# Wykorzystanie Technik Analizy Danych do Badania Lotnych Związków Organicznych w Próbkach Perfum oraz ich Zamienników

# Summary

This article presents the application of data analysis techniques for measuring the gas resistance of volatile organic compounds of perfume samples. The study was performed using a BME688 electronic nose. The paper focuses on exploratory data analysis, clustering and classification of fragrance samples. The paper shows how to build a highly effective predictive model, distinguishing between odours that are very similar to the human sense of smell. This approach has potential applications in the study of perfume quality and authenticity, enabling effective detection of imitations using statistical methods.

# Key words – DATA SCIENCE, VOLATILE ORGANIC COMPOUNDS, PERFUME

# Streszczenie

W niniejszym artykule zaprezentowano zastosowanie technik analizy danych dotyczących pomiarów rezystancji gazowej lotnych związków organicznych próbek perfum. Badanie wykonano za pomocą elektronicznego nosa BME688. W pracy skoncentrowano się na eksploracyjnej analizie danych, klastrowaniu oraz klasyfikacji próbek zapachowych. Artykuł ukazuje sposób budowy wysoce skutecznego modelu predykcyjnego, odróżniającego zapachy, które dla ludzkiego zmysłu węchu są bardzo podobne. Takie podejście ma potencjalne zastosowanie w badaniu jakości i autentyczności perfum, umożliwiając skuteczne wykrywanie imitacji przy użyciu metod statystycznych.

# Słowa kluczowe – ANALIZA DANYCH, LOTNE ZWIĄZKI ORGANICZNE, PERFUMY

# Wstęp

Perfumy, będące przedmiotem codziennego użytku, to mieszanina związków zapachowych, środków homogenizujących i wzmacniaczy oraz rozpuszczalników. W zależności od ilości rozpuszczalnika rozróżnia się perfumy właściwe (ekstrakty), wody perfumowane i wody toaletowe. Kompozycja zapachowa, decydująca o specyficznym zapachu, składa się z trzech akordów: bazowego, średniego (serca) oraz wysokiego (głowy). Zapachy dzielą się na kategorie takie jak cytrusowe, kwiatowe, orientalne czy drzewne. Różnorodność i złożoność zapachów sprawiają trudność ludzkiemu zmysłowi węchu, przy próbie ich rozróżniania. Rozwiązaniem tego problemu jest elektroniczny nos.

Elektroniczny nos to zespół czujników gazów, mierzących takie cechy jak rezystancja czy wilgotność. Takie urządzenia, w połączeniu z uczeniem maszynowym, pozwalają na rozwiązywanie wielu problemów. Elektroniczne nosy znajdują zastosowanie w badaniach zanieczyszczeń powietrza [1], klasyfikacji produktów spożywczych [2] czy analizie oddechu pacjentów w celu predykcji nowotworów płuc [3].

W pracy przedstawiono zastosowanie technik statystycznych oraz uczenia maszynowego w zadaniach związanych z elektronicznym nosem, pozwalając   
m. in. na odkrycie interesujących zależności stojących za zapachami,   
a w konsekwencji, zbudowanie modeli klasyfikacyjnych, skutecznie odpowiadających na zadane problemy.

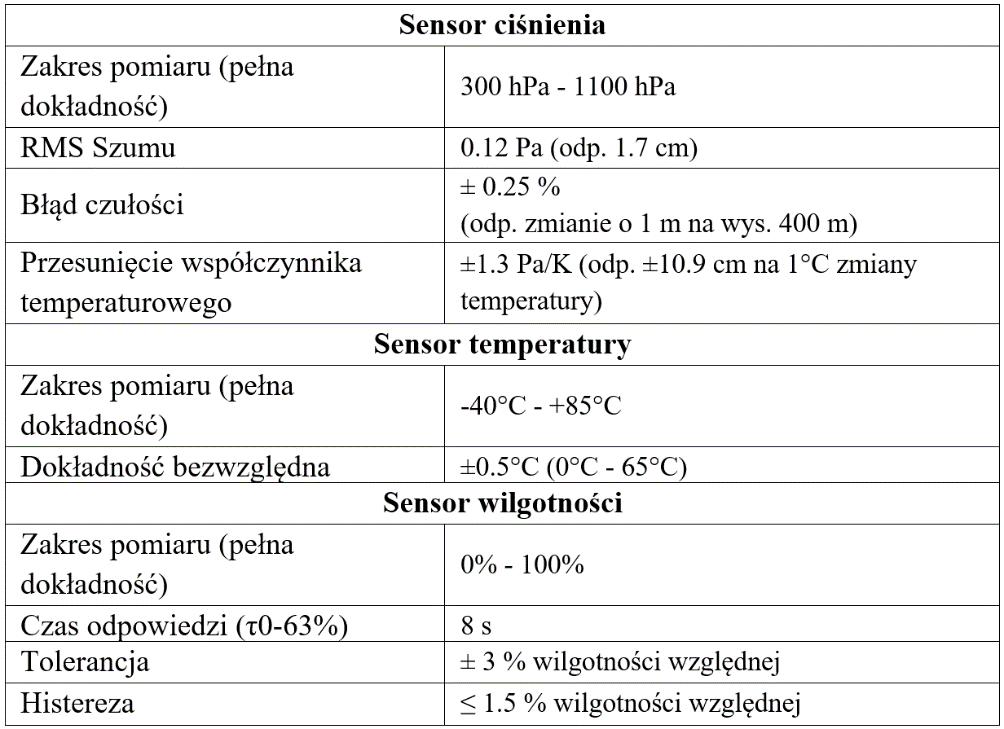
Praca składa się z czterech części. Pierwsza opisuje narzędzie i metodologię zbierania danych. Druga część przedstawia analizę eksploracyjną i przygotowanie danych do modelu predykcyjnego. Trzecia część omawia zastosowane metody. Ostatnia część prezentuje wyniki i wnioski z badania.

