Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej

Projekt dyplomowy

Automatyczne identyfikowanie gatunków ptaków na podstawie nagrań ich wokalizacji

Automatic identification of bird species based on recordings of their vocalizations

Autor: Michał Rola

Kierunek studiów: Automatyka i Robotyka

Opiekun pracy: Dr Inż. Andrzej Izworski

Kraków, rok 2023/2024

Spis treści

1 Wstęp

(1.5 strony, czego dotyczy praca)

2 Przedstawienie problemu badawczego

(3-4 strony)

3 Przegląd wybranych dostępnych rozwiązań

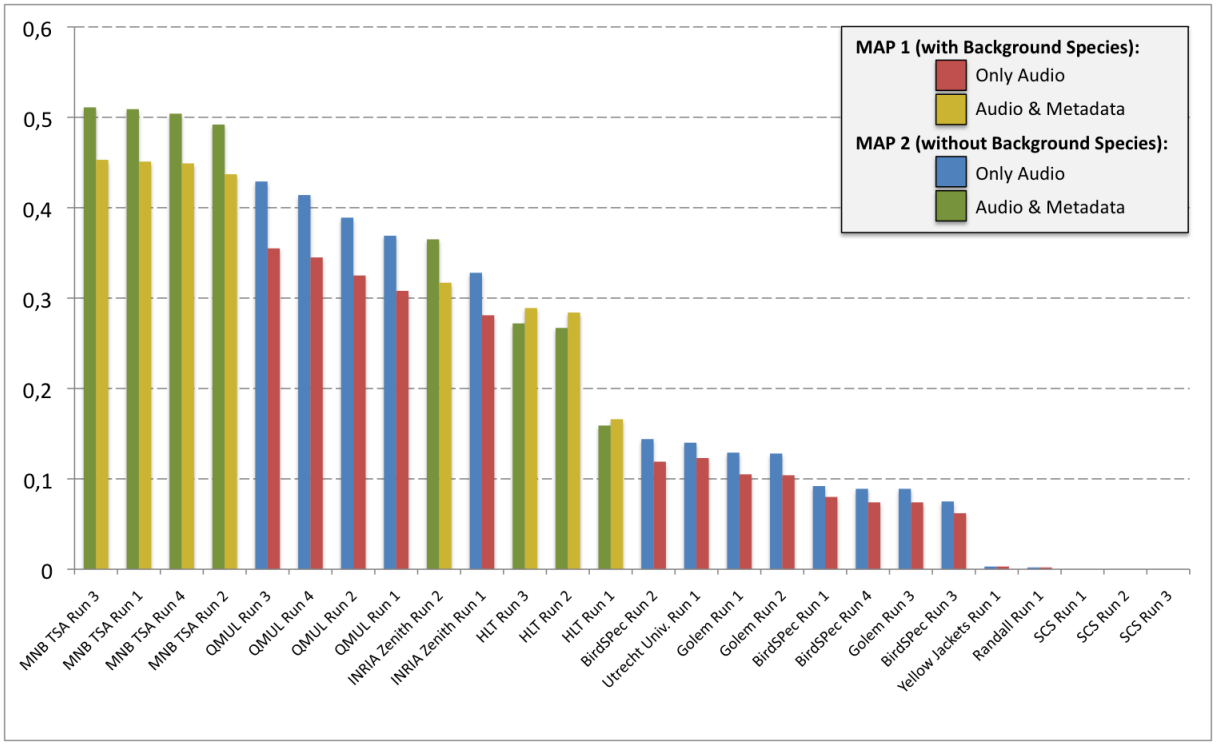
Na rynku dostępne jest wiele aplikacji specjalizujących się w rozpoznawaniu gatunków ptaków na podstawie nagrań audio. Rozdział ten skupia się na najpopularniejszych z nich i posiadających obszerną dokumentację.

3.1 Warblr

Aplikacja Warblr została wymyślona przez Florence Wilkinson oraz Dana Stowell, który posiada tytuł Profesora Nadzwyczajnego na Uniwersytecie Królowej Marii w Londynie (tłumaczenie własne, ang. Queen Mary University in London (QMUL)). Model wykorzystany w aplikacji jest w stanie rozpoznać ponad 80 gatunków ptaków występujących w Wielkiej Brytanii oraz posiada dokładność na poziomie 95% przy optymalnych warunkach [1].

Model ten wykorzystuje uczenie się cech charakterystycznych (tłumaczenie własne, ang. feature/representation learning), które polega na automatycznej identyfikacji przez model właściwości obiektu na podstawie nieprzetworzonych danych wejściowych. W celu zaimplementowania tej metody Mel spektrogram zostaje podzielony na odcinki liczące ułamek sekundy i na nich zostaje użyty sferyczny algorytm k-średnich. Algorytm ten różni się od standardowego algorytmu k-średnich w ten sposób, że zamiast znajdować centroidy minimalizując odległość Euklidesową, minimalizuje odległość kątową pomiędzy centroidami w formie wektorów i znanymi punktami [2].

