



EMOTION RECOGNITION

ROZPOZNAWANIE EMOCJI Z NAGRAŃ AUDIO

Natalia Gotkiewicz | Angelika Popiela | Jagoda Spychała

PLAN PREZENTACJI

- O1. CEL PROJEKTU**
- O2. DANE**
- O3. EDA**
- O4. MODELOWANIE**
- O5. ZAAWANSOWANA ANALIZA**
- O6. APLIKACJA WEBOWA**
- O7. WNIOSKI**

CEL PROJEKTU



PROBLEM BADAWCZY

Projekt dotyczy rozpoznawania emocji w mowie, co stanowi wyzwanie w kontekście analizy dźwięku, ponieważ emocje mogą być subtelne, różne i zależne od kontekstu.



ISTOTA PROJEKTU

Rozpoznawanie emocji z mowy pozwala tworzyć bardziej empatyczne technologie – wspiera diagnozowanie zaburzeń, ulepsza interakcje człowiek-komputer i zwiększa użyteczność systemów głosowych.



CELE I ZAŁOŻENIA

Celem projektu było stworzenie modelu uczenia maszynowego do klasyfikacji emocji z nagrań głosowych oraz zbudowanie aplikacji, która umożliwia jego zastosowanie w praktyce.

DANE



POCHODZENIE

Zbiór danych **nEMO** pochodzi z badań nad rozpoznawaniem emocji w mowie. Został opracowany przez zespół z Uniwersytetu Adama Mickiewicza w Poznaniu.



CHARAKTERYSTYKA

- **Liczba próbek:** 4 481 nagrani
- **Emocje:** radość, smutek, złość, strach, zaskoczenie, neutralność
- **Mówcy:** 9 aktorów
- Dane czyste, bez jakichkolwiek braków, wartości zerowych itp.

EDA

REPREZENTACJE SYGNAŁU AUDIO



Dlaczego przekształcamy surowy sygnał audio?

Surowe próbki dźwięku nie zawierają bezpośrednio informacji o cechach istotnych dla rozpoznawania emocji. Transformacje sygnału podkreślają różne aspekty dźwięku ważne w analizie mowy emocjonalnej.

REPREZENTACJE SYGNAŁU AUDIO



Reprezentacje czasowo - częstotliwościowe

- Spektrogram
- Spektrogram w skali Melowej
- CQT (Constant-Q Transform)



Cechy sygnału audio

CEPSTRALNE

- MFCC
(Mel-Frequency Cepstral Coefficients)

TONALNE

- Cechy chromatyczne (Chroma)
- Tonnetz (Siatka Tonalna)



Techniki dekompozycji

- Separacja Harmoniczno-Perkusyjna

DOMENY CZASU

- RMS (Root Mean Square)
- ZCR (Zero Crossing Rate)

WIDMOWE

- Spectral Contrast

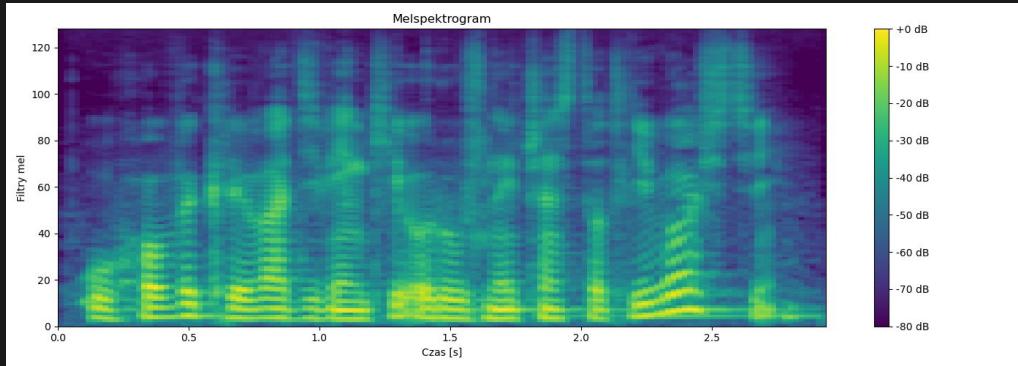
TEMPORALNE (RYTM I TEMPO)

- Tempogram
- Delta tempogram

NAJLEPSZE REPREZENTACJE

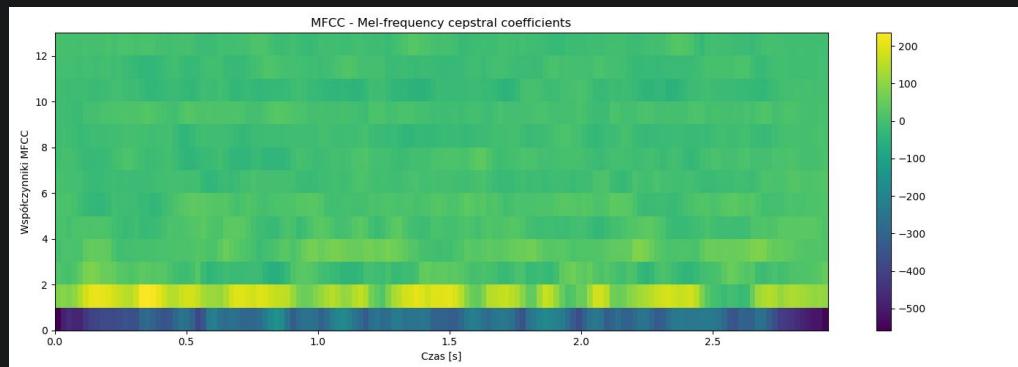
Spektrogram Mela

Jest to przekształcony spektrogram, który odwzorowuje percepcję częstotliwości przez człowieka. Dzięki zwiększonej rozdzielczości w niskich tonach jest szczególnie przydatny w analizie mowy i emocji głosowych.

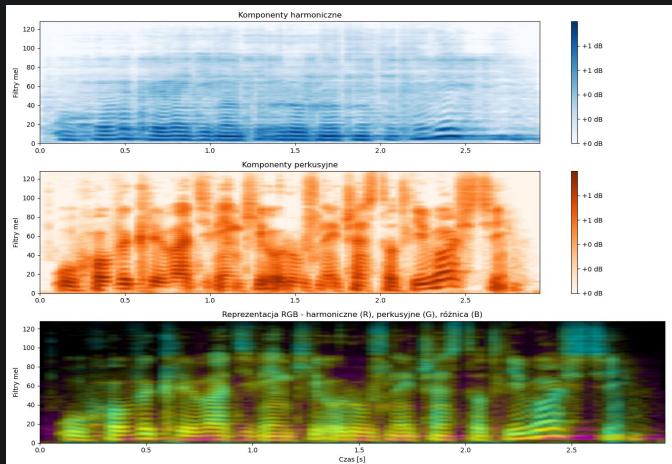


MFCC

Mel-Frequency Cepstral Coefficients to zwięzła reprezentacja cech głosu, naśladowająca sposób, w jaki ludzki układ słuchowy przetwarza dźwięki. Powstają przez przekształcenie sygnału dźwiękowego w zestaw wartości liczbowych opisujących jego strukturę spektralną.



NAJLEPSZE REPREZENTACJE

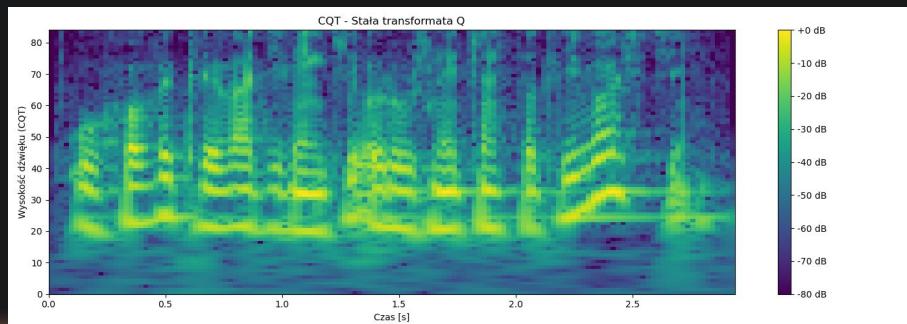


Separacja Harmoniczno-Perkusyjna

HPS rozdziela dźwięk na:

Harmoniczny: Niesie informacje o intonacji i barwie głosu, kluczowe dla emocji (np. podwyższona intonacja w radości).