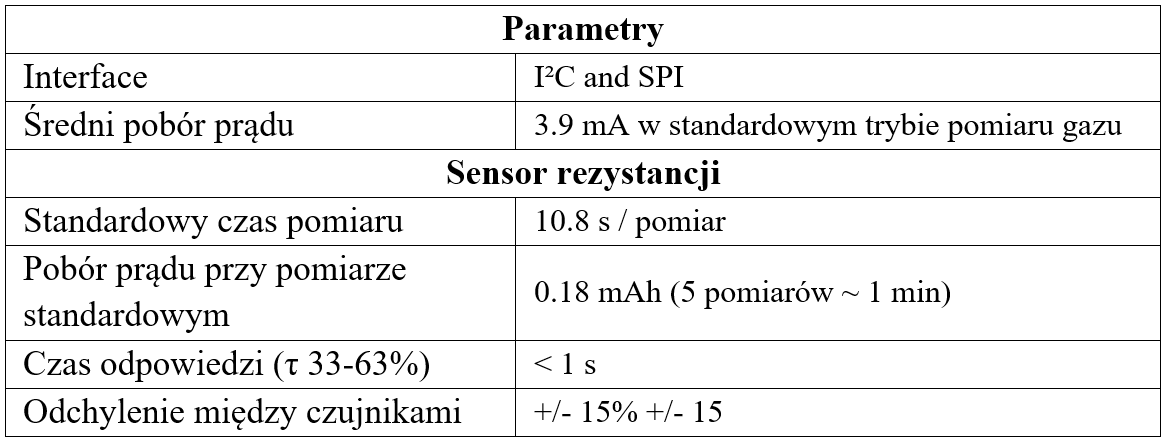
# Sprzęt oraz metodologia zbierania danych

Ta część poświęcona jest opisowi używanego sprzętu oraz metodologii zbierania danych do naszego badania. Przedstawiamy właściwości elektronicznego nosa, którym posłużyliśmy się w naszej pracy oraz szczegółowo omawiamy metodologię zbierania próbek lotnych związków organicznych pochodzących z perfum. Ostatnia sekcja prezentuje niektóre funkcjonalności programu BME AI-Studio, dedykowanego sensorom.

# Sprzęt

Na potrzeby naszych badań wykorzystaliśmy zespół Evaluation Kit Board BME688, będący połączeniem płytki deweloperskiej Adafruit HUZZAH32   
z wbudowanym modułem Wi-Fi ESP32, oraz płytki deweloperskiej BME688 zawierającej osiem sensorów BOSCH BME688. Sensory są zdolne do rejestracji dziesięciu punktów pomiarowych rezystancji gazu, ciśnienia atmosferycznego   
w zakresie od 300 hPa do 1100 hPa, temperatury w szerokim zakresie od -40°C do +85°C oraz wilgotności powietrza w pełnym zakresie. W tabeli poniżej przedstawiono właściwości używanego urządzenia.

Tabela 1.Właściwości Evaluation Kit Board BME688



Źródło: Opracowanie własne

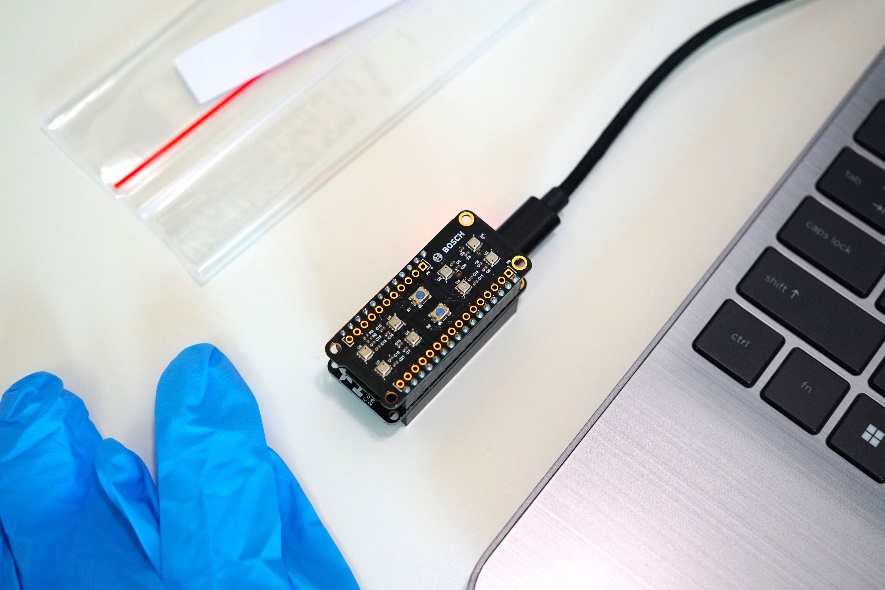
# Metodologia

Podczas zbierania próbek bardzo istotnym było zachowanie szczególnej ostrożności, by nie zanieczyścić płytki z sensorami, gdyż nawet niewielkie zabrudzenie powierzchni sensora mogłoby zaburzyć jego odczyty.

