# 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, składają się z mieszaniny związków zapachowych, środków homogenizujących i wzmacniających oraz rozpuszczalnika. W zależności od ilości rozpuszczalnika rozróżnia się perfumy właściwe (ekstrakty perfum), wody perfumowane i wody toaletowe. Mieszanina związków zapachowych oraz środków wzmacniających jest nazywana często kompozycją zapachową, która decyduje o specyficznym zapachu. Kompozycje te są mieszaninami składników, które dzieli się na akordy: bazowy, średni (serca) oraz wysoki (głowy). Składniki akordu bazowego uwalniają się do powietrza najwolniej oraz nie pachną intensywnie. Wśród nich można wyróżnić m. in. ambrę czy piżmo. Akord średni, którego reprezentantami są składniki kwiatowe czy korzenne charakteryzują się umiarkowaną lotnością i intensywnością zapachu. Składniki akordu wysokiego, takie jak cytrusy czy zioła, wyróżnia największa lotność oraz najintensywniejszy zapach. Istnieją także kategorie zapachów, wśród których można wymienić cytrusowe, kwiatowe, orientalne czy drzewne. Szeroki wachlarz wymienionych zapachów, oraz ich zmienna intensywność, sprawiają trudność ludzkiemu zmysłowi węchu, przy próbie rozróżniania ich między sobą. Odpowiedzią na powyższy problem jest popularna w ostatnich latach płytka zwana powszechnie elektronicznym nosem.

Elektroniczny nos to zespół sensorów odczytujący pomiary lotnych związków organicznych zawartych w otoczeniu. Czujnik gazów, poprzez pomiary takich cech jak rezystancja czy wilgotność, pozwala na przybliżenie zmysłu węchu. Technologia stojąca za czujnikami gazu, takim jak BME688, w połączeniu z uczeniem maszynowym może stanowić krok ku rozwiązaniu bardzo wielu intrygujących zagadnień. Sensory te zostały wykorzystane między innymi w badaniach zanieczyszczeń powietrza[1], klasyfikacji czarnej herbaty [2], czy analizie oddechu pacjentów w celu predykcji nowotworów płuc [3].

W niniejszej pracy przedstawiono zastosowanie technik oraz metod analizy danych dla pomiarów zapachów perfum zebranych za pomocą elektronicznego nosa BME688. Praca koncentruje się na eksploracyjnej analizie danych, klastrowaniu oraz klasyfikacji próbek zapachowych perfum oryginalnych oraz ich falsyfikatów. 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, dostrzeżenie, co przyczynia się do tego, jak ludzki zmysł węchu odczuwa dany zapach oraz, w konsekwencji, na zbudowanie modeli klasyfikacyjnych, skutecznie odpowiadających na zadane problemy.

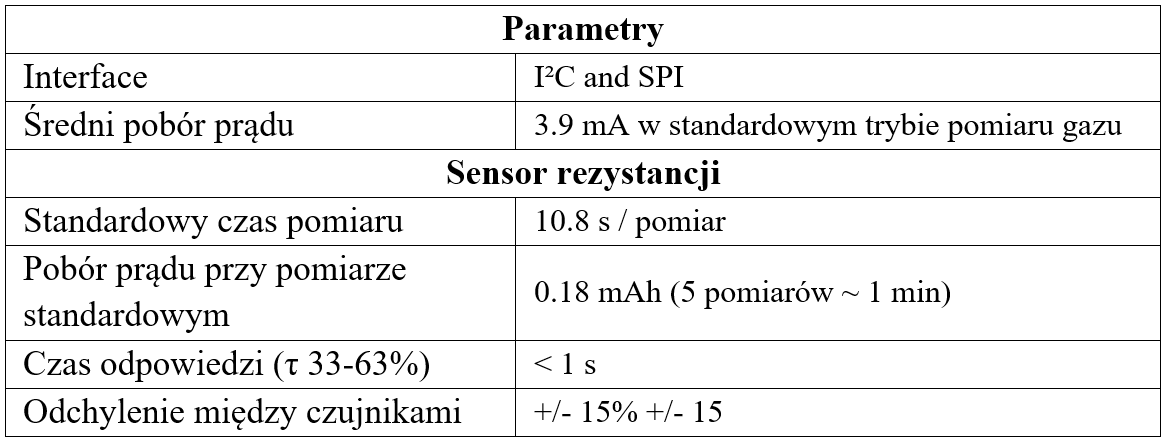
Praca składa się z czterech głównych części. Pierwsza przedstawia narzędzie służące przedstawionemu badaniu oraz metodologię zbierania danych. Druga dotyczy zebranych danych. Przeprowadzono tam analizę eksploracyjną oraz opisano preparację danych do modelu predykcyjnego. W części trzeciej przedstawiono zastosowane przez nas metody. Ostatnia część zawiera rezultaty oraz wnioski płynące z przeprowadzonego badania.