Perkusyjny: Odpowiada za rytm i energię mowy, również istotne dla emocji (np. szybsze tempo w złości). Dzięki temu model analizuje emocje z "melodii" i "rytmu" głosu osobno, co zwiększa jego skuteczność.



Constant-Q Transform (CQT)

CQT to reprezentacja czasowo-częstotliwościowa dźwięku, która dzięki logarytmicznej skali częstotliwości oferuje lepszą rozdzielcość dla niskich częstotliwości i lepszą rozdzielcość czasową dla wysokich. Skuteczniej niż standaryzowany spektrogram oddaje niuanse intonacji i harmonicznych, kluczowe w analizie emocji.

MODELOWANIE

ZASTOSOWANE MODELE



RODZAJE

W projekcie wykorzystano różne modele głębokiego uczenia:

ResNet18 dla jednej reprezentacji oraz jego wersję **ensemble**, w której każda sieć przetwarza inną reprezentację dźwięku. Użyto również modelu **VGG16**, znanego z dobrej analizy danych obrazowych, oraz dwóch wariantów konwolucyjnych sieci: **Simple CNN** i **autorskiej CNN**. Do sekwencyjnych danych czasowych zastosowano **Conv1D_RNN**, łączący cechy konwolucyjne z pamięcią czasową.



DLACZEGO?

Zróżnicowane modele pozwolłyły lepiej uchwycić złożone cechy mowy. Sieci konwolucyjne dobrze wykrywają lokalne wzorce w spektrogramach, a rekurencyjne ułatwiają analizę zmian w czasie. Ensemble ResNet łączy zalety różnych reprezentacji, co zwiększa dokładność klasyfikacji emocji. Dzięki temu projekt lepiej radzi sobie z różnorodnością sygnałów audio.

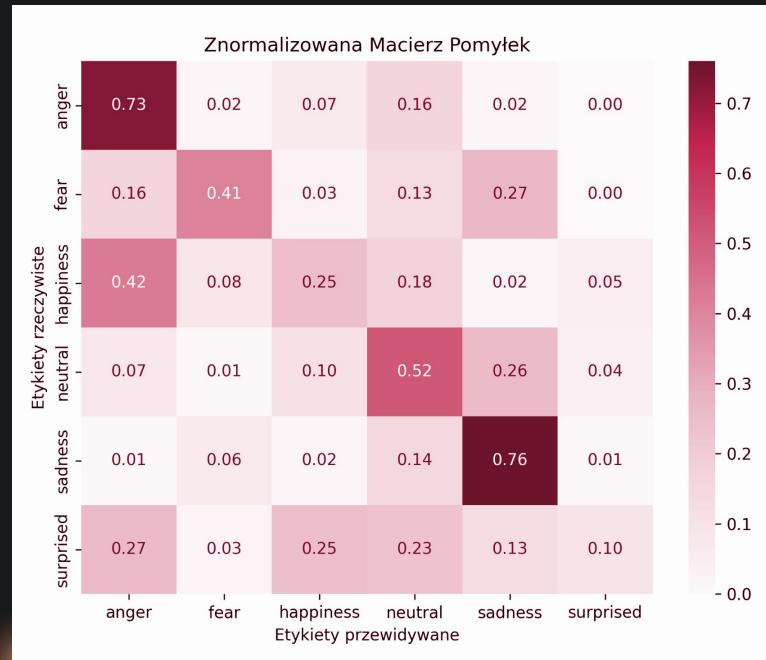
KONWOLUCYJNA SIEĆ NEURONOWA (CNN)

WŁASNA ARCHITEKTURA

Sieć konwolucyjna (CNN) ekstrahuje cechy z obrazów spektrogramów Mel'a, wyszukując charakterystyczne wzorce akustyczne dla poszczególnych emocji. Składa się z trzech bloków konwolucyjnych (Conv2D, MaxPooling2D, Dropout 0.3) z rosnącą liczbą filtrów $32 \rightarrow 64 \rightarrow 128$.

Po ekstrakcji cech dane redukuje warstwa GlobalAveragePooling2D, następnie analizuje warstwa Dense (128 neuronów, Dropout 0.3), a klasyfikacji do sześciu klas dokonuje warstwa wyjściowa z aktywacją Softmax.

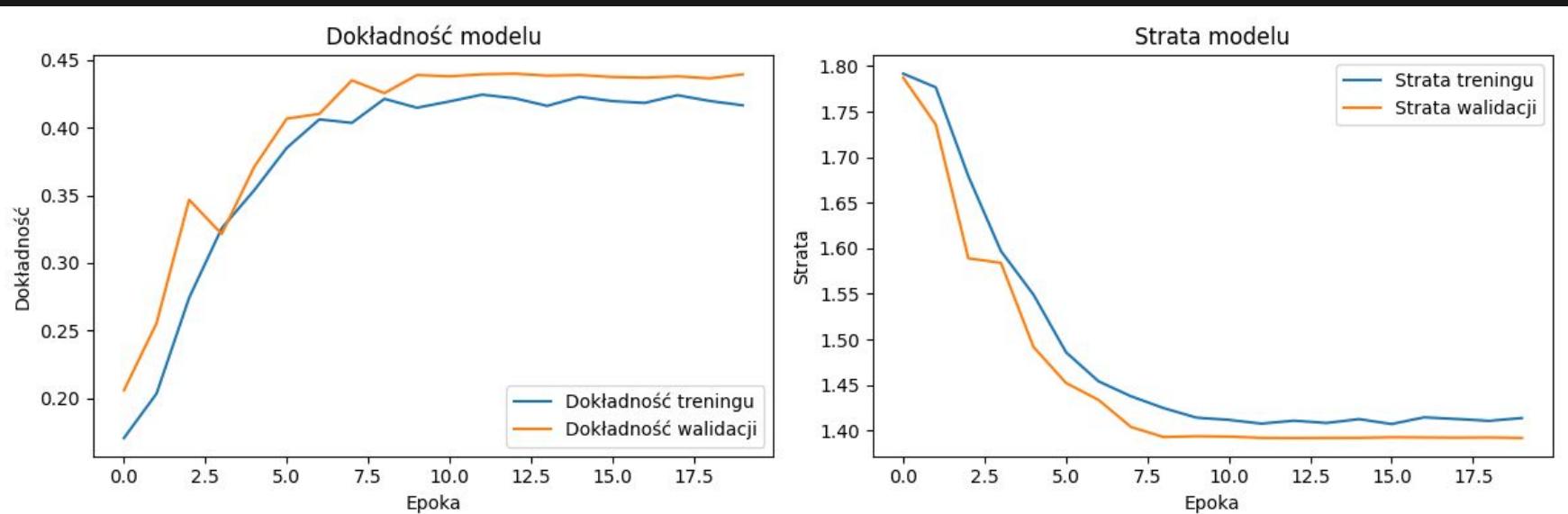
Ogólna dokładność klasyfikacji wyniosła 47%. Najwyższą skuteczność osiągnięto w rozpoznawaniu emocji smutek (76%) i złość (73%), natomiast największe trudności pojawiły się przy klasyfikacji zaskoczenia (10%) i szczęścia (25%).



KONWOLUCYJNA SIEĆ NEURONOWA (CNN)



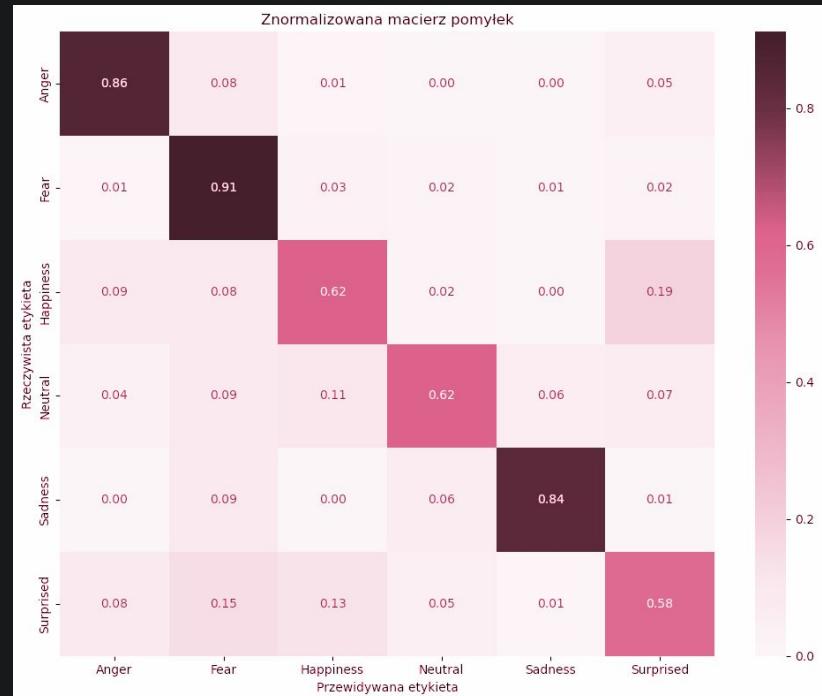
HISTORIA TRENINGU



VGG16 - KONWOLUCYJNA SIEĆ NEURONOWA

VGG16 to głęboka sieć konwolucyjna (CNN) o 16 warstwach, wykorzystująca stosy filtrów 3x3. Przetwarza ona dwuwymiarowe reprezentacje audio (np. spektrogramy) jako obrazy. Dzięki temu VGG16 efektywnie wydobywa hierarchiczne cechy czasowo-częstotliwościowe, kluczowe dla uczenia się wzorców emocjonalnych, mimo że istnieją nowsze, wydajniejsze architektury.

