9314	128	0,343	0,0033	3,206E-06	2	GELU	AdamW
144	0,9377	0,9070	128	0,292	0,0051	4,490E-06	4	GELU	Adam
198	0,9377	0,9152	512	0,319	0,0012	9,142E-06	2	GELU	AdamW
192	0,9374	0,8992	512	0,341	0,0013	7,187E-06	2	GELU	AdamW
32	0,9374	0,9091	128	0,325	0,0053	1,735E-06	32	GELU	AdamW
106	0,9369	0,9201	128	0,327	0,0047	4,421E-06	4	GELU	Adam
162	0,9368	0,9207	128	0,311	0,0050	4,554E-06	4	GELU	Adam
80	0,9367	0,9220	128	0,308	0,0040	1,558E-06	2	GELU	Adam
33	0,9366	0,8914	128	0,325	0,0030	1,725E-06	16	GELU	AdamW
102	0,9365	0,9242	128	0,293	0,0053	4,425E-06	4	GELU	Adam
134	0,9363	0,9333	128	0,317	0,0062	5,275E-06	4	GELU	Adam
93	0,9359	0,9280	128	0,285	0,0055	3,923E-06	4	GELU	Adam
130	0,9358	0,9333	128	0,322	0,0049	2,455E-06	4	GELU	Adam
135	0,9352	0,9161	128	0,317	0,0062	3,288E-06	4	GELU	Adam
112	0,9352	0,9087	128	0,312	0,0053	3,079E-06	4	GELU	Adam
187	0,9349	0,9176	512	0,347	0,0058	3,789E-06	2	GELU	AdamW
57	0,9346	0,9032	128	0,394	0,0036	2,291E-06	8	GELU	AdamW

111:6242728616

163	0,9343	0,9324	128	0,287	0,0053	4,289E-06	4	GELU	Adam
191	0,9338	0,8994	512	0,356	0,0009	7,191E-06	2	GELU	AdamW
85	0,9334	0,8884	128	0,303	0,0060	4,017E-06	2	GELU	Adam
87	0,9332	0,8973	128	0,305	0,0067	3,425E-06	2	GELU	Adam
133	0,9332	0,9332	128	0,299	0,0055	4,675E-06	4	GELU	Adam
148	0,9329	0,9264	128	0,302	0,0032	3,691E-06	4	GELU	Adam
143	0,9328	0,9119	128	0,292	0,0055	4,409E-06	4	GELU	Adam
95	0,9325	0,9108	128	0,278	0,0058	3,824E-06	4	GELU	Adam
184	0,9325	0,9280	512	0,329	0,0052	5,546E-06	2	GELU	AdamW
114	0,9324	0,9166	128	0,314	0,0041	4,450E-06	4	GELU	Adam
139	0,9323	0,9016	128	0,332	0,0043	1,403E-05	4	GELU	Adam
89	0,9321	0,9226	128	0,291	0,0066	2,803E-06	4	GELU	Adam
103	0,9321	0,9156	128	0,319	0,0038	5,824E-06	4	GELU	Adam
37	0,9320	0,9365	128	0,312	0,0067	3,332E-06	16	GELU	Adam
128	0,9318	0,9283	128	0,311	0,0077	1,691E-06	4	GELU	Adam
91	0,9318	0,9198	128	0,287	0,0085	5,753E-06	2	GELU	Adam

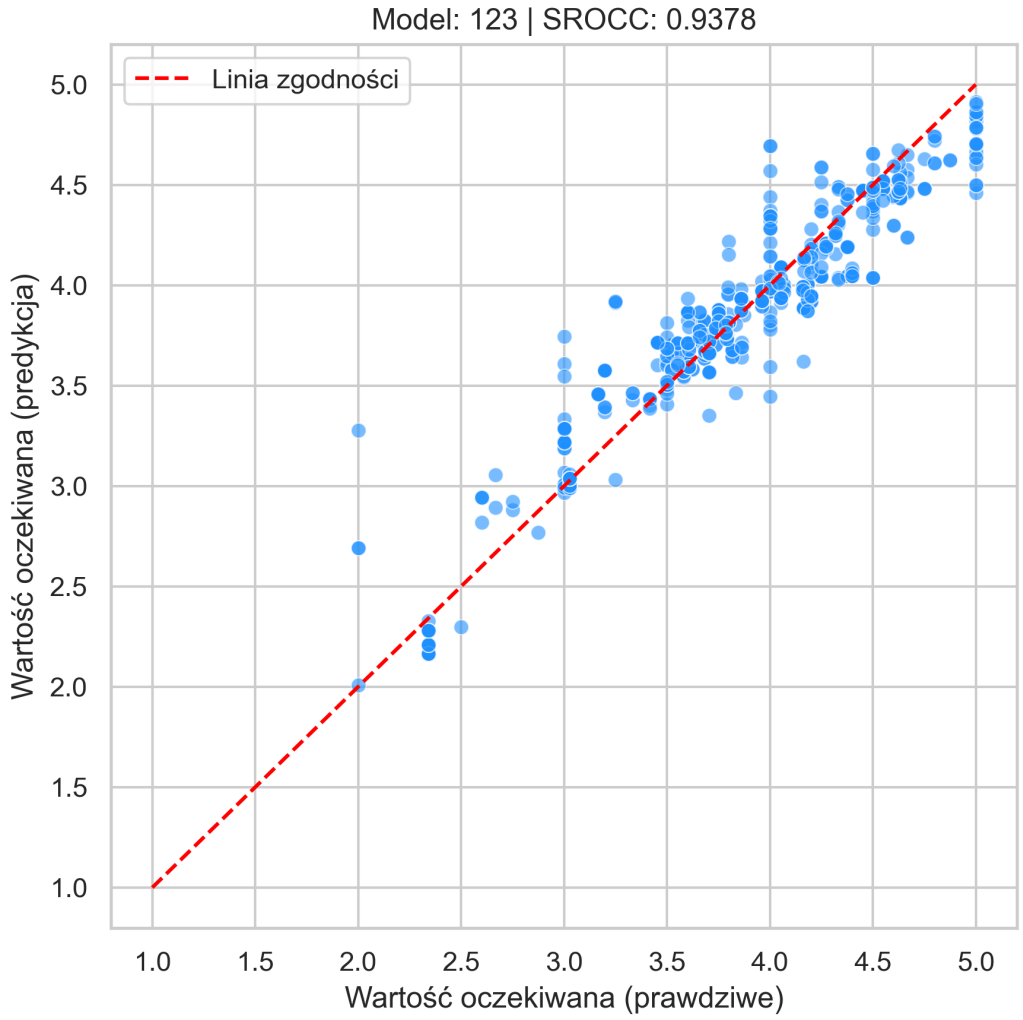
Druga faza obejmowała wytrenowanie 50-ciu najlepszych modeli spośród 192 algorytmów trenowanych w trakcie pierwszej fazy. W kolumnie *SROCC 2* tabeli 5.2 zawarto uśrednioną wartość współczynnika korelacji Spearmana dla 10 iteracji uczenia każdego modelu realizowanego w ramach 10-cio krotnej walidacji krzyżowej.
W tej fazie najwyższe wartości współczynnika SROCC uzyskano dla prób 123, 37 oraz 134:

- Model 123 – SROCC: 0,9378,
- Model 37 – SROCC: 0,9365,
- Model 134 – SROCC: 0,9333.

Wartość SROCC wynosiła powyżej 0,93 dla każdej z prób. Wspólne hiperparametry to liczba wejść warstwy ukrytej wynosząca $N = 128$, funkcja aktywacji *GELU* oraz optymalizator *Adam*. Model z próby 37 korzystał z 16 próbek w jednym kroku propagacji, natomiast pozostałe operowały na 4 próbkach. W kolejnym podrozdziale przedstawiono szczegółowe wyniki dla modeli z wybranych prób.

112:1156080170

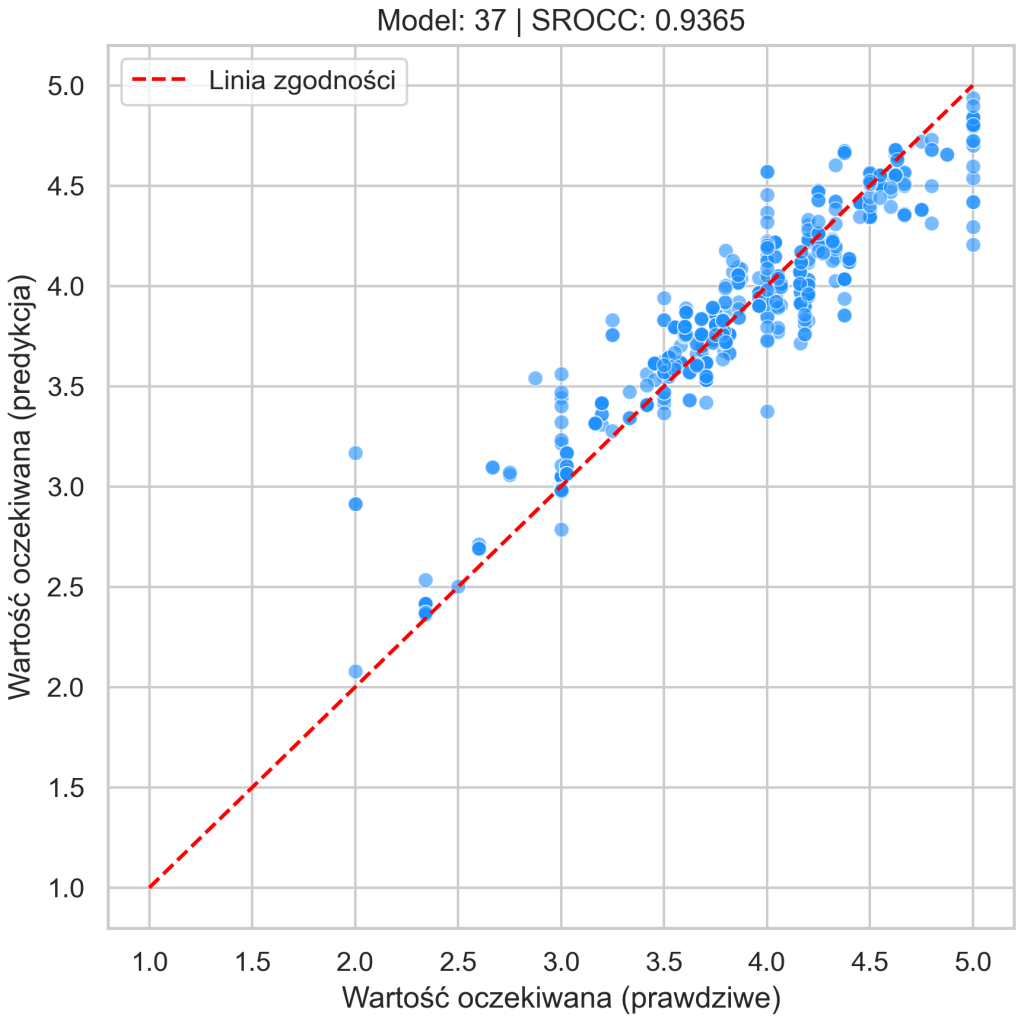
5.2 Wyniki dla modeli z jedną warstwą ukrytą



Rysunek 5.1 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Zoptymalizowany model z pojedynczą warstwą ukrytą z próby 123

Model z próby 123 uzyskał SROCC na poziomie 0,9378. Na wykresie 5.1 przedstawiono rozproszenie wartości oczekiwanych dla wektorów wartości prawdziwych i predykcji. Widać na nim, że wartości szacowane nieco odstają od czerwonej linii zgodności, jednak skupienie punktów jest mniej rozproszone niż dla modeli z poprzedniego rozdziału.

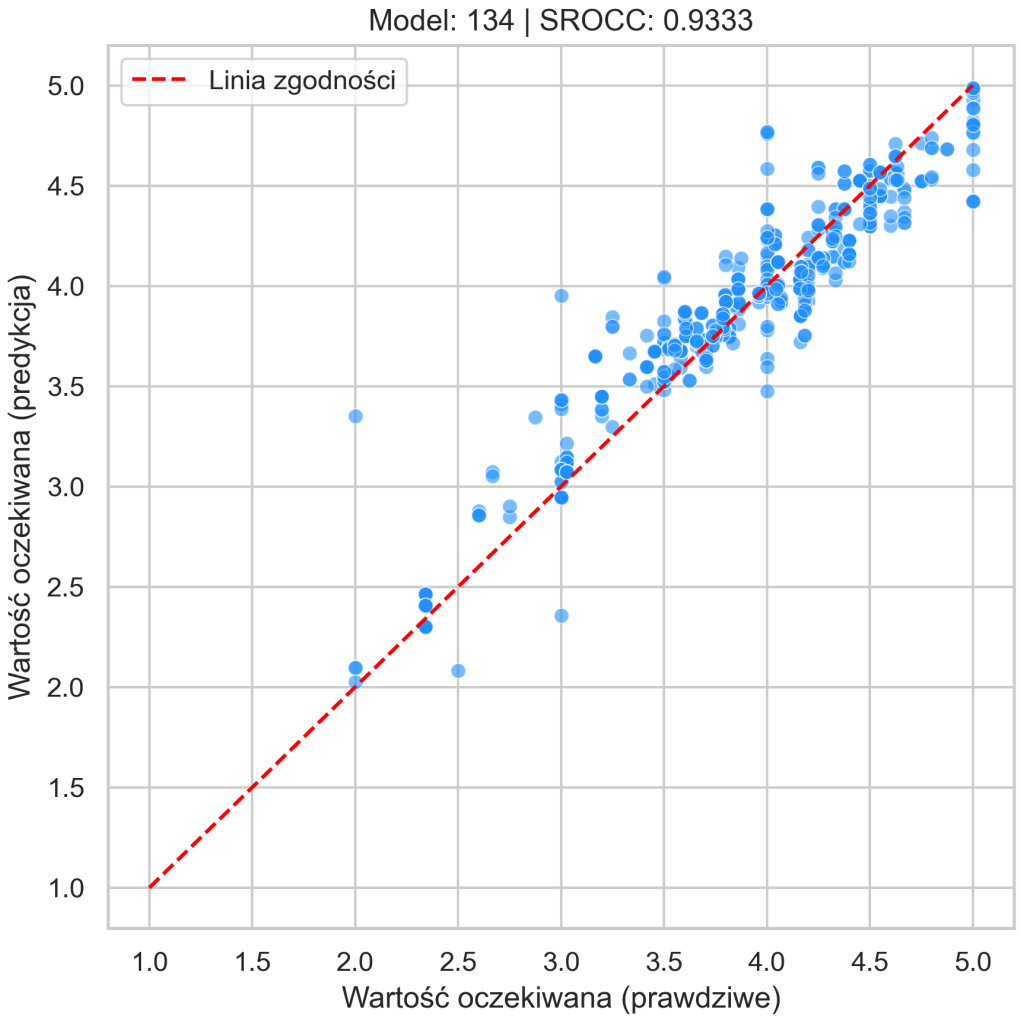
113:1132753928



Rysunek 5.2 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Zoptymalizowany model z pojedynczą warstwą ukrytą z próby 37

Na wykresie 5.2 przedstawiono rozproszenie dla modelu z próby 37 i drugim najwyższym wskaźniku SROCC, wynoszącym 0,9365. Rozproszenie predykcji wartości oczekiwanych jest zbliżone do poprzedniego modelu. Model ten nieco gorzej radzi sobie z najwyższymi ocenami.

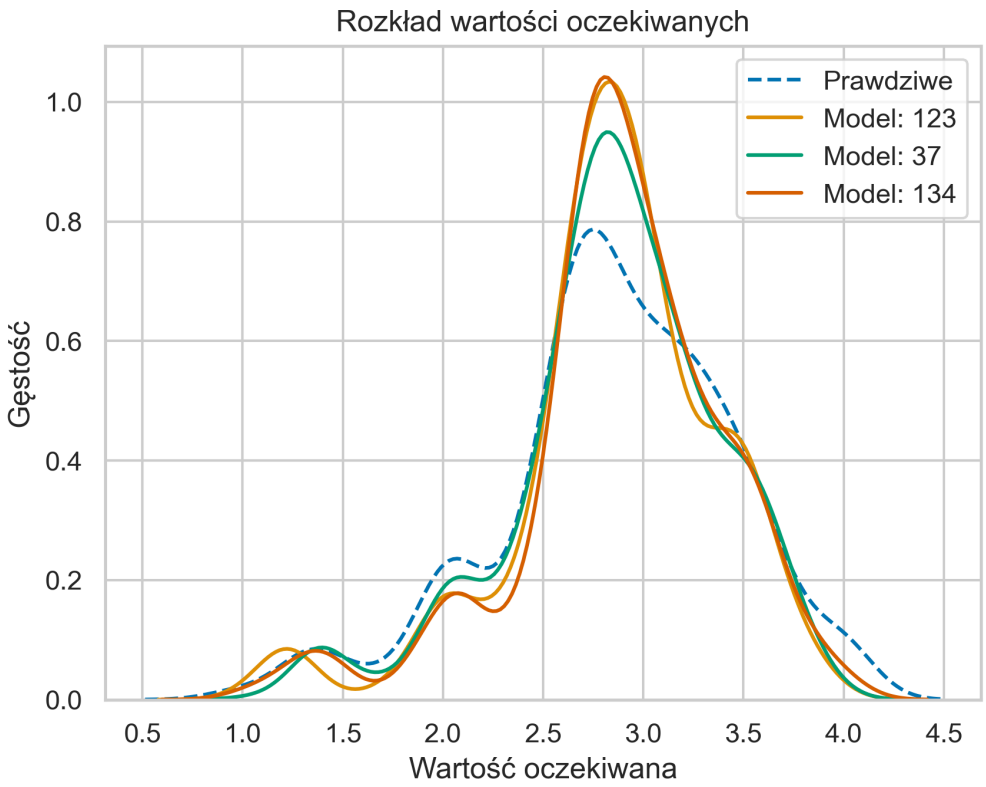
114:1015313379



Rysunek 5.3 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Zoptymalizowany model z pojedynczą warstwą ukrytą z próby 134

Wykres rozproszenia dla trzeciego modelu z próby 134 o wartości SROCC 0,9333 przedstawiono na wykresie 5.3. Model ten ma niższą skuteczność predykcji dla ocen w okolicy oceny 4.

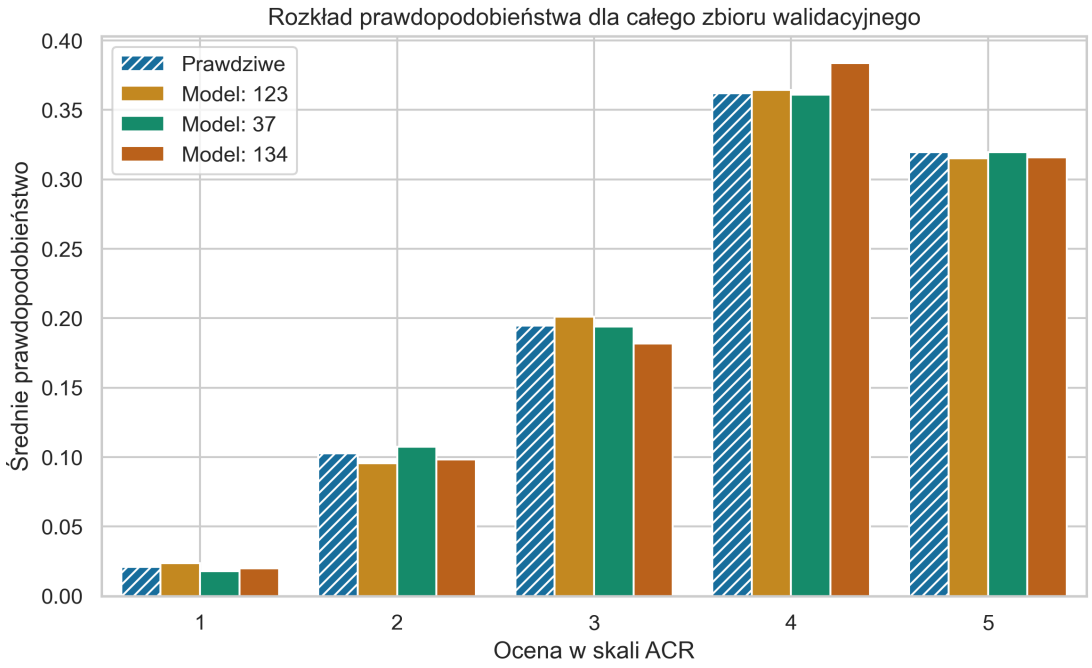
115:5812493312



Rysunek 5.4 Gęstość rozkładu wartości oczekiwanych dla zoptymalizowanych modeli z pojedynczą warstwą ukrytą

Na wykresie gęstości rozkładu wartości oczekiwanych 5.4 widać, że charakterystyki są do siebie podobne. Rozkład wartości oczekiwanych dla rzeczywistych prawdopodobieństw zaznaczono linią przerywaną. Gestości dla modeli zoptymalizowanych zbliżyły się do gęstości wartości oczekiwanych rzeczywistych rozkładów. Nadal widoczne jest przeszacowanie dla zakresu od 2,0 do 3,5 jednak jest ono niższe niż dla modeli z poprzedniego rozdziału. Obecne jest także niedoszacowanie ocen w okolicach wartości oczekiwanej wynoszącej 4,0.

116:1138517240



Rysunek 5.5 Rozkład uśrednionego prawdopodobieństwa każdej klasy dla zoptymalizowanych modeli z pojedynczą warstwą ukrytą

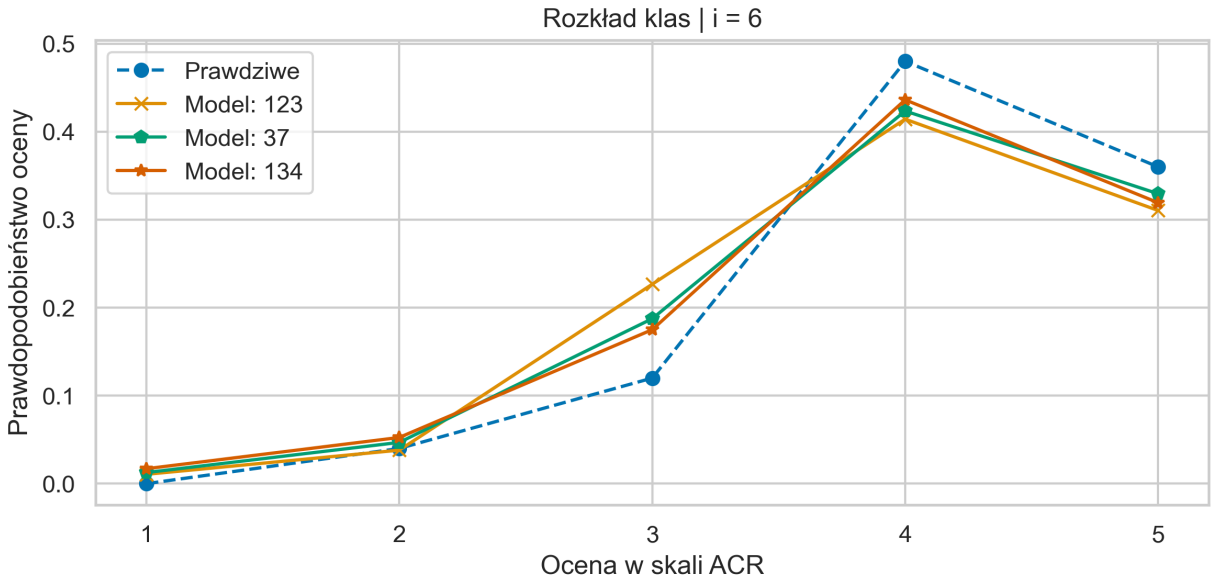
Rozkład uśrednionego prawdopodobieństwa każdej klasy przedstawiono na wykresie 5.5. Przeciętnie modele dość dobrze odwzorowują prawdopodobieństwo etykiet dla rzeczywistych ocen. Widoczny jest brak zbalansowania etykiet spowodowany subiektywnością danych. Modele potrafią odwzorować niezbalansowanie danych i nie dążą do uśredniania wyników, co jest oczekiwanym efektem. Gdyby rozkłady uśrednionego prawdopodobieństwa dla predykcji były mniej więcej na takim samym poziomie dla każdej oceny, oznaczałoby to brak decyzyjności modeli.

Na wykresach 5.6 - 5.11 przedstawiono wyniki predykcji dla przykładowych próbek (takich samych jak w rozdziale poprzednim). Każda próbka składa się z sekwencji 2 filmów po 30 sekund. Wykresy przedstawiono w formie liniowej ze względu na lepszą czytelność. W tabelach 5.3 - 5.8 zawarto podsumowanie parametrów próbek oraz skojarzone wektory prawdopodobieństw [p_1 - p_5]. Pierwszy wiersz *numer próbki - true* w każdej tabeli zawiera właściwości danej próbki oraz rzeczywiste wartości prawdopodobieństw. Kolejne wiersze to predykcja modeli z próby (ang. *trial*). Przykładowo wiersz *6 - 123* oznacza wyniki predykcji dla próbki $i = 6$ przez model z próby 123. Nagłówki tabel przedstawione jako akronimy objaśniono poniżej:

- i - trial – numer próbki oraz modelu, dla przypadku *i - true* są to wartości prawdziwe,
- c_a – current_assessment - subiektywna ocena w skali od 1 do 100 aktualnie odtwarzanego materiału wideo (materiał B z pary A i B),
- p_a – previous_assessment - subiektywna ocena w skali od 1 do 100 poprzednio odtwarzanego materiału wideo (materiał A z pary A i B),
- c_b – current_bitrate - przepływność aktualnie odtwarzanego materiału wideo,

117:2644403323

- p_b – previous_bitrate - przepływność poprzednio odtwarzanego materiału wideo,
- cdc – codec - typ kodeka,
- t_v_s – zakres rozmiaru ekranu TV w calach,
- a_r – zakres wieku oglądającego,
- p_1, p_2, p_3, p_4, p_5 – wektor prawdopodobieństw przydzielenia danej oceny w skali ACR.