Aby uniknąć bezpośredniej ekspozycji sensorów na estry zawarte w badanych wodach perfumowanych, postanowiliśmy nakładać zapachy na paski z gęstego, niezadrukowanego papieru, pochodzącego z tej samej płachty papieru dla maksymalnej powtarzalności testu. W ten sposób przygotowany preparat pozostawiany był do odparowania alkoholu przez minutę. Następnie, wraz z płytką, preparat umieszczany był w szczelnym worku wykonanym z wysokiej gęstości polietylenu. Po trwającej 30 minut (±1 min.) sesji poboru głównej próbki, pobierany był pomiar poza workiem, samego otoczenia pozbawionego preparatu perfum. Trwający 10 minut (±0.5 min.) pomiar miał na celu umożliwienie późniejszego odfiltrowania rezydualnych cząstek zapachowych, które podczas pomiaru głównego mogły osadzić się na powierzchni zespołu.

Po zakończeniu danej sesji pomiarów, cały zespół sensorów pozostawiano do wywietrzenia przez okres ok. 2 godz. Warto zaznaczyć, że zadbaliśmy o to, by pomiary próbek oryginału i imitacji danego zapachu zawsze były oddzielone pomiarem wody perfumowanej innego producenta. W ten sposób ograniczyliśmy scenariusz, w którym wyniki badań wskazywałyby na wyższe podobieństwo zapachów, niż obserwowane w rzeczywistości, na przykład z powodu rezydualnych śladów substancji zapachowych perfum oryginalnych, obecnych   
w trakcie badania perfum nieoryginalnych.

Kolejna sesja pomiarów odbywała się przy użyciu wywietrzonego zespołu czujników, w wywietrzonym pomieszczeniu oraz z wykorzystaniem nowego worka.



Rys. 1. Evaluation Kit Board BME688 podczas zbierania próbek perfum  
Źródło: Opracowanie własne

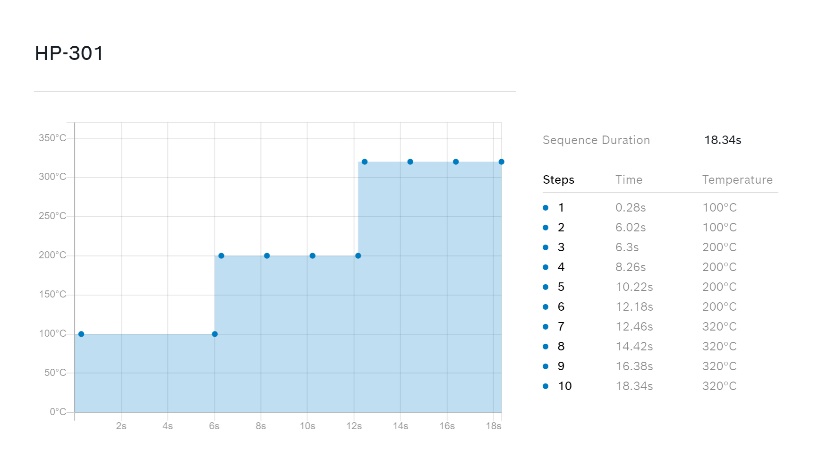
# Program

Zespół układów scalonych i sensorów współpracuje z oprogramowaniem BME AI-Studio, umożliwiającym dostosowanie sposobu, w jaki urządzenie będzie zbierało dane z otoczenia.

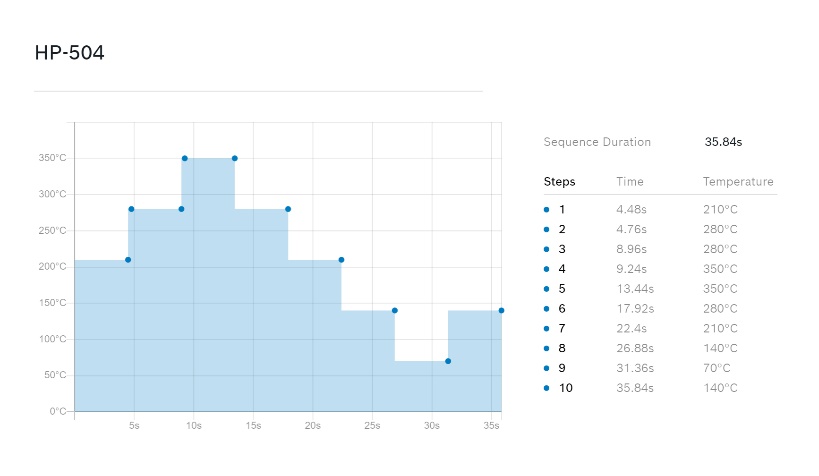
Ustalając szczegółowy program, szukaliśmy złotego środka pomiędzy różnorodnością i szerokością zakresu pomiarowego, a powtarzalnością   
i niezawodnością. Połączenie tych cech zapewniła kombinacja dwóch   
profili: HP-301 oraz HP-504, rozłożonych równomiernie na 4 pary sensorów.

Decydując o cyklu pracy układu wybraliśmy ustawienie RDC-1-0, zapewniające stałe próbkowanie przez cały okres trwania pomiaru.