W pierwszej edycji zawodów BirdCLEF (2014), których organizatorem jest instytucja CLEF (ang. Conference and Labs of the Evaluation Forum), model wykorzystany w aplikacji Warblr był najlepszym zaproponowanym rozwiązaniem, jeżeli chodzi o rozpoznawanie gatunków ptaków wyłącznie na podstawie danych audio. Baza danych wykorzystana w konkursie składała się z 501 gatunków ptaków występujących w Brazylii i zawierała 14 027 nagrań. Model wykorzystany w aplikacji Warblr był w stanie rozpoznać średnio 40% gatunków na nagraniach (33.33% jeżeli na nagraniu występowały odgłosy innych gatunków w tle) [3]. Wykres przedstawiający jak poradziły sobie wszystkie zespoły startujące w danej edycji BirdCLEF znajduje się na rysunku 3.1.



Rysunek 3.1 Wykres przedstawiający dokładności uzyskane przez poszczególne zespoły. Próby zespołu odpowiedzialnego za aplikację Warblr zaczynają się na QMUL (od uczelni, którą reprezentował zespół) [3]

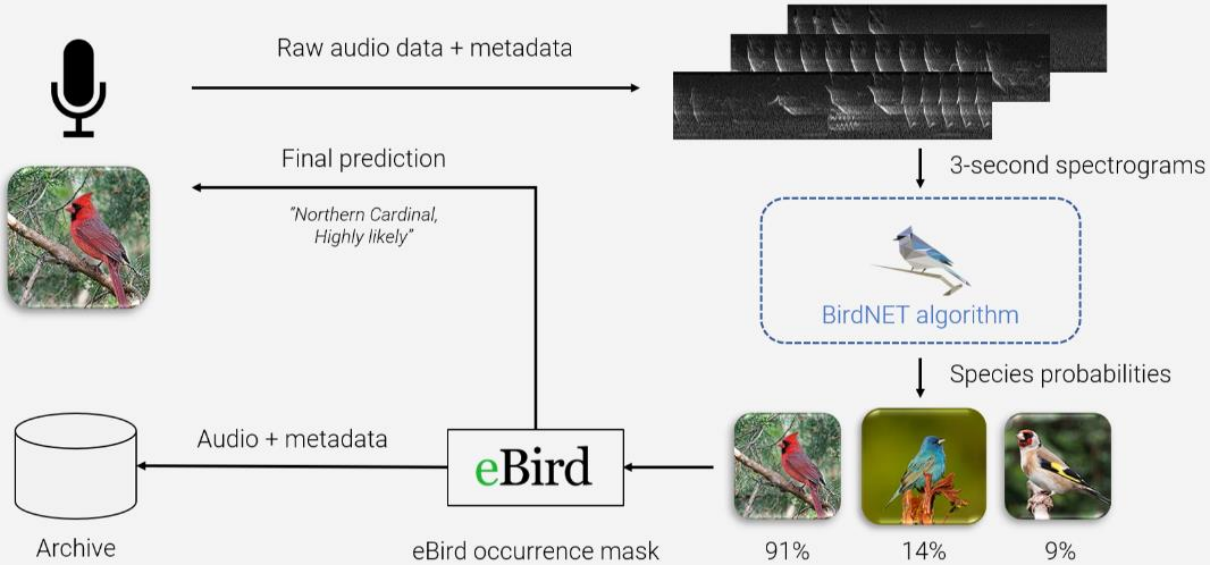
3.2 BirdNET Sound ID

Jest to aplikacja stworzona przez Centrum Bioakustyki Konserwacyjnej im. K. Lisa Yang należącego do Laboratorium Ornitologiczne Cornell (tłumaczenie własne, ang. The K. Lisa Yang Center for Conservation Bioacoustics at the Cornell Lab of Ornithology) we współpracy z Katedrą Informatyki Medialnej na Politechnice w Chemnitz (tłumaczenie własne, ang. the Chair of Media Informatics at Chemnitz University of Technology).

Ich aplikacja wspiera szeroką gamę urządzeń oraz systemów operacyjnych, między innymi: mikrokontrolery Arduino oraz Raspberry Pi, smartfony, przeglądarki internetowe, komputery oraz usługi w chmurze. W momencie pisania tej pracy BirdNET jest w stanie zidentyfikować około 3 000 najpopularniejszych gatunków ptaków występujących na świecie.

