Science, Technology and Innovation
ISSN 2544-9125 ▪ 2026; 24 (1): 32-45
doi: 10.55225/sti.707

STI • 2026; 24 (1): 32-45

Automatyczna klasyfikacja wybranych gatunków ptaków z użyciem sieci LSTM i mechanizmu uwagi na podstawie cyfrowych nagrań wokalizacji

Automatic classification of selected bird species using LSTM network and attention mechanism based on digital vocalization recordings

Marcin Skobel[1] A,C,D-G, Robert Wielgat[2] A,B,D,E,G

[1] Akademia Tarnowska, Wydział Nauk Technicznych, Katedra Informatyki, ul. Mickiewicza 8, 33-100 Tarnów, Polska
[2] Akademia Tarnowska, Wydział Nauk Technicznych, Katedra Elektroniki i Technologii Inteligentnych, ul. Mickiewicza 8, 33-100 Tarnów, Polska

Artykuł oryginalny

Abstrakt

Cel: Niniejsza praca koncentruje się na opracowaniu i ewaluacji modelu uczenia głębokiego przeznaczonego do klasyfikacji sygnałów dźwiękowych 31 gatunków ptaków. Głównym celem było zbadanie wpływu mechanizmu uwagi typu Luong na zdolność sieci do generalizacji wzorców w danych sekwencyjnych.

Materiał i metody: Zastosowano architekturę hybrydową łączącą warstwy długoterminowej pamięci krótkotrwałej (LSTM – Long Short-Term Memory) z globalnym mechanizmem uwagi. Proces przetwarzania danych obejmował ekstrakcję cech mel-cepstralnych (MFCC – Mel-Frequency Cepstral Coefficients) oraz dopasowanie danych do jednolitego interwału czasowego. Model został poddany trzykrotnej procedurze uczenia i testowania na losowo wybieranych podzbiorach danych, co pozwoliło na rzetelną ocenę stabilności wyników. Skuteczność klasyfikacji mierzono za pomocą metryk: dokładności, F1-score oraz średniej wartości AUC.

Wyniki i wnioski: Badanie wykazało, że zastosowanie mechanizmu uwagi zamiast kolejnej warstwy LSTM znacząco redukuje wymiarowość modelu przy jednoczesnej poprawie dokładności i jakości klasyfikacji. Uzyskano średnią dokładność na poziomie 0,9787, F1-score na poziomie 0,9493 oraz średnią wartość AUC wynoszącą 0,9991. Połączenie warstw LSTM z mechanizmem uwagi stanowi skuteczne narzędzie klasyfikacji sygnałów akustycznych wydawanych przez różne gatunki ptaków.

Słowa kluczowe

- rozpoznawanie głosów ptaków
- klasyfikacja dźwięku
- sieci neuronowe
- uczenie głębokie
- mechanizm uwagi
- sygnały akustyczne

Udziały autorów

A – przygotowanie badań
B – gromadzenie danych
C – analiza statystyczna uzyskanych wyników
D – interpretacja uzyskanych wyników
E – przygotowanie pierwotnej wersji tekstu
F – przegląd literatury
G – korekta i rewizja tekstu

Informacje o artykule

Historia artykułu (Article history)

- Otrzymano (Received): 2026-01-08
- Zaakceptowano (Accepted): 2026-05-22
- Opublikowano (Published): 2026-06-30

Wydawca (Publisher)

Akademia Tarnowska
University of Applied Sciences in Tarnow
ul. Mickiewicza 8, 33-100 Tarnow, Poland

Finansowanie (Financing)

Badania nie zostały sfinansowane z grantów pochodzących ze środków publicznych, organizacji komercyjnych lub non-profit.

Konflikt interesów (Conflict of interest)

Nie zadeklarowano konfliktu interesów.

Abstract

Aim: This paper focuses on the development and evaluation of a deep learning model designed for the classification of audio signals from 31 bird species. The main objective of the study was to investigate the impact of the Luong attention mechanism on the network's ability to generalize patterns in sequential data.

Material and methods: A hybrid neural network architecture combining Long Short-Term Memory (LSTM) layers with a global attention mechanism was employed. The data preprocessing pipeline included MFCC (Mel-Frequency Cepstral Coefficients) feature extraction and the alignment of signals to a uniform time interval. The model was subjected to a threefold training and testing procedure on randomly selected data subsets, enabling a reliable assessment of result stability. Classification performance was evaluated using the following metrics: accuracy, F1-score, and mean AUC.

Results and conclusion: The results demonstrated that replacing an additional LSTM layer with an attention mechanism significantly reduces model dimensionality while simultaneously improving classification accuracy and overall performance. The proposed approach achieved a mean accuracy of 0.9787, an F1-score of 0.9493, and a mean AUC value of 0.9991. The combination of LSTM layers with an attention mechanism constitutes an effective tool for the classification of acoustic signals produced by different bird species.

Korespondencja

Robert Wielgat
e-mail: r_wielgat@atar.edu.pl
Akademia Tarnowska
Wydział Nauk Technicznych
Katedra Elektroniki i Technologii Inteligentnych
ul. Mickiewicza 8
33-100 Tarnów, Polska

Wprowadzenie

Wokalizacja ptaków stanowi jeden z kluczowych atrybutów wyróżniających poszczególne gatunki. Unikalne wzorce dźwiękowe charakterystyczne dla każdego gatunku kształtowały się w trakcie milionów lat ewolucji i stanowią podstawę ich rozpoznawania. Różnorodność dźwięków wynika z fizycznej budowy krtani dolnej u ptaków (ang. *syrinx*) i wpływa na możliwości głosowe poszczególnych gatunków. W świecie ptaków głos odgrywa kluczową rolę w procesach godowych, natomiast z punktu widzenia ochrony przyrody stanowi istotny wzorzec identyfikacyjny.

Rejestracja dźwięków w postaci cyfrowej otwiera szerokie pole do badań związanych z rozpoznawaniem poszczególnych gatunków ptaków oraz identyfikacją i monitorowaniem gatunków zagrożonych. W niniejszej pracy podjęto zadanie wykorzystania sieci rekurencyjnych w procesie klasyfikacji ptaków na podstawie cyfrowych nagrań. Przeprowadzone eksperymenty mają na celu ocenę skuteczności sieci rekurencyjnych w zadaniu klasyfikacji gatunków ptaków. Ponadto badanie ma na celu weryfikację, czy połączenie warstw rekurencyjnych z mechanizmem uwagi umożliwia redukcję liczby parametrów modelu w porównaniu z architekturą w pełni rekurencyjną, przy jednoczesnym zachowaniu wysokiej dokładności klasyfikacji.

Badania przeprowadzono na podstawie materiału opisanego w pracy [1]. Materiał ten obejmuje łącznie nagrania wokalizacji 31 gatunków ptaków, przy czym w artykule [1] wykorzystano nagrania 30 gatunków, natomiast w niniejszych badaniach uwzględniono wszystkie 31 gatunków. W pracy [1] zastosowano klasyfikator oparty na ukrytych modelach Markowa (ang. Hidden Markov Models – HMM) z wykorzystaniem parametrów mel-cepstralnych (ang. mel-frequency cepstral coefficients, MFCC) uzyskując dokładność klasyfikatora na poziomie 93,33%. Parametry MFCC stanowią obecnie standard w badaniach związanych z klasyfikacją gatunków ptaków na podstawie wokalizacji [1–6]. Warto podkreślić, że współczynniki MFCC zostały pierwotnie opracowane na potrzeby systemów automatycznego rozpoznawania mowy (ASR), a ich konstrukcja odzwierciedla właściwości percepcyjne ludzkiego układu słuchowego, w szczególności nieliniową skalę częstotliwości odpowiadającą czułości ucha ludzkiego [7–9]. Z tego względu stosowanie cech MFCC do analizy wokalizacji ptaków — sygnałów generowanych i odbieranych przez organizmy o odmiennych mechanizmach percepcji akustycznej – nie musi być optymalne i wymaga każdorazowego potwierdzenia eksperymentalnego.