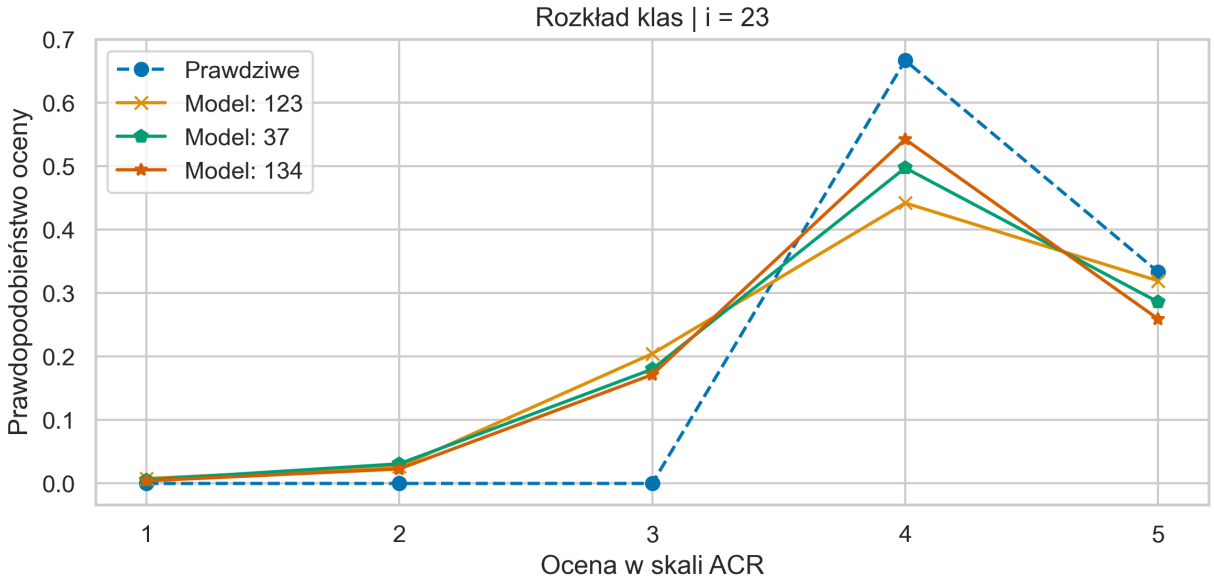
Rysunek 5.6 Wyniki predykcji dla próbki walidacyjnej i = 6 dla zoptymalizowanych modeli z pojedynczą warstwą ukrytą

Tabela 5.3 Parametry dla próbki $i = 6$

i – trial	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
6 – true	79	79	3298	5342	H.265	31 - 42	18 - 25	0,000	0,040	0,120	0,480	0,360
6 – 123	–	–	–	–	–	–	–	0,011	0,038	0,227	0,414	0,310
6 – 37	–	–	–	–	–	–	–	0,013	0,047	0,187	0,424	0,329
6 – 134	–	–	–	–	–	–	–	0,017	0,053	0,175	0,436	0,319

Dla próbki $i = 6$ rzeczywistą oceną byłoby 4. Na wykresie 5.6 oraz w tabeli 5.3 widoczna jest zgodność modeli co do kolejności prawdopodobieństw ocen. Dla ocen 3-5 istnieje nieznaczny błąd predykcji. Ocena 3 została nieznacznie przeszacowana, natomiast wartości 4 i 5 są niedoszacowane.

118:4509142865



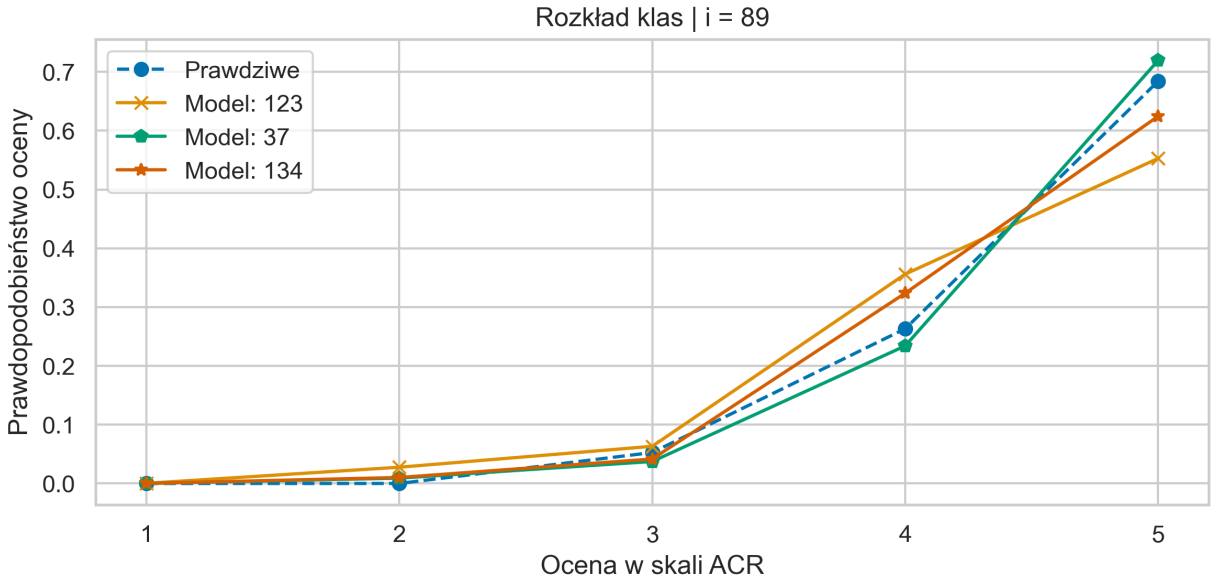
Rysunek 5.7 Wyniki predykcji dla próbki walidacyjnej i = 23 dla zoptymalizowanych modeli z pojedynczą warstwą ukrytą

Tabela 5.4 Parametry dla próbki $i = 23$

i – trial	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
23 – true	68	90	3000	6000	H.264	> 56	40 - 59	0,000	0,000	0,000	0,667	0,333
23 – 123	–	–	–	–	–	–	–	0,008	0,026	0,204	0,442	0,319
23 – 37	–	–	–	–	–	–	–	0,006	0,031	0,180	0,497	0,286
23 – 134	–	–	–	–	–	–	–	0,004	0,023	0,171	0,542	0,259

Predykcja dla próbki $i = 23$ widoczna na wykresie 5.7 oraz w tabeli 5.4 pokazuje, że podobnie jak poprzednio, modele prawidłowo przydzieliły najwyższe prawdopodobieństwo ocenie 4. Widoczne jest mocne przeszacowanie oceny 3 oraz niedoszacowanie 4. Rzeczywiste wartości prawdopodobieństw są niezerowe tylko dla próbek 4 i 5. Predykcja jest bardziej rozmyta, gdyż każda ocena ma niezerowe prawdopodobieństwo.

119:1401369613



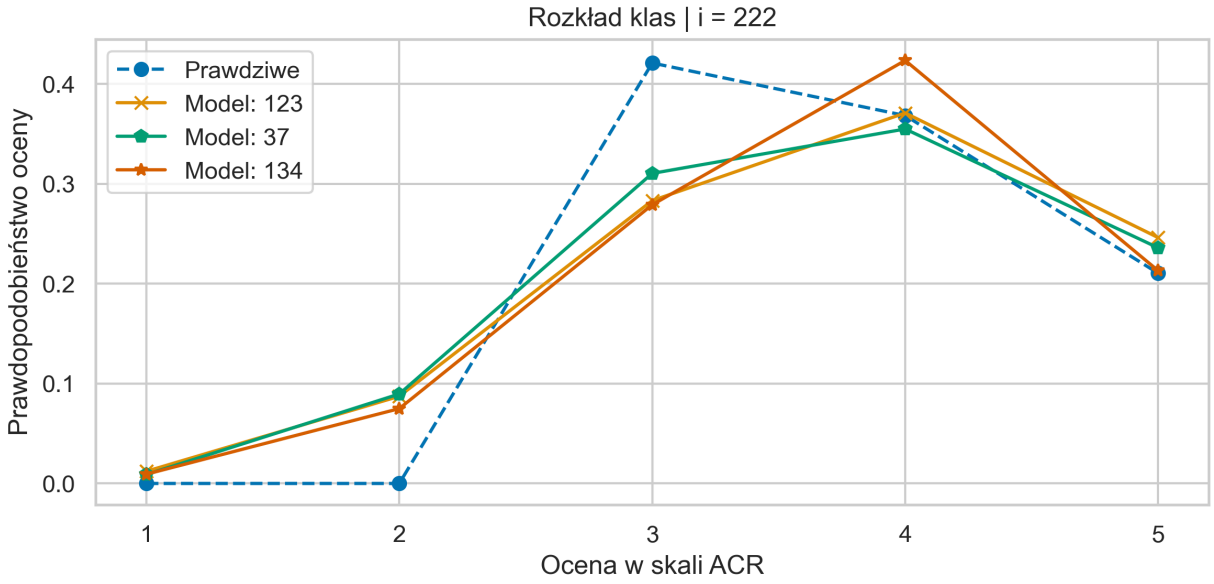
Rysunek 5.8 Wyniki predykcji dla próbki walidacyjnej i = 89 dla zoptymalizowanych modeli z pojedynczą warstwą ukrytą

Tabela 5.5 Parametry dla próbki $i = 89$

i – trial	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
89 – true	85	75	10000	1500	H.264	43 - 55	26 - 39	0,000	0,000	0,053	0,263	0,684
89 – 123	–	–	–	–	–	–	–	0,000	0,028	0,063	0,356	0,553
89 – 37	–	–	–	–	–	–	–	0,000	0,009	0,037	0,234	0,720
89 – 134	–	–	–	–	–	–	–	0,000	0,010	0,042	0,324	0,625

Wartości dla próbki $i = 89$ widoczne na wykresie 5.8 oraz w tabeli 5.5 wskazują celne odwzorowanie predykcji przez wszystkie modele. Wszystkie predykcje dla oceny 1 są zerowe. Model z próby 37 odwzorował rzeczywiste prawdopodobieństwa z minimalnym błędem. Modele z prób 123 i 134 wskazały ocenę 5 z najwyższym prawdopodobieństwem, jednak widoczne jest nideoszacowanie tej oceny kosztem oceny niższej.

120:4565667724



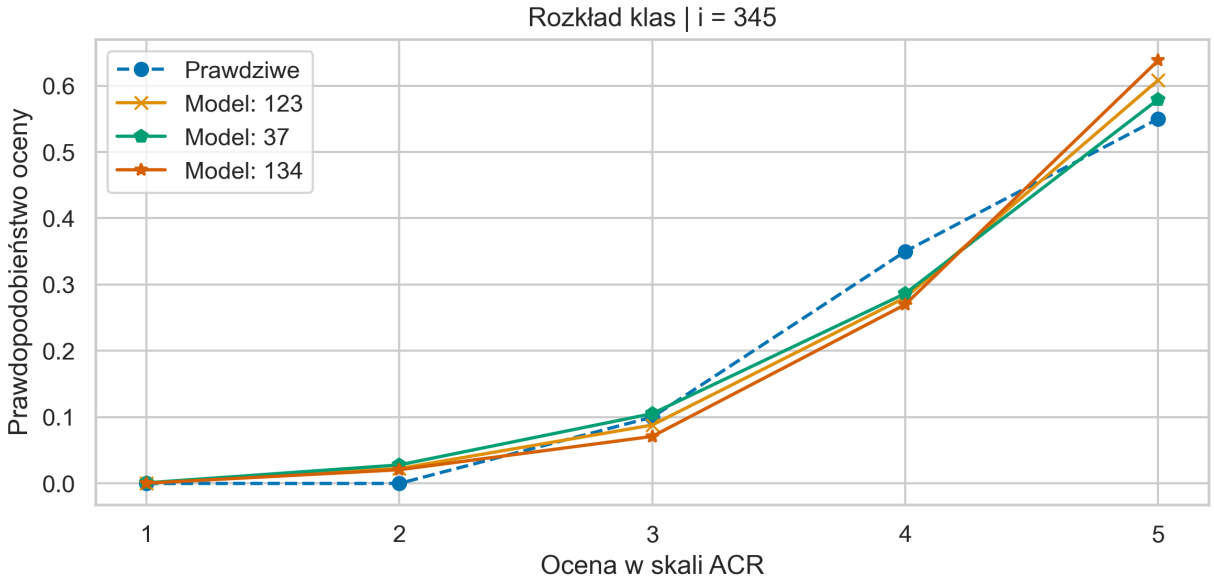
Rysunek 5.9 Wyniki predykcji dla próbki walidacyjnej i = 222 dla zoptymalizowanych modeli z pojedynczą warstwą ukrytą

Tabela 5.6 Parametry dla próbki $i = 222$

i – trial	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
222 – true	63	60	6000	1500	H.264	43 - 55	26 - 39	0,000	0,000	0,421	0,368	0,211
222 – 123	–	–	–	–	–	–	–	0,012	0,087	0,283	0,371	0,246
222 – 37	–	–	–	–	–	–	–	0,009	0,090	0,311	0,355	0,236
222 – 134	–	–	–	–	–	–	–	0,009	0,075	0,279	0,424	0,213

Parametry oraz predykcje dla próbki $i = 222$ widoczne na wykresie 5.9 i w tabeli 5.6 wskazują na błędne oszacowanie oceny przez wszystkie modele. Subiektywne oceny wskazują na najbardziej prawdopodobną ocenę o wartości 3, następnie 4 i 5. Oceny 1 i 2 są wyzerowane. Modele wskazują najwyższe prawdopodobieństwo dla oceny 4. Oceny 3 i 5 są odpowiednio drugie i trzecie w kolejności.

121:1047113352



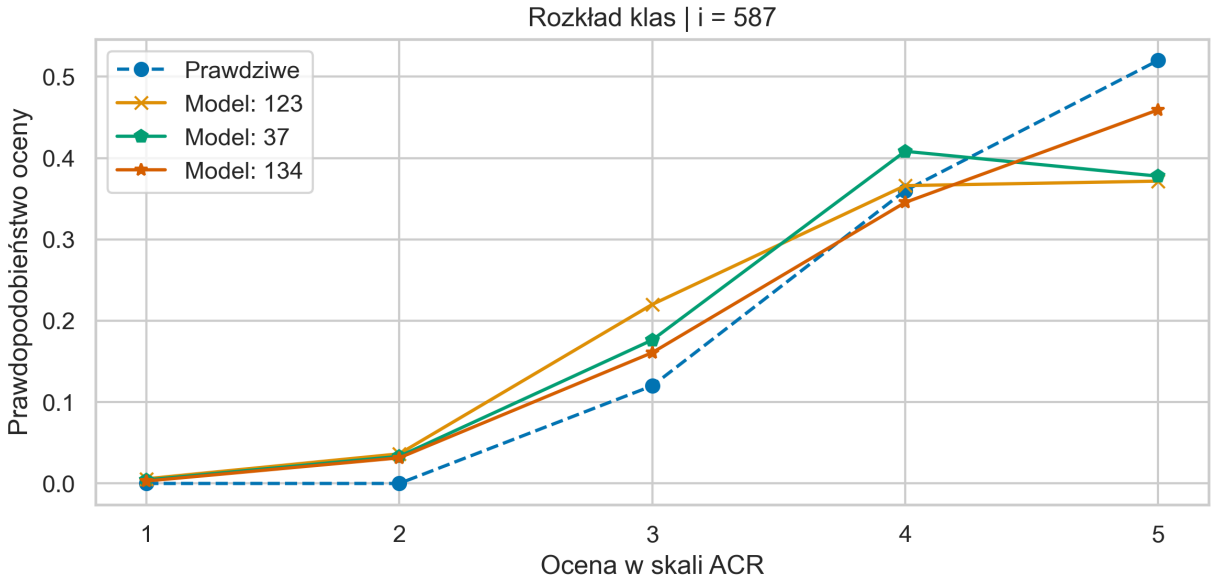
Rysunek 5.10 Wyniki predykcji dla próbki walidacyjnej i = 345 dla zoptymalizowanych modeli z pojedynczą warstwą ukrytą

Tabela 5.7 Parametry dla próbki $i = 345$

i – trial	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
345 – true	77	88	3298	6364	H.265	43 - 55	26 - 39	0,000	0,000	0,100	0,350	0,550
345 – 123	–	–	–	–	–	–	–	0,001	0,023	0,088	0,280	0,608
345 – 37	–	–	–	–	–	–	–	0,001	0,028	0,105	0,286	0,579
345 – 134	–	–	–	–	–	–	–	0,000	0,021	0,071	0,270	0,638

Oceny dla próbki $i = 345$ widoczne na wykresie 5.10 oraz w tabeli 5.7 są oszacowane prawidłowo. Monotoniczność wartości prawdopodobieństw jest taka sama jak dla próbki $i = 89$. Podobnie jak dla poprzedniej próbki widoczne są nieznaczne błędy predykcji dla ocen 4 i 5.

122:1014505920



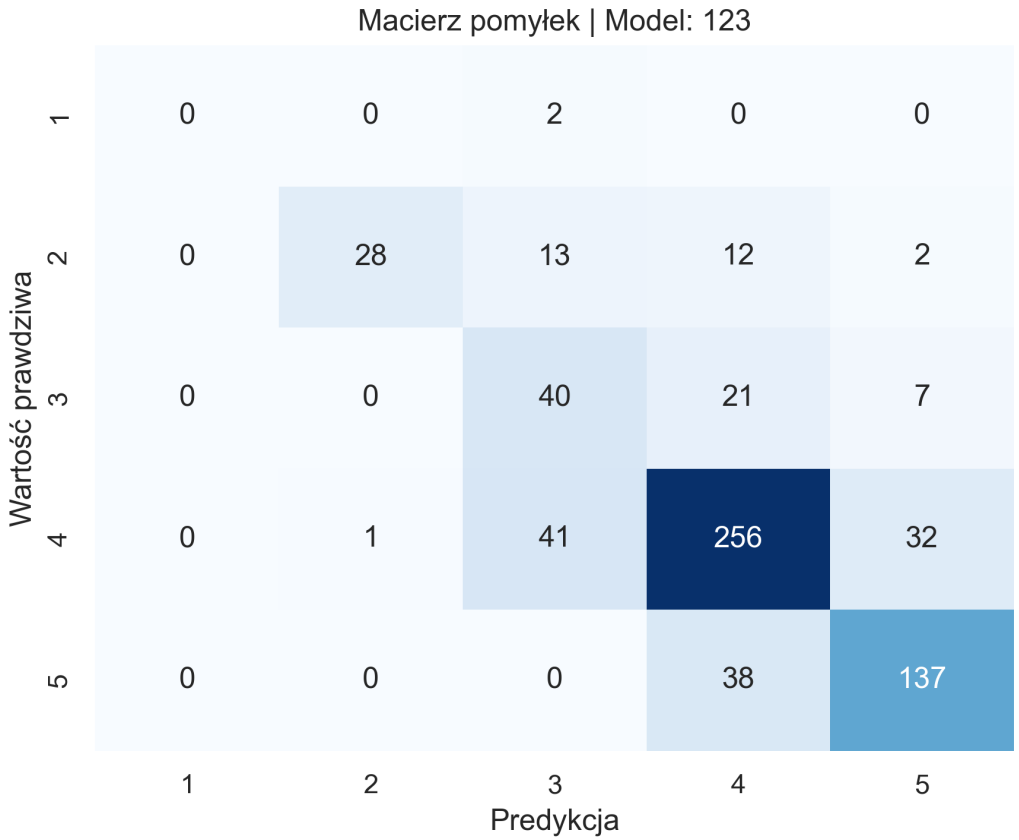
Rysunek 5.11 Wyniki predykcji dla próbki walidacyjnej i = 587 dla zoptymalizowanych modeli z pojedynczą warstwą ukrytą

Tabela 5.8 Parametry dla próbki $i = 587$

i – trial	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
587 – true	93	81	5342	2276	H.265	31 - 42	18 - 25	0,000	0,000	0,120	0,360	0,520
587 – 123	–	–	–	–	–	–	–	0,006	0,036	0,220	0,366	0,372
587 – 37	–	–	–	–	–	–	–	0,004	0,033	0,176	0,408	0,378
587 – 134	–	–	–	–	–	–	–	0,003	0,031	0,161	0,345	0,459

Dla próbki $i = 587$ tylko model z próby 134 jednoznacznie wskazał poprawną kolejność prawdopodobieństw. Na wykresie 5.11 oraz w tabeli 5.8, widać rozbieżność pomiędzy predykcjami. Model z próby 37 przydzieliłby ocenę 4. Model z próby 123 wskazał prawie takie same prawdopodobieństwo dla ocen 4 i 5, z minimalną przewagą dla oceny 5. Wartość różnicy wskazuje, że model może nie być pewny decyzji dla tej próbki.

123:1311017954



Rysunek 5.12 Macierz pomyłek dla zoptymalizowanego modelu z pojedynczą warstwą ukrytą z próby 123

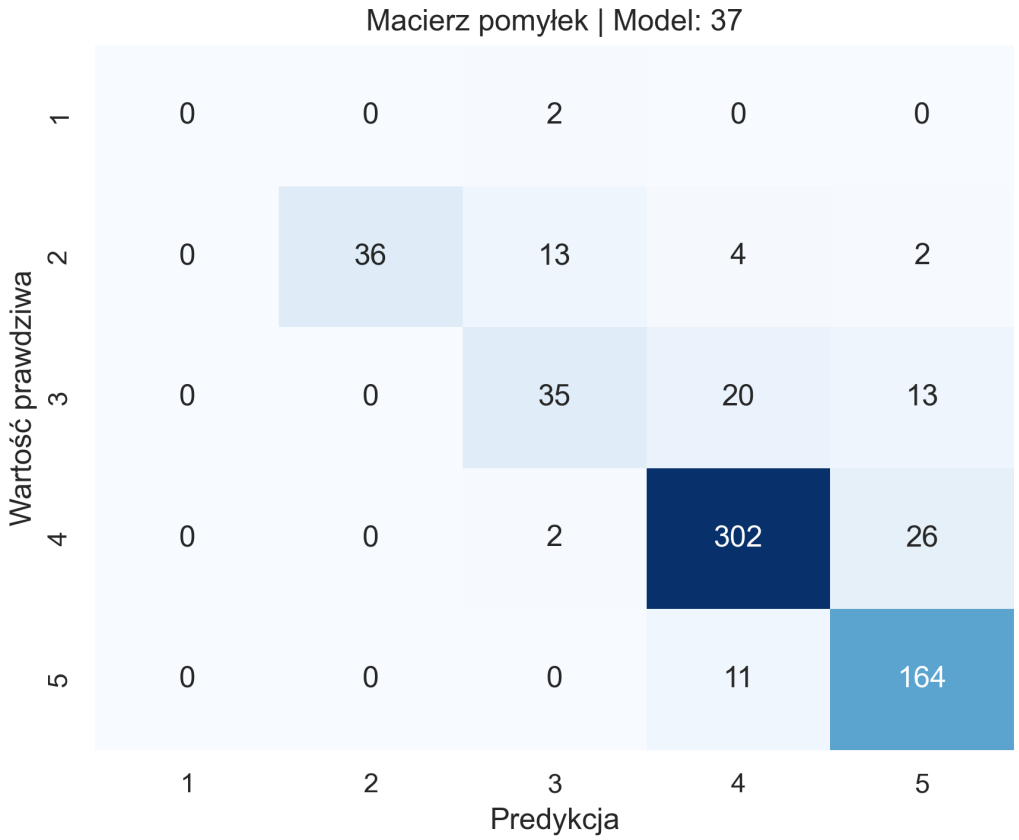
Tabela 5.9 Metryki dla zoptymalizowanego modelu z pojedynczą warstwą ukrytą z próby 123

Model	SROCC	Accuracy	Precision	Recall	F1
Model 123	0,938	0,732	0,587	0,531	0,542

Podobnie jak w poprzednim rozdziale, macierze pomyłek zostały wygenerowane bazując na ocenach z najwyższym prawdopodobieństwem czyli $argmax\, y(i)$. Macierz pomyłek dla modelu z próby 123 została przedstawiona na wykresie 5.12. Model ten nie wskazał najniższej oceny dla żadnej próbki, zamiast tego, została wskazana ocena 3. Ocena 2, dla której 28 próbek zostało oznaczone prawidłowo jest często mylona z ocenami 3 i 4, odpowiednio 13 i 12 błędów. 40 przypadków zostało prawidłowo sklasyfikowanych jako ocena 3. Klasa ta została pomylona 21 razy z oceną 4 i 7 razy z oceną 5. W subiektywnym zbiorze jest najwięcej ocen o wartości 4. 256 razy została prawidłowo wskazana ocena 4. Dla tej klasy wystąpiło 41 pomyłek z oceną 3 i 32 z oceną 5. Ocena 5 została wskazana prawidłowo w 137 przypadkach, natomiast 38 razy przypisano 4. Z macierzy pomyłek widać, że model z próby 123 dość często myli oceny z klasami sąsiednimi. Ze względu na minimalną liczbę negatywnych ocen - ocena 1, model nie był w stanie nauczyć się rozpoznawać tej klasy.

124:8068095573

W tabeli 5.9 zawarto wartości dla metryk określających skuteczność modelu. Model z próby 123 znacząco dobrze odwzorowuje kolejności prawdopodobieństw w wektorach, na co wskazuje wartość współczynnika korelacji rang Spearmana. Pozostałe metryki tj. Accuracy, Precision, Recall i F1 są na dość niskim poziomie. Dokładność (Accuracy) wskazuje, że 72,3% próbek zostało przypisanych do właściwej klasy. Niska precyzja wynosząca 58,7% jest powiązana z wskazywaniem błędnie sąsiednich klas. Czułość (Recall) i średnia harmoniczna F1 na poziomie nieco powyżej 50% wskazuje, że model często zachowuje trend, jednak ma problemy z precyzyjnym wskazaniem prawidłowej oceny.