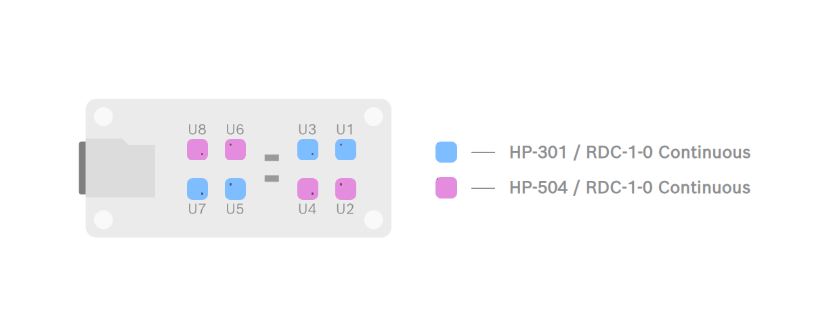
Konfiguracja sensorów oraz przebiegi wybranych do badania profili termicznych, prezentują pochodzące z programu BME AI-Studio grafiki zamieszczone poniżej.



Rys. 2. Przebieg profilu HP-301  
Źródło: Program BME AI-Studio



Rys. 3. Przebieg profilu HP-504  
Źródło: Program BME AI-Studio



Rys. 4. Układ sensorów BME688  
Źródło: Program BME AI-Studio

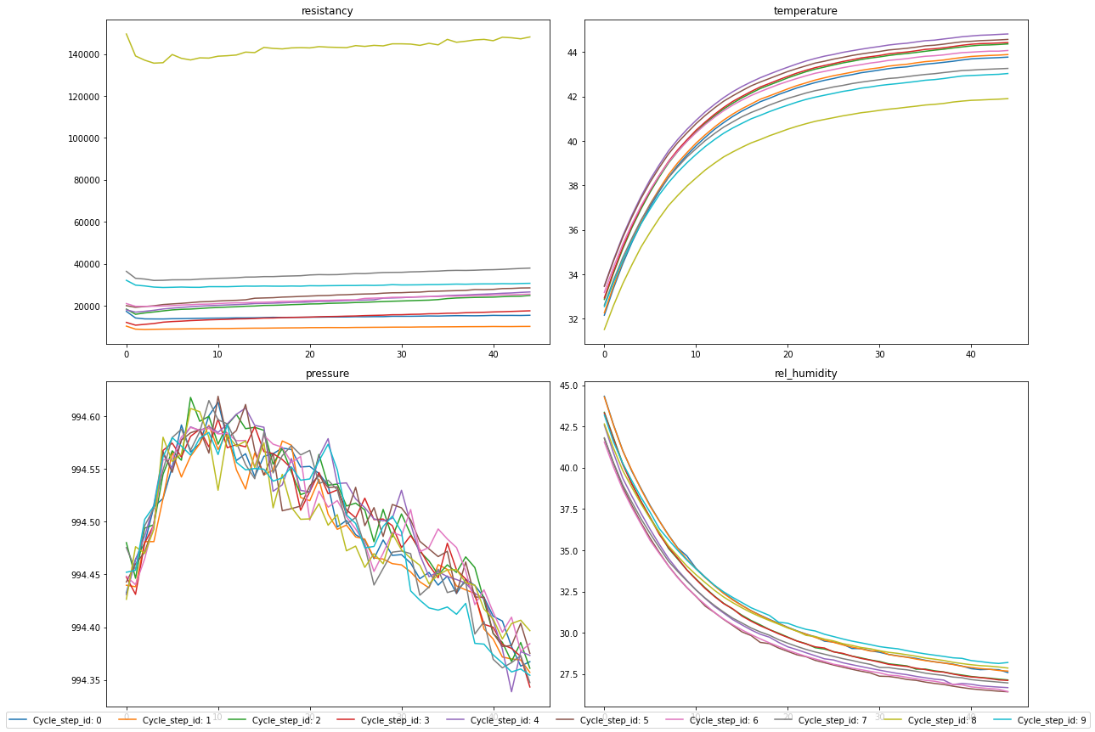
# Wykorzystanie i modelowanie danych

W tej części pracy zostaną przedstawione zebrane dane, zaprezentowana zostanie eksploracyjna analiza danych oraz porównanie zebranych danych w kontekście postawionego przedmiotu badania.

# Charakterystyka zebranych danych

W badaniu uwzględniono 4 pary perfum oryginalnych oraz fałszywych.   
Są to: *Euphoria* (*Calvin Klein*), *Good Girl* (*Carolina Herrera*), *The One* (*Dolce & Gabbana*) oraz *Y* (*Yves Saint Laurent*).

Elektroniczny nos zbiera dane w dziesięciu krokach pomiarowych dla czterech kanałów: rezystancji (w Ohmach), temperatury (w °C), ciśnienia (w hPa) oraz wilgotności (w %).

Jako przykład opracowania zebranych danych posłużą falsyfikaty perfum *Euphoria*.