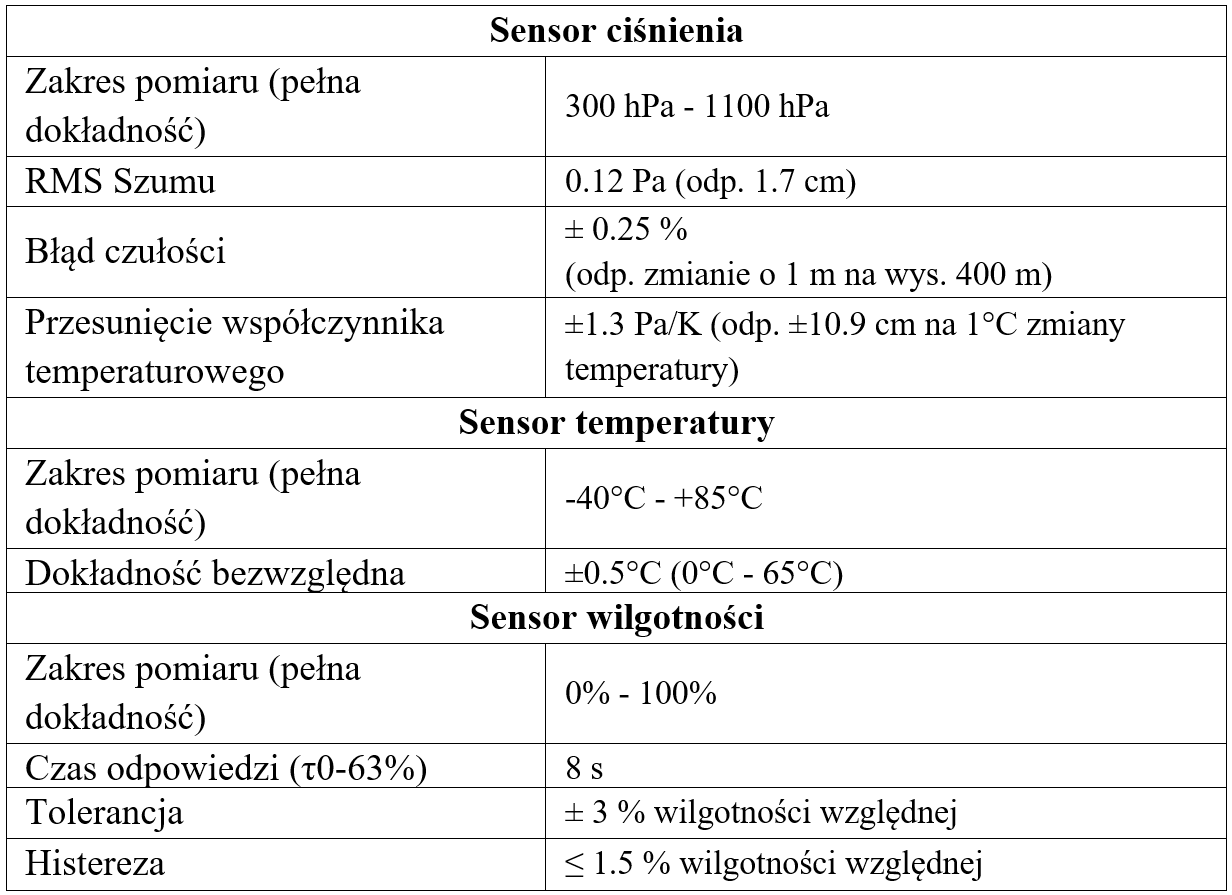
# Sprzęt oraz metodologia zbierania danych

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.