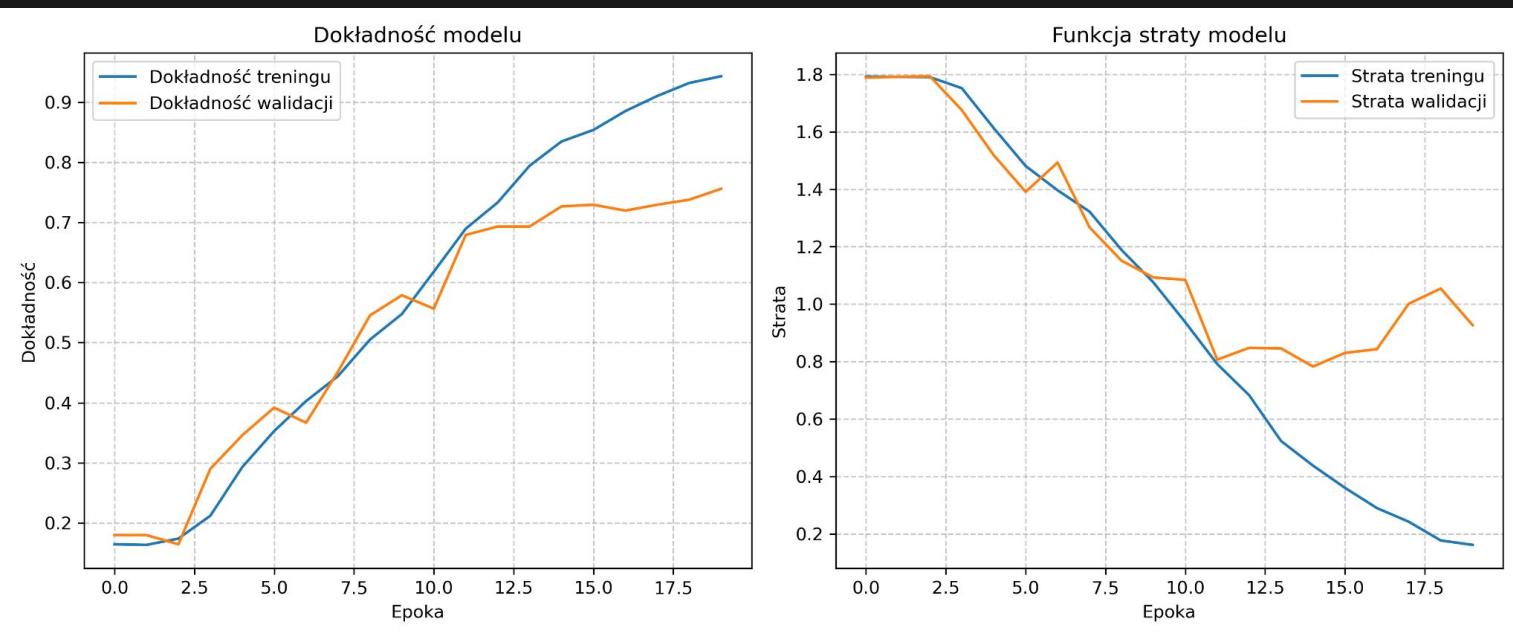
Ogólna dokładność klasyfikacji wyniosła 74%. Najwyższą skuteczność osiągnięto w rozpoznawaniu emocji negatywnych takich jak strach (91%), złość (86%) i smutek (84%) natomiast największe trudności pojawiły się przy klasyfikacji zaskoczenia (58%).



VGG16 - KONWOLUCYJNA SIEĆ NEURONOWA



HISTORIA TRENINGU



SIMPLE CNN (SUROWE AUDIO)

Konwolucyjna sieć neuronowa 1D zaprojektowana do bezpośredniego przetwarzania surowych sygnałów audio, z pominięciem etapu konwersji na reprezentacje obrazowe. Zawiera cztery bloki konwolucyjne z progresywnie rosnącą liczbą filtrów ($32 \rightarrow 64 \rightarrow 128 \rightarrow 256$), każdy z normalizacją wsadową i max-poolingiem, zakończone globalnym poolingiem adaptacyjnym.

Model wykorzystuje zaawansowane techniki regularyzacji (dwupoziomowy dropout, normalizacja wsadowa) oraz architekturę "bottleneck" w części klasyfikacyjnej ($256 \rightarrow 512 \rightarrow 6$).

Osiąga 82.76% dokładności ogólnej na zbiorze testowym, efektywnie ucząc się wzorców emocjonalnych bezpośrednio z próbek czasowych sygnału dźwiękowego.

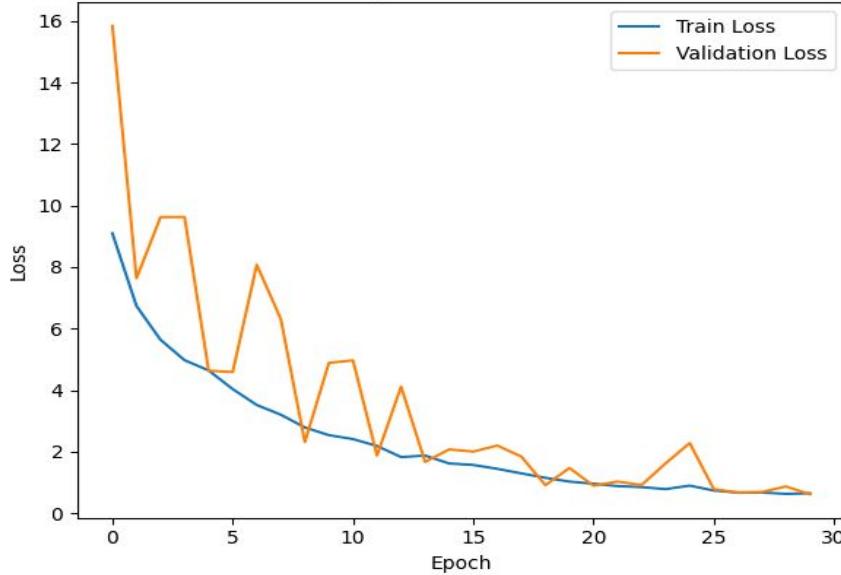


SIMPLE CNN (SUROWE AUDIO)

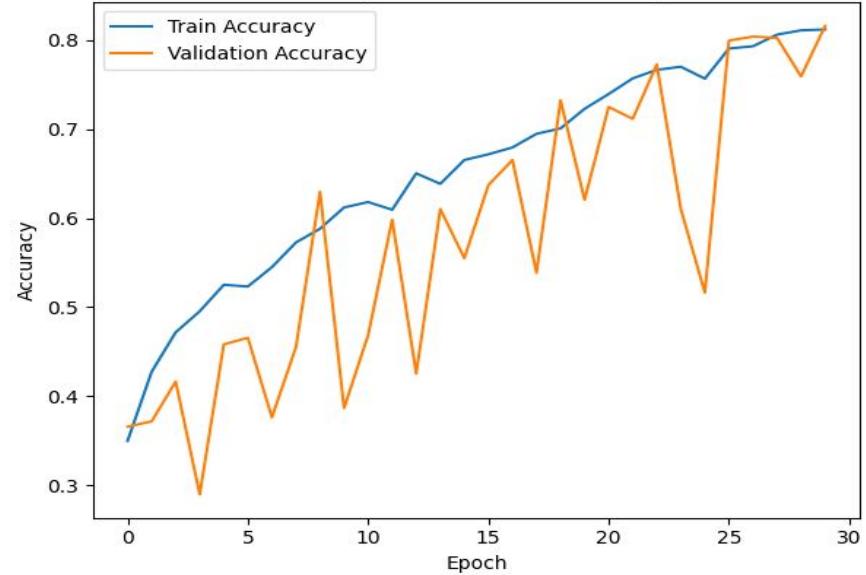


HISTORIA
TRENINGU

Training and Validation Loss



Training and Validation Accuracy



ResNET-18 (SPEKTRGRAMY MELA)

Konwolucyjna sieć neuronowa oparta na architekturze ResNet-18, dostosowana do analizy dźwięku poprzez modyfikację warstwy wejściowej na jednokanałową. Bloki rezydualne z połączeniami skrótowymi umożliwiają efektywne uczenie głębszej struktury sieci i rozpoznawanie złożonych wzorców akustycznych.