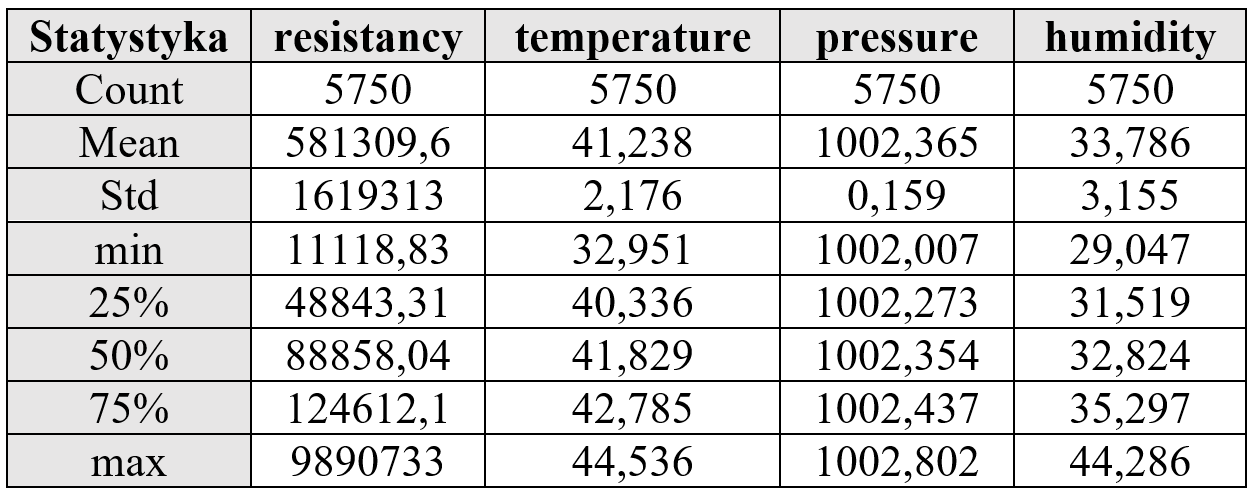
Rys. 5. Odczyty 4 kanałów w dziesięciu krokach dla falsyfikatów perfum Euphoria  
Źródło: Opracowanie własne

Na rysunku Rys. 5 widzimy tendencję badanych zmiennych dla przedstawionych perfum - zwiększa się temperatura otoczenia pod wpływem pracy układu, zmniejsza się wilgotność powietrza. Wraz z upływem czasu zbierania próbek, zwiększają się wartości rezystancji.

# Eksploracyjna analiza danych

W tabeli poniżej zaprezentowano statystyki opisowe danych odczytanych z czujnika gazów na przykładzie wyżej wymienionego zapachu.

*Tabela 2. Statystyki opisowe czterech zmiennych dla wybranego przykładu perfum*

*  
Źródło: Opracowanie własne*

Na podstawie powyższych statystyk opisowych czterech kanałów zebranych przez elektroniczny nos, dostrzec można, że układ scalony nagrzewał się podczas badania - od początkowej temperatury 32°C przeszedł do 44°C. Ciśnienie zmierzone w tym samym czasie pozostawało na względnie stałym poziomie. Wilgotność, podobnie jak i temperatura, ulegała zmianie podczas badania. Zebrane dane różnią się znacząco od siebie zakresami zmienności, co skłoniło nas ku decyzji o ich standaryzacji, według poniższej formuły:

Warto także zwrócić uwagę na minimalne oraz maksymalne wartości rezystancji – różnią się one od siebie niemal 900 razy. Wynika to z faktu, że przebieg rezystancji jest skalowany w zależności od kroku pomiarowego, co obserwujemy na Rys. 5.

# Porównanie oryginałów oraz falsyfikatów perfum

# 

*Rys. 6. Porównanie odczytów rezystancji dla oryginalnych perfum oraz ich falsyfikatów*