Pokrewną do MFCC grupą parametrów są współczynniki HFCC (ang. *human factor cepstral coefficients*), które w jeszcze większym stopniu odzwierciedlają fizjologię ludzkiego słuchu [10–15]. Parametry HFCC zostały skutecznie wykorzystane w zagadnieniach związanych z klasyfikacją wokalizacyjną ptaków [5, 16]. Interesującą alternatywę stanowią także cechy Chroma i Tonnetz, pierwotnie wykorzystywane w rozpoznawaniu dźwięków instrumentów muzycznych [17, 18]. Raportowane wyniki badań wskazują, że cechy te mogą charakteryzować się wysoką skutecznością, zwłaszcza w przypadku gatunków ptaków o złożonej strukturze

tonalnej śpiewu [19, 20]. W prezentowanych w niniejszym artykule eksperymentach zastosowano cechy MFCC, ponieważ celem badań było porównanie klasyfikatora opartego na sieci LSTM z klasyfikatorem HMM opisanym w pracy [1], w której również wykorzystano parametry MFCC.

W ostatnich latach popularność zdobywają metody klasyfikacji ptaków na podstawie wokalizacji z użyciem sieci rekurencyjnych. Standardem staje się stosowanie sieci z warstwami GRU [21] w tym model typu BiGRU z potwierdzoną dokładnością na poziomie 84% [6] na zbiorze Xeno-Canto [22]. Wysoką skuteczność w klasyfikacji głosów ptaków wykazują metody oparte o warstwy LSTM, które w połączeniu ze splotowymi sieciami neuronowymi uzyskują niemal idealną dokładność na małych zbiorach danych [23].

Kolejnym interesującym elementem współczesnych sieci neuronowych służących do rozpoznawania ptaków po głosie jest zastosowanie mechanizmu uwagi. Badania wykazują (na przykład model MFF-ScSEnet [24]), że mechanizmy uwagi pozwalają na lepszą identyfikację ptaków w trudnych warunkach akustycznych [24]. Dodatkowo badania wskazują, że stosowanie hierarchicznego mechanizmu uwagi pozwala na detekcję drobnych różnic w zakresie rozpoznawania spokrewnionych gatunków ptaków [25]. Mechanizm uwagi został również zaadoptowany w sieci typu Wav2vec [26], która po odpowiednim dostrojeniu i modyfikacji może stanowić skuteczne narzędzie do rozpoznawania ptaków na podstawie nagrań obfitujących w wielogatunkowe wokalizacje [27]. Niemniej jednak w pewnych uzasadnionych sytuacjach stosowanie tak rozbudowanych modeli może prowadzić do przeuczenia, dlatego odpowiedni dobór materiału badawczego oraz skutecznie podawanie istotnych fragmentów nagrań w procesie uczenia sieci neuronowej może przynieść korzyść w scenariuszach o małej próbie danych [28]. Co więcej, w specyficznych warunkach, dla niewielkich zbiorów danych, odpowiednio zoptymalizowane, klasyczne metody uczenia maszynowego (np. SVM – *Support Vector Machine*) wspomagane przez precyzyjną inżynierię cech mogą prowadzić do równie dobrych lub nawet lepszych rezultatów niż sieci neuronowe [29]. Ważnym więc aspektem projektowania sieci neuronowej w zaproponowanym w tym artykule rozwiązaniu było dopasowanie optymalnej architektury modelu do posiadanych danych oraz zastosowanie odpowiedniego mechanizmu uwagi.

W niniejszej pracy zaprezentowano wyniki klasyfikacji modeli opartych na hybrydowych architekturach LSTM, łączących warstwy LSTM z mechanizmem uwagi (LSTM + MU). Dodatkową zaletą zaproponowanego podejścia jest redukcja liczby parametrów modelu, co umożliwia jego potencjalne zastosowanie w rozwiązaniach mobilnych oraz systemach o ograniczonych zasobach obliczeniowych.

Materiały i metody

Przygotowany zbiór nagrań wokalizacji ptaków obejmuje 31 gatunków. Rejestracji dźwięków dokonano z częstotliwością próbkowania 48 kHz. Zdecydowana większość gatunków ptaków, których głosy analizowano, objęta jest ochroną gatunkową. W aktualnie badanym zbiorze gatunek *Acrocephalus paludicola* (wodniczka) zaliczany jest do grupy gatunków narażonych na wyginięcie [30]. Niemniej jednak status ochrony poszczególnych gatunków może ulegać okresowym zmianom, w związku z czym nie jest możliwe jednoznaczne określenie, które z nich w przyszłości będą wymagały objęcia czynną ochroną gatunkową. Należy podkreślić, że analizowany zbiór nagrań nie jest w pełni reprezentatywny dla awifauny Polski, gdzie stwierdzono występowanie około 492 gatunków ptaków [31]. Niemniej jednak obejmuje on głosy 18 pospolitych, bardzo często spotykanych rodzimych gatunków ptaków, 9 gatunków średnio licznych oraz 4 gatunków rzadkich (w tym wspomnianej wcześniej wodniczki). W sieci dostępne są obecnie znacznie obszerniejsze zbiory nagrań ptaków z terenu Polski, takie jak Xeno-Canto [22], Macaulay Library [32] czy Tierstimmenarchiv [33]. Przykładowo baza Xeno-Canto zawiera około 24 000 nagrań audio 370 gatunków ptaków z Polski, baza Macaulay Library około 12 700 nagrań (brak informacji o liczbie gatunków), natomiast baza Tierstimmenarchiv obejmuje 401 nagrań 45 gatunków. Nagrania dostępne w wymienionych bazach nie zawsze są jednak poprawnie oznaczone, a ponadto z reguły nie są bezpośrednio przygotowane do eksperymentów z zakresu automatycznego rozpoznawania głosów ptaków – w przeciwieństwie do analizowanego w niniejszym artykule, starannie opracowanego zbioru nagrań.

Dane zostały podzielone na zestaw uczący 60%, zestaw walidacyjny 20% oraz testowy 20%. Analiza liczebności danych z tabeli 1 wskazuje, iż niektóre gatunki są reprezentowane w bardzo niewielkim stopniu. W konsekwencji *Passer domesticus* (wróbel zwyczajny) posiadający tylko 3 obserwacje został pominięty w dalszych badaniach. Ponadto *Anas platyrhynchos* (krzyżówka

zwyczajna) oraz *Sylvia curruca* (pieżga) posiadają jedynie po 1 obserwacji w zestawie walidacyjnym i testowym.

Tabela 1. Charakterystyka zbioru danych użytych w badaniach

Gatunek	Liczba nagrań	Czas min. [s]	Czas maks. [s]	Czas średni [s]
Acrocephalus paludicola	73	0,24	0,75	0,40
Anas platyrhynchos	7	0,26	0,55	0,37
Anthus pratensis	39	0,29	0,31	0,30
Asio otus	121	0,55	0,83	0,69
Buteo buteo	34	0,32	1,31	0,78
Carpodacus erythrinus	14	1,10	1,34	1,21
Corvus frugilegus	104	0,39	0,97	0,54
Corvus monedula	33	0,25	0,44	0,33
Crex crex	126	0,37	0,44	0,41
Cuculus canorus	118	0,29	0,45	0,36
Dendrocopos major	48	0,23	0,44	0,25
Emberiza hortulana	9	1,30	1,56	1,43
Fringilla coelebs	291	0,27	3,26	0,69
Garrulus glandarius	70	0,46	2,62	1,01
Hirundo rustica	18	0,23	0,29	0,26
Jynx torquilla	204	0,29	0,37	0,33
Lanius collurio	17	0,31	0,37	0,34
Luscinia luscinia	109	0,22	0,55	0,29
Parus ater	47	0,35	0,39	0,38
Parus major	52	0,49	1,09	0,73
Passer domesticus[a]	3	0,39	0,46	0,43
Passer montanus	59	0,23	0,42	0,29
Phylloscopus collybita	151	0,25	0,41	0,36
Phylloscopus trochilus	150	0,11	0,39	0,30
Pica pica	15	0,28	2,08	0,68
Sitta europaea	36	0,45	0,50	0,47
Strix aluco	19	1,04	3,07	1,76
Strix uralensis	80	0,60	1,76	1,08
Sturnus vulgaris	23	0,88	1,62	1,21
Sylvia curruca	6	1,15	1,61	1,41
Turdus pilaris	13	0,25	0,27	0,26
Upupa epops	106	0,34	0,79	0,57
RAZEM	2195			

[a] *Passer domesticus* został wykluczony w dalszych eksperymentach.