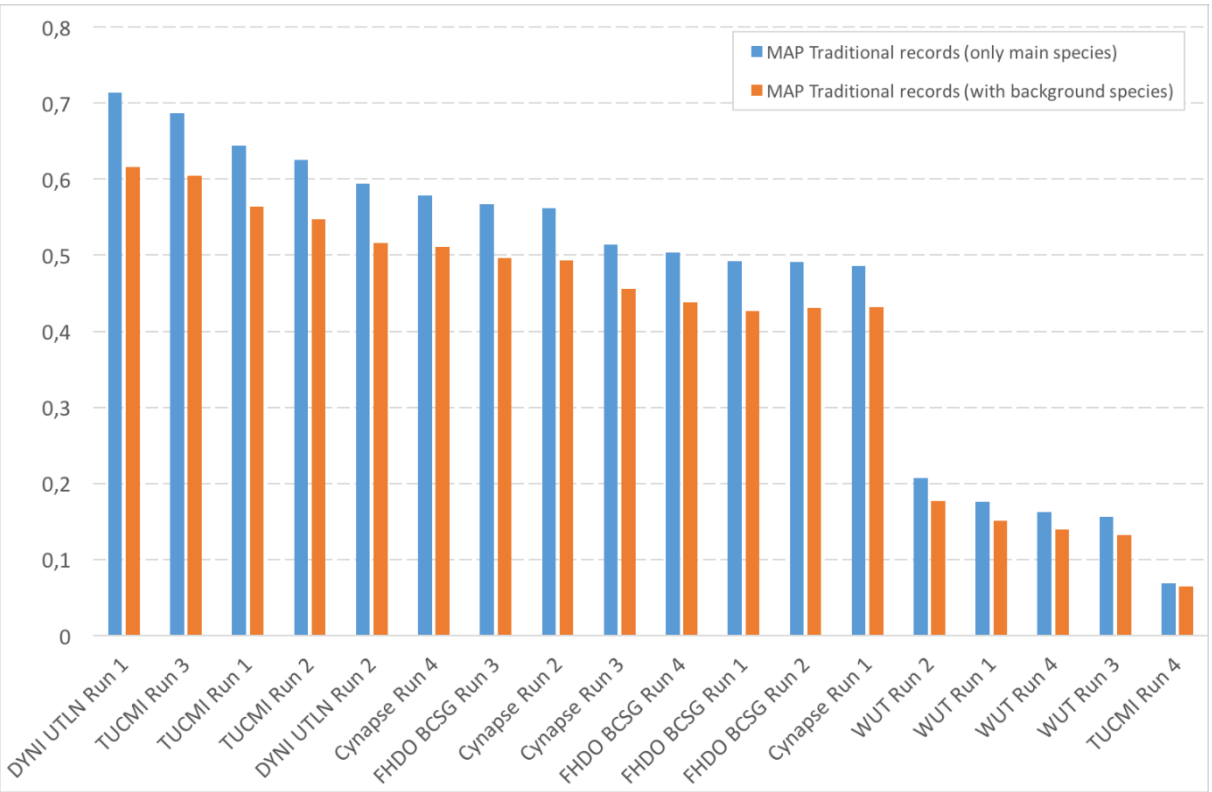
Na rysunku 3.2 przedstawiono schemat w jaki sposób z danych audio oraz metadanych (lokalizacji oraz okresu dokonania obserwacji) algorytm jest w stanie zidentyfikować gatunek słyszany na nagraniu. W pierwszej kolejności nagranie zostaje przetworzone na spektrogram i podzielone na 3 sekundowe fragmenty. Tak przetworzone dane zostają wprowadzone do modelu, który zwraca prawdopodobieństwa wystąpienia gatunku na nagraniu. Prawdopodobieństwa te zostają porównane z bazą występowania eBird (witrynę, która zrzesza zarówno amatorów jak i fanatyków ptasiarstwa oraz wykorzystuję ich wiedzę w celu tworzenia

ogólnodostępnej mapy obserwacji, jak i również bazę zdjęć oraz nagrań audio) [4]. Następnie klientowi zostaje zwrócony wynik wraz z prawdopodobieństwem, a do archiwum zostaje zapisany pomiar dostarczony przez klienta [5]. Pomiary zebrane w ten sposób mogłyby pomóc ekologom w ustalaniu wzorców migracyjnych oraz jak te wzorce będą się zmieniały wraz ze zmianami klimatycznymi [6]. Dokładność tego modelu w chwili pracy nad tym projektem jest na poziomie 80% [7].



Rysunek 3.2 Schemat przedstawiający działanie aplikacji BirdNET [5]

Osoby, które są odpowiedzialne za stworzenie algorytmu wykorzystanego w BirdNET również wzięły udział w BirdCLEF, ale w edycji z 2017 roku. Baza danych wykorzystana w konkursie składała się z 1 500 gatunków ptaków występujących w północnej części Ameryki Południowej i zawierała 39 496 nagrań. Na Rysunku 3.3 znajduje się wykres przedstawiający jak poradziły sobie wszystkie zespoły startujące w BirdCLEF w 2017 roku [8].