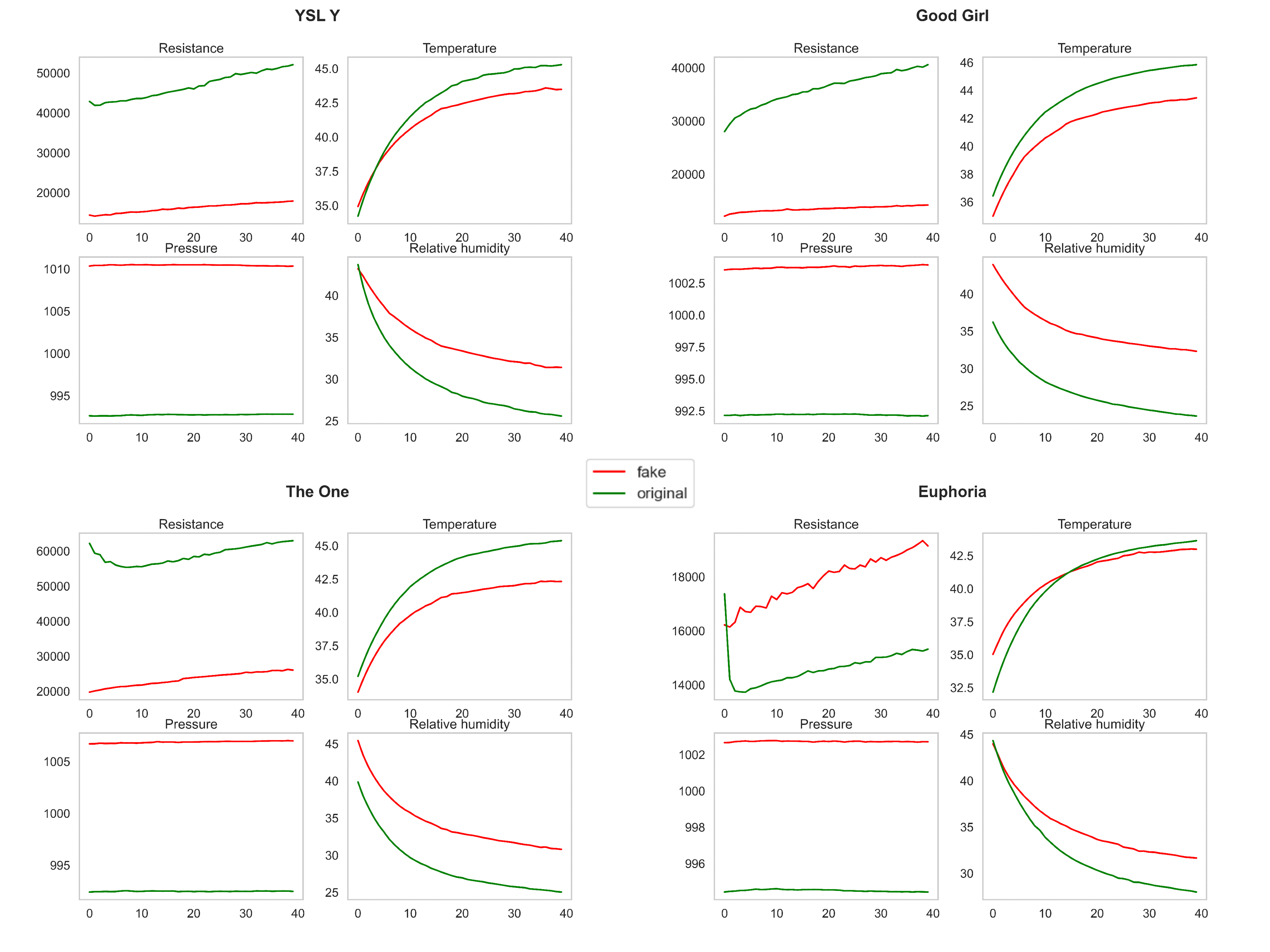
*Źródło: Opracowanie własne*

Na wykresie na Rys. 6. widzimy, że dla zapachu *Euphoria* wartość rezystancji oryginalnych perfum jest mniejsza niż dla pozostałych próbek. Jednak dla trzech pozostałych próbek, wartości rezystancji oryginałów znacznie przewyższają ich odpowiedniki.



Rys. 7. Zestawienie logarytmów średniej rezystancji dla oryginalnych zapachów oraz ich podróbek dla każdego kroku cyklu  
Źródło: Opracowanie własne

Na rysunku *Rys. 7.*, dla wszystkich zapachów oprócz *Euphoria* (*CK*)*,* obserwujemy wyższe wartości odczytów rezystancji dla oryginalnych perfum niż dla ich podróbek.



*Rys.8. Porównanie odczytów z czterech kanałów oryginałów oraz falsyfikatów dla wybranego sensora i kroku cyklu*

*Źródło: Opracowanie własne*

Na rysunku Rys. 8. widać wyraźne różnice między oryginałami perfum a ich imitacjami. *The One*, *Y* i *Good Girl* mają wyższe wartości rezystancji niż ich odpowiedniki, natomiast w przypadku zapachu *Euphoria* to imitacja ma wyższe wartości rezystancji gazowej. Odczyty ciśnienia różnią się znacząco dla każdej pary perfum, podczas gdy temperatura i wilgotność pozostają na podobnym poziomie dla wszystkich próbek.

# Zastosowane metody

Ten rozdział poświęcony jest opisowi metod analizy danych wykorzystanych w badaniu. Przedstawiamy w nim każdą z wykorzystanych metod, omawiając powód ich zastosowania oraz opisujemy ogólną zasadę działania wraz   
z przedstawieniem szczegółowych wartości hiperparametrów użytych modeli.

# Analiza skupień

W celu odkrycia wzorców i podobieństw w zbadanych próbkach zapachowych, sięgnęliśmy po algorytmy analizy skupień. Pomimo że analiza skupień jest jedną z metod nienadzorowanego uczenia maszynowego, w naszym przypadku, gdzie posiadaliśmy oznaczone próbki, mogliśmy porównać wyniki   
z rzeczywistymi etykietami.

Ze względu na sekwencyjny charakter danych, w naszym badaniu nie mogliśmy wykorzystać klasycznych metod klastrowania – rekordy traktowane są przez nie jak punkty w wielowymiarowej przestrzeni, a nie jako strumień danych w czasie. Wobec tego konieczne było dostosowanie algorytmów klastrowania do sekwencyjnego układu danych, traktując poszczególne próbki jako szeregi czasowe. Dzięki temu możliwe było grupowanie podobnych przebiegów, uwzględniając ich kształt, amplitudę i zmienność.

Grupowaniu poddany został sygnał z kanału rejestrującego rezystancję. Wykorzystaliśmy do tego adaptację algorytmu *k*-średnich dla szeregów czasowych [8]. Zastosowanie tego algorytmu umożliwiło uwzględnienie dynamicznego charakteru sygnału i pogrupowanie obiektów ze względu na przebieg rezystancji w czasie.

W kontekście oceny podobieństwa między próbkami szeregów czasowych wykorzystujemy metrykę *Dynamic Time Warping* *(DTW)* [10], polegającą na znalezieniu optymalnej ścieżki dopasowania pomiędzy dwoma szeregami, minimalizującej sumaryczną odległość między odpowiadającymi sobie punktami czasowymi. *DTW* dane wzorem (2) jest szczególnie użyteczne w sytuacjach, kiedy analizowane sygnały mogą różnić się szybkością lub przesunięciami czasowymi. Dzięki temu algorytmowi możliwe jest elastyczne dopasowanie sygnałów, które są podobne w kształcie, ale występują w różnych momentach czasowych.

Rys. 9. Porównanie dwóch sposobów obliczania odległości szeregów czasowych  
Źródło: R. Tavenard, An introduction to Dynamic Time Warping, https://rtavenar.github.io/blog/dtw.html, dostęp dnia 31.05.2024 roku

Ze względu na sposób zbierania danych opisany w rozdziale **Metodologia***,* zdecydowaliśmy się na szesnaście skupień. Taki podział zapewnia wysoki stosunek zmienności międzygrupowej do wewnątrzgrupowej i jednocześnie pozwala na dokładną interpretację wyników.