Gatunki ptaków różnią się ponadto długością czasu wokalizacji, co w konsekwencji wpływa na długość zarejestrowanych sygnałów akustycznych. Średni czas nagrania dla dzięcioła dużego (*Dendrocopos major*) wynosi 0,25 sekundy, podczas gdy dla puszczyka (*Strix aluco*) wynosi 1,76 sekundy. Ta różnorodność ma wpływ na dalsze postępowanie z pozyskanymi sygnałami akustycznymi.

Wstępne przetwarzanie sygnałów

Nagrania wokalizacyjne ptaków przed wprowadzeniem do sieci neuronowej wymagają wstępnego przetworzenia. Standardowym podejściem w przypadku danych akustycznych jest transformacja surowego sygnału do reprezentacji liczbowej w postaci sekwencji współczynników MFCC. W niniejszych badaniach do ekstrakcji współczynników MFCC z sygnału akustycznego wykorzystano funkcję librosa.feature.mfcc dostępną w bibliotece Librosa dla języka Python.

W badaniu wykonano wstępne badania dotyczące dwóch parametrów sygnału akustycznego: czasu trwania oraz liczby ekstrahowanych współczynników MFCC.

Czas trwania sygnału został sprawdzony w trzech konfiguracjach: 1, 2 oraz 3 sekundy. W poprzednim rozdziale wykazano, że czas trwania większości nagrań mieści się w przedziale od 1 do 3 sekund, więc można z wysokim prawdopodobieństwem założyć, iż kluczowe cechy akustyczne mieszczą się w tym przedziale. Przed wprowadzeniem sygnałów do klasyfikatora wykonywany jest algorytm badający długość nagranego sygnału. Dane wejściowe zawierają dwie zmienne: A – tablica próbek dźwiękowych oraz L – referencyjna długość sygnału. Algorytm składa się z 3 głównych kroków:

1. Pobranie aktualnej długości nagranego sygnału (N).
2. Badanie relacji pomiędzy N oraz L:
   a) Przypadek I: N = L Nagranie trwa dokładnie L. Przejdź do ostatniego kroku.
   b) Przypadek II: N < L Nagranie krótsze od długości referencyjnej. Oblicz, ile zer należy dopisać do próbki R = L – N. Przygotuj tablicę zer o długości R i dołącz do nagrania A = A + [0, 0, ..., 0].
   c) Przypadek III: N > L Nagranie przekracza referencyjną długość. Przytnij do długości L. Tablica A = A [0 : L]
3. Zwróć wynik w postaci tablicy A o długości L.

Wśród nagrań błędnie rozpoznanych było 5 nagrań o czasie trwania krótszym niż 1s oraz 3 nagrania o czasie trwania powyżej 1 sekundy. Ponadto liczba dobrze rozpoznanych przykładów w zbiorze testowym wśród gatunków u których wystąpiły błędy rozpoznawania wynosiła {1, 2, 2, 9, 11, 21, 23, 57} z medianą równą 10. Wpływ stałej długości okna czasowego na skuteczność klasyfikacji poszczególnych gatunków ptaków różni się. Niektóre gatunki, charakteryzujące się, krótką i zwartą wokalizacją pozwalają na przetwarzanie całej struktury zawołania. Dla części nagrań gatunków o dłuższych i złożonych wokalizacjach zmniejszone okno czasowe (1 lub 2 sekundy) może powodować, że sygnał wejściowy przekaże jedynie fragment zawołania. W praktyce zaproponowany model zawiera dwuetapowy proces wyboru najistotniejszych cech nagrania. Pierwszy mechanizm w postaci warstwy LSTM pozwala na zachowywanie najważniejszych elementów sygnału w procesie uczenia modelu. Drugi etap w postaci mechanizmu uwagi Luonga pozwala modelowi na selektywne skupienie się na cechach spektralno-czasowych obecnych w analizowanym oknie niezależnie od faktu, czy stanowi ona całą wokalizację, czy jedynie jej część. W konsekwencji okazuje się, że długość nagrania nie jest kluczowym czynnikiem determinującym jakość klasyfikacji dla wybranych gatunków ptaków.

Drugi analizowany parametr, określony jako liczba ekstrahowanych współczynników MFCC, miał na celu ocenę wpływu tej wielkości na skuteczność procesu klasyfikacji wokalizacji ptaków. W niniejszych badaniach rozpatrzono dwa warianty: 20 oraz 30 współczynników MFCC. Pozostałe parametry ekstrakcji cech MFCC ustawiono na wartości domyślne funkcji librosa.feature.mfcc dostępnej w bibliotece Librosa dla języka Python. Szczegółowe wartości parametrów transformacji sygnału do reprezentacji MFCC przedstawiono w tabeli 2.

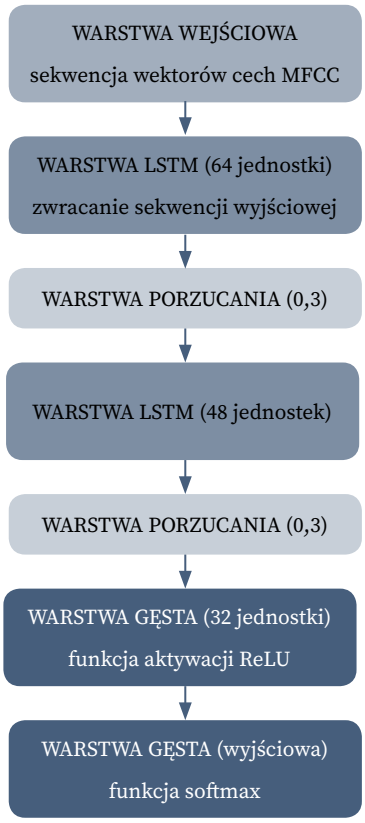
Tabela 2. Parametry transformacji sygnału akustycznego do reprezentacji MFCC

Parametr	Wartość	Opis
Parametr	Wartość	Opis
duration	1, 2 lub 3	Czas trwania sygnału (sekundy)
sr	48000	częstotliwość próbkowania sygnału (Hz)
N_fft	2048	Długość okna FFT – długość ramki (okna czasowego) w próbkach
hop_length	512	liczba próbek między kolejnymi ramkami
window	‘hann’	Typ okna czasowego, wybrano okno Hanna
power	2	Moduł widma FFT podniesiony do potęgi 2
n_mels	128	Liczba pasm (filtrów) melowych
mel_norm	‘slaney’	Normalizacja filtrów melowych według metody Slaney’a
lifter	0	Brak lifteringu
dct_type	2	Typ transformaty DCT, wybrano DCT typu II
norm	‘ortho’	Dokonano orto-normalizacji macierzy przekształcenia DCT
n_mfcc	20 lub 30	liczba współczynników MFCC

Liczba filtrów (pasm) melowych to liczba podasm częstotliwościowych, na których wykonuje się transformację DCT. Wynikiem tej operacji jest 20 lub 30 współczynników DCT, które stanowią obliczone parametry MFCC. Te parametry są również cechami podawanymi na wejście sieci neuronowej.

Modele sieci neuronowych

W ramach badań przygotowano modele sieci neuronowych opartych na warstwach rekurencyjnych LSTM.