Rysunek 5.13 Macierz pomyłek dla zoptymalizowanego modelu z pojedynczą warstwą ukrytą z próby 37

Tabela 5.10 Metryki dla zoptymalizowanego modelu z pojedynczą warstwą ukrytą z próby 37

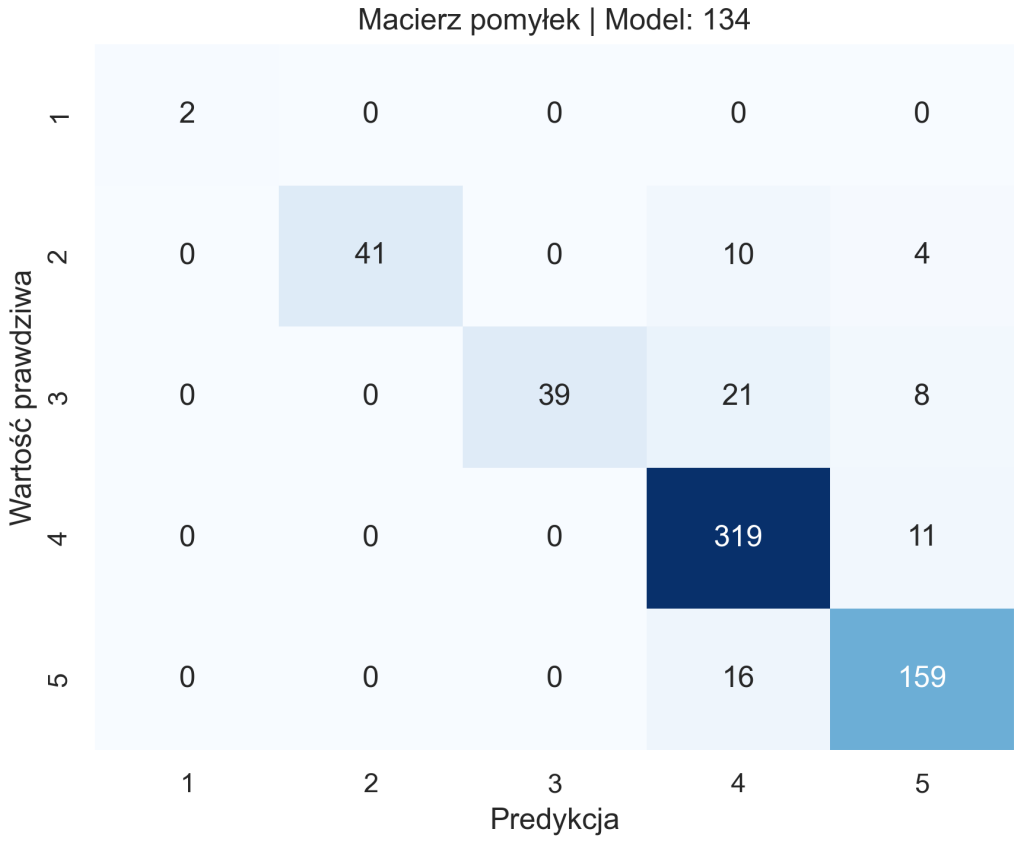
Model	SROCC	Accuracy	Precision	Recall	F1
Model 37	0,937	0,852	0,674	0,604	0,629

Macierz pomyłek dla modelu z próby 37 została przedstawiona na rysunku 5.13. Model, podobnie jak poprzedni, nie był w stanie rozpoznać próbek z najniższą oceną. Zostały one przypisane do oceny o wartości 3. Prawidłowe przypisanie dla klasy 2 wystąpiło 36 razy. Ocena 2 została sumarycznie 19 razy błędnie przypisana do wyższych klas. Dla tego przypadku model ma tendencję do zawyżania oceny. Środkowa ocena o wartości 3 została zidentyfikowana prawidłowo dla 35 przypadków. 20

125:5260253107

próbek zostało błędnie przypisanych do oceny 4, a 13 razy do oceny 5. W tym przypadku także istnieje tendencja do zawyżania ocen. Ocena 4 została poprawnie sklasyfikowana 302 razy. Grupa ta jest najliczniejsza i jest to najpewniej rozpoznawana klasa. Model przypisał błędnie ocenę o wartości 3 dwukrotnie, natomiast ocena o wartości 5 została błędnie przypisana 26 razy. Próbki w najwyższej klasie zostały sklasyfikowane poprawnie 164 razy. Klasa 4 została błędnie przypisana 11 razy. Model z próby 37 nie jest w stanie wykryć próbek w najniższej klasie, podobnie jak poprzedni model. Dla niższych klas model ma tendencję do zawyżania ocen.

W tabeli 5.10 zawarto wartości metryk oceniających skuteczność modelu. Współczynnik korelacji rang Spearmana na poziomie 0,937 wskazuje, że model bardzo dobrze zachowuje porządek między prawdopodobieństwami. Dokładność na poziomie 0,852 określa, że ponad 85% próbek zostało poprawnie sklasyfikowanych. Jest to wartość wyższa o około 12% w stosunku do modelu z próby 123. Precyzja modelu to niewiele ponad 67%, natomiast czułość to 60,4%. Poprawa względem poprzedniego modelu jest niewielka. Średnia harmoniczna precyzji i czułości wyniosła 62,9%. Model jest w stanie zachowywać porządek ocen, dokładność na poziomie 85,2% wskazuje, że większość próbek jest klasyfikowana poprawnie. Model ma niską precyzję i czułość, ze względu na pomyłki, szczególnie dla niższych klas.



Rysunek 5.14 Macierz pomyłek dla zoptymalizowanego modelu z pojedynczą warstwą ukrytą z próby 134

126:6963526661

Tabela 5.11 Metryki dla zoptymalizowanego modelu z pojedynczą warstwą ukrytą z próby 134

Model	SROCC	Accuracy	Precision	Recall	F1
Model 134	0,933	0,889	0,949	0,839	0,878

Dla modelu z próby 134 macierz pomyłek została przedstawiona na rysunku 5.14. Ocena 1 została prawidłowo wskazana dla obu próbek. Model był w stanie rozpoznać najniższą ocenę, pomimo bardzo niskiej reprezentacji subiektywnych ocen o wartości 1 w zbiorze danych. Ocena wyższa została sklasyfikowana prawidłowo 41 razy, błędnie 14 razy, każdy błąd to zawyżenie oceny. 39 próbek zostało poprawnie sklasyfikowanych dla środkowej oceny. Niestety, model wskazał błędną ocenę 29 razy. Zawyżając ją za każdym razem. Najliczniejsza klasa została prawidłowo wskazana 319 razy, tylko 11 razy została zawyżona do oceny wyższej. Najwyższa ocena została wskazana poprawnie 159 razy, a 16 razy ocena została niedoszacowana. Model z próby 134 jako jedyny do tej pory poprawnie sklasyfikował próbki z najniższą oceną. Nadal występuje problem dla klas 2 i 3, które są zawyżane. Model bardzo dobrze radzi sobie z predykcją wyższych ocen.

Metryki dla modelu z próby 134 zostały przedstawione w tabeli 5.11. Podobnie jak dla poprzednich modeli, współczynnik SROCC jest powyżej wartości 0,93. Dokładność na poziomie 0,899 to wartość wyższa niż dla dwóch poprzednich modeli. Precyzja o wartości 0,949, oznacza, że model zazwyczaj prawidłowo wskazuje ocenę. Jest to znacznie wyższa wartość w odniesieniu do poprzednich modeli. Czułość na poziomie 0,839 oznacza, że prawie 84% próbek zostało prawidłowo odnalezionych. Pozostałe oceny to błędne klasyfikacje powiązane z sąsiednimi klasami. Średnia harmoniczna F1 na poziomie 0,878 wskazuje, że model jest nieco bardziej precyzyjny i zbalansowany do predykcji ocen bazujących na danych subiektywnych niż poprzednie modele.

Pierwsza próba optymalizacji umożliwiła na uzyskanie nieco lepszych wyników dla SROCC niż wartości dla modeli bazowych. Modele z prób 123 i 37 uzyskały niezadowalający poziom wartości dokładności (ang. *Accuracy*), precyzji (ang. *Precision*), czułości (ang. *Recall*) oraz średniej harmonicznej F1. Trzeci wybrany model, z próby 134 uzyskał znacznie lepsze wyniki wskazanych metryk. Wartości zbliżone do 90% sugerują, że może istnieć przestrzeń do poprawy skuteczności modeli.

5.3 Optymalizacja dla modeli z wieloma warstwami ukrytymi

W tym podrozdziale opisano dalsze kroki optymalizacyjne dla modeli szacujących poziom QoE, bazujących na subiektywnym zbiorze danych. Modele nadal były optymalizowane pod kątem SROCC, tak jak to opisano w poprzednim rozdziale 5.1. Tym razem budowano jednak modele głębsze, z dodatkowymi warstwami ukrytymi. W tabeli 5.12 przedstawiono 4 rodzaje architektur testowanych modeli. Modele płytkie to jedna warstwa ukryta z liczbą wejść $N = 64$, modele średnie posiadają dwie warstwy ukryte, pierwsza z nich posiada liczbę wejść $N = 128$, natomiast druga $N = 64$. Modele głębokie posiadają trzy warstwy ukryte z liczbą wejść N wynoszącą odpowiednio $256, 128$ i 64. Modele bardzo głębokie posiadają pięć warstw ukrytych, z liczbą wejść N wynoszącą odpowiednio $512, 256, 128, 64, 32$. Pozostałe hiperparametry pozostały bez zmian zgodnie z tabelą 5.1.

127:9673446327

Tabela 5.12 Typy modeli testowanych w drugim podejściu optymalizacyjnym

Typ architektury	Liczba warstw	Liczba wejść kolejnych warstw ukrytych
Płytka	1	64
Średnia	2	128, 64
Głęboka	3	256, 128, 64
Bardzo głęboka	5	512, 256, 128, 64, 32

Algorytm 5: Optymalizacja sieci neuronowej dla modeli z różną liczbą warstw ukrytych

Dane wyjściowe: Lista prób: *trials_list*, Zbiór metryk prób: *trials_metrics*
trials_list ← []
trials_metrics ← []
Dla *trial* ← 1 *do* 250 **wykonaj**
  arch_options ← wybierz typ architektury z: [*płytka*, *średnia*, *głęboka*, *bardzo głęboka*]
  arch_dims ← wybierz liczbę wejść w zależności od wybranej architektury:
  dropout ← wybierz wartość dropout z zakresu: $0,1 - 0,5$
  learning_rate ← wybierz wartość współczynnika uczenia z zakresu: $0,0001 - 0,01$
  weight_decay ← wybierz wartość współczynnika regularyzacji L2 z zakresu: $0,000001 - 0,001$
  batch_size ← wybierz wartość liczby próbek z: $[2, 4, 8, 16, 32, 64]$
  activation_func ← wybierz funkcję aktywacji z: [*ReLU*, *LeakyReLU*, *Tanh*, *GELU*]
  optimizer ← wybierz optymalizator z: [*Adam*, *AdamW*, *SGD*, *RMSProp*]
  H ← przypisz do zbioru wybrane hiperparametry:
    $arch_options, arch_dims, dropout, learning_rate, weight_decay, batch_size, activation_func, optimizer$
  train_loader ← utwórz iterator zbioru treningowego
  validation_loader ← utwórz iterator zbioru walidacyjnego
  model ← utwórz instancję modelu bazując na hiperparametrach H
  trial_model, *trial_metric* ← wywołaj *Algorytm 3* z argumentami:
    model, *train_loader*, *validation_loader*
  trials_list ← dodaj do listy *trial_model*
  trials_metrics ← dodaj do listy *trial_metric*
Koniec
Zwróć *trials_list*, *trials_metrics*

Poszukiwanie optymalnego modelu zostało podzielone na dwie fazy podobnie jak w poprzednim przypadku. Główna pętla algorytmu 5 odpowiada za 250 prób optymalizacji wywołując algorytm 3 uczący modele wielowarstwowe. Podobnie jak w poprzednim przypadku podzielono zbiór danych na zbiory: treningowy, walidacyjny i testowy w proporcjach 3:1:1.

Zbiór 50-ciu najlepszych modeli został przedstawiony w tabeli 5.13. Akronimy w kolumnach tabeli zawierają hiperparametry modeli z 50-ciu najlepszych prób:

- *Trial* – próba optymalizacyjna,
- *L* – liczba wejść warstw ukrytych, wskazuje także na architekturę modelu,
- *DO* – wartość współczynnika dropout z zakresu 0,1 - 0,5,
- *LR* – wartość współczynnika uczenia z zakresu 0,0001 - 0,01,
- *WD* – wartość współczynnika regularyzacji L2 z zakresu 0,000001 - 0,001,

128:8523388868

- *BS* – liczba próbek w jednym kroku propagacji z zakresu 2 - 64,
- *Act* – funkcja aktywacji,
- *Opt* – zastosowany optymalizator.

Na zbiorze 50-ciu najlepszych modeli przeprowadzono 10-krotną walidację krzyżową zgodnie z algorytmem 4. Modele zostały wytrenowane ponownie, bez wcześniejszego zapisu stanu. Dla wytrenowanych modeli obliczono SROCC, dokładność, precyzję, czułość i średnią harmoniczną F1 dla wartości oczekiwanych rozkładu prawdopodobieństw rzeczywistych i predykcji. Wyniki zostały zawarte w tabeli 5.14. Modele w tabeli zostały posortowane malejąco według wartości SROCC.

Tabela 5.13 50 najlepszych modeli spośród 200 prób optymalizacji dla architektury wielowarstwowej. Modele zostały posortowane malejąco według uśrednionych metryk

Trial	L	DO	LR	WD	BS	Act	Opt
225	[128, 64]	0,141	0,0100	2,259E-05	4	ReLU	SGD
226	[128, 64]	0,141	0,0093	2,262E-05	4	ReLU	SGD
245	[128, 64]	0,145	0,0095	2,693E-05	4	ReLU	SGD
200	[128, 64]	0,126	0,0067	1,223E-05	4	ReLU	SGD
244	[128, 64]	0,145	0,0085	1,506E-05	4	ReLU	SGD
204	[128, 64]	0,144	0,0068	1,136E-05	4	ReLU	SGD
217	[128, 64]	0,141	0,0089	1,663E-05	4	ReLU	SGD
215	[128, 64]	0,140	0,0084	1,622E-05	4	ReLU	SGD
143	[512, 256, 128, 64, 32]	0,117	0,0053	1,243E-06	4	LeakyReLU	SGD
223	[128, 64]	0,138	0,0083	2,088E-05	4	ReLU	SGD
233	[128, 64]	0,139	0,0093	2,081E-05	4	ReLU	SGD
247	[128, 64]	0,153	0,0098	2,713E-05	4	ReLU	SGD
193	[128, 64]	0,116	0,0056	1,227E-05	4	ReLU	SGD
218	[128, 64]	0,142	0,0083	1,716E-05	4	ReLU	SGD
208	[128, 64]	0,145	0,0083	1,417E-05	4	ReLU	SGD
236	[128, 64]	0,162	0,0091	2,884E-05	4	ReLU	SGD
221	[128, 64]	0,143	0,0094	2,025E-05	4	ReLU	SGD
155	[128, 64]	0,134	0,0076	1,006E-06	4	ReLU	SGD
170	[128, 64]	0,115	0,0073	1,548E-06	4	ReLU	SGD
184	[128, 64]	0,123	0,0067	2,270E-05	4	ReLU	SGD
229	[128, 64]	0,159	0,0089	2,412E-05	4	ReLU	SGD
231	[128, 64]	0,159	0,0090	2,726E-05	4	ReLU	SGD
181	[128, 64]	0,115	0,0067	1,048E-05	4	ReLU	SGD
212	[128, 64]	0,127	0,0083	1,368E-05	4	ReLU	SGD
182	[128, 64]	0,122	0,0068	1,422E-05	4	ReLU	SGD

129:6764137341

10	[512, 256, 128, 64, 32]	0,116	0,0014	8,737E-06	16	LeakyReLU	Adam
214	[128, 64]	0,139	0,0084	1,608E-05	4	ReLU	SGD
216	[128, 64]	0,141	0,0082	1,597E-05	4	ReLU	SGD
199	[128, 64]	0,122	0,0067	1,722E-05	4	ReLU	SGD
224	[128, 64]	0,141	0,0095	2,352E-05	4	ReLU	SGD
163	[128, 64]	0,110	0,0075	1,147E-06	4	ReLU	SGD
213	[128, 64]	0,128	0,0068	1,673E-05	4	ReLU	SGD
173	[128, 64]	0,103	0,0063	1,554E-06	4	ReLU	SGD
201	[128, 64]	0,126	0,0055	1,822E-05	4	ReLU	SGD
227	[128, 64]	0,142	0,0098	2,146E-05	4	ReLU	SGD
171	[128, 64]	0,104	0,0064	1,567E-06	4	ReLU	SGD
140	[512, 256, 128, 64, 32]	0,115	0,0066	1,001E-06	4	LeakyReLU	SGD
183	[128, 64]	0,124	0,0068	1,496E-05	4	ReLU	SGD
205	[128, 64]	0,124	0,0067	1,184E-05	4	ReLU	SGD
159	[128, 64]	0,135	0,0075	1,009E-06	4	ReLU	SGD
195	[128, 64]	0,120	0,0056	1,163E-05	4	ReLU	SGD
230	[128, 64]	0,152	0,0087	2,643E-05	4	ReLU	SGD
246	[128, 64]	0,150	0,0100	2,532E-05	4	ReLU	SGD
142	[512, 256, 128, 64, 32]	0,109	0,0067	1,257E-06	4	LeakyReLU	SGD
207	[128, 64]	0,147	0,0063	1,452E-05	4	ReLU	SGD
172	[128, 64]	0,103	0,0061	1,525E-06	4	ReLU	SGD
222	[128, 64]	0,153	0,0084	1,894E-05	4	ReLU	SGD
234	[128, 64]	0,150	0,0083	2,002E-05	4	ReLU	SGD
188	[128, 64]	0,112	0,0068	1,273E-05	4	ReLU	SGD
180	[128, 64]	0,109	0,0049	1,564E-05	4	ReLU	SGD

Tabela 5.14 Wartości metryk zoptymalizowanych wielowarstwowych modeli po 10-krotnej walidacji krzyżowej. Wartości obliczone dla wartości oczekiwanej wartości prawdziwych i predykcji

Trial	SROCC 1	SROCC 2	Accuracy	F1	Recall
225	0,9839	0,9797	0,8667	0,6699	0,6432
226	0,9832	0,9786	0,9254	0,9091	0,8854
245	0,9821	0,9844	0,9222	0,7189	0,7010
200	0,9798	0,9809	0,8698	0,6524	0,6456
244	0,9793	0,9832	0,8984	0,6776	0,6633
204	0,9790	0,9841	0,9349	0,7459	0,7371
217	0,9785	0,9886	0,9143	0,7292	0,7208
215	0,9785	0,9885	0,8857	0,6572	0,6295

130:6830213648

143	0,9784	0,9875	0,8492	0,6181	0,5997
223	0,9784	0,9825	0,8746	0,6463	0,6242
233	0,9781	0,9851	0,8762	0,6527	0,6140
247	0,9780	0,9837	0,9381	0,7481	0,7396
193	0,9774	0,9784	0,9175	0,6723	0,6566
218	0,9773	0,9857	0,9270	0,7256	0,7138
208	0,9773	0,9793	0,9063	0,6980	0,6749
236	0,9772	0,9822	0,9190	0,9179	0,8904
221	0,9770	0,9833	0,8635	0,6302	0,6126
155	0,9769	0,9805	0,8952	0,6778	0,6795
170	0,9769	0,9825	0,8968	0,9102	0,8957
184	0,9760	0,9844	0,9000	0,6609	0,6366
229	0,9758	0,9786	0,8794	0,6549	0,6474
231	0,9757	0,9811	0,9317	0,7214	0,7033
181	0,9754	0,9858	0,8873	0,6681	0,6377
212	0,9752	0,9828	0,9111	0,7141	0,7014
182	0,9746	0,9853	0,9095	0,7128	0,6971
10	0,9743	0,9849	0,8794	0,7033	0,6933
214	0,9740	0,9840	0,8984	0,7007	0,6818
216	0,9734	0,9820	0,9127	0,7098	0,6887
199	0,9731	0,9790	0,9063	0,7015	0,6865
224	0,9729	0,9809	0,8952	0,7010	0,6840
163	0,9728	0,9850	0,9000	0,7155	0,7003
213	0,9727	0,9811	0,9048	0,8899	0,8704
173	0,9724	0,9804	0,8984	0,6732	0,6516
201	0,9724	0,9764	0,9413	0,7241	0,7110
227	0,9723	0,9857	0,8762	0,6852	0,6666
171	0,9718	0,9845	0,9222	0,7149	0,6947
140	0,9716	0,9884	0,8413	0,6408	0,6408
183	0,9707	0,9805	0,9063	0,7183	0,7069
205	0,9704	0,9797	0,8683	0,6447	0,6321
159	0,9704	0,9836	0,8889	0,6685	0,6559
195	0,9704	0,9812	0,9111	0,7200	0,7046
230	0,9704	0,9831	0,8857	0,7059	0,6933
246	0,9702	0,9850	0,8984	0,6589	0,6387
142	0,9700	0,9870	0,8571	0,6562	0,6431
207	0,9699	0,9757	0,8794	0,6534	0,6505

131:1066048461

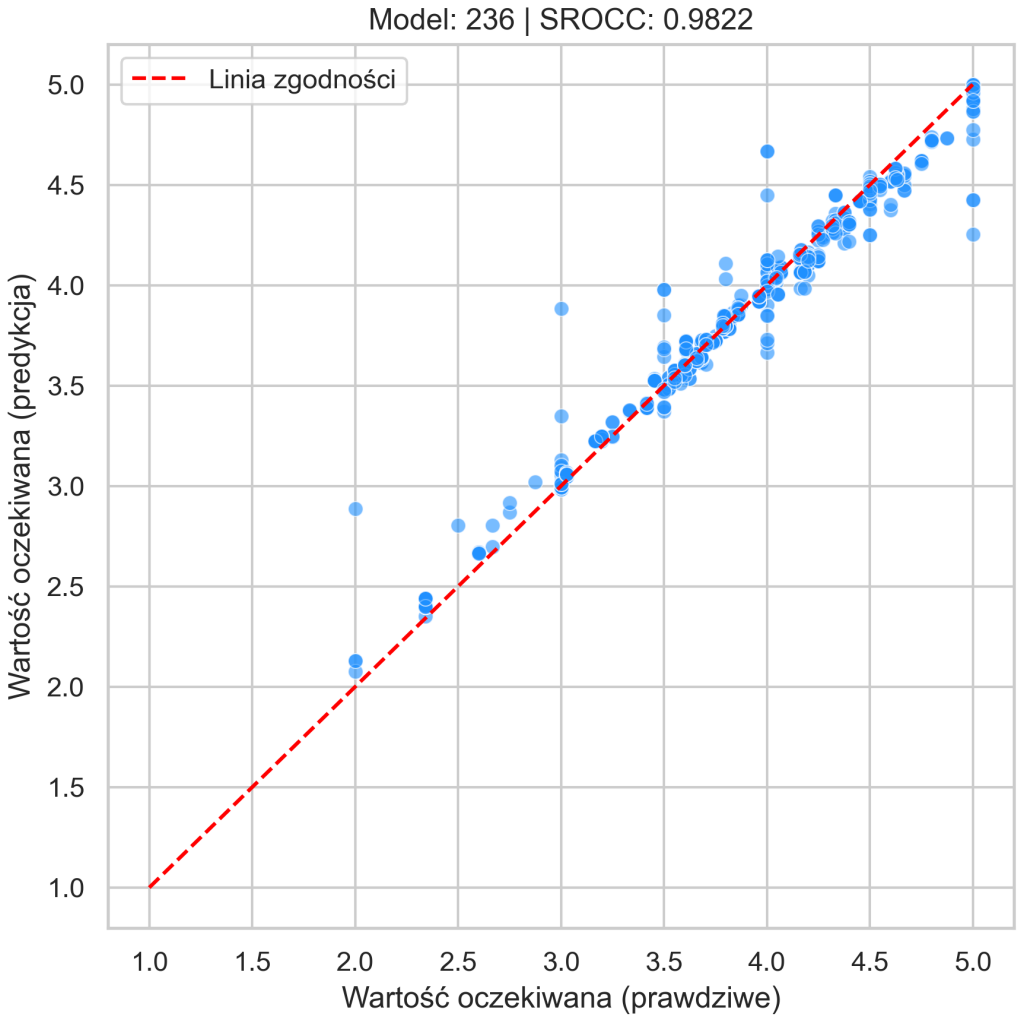
172	0,9697	0,9764	0,9302	0,7374	0,7272
222	0,9693	0,9862	0,8571	0,6501	0,6349
234	0,9692	0,9805	0,8587	0,6402	0,6269
188	0,9690	0,9820	0,9095	0,6707	0,6584
180	0,9687	0,9713	0,9127	0,7095	0,6876

Ostatecznie do dalszej analizy wybrano 3 modele. Wybór został dokonany bazując na złożoności modelu, wartości SROCC oraz średniej dokładności, precycji, czułości i średniej harmonicznej F1. Decyzja o wyborze jest powiązana z wynikami analizy zoptymalizowanych modeli jednowarstwowych. W poprzednim podejściu dwa modele o najwyższym SROCC nie były wystarczająco dokładne, precyzyjne i czułe. Model trzeci z próby 134 uzyskał nieco lepsze wyniki. Finalnie wytypowano trzy modele z prób: 236, 226 oraz 170. W kolejnym podrozdziale przedstawiono szczegółowe wyniki dla wybranych modeli.