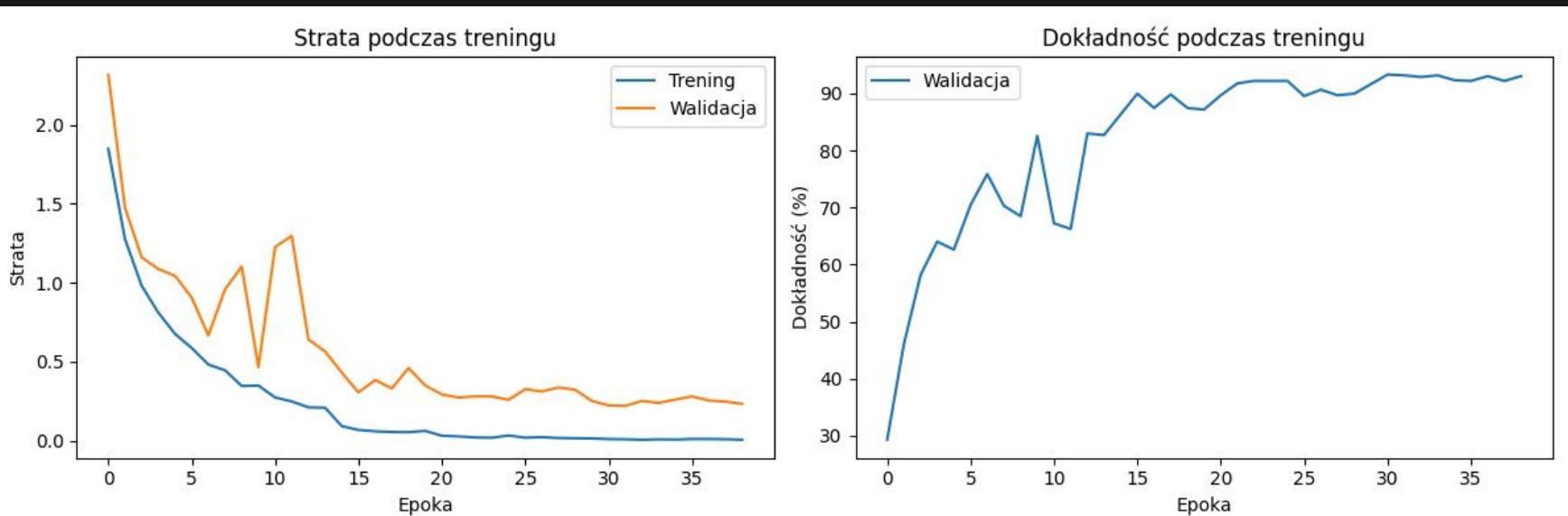
Model osiąga 91.3% dokładności ogólnej, szczególnie wysokie wyniki dla emocji: smutek (98% recall), strach (96% F1) i stan neutralny (95% recall).