Rysunek 1. Struktura zastosowanej sieci LSTM

Badana sieć LSTM składa się z kilku warstw połączonych ze sobą jak na rys. 1. Sieć zawiera warstwę wejściową, która przyjmuje dane akustyczne w postaci sekwencji wektorów cech MFCC. Następnie dane przekazywane są do warstwy LSTM z 64 neuronami, w której zwracana jest sekwencja wyjściowa (return sequences). W celu zwiększenia odporności sieci na przeuczenie zastosowano warstwę porzucania (dropout) o eksperymentalnie dobranej wartości 0,3. Kolejnym elementem architektury jest warstwa LSTM z 48 neuronami, po której ponownie zastosowano warstwę porzucania o wartości 0,3. Przedostatnia warstwa składa się z 32 neuronów z funkcją aktywacji w postaci jednostronnie obciętej funkcji liniowej (ReLU):

$$ReLU(x) = \begin{cases} 0 \; dla \; x < 0 \\ x \; dla \; x \geq 0 \end{cases} \quad (1)$$

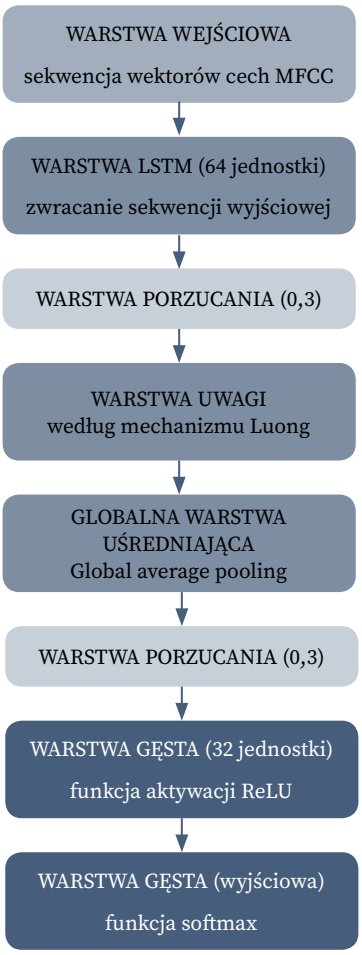
Ostatnią warstwą sieci LSTM jest warstwa wyjściowa, która zawiera liczbę neuronów odpowiadającą liczbie kategorii głosów ptaków czyli 31. W warstwie tej zastosowano znormalizowana funkcję wykładniczą (softmax):

$$Softmax(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (2)$$

gdzie:
z_i – wartość wyjściowa dla danej klasy;
e^{z_i} – funkcja wykładnicza o wartości dodatniej;
$\sum_{j=1}^{K} e^{z_j}$ – suma wartości wykładniczych dla wszystkich 31 kategorii ptaków.

W praktyce funkcja softmax ma za zadanie zwrócić prawdopodobieństwo przynależności próbki akustycznej do konkretnej klasy. W takiej konfiguracji sieć posiada 46 047 parametrów, przy czym wszystkie są używane w uczeniu modelu.

Wprowadzona modyfikacja polegała przede wszystkim na wprowadzeniu mechanizmu uwagi i redukcji liczby parametrów modelu. Zmodyfikowany układ sieci neuronowej został pokazany na rysunku 2.



Rysunek 2. Struktura zastosowanej sieci z mechanizmem uwagi (LSTM + MU)

Architektura sieci składa się z liczby warstw zbliżonej do tej zastosowanej w poprzednim modelu. Pierwszą z nich jest warstwa wejściowa, która służy do przyjęcia danych wejściowych w postaci sekwencji wektorów cech MFCC. Po niej występuje warstwa rekurencyjna z 64 neuronami w stylu LSTM, po której następuje warstwa porzucania o wartości 0,3. Następnie potok przetworzonych danych jest podawany na warstwę uwagi według mechanizmu Luong [34]. Zadaniem tej warstwy jest przegląd wszystkich stanów ukrytych warstwy LSTM naraz, aby zdecydować, które fragmenty sekwencji wejściowej są najważniejsze w danym momencie. Model sprawdza, które części nagrania dźwiękowego (w postaci ciągu wektorów MFCC) korelują z częściami tego samego nagrania. Kolejną warstwą jest globalna warstwa uśredniająca (*Global Average Pooling*). Zadaniem tej warstwy jest redukcja wymiarowości danych przy zachowaniu kluczowych informacji o cechach wyekstrahowanych przez poprzednie warstwy modelu. Dwie ostatnie warstwy to podobnie jak w przypadku pierwszej sieci, tak zwane warstwy gęste. Ich zadaniem jest dokonanie ostatecznej klasyfikacji ptaków na podstawie cech wydobytych w poprzednich warstwach modelu. Ta konfiguracja modelu posiada zaledwie 27 423 parametry. Jest to znaczna redukcja względem poprzedniej sieci LSTM.

W proponowanej architekturze (rysunek 2) zastosowano mechanizm uwagi zainspirowany koncepcją globalnej atencji Luonga [34], dostosowany do pracy w sieci typu jednokierunkowego (bez oddzielnego bloku dekodera). Klasyczny mechanizm Luonga wylicza wektor kontekstowy na podstawie relacji między stanem dekodera a stanami enkodera. W niniejszej pracy mechanizm ten został zaimplementowany jako samouwaga (ang. *self-attention*). Proces ten można rozpisać w kolejnych krokach:

1. Warstwa LSTM przetwarza wejściową sekwencję współczynników MFCC, zwracając pełną sekwencję stanów ukrytych $H = [h_1, h_2, ..., h_T]$ gdzie T oznacza liczbę kroków czasowych.
2. W procesie obliczania wag zastosowano funkcję punktową do wyznaczenia relacji między każdym krokiem czasowym. W tej konfiguracji stany LSTM pełnią rolę zapytania (Q) oraz klucza (K). Wynik dopasowania obliczany jest jako:

$$score(h_t, h_s) = h_t^T h_s \tag{3}$$

3. Wagi są normalizowane przy użyciu funkcji softmax, tworząc mapę uwagi określającą, które fragmenty nagrania mają kluczowe znaczenie dla klasyfikacji:

$$\propto_{ts} = \frac{\exp\left(score(h_t, h_s)\right)}{\sum_{i=1}^{T} \exp\left(score(h_t, h_i)\right)} \tag{4}$$

4. Wyjściowy tensor atencji jest ważoną sumą wektorów wartości (V), co pozwala na agregację istotnych informacji z całej sekwencji w jedną, skondensowaną reprezentację, która następnie poddawana jest globalnemu uśrednianiu (*Global Average Pooling 1D*).

Zastosowanie wariantu globalnego oznacza, że model bierze pod uwagę wszystkie kroki czasowe w oknie analizy (3 sekundy sygnału), co jest szczególnie istotne w rozpoznawaniu aktywności akustycznej, gdzie kluczowe cechy mogą pojawiać się w różnych momentach nagrania. Rezygnacja z lokalnego wariantu atencji wynika z relatywnie krótkiego czasu trwania próbek, co eliminuje potrzebę ograniczania okna uwagi.

Przebieg eksperymentu

Hiperparametry zastosowanych modeli otrzymały identyczne wartości. Trening prowadzono z parametrem maksymalnej liczby epok wynoszącym 1500 przy rozmiarze wsadu (*batch size*) równym 32. Wykorzystano optymalizator Adam [35] z krokiem uczenia (*learning rate*) na poziomie 0,001 (pozostałe parametry algorytmu pozostały domyślne) oraz funkcję straty w postaci rzadkiej kategorialnej entropii krzyżowej (*sparse categorical crossentropy*). W celu optymalizacji procesu zastosowano mechanizm wczesnego zatrzymania (*early stopping*) z parametrem cierpliwości (*patience*) wynoszącym 300 epok, monitorując wartość funkcji straty na zbiorze walidacyjnym. W konsekwencji zbiór walidacyjny pełnił potrójną rolę: służył do bieżącego monitorowania zjawiska przeuczenia, umożliwił wybór najlepszej wersji modelu oraz pozwolił na automatyczne zakończenie obliczeń. Dobór hiperparametrów został przeprowadzony drogą eksperymentalną.

Badania zrealizowano w środowisku obliczeniowym Anaconda, wykorzystując środowisko programistyczne (IDE) Spyder. Implementację modeli oraz proces przetwarzania danych przeprowadzono w języku Python przy użyciu bibliotek: TensorFlow [36] w wersji 2.10.0 oraz Keras [37] w wersji 2.9.0 (budowa i trenowanie sieci), Librosa [38] w wersji 0.11.0 (ekstrakcja cech sygnałów dźwiękowych), NumPy [39] w wersji 1.23.5 (operacje macierzowe) oraz Scikit-learn [40] w wersji 1.1.3 (podział danych i metryki). Wizualizację wyników wykonano z wykorzystaniem bibliotek Matplotlib [41] w wersji 3.5.2 oraz Seaborn [42] w wersji 0.12.2.

Obliczenia zostały przeprowadzone na stacji roboczej wyposażonej w kartę graficzną NVIDIA GeForce RTX 2060 (12 GB VRAM), co pozwoliło na wykorzystanie akceleracji sprzętowej CUDA w procesie uczenia.