- Model 236 – SROCC: 0,9822
- Model 226 – SROCC: 0,9786
- Model 170 – SROCC: 0,9825

132:8143596162

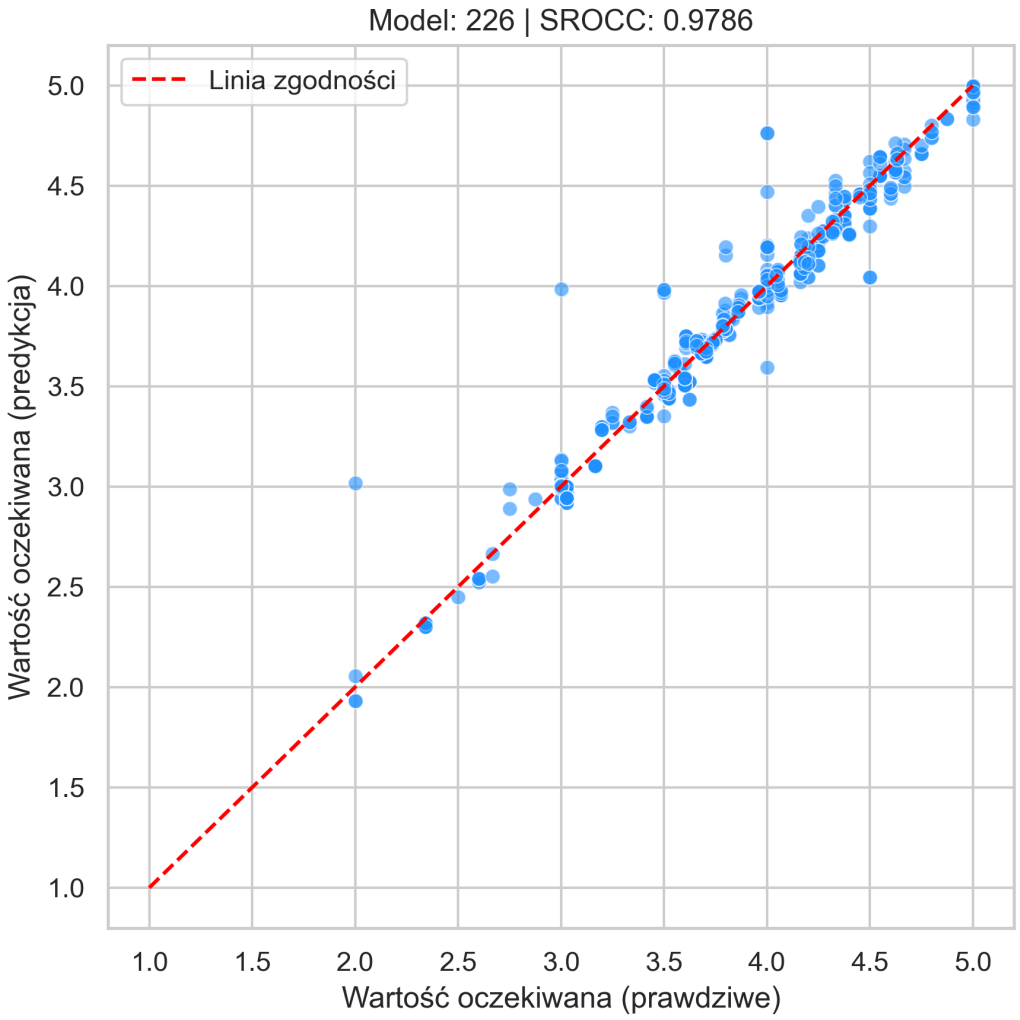
5.4 Wyniki dla modeli z wieloma warstwami ukrytymi



Rysunek 5.15 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Zoptymalizowany wielowarstwowy model z próby 236

Na wykresie 5.15 przedstawiono wykres rozproszenia wartości oczekiwanej dla modelu z próby 236. Model ten uzyskał SROCC na poziomie 0,9822. Na wykresie widać że predykcja nieco lepiej oddaje stan rzeczywisty w porównaniu do poprzednich modeli. Widoczne są pojedyncze elementy odstające, szczególnie dla dwóch najwyższych ocen.

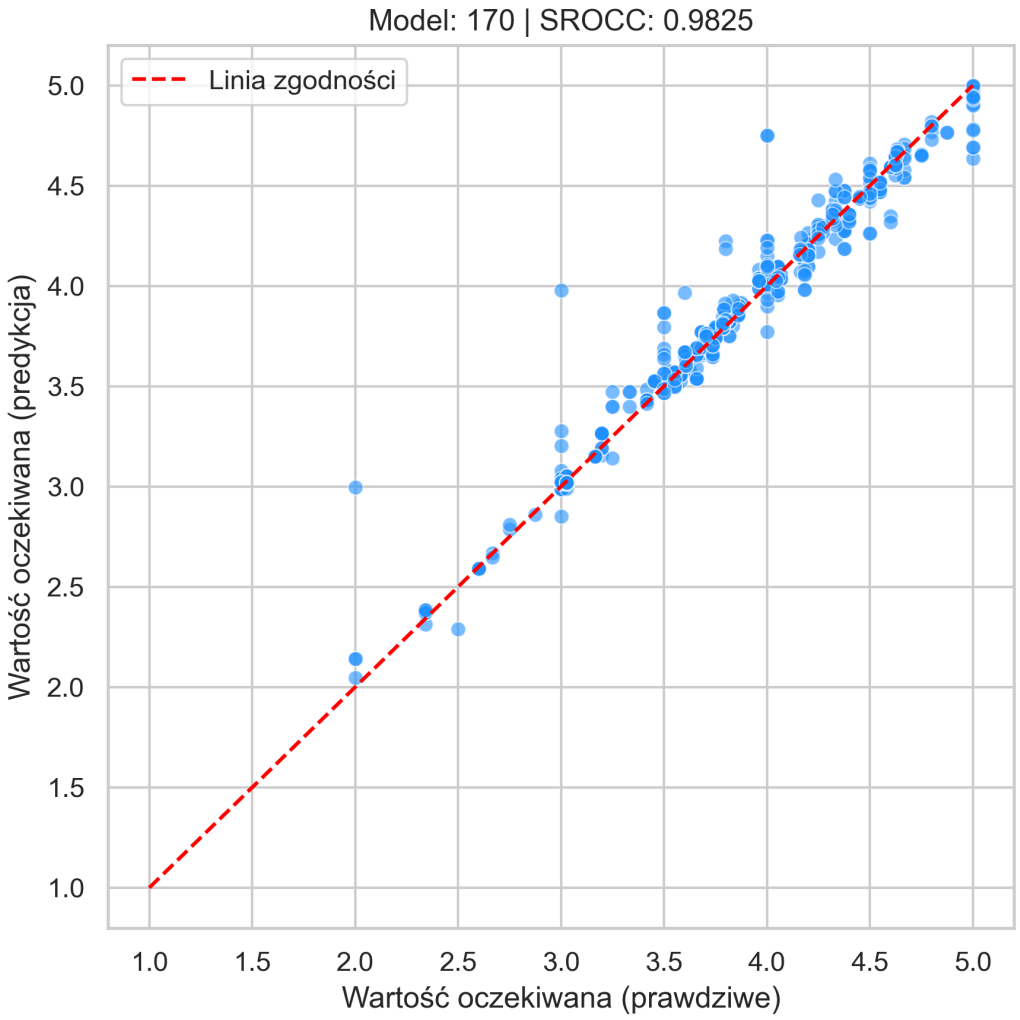
133:9710401272



Rysunek 5.16 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Zoptymalizowany wielowarstwowy model z próby 226

Rozproszenie wartości oczekiwanej dla modelu z próby 226 zostało przedstawione na wykresie 5.16. W przypadku tego modelu wartość oczekiwana z predykcji dla najwyższej oceny ma wyższą zgodność niż poprzedni model. Pojedyncze punkty, reprezentujące wartość oczekiwaną z predykcji są przesunięte w kierunku wyższej oceny. Model z próby 226 osiągnął SROCC na poziomie 0,9786.

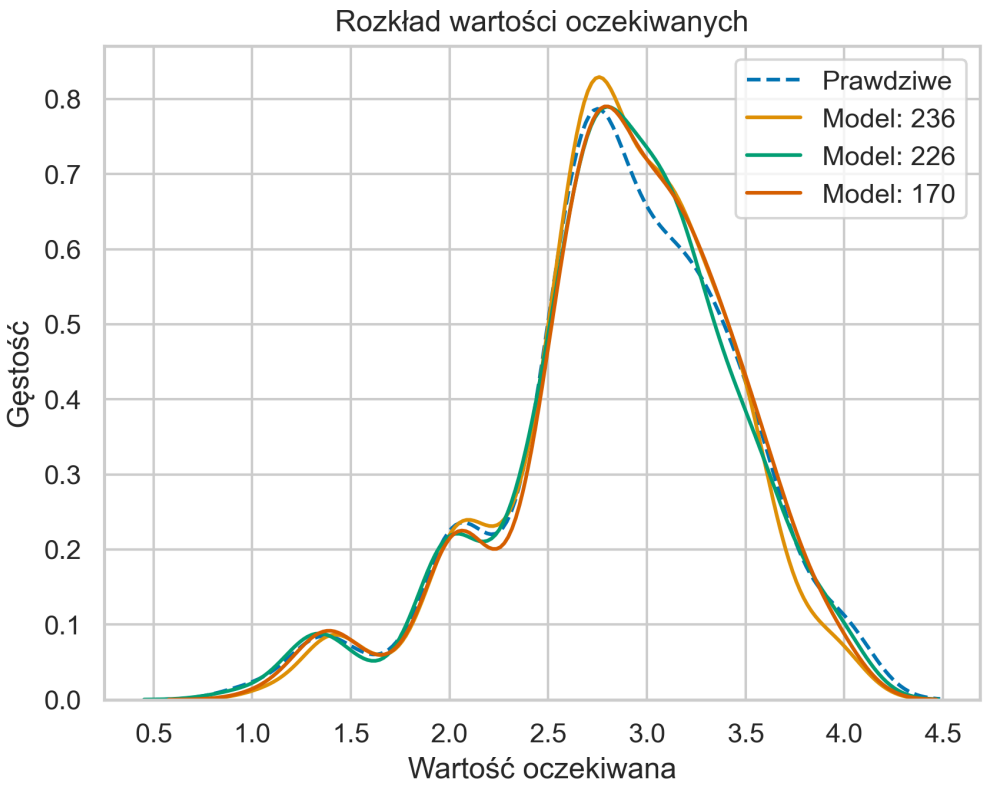
134:6931263320



Rysunek 5.17 Rozproszenie wartości oczekiwanych rozkładu ocen ACR dla wektorów wartości prawdziwych i predykcji. Zoptymalizowany wielowarstwowy model z próby 170

Model z próby 170 osiągnął SROCC o wartości wynoszącej 0,9825. Podobnie jak w przypadku poprzednich modeli widoczne są pojdyncze, przewidziane wartości oczekiwane, które są przesunięte w kierunku wyższej oceny. Wykresy rozproszenia wartości oczekiwanej dla modeli wielowarstwowych są bardziej spłaszczone dookoła linii zgodności niż wykresy dla modeli jednowarstwowych. Jeżeli pojawiają się widoczne punkty odstające to zazwyczaj są to pojedyncze wartości. Modele wielowarstwowe osiągnęły znacząco wyższy współczynnik korelacji rang Spearmana.

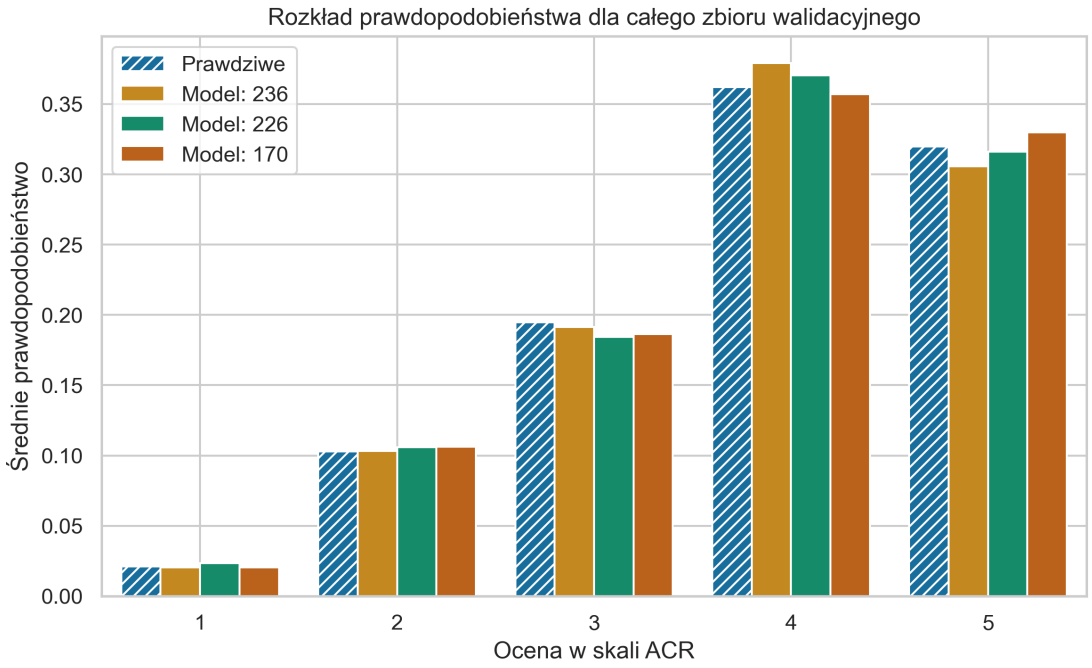
135:2894416398



Rysunek 5.18 Gęstość rozkładu wartości oczekiwanych dla zoptymalizowanych modeli wielowarstwowych

Na wykresie gęstości wartości oczekiwanych 5.18 widać, że rzeczywisty rozkład wartości oczekiwanych jest znacząco zbliżony do rozkładów wartości oczekiwanych przewidzianych ocen przez wszystkie modele. Dla poprzednich modeli, widoczna była znacząca różnica w okolicach środkowej oceny.

136:6708644262



Rysunek 5.19 Rozkład uśrednionego prawdopodobieństwa każej klasy dla zoptymalizowanych modeli wielowarstwowych

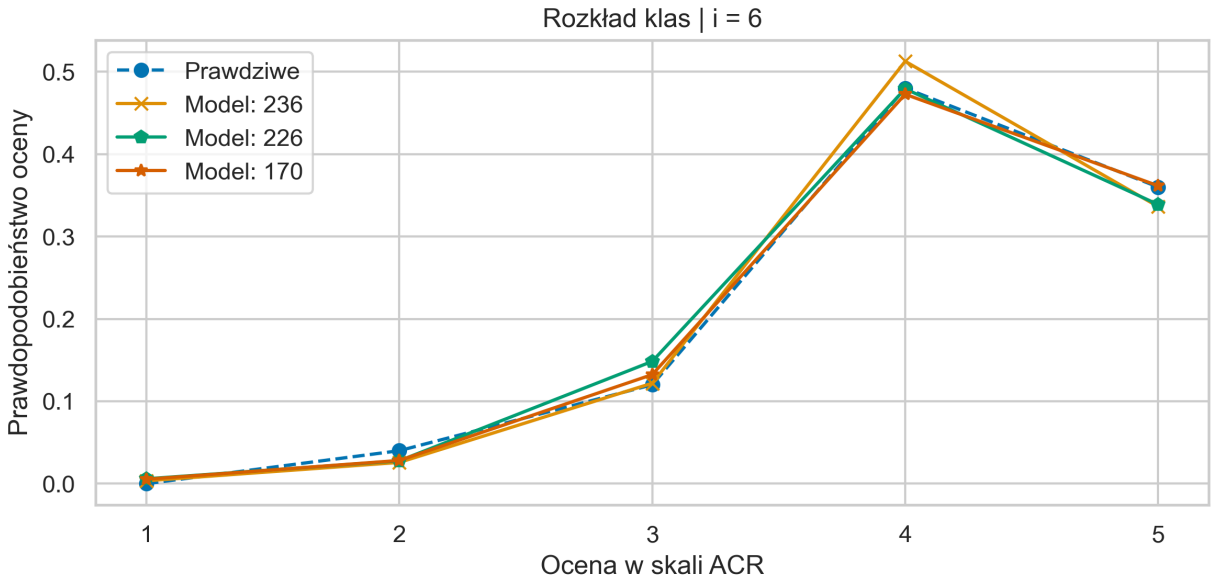
Na rozkładzie prawdopodobieństwa dla całego zbioru walidacyjnego 5.19 widać, że modele zachowują odpowiednie proporcje średniego prawdopodobieństwa ocen w stosunku do wartości rzeczywistych.

Na wykresach 5.20 - 5.25 zaprezentowano wyniki predykcji dla trzech modeli wielowarstwowych dla takich samych próbek jak w poprzednich podrozdziałach. W tabelach 5.15 - 5.20 zamieszczono podsumowanie parametrów próbek oraz skojarzone wektory prawdopodobieństw [p_1 - p_5]. Podobnie jak poprzednio, pierwszy wiersz *numer próbki* – *true* w każdej tabeli to próbka z rzeczywistymi wartościami prawdopodobieństw. Kolejne wiersze to predykcja modeli z danej próby. Nagłówki tabel przedstawione jako akronimy objaśniono poniżej:

- i - trial - numer próbki oraz modelu, dla przypadku *i* - *true* są to wartości prawdziwe
- c_a – current_assessment - subiektywna ocena w skali od 1 do 100 aktualnie odtwarzanego materiału wideo (materiał B z pary A i B),
- p_a – previous_assessment - subiektywna ocena w skali od 1 do 100 poprzednio odtwarzanego materiału wideo (materiał A z pary A i B),
- c_b – current_bitrate - przepływność aktualnie odtwarzanego materiału wideo,
- p_b – previous_bitrate - przepływność poprzednio odtwarzanego materiału wideo,
- cdc – codec - typ kodeka,
- t_v_s – zakres rozmiaru ekranu TV w calach,

137:5608230687

- a_r – zakres wieku oglądającego,
- p_1, p_2, p_3, p_4, p_5 – wektor prawdopodobieństw przydzielenia danej oceny w skali ACR.



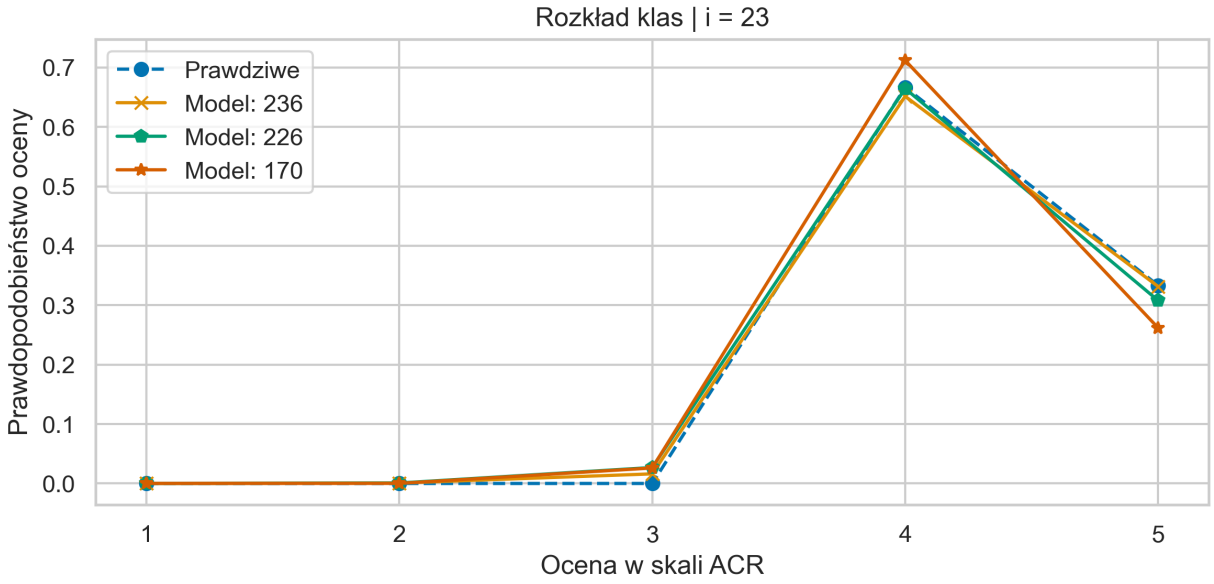
Rysunek 5.20 Wyniki predykcji dla próbki walidacyjnej i = 6 dla zoptymalizowanych modeli wielowarstwowych

Tabela 5.15 Parametry dla próbki $i = 6$

i – trial	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
6 – true	79	79	3298	5342	H.265	31 - 42	18 - 25	0,000	0,040	0,120	0,480	0,360
6 – 236	–	–	–	–	–	–	–	0,003	0,026	0,122	0,513	0,336
6 – 226	–	–	–	–	–	–	–	0,006	0,028	0,148	0,479	0,339
6 – 170	–	–	–	–	–	–	–	0,005	0,028	0,132	0,473	0,362

Wykres 5.20 pokazuje, że wszystkie trzy modele w zbliżony sposób dokonały predykcji dla próbki $i = 6$. Model z próby 236 nieznacznie przeszacował prawdopodobieństwo oceny 4. Kolejność prawdopodobieństw została zachowana, tak jak w przypadku wartości rzeczywistych. W tabeli 5.15 widać, że wszystkie modele przypisały minimalne, niezerowe prawdopodobieństwo najniższej ocenie.

138:8418946174



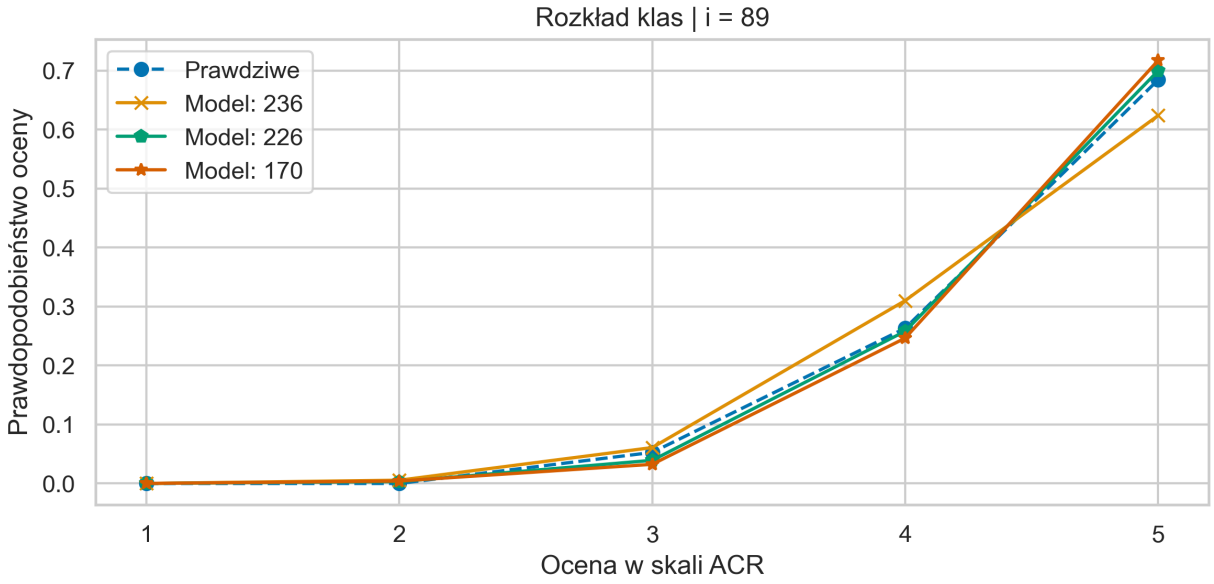
Rysunek 5.21 Wyniki predykcji dla próbki walidacyjnej i = 23 dla zoptymalizowanych modeli wielowarstwowych

Tabela 5.16 Parametry dla próbki $i = 23$

i – trial	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
23 – true	68	90	3000	6000	H.264	> 56	40 - 59	0,000	0,000	0,000	0,667	0,333
23 – 236	–	–	–	–	–	–	–	0,000	0,001	0,016	0,652	0,331
23 – 226	–	–	–	–	–	–	–	0,000	0,001	0,027	0,665	0,308
23 – 170	–	–	–	–	–	–	–	0,000	0,000	0,026	0,712	0,262

Próbka $i = 23$ składa się z dwóch sekwencji wideo reprezentujących przejście ze strumienia 3000 Kbit/s do 6000 Kbit/s. Na wykresie 5.21 oraz w tabeli 5.16 widać, że prawidłowo została oszacowana klasa 4, jako ocena z najwyższym prawdopodobieństwem.

139:6351691484



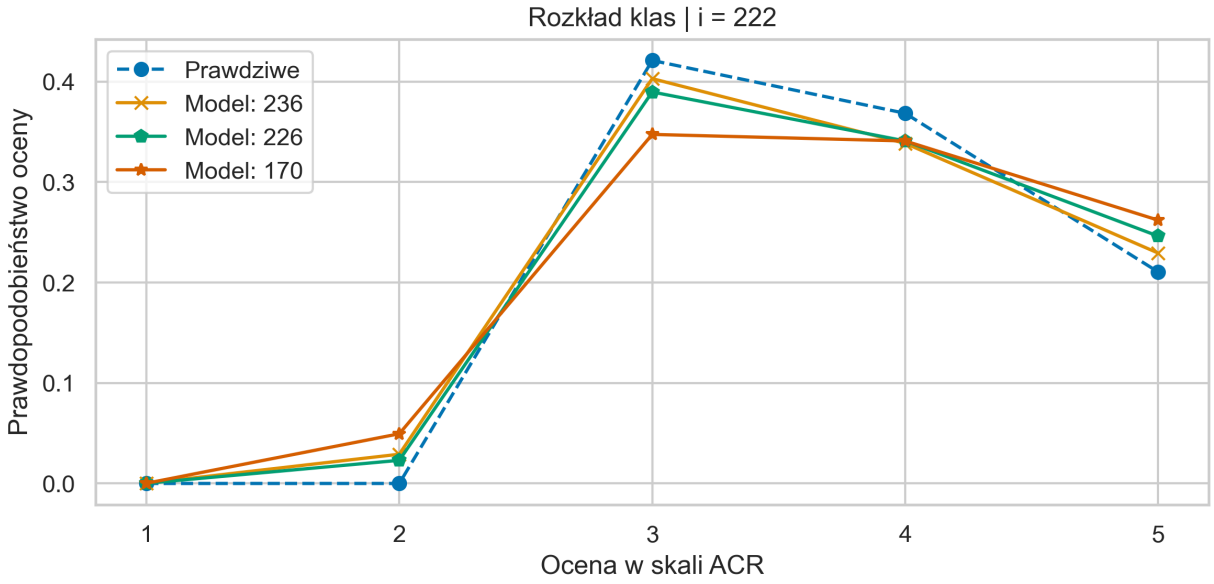
Rysunek 5.22 Wyniki predykcji dla próbki walidacyjnej i = 89 dla zoptymalizowanych modeli wielowarstwowych

Tabela 5.17 Parametry dla próbki $i = 89$

i – trial	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
89 – true	85	75	10000	1500	H.264	43 - 55	26 - 39	0,000	0,000	0,053	0,263	0,684
89 – 236	–	–	–	–	–	–	–	0,000	0,006	0,061	0,310	0,624
89 – 226	–	–	–	–	–	–	–	0,000	0,003	0,039	0,258	0,700
89 – 170	–	–	–	–	–	–	–	0,000	0,004	0,032	0,246	0,717

W tabeli 5.17 zawarto oceny dla próbki $i = 89$ reprezentującej wzrost przepływności strumienia z 1500 Kbit/s do 10000 Kbit/s. Segmenty wideo zostały średnio ocenione na 75 punktów - 1500 Kbit/s oraz 85 punktów - 10000 Kbit/s. Pomimo dużej różnicy przepływności, strumienie mają zbliżone oceny w skali od 1 do 100. Na wykresie 5.22 widać, że kolejność prawdopodobieństw została prawidłowo odwzorowana. Z największym prawdopodobieństwem została wskazana ocena o najwyższej klasie.