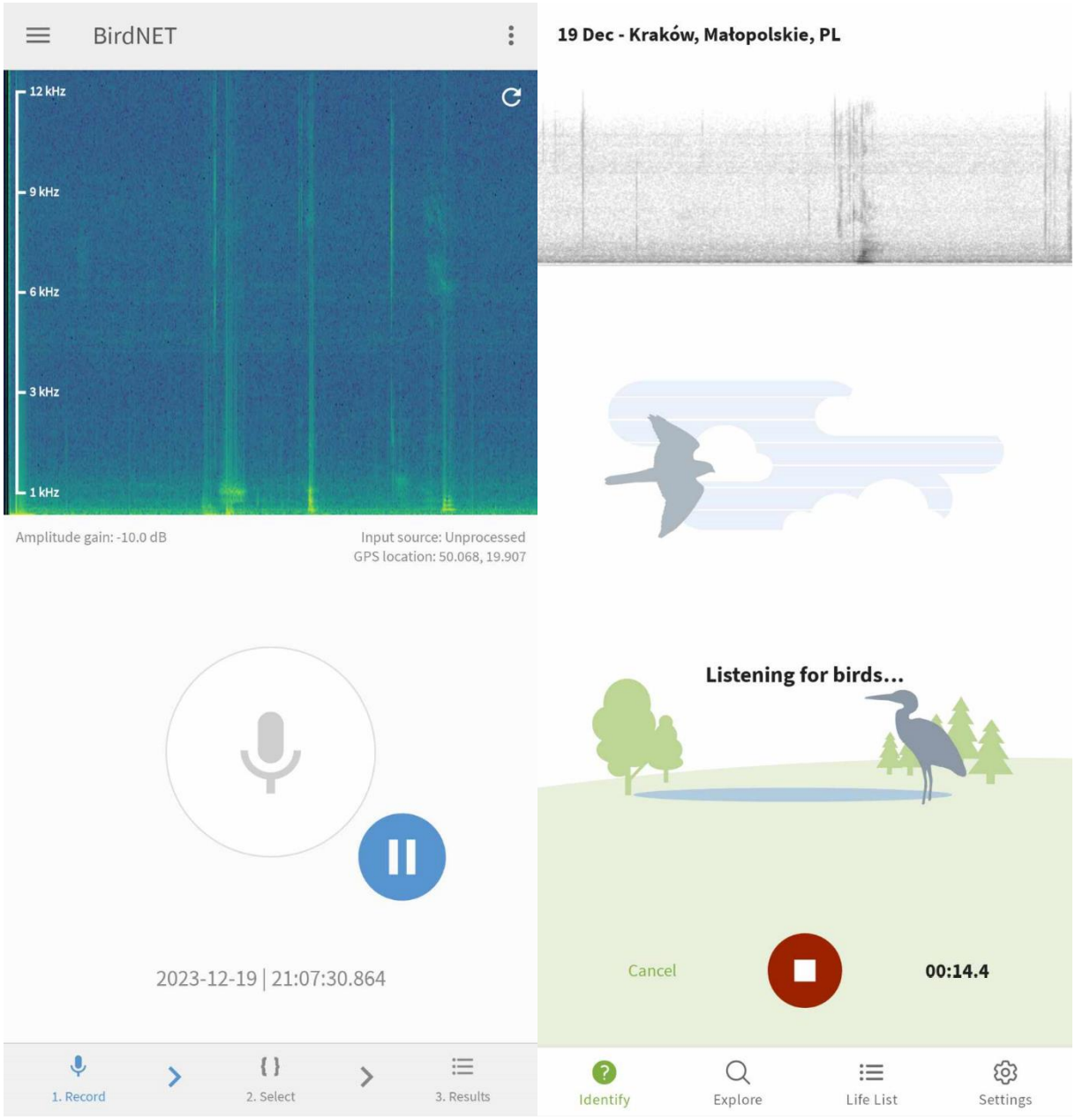


Rysunek 3.3 Wykres przedstawiający dokładności uzyskane przez poszczególne zespoły. Próby zespołu, który stworzył aplikacje BirdNET zaczynają się na TUCMI (niem. Technische Universität Chemnitz) [8]

3.3 Merlin Bird ID

Aplikacja ta została wydana w 2014 roku również przez zespół należący do Laboratorium Ornitologiczne Cornell. Mimo to aplikacja ta znacznie się różni od BirdNET. Przede wszystkim jest w stanie rozpoznawać nie tylko gatunki po wokalizacjach, ale również po zdjęciach oraz na podstawie odpowiedzi na serię pytań, jakie by zadali doświadczeni ornitolodzy w celu pomocy w identyfikacji zaobserwowanego gatunku [9]. W momencie pisania tej pracy aplikacja Merlin jest w stanie rozpoznać 1 054 gatunków na podstawie wokalizacji [10].

Wizualne porównanie interfejsów służących do identyfikacji gatunków po wokalizacji znajduje się na rysunku 3.4. Jak można zauważyć obydwie aplikacje wykorzystują lokalizację urządzenia oraz informacje o lokalnym czasie. Spektrogram aplikacji BirdNET ma zakres od 1kHz do 12kHz (jest to spowodowane tym, iż wokalizacja wielu gatunków znajduje się w tym zakresie [11]). Aplikacja Merlin nie posiada skali, ale doświadczalnie udało się ustalić, że dolny limit to około 100Hz, a górny około 11kHz.



Rysunek 3.4 Interfejsy identyfikacji ptaków po wokalizacji aplikacji BirdNET (po lewej) i Merlin (po prawej) [12], [13]