W celu zapewnienia wiarygodności wyników, proces uczenia i ewaluacji modelu przeprowadzono w trzech niezależnych iteracjach (zestawach). W każdej z nich dokonano losowego podziału zbioru danych na trzy rozłączne podzbiory: treningowy, walidacyjny oraz testowy. Takie podejście umożliwia weryfikację, która z analizowanych metod wykazuje ogólną tendencję do uzyskiwania lepszych wyników, a także pozwala potwierdzić stabilność działania modeli niezależnie od konkretnego doboru zbiorów treningowych, walidacyjnych i testowych.

Ocena skuteczności zaproponowanego modelu została dokonana w oparciu o zestaw standardowych metryk klasyfikacyjnych. Kluczowymi wskaźnikami poddanymi analizie były: dokładność (ang. *accuracy*), miara *F*1 (*F*1-Score) oraz średni obszar pod krzywą charakterystyki operacyjnej odbiornika (Śr. AUC).

Ze względu na wieloklasowy charakter analizowanego problemu, skuteczność modelu oceniono stosując strategię jeden kontra reszta (*One-vs-Rest*). W tym podejściu metryki są obliczane oddzielnie dla każdej z k klas, traktując wybraną klasę jako pozytywną, a wszystkie pozostałe jako negatywne. W celu uzyskania pojedynczej wartości opisującej ogólną wydajność modelu, zastosowano uśrednianie makroskopowe (*macro-averaging*). Metoda ta przypisuje taką samą wagę każdej kategorii, co pozwala na rzetelną ocenę zdolności predykcyjnych modelu również w odniesieniu do klas mniej licznych. W kontekście klasyfikacji wieloklasowej poszczególne elementy macierzy pomyłek przyjmują oznaczenia:

- TP_i (*True Positives*) – liczba prawdziwie dodatnich klasyfikacji dla klasy i;
- TN_i (*True Negatives*) – liczba prawdziwie ujemnych klasyfikacji dla klasy i;
- FP_i (*False Positives*) – liczba błędnie dodatnich klasyfikacji dla klasy i;
- FN_i (*False Negatives*) – liczba fałszywie ujemnych klasyfikacji dla klasy i.

Dokładność (*accuracy*) definiowana jest jako stosunek liczby poprawnych predykcji modelu do całkowitej liczby próbek i wyrażana jest wzorem:

$$Dokładność = \frac{\sum_{i=1}^{k} TP_i}{N} \tag{5}$$

gdzie:
k – oznacza liczbę klas; natomiast N – stanowi całkowitą liczbę wszystkich próbek w zbiorze testowym.

*F*1-Score to metryka stanowiąca średnią harmoniczną precyzji i czułości modelu. Jest stosowana jako alternatywa dla dokładności w przypadku niezbalansowanych zbiorów danych. Precyzja obliczana jest na podstawie wzoru:

$$Precyzja_{macro} = \frac{1}{k}\sum_{i=1}^{k} \frac{TP_i}{TP_i+FP_i} \tag{6}$$

Natomiast czułość (*recall*) wyraża się wzorem:

$$Czułość_{macro} = \frac{1}{k}\sum_{i=1}^{k} \frac{TP_i}{TP_i+FN_i} \tag{7}$$

Stąd współczynnik F1-Score najczęściej w literaturze przyjmuje postać:

$$F1 - Score_{macro} = 2 \cdot \frac{Precyzja_{macro} \cdot Czułość_{macro}}{Precyzja_{macro} + Czułość_{macro}} \tag{8}$$

W przypadku analizy skuteczności modeli uczonych na niezbalansowanych danych stosuje się krzywe charakterystyki operacyjnej odbiornika (ROC), a następnie wyznacza się obszar pod krzywą (AUC). Obszar pod krzywą ROC można formalnie opisać następującym wzorem:

$$AUC_i = \int_0^1 TPR_i(FPR_i)dFPR_i \tag{9}$$

gdzie:
TPR_i (ang. *True Positive Rate*) – czułość dla klasy i,
FPR_i (ang. *False Positive Rate*) – współczynnik fałszywych alarmów dla klasy i.

Przyjmując, że k to liczba klas możemy wyznaczyć średnią wartość obszarów AUC:

$$Śr.\ AUC = \frac{\sum_i^k AUC_i}{k} \tag{10}$$

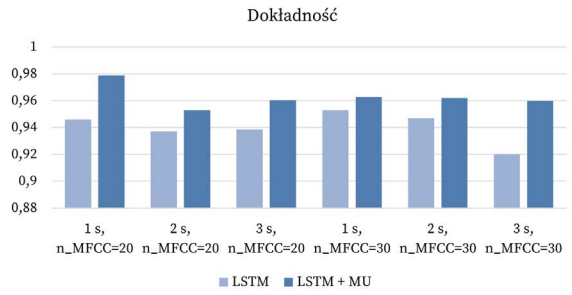
Wyniki i dyskusja

Wszystkie uzyskane wyniki badań umieszczone zostały w tabeli 3. Tabela zawiera wyniki ewaluacji modeli wyłącznie na danych testowych, czyli takich, które nawet w minimalnym stopniu nie uczestniczyły w procesie uczenia modeli.

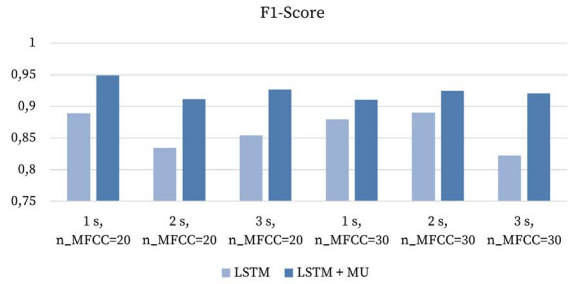
Porównanie wartości parametrów otrzymanych za pomocą sieci LSTM oraz sieci LSTM z dodanym mechanizmem uwagi (LSTM + MU) zostało zaprezentowane na rysunkach 3, 4 i 5.

Tabela 3. Wyniki eksperymentu na dla danych testowych

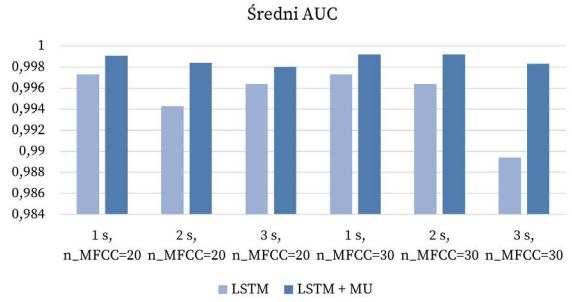
	LSTM			LSTM + MU		
Użyte dane	**Dokładność**	***F*1-Score**	**Śr. AUC**	**Dokładność**	***F*1-Score**	**Śr. AUC**
Parametry sygnału: Czas trwania 1 sekunda; liczba MFCC = 20						
Zestaw I	0,9544	0,8995	0,9976	0,9772	0,9343	0,9981
Zestaw II	0,9385	0,8833	0,9977	0,9772	0,9587	0,9995
Zestaw III	0,9453	0,8838	0,9967	0,9818	0,9548	0,9996
Średnia	0,9461	0,8889	0,9973	0,9787	0,9493	0,9991
Parametry sygnału: Czas trwania 2 sekundy; liczba MFCC = 20						
Zestaw I	0,9476	0,8651	0,9966	0,9590	0,9125	0,9987
Zestaw II	0,9248	0,8151	0,9932	0,9544	0,9146	0,9981
Zestaw III	0,9385	0,8220	0,9932	0,9453	0,9065	0,9984
Średnia	0,9370	0,8341	0,9943	0,9529	0,9112	0,9984
Parametry sygnału: Czas trwania 3 sekundy; liczba MFCC = 20						
Zestaw I	0,9134	0,8214	0,9953	0,9567	0,9099	0,9981
Zestaw II	0,9453	0,8653	0,9963	0,9636	0,9253	0,9965
Zestaw III	0,9567	0,8749	0,9977	0,9613	0,9453	0,9995
Średnia	0,9385	0,8539	0,9964	0,9605	0,9268	0,9980
Parametry sygnału: Czas trwania 1 sekunda; liczba MFCC = 30						
Zestaw I	0,9590	0,8924	0,9985	0,9727	0,9206	0,9988
Zestaw II	0,9408	0,8436	0,9939	0,9613	0,8937	0,9994
Zestaw III	0,9590	0,9026	0,9995	0,9544	0,9162	0,9996
Średnia	0,9529	0,8795	0,9973	0,9628	0,9102	0,9992
Parametry sygnału: Czas trwania 2 sekundy; liczba MFCC = 30						
Zestaw I	0,9590	0,9025	0,9967	0,9590	0,8962	0,9995
Zestaw II	0,9408	0,8871	0,9978	0,9658	0,9381	0,9994
Zestaw III	0,9408	0,8808	0,9946	0,9613	0,9389	0,9986
Średnia	0,9469	0,8901	0,9964	0,9620	0,9244	0,9992
Parametry sygnału: Czas trwania 3 sekundy; liczba MFCC = 30						
Zestaw I	0,9112	0,7667	0,9912	0,9567	0,9125	0,9976
Zestaw II	0,9089	0,8206	0,9873	0,9590	0,9122	0,9994
Zestaw III	0,9408	0,8788	0,9896	0,9636	0,9372	0,9978
Średnia	0,9203	0,8220	0,9894	0,9598	0,9206	0,9983