140:2507795448



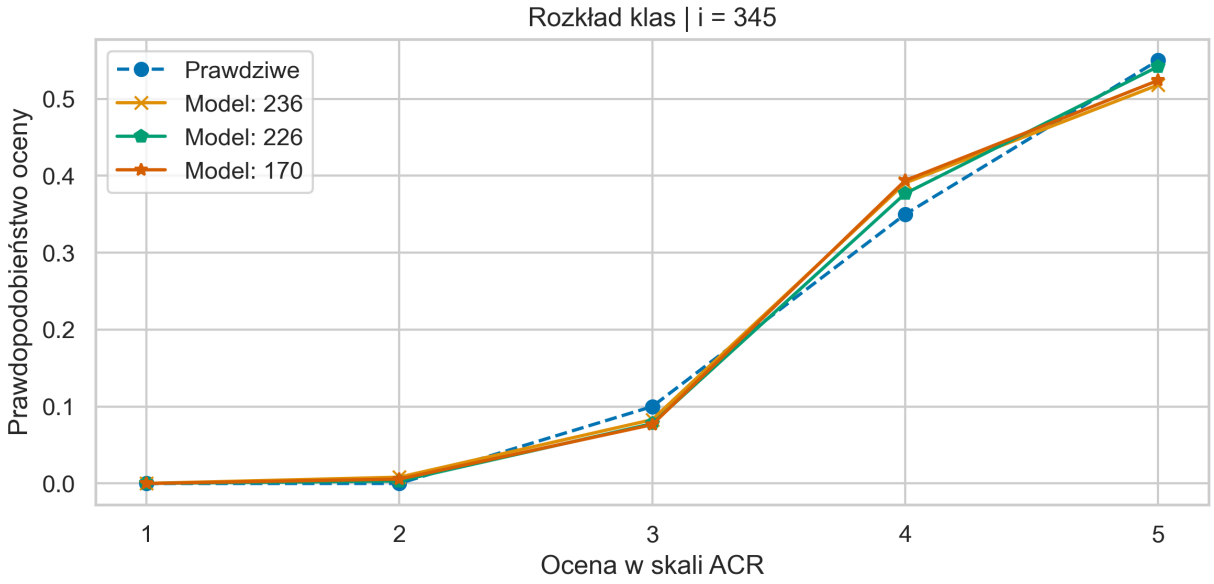
Rysunek 5.23 Wyniki predykcji dla próbki walidacyjnej i = 222 dla zoptymalizowanych modeli wielowarstwowych

Tabela 5.18 Parametry dla próbki $i = 222$

i – trial	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
222 – true	63	60	6000	1500	H.264	43 - 55	26 - 39	0,000	0,000	0,421	0,368	0,211
222 – 236	–	–	–	–	–	–	–	0,000	0,029	0,403	0,338	0,229
222 – 226	–	–	–	–	–	–	–	0,000	0,023	0,390	0,341	0,246
222 – 170	–	–	–	–	–	–	–	0,000	0,049	0,347	0,341	0,262

Próbka $i = 222$ widoczna na wykresie 5.23 i opisana w tabeli 5.18 to para segmentów wideo o przepływnościach 6000 Kbit/s oraz 1500 Kbit/s. Próbka określa wzrost przepływności o 4500 Kbit/s. Segmenty zostały subiektywnie ocenione na podobnym poziomie, czyli w okolicach 60 punktów. Modele prawidłowo wskazały ocenę 3, jako klasę o najwyższym prawdopodobieństwie. Model 170 wskazał oceny 3 i 4 z bardzo zbliżonym prawdopodobieństwem. Kolejność prawdopodobieństw została odwzorowana prawidłowo.

141:1731253440



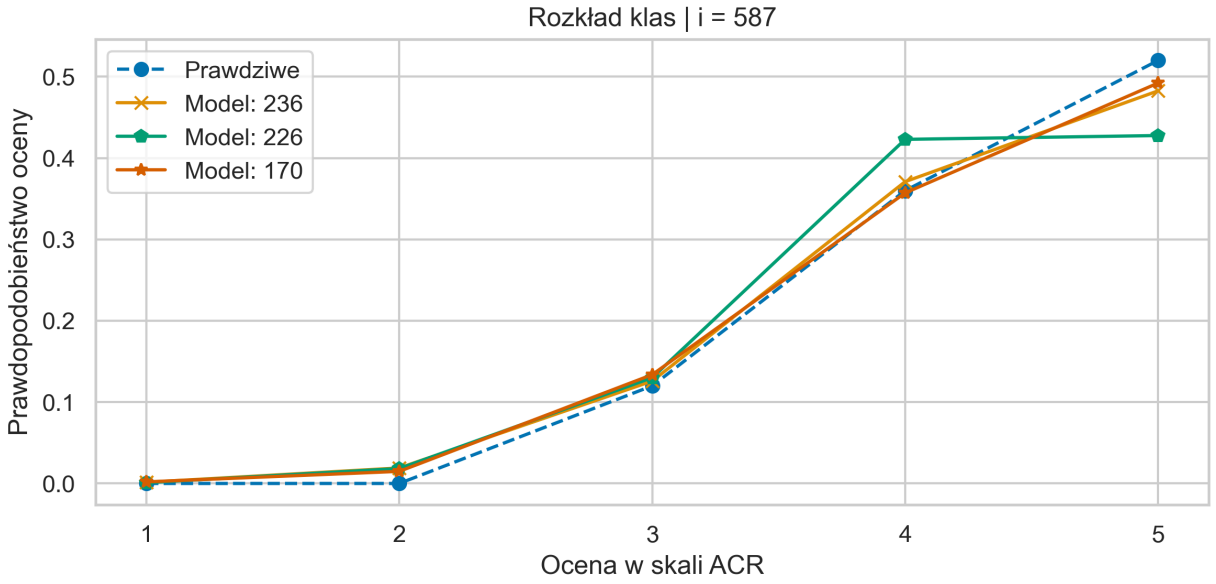
Rysunek 5.24 Wyniki predykcji dla próbki walidacyjnej i = 345 dla zoptymalizowanych modeli wielowarstwowych

Tabela 5.19 Parametry dla próbki $i = 345$

i – trial	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
345 – true	77	88	3298	6364	H.265	43 - 55	26 - 39	0,000	0,000	0,100	0,350	0,550
345 – 236	–	–	–	–	–	–	–	0,000	0,008	0,083	0,391	0,518
345 – 226	–	–	–	–	–	–	–	0,000	0,003	0,078	0,377	0,542
345 – 170	–	–	–	–	–	–	–	0,000	0,006	0,076	0,394	0,524

Próbka $i = 345$ widoczna na wykresie 5.24 to para segmentów opisująca spadek jakości o około 3000 Kbit/s dla materiałów wideo kodowanych przez H.265. Pomimo spadku jakości, rzeczywista ocena z najwyższym prawdopodobieństwem to 5. Także i w tym przypadku, modele wielowymiarowe prawidłowo uszeregowały kolejność prawdopodobieństw. W tabeli 5.19 widać, że dla dwóch najniższych ocen przypisano zerowe lub nieznaczne prawdopodobieństwo, zgodnie z rzeczywistym stanem.

142:1030070293



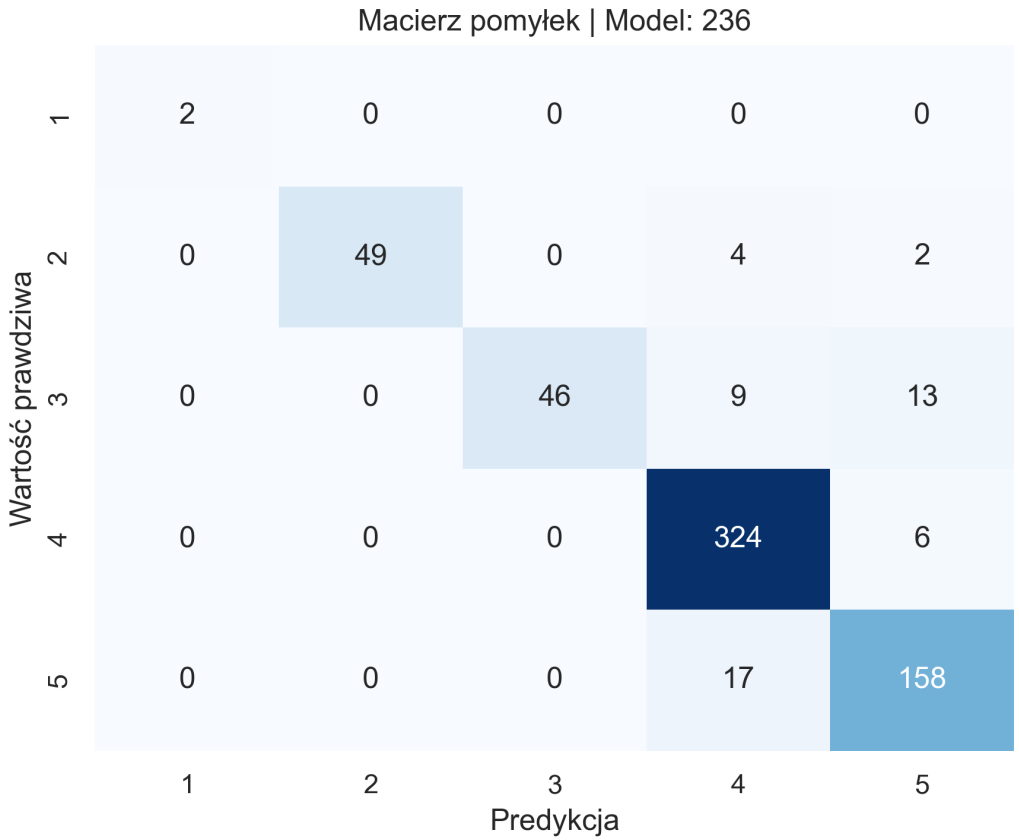
Rysunek 5.25 Wyniki predykcji dla próbki walidacyjnej i = 587 dla zoptymalizowanych modeli wielowarstwowych

Tabela 5.20 Parametry dla próbki $i = 587$

i – trial	c_a	p_a	c_b	p_b	cdc	t_v_s	a_r	p_1	p_2	p_3	p_4	p_5
587 – true	93	81	5342	2276	H.265	31 - 42	18 - 25	0,000	0,000	0,120	0,360	0,520
587 – 236	–	–	–	–	–	–	–	0,001	0,019	0,126	0,371	0,483
587 – 226	–	–	–	–	–	–	–	0,001	0,018	0,130	0,423	0,428
587 – 170	–	–	–	–	–	–	–	0,002	0,015	0,134	0,357	0,492

Na wykresie 5.25 widoczna jest próbka dla wzrostu przepływności o około 3000 Kbit/s dla segmentów kodowanych H.265. W tym przypadku wzrost jest od przepływności nieznacznie przekraczającej 2200 Kbit/s do ponad 5300 Kbit/s. Uczestnicy badania wskazali, że średnia subiektywna ocena to 5 dla obu treści. Dla H.265 segmenty wideo często były oceniane w górnej części skali ACR, więc pomimo niskiej przepływności początkowej, segmenty były subiektywnie oceniane na podobnym poziomie 5.20. Modele prawidłowo odwzorowały kolejność prawdopodobieństw. Model 226 wskazał zbliżone wartości prawdopodobieństw dla dwóch najwyższych ocen. Pozostałe modele wskazały ocenę 5 jednoznacznie.

143:1041701592



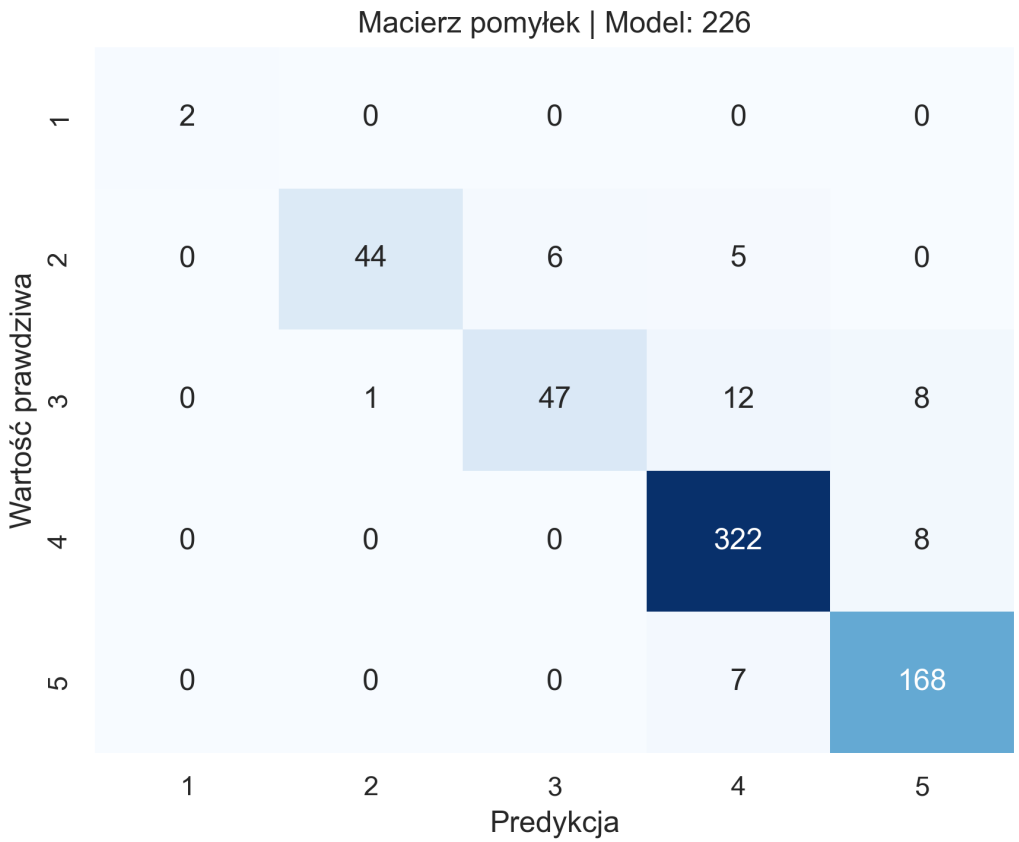
Rysunek 5.26 Macierz pomyłek dla zoptymalizowanego modelu wielowarstwowego z próby 236

Tabela 5.21 Metryki dla zoptymalizowanego modelu wielowarstwowego z próby 236

Model	SROCC	Accuracy	Precision	Recall	F1
Model 236	0,982	0,919	0,960	0,890	0,918

Macierz pomyłek dla modelu z próby 236 została przestawiona na wykresie 5.26. Model prawidłowo rozpoznał przypadki dla najniższej i najrzadziej występującej oceny. Dla oceny 2 prawidłowo zostało sklasyfikowanych 49 próbek, a 6 próbek zostało przeszacowanych. Dla środkowej oceny prawidłowo zostało sklasyfikowanych 46 próbek, a 22 próbki zostały przeszacowane. Najliczniejsza klasa 4 została prawidłowo oszacowana 324 razy, a 6 razy została zawyżona. Dla najwyższej klasy model prawidłowo wskazał 158 próbek, natomiast 17 zostało niedoszacowych. Model pewnie rozpoznaje dominującą klasę 4. Nie ma także problemu z oceną najniższą i prawidłowo sklasyfikował 2 próbki należące do tej klasy. Klasy pośrednie są rozpoznawane z tendencją do zawyżania oceny. W tabeli 5.21 zawarto wartości metryk określających skuteczność modelu z próby 236. SROCC na poziomie 0,982 wskazuje, że model bardzo dobrze odwzorowuje kolejność prawdopodobieństw. Dokładność na poziomie 91,9% to poprawa względem poprzednich modeli. Precyzja na poziomie 96% oznacza, że model zazwyczaj przypisuje poprawną klasę. Model poprawnie wykrywa około 89% wszystkich przykładów o czym świadczy wartość czułości. Średnia harmoniczna F1 o wartości 0,918% wskazuje, na znaczną precyzje i kompletność modelu.

144:4473043511



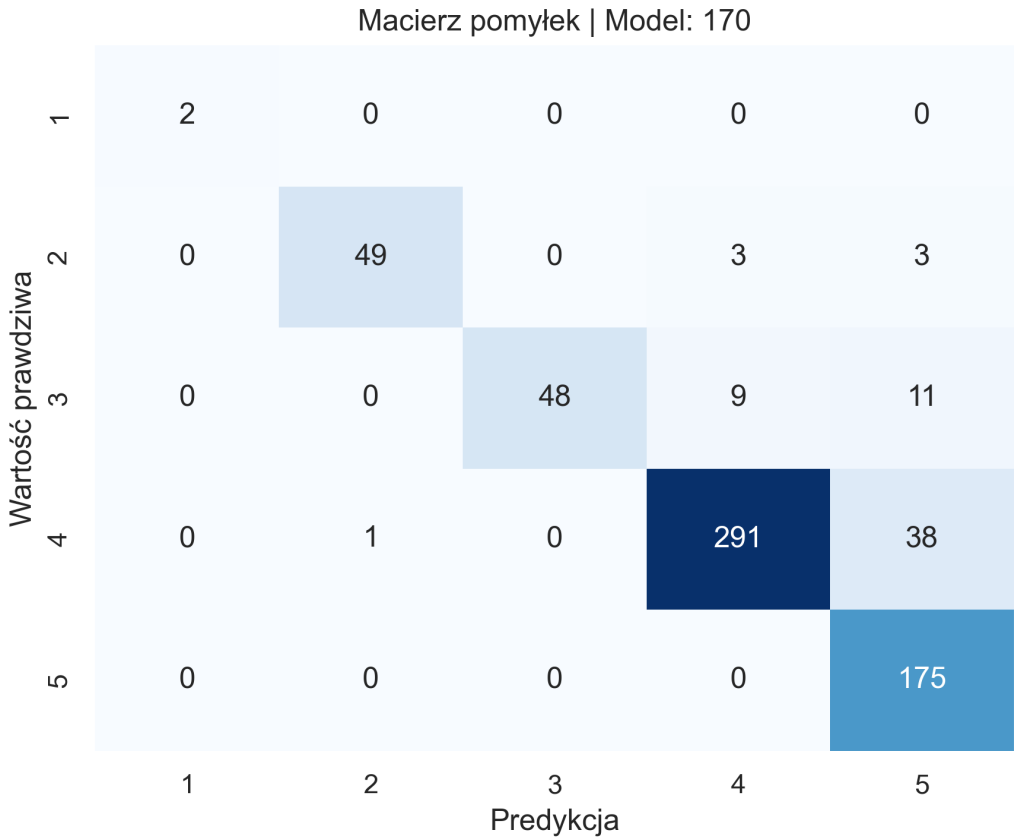
Rysunek 5.27 Macierz pomyłek dla zoptymalizowanego modelu wielowarstwowego z próby 226

Tabela 5.22 Metryki dla zoptymalizowanego modelu wielowarstwowego z próby 226

Model	SROCC	Accuracy	Precision	Recall	F1
Model 226	0,979	0,925	0,942	0,885	0,909

Dla modelu z próby 226 przedstawiono macierz pomyłek na wykresie 5.27. Próbki z najniższą oceną zostały wykryte prawidłowo. Dla oceny 2 prawidłowo sklasyfikowano 44 próbki, 11 próbek zostało przeszacowanych. Ocena środkowa została wykryta prawidłowo dla 47 próbek, 1 z nich została niedoszacowana, a 20 przeszacowanych. Dla oceny 4 sklasyfikowano poprawnie 322 próbki a 8 zostało przeszacowanych. Dla najwyższej klay prawidłowo wskazano 168 próbek, natomiast 7 zostało niedoszacowanych. W tabeli 5.22 zawarto wartości metryk modelu. SROCC dla tego modelu wynosi 0,979, jest to wartość zbliżona do podobnego modelu. Dokładność to 92,5%, precyzja jest na poziomie 94,2%, natomiast wartość czułości to około 88,5%. Średnia harmoniczna F1 wynosi 90,9%. Predykcje modelu są trafne, jednak istnieje tendencja do zawyżania ocen, głównie w kierunku wyższych klas.

145:3197474848



Rysunek 5.28 Macierz pomyłek dla zoptymalizowanego modelu wielowarstwowego z próby 170

Tabela 5.23 Metryki dla zoptymalizowanego modelu wielowarstwowego z próby 170

Model	SROCC	Accuracy	Precision	Recall	F1
Model 170	0,983	0,897	0,942	0,896	0,910

Macierz pomyłek dla modelu z próby 170 została przedstawiona na wykresie 5.28. Dla najniższej klasy obie próbki zostały sklasyfikowane poprawnie. Dla oceny 2 prawidłowo oszacowano 49 próbek, natomiast 6 z nich zostało przeszacowanych. Model wykonał 48 poprawnych klasyfikacji dla oceny 3, a 20 próbek zostało błędnie zidentyfikowanych jako klasy 4 lub 5. 291 próbek zostało poprawnie sklasyfikowanych jako ocena 4. 39 próbek zostało błędnie uznanych za najwyższą klasę, a 1 z nich została błędnie wskazana jako ocena 2. Dla klasy najwyższej wszystkie próbki zostały poprawnie wskazane. W tabeli 5.23 zawarto metryki porównawcze dla modelu z tej próby. Wartość SROCC wynosi około 0,983. Jest to podobna wartość do poprzednich modeli wielowarstwowych. Dokładność wynosi około 89,7% i jest nieznacznie niższa w stosunku do poprzednich modeli. Precyzja wynosi 94,2%, a czułość to około 89,6%. Średnia harmoniczna F1 dla tego modelu wyniosła 91%.

146:1037732192

5.5 Podsumowanie

W rozdziale przedstawiono podejście do optymalizacji modeli uczenia maszynowego w celu uzyskania rozwiązania możliwego do wdrożenia do oceny jakości doświadczeń w produkcyjnym systemie telewizji Hiway, bazującym na transmisji strumieniowej HTTP. Łącznie wytrenowano 442 modele. Dla weryfikacji ponownie przetrenowano 100 najlepszych modeli (dla architektur jednowarstwowych i wielowarstwowych), wykorzystując walidację krzyżową.

Tabela 5.24 Porównanie wybranych modeli przedstawionych w rozdziale

Próba	SROCC	Accuracy	Precision	Recall	F1
Model 123	0,938	0,732	0,587	0,531	0,542
Model 37	0,937	0,852	0,674	0,604	0,629
Model 134	0,933	0,889	0,949	0,839	0,878
Model 236	0,982	0,919	0,960	0,890	0,918
Model 226	0,979	0,925	0,942	0,885	0,909
Model 170	0,983	0,897	0,942	0,896	0,910

Spośród 100 najlepszych modeli wybrano po trzy modele dla architektury jednowarstwowej i wielowarstwowej. Metryki porównawcze tych modeli zestawiono w tabeli 5.24.

- **Model z próby 123** — jedna warstwa ukryta, liczba wejść $N = 128$, funkcja aktywacji GELU, optymalizator Adam, 2 próbki na wsad. Model osiągnął wysoki SROCC równy 0,938. Dokładność wyniosła 73,2%, precyzja 58,7%, czułość 53,1%, a miara F1 — 54,2%. Wśród modeli jednowarstwowych był to jeden z najlepszych wyników SROCC (prawidłowe porządkowanie prawdopodobieństw), jednak pozostałe metryki są niskie, co ogranicza użyteczność praktyczną. Może stanowić punkt odniesienia ze względu na SROCC.

- **Model z próby 37** — jedna warstwa ukryta, $N = 128$, funkcja aktywacji GELU, optymalizator Adam, 16 próbek na wsad. SROCC porównywalny z poprzednim modelem; pozostałe metryki uległy poprawie, choć nadal są poniżej oczekiwań.

- **Model z próby 134** — jedna warstwa ukryta, $N = 128$, funkcja aktywacji GELU, optymalizator Adam, 4 próbki na wsad. SROCC zbliżony do poprzednich; znacząco lepsze metryki klasyfikacyjne: dokładność 88,9%, precyzja 94,9%, czułość 83,9%, F1 87,8%.

- **Model z próby 236** — dwie warstwy ukryte ($N = 128$ oraz $N = 64$), funkcja aktywacji ReLU, optymalizator SGD, 4 próbki na wsad. SROCC wyższy o ok. 5% względem najlepszych modeli jednowarstwowych; dokładność 91,9%. Precyzja 96% (najwyższa ze wszystkich), czułość 89%, F1 91,8% (najwyższa w zestawieniu).

- **Model z próby 226** — dwie warstwy ukryte ($N = 128$ oraz $N = 64$), funkcja aktywacji ReLU, optymalizator SGD, 4 próbki na wsad. SROCC 97,9% (nieco niższy niż w próbie 236). Najwyższa dokładność: 92,5%. Precyzja 94,2%, czułość 88,5%, F1 90,9% — nieznacznie niższe niż w próbie 236.

147:1994608612

- **Model z próby 170** — dwie warstwy ukryte ($N = 128$ oraz $N = 64$), funkcja aktywacji ReLU, optymalizator SGD, 4 próbki na wsad. Najwyższy SROCC: 98,3%. Dokładność niższa (poniżej 90%). Precyzja taka jak w próbie 226, czułość najwyższa (89,6%), F1 $\approx$ 91% (zbliżona do próby 226).

Z wyjątkiem próby 134, modele z pojedynczą warstwą ukrytą wyraźnie ustępują modelom z dwiema warstwami. Dodanie warstwy ukrytej poprawiło metryki i stanowi korzystny kompromis między złożonością a skutecznością. Model z próby 236 osiągnął najwyższą wartość F1 (uwzględniając precyzję i czułość), zadowalającą dokładność oraz bardzo wysoki współczynnik SROCC, co wskazuje na niemal idealne odwzorowanie porządku prawdopodobieństw w wektorach wyjściowych. Model z próby 226 może być alternatywą, jeśli priorytetem jest maksymalizacja dokładności. Próba 170 oferuje najwyższy SROCC kosztem nieco niższej dokładności.

Modele z dwiema warstwami ukrytymi są do siebie bardzo zbliżone pod względem wyników. Do wdrożenia wybrano model z próby 236. Średnia trzech kluczowych miar (SROCC, Accuracy, F1) jest dla tej próby najwyższa i wynosi 94%. Model z próby 226 uzyskał średnią 93,8%, a model z próby 170 — 93%.