4 Wykorzystane technologie oraz narzędzia

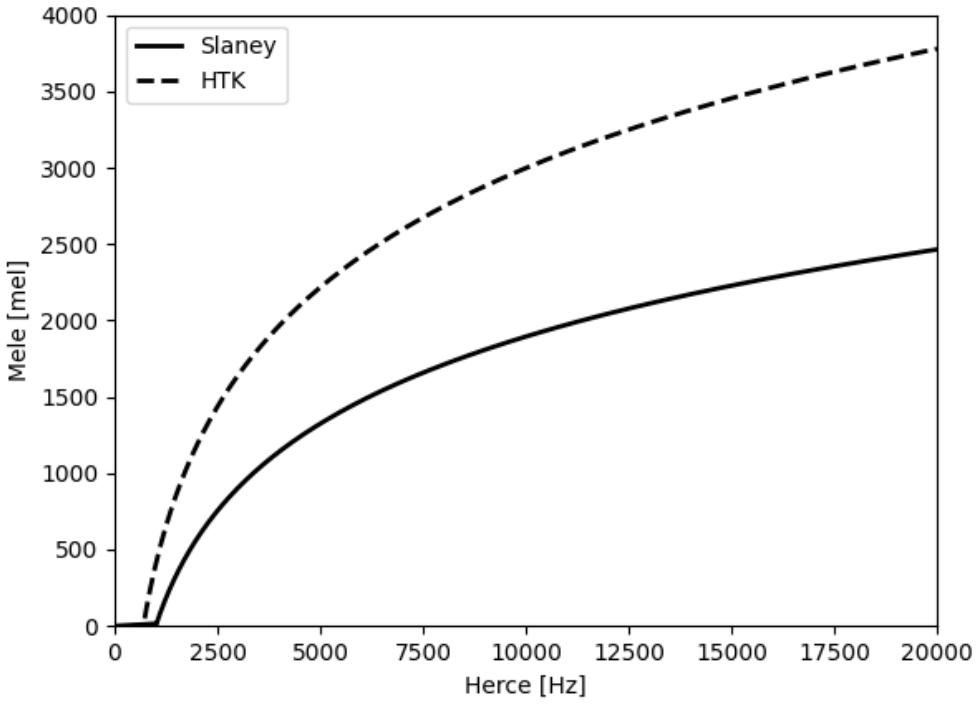
4.1 Mel Spektrogramy

Mel Spektrogramy to spektrogramy posiadające skale Melową na osi OY zamiast liniowo rosnącej częstotliwości. Skala Melowa powstała dzięki badaniom polegającym na ocenie przez słuchaczy do jakiej częstotliwości należy dany ton. Jest więc to miara subiektywna i posiada wiele równań zależnie od interpretacji tych danych, jednak znacznie lepiej przedstawia w jaki sposób ludzkie ucho postrzega dźwięki. Może to pozytywnie wpłynąć na działanie modelu, ponieważ wokalizacje wielu gatunków mieszczą się w częstotliwości do 10kHz i właśnie te częstotliwości będą najbardziej widoczne na Mel Spektrogramach. W pracy porównane zostały efekty uczenia modelu na Mel Spektrogramach uzyskanych na podstawie wzorów Slaney'a (4.1) oraz HTK (4.2) [14].

$$m(f) = \begin{cases} \frac{3f}{200}, & f < 1000 \\ 1527 \log_{6,4}\left(\frac{f}{1000}\right), & f \geq 1000 \end{cases} \tag{4.1}$$

$$m(f) = 2595 * \log_{10}(\frac{1+f}{700}) \tag{4.2}$$

Wykresy przedstawiające porównanie tych skali na osi częstotliwości wyskalowanej w Hercach przedstawia rysunek 4.1.



Rysunek 4.1 Porównanie skali Slaney oraz HTK na osi częstotliwości wyskalowanej w Hercach [źródło: opracowanie własne]

4.2 Python

Język programowania Python pozwala na szybką oraz względnie prostą implementację kodu. W trakcie tworzenia tego projektu dyplomowego została wykorzystana wersja 3.11 Pythona.

W tabeli 4.1 znajdują się informacje na temat wykorzystanych bibliotek.

Tabela 4.1 Użyte biblioteki wraz z opisami oraz informacjami o zastosowaniach ich w projekcie dyplomowym [źródło: opracowanie własne]

Nazwa biblioteki	**Wersja**	**Opis i zastosowanie w projekcie dyplomowym**
Keras	2.15	Służy do łatwego tworzenia modelów Uczenia Maszynowego i do tego został wykorzystany w tej pracy [15].
Librosa	0.10.1	Biblioteka ta służy do analizy plików audio. Zapewnia również elementy niezbędne do tworzenia systemów wyszukiwania informacji muzycznych [16]. W tym projekcie dyplomowym została wykorzystana do wczytywania plików audio oraz przetwarzaniu ich na Mel Spektrogramy.
Matplotlib	3.8.1	Biblioteka ta pozwala na szybkie i łatwe tworzenie statycznych, interaktywnych lub animowanych wizualizacji danych [17]. Pomogła ona przy tworzeniu wizualizacji wykorzystanych w tej pracy oraz przy zapisie oraz wczytywaniu Mel Spektrogramów.
NumPy	1.26.2	Biblioteka ta umożliwia przeprowadzanie wszelkiego rodzaju działań na macierzach w prosty sposób, zarówno numerycznych, jak i ich przetwarzania (dodawanie wierszy/rzędów/wymiarów, zamiana ich miejscami). Pozwala również na zapis oraz wczytywanie macierzy, co również zostało wykorzystane w tej pracy, w celu stworzenia zbioru danych [18].
scikit-learn	1.3.2	Biblioteka ta powstała z wykorzystaniem takich bibliotek, jak NumPy, SciPy oraz Matplotlib i pozwala na szybki dostęp do najpopularniejszych technik wykorzystywanych w Uczeniu Maszynowym [19], [20]. W pracy tej biblioteka ta zostanie wykorzystana przede wszystkim do wstępnego przetworzenia danych oraz oceny modelu.

5 Zbiór danych próbek audio

Stworzenie zbioru treningowego zawierającego dużo wysokiej jakości próbek jest jednym z ważniejszych i bardziej czasochłonną częścią tworzenia dobrego modelu.