Rysunek 3. Porównanie wyników klasyfikacji dla sieci LSTM oraz LSTM + MU – parametr *Dokładność*



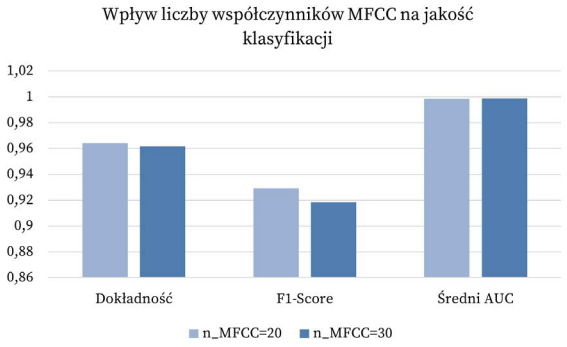
Rysunek 4. Porównanie wyników klasyfikacji dla sieci LSTM oraz LSTM + MU – parametr *F1-Score*



Rysunek 5. Porównanie wyników klasyfikacji dla sieci LSTM oraz LSTM + MU – parametr *Średni AUC*

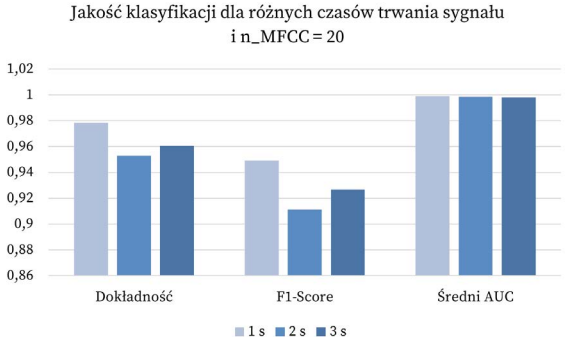
Wyniki zaprezentowane w tabeli 3 oraz na rysunkach 3, 4 i 5 wskazują, że wprowadzenie mechanizmu uwagi nie tylko zmniejszyło liczbę parametrów modelu, lecz również skutecznie poprawiło jakość klasyfikacji dla każdej badanej kombinacji parametrów przetwarzania sygnału (czas trwania + liczba współczynników MFCC). Uzyskana maksymalna dokładność klasyfikatora LSTM + MU jest również wyższa o 4,54% od dokładności uzyskanej za pomocą klasyfikatora HMM opisanego w pracy [1], który był trenowany na tym samym zbiorze nagrań poddanych bardzo zbliżonej procedurze wstępnego przetwarzania sygnału. Dalsza część analizy wyników została zatem ograniczona do przypadku klasyfikacji za pomocą modeli LSTM + MU.

Nieznacznie lepsze średnie rezultaty klasyfikacji uzyskano dla liczby współczynników MFCC = 20 w porównaniu z liczbą współczynników MFCC = 30. Dla parametru Dokładność i *F*1-Score nastąpiła poprawa odpowiednio o 0,25% i 1,07 %, natomiast dla parametru średni AUC nieznaczne pogorszenie o 0,04% (rysunek 6).

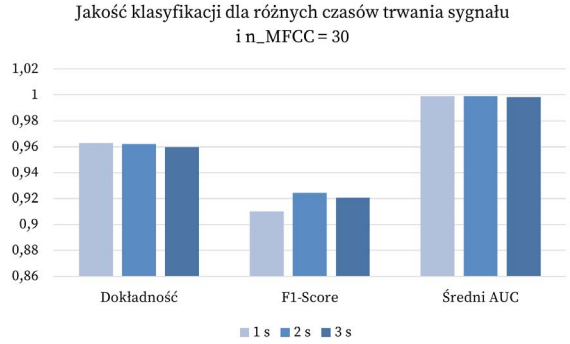


Rysunek 6. Porównanie średnich wyników klasyfikacji dla sieci LSTM + MU dla różnej liczby współczynników MFCC

Natomiast czas trwania daje niejednoznaczne rezultaty, gdyż większa liczba czynników MFCC wpływa lepiej na klasyfikację sygnałów 2 sekundowych, natomiast mniejsza liczba MFCC lepiej działa dla sygnałów 1-sekundowych. Zwiększenie długości sygnału do 3 sekund nie poprawia rezultatów. Może to wynikać zarówno z braku istotnych informacji we fragmentach nagrań dłuższych jak i z powodu przyjętej metody uzupełniania brakujących fragmentów sygnału (wypełniania zerami). W przyszłości warto by było skupić się na temacie przetestowania innych metod uzupełniania sygnału. Wpływ czasu trwania sygnału na jakość klasyfikacji został zaprezentowany na rysunkach 7 i 8.

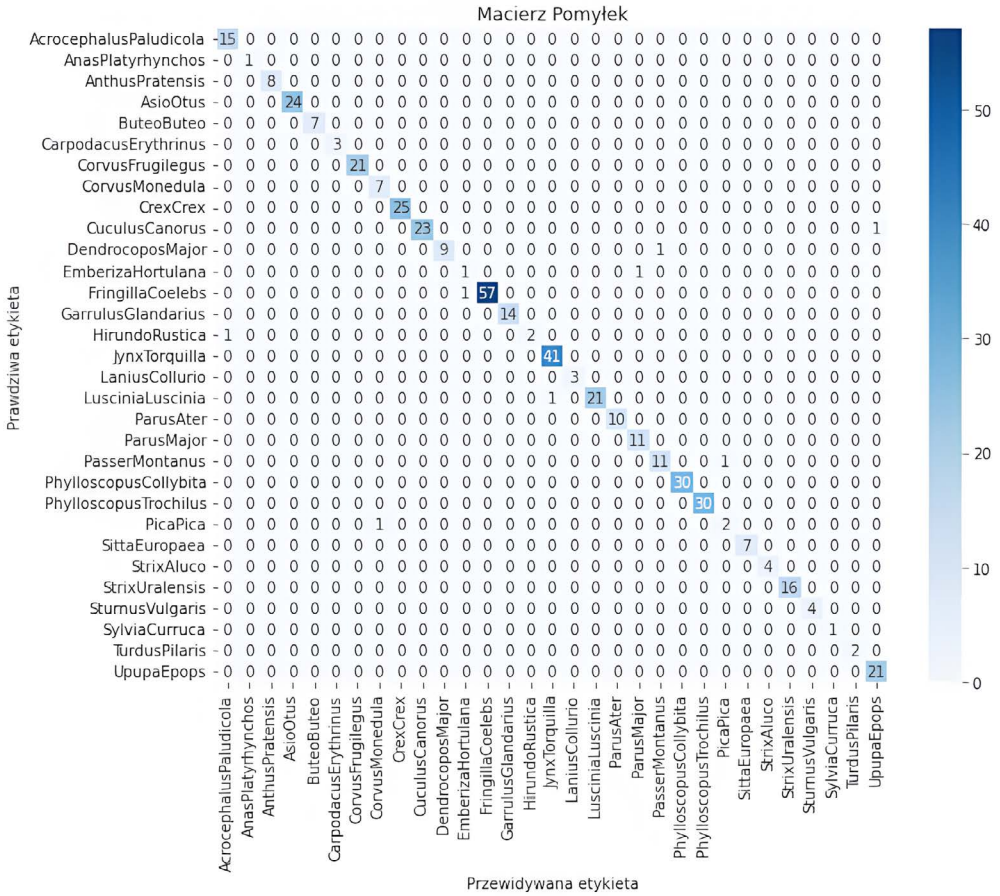


Rysunek 7. Porównanie średnich wyników klasyfikacji dla sieci LSTM + MU dla różnego czasu trwania sygnału i liczby współczynników MFCC = 20