ResNET-18

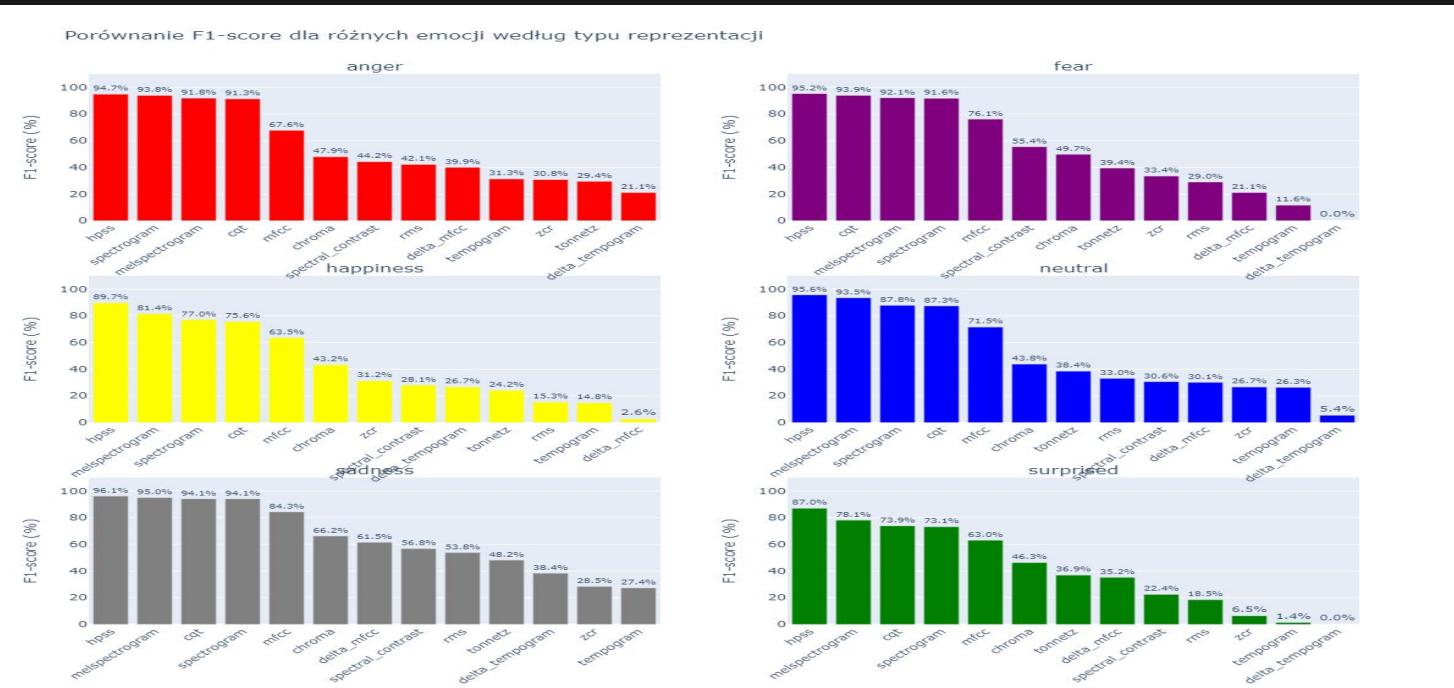


HISTORIA TRENINGU



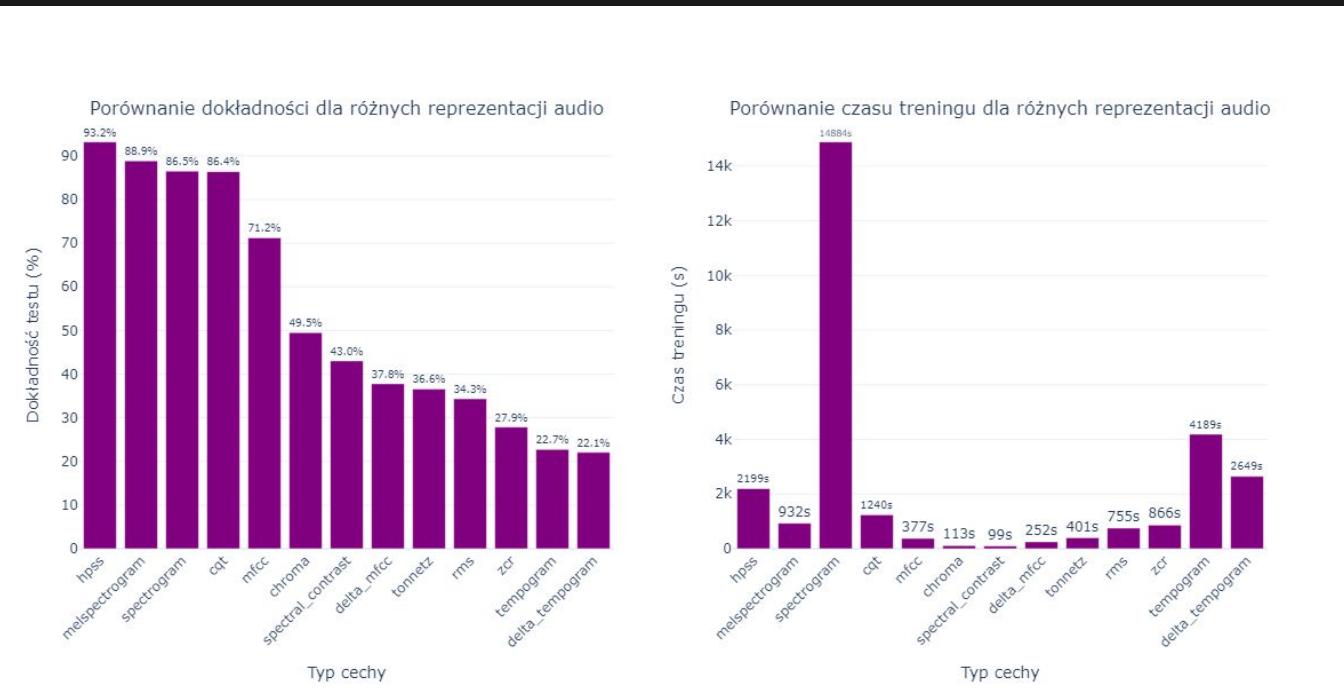
MODEL ResNET

Wyniki treningu każdej z 13 reprezentacji dźwięku



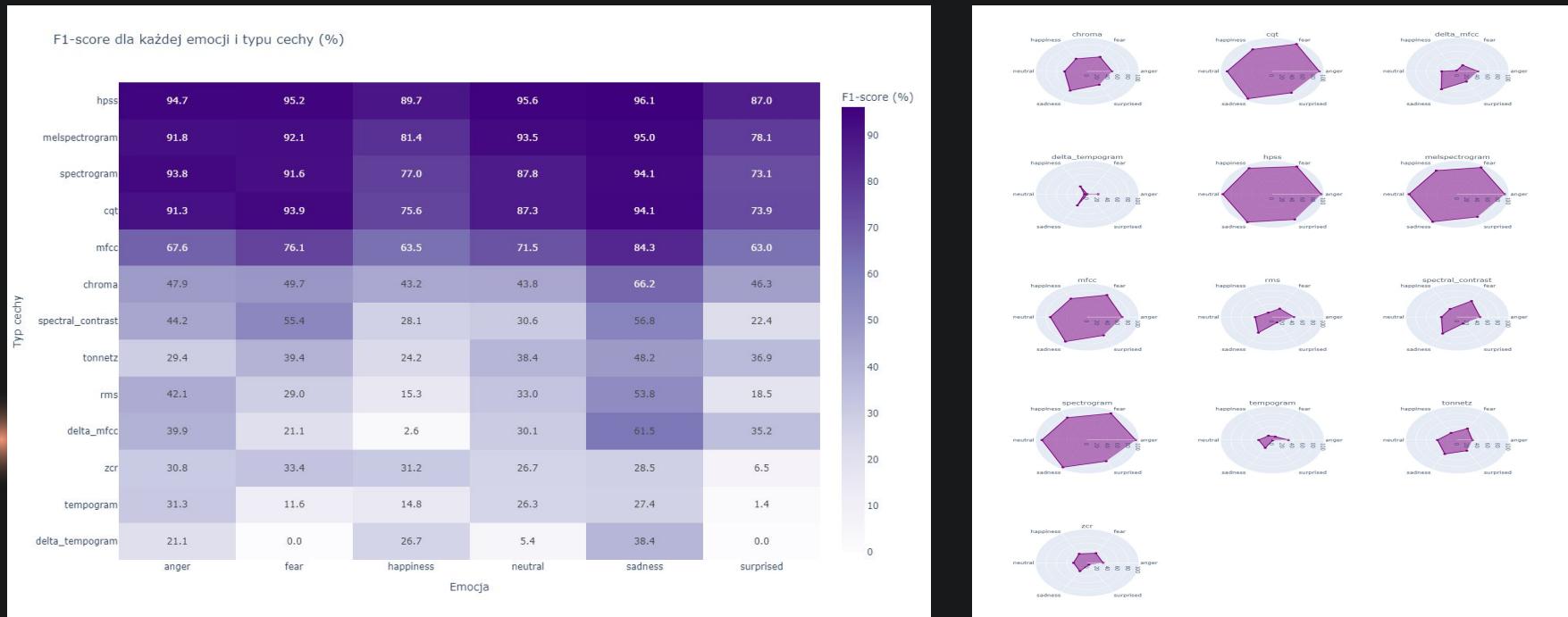
MODEL ResNET

Wyniki treningu każdej z 13 reprezentacji dźwięku



MODEL ResNET

Wyniki treningu każdej z 13 reprezentacji dźwięku



MODEL ENSEMBLE



Architektura

Model ensemble z ważonym uśrednianiem ([weighted averaging](#)) łączy predykcje pięciu niezależnych sieci AudioResNet trenowanych na komplementarnych reprezentacjach danych audio:

- [Reprezentacji harmoniczno-perkusyjnej](#) (waga: 0.3595)
- [Tempogram](#) (waga: 0.1812)
- [MFCC](#) (waga: 0.1552)
- [Chroma](#) (waga: 0.1546)
- [Spektrogram Melq](#) (waga: 0.1495)



Mechanizm działania

- **Ważone sumowanie predykcji** z optymalizacją wag przez Optunę
- **Kalibracja temperatury** ($T=1.0$) do kontroli "ostrości" rozkładów prawdopodobieństwa
- **Regularizacja L1** ($\lambda=0.01$) promującą rzadsze wagi i redukującą przeuczenie
- **Automatyczne ładowanie** najlepszych wersji modeli bazowych

MODEL ENSEMBLE

Model osiąga 94.65% dokładności na zbiorze testowym, przewyższając każdy z modeli składowych. Szczegółowe wskaźniki F1 dla poszczególnych emocji pokazują najwyższą skuteczność w rozpoznawaniu smutku (97,11% F1), złości (96,99% F1) i strachu (95,92% F1). Najniższe wartości F1 obserwujemy dla radości (92,00% F1) oraz zaskoczenia (88,39% F1).

Efekt synergii potwierdza wyraźną przewagę podejścia zespołowego nad indywidualnymi reprezentacjami audio, podkreślając wartość łączenia komplementarnych cech w rozpoznawaniu emocji.



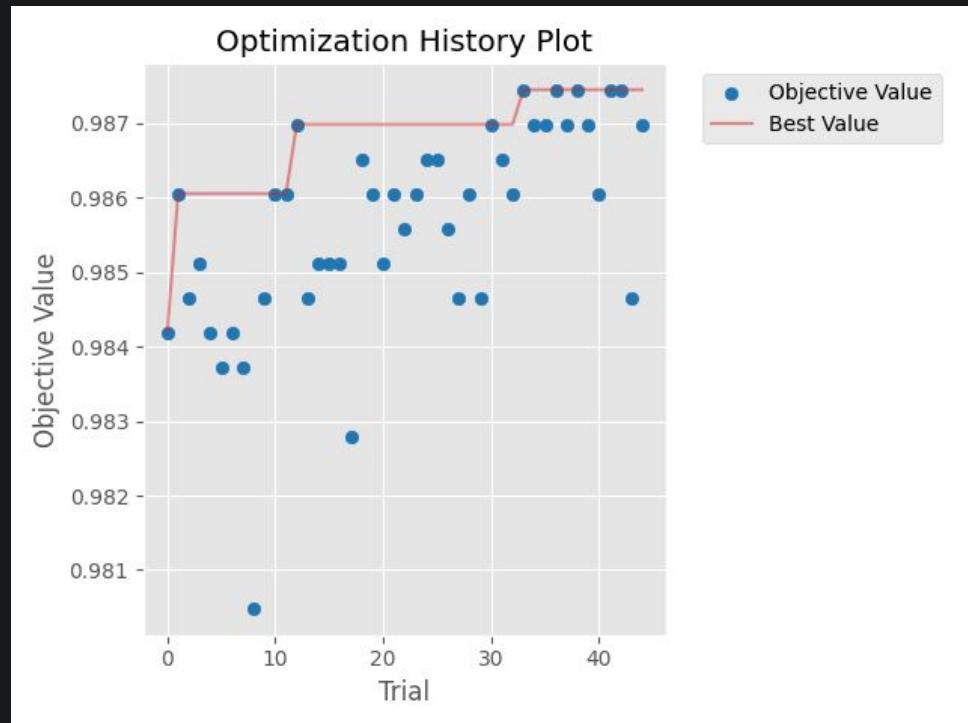
MODEL ENSEMBLE

Historia procesu optymalizacji wag modelu ensemble przy użyciu biblioteki Optuna dla pięciu reprezentacji audio: spektrogramu Mela, MFCC, HPSS, Chroma oraz Tempogramu.