Przy tworzeniu zbioru danych wykonano następujące kroki:

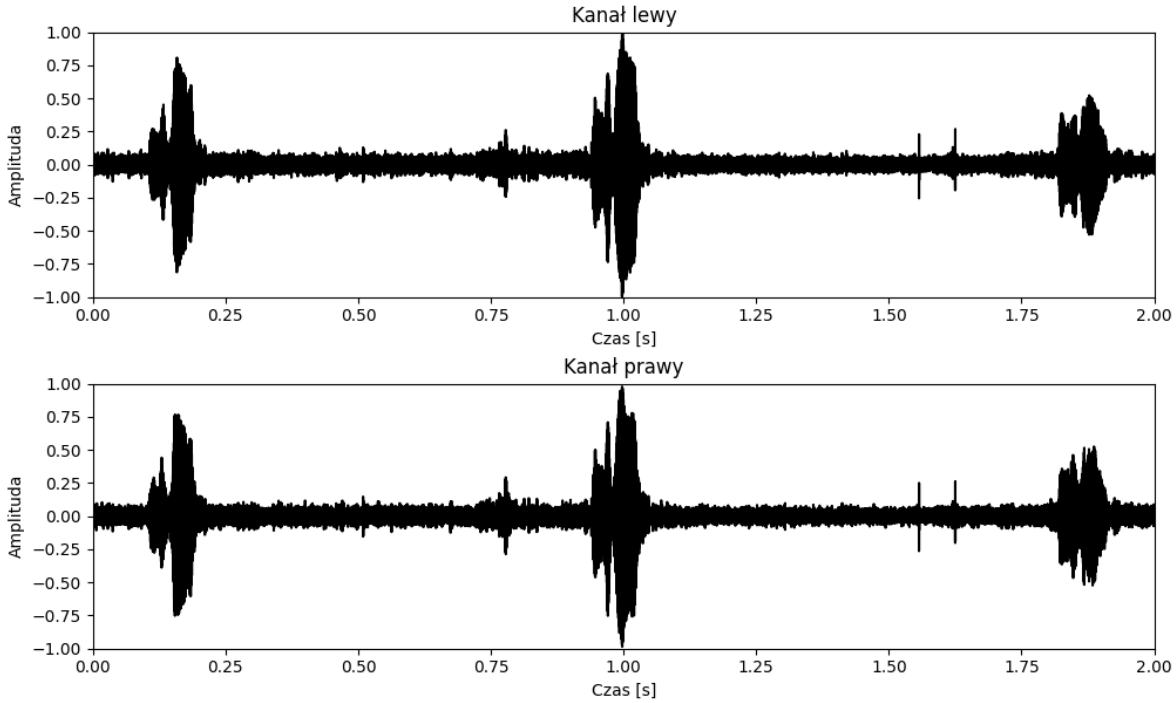
1. Pobranie odpowiedniej liczby nagrań audio do stworzenia zbioru danych,
2. Wydzielenie próbek audio,
3. Stworzenie Mel Spektrogramów z próbek audio,
4. Zapis zbioru danych składających się z Mel Spektrogramów na dysku w postaci macierzy oraz tablicy składającej się z nazw gatunków, przypisanych do obrazów.

5.1 Akwizycja próbek audio

Próbki audio pobrano 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.

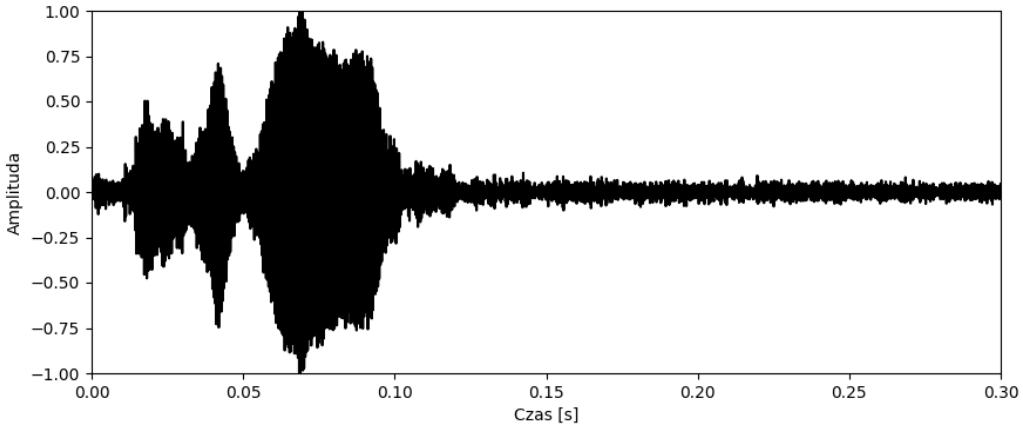
Postanowiono stworzyć zbiór danych składającą się z wokalizacji 10 gatunków ptaków, po 50 nagrań na gatunek dla zbioru treningowego oraz po 10 nagrań na gatunek dla zbioru testowego.

Przy doborze nagrań do zbioru treningowego 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 znajduje się na rysunku 5.1). Nagrania do zbioru testowego przeszły mniej rygorystyczną selekcję w celu sprawdzenia, jak model poradzi sobie z próbkami o jakości bliższej realnym warunkom.