Rysunek 8. Porównanie średnich wyników klasyfikacji dla sieci LSTM + MU dla różnego czasu trwania sygnału i liczby współczynników MFCC = 30

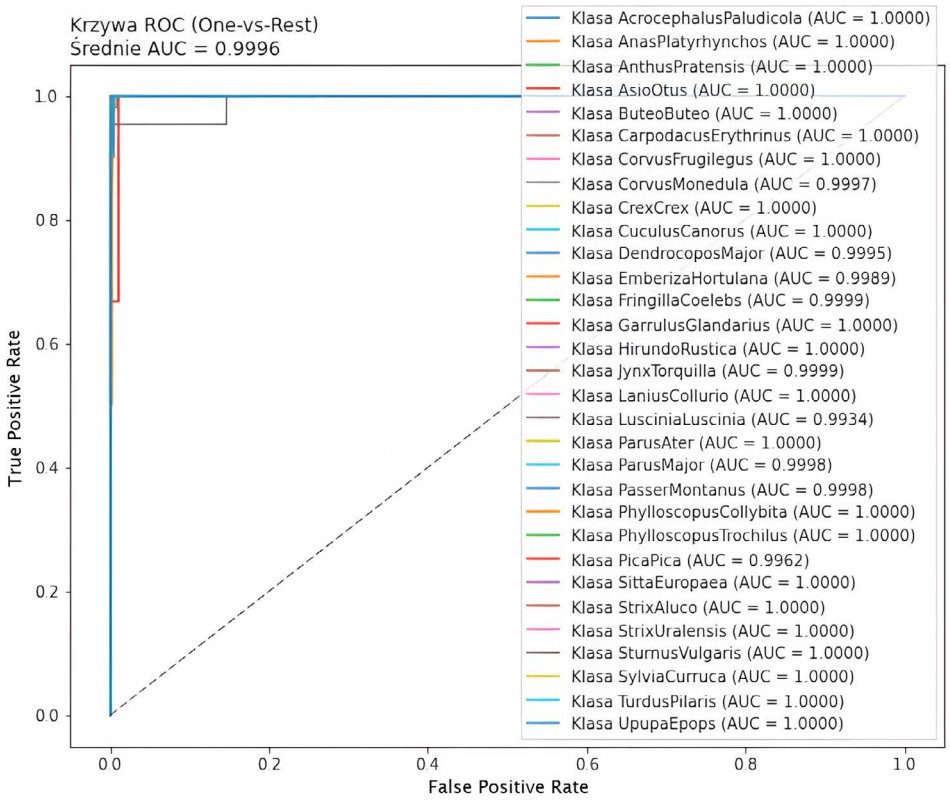
Średni najlepszy wynik uzyskany dla modeli LSTM + MU wyniósł odpowiednio 0,9787 dla współczynnika dokładności, 0,9493 dla *F*1-Score oraz 0,9991 dla Śr. AUC dla przypadku klasyfikacji nagrań 1-sekundowych przy 20 współczynnikach MFCC. Wynik ten jest dość zaskakujący ze względu na to, że ograniczenie czasu trwania sygnału akustycznego do 1 sekundy spowodowało wycięcie części sygnału i utratę informacji w nagraniach dłuższych niż 1 sekunda, a mimo to jakość klasyfikacji poprawiła się. Natomiast najlepsze wyniki klasyfikacji uzyskał model zbudowany na interwale 1 sekundy dla 20 współczynników MFCC oraz wylosowanych danych Zestawu III. W tym przypadku metryki *Dokładność* i *Średni AUC* osiągnęły najwyższe wartości we wszystkich badaniach, odpowiednio 0,9818 oraz 0,9996, natomiast metryka *F1-Score* osiągnęła drugi najlepszy wynik na poziomie 0,9548. Macierz pomyłek dla przedstawionego powyżej najlepszego modelu (rysunek 9) wskazuje, że ogółem pomylił się on jedynie w 7 przypadkach.



Rysunek 9. Macierz pomyłek modelu LSTM+MU dla Zestawu III, 1 sekundy oraz 20 MFCC

Również krzywe ROC wskazują, na wysoką dokładność wybranego modelu (rysunek 10) oraz że niemal wszystkie klasy są idealnie sklasyfikowane. Biorąc pod uwagę poniższe wyniki jedynie krzywa ROC dla słowika szarego (*Luscinia luscinia*) nieco odbiega od pozostałych uzyskując jednakże wysoki wynik 0,9934 AUC.

Wyniki eksperymentów potwierdzają założenie dotyczące niepogorszenia wyników klasyfikacji w wyniku zastosowania mechanizmu uwagi względem tradycyjnej sieci LSTM. Zaobserwowano, że uzyskiwane wyniki są w dużej mierze determinowane przez specyfikę wylosowanych podzbiorów danych (uczących, walidacyjnych i testowych). Wskazuje to na dużą wrażliwość modelu na strukturę danych wejściowych, co uzasadnia zastosowanie wielokrotnego powtórzenia procesu podziału i uczenia w celu uśrednienia metryk i uzyskania bardziej obiektywnej oceny skuteczności.



Rysunek 10. Zbiorcze zestawienie ROC modelu LSTM+MU dla Zestawu III, 1 sekundy oraz 20 MFCC

Wnioski

Badania wykazały zdecydowaną poprawę wyników klasyfikacji nagrań wokalizacji ptaków za pomocą sieci LSTM z wprowadzonym mechanizmem uwagi w porównaniu ze standardową siecią LSTM. Są to również wyniki zdecydowanie przewyższające dokładność klasyfikatora opartego na ukrytych modelach Markowa (HMM) opisywanego w pracy [1]. Mechanizm uwagi stanowi zatem podstawę do dalszego rozwoju metod klasyfikacji ptaków na podstawie danych wokalizacyjnych w szczególności do opracowania metody klasyfikacji sygnałów w oparciu o większe zbiory nagrań wokalizacji ptaków takie jak baza Xeno-Canto.

Skuteczność modeli zbudowanych na 1-sekundowych sygnałach pozwala ponadto na budowę bardzo szybkich urządzeń do klasyfikacji i detekcji ptaków w przyszłości. Co więcej zastosowanie mechanizmu uwagi znacznie redukuje liczbę parametrów sieci, co z kolei pozytywnie wpływa na potencjalne zastosowania w urządzeniach mobilnych.

Informacje uzupełniające

Plik S1. Przetworzone dane w postaci modelu zapisane w pliku w formacie HDF5 są dostępne wraz z artykułem. Dane te są udostępniane na licencji Creative Commons

Bibliografia

[1] Wielgat R, Potempa T, Świętojański P, Król D. On using prefiltration in HMM-based bird species recognition. W: 2012 International Conference on Signals and Electronic Systems (ICES). Wroclaw: IEEE; 2012:1–5. https://doi.org/10.1109/ICSES.2012.6382258.

[2] Kwan C, Ho, K, Mei G, Li Y, Ren Z, Xu R, Zhang Y, Lao D, Stevenson M, Stanford V, Rochet C. An automated acoustic system to monitor and classify birds. EURASIP Journal on Advances Signal Processing. 2006:096706(2006). https://doi.org/10.1155/ASP/2006/96706.

[3] Fagerlund S. Bird species recognition using support vector machines. EURASIP Journal on Advances in Signal Processing. 2007:038637(2007). https://doi.org/10.1155/2007/38637.

[4] Cai J, Ee D, Pham B, Roe P, Zhang J. Sensor network for the monitoring of ecosystem: Bird species recognition. W: 3rd International Conference on Intelligent Sensors, Sensor Networks and Information. Melbourne; 2007:293–298. https://doi.org/10.1109/ISSNIP.2007.4496859.