148:1142138958

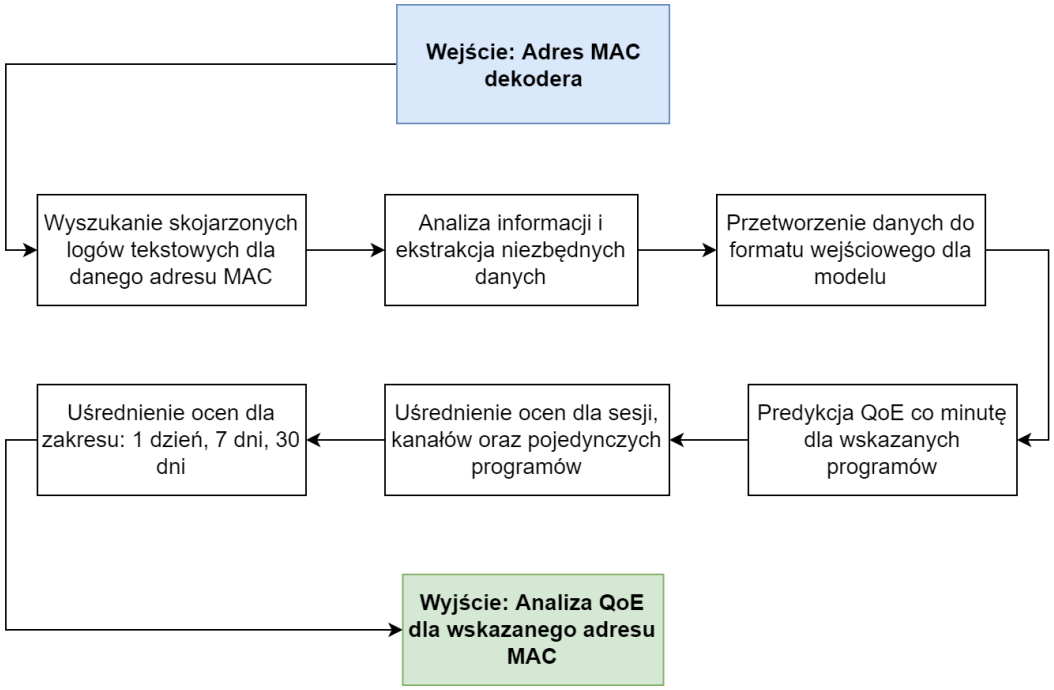
Rozdział 6

Implementacja narzędzia do analizy QoE

Model z próby 236, z hiperparametrami zestawionymi w tabeli 6.1, został wdrożony jako narzędzie do analizy jakości doświadczeń (QoE) w usłudze Hiway. Implementację wykonano w języku Python, a podstawowym elementem rozwiązania jest framework Django. Narzędzie jest dostępne z poziomu przeglądarki WWW.

Tabela 6.1 Parametry wdrożonego modelu z próby 236

Liczba wejść	Dropout	Learning_rate	Weight_decay	Batch_size	Aktywacja	Optymalizator
[128, 64]	0,162	0,0091	2,884E-05	4	ReLU	SGD



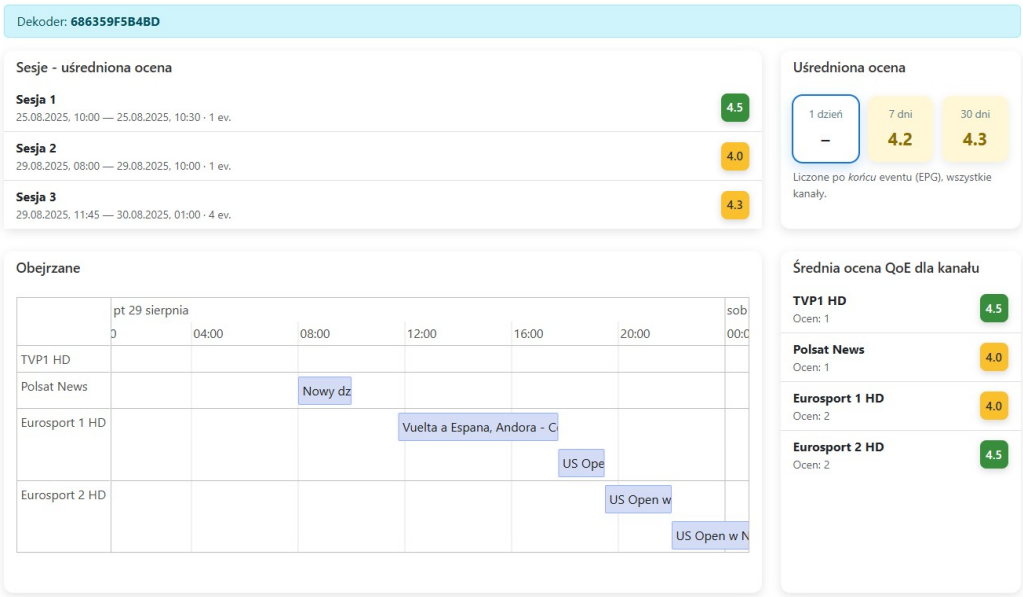
Rysunek 6.1 Schemat narzędzia do analizy Quality of Experience dla treści prezentowanych przez dekodery

149:5203289312

Celem wdrożenia było przygotowanie aplikacji umożliwiającej sprawdzenie jakości doświadczeń QoE na wybranym dekoderze usługi Hiway. Narzędzie zaprojektowano tak, aby stanowiło rozszerzenie zestawu istniejących narzędzi monitorujących usługę, opisanych w ramach wielowarstwowego modelu analizy jakości w [1]. Na rysunku 6.1 przedstawiono uproszczony schemat działania aplikacji.

Wejściem systemu jest adres MAC dekodera, który ma zostać poddany analizie. Na tej podstawie wyszukiwane są logi generowane przez urządzenie. Logi opisują m.in. uruchomienie i wyłączenie dekodera, rozpoczęcie odtwarzania kanału, wystąpienie błędów, zmianę jakości strumienia czy skorzystanie z funkcji innych niż telewizja na żywo. Duża część logów jest nieistotna z punktu widzenia analizy jakości, dlatego poddawane są one filtracji i ekstrakcji niezbędnych danych. Następnie wykonywana jest standaryzacja i kodowanie cech, po czym wyznaczane są obejrzane audycje i aktualnie odtwarzany program.

Dla każdej z audycji model dokonuje predykcji jakości doświadczeń w skali ACR, aktualizowanej co minutę. Możliwe jest również uwzględnienie w analizie przedziału wiekowego odbiorcy i rozmiaru ekranu. Wyniki predykcji są uśredniane dla całej audycji, kanału, a także całej sesji oglądania. Aplikacja umożliwia również prezentację średnich ocen w ujęciu dobowym, 7-dniowym i 30-dniowym. Dane przedstawiane są w formie graficznej w dedykowanych widokach aplikacji.



Rysunek 6.2 Narzędzie do analizy QoE — widok sesji oglądania i średnich ocen dla kanału

Na rysunku 6.2 przedstawiono przykład analizy całej sesji oglądania — od momentu uruchomienia dekodera do jego wyłączenia. Dodatkowo narzędzie prezentuje uśrednioną ocenę dla wybranego kanału oraz wartości skumulowane dla dnia, 7- i 30-dniowego okresu.

150:3036506964



Rysunek 6.3 Narzędzie do analizy QoE — widok listy audycji

Widok zaprezentowany na rysunku 6.3 pozwala na przegląd listy audycji obejrzanych na danym dekoderze wraz z przypisanymi do nich uśrednionymi ocenami.



Rysunek 6.4 Narzędzie do analizy QoE — widok szczegółowy audycji

Przejście do szczegółowego widoku (rys. 6.4) umożliwia analizę ocen predykowanych co minutę dla wybranego programu, zgodnie z wynikami modelu z próby 236.

Celem przygotowanego narzędzia było umożliwienie szybkiej weryfikacji jakości doświadczeń użytkowników, szczególnie w sytuacjach zgłaszania problemów z działaniem usługi TV. W kolejnym rozdziale przedstawiono podsumowanie pracy i wnioski z badań.

151:7399906886

Rozdział 7

Podsumowanie

Analiza doświadczeń klienta oraz wynikająca z niej dążność do zaspokojenia jego oczekiwań stanowi dziś kluczowy element działalności każdej firmy świadczącej usługi. Jest to również istotny aspekt funkcjonowania sieci operatorskich, determinujący rozwój nowoczesnych usług.

Szczególnie wyraźnie widać to w przypadku usług telewizyjnych, w których zapewnienie użytkownikom możliwie najlepszych wrażeń z odbioru treści oraz jednoczesne utrzymanie ich lojalności stanowi jedno z najtrudniejszych wyzwań dla dostawców. Podstawowym produktem oferowanym w ramach usługi telewizyjnej jest treść wideo. Sposób jej dystrybucji zależy od technologii stosowanej w danym systemie operatorskim.

Klasyczne systemy *IPTV* (ang. *Internet Protocol Television*) dostarczały liniowy sygnał wideo w technologii *IP Multicast*, cechującej się optymalnym wykorzystaniem zasobów sieciowych. Funkcje dodatkowe w takich systemach były ograniczone do podstawowych, takich jak wideo na żądanie *VOD* (ang. *Video on Demand*) czy nagrywanie programów na lokalnym dysku dekodera.

Obecnie systemy telewizyjne oparte na strumieniowaniu adaptacyjnym HTTP wykorzystują zaawansowane urządzenia *STB* (ang. *Set-top-box*), stale podłączone do sieci. Odtwarzanie treści liniowych stanowi tylko jedną z wielu oferowanych funkcjonalności. Nowoczesne dekodery i oprogramowanie umożliwiają m.in. nagrywanie w chmurze, cofanie programów, nagrywanie seryjne, profilowanie użytkowników oraz rekomendowanie audycji przy użyciu silników rekomendacji. Dekodery pozwalają również na korzystanie z usług VOD w formie aplikacji instalowanych bezpośrednio na urządzeniach, co zapewnia dostęp do niemal nieograniczonego katalogu treści. Równolegle rozwijane są systemy monitorowania jakości i gromadzenia danych, umożliwiające operatorom szybkie wykrywanie i usuwanie problemów wpływających na jakość odbioru.

Niezmiennie jednak kluczowym elementem pozostaje treść wideo. Niska jakość obrazu lub brak oczekiwanych kanałów może sprawić, że pozostałe funkcje stracą dla użytkownika znaczenie, a jego niezadowolenie przełoży się na negatywną ocenę całej usługi oraz operatora.

W rozprawie przedstawiono metodę odwzorowania subiektywnej oceny jakości usługi telewizyjnej platformy Hiway, w ramach której firma INEA oferuje swoim abonentom m.in. usługę telewizji. W tym celu przeprowadzono badania umożliwiające zebranie ocen użytkowników. Dużym wyzwaniem było przeprowadzenie badań na rzeczywistych urządzeniach STB platformy Hiway, ze względu na stosowane w nich algorytmy adaptacji strumienia, które mogą powodować odmienne zachowania

152:3269126905

w porównaniu z typowymi odtwarzaczami *HAS* (ang. *HTTP Adaptive Streaming*). Trudność stanowiło również opracowanie sposobu zbierania ocen w trakcie rzeczywistego oglądania treści. Tradycyjny formularz papierowy, zgodny z zaleceniami ITU, okazał się niewystarczający ze względu na niską precyzję i niedostosowaną formę. Skutecznym rozwiązaniem była specjalnie przygotowana aplikacja mobilna, udostępniana abonentom po zeskanowaniu kodu QR.

Wyzwaniem była także subiektywność i rozbieżność ocen. Dane były silnie niezrównoważone, ponieważ większość ocen znajdowała się w górnych decylach skali, a liczba ocen najniższych była znikoma. Zdecydowano się jednak nie stosować metod równoważenia zbioru, aby w pełni zachować autentyczność danych użytkowników.

Do predykcji ocen zastosowano perceptron wielowarstwowy w konfiguracji miękkiego klasyfikatora (ang. *soft classifier*), którego wyjście stanowi rozkład prawdopodobieństwa poszczególnych ocen. Rozwiązanie to wybrano ze względu na charakter danych — klasyczne modele klasyfikacyjne nie były w stanie uchwycić złożonych wzorców i popełniały zbyt wiele błędów, aby mogły zostać wdrożone w środowisku produkcyjnym. Skuteczność modeli oceniano za pomocą współczynnika korelacji rang Spearmana (SROCC), powszechnie stosowanego w badaniach jakości wideo, a także standardowych metryk klasyfikacyjnych. Do predykcji przyjmowano ocenę o najwyższym prawdopodobieństwie.

Ze względu na wdrożeniowy charakter pracy przeprowadzono optymalizację modeli, przeszukując przestrzeń hiperparametrów i porównując różne architektury klasyfikatora. Na każdym etapie wyłoniono trzy najlepsze modele, które poddano szczegółowej walidacji.

Zweryfikowana hipoteza badawcza zakłada, że możliwa jest estymacja jakości usług telewizyjnych na podstawie analizy szerokiego zakresu parametrów dostarczanych przez system *OTT* (ang. *over-the-top*), w tym danych z urządzeń końcowych, z dokładnością przekraczającą 90%. Warunek ten został spełniony poprzez:

- Dobór liczebności próby tak, aby błąd maksymalny nie przekraczał 10% przy 95% poziomie ufności, z uwzględnieniem populacji systemu.
- Osiągnięcie współczynnika korelacji rang Spearmana na poziomie co najmniej 0,9 względem ocen uzyskanych w badaniach subiektywnych.
- Uzyskanie średnich wartości metryk klasyfikacyjnych (dokładności, precyzji, czułości i miary harmonicznej F1) na poziomie co najmniej 90%.

W wyniku przeprowadzonych badań i zdobytych doświadczeń zaproponowano wdrożenie systemu umożliwiającego analizę danych i predykcję jakości doświadczeń QoE dla wybranych zdarzeń, kanałów liniowych oraz całych sesji oglądania — od włączenia do wyłączenia dekodera.

Do najważniejszych osiągnięć rozprawy należy zaliczyć:

1. Analizę architektury systemu i danych raportowanych przez urządzenia końcowe STB.
2. Przegląd i analizę metod subiektywnej oceny jakości treści wideo w strumieniowaniu adaptacyjnym.
3. Przygotowanie i wdrożenie badań subiektywnej oceny jakości z wykorzystaniem aplikacji mobilnej.

153:8913507916

4. Analizę wyników badań i utworzenie zbioru danych łączącego parametry z dekoderów z ocenami subiektywnymi.

5. Wytrenowanie bazowych modeli miękkiej klasyfikacji wyznaczających kierunek dalszych prac.

6. Wytrenowanie i optymalizację końcowych modeli klasyfikacyjnych spełniających założone wymagania.

7. Opracowanie i rekomendację wdrożenia systemu predykcji jakości doświadczeń QoE.

Dalsze kierunki badań mogą obejmować:

1. Rozszerzenie badań nad pozyskiwaniem subiektywnych ocen jakości w różnych scenariuszach korzystania z usługi.

2. Uwzględnienie wpływu dynamicznego wstrzykiwania reklam *DAI* (ang. *Dynamic Ad Insertion*) na postrzeganą jakość.

3. Rozwój metod proaktywnego wykrywania spadków jakości doświadczeń i automatyzacji działań poprawiających QoE.

Podsumowując, w pracy przeprowadzono badania percepcji subiektywnej jakości wideo, wytrenowano i zoptymalizowano algorytmy uczenia maszynowego, a następnie zweryfikowano ich skuteczność. Uzyskane wyniki potwierdzają, że postawiony cel został osiągnięty, a zaproponowane rozwiązania mogą stanowić podstawę dalszego rozwoju systemów monitorowania i poprawy jakości usług telewizyjnych.

154:1052207000

Dodatek A

Treść wideo wykorzystana w sekwencji nr 1

Tabela A.1 Materiał audio wykorzystany w trakcie sekwencji badawczej nr 1

#	**Nazwa filmu**	**Prawa autorskie**	**URL**
1	Stockaudios - Upbeat Acoustic	Music by Ruwin from Pixabay	https://pixabay.com/music/acoustic-group-upbeat-acoustic-113986/

Tabela A.2 Treści wideo wykorzystane w sekwencji badawczej nr 1 dla kodowania zgodnie ze standardem H.264

Para	**Nazwa filmu**	**Prawa**	**URL**
Test 1	Elephant, Tusks, Family	Video by Tesh Lad from Pixabay	https://pixabay.com/videos/elephant-tusks-family-animal-134761/
Test 2	Autumn, Forest, Trees	Video by Fundacja Dla Biebrzy from Pixabay	https://pixabay.com/videos/autumn-forest-trees-nature-53580/
1	Road By The Riverside	Video by Kelly	https://www.pexels.com/video/road-by-the-riverside-2356321/
2	Drone Footage On The Island Of Maldives	Video by Videographer Shiyaz	https://www.pexels.com/video/maldives-free-download-drone-video-beach-island-4010941/

155:5049761230

3	Drone Footage Of Sakhalin Lighthouse In Russia	Video by Yaroslav Shuraev	https://www.pexels.com/video/sea-dark-storm-seagulls-4167691/
4	Traffic On An Intersection Road In A City	Video by Taryn Elliott	https://www.pexels.com/video/traffic-on-an-intersection-road-in-a-city-3121459/
5	Waves Crashing To The Rocks	Video by Engin Akyurt	https://www.pexels.com/video/waves-crashing-to-the-rocks-2141798/
6	Traffic Flow In The Highway	Video by Mike Bird	https://www.pexels.com/video/traffic-flow-in-the-highway-2103099/
7	Elephants And Their Habitat	Video by Magda Ehlers	https://www.pexels.com/video/elephants-and-their-habitat-3842814/
8	A Motor Boat Traversing A River	Video by Tom Fisk	https://www.pexels.com/video/a-motor-boat-traversing-a-river-2647736/
9	Man Paddleboarding in Placid River	Video by Be The Observer	https://www.pexels.com/video/man-paddleboarding-in-placid-river-4791166/
10	Video of a Ducks on Water	Video by Ambient_Nature_Atmosphere	https://www.pexels.com/video/video-of-a-ducks-on-water-6235636/
11	Cerro de Monserrate 01	Video by Ricardo Perez	https://www.pexels.com/video/cerro-de-monserrate-01-18092718/
12	People Hanging Out And Walking Around Washington Square On a Breezy Day	Video by Joshua Santos	https://www.pexels.com/video/people-hanging-out-and-walking-around-washington-square-on-a-breezy-day-2985407/

156:4554107244

13	Mountains, Alps, Fog	Video by FelixMittermeier from Pixabay	https://pixabay.com/pl/videos/g%C3%B3ry-alpy-mg%C5%82a-%C5%9Bnieg-chmury-zima-91545/
14	Hong-Kong, Cruise, Harbor	Video by Bellergy BC from Pixabay	https://pixabay.com/videos/hong-kong-cruise-harbour-transport-2849/
15	Waterfall, River, Cascade	Video by Engin Akyurt from Pixabay	https://pixabay.com/videos/waterfall-river-cascade-that-126442/

157:7379945674

Dodatek B

Treść wideo wykorzystana w sekwencji nr 2

Tabela B.1 Materiał audio wykorzystany w trakcie sekwencji badawczej nr 2

#	**Nazwa filmu**	**Prawa autorskie**	**URL**
1	Stockaudios - Upbeat Acoustic	Music by Ruwin from Pixabay	https://pixabay.com/music/acoustic-group-upbeat-acoustic-113986/

Tabela B.2 Treści wideo wykorzystane w sekwencji badawczej nr 2 dla kodowania zgodnie ze standardem H.265

Para	**Nazwa filmu**	**Prawa**	**URL**
Test 1	Elephant, Tusks, Family	Video by Tesh Lad from Pixabay	https://pixabay.com/videos/elephant-tusks-family-animal-134761/
Test 2	Autumn, Forest, Trees	Video by Fundacja Dla Biebrzy from Pixabay	https://pixabay.com/videos/autumn-forest-trees-nature-53580/
1	Bird, Grouse, Rock	Video by Kanenori from Pixabay	https://pixabay.com/videos/bird-grouse-rock-ptarmigan-174406/
2	Trees, Sunrise, Farm	Video by D M from Pixabay	https://pixabay.com/videos/trees-sunrise-farm-nature-117745/
3	Work, Office, People	Video by Oybek Ostanov from Pixabay	https://pixabay.com/videos/work-office-people-working-87593/

158:4379577587

5	Bike Ride, Bike Path, Park	Video by dae jeung kim from Pixabay	https://pixabay.com/videos/bike-ride-bike-path-park-sports-171960/
6	Ski, Winter, Snow	Video by heimseiten WebdesignKoeln from Pixabay	https://pixabay.com/videos/ski-winter-snow-mountains-8558/
7	Zebra, Wild, Animal	Video by Tham Nguyen from Pixabay	https://pixabay.com/videos/zebra-wild-animal-grass-field-152527/
8	Road, Highway, Car	Video by Design Scape Studio from Pixabay	https://pixabay.com/videos/road-highway-car-mountain-42483/
9	Sea, View, Sky	Video by Engin Akyurt from Pixabay	https://pixabay.com/videos/sea-view-sky-ocean-nature-kennedy-122450/
10	Rabbits, Feeder, Mammals	Video by Engin Akyurt from Pixabay	https://pixabay.com/videos/rabbits-feeder-mammals-animal-farm-172613/
11	Road, Farmland, Scooter	Video by Matthias Groeneveld from Pixabay	https://pixabay.com/videos/road-farmland-scooter-plot-forest-167768/
12	Mountain, River, Nature	Video by Paul Diaconu from Pixabay	https://pixabay.com/videos/mountain-river-nature-landscape-34545/
13	Rims, Worker, Manufacturing	Video by Alex Kopeykin from Pixabay	https://pixabay.com/videos/rims-worker-manufacturing-hard-work-78331/
14	Cell Tower, Tower, Steel	Video by Jerry Turner from Pixabay	https://pixabay.com/videos/cell-tower-tower-steel-construction-58484/
15	Construction, Equipment, Tools	Video by Rodney Eastman from Pixabay	https://pixabay.com/videos/construction-equipment-tools-139175/

159:2216015080

Dodatek C

Plansze informacyjne wykorzystane w sekwencjach badawczych

C.1 Grafiki informacyjne dla sekwencji badawczych

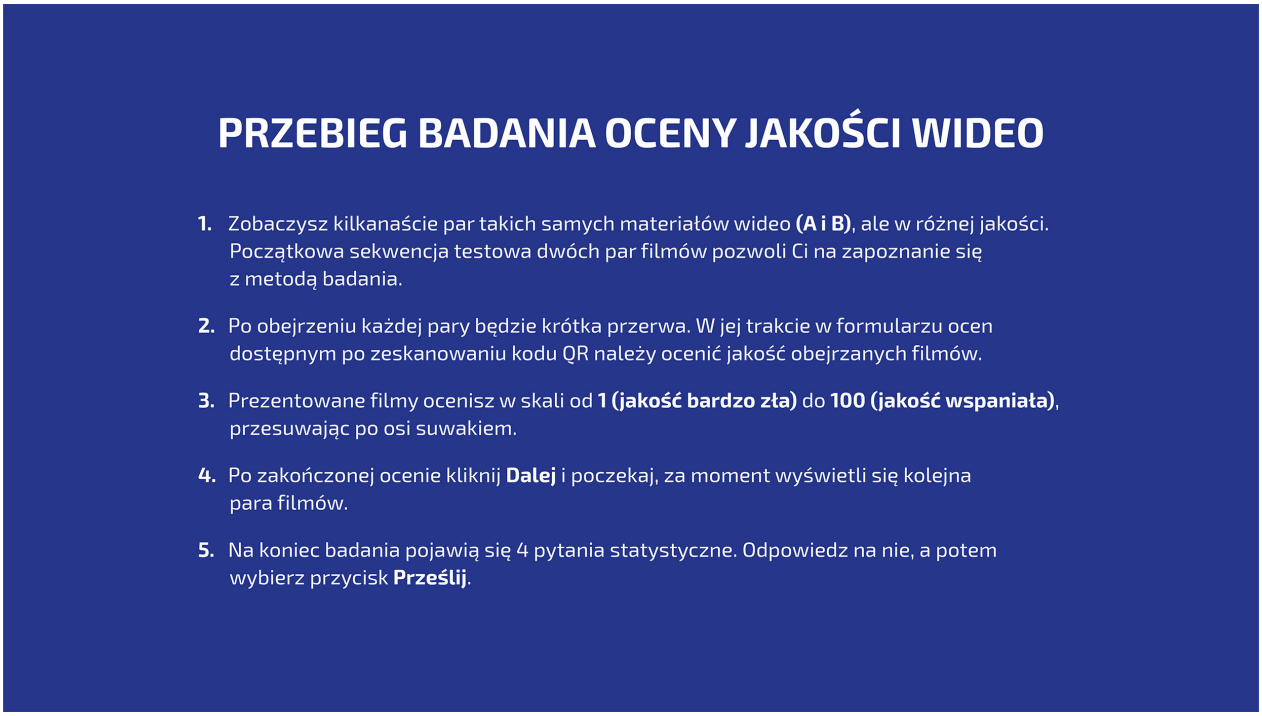
W tej sekcji zostały przedstawione grafiki według kolejności wyświetlania dla sekwencji numer 1. Dla drugiej sekwencji plansza C.1 zostaje zamienione na C.10, natomiast plansza C.3 na C.11.

Grafika C.1 prezentuje informację powitalną, która zostaje wyświetlona po uruchomieniu sekwecji na dekoderze. Grafika wyświetlana jest przez 17 sekund i zawiera informację o anonimowości badania oraz czasie trwania badania nie przekraczającym 25 minut. Zeskanowanie kodu QR powoduje uruchomienie formularza ocen na telefonie.