Po znormalizowaniu udział poszczególnych reprezentacji w modelu ensemble wyniósł:

- HPSS - 35,95%,
- Tempogram - 18,11%,
- MFCC - 15,52%,
- Chroma - 15,46%,
- Melspektrogram - 14,85%.

Analiza ważności parametrów wykazała, że największy wpływ na wartość funkcji celu miał HPSS (0.57), a w następnej kolejności MFCC (0.18), Chroma (0.16), Melspektrogram (0.07) i Tempogram (0.02).



MODELOWANIE

MODEL	OPIS	Acc	F1	Prec	Rec
ResNET - ensembled	Model zespołowy wykorzystujący ważone uśrednianie predykcji z trzech niezależnych sieci ResNET, trenowanych na danych audio: melspekrogramach, MFCC, CQT oraz reprezentacji harmoniczno-perkusyjnej.	95%	95%	95%	95%
ResNET - 18	Konwolucyjna sieć neuronowa stworzona do przetwarzania obrazów, wytrenowana na spektrogramach Mel'a.	91%	91%	91%	91%
Simple CNN	Konwolucyjna sieć neuronowa 1D do przetwarzania surowych sygnałów audio. Wykorzystuje cztery bloki konwolucyjne z rosnącą liczbą filtrów, normalizacją wsadową i poolingiem.	83%	82%	82%	82%
VGG16	Głęboka sieć konwolucyjna składająca się z 16 warstw, wykorzystująca małe filtry 3x3 do skutecznej analizy obrazów.	74%	74%	75%	74%
CNN	Własna architektura konwolucyjnej sieci neuronowej trenowana na spektrogramach Mel'a.	47%	43%	47%	46%
Conv1D_RNN	Hybrydowa architektura sieci neuronowej, łącząca jednowymiarowe warstwy konwolucyjne (Conv 1D) oraz dwukierunkowe sieci rekurencyjne (BiLSTM, Bidirectional Long Short-Term Memory).	34%	32%	37%	34%

XAI

EXPLAINABLE AI

CZYM JEST

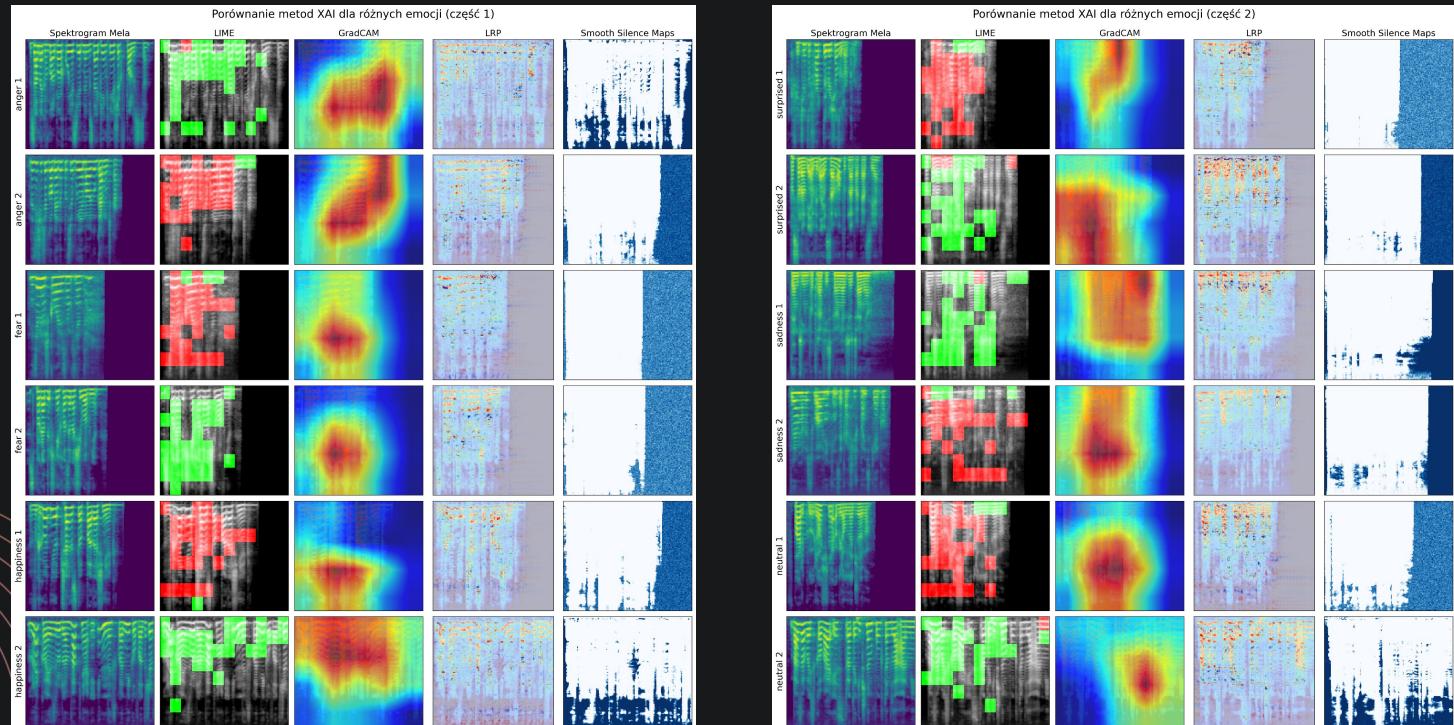
XAI (Explainable AI) to podejście, które pozwala zrozumieć, jak sztuczna inteligencja podejmuje decyzje. Dzięki XAI możemy sprawdzić, co miało wpływ na wynik modelu.

METODY

- **LIME:** Losowo zaburza fragmenty wejścia (np. kawałki spektrogramu) i sprawdza, jak zmienia się predykcja. Na tej podstawie szacuje, które fragmenty miały największy wpływ na wynik.
- **GradCAM:** Działa na konwolucyjnych sieciach neuronowych. Używa gradientów, aby wskazać, które obszary wejścia były najważniejsze dla danej klasy – tworzy „mapę ciepła” na spektrogramie.
- **LRP:** Przechodzi przez model „od końca do początku” i pokazuje, jak bardzo każdy element wejścia przyczynił się do ostatecznego wyniku.
- **Smooth Saliency Maps:** metoda, która pokazuje, na które fragmenty wejścia model silnie reaguje — co pozwala zauważać zarówno istotne cechy, jak i niepokojące wrażliwości na elementy niezwiązane z emociją.