Rysunek 5.1 Przykładowy fragment przebiegu sygnału przed obróbką
[źródło: opracowanie własne na podstawie [21]]

Po zebraniu wystarczającej ilości nieprzetworzonych nagrań wyizolowano wokalizacje poszczególnych gatunków. Uczono model na nagraniach posiadających po 1 wokalizacji o różnej długości (od 0,2-1,0 sekundy), z czego większość próbek trwa 0,3 sekundy. Zrezygnowano z jakości stereo. Dalsze prace prowadzono na sygnale mono w celu zmniejszenia objętości plików, przyśpieszenia późniejszej obróbki oraz aby model nie musiał analizować dwóch Mel Spektrogramów zawierających bardzo zbliżone do siebie sygnały (przykładowy kształt wyizolowanego sygnału zamieszczono na rysunku 5.2).

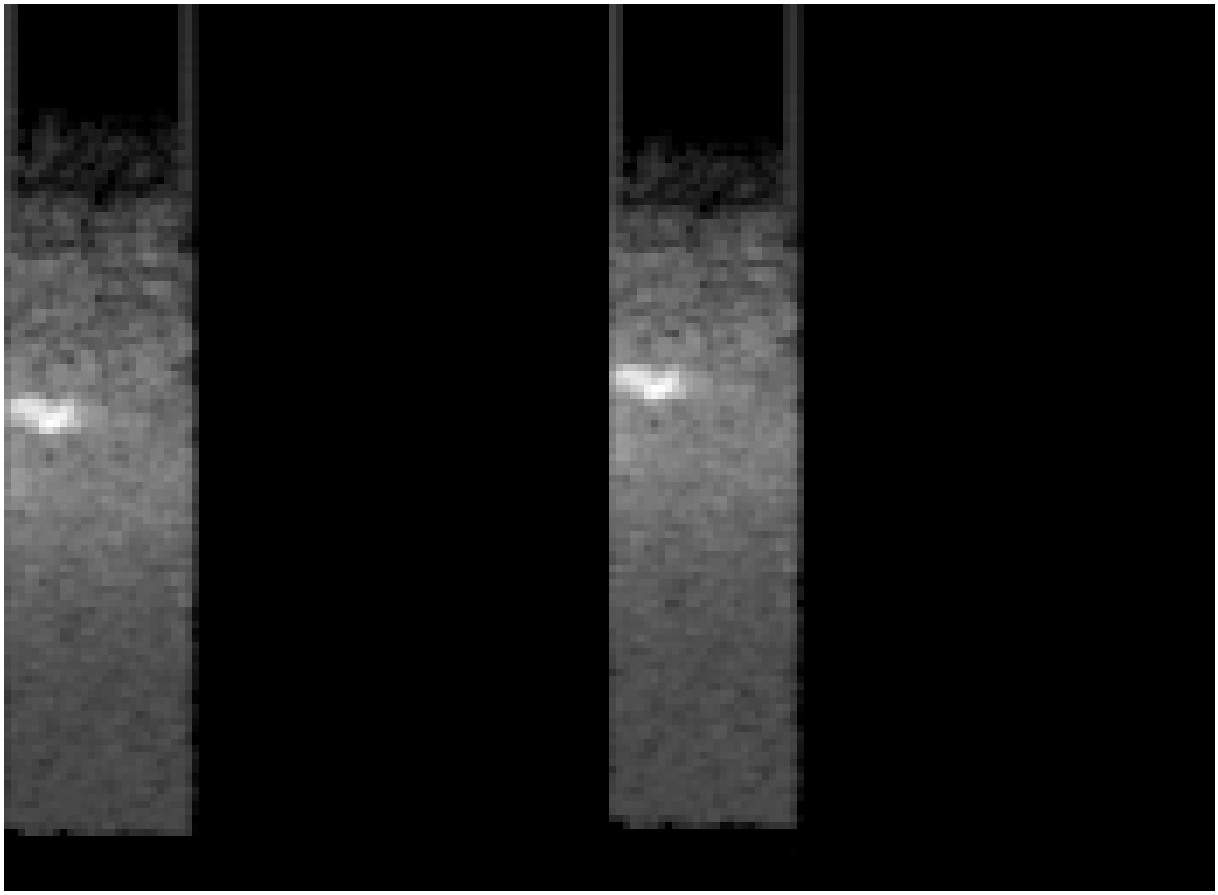


Rysunek 5.2 Sygnał z rysunku 5.1 po wyizolowaniu wokalizacji ptaka oraz zmniejszeniu ilości kanałów do jednego [źródło: opracowanie własne na podstawie [21]]

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

5.2 Przetwarzanie cyfrowe próbek audio

Z pomocą biblioteki Librosa przetworzono próbki audio z formatu WAV (ang. *Waveform Audio Format*) na obrazy Mel Spektrogramów. Zbyt krótkie nagrania uzupełniono wartościami zerowymi w celu ujednolicenia długości nagrań. Tak przetworzone obrazy zostały zapisane w formacie PNG (ang. *Portable Network Graphics*), który posiada bezstratną kompresję. Obrazy zapisano w odcieniach szarości, w celu ograniczenia zajmowanego przez nie miejsca, ponieważ zamiast trzech kanałów na kolory wymagany jest tylko jeden. Przykłady formatu obrazów zastosowanych do uczenia sieci neuronowej przedstawiono na rysunku 5.3.



Rysunek 5.3 Zarys Mel Spektrogramu sygnału z rysunku 5.2 z wykorzystaniem wzorów: Slaney'a (po lewej) oraz HTK (po prawej) [źródło: opracowanie własne]

5.3 Przygotowanie zbiorów danych

Dane podzielono na cechy (macierze reprezentujące wartości pikseli na Mel Spektrogramach w skali od 0-1) oraz tabelę z etykiety (nazwami gatunków) i zapisano w formacie NPY. Zdecydowano się na ten format, ponieważ pozwala na szybkie wczytywanie i zapis danych.