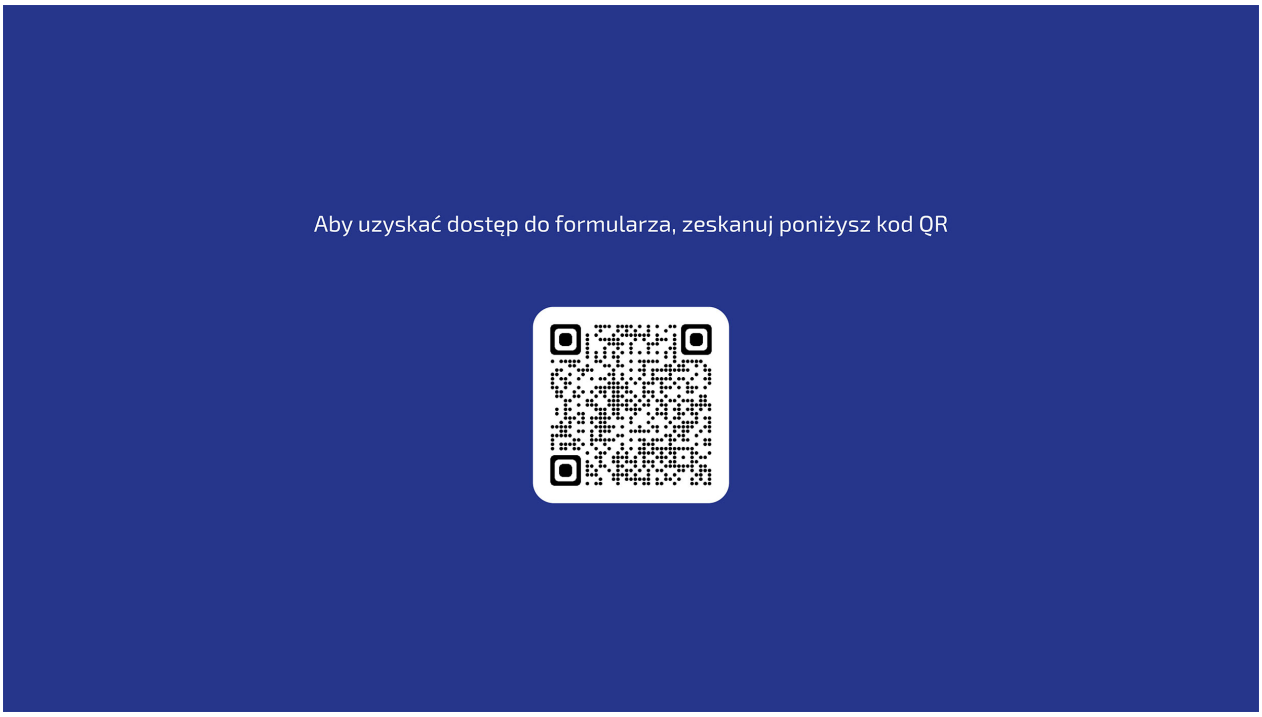


Rysunek C.1 Informacja powitalna dla sekwencji numer 1 wyświetlana 17 sekund.

160:5495860852

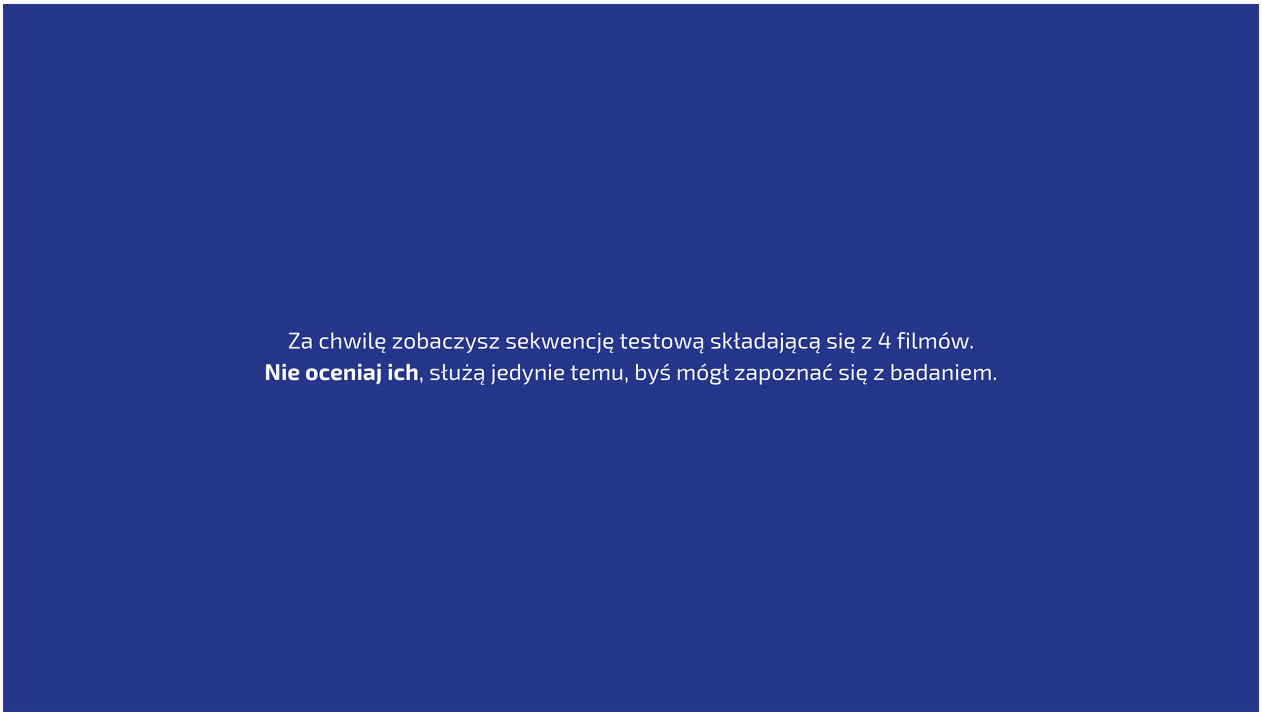


Rysunek C.2 Plansza ze szczegółową instrukcją wyświetlaą przez 40 sekund. W instrukcji przebiegu badania opisano procedurę oceniania treści oraz dostępną skalę ocen.



Rysunek C.3 Powtórzenie kodu QR na planszy trwającej 10 sekund. Sekwencje mają osobne formularze ze względów organizacyjnych związanych z gromadzeniem ocen.

161:6125193886



Rysunek C.4 Informacja o rozpoczęciu sekwencji demonstracyjnej składajcej się z 2 par filmów. Plansza wyświetlana przez 8 sekund.



Rysunek C.5 Informacja o rozpoczęciu prezentacji pierwszego filmu testowego w ramach sesji demonstracyjnej. Każdy z 4 filmów posiada osobną planszę wyświetlaną przez 4 sekundy.

162:1987012127



Rysunek C.6 Informacja o rozpoczęciu właściwej części badania, czyli prezentacji 15 par materiałów testowych. Plansza wyświetlana przez 10 sekund.



Rysunek C.7 Informacja o rozpoczęciu prezentacji filmu 1A. Każdy z 30 filmów posiada własną planszę wyświetlaną przez 4 sekundy.

163:1962574145



Rysunek C.8 Prośba o dokonanie oceny za pomocą formularza. Plansza wyświetlana przez 7 sekund po każej parze filmów.



Rysunek C.9 Podziękowanie za wzięcie udziału w badaniu. Plansza wyświetlana przez 15 sekund na zakończenie sekwencji.

164:3213167774

C.2 Plansze zamienne dla sekwencji badawczej nr 2



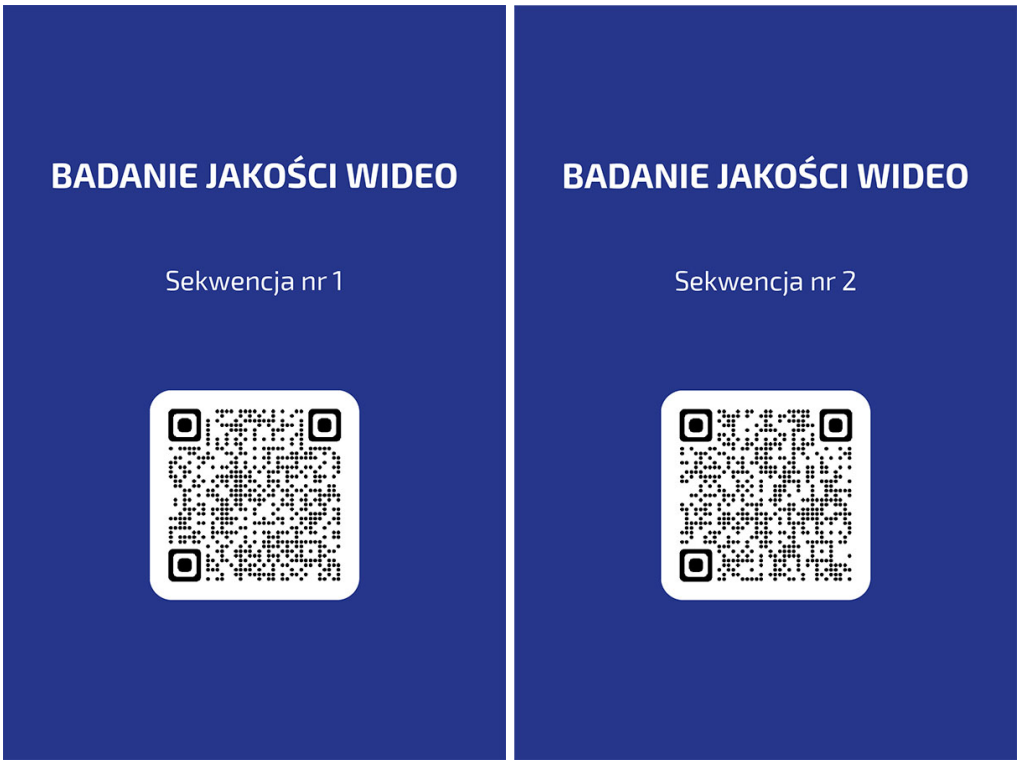
Rysunek C.10 Informacja powitalna dla sekwencji numer 2 wyświetlana 17 sekund.



Rysunek C.11 Ponowienie kodu QR na planszy trwającej 10 sekund.

165:6563140345

C.3 Implementacja badania na dekoderze

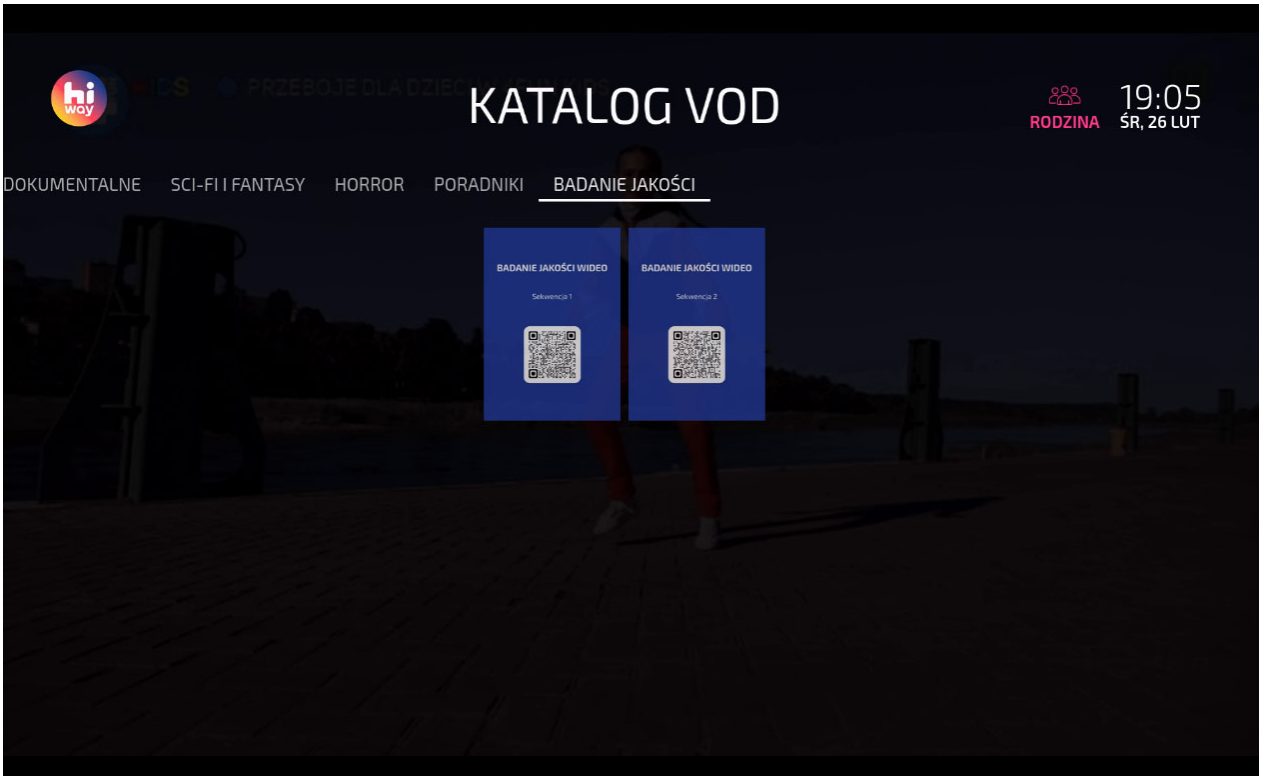


(a) Sekwencja kodowana zgodnie ze standardem H.264

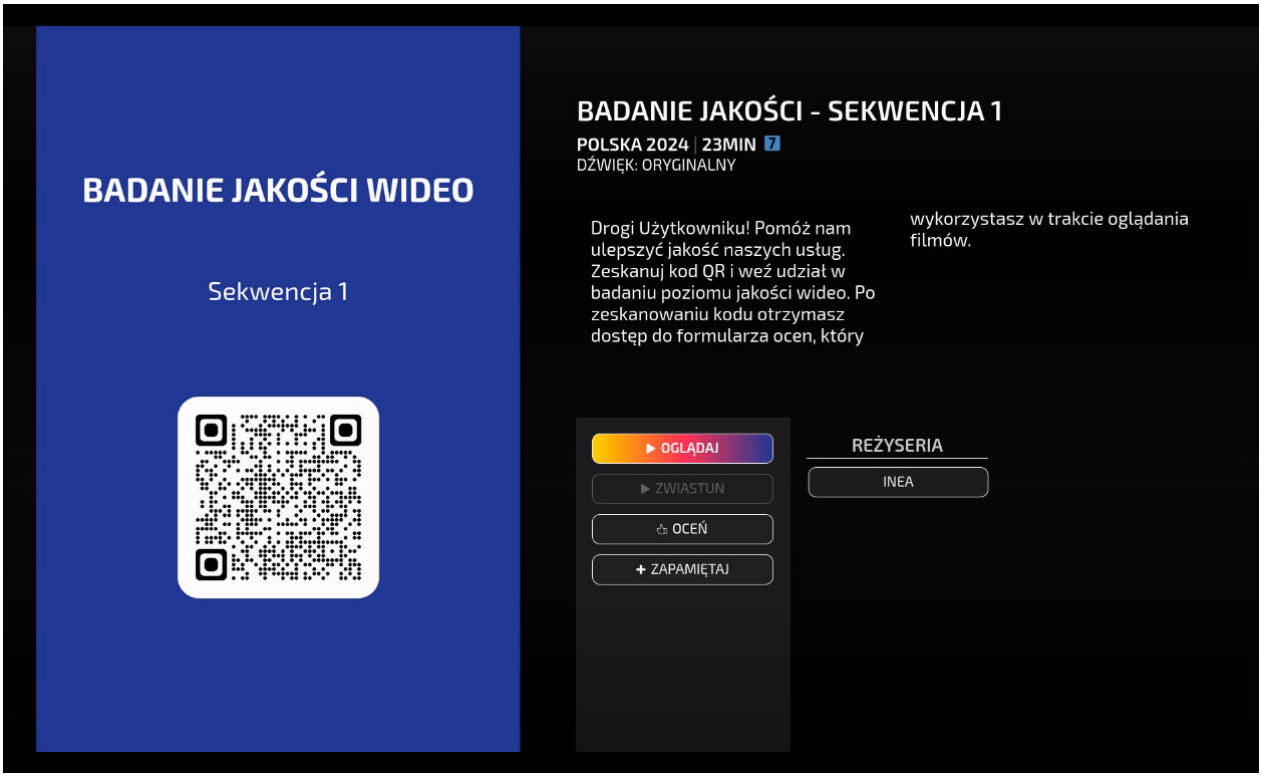
(b) Sekwencja kodowana zgodnie ze standardem H.265

Rysunek C.12 Okładki VOD dla sekwencji badawczych

166:8493223036



Rysunek C.13 Sekwencje badawcze dostępne w bibliotece VOD usługi Hiway.



Rysunek C.14 Film VOD widoczny na dekoderze. Przycisk **Oglądaj** kieruje użytkownika do badania.

167:4394404903

Dodatek D

Szczegółowy zbiór ocen dla badań subiektywnej jakości wideo

168:1149316959

Tabela D.1 Oceny uzyskane od uczestników badania dla sekwencji z materiałami wideo kodowanymi H.264. Skrajna lewa kolumna określa numer pary, natomiast nagłówek A i B określa materiał w ramach danej pary.

	A							B						
#	MIN	Q1	Q2	Q3	95%	AVG	σ	MIN	Q1	Q2	Q3	95%	AVG	σ
1	15,0	64,3	74,0	81,0	99,7	72,5	14,9	13,0	32,0	45,5	65,5	80,9	49,3	20,7
2	18,0	51,3	70,0	81,0	95,8	66,3	20,7	1,0	50,0	65,0	78,8	97,9	61,1	24,5
3	13,0	59,0	70,0	86,8	95,8	70,0	19,8	25,0	63,0	80,0	87,8	100,0	74,9	19,3
4	8,0	55,0	70,5	80,0	90,0	66,6	18,5	1,0	24,3	35,0	60,0	83,9	41,5	24,0
5	27,0	58,3	72,5	81,0	95,0	68,5	18,9	11,0	60,3	70,0	78,8	98,0	68,7	18,1
6	9,0	50,0	63,0	70,8	84,9	60,3	17,6	12,0	60,5	78,0	85,0	99,7	71,8	20,2
7	15,0	51,0	71,0	81,5	93,0	66,3	20,7	8,0	59,0	70,0	80,0	93,8	66,7	19,9
8	12,0	61,5	74,0	83,8	91,0	70,7	17,2	37,0	55,3	70,0	80,0	93,0	67,9	16,9
9	1,0	14,3	33,5	63,5	85,0	40,0	28,6	39,0	66,8	79,5	90,0	100,0	77,6	16,2
10	13,0	52,3	72,0	85,0	94,0	67,8	21,9	11,0	59,3	73,0	80,8	99,7	69,1	19,0
11	8,0	68,5	78,0	85,0	94,9	72,9	18,5	11,0	34,3	56,0	70,8	86,9	53,8	22,7
12	13,0	45,0	68,0	78,0	87,0	62,0	20,6	14,0	47,3	70,0	80,8	90,0	64,4	21,6
13	24,0	72,3	83,0	90,0	100,0	78,9	16,0	16,0	45,5	64,0	75,8	99,7	61,5	22,5
14	7,0	48,0	63,5	79,5	89,0	60,8	20,4	15,0	50,0	65,0	77,8	93,8	62,6	21,5
15	7,0	41,5	68,0	76,8	99,7	62,2	22,6	25,0	60,3	74,5	85,0	95,9	70,8	17,9

Tabela D.2 Oceny uzyskane od uczestników badania dla sekwencji z materiałami wideo kodowanymi H.265. Skrajna lewa kolumna określa numer pary, natomiast nagłówek A i B określa materiał w ramach danej pary.

	A							B						
#	MIN	Q1	Q2	Q3	95%	AVG	σ	MIN	Q1	Q2	Q3	95%	AVG	σ
1	30,0	65,5	76,0	87,5	95,9	74,7	16,1	28,0	67,0	82,0	90,0	99,5	77,6	17,7
2	33,0	60,5	70,0	77,5	90,0	67,8	16,6	20,0	55,0	70,0	79,0	89,9	65,6	18,6
3	20,0	58,5	71,0	84,5	92,9	70,1	18,6	15,0	66,0	77,0	87,0	93,0	72,4	19,6
4	32,0	69,5	80,0	89,5	94,0	76,0	16,2	29,0	65,0	77,0	84,0	94,0	74,2	15,2
5	25,0	70,0	80,0	88,0	98,7	76,7	16,1	36,0	66,0	80,0	87,5	94,0	74,7	16,6
6	40,0	66,5	75,0	90,0	94,0	75,6	14,8	20,0	50,0	71,0	86,0	94,9	68,7	21,0
7	38,0	67,5	79,0	81,5	97,6	74,4	14,4	33,0	68,5	79,0	87,0	99,6	76,7	15,7
8	3,0	59,0	72,0	84,0	92,8	68,2	20,1	23,0	56,0	71,0	84,0	93,9	68,0	19,7
9	20,0	69,0	81,0	88,0	99,7	75,9	18,3	20,0	65,0	77,0	88,5	100,0	74,2	19,8
10	25,0	75,5	85,0	94,5	100,0	82,5	15,8	35,0	73,0	80,0	88,0	99,9	78,5	14,8
11	18,0	58,0	68,0	77,0	87,8	65,4	17,8	18,0	40,5	56,0	72,5	85,4	55,4	19,6
12	1,0	62,0	73,0	81,5	93,0	68,7	20,6	20,0	61,5	80,0	89,5	99,8	75,9	18,3
13	1,0	36,5	61,0	75,0	88,6	57,0	23,1	1,0	37,5	54,0	70,0	86,8	54,6	22,8
14	1,0	52,5	66,0	79,0	92,9	65,1	19,3	1,0	43,0	62,0	78,5	90,9	59,1	25,5
15	34,0	59,0	71,0	85,5	93,8	71,2	15,8	3,0	53,0	66,0	78,5	94,9	64,9	20,9

169:1380129551

Tabela D.3 Różnice ocen pomiędzy materiałami A i B dla danej pary określonej przez skrajną lewą kolumnę. Wartość ujemna oznacza spadek oceny dla materiału B w stosunku do A. Wartość dodatnia to wzrost oceny. Wartości dla sekwencji kodowanej H.264.

#	MIN	Q1	Q2	Q3	95%	AVG	σ
1	-71,0	-40,0	-24,0	-12,0	13,8	-23,2	22,4
2	-78,0	-16,8	-4,5	6,0	38,8	-5,2	23,1
3	-62,0	-12,3	4,0	21,8	43,7	4,9	27,3
4	-78,0	-47,5	-28,0	-7,8	23,6	-25,1	29,1
5	-89,0	-12,5	-0,5	11,5	43,8	0,2	22,3
6	-73,0	1,0	13,5	27,3	44,9	11,6	26,5
7	-85,0	-10,0	0,0	10,0	30,0	0,4	20,4
8	-53,0	-13,0	-3,0	6,8	29,6	-2,8	17,2
9	-21,0	20,8	36,0	59,8	88,8	37,7	31,0
10	-46,0	-11,0	0,0	7,0	42,9	1,3	20,5
11	-86,0	-42,8	-16,0	-0,5	28,7	-19,0	31,4
12	-69,0	-10,0	4,0	19,8	40,0	2,4	25,7
13	-74,0	-34,0	-15,5	1,8	20,7	-17,4	25,8
14	-63,0	-7,5	0,0	9,5	36,7	1,8	19,2
15	-51,0	-9,8	4,5	23,5	49,9	8,6	25,2

Tabela D.4 Różnice ocen pomiędzy materiałami A i B dla danej pary określonej przez skrajną lewą kolumnę. Wartość ujemna oznacza spadek oceny dla materiału B w stosunku do A. Wartość dodatnia to wzrost oceny. Wartości dla sekwencji kodowanej H.265.

#	MIN	Q1	Q2	Q3	95%	AVG	σ
1	-45,0	-5,0	0,0	13,0	25,4	2,9	16,5
2	-48,0	-13,5	0,0	8,0	21,4	-2,2	17,0
3	-49,0	-1,5	0,0	9,0	24,7	2,3	15,4
4	-46,0	-10,0	0,0	8,0	15,0	-1,7	15,8
5	-41,0	-6,0	0,0	5,5	15,9	-2,0	13,8
6	-47,0	-14,5	-4,0	0,5	14,8	-6,9	15,2
7	-40,0	0,0	2,0	10,0	24,9	2,3	14,4
8	-43,0	-10,5	0,0	10,0	20,0	-0,1	16,1
9	-46,0	-7,0	0,0	1,5	16,0	-1,7	16,3
10	-42,0	-7,5	-2,0	0,0	9,8	-4,0	13,2
11	-49,0	-16,0	-10,0	0,0	10,0	-10,0	16,5
12	-28,0	-1,5	5,0	12,0	25,4	7,2	19,6
13	-47,0	-8,5	0,0	5,0	16,6	-2,3	14,9
14	-71,0	-12,0	-2,0	3,5	23,9	-6,0	20,5
15	-70,0	-15,0	-1,0	1,0	10,9	-6,3	16,5

170:1825437708

Dodatek E

Dorobek publikacyjny

W ramach realizacji rozprawy doktorskiej autor opublikował szereg prac naukowych, obejmujących artykuły w czasopismach, rozdziały w monografiach oraz publikacje konferencyjne. Poniżej przedstawiono szczegółowy wykaz dorobku wraz z punktacją (MEiN oraz punktacją czasopism z tzw. listy CORE) oraz zestawienie podsumowujące.