# Klasyfikacja rodzajów perfum

Celem podjęcia problemu klasyfikacji rodzajów perfum było sprawdzenie, czy za pomocą wykorzystanego przez nas urządzenia, wraz z technikami uczenia maszynowego, możliwe jest rozróżnienie oryginałów i odpowiedników zapachów.

Do klasyfikacji wykorzystaliśmy płytką sieć neuronową, składającą się z jednej warstwy LSTM, jednej warstwy gęstej oraz warstwy wyjściowej, w skład której wchodzi 9 neuronów. Każdy z ośmiu odpowiada jednej z wyjściowych klas perfum, a dziewiąty reprezentuje próbkę czystego powietrza.

*Rys. 10. Architektura zbudowanego klasyfikatora  
Źródło: Opracowanie własne*

Wejście do modelu stanowiły zestandaryzowane sygnały pochodzące ze wszystkich kanałów pomiarowych – tensor rozmiaru (4, 101).

Dane podzielone zostały na trzy podzbiory: treningowy, walidacyjny oraz testowy. Zbiór walidacyjny posłużył monitorowaniu, czy nie wystąpiło zjawisko przeuczenia, a zbiór testowy posłużył finalnej ewaluacji dopasowania modelu.

Aby zwiększyć liczbę danych treningowych, tym samym poprawiając zdolność modelu do generalizacji, zastosowaliśmy technikę zaszumienia próbek treningowych poprzez dodanie do nich szumu gaussowskiego [9] o średniej wartości 0 oraz odchyleniu standardowym równym ⅛ odchylenia standardowego oryginalnych danych. W wyniku tego otrzymano 1426 obserwacji treningowych oraz po 204 próbki ze zbioru walidacyjnego oraz testowego.

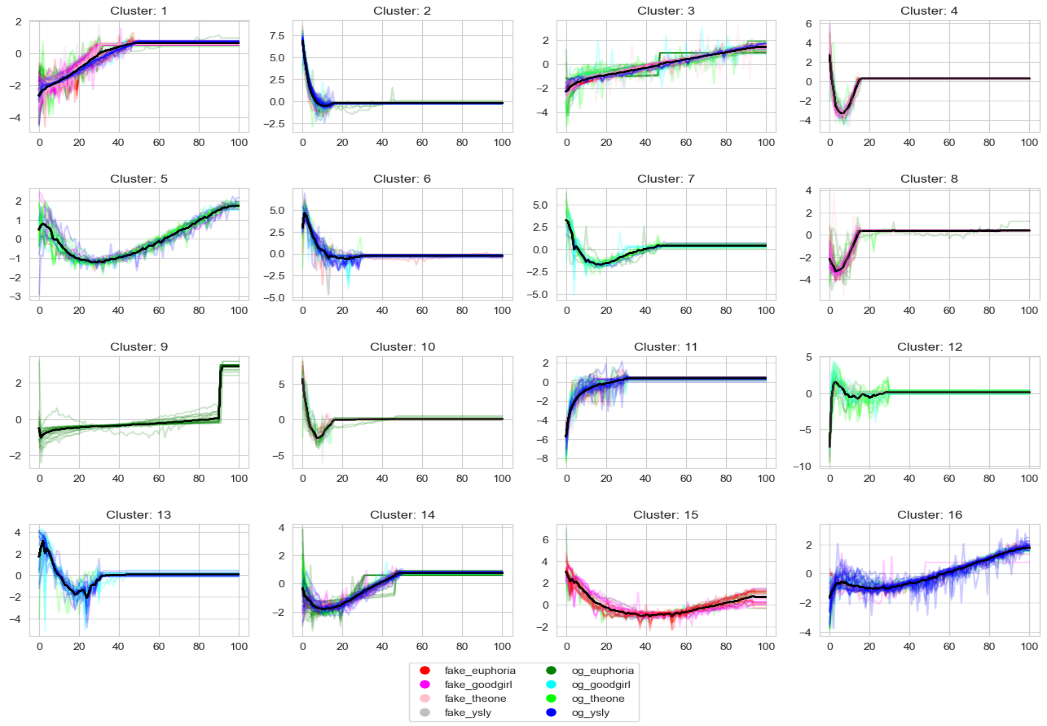
Proces uczenia modelu trwał przez ponad 150 epok, gdzie każda epoka obejmowała cykl aktualizacji wewnętrznych parametrów modelu za pomocą algorytmu AdamW [11] poprzez optymalizację funkcji straty kategorycznej entropii krzyżowej (ang. *Category Cross Entropy Loss Function*) [12].

# Wyniki oraz wnioski

Ta część pracy stanowi podsumowanie przeprowadzonego badania. Opisane zostały wyniki klastrowania oraz efekty zastosowania modelu klasyfikacyjnego.

# Wyniki klastrowania

RysunekRys. 11. przedstawia wyniki grupowania sygnałów rezystancji zebranych próbek ze względu na kształt przebiegu sygnału w czasie.