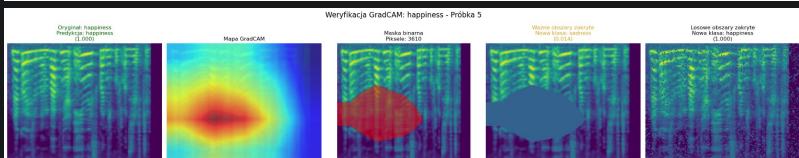
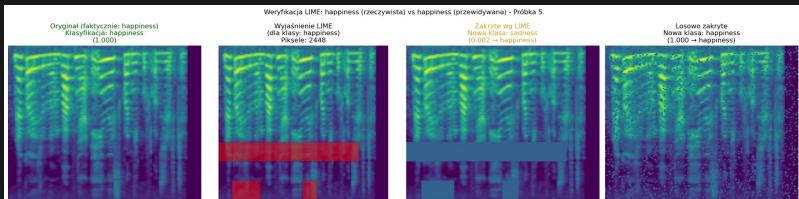
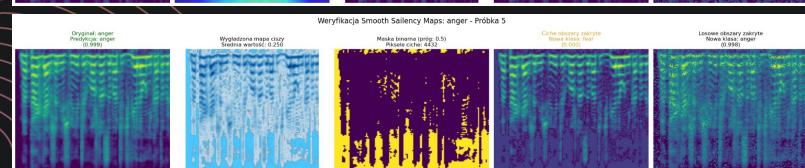
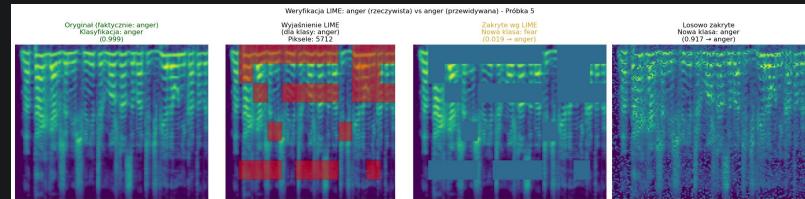
XAI

PORÓWNANIE



XAI

SPRAWDZENIE

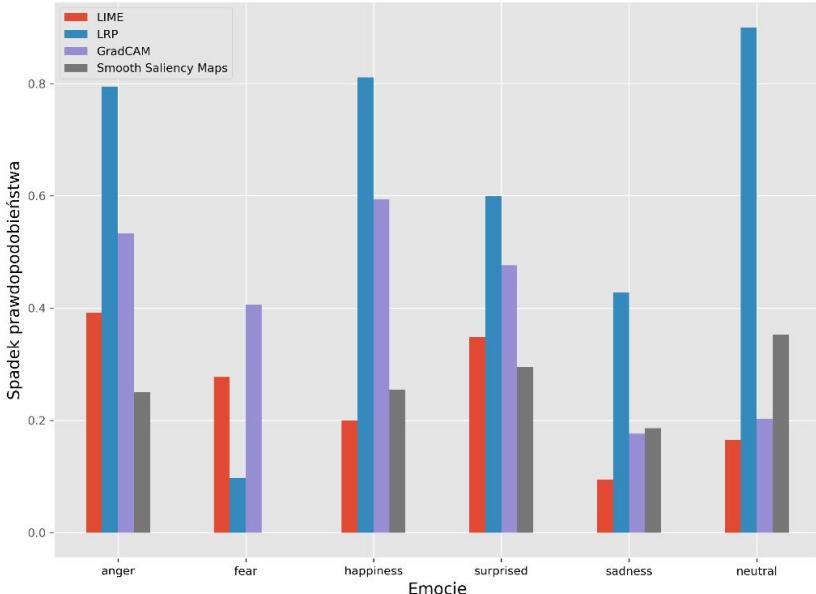


XAI

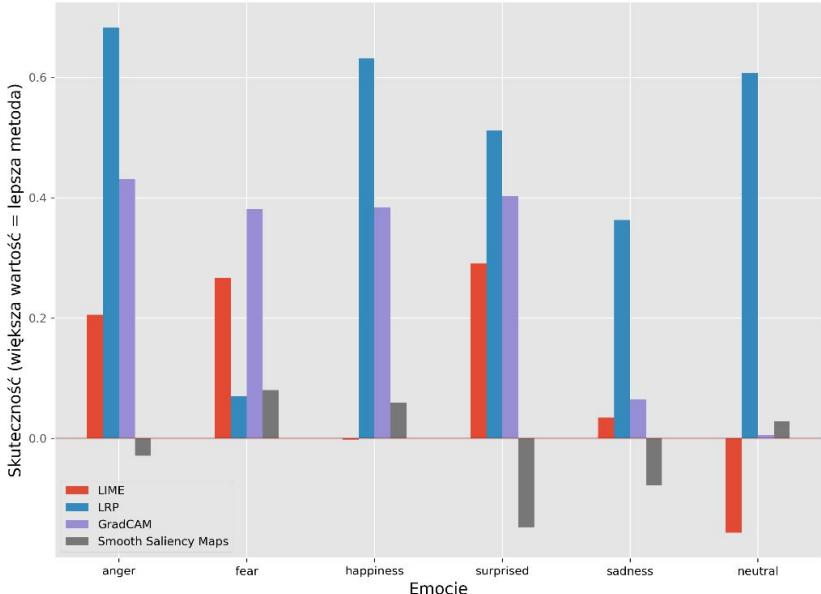
WYNIKI

Porównanie metod XAI dla rozpoznawania emocji

Porównanie spadku prawdopodobieństwa
dla obszarów istotnych

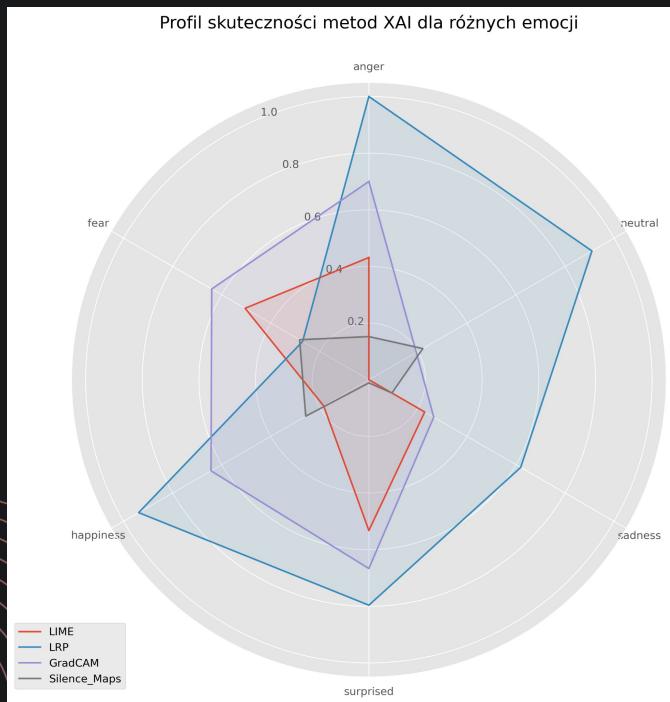


Skuteczność metod XAI
dla każdej emocji



XAI

WNIOSKI



Podsumowanie skuteczności metod

XAI:

LIME: 0.1064

LRP: 0.4779

GradCAM: 0.2784

Smooth Silence Maps: -0.0145

Najlepsza metoda dla tego modelu: LRP

ATAK ADWERSALNY NA SIEĆ ReSNET-18

CZYM JEST

Atak adwersalny (ang. adversarial attack) na sieć neuronową to technika, która polega na celowym modyfikowaniu danych wejściowych w taki sposób, aby wprowadzić w błąd model uczenia maszynowego, przy czym zmiany te są często niemal niewidoczne dla człowieka.

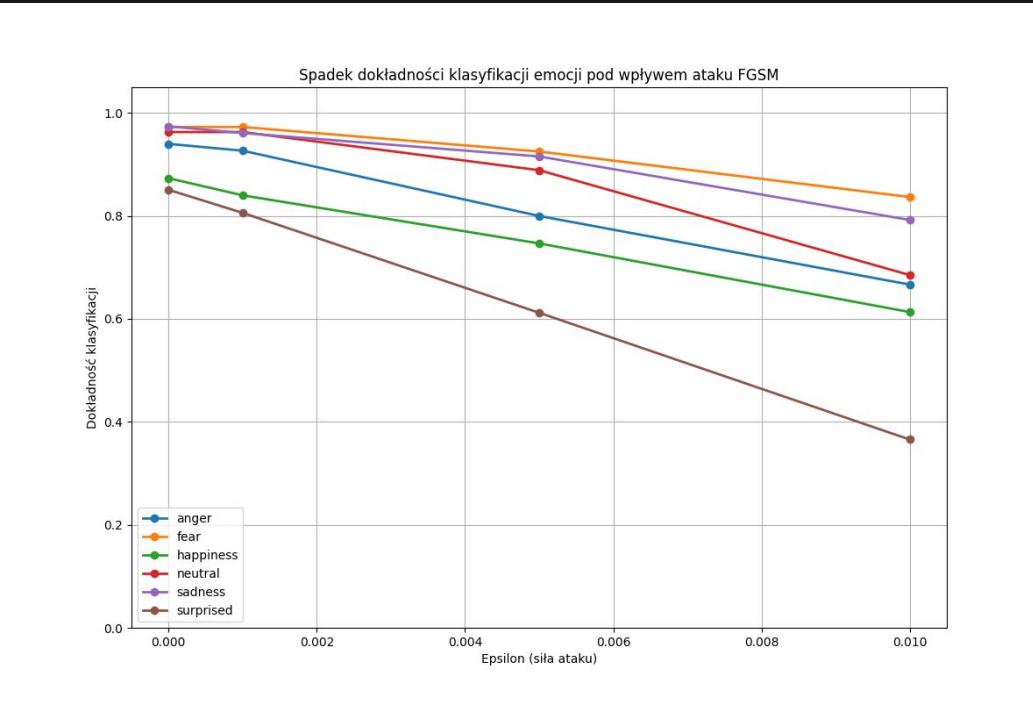
METODA

Jedną z technik wykorzystywanych do przeprowadzenia ataku adwersalnego jest Fast Gradient Sign Method (FGSM). Polega na dodaniu do oryginalnego sygnału niewielkiego zakłócenia (ϵ - epsilon) obliczonego na podstawie gradientu funkcji straty względem danych wejściowych.