Ponad to, dane zapisane w ten sposób można w prosty sposób poddać procesowi inżynierii wstecznej. Co jest przydatne, jako że zbiory danych często są w stanie przetrwać przydatność programów, na potrzeby których zostały stworzone. Mogą one jednak przechowywać tylko szeregi o tych samych wymiarach, co utrudniło zapis cech oraz etykiet w jednym pliku i dlatego zdecydowano się na ich podział [22].

Tak przygotowane dane były gotowe do wprowadzenia do modelu.

6 Wybór i konstrukcja sieci neuronowej

6.1 Wybór modelu oraz uzasadnienie

6.2 Projektowanie oraz uczenie modelu

7 Prezentacja i analiza wyników

8 Podsumowanie

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

Bibliografia

[1] ‘Technology’, Warblr. Accessed: Dec. 18, 2023. [Online]. Available: https://www.warblr.co.uk/our-technology

[2] D. Stowell and M. D. Plumbley, ‘Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning’, *PeerJ*, vol. 2, p. 31, Jul. 2014, doi: 10.7717/peerj.488.

[3] ‘Bird task | ImageCLEF / LifeCLEF - Multimedia Retrieval in CLEF’. Accessed: Dec. 18, 2023. [Online]. Available: https://www.imageclef.org/2014/lifeclef/bird

[4] ‘About eBird - eBird’. Accessed: Dec. 19, 2023. [Online]. Available: https://ebird.org/ebird/about

[5] ‘BirdNET Sound ID – The easiest way to identify birds by sound.’ Accessed: Dec. 18, 2023. [Online]. Available: https://birdnet.cornell.edu/

[6] ‘What’s that bird song? ID birds by sound with BirdNET - YouTube’. Accessed: Dec. 18, 2023. [Online]. Available: https://youtu.be/MQHunTLt1TI?t=1302

[7] *What’s that bird song? ID birds by sound with BirdNET*, (2020). Accessed: Dec. 19, 2023. [Online Video]. Available: https://youtu.be/MQHunTLt1TI?t=2807

[8] ‘BirdCLEF 2017 | ImageCLEF / LifeCLEF - Multimedia Retrieval in CLEF’. Accessed: Dec. 18, 2023. [Online]. Available: https://www.imageclef.org/lifeclef/2017/bird

[9] ‘The Story’, Merlin Bird ID - Free, instant bird identification help and guide for thousands of birds. Accessed: Dec. 19, 2023. [Online]. Available: https://merlin.allaboutbirds.org/the-story/

[10] ‘Identify Bird Songs and Calls with Sound ID’, Merlin Bird ID - Free, instant bird identification help and guide for thousands of birds. Accessed: Dec. 19, 2023. [Online]. Available: https://merlin.allaboutbirds.org/sound-id/

[11] ‘Do bird songs have frequencies higher than humans can hear?’, All About Birds. Accessed: Dec. 19, 2023. [Online]. Available: https://www.allaboutbirds.org/news/do-bird-songs-have-frequencies-higher-than-humans-can-hear/

[12] ‘BirdNET – Aplikacje w Google Play’. Accessed: Dec. 19, 2023. [Online]. Available: https://play.google.com/store/apps/details?id=de.tu_chemnitz.mi.kahst.birdnet&hl=pl

[13] ‘Merlin Bird ID by Cornell Lab – Aplikacje w Google Play’. Accessed: Dec. 19, 2023. [Online]. Available: https://play.google.com/store/apps/details?id=com.labs.merlinbirdid.app&hl=pl

[14] ‘librosa.mel_frequencies — librosa 0.10.1 documentation’. Accessed: Dec. 20, 2023. [Online]. Available: https://librosa.org/doc/main/generated/librosa.mel_frequencies.html

[15] ‘Keras: Deep Learning for humans’. Accessed: Dec. 20, 2023. [Online]. Available: https://keras.io/

[16] ‘librosa — librosa 0.10.1 documentation’. Accessed: Dec. 20, 2023. [Online]. Available: https://librosa.org/doc/latest/index.html

[17] ‘Matplotlib — Visualization with Python’. Accessed: Dec. 20, 2023. [Online]. Available: https://matplotlib.org/

[18] ‘NumPy’. Accessed: Dec. 20, 2023. [Online]. Available: https://numpy.org/

[19] ‘scikit-learn: machine learning in Python — scikit-learn 1.3.2 documentation’. Accessed: Dec. 20, 2023. [Online]. Available: https://scikit-learn.org/stable/index.html

[20] F. Pedregosa *et al.*, ‘Scikit-learn: Machine Learning in Python’, *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

[21] ‘How to Visualize Sound in Python’, LearnPython.com. Accessed: Dec. 18, 2023. [Online]. Available: https://learnpython.com/blog/plot-waveform-in-python/

[22] ‘numpy.lib.format — NumPy v2.0.dev0 Manual’. Accessed: Dec. 20, 2023. [Online]. Available: https://numpy.org/devdocs/reference/generated/numpy.lib.format.html

Dodatek 1 – Dziedzinowy słownik pojęć

Dodatek 2 – Przykładowe próbki wokalizacji

(link)