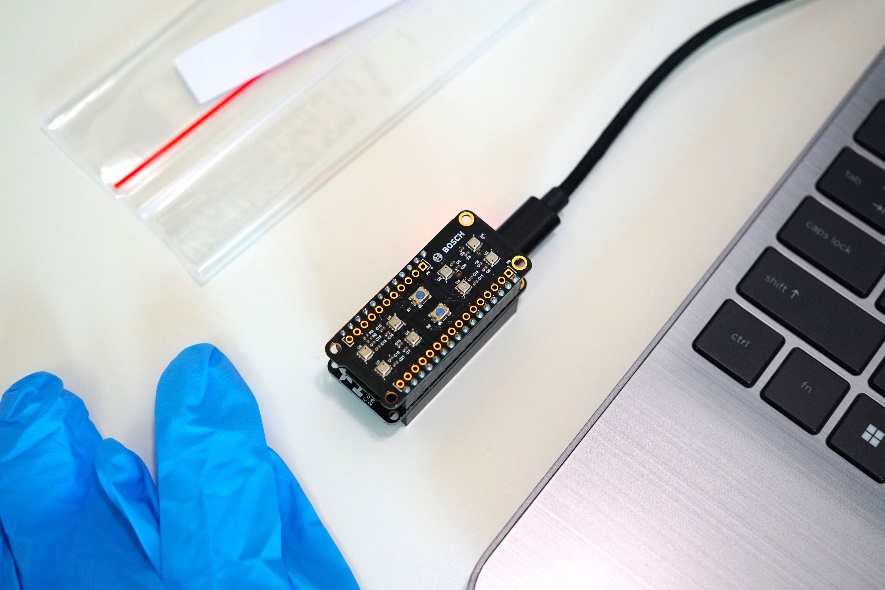
# Metodologia

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

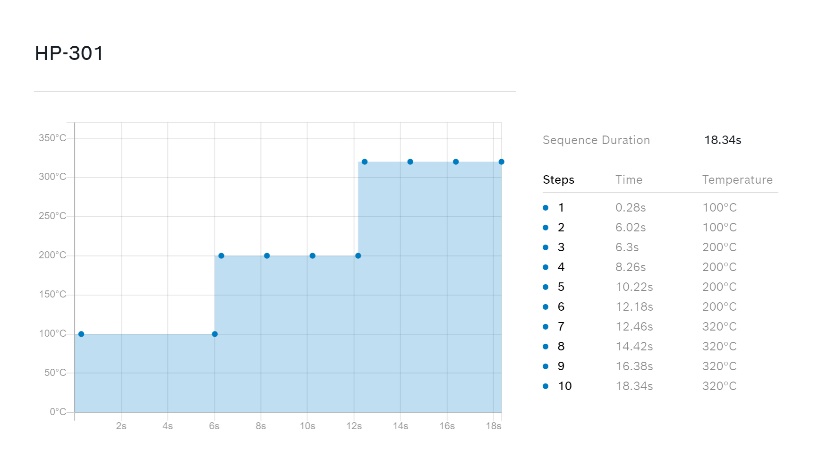
# Program

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

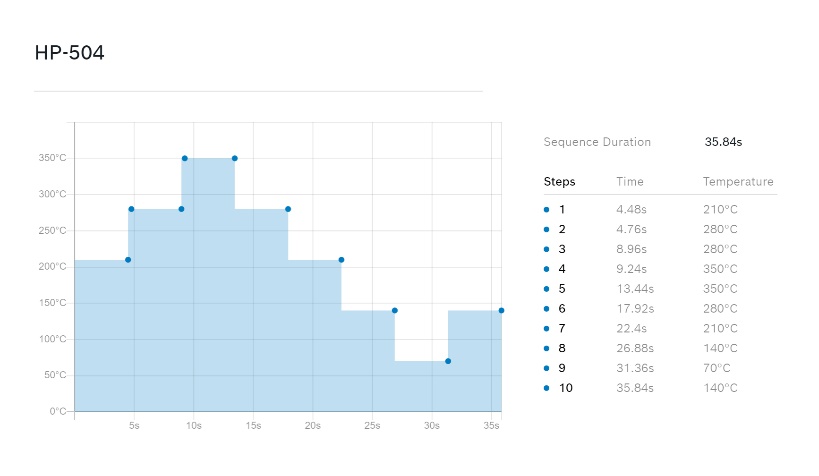
Ustalając szczegółowy program, szukaliśmy złotego środka pomiędzy różnorodnością i szerokością zakresu pomiarowego, a powtarzalnością   
i niezawodnością. Połączenie tych cech zapewniła kombinacja dwóch   
profili: HP-301 oraz HP-504, rozłożonych równomiernie na 4 pary sensorów.

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

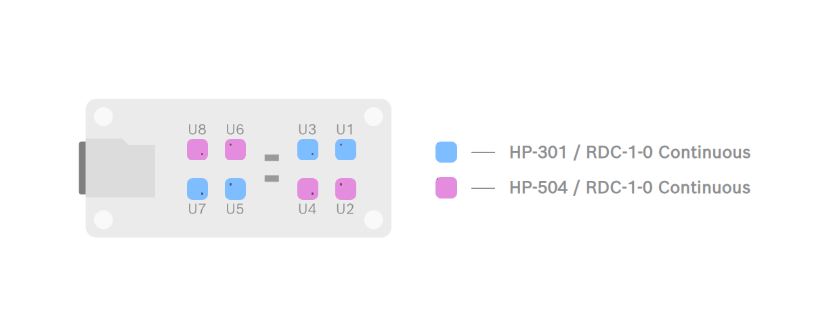
Konfiguracja sensorów oraz przebiegi wybranych do badania profili termicznych, prezentują pochodzące z programu BME AI-Studio grafiki zamieszczone poniżej.