Artykuły w czasopismach naukowych

1. Paweł Andruloniw, Piotr Zwierzykowski, Karol Kowalik: *Subiektywne postrzeganie pojedynczych zniekształceń w telewizji opartej o strumieniowanie adaptacyjne*, *Przegląd Telekomunikacyjny – Wiadomości Telekomunikacyjne*, 2024, nr 4, s. 249–252. **(20 pkt MEiN)**

2. Paweł Andruloniw, Karol Kowalik, Bartosz Partyka, Piotr Zwierzykowski: *Multi-layer QoE Learning System Implemented by Fiberhost*, *Applied Sciences*, 2023, vol. 13, iss. 4, s. 2300-1–2300-23. **(100 pkt MEiN, IF=2.5)**

3. Karol Kowalik, Paweł Andruloniw, Bartosz Partyka, Piotr Zwierzykowski: *Telecom Operator's Approach to QoE*, *Journal of Telecommunications and Information Technology*, 2022, nr 2, s. 26–34. **(40 pkt MEiN)**

4. Paweł Andruloniw, Karol Kowalik, Piotr Zwierzykowski: *Unsupervised Learning Data-Driven Continuous QoE Assessment in Adaptive Streaming-Based Television System*, *Applied Sciences*, 2022, vol. 12, iss. 16, s. 8288-1–8288-23. **(100 pkt MEiN, IF=2.7)**

Rozdziały w monografiach i publikacje konferencyjne

1. Paweł Andruloniw, Piotr Zwierzykowski, Karol Kowalik: *Quality of Experience Studies in TV Systems: Insights and Lessons Learned*, w: *Artificial Intelligence and Machine Learning. 43rd IBIMA Conference, Madrid, 2024*, Springer Nature, 2025, s. 267–273. **(20 pkt MEiN, 70 pkt CORE)**

2. Paweł Andruloniw, Piotr Zwierzykowski, Karol Kowalik: *Quality of Experience Studies in TV Systems: Insights and Lessons Learned*, w: *Artificial Intelligence and Computer Security in*

171:7302959150

Modern Computing: Proceedings of the 43rd IBIMA Conference, 2024, IBIMA Publishing, 2024, s. 474–478. **(5 pkt MEiN, 70 pkt CORE)**

3. Paweł Andruloniw, Piotr Zwierzykowski, Karol Kowalik: *Subiektywna ocena jakości wideo w sieci dostawcy dla usługi telewizyjnej opartej o strumieniowanie adaptacyjne*, w: *Perspektywy rozwoju w naukach inżynieryjno-technicznych – trendy, innowacje i wyzwania. Tom 1*, Wyd. Naukowe TYGIEL, 2024, s. 104–114. **(20 pkt MEiN)**

4. Paweł Andruloniw, Piotr Zwierzykowski, Karol Kowalik: *Quality of Experience Data Gathering in OTT Television Service*, w: *Proceedings of the 41st IBIMA Conference, Granada, 2023*, IBIMA Publishing, 2023, s. 119–122. **(5 pkt MEiN, 70 pkt CORE)**

5. Paweł Andruloniw, Karol Kowalik, Piotr Zwierzykowski: *Analiza skupień dla skończonego oraz ciągłego strumienia danych*, w: *Rozwiązania technologiczne XXI wieku – skutki i perspektywy rozwoju. Tom 1*, Wyd. Naukowe TYGIEL, 2022, s. 7–22. **(20 pkt MEiN)**

6. Paweł Andruloniw, Piotr Zwierzykowski, Karol Kowalik: *Analysis of Adaptive Video Streaming Degradation Based on Clustering Approach*, w: *Proceedings of the 39th IBIMA Conference, Granada, 2022*, IBIMA Publishing, 2022, s. 192–199. **(5 pkt MEiN, 70 pkt CORE)**

7. Paweł Andruloniw, Piotr Zwierzykowski: *W jaki sposób metody uczenia maszynowego mogą pomóc w wykrywaniu degradacji i szacowaniu jakości usług wideo w sieciach telekomunikacyjnych?*, w: *Technologie XXI wieku – aktualne problemy i nowe wyzwania. Tom 3*, Wyd. Naukowe TYGIEL, 2021, s. 132–141. **(20 pkt MEiN)**

Podsumowanie punktacji

Tabela E.1 Podsumowanie punktacji publikacji autora: punkty MEiN oraz punkty po uwzględnieniu wyższej wartości w przypadku podwójnej punktacji.

Kategoria	**Punkty MEiN**	**Punkty (z wyższą punktacją)**
Artykuły w czasopismach	260	260
Rozdziały i konferencje	95	340
Razem	**355**	**600**

Przy uwzględnieniu wyłącznie punktacji MEiN dorobek wynosi **355 pkt**, z czego 73,2% przypada na artykuły w czasopismach, a 26,8% na rozdziały i konferencje. Po uwzględnieniu punktacji z konferencji umieszonych na tzw. liście CORE całkowita liczba punktów wynosi **600**.

172:4403560664

Bibliografia

[1] Andruloniw, Paweł and Kowalik, Karol and Partyka, Bartosz and Zwierzykowski, Piotr. Multi-layer QoE learning system implemented by Fiberhost. *Applied Sciences*, 13(4):2300–1–2300–23, 2023.

[2] Andruloniw, Paweł and Kowalik, Karol and Zwierzykowski, Piotr. Unsupervised learning data-driven continuous QoE assessment in adaptive streaming-based television system. *Applied Sciences*, 12(16):8288–1–8288–23, 2022.

[3] *BT Series: Broadcasting service (television) – Methodologies for the subjective assessment of the quality of television images*, 2023. Recommendation ITU-R BT.500-15.

[4] *Series P: Telephone Transmission Quality – Methods for objective and subjective assessment of quality*, 1996. Recommendation ITU-T P.800.

[5] Netflix. https://www.netflix.com/pl/.

[6] Disney Plus. https://www.disneyplus.com/pl-pl.

[7] HBO Max. https://www.hbomax.com/pl/pl.

[8] Amazon Prime. https://www.primevideo.com/.

[9] R. Giuliano, F. Mazzenga, and A. Vizzarri. Integration of Broadcaster and Telco Access Networks for Real Time/Live Events. *IEEE Transactions on Broadcasting*, 66(3):667–675, 2020.

[10] N.-N. Dao, A.-T. Tran, N. H. Tu, T. T. Thanh, V. N. Q. Bao, and S. Cho. A Contemporary Survey on Live Video Streaming from a Computation-Driven Perspective. *ACM Computing Surveys*, 54:1 – 38, 2022.

[11] L. Peroni and S. Gorinsky. An End-to-End Pipeline Perspective on Video Streaming in Best-Effort Networks: A Survey and Tutorial. *ACM Computing Surveys*, 2024.

[12] *Multi-access Edge Computing (MEC) Framework and Reference Architecture*, 2019. ETSI GS MEC 003.

[13] X. Jiang, F. Yu, T. Song, and V. C. M. Leung. A Survey on Multi-Access Edge Computing Applied to Video Streaming: Some Research Issues and Challenges. *IEEE Communications Surveys & Tutorials*, 23:871–903, 2021.

[14] M. Huang, Q. Shen, Z. Ma, A. C. Bovik, P. Gupta, R. Zhou, and X. Cao. Modeling the Perceptual Quality of Immersive Images Rendered on Head Mounted Displays: Resolution and Compression. *Trans. Img. Proc.*, 27(12):6039–6050, December 2018.

[15] R. Shafi, W. Shuai, and M. Younus. 360-Degree Video Streaming: A Survey of the State of the Art. *Symmetry*, 12:1491, 2020.

[16] A. Yaqoob, T. Bi, and G.-M. Muntean. A Survey on Adaptive 360° Video Streaming: Solutions, Challenges and Opportunities. *IEEE Communications Surveys & Tutorials*, 22:2801–2838, 2020.

[17] D. A. Komathi, A. Suresh, D. J. Lenin, M. Suguna, S. Asha, and C. Srinivasan. Cloud Computing's Effect on Video Games Streaming. *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pages 401–406, 2023.

[18] A. Saha, Y.-C. Chen, C. Davis, B. Qiu, X. Wang, R. Gowda, I. Katsavounidis, and A. Bovik. Study of Subjective and Objective Quality Assessment of Mobile Cloud Gaming Videos. *IEEE Transactions on Image Processing*, 32:3295–3310, 2023.

173:8729344317

[19] N. Barman, E. Jammeh, S. Ghorashi, and M. Martini. No-Reference Video Quality Estimation Based on Machine Learning for Passive Gaming Video Streaming Applications. *IEEE Access*, 7:74511–74527, 2019.

[20] D. Ding, Z. Ma, D. Chen, Q. Chen, Z. Liu, and F. Zhu. Advances in Video Compression System Using Deep Neural Network: A Review and Case Studies. *Proceedings of the IEEE*, 109:1494–1520, 2021.

[21] G. Kougioumtzidis, V. Poulkov, Z. Zaharis, and P. Lazaridis. A Survey on Multimedia Services QoE Assessment and Machine Learning-Based Prediction. *IEEE Access*, 10:19507–19538, 2022.

[22] J. Luo, F. R. Yu, Q. Chen, and L. Tang. Adaptive Video Streaming With Edge Caching and Video Transcoding Over Software-Defined Mobile Networks: A Deep Reinforcement Learning Approach. *IEEE Transactions on Wireless Communications*, 19:1577–1592, 2020.

[23] H. E. Aouifi, M. Hajji, Y. Es-saady, and H. Douzi. Video-Based Learning Recommender Systems: A Systematic Literature Review. *IEEE Transactions on Learning Technologies*, 17:485–497, 2024.

[24] IBC - International Broadcasting Convention. https://show.ibc.org/.

[25] IBC2025 Explores Future of Media Technology. https://www.tvtechnology.com/news/ibc2025-explores-future-of-media-technology.

[26] *Digital Video Broadcasting (DVB); Specification for Service Information (SI) in DVB systems*, 2022. ETSI EN 300 468.

[27] *Advanced video coding for generic audiovisual services*, 2024. ITU-T H.264 (V15).

[28] *High efficiency video coding*, 2021. ITU-T H.265 (V8).

[29] *Information technology — Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s Part 3: Audio*, 1993. ISO/IEC 11172-3.

[30] *Information technology — Generic coding of moving pictures and associated audio information Part 7: Advanced Audio Coding (AAC)*, 2006. ISO/IEC 13818-7.

[31] *Information technology — Coding of audio-visual objects*, 2019. ISO/IEC 14496-3.

[32] *Digital Audio Compression (AC-3, Enhanced AC-3) Standard*, 2017. ETSI TS 102 366.

[33] S. Nadas. *RFC 5798: Virtual Router Redundancy Protocol (VRRP) Version 3 for IPv4 and IPv6*. USA, 2010.

[34] IEEE. *IEEE Standard for Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements—Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications—Amendment: Aggregation of Multiple Link Segments*, 2000.

[35] P. Pan, G. Swallow, and A. Atlas. *RFC 4090: Fast Reroute Extensions to RSVP-TE for LSP Tunnels*. USA, 2005.

[36] G. LLC. *Google Widevine Digital Rights Management*, 2023. Available at: https://www.widevine.com/.

[37] M. Corporation. *Microsoft PlayReady Documentation*, 2023. Available at: https://www.microsoft.com/playready/.

[38] A. Inc. *FairPlay Streaming Overview*, 2023. Available at: https://developer.apple.com/documentation/fairplaystreaming.

[39] Kowalik, Karol and Andruloniw, Paweł and Partyka, Bartosz and Zwierzykowski, Piotr Leszek. Telecom Operator's Approach to QoE. *Journal of Telecommunications and Information Technology*, (2):26–34, 2022.

[40] D. Z. Rodriguez, Z. Wang, R. Rosa, and G. Bressan. The impact of video-quality-level switching on user quality of experience in dynamic adaptive streaming over HTTP. *EURASIP Journal on Wireless Communications and Networking*, 2014:216, 12 2014.

174:4553229163

[41] H. Nam, K.-H. Kim, and H. Schulzrinne. QoE matters more than QoS: Why people stop watching cat videos. pages 1–9, 04 2016.

[42] L. Tu. Density-based clustering for real-time stream data. pages 133–142, 08 2007.

[43] Telewizja Hiway. https://hiway.pl/czym-jest-hiway/.

[44] Y. Sugito and Y. Kusakabe. A comparison of non-experts and experts using DSIS method. *Electronic Imaging*, 34(11):131–1–131–1, 2022.

[45] M. H. Pinson and S. Wolf. Comparing subjective video quality testing methodologies. 5150, 2003.

[46] T. Rahim and S. Y. Shin. Subjective Evaluation of Ultra-high Definition (UHD) Videos. *KSII Trans. Internet Inf. Syst.*, 14:2464–2479, 2020.

[47] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi. Performance comparisons of subjective quality assessment methods for mobile video. *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 82–87, 2010.

[48] *SERIES P: Telephone transmission quality, telephone installations, local line networks – Subjective video quality assessment methods for multimedia applications*, 2023. Recommendation ITU-T P.910.

[49] A. Pastor, P. David, I. Katsavounidis, L. Krasula, A. Norkin, H. Tmar, and P. Callet. "Discriminability–Experimental Cost" Tradeoff in Subjective Video Quality Assessment of Codec: DCR with EVP Rating Scale Versus ACR–HR. *2024 Picture Coding Symposium (PCS)*, pages 1–5, 2024.

[50] J. Gutiérrez, P. Pérez, M. Orduna, A. Singla, C. Cortés, P. Mazumdar, I. Viola, K. Brunnström, F. Battisti, N. Cieplińska, D. Juszka, L. Janowski, M. Leszczuk, A. Adeyemi-Ejeye, Y. Hu, Z. Chen, G. V. Wallendael, P. Lambert, C. Díaz, J. Hedlund, O. Hamsis, S. Fremerey, F. Hofmeyer, A. Raake, P. César, M. Carli, and N. García. Subjective Evaluation of Visual Quality and Simulator Sickness of Short 360° Videos: ITU-T Rec. P.919. *IEEE Transactions on Multimedia*, 24:3087–3100, 2021.

[51] X. Liu, W. Song, Q. He, M. Mauro, and A. Liotta. Speeding Up Subjective Video Quality Assessment via Hybrid Active Learning. *IEEE Transactions on Broadcasting*, 69:165–178, 2023.

[52] Z. Li, C. Bampis, L. Janowski, and I. Katsavounidis. A Simple Model for Subject Behavior in Subjective Experiments. *ArXiv*, abs/2004.02067, 2020.

[53] M. Cheon and J.-S. Lee. Subjective and Objective Quality Assessment of Compressed 4K UHD Videos for Immersive Experience. *IEEE Transactions on Circuits and Systems for Video Technology*, 28:1467–1480, 2018.

[54] J. He, J. Yuan, and F. Xing. DVL2024: A Larger Ultra-High-Definition Video Dataset for Perceptual Quality Study. *IEEE Access*, 13:40429–40437, 2025.

[55] Y. Li, S. Meng, X. Zhang, Y. Wang, and S. Ma. User-Generated Video Quality Assessment: A Subjective and Objective Study. *IEEE Transactions on Multimedia*, 25:154–166, 2020.

[56] B. Naderi and R. Cutler. A Crowdsourcing Approach to Video Quality Assessment. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2810–2814, 2022.

[57] X. Guan, F. Li, Z. Huang, and H. Liu. Study of Subjective and Objective Quality Assessment of Night-Time Videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 32:6627–6641, 2022.

[58] L. Lin, Z. Wang, J. He, W. Chen, Y. Xu, and T. Zhao. Deep Quality Assessment of Compressed Videos: A Subjective and Objective Study. *IEEE Transactions on Circuits and Systems for Video Technology*, 33:2616–2626, 2022.

[59] A. Trioux, G. Valenzise, M. Cagnazzo, M. Kieffer, F. Coudoux, P. Corlay, and M. Gharbi. Subjective and Objective Quality Assessment of the SoftCast Video Transmission Scheme. *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 96–99, 2020.

175:1044847358

[60] S. Jakubczak and D. Katabi. SoftCast: one-size-fits-all wireless video. In *Proceedings of the ACM SIGCOMM 2010 Conference*, SIGCOMM '10, page 449–450, New York, NY, USA, 2010. Association for Computing Machinery.

[61] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[62] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. Toward A Practical Perceptual Video Quality Metric. https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652, 2016.

[63] Z. Wang, E. Simoncelli, and A. Bovik. Multiscale structural similarity for image quality assessment. volume 2, pages 1398 – 1402 Vol.2, 12 2003.

[64] P. Dymarski, S. Kula, and T. N. Huy. QoS Conditions for VoIP and VoD. *Journal of Telecommunications and Information Technology*, 2023.

[65] M. Uhrina, A. Holesova, J. Bienik, and L. Sevcik. Impact of Scene Content on High Resolution Video Quality. *Sensors (Basel, Switzerland)*, 21, 2021.

[66] D. Keller, P. R. Frank, S. Goering, and A. Raake. Investigating the Impact of High Frame Rate on Video Quality: A SAMVIQ Approach. *2024 International Symposium on Multimedia (ISM)*, pages 143–144, 2024.

[67] S. Bae, J. il Kim, M. Kim, S. Cho, and J. Choi. Assessments of Subjective Video Quality on HEVC-Encoded 4K-UHD Video for Beyond-HDTV Broadcasting Services. *IEEE Transactions on Broadcasting*, 59:209–222, 2013.

[68] E. Pei, H. Lee, E. Fedorovskaya, and S. Farnand. Evaluation of Subjective Video Quality of Television Displays. *Electronic Imaging*, 2024.

[69] N. Staelens, S. Moens, W. V. D. Broeck, I. Mariën, B. Vermeulen, P. Lambert, R. Walle, and P. Demeester. Assessing Quality of Experience of IPTV and Video on Demand Services in Real-Life Environments. *IEEE Transactions on Broadcasting*, 56:458–466, 2010.

[70] Review: Harmonic ProMedia Xpress. https://www.streamingmedia.com/articles/editorial/featured-articles/review-harmonic-promedia-xpress-90646.aspx.

[71] Movie Maker - Video Editor. https://apps.microsoft.com/detail/9mvfq4lmz6c9?hl=en-US&gl=US.

[72] FFmpeg - A complete, cross-platform solution to record, convert and stream audio and video. https://www.ffmpeg.org/.

[73] *CableLabs® Content 3.0 Specification*, 2015. Metadata Specifications Metadata 3.0.

[74] Jotform. https://www.jotform.com/.

[75] W. Cochran. *Sampling Techniques*. Wiley Series in Probability and Statistics. Wiley, 1977.

[76] A. Mood, F. Graybill, and D. Boes. *Introduction to the Theory of Statistics*. International Student edition. McGraw-Hill, 1973.

[77] Support Vector Machines. https://scikit-learn.org/stable/modules/svm.html.

[78] H. Sheikh and A. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006.

[79] S. Li, F. Zhang, L. Ma, and K. N. Ngan. Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments. *IEEE Transactions on Multimedia*, 13(5):935–949, 2011.

[80] S. Wolf and M. H. Pinson. Video Quality Model for Variable Frame Delay (VQM_VFD). *Technical Memorandum NTIA TM-11-482*, 2011.

[81] NTIA. https://www.ntia.gov/.

176:1132651284

[82] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz. Quantification of YouTube QoE via Crowdsourcing. In *2011 IEEE International Symposium on Multimedia*, pages 494–499, 2011.

[83] YouTube. https://www.youtube.com/.

[84] D. Z. Rodriguez, J. Abrahao, D. C. Begazo, R. L. Rosa, and G. Bressan. Quality metric to assess video streaming service over TCP considering temporal location of pauses. *IEEE Transactions on Consumer Electronics*, 58(3):985–992, 2012.

[85] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, and Z. Wang. A Quality-of-Experience Index for Streaming Video. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):154–166, 2017.

[86] C. G. Bampis and A. C. Bovik. Learning to Predict Streaming Video QoE: Distortions, Rebuffering and Memory, 2017.

[87] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli. A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP. *SIGCOMM Comput. Commun. Rev.*, 45(4):325–338, August 2015.

[88] R. K. P. Mok, X. Luo, E. W. W. Chan, and R. K. C. Chang. QDASH: a QoE-aware DASH system. In *Proceedings of the 3rd Multimedia Systems Conference*, MMSys '12, page 11–22, New York, NY, USA, 2012. Association for Computing Machinery.

[89] J. Klink, M. Łuczyński, and S. Brachmański. Video Quality Modelling—Comparison of the Classical and Machine Learning Techniques. *Applied Sciences*, 14(16), 2024.

[90] A. Bentaleb, A. C. Begen, and R. Zimmermann. SDNDASH: Improving QoE of HTTP Adaptive Streaming Using Software Defined Networking. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, page 1296–1305, New York, NY, USA, 2016. Association for Computing Machinery.

[91] D. Ghadiyaram, J. Pan, and A. C. Bovik. Learning a Continuous-Time Streaming Video QoE Model. *IEEE Transactions on Image Processing*, 27(5):2257–2271, 2018.

[92] Z. Duanmu, W. Liu, D. Chen, Z. Li, Z. Wang, Y. Wang, and W. Gao. A Knowledge-Driven Quality-of-Experience Model for Adaptive Streaming Videos, 2019.

[93] Z. Zhou, Y. Dong, L. Song, R. Xie, L. Li, and B. Zhou. Quality of Experience Evaluation for Streaming Video Using CGNN. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 285–288, 2020.

[94] R. Oliver. Effect of Expectation and Disconfirmation on Postexposure Product Evaluations: An Alternative Interpretation. *Journal of Applied Psychology*, 62:480–486, 08 1977.

[95] R. L. Oliver. A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions. *Journal of Marketing Research*, 17:460 – 469, 1980.

[96] Z. Duanmu, K. Ma, and Z. Wang. Quality-of-Experience for Adaptive Streaming Videos: An Expectation Confirmation Theory Motivated Approach. *IEEE Transactions on Image Processing*, 27(12):6135–6146, 2018.

[97] N. Eswara, S. Ashique, A. Panchbhai, S. Chakraborty, H. P. Sethuram, K. Kuchi, A. Kumar, and S. S. Channappayya. Streaming Video QoE Modeling and Prediction: A Long Short-Term Memory Approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(3):661–673, 2020.

[98] C. G. Bampis, Z. Li, and A. C. Bovik. Continuous Prediction of Streaming Video QoE Using Dynamic Networks. *IEEE Signal Processing Letters*, 24(7):1083–1087, 2017.

[99] N. Eswara, H. P. Sethuram, S. Chakraborty, K. Kuchi, A. Kumar, and S. S. Channappayya. Modeling Continuous Video QoE Evolution: A State Space Approach. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2018.

[100] C. G. Bampis, Z. Li, I. Katsavounidis, and A. C. Bovik. Recurrent and Dynamic Models for Predicting Streaming Video Quality of Experience. *IEEE Transactions on Image Processing*, 27(7):3316–3331, 2018.

177:7565485404

[101] H. Zhang, L. Dong, G. Gao, H. Hu, Y. Wen, and K. Guan. DeepQoE: A Multimodal Learning Framework for Video Quality of Experience (QoE) Prediction. *IEEE Transactions on Multimedia*, 22(12):3210–3223, 2020.

[102] M. De Filippo De Grazia, D. Zucchetto, A. Testolin, A. Zanella, M. Zorzi, and M. Zorzi. QoE Multi-Stage Machine Learning for Dynamic Video Streaming. *IEEE Transactions on Cognitive Communications and Networking*, 4(1):146–161, 2018.

[103] Y. Zhu, T. Sun, Q. Li, L. Lu, X. Duan, and W. Li. Machine Learning based User QoE Evaluation for Video Streaming over Mobile Network. In *2020 IEEE International Conference on Smart Data Services (SMDS)*, pages 18–25, 2020.

[104] Y.-T. Lin, E. M. R. Oliveira, S. Ben Jemaa, and S. E. Elayoubi. Machine learning for predicting QoE of video streaming in mobile networks. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6, 2017.

[105] G. Huang, O. Ercetin, H. Gokcesu, and G. Kalem. Deep Learning-Based QoE Prediction for Streaming Services in Mobile Networks. In *2022 18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 327–332, 2022.

[106] S. Schwarzmann, C. Cassales Marquezan, M. Bosk, H. Liu, R. Trivisonno, and T. Zinner. Estimating Video Streaming QoE in the 5G Architecture Using Machine Learning. In *Proceedings of the 4th Internet-QoE Workshop on QoE-Based Analysis and Management of Data Communication Networks*, Internet-QoE'19, page 7–12, New York, NY, USA, 2019. Association for Computing Machinery.

[107] R. Ul Mustafa, S. Ferlin, C. Esteve Rothenberg, D. Raca, and J. J. Quinlan. A Supervised Machine Learning Approach for DASH Video QoE Prediction in 5G Networks. In *Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks*, Q2SWinet '20, page 57–64, New York, NY, USA, 2020. Association for Computing Machinery.

[108] M. Ghosh, D. C. Singhal, and R. Wayal. DeSVQ: Deep Learning Based Streaming Video QoE Estimation. In *Proceedings of the 23rd International Conference on Distributed Computing and Networking*, ICDCN '22, page 19–25, New York, NY, USA, 2022. Association for Computing Machinery.

[109] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM, August 2016.

[110] S. Wassermann, N. Wehner, and P. Casas. Machine Learning Models for YouTube QoE and User Engagement Prediction in Smartphones. 46(3):155–158, January 2019.

[111] S. Altamimi and S. Shirmohammadi. QoE-Fair DASH Video Streaming Using Server-side Reinforcement Learning. 16(2s), June 2020.

[112] *Information technology — Dynamic adaptive streaming over HTTP (DASH). Part 1: Media presentation description and segment formats*, 2022. ISO/IEC 23009-1:2022.

[113] G. Dimopoulos, I. Leontiadis, P. Barlet-Ros, and K. Papagiannaki. Measuring Video QoE from Encrypted Traffic. In *Proceedings of the 2016 Internet Measurement Conference*, IMC '16, page 513–526, New York, NY, USA, 2016. Association for Computing Machinery.

[114] C. Gutterman, K. Guo, S. Arora, T. Gilliland, X. Wang, L. Wu, E. Katz-Bassett, and G. Zussman. Requet: Real-Time QoE Metric Detection for Encrypted YouTube Traffic. 16(2s), July 2020.

[115] M. Muraszkiewicz and R. Nowak. *Sztuczna inteligencja dla inżynierów. Metody ogólne.* Oficyna Wydawnicza Politechniki Warszawskiej, 2022.

[116] D. J. C. MacKay. *Information Theory, Inference & Learning Algorithms.* Cambridge University Press, USA, 2002.

[117] J. Koronacki and J. Mielniczuk. *Statystyka dla studentów kierunków technicznych i przyrodniczych.* Wydawnictwo Naukowo-Techniczne, 2001.

[118] M. Szeliga. *Praktyczne uczenie maszynowe.* Wydawnictwo Naukowe PWN, Warszawa, 2019.

[119] scikit-learn Machine Learning in Python. https://scikit-learn.org/.

178:3281288397

[120] J. Koronacki and J. Ćwik. *Statystyczne systemy uczące się. Wydanie drugie.* Akademicka Oficyna Wydawnicza EXIT Andrzej Lang, 2015.

[121] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. 15(1):1929–1958, jan 2014.

[122] JupyterLab: A Next-Generation Notebook Interface. https://jupyter.org/.

[123] PyTorch. https://pytorch.org/.

[124] NumPy - The fundamental package for scientific computing with Python. https://numpy.org/.

[125] SciPy - Fundamental algorithms for scientific computing in Python. https://scipy.org/.

[126] Pandas – Python Data Analysis Library. https://pandas.pydata.org/.

[127] Optuna - A hyperparameter optimization framework. https://optuna.org/.

[128] M. Kendall and J. D. Gibbons. *Rank Correlation Methods.* A Charles Griffin Title, 1990.

[129] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June 2002.

179:2313920735