WNIOSKI

- Wartość $\epsilon=0.02$ jest punktem krytycznym, przy którym dokładność modelu spada z >90% poniżej 50% (dokładnie do 31%).
- Klasy zaskoczenie, złość i szczęście są znacznie bardziej podatne na ataki niż strach, smutek i neutralność.
- Niewielkie perturbacje rzędu $\epsilon=0.01$ powodują spadek dokładności o 30%, co wskazuje na istotną podatność modelu ResNet-18 na ataki adwersalne.

ATAK ADWERSALNY NA SIEĆ ReSNET-18



Spadek dokładności klasyfikacji emocji w wyniku ataku FGSM.

APLIKACJA

GŁÓWNE FUNKCJE

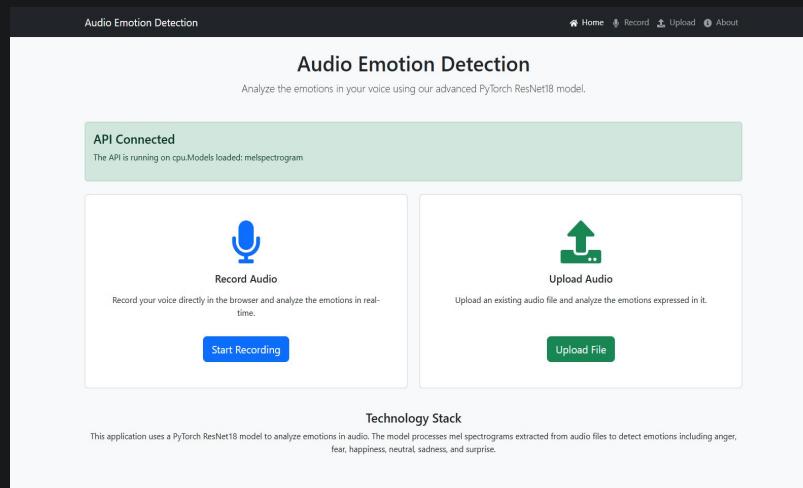
Nagrywanie dźwięku
w czasie rzeczywistym przez
przeglądarkę
(z limitem do 10 sekund)

Wczytywanie plików audio
z dysku lokalnego
użytkownika

Analiza emocji przy użyciu
modelu ResNet18
trenowanego na MFCC

Obsługa popularnych
formatów audio (mp3, wav,
ogg, flac, m4a, webm)

Responsywny interfejs



ARCHITEKTURA

Backend:

FastAPI, PyTorch, Librosa

Frontend:

React, Bootstrap

Wdrożenie:

Docker, konteneryzacja
całego stosu
technologicznego

APLIKACJA

GŁÓWNE FUNKCJE

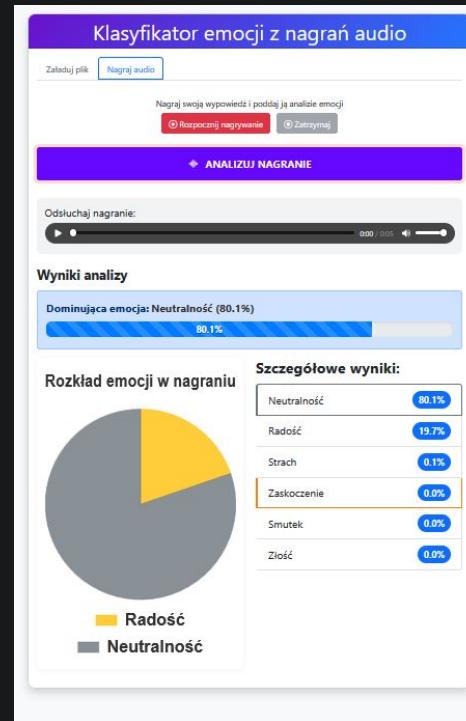
Nagrywanie dźwięku
w czasie rzeczywistym
przez przeglądarkę
(z limitem do 5 sekund)

Wczytywanie plików audio
z dysku lokalnego użytkownika

Analiza emocji przy użyciu
modelu ResNet18 trenowanego
na melspekrogramach

Obsługa popularnych formatów
audio (mp3, wav)

Możliwość odsłuchania
załadowanego pliku



ARCHITEKTURA

Backend:

FastAPI, PyTorch, Librosa

Frontend:

HTML + CSS + vanilla JS

Wdrożenie:

Uvicorn, czyli lokalny
serwer ASGI (Asynchronous
Server Gateway Interface)
dla aplikacji FastAPI

WNIOSKI

SKUTECZNOŚĆ MODELU RESNET ENSEMBLE

Zrealizowano główny cel projektu, osiągając **95%** dokładności w klasyfikacji emocji za pomocą modelu ResNet Ensemble (TBC).

TENDENCJE MODELU

Model wykazuje wyraźną tendencję do skuteczniejszego rozpoznawania **negatywnych emocji**, co może mieć praktyczne zastosowania w systemach wykrywania stresu czy kryzysu emocjonalnego.

NAJLEPSZE REPREZENTACJE

Dla modelu ResNet18 przetestowano różne reprezentacje sygnału audio, identyfikując te najbardziej efektywne w pojedynczym modelu: HPSS 93.2%, MelSpect 89.9%, Spektrogram 86.5%, CQT 86.4%.

WYJAŚNIENIA PREDYKCJI Z XAI

Zastosowanie metod XAI pozwoliło określić, które fragmenty Mel-spektrogramów są kluczowe dla rozpoznawania każdej z emocji, pogłębiając interpretowalność działania modelu.

PODATNOŚĆ NA ATAKI

Dla modelu ResNet + Mel-spektrogram przeprowadzono atak adwersarialny, który przy perturbacjach rzędu **0.01** powodował spadek dokładności o **30%** — najbardziej podatne były emocje: zaskoczenie, złość, szczęście.

IMPLEMENTACJA APLIKACJI

Stworzono dwie aplikacje wykorzystujące model ResNet dla Mel-spektrogramów, które umożliwiają rozpoznawanie emocji z mowy — rozwiązania te mogą stanowić bazę pod dalszy rozwój i personalizację systemu.

DALSZY ROZWÓJ



OPTYMALIZACJA MODELU

Planowane jest przeprowadzenie eksperymentów z różnymi kombinacjami reprezentacji dźwięku w celu identyfikacji zestawów cech umożliwiających jak najlepsze uchwycenie charakterystyk emocjonalnych w sygnale audio. Celem jest stworzenie zoptymalizowanego modelu ensemble, który będzie w stanie korzystać z uzupełniających się reprezentacji akustycznych.



UPGRADE APLIKACJI

Zintegrowanie najlepszych, najnowszych wersji modeli ensemble z aplikacją w celu poprawy trafności detekcji emocji oraz ogólnej wydajności systemu.



ROZSzerZENIE BAZY

Wprowadzenie funkcji pozwalającej użytkownikom na dobrowolne udostępnianie nagrani głosowych oraz ocenę poprawności detekcji emocji (np. poprzez wybór jednej z kilku proponowanych emocji, jeśli model się pomyli). Dzięki temu możliwe będzie ciągłe wzbogaczanie zbioru danych o różnorodne przypadki, co pozwoli na dalsze uczenie modeli i zwiększenie ich odporności na nietypowe lub trudne przykłady.



XAI + ATAK ADWERSALNY

Wykorzystanie różnych metod XAI do precyzyjnego wskazywania wrażliwych obszarów danych, które mogą być celem ataków adwersarialnych — co pozwoli na skuteczniejsze testowanie i wzmacnianie odporności modelu.

**DZIĘKUJEM
Y
ZA UWAGĘ!**