[5] Wielgat R, Zieliński TP, Potempa T, Lisowska-Lis A, Król D. HFCC based recognition of bird species. Signal Processing Algorithms, Architectures, Arrangements, and Applications SPA 2007, 7 Sept. 2007 – 7 Sept. 2007, Poznan, Poland. Poznan; 2007:129–134. https://doi.org/10.1109/SPA.2007.5903313.

[6] Noumida A, Rajan R. Multi-label bird species classification from audio recordings using attention framework. Applied Acoustics. 2022;197:108901. https://doi.org/10.1016/j.apacoust.2022.108901.

[7] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. W: IEEE Transactions on Acoustics, Speech and Signal Processing. 1980;28(4):357–366. http://dx.doi.org/10.1109/TASSP.1980.1163420.

[8] Rabiner L, Juang B-H. Fundamental of Speech Recognition. Englewood Cliffs: Prentice-Hall; 1993.

[9] Mermelstein P. Distance measures for speech recognition, psychological and instrumental. Pattern Recognition and Artificial Intelligence. 1976:374–388.

[10] Skowronski MD, Harris JG. Human factor cepstral coefficients. The Journal of the Acoustical Society of America. 2002;112(5)(Suppl.):2279. https://doi.org/10.1121/1.4779137.

[11] Wielgat R, Zieliński TP, Woźniak T, Grabias S, Król D. Automatic recognition of pathological phoneme production. Folia Phoniatrica et Logopaedica. 2008;60(6):323–331. https://doi.org/10.1159/000170083.

[12] Grzybowska J, Kłaczyński M. Computer-assisted HFCC-based learning system for people with speech sound disorders, XXII Annual Pacific Voice Conference (PVC). Krakow: IEEE; 2014:1–5. https://doi.org/10.1109/PVC.2014.6845423.

[13] Benba A, Jilbab A, Hammouch A. Using human factor cepstral coefficient on multiple types of voice recordings for detecting patients with Parkinson's disease. IRBM. 2017;38(6):346–351. https://doi.org/10.1016/j.irbm.2017.10.002.

[14] Gmyrek S, Libal U, Hossa R. The impact of training strategies on overfitting in vowel classification using PS-HFCC parametrization for automatic speech recognition. Archives of Acoustics. 2025;50(3):371–382. https://doi.org/10.24425/aoa.2025.154823.

[15] Zouhir Y, Zarka M, El Amraoui L, Ouni K. Auditory feature extraction approach for robust pathological voice recognition. Journal of Voice. 2026;[in press], https://doi.org/10.1016/j.jvoice.2025.12.031.

[16] Stastny J, Munk M, Juranek L. Automatic bird species recognition based on birds vocalization. EURASIP Journal on Audio, Speech, and Music Processing. 2018;2018:19. https://doi.org/10.1186/s13636-018-0143-7.

[17] Müller M. Fundamentals of Music Processing. Cham: Springer; 2015: 123–130. https://doi.org/10.1007/978-3-319-21945-5.

[18] Harte C, Sandler M, Gasser M. 2006. Detecting harmonic change in musical audio. W: AMCMM '06: Proceedings of the 1st ACM workshop on Audio and music computing multimedia. New York: Association for Computing Machinery; 2006:21–26. https://doi.org/10.1145/1178723.1178727.

[19] Shehab SA, Darwish A, Hassanien AE. (2024). Classifying bird songs based on chroma and spectrogram feature extraction. W: Hassanien AE, Darwish A, Elghamrawy SM, editors. Artificial Intelligence for Environmental Sustainability and Green Initiatives. Cham: Springer; 2024:105–126. https://doi.org/10.1007/978-3-031-63451-2_7.

[20] Zhang S, Gao Y, Cai J, Yang H, Zhao Q, Pan F. A novel bird sound recognition method based on multifeature fusion and a transformer encoder. Sensors. 2023;23(19):8099. https://doi.org/10.3390/s23198099.

[21] Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. W: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: Association for Computational Linguistics; 2014:1724–1734. https://doi.org/10.3115/v1/D14-1179.

[22] Vellinga WP, Planqué B, Vellinga W. Xeno-canto – bird sounds from around the world. Xeno-canto Foundation for Nature Sounds. Occurrence Dataset. [Internet]

5 listopada 2018 [cytowane 21 grudnia 2025]. Dostępne na: GBIF.org. https://doi.org/10.15468/qv0ksn.

[23] Han X, Peng J. Multi-label bird species classification using transfer learning network. Archives of Acoustics. 2025;50(2):223–233. https://doi.org/10.24425/aoa.2025.154812.

[24] Hu S, Chu Y, Wen Z, Zhou G, Sun Y, Chen A. Deep learning bird song recognition based on MFF-ScSEnet. Ecological Indicators. 2023;154:110844. https://doi.org/10.1016/j.ecolind.2023.110844.

[25] Wang Q, Song Y, Du Y, Yang Z, Cui P, Luo B. Hierarchical-taxonomy-aware and attentional convolutional neural networks for acoustic identification of bird species: A phylogenetic perspective. Ecological Informatics. 2024;80:102538. https://doi.org/10.1016/j.ecoinf.2024.102538.

[26] Schneider S, Baevski A, Collobert R, Auli M. Wav2vec: Unsupervised pre-training for speech recognition. W: Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech 2019). Graz: ISCA; 2019:3465–3469. https://doi.org/10.21437/Interspeech.2019-1873.

[27] Swaminathan B, Jagadeesh M, Vairavasundaram S. Multi-Label classification for acoustic bird species detection using transfer learning approach. Ecological Informatics. 2024;80:102471. https://doi.org/10.1016/j.ecoinf.2024.102471.

[28] Ilyass M, Farrugia N, Serizel R. Self-supervised learning for few-shot bird sound classification. W: 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). Seoul: IEEE; 2024:600–604. https://doi.org/10.1109/ICASSPW62465.2024.10627576.

[29] Han X, Jianxin P. Bird sound classification based on ECOC-SVM. Applied Acoustics. 2023;204:109245. https://doi.org/10.1016/j.apacoust.2023.109245.

[30] BirdLife International. Acrocephalus paludicola. The IUCN Red List of Threatened Species; 2022:e.T22714696A176687364.

[31] Avibase – światowy wykaz ptaków: Polska. [Internet, cytowane 27 kwietnia 2026]. Dostępne na: https://avibase.bsc-eoc.org/checklist.jsp?region=PL.

[32] Macaulay Library. Your wildlife media archive since 1929: Explore birds, amhibians, mammals, and more. [Internet, cytowane 27 kwietnia 2026]. Dostępne na: https://www.macaulaylibrary.org/.

[33] Museum für Naturkunde. Animal Sound Archive. [Internet, cytowane 27 kwietnia 2026]. Dostępne na: https://www.museumfuernaturkunde.berlin/forschung/sammlung/tierstimmenarchiv.

[34] Luong T, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. W: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics; 2015:1412–1421. https://doi.org/10.18653/v1/D15-1166.

[35] Kingma DP, Jimmy B. Adam: A method for stochastic optimization. ICLR. [Internet] 2015 [cytowane 29 kwietnia 2026]. Dostępne na: https://www.intel.com/content/dam/www/public/us/en/ai/documents/1412.6980.pdf.

[36] Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Software. White Paper. 9 listopada 2015 [cytowane 29 kwietnia 2026]. Dostępne na: https://www.tensorflow.org/extras/tensorflow-whitepaper2015.pdf.

[37] Chollet F. 2018. Keras: The python deep learning library. [Internet] 2018 [cytowane 29 kwietnia 2026]. Dostępne na: https://api.semanticscholar.org/CorpusID:215844202.

[38] McFee B, McVicar M, Faronbi D, et al. Librosa/librosa: 0.10.1. Zenodo [Internet] 2023 [cytowane 29 kwietnia 2026]. Dostępne na: https://doi.org/10.5281/zenodo.8252662.

[39] Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. Nature. 2020;585:357–362. https://doi.org/10.1038/s41586-020-2649-2.

[40] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12(85):2825–2830.

[41] Hunter JD. Matplotlib: A 2D graphics environment. Computing in Science and Engineering. 2007;9(3):\ 90–95. https://doi.org/10.1109/MCSE.2007.55.

[42] Waskom ML. Seaborn: Statistical data visualization. Journal of Open Source Software. 2021;6(60):3021. https://doi.org/10.21105/joss.03021.