Zauważyć można, że uzyskane skupienia nie są całkowicie jednorodne ze względu na dany zapach. Oznacza to, że kształty sygnałów dla różnych perfum nie wykazują pełnej unikalności każdego z badanych wariantów. Wyjątek stanowią tutaj oryginały zapachów *Euphoria* (*CK*) oraz *The One* (*D&G*)–odpowiednio klastry 9. i 12. Pomimo to, wyniki wskazują że zapachy mają tendencję do grupowania się na jednorodne grupy w podziale na *fake*-*original*, co zaobserwować można na przykładzie skupień o indeksach 2, 5, 7, 13 oraz 15.

Rys. 11. Znalezione skupienia wraz z przypisanymi do nich obiektami  
Źródło: Opracowanie własne

# Wyniki klasyfikacji

Wytrenowany model klasyfikacyjny cechuje się wysoką ogólną skutecznością w rozróżnianiu badanych próbek zapachowych. Świadczy o tym wysoka wartość *accuracy*, przedstawiona w tabeli *Tabela 3.*

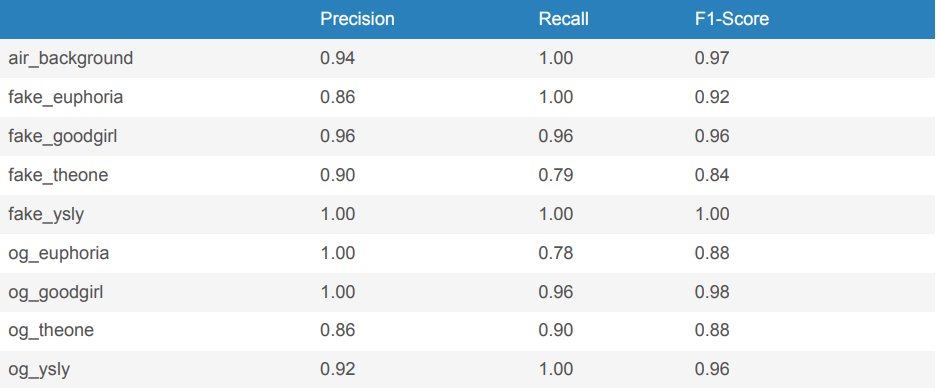
Tabela 3. Wartości ogólnych miar dopasowania na zbiorze testowym



Źródło: Opracowanie własne

Z metryk zestawionych w tabeli *Tabela 4.* wynika, że najniższe wartości *recall* dotyczą odpowiednika zapachu *The One* (*D&G*)oraz oryginału *Euphoria* (*CK*)*.* To oznacza, że model nie jest w stanie rozpoznać tych klas dla wszystkich przypadków, choć jednocześnie wysokie *precision* świadczy o dobrej jakości predykcji modelu.

Tabela 4. Metryki ewaluacyjne dla każdej z rozpatrywanych klasy na zbiorze testowym



Źródło: Opracowanie własne

Z macierzy konfuzji przedstawionej na rysunku *Rys. 12.* odczytać można, że błędne klasyfikacje prawdziwego zapachu jako podróbki to pojedyncze przypadki.

# 

Rys. 12. Macierz konfuzji dla klasyfikacji rodzajów perfum oraz ich podróbek dla zbioru testowego  
Źródło: Opracowanie własne

# Literatura

[1] L. Dang et al., *A novel classifier ensemble for recognition of multiple indoor air contaminants by an electronic nose*, Sensors Actuators A Phys., vol. 207, pp. 67–74, Mar. 2014.

[2] N. Bhattacharyya et al., *Electronic nose for black tea classification and correlation of measurements with ‘tea taster’ marks*, IEEE Trans. Instrum. Meas., vol. 57, no. 7, pp. 1313–1321, Jul. 2008.

[3] C. Di Natale et al., *Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors*, Biosensors Bioelectron., vol. 18, no. 10, pp. 1209–1218, 2003.

[4] [https://www.adafruit.com/category/946 dostęp dnia 15.04.2024](https://www.adafruit.com/category/946%20dostęp%20dnia%2015.04.2024) roku

[5] BME 688 DEV-KIT: BME688 Development kit at reichelt elektronik, <https://www.reichelt.com/pl/en/bme688-development-kit-bme-688-dev-kit-p310709.html>,   
dostęp dnia 15.04.2024 roku

[6] <https://www.bosch-sensortec.com/products/environmental-sensors/gas-sensors/bme688/> dostęp dnia 15.04.2024 roku

[7] <https://www.bosch-sensortec.com/media/boschsensortec/downloads/datasheets/bst-bme688-ds000.pdf> dostęp dnia 15.04.2024 roku

[8] Tavenard R., Faouzi J., Vandewiele G., Divo F., Androz G., Holtz C., Payne M., Yurchak R., Rußwurm M., Kolar K., Woods E., *Tslearn, A Machine Learning Toolkit for Time Series Data*. www.jmlr.org/papers/v21/20-091.html, 2020.

[9] Benegui C., Ionescu R.T., *To Augment or Not to Augment? Data Augmentation in User Identification Based on Motion Sensors*, Springer, 2020

[10] R. Tavenard, *An introduction to Dynamic Time Warping*, <https://rtavenar.github.io/blog/dtw.html,> dostęp dnia 31.05.2024 roku

[11] DS & AI Solutions, *Understanding Adam and AdamW*, <https://www.linkedin.com/pulse/understanding-adam-adamw-dsaisolutions-ileof,> dostęp dnia 31.05.2024 roku

[12] Gómez R., *Understanding Categorical Cross-Entropy Loss, Binary Cross-Entropy Loss, Softmax Loss, Logistic Loss, Focal Loss and all those confusing names*, <https://gombru.github.io/2018/05/23/cross_entropy_loss/,> 2018