**AKADEMIA GÓRNICZO-HUTNICZA**

**im. Stanisława Staszica w Krakowie**

**Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej**

**Praca dyplomowa  
inżynierska**

**Michał Rola**

*Imię i nazwisko*

**Automatyka i Robotyka**

*Kierunek studiów*

**Automatyczne identyfikowanie gatunków ptaków  
na podstawie nagrań ich wokalizacji**

*Temat pracy dyplomowej*

**Dr. Inż. Andrzej Izworski**

*Promotor pracy*

……………………

*Ocena*

**Kraków, rok 2023/2024**

Spis treści

[1 Wstęp 3](#_Toc151656238)

[2 Przedstawienie problemu badawczego 3](#_Toc151656239)

[3 Baza danych próbek audio 4](#_Toc151656240)

[3.1 Akwizycja próbek audio 4](#_Toc151656241)

[3.2 Przetwarzanie cyfrowe próbek audio 5](#_Toc151656242)

[3.3 Przygotowanie zbiorów danych 5](#_Toc151656243)

[4 Wybór i konstrukcja sieci neuronowej 6](#_Toc151656244)

[4.1 Przegląd dostępnych rozwiązań 6](#_Toc151656245)

[4.2 Wykorzystane technologie oraz narzędzia 6](#_Toc151656246)

[4.3 Wybór modelu oraz uzasadnienie 6](#_Toc151656247)

[4.4 Projektowanie oraz uczenie modelu 6](#_Toc151656248)

[5 Prezentacja i analiza wyników 6](#_Toc151656249)

[6 Podsumowanie 6](#_Toc151656250)

[Bibliografia 6](#_Toc151656251)

[Dodatek 1 – Dziedzinowy słownik pojęć 6](#_Toc151656252)

[Dodatek 2 – Przykładowe próbki wokalizacji 6](#_Toc151656253)

# Wstęp

(1.5 strony, czego dotyczy praca)

# Przedstawienie problemu badawczego

(3-4 strony)

# Zbiór danych próbek audio

Stworzenie zbioru treningowego zawierającego dużo wysokiej jakości próbek jest najważniejszą i najbardziej czasochłonną częścią tworzenia dobrego modelu.

## Akwizycja próbek audio

Próbki audio zdecydowano się pobrać ze strony https://xeno-canto.org. Jest to strona, której użytkownicy z całego świata wspólnie zbierają, identyfikują oraz dzielą się doświadczeniem dotyczącym nagrań zwierząt. Poza nagraniami wokalizacji ptaków, które stanowią lwią część zbiorów strony, można znaleźć również nagrania odgłosów koników polnych oraz nietoperzy.

Postanowiono stworzyć zbiór danych składającą się z wokalizacji 10 gatunków ptaków, po 50 nagrań na gatunek dla zbioru treningowego oraz po 10 nagrań na gatunek dla zbioru testowego. Postarano się przy tym zadbać o to, aby zawartość nagrań sprawdziła jak najszerszy zakres umiejętności modelu. Od rozpoznawania gatunków o podobnych wokalizacjach po rozpoznawaniu gatunku po różnych jego odgłosach.

Przy doborze nagrań skupiono się na tym, aby posiadały jak najmniej zakłóceń i żeby zawierały tylko odgłosy jednego gatunku. Ponadto nagranie musiało być nagrane z częstotliwością próbkowania równą co najmniej 32.0kHz oraz przepływnością stałą na poziomie minimum 320kb/s (przykładowy fragment przebiegu sygnału można znaleźć na rysunku 3.1).

A screenshot of a sound wave

Description automatically generated

Rysunek . Przykładowy fragment przebiegu sygnału przed obróbką [źródło: opracowanie własne]

Po zebraniu wystarczającej ilości nieprzetworzonych nagrań przystąpiono do izolowania wokalizacji ptaków. Postanowiono uczyć model na nagraniach posiadających po 1 wokalizacji o różnej długości (od 0.2 sek. do 1.0 sek), jednak starano się, aby większość próbek była tej samej długości (0.3 sek). Zmieniono również ilość kanałów ze stereo na mono w celu zmniejszenia objętości plików, przyśpieszenia późniejszej obróbki oraz aby model nie musiał analizować 2 spektrogramów zawierających bardzo zbliżone do siebie sygnały (przykładowy wyizolowany sygnał zamieszczono na rysunku 3.2).

A screen shot of a sound wave

Description automatically generated

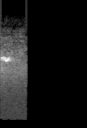
Rysunek . Sygnał z rysunku 3.1 po wyizolowaniu wokalizacji ptaka oraz zmniejszeniu ilości kanałów do 1 [źródło: opracowanie własne]

Po zebraniu odpowiedniej ilości wstępnie przygotowanych próbek można było przejść do ich dalszego przetwarzania.

## Przetwarzanie cyfrowe próbek audio

Z pomocą biblioteki Librosa przetworzono próbki audio z formatu WAV  
(ang. *Waveform Audio Format*) na obrazy Mel Spektrogramów, czyli Spektrogramów posiadających skale Mela na osi OY zamiast liniowo rosnącej częstotliwości (skala Mela jest skalą logarytmiczną i dzięki temu lepiej reprezentuje w jaki sposób ludzkie ucho postrzega dźwięk – dźwięki o niskiej częstotliwości są łatwiejsze do rozróżnienia od dźwięków o wysokich częstotliwościach).

Tak przetworzone obrazy zostały zapisane w formacie PNG (ang. *Portable Network Graphics*) Zdecydowano się na ten format, ponieważ jest to format o bezstratnej kompresji, więc jest możliwe odtworzenie informacji pierwotnych z postaci skompresowanej. Obrazy zapisano w odcieniach szarości, które badania pokazały są łatwiejsze do interpretacji przez modele sztucznej inteligencji. Pozwoliło to również ponownie zaoszczędzić miejsce, ponieważ zamiast 3 kanałów na kolory jest potrzebny tylko 1. Przykład obrazu, na którym model będzie się uczył przedstawiono na rysunku 3.3.



Rysunek . Mel Spektrogram sygnału z rysunku 3.2 [źródło: opracowanie własne]

Później odkryto, że należy sygnały, o zbyt krótkim czasie trwania, uzupełnić o ciszę do odpowiedniej długości, aby być w stanie stworzyć szereg (ang. *Array*), w którym zapisywano cały zbiór danych (ang. *Data Set*).

## Przygotowanie zbiorów danych

Kod odpowiedzialny za tworzenie zbioru danych przedstawiono poniżej.



Tak przygotowane dane były gotowe do wprowadzenia do modelu.

# Wybór i konstrukcja sieci neuronowej

## Przegląd dostępnych rozwiązań

## Wykorzystane technologie oraz narzędzia

## Wybór modelu oraz uzasadnienie

## Projektowanie oraz uczenie modelu

# Prezentacja i analiza wyników

# Podsumowanie

(1.5 co zostało zrobione, co nie wyszło i dlaczego, co można poprawić)

# Bibliografia

Tyle pozycji, aby przejść na kolejną stronę.

# Dodatek 1 – Dziedzinowy słownik pojęć

# Dodatek 2 – Przykładowe próbki wokalizacji

(link)