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

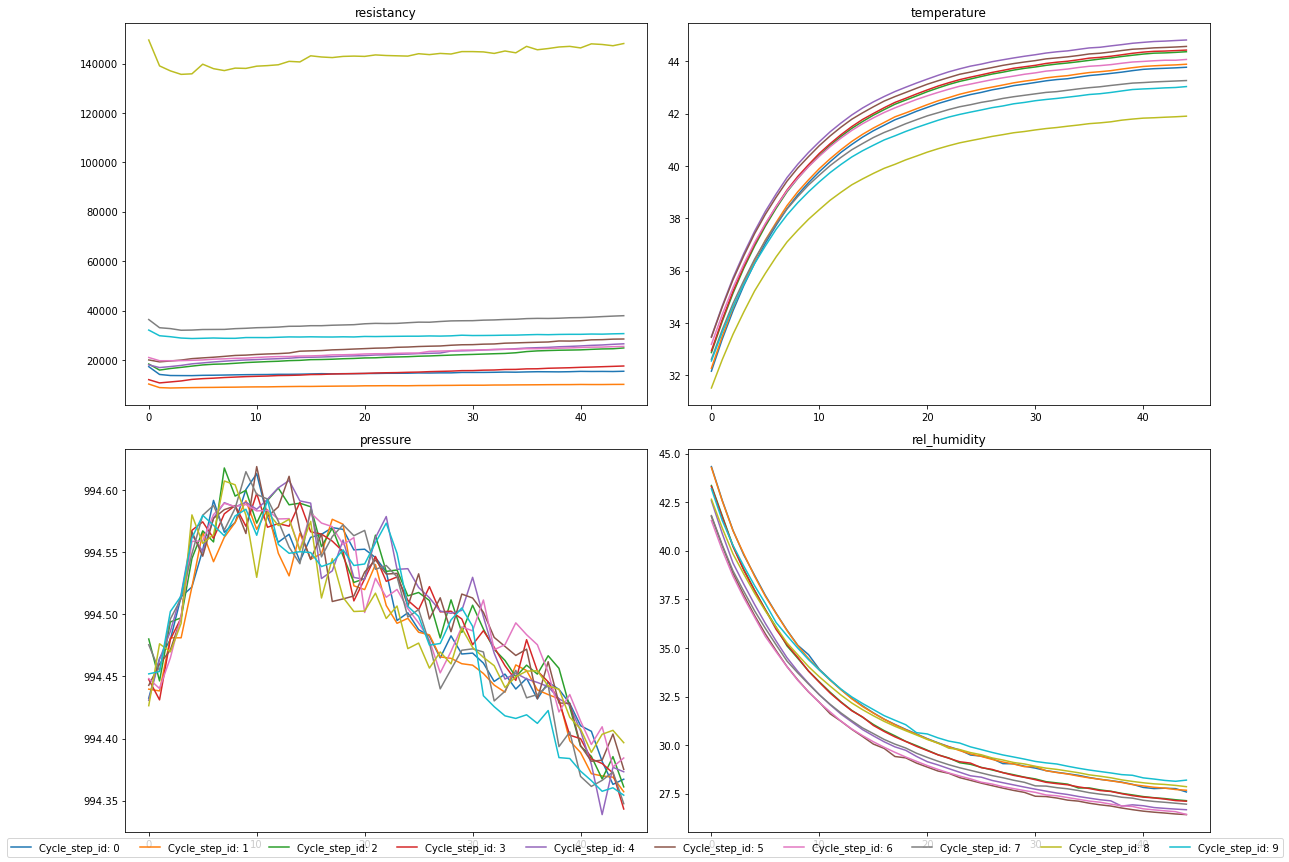
W tej części pracy zostaną przedstawione zebrane dane, zaprezentowana zostanie eksploracyjna analiza danych oraz porównanie zebranych danych w kontekście postawionego przedmiotu badania.

# 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).

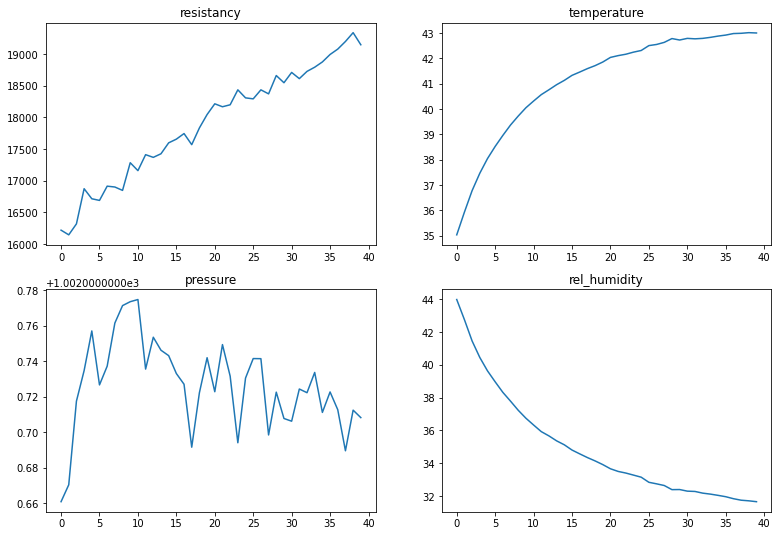
Jak wspomniano w poprzedniej części, dane dotyczące perfum zostały zebrane za pomocą 8 sensorów BME688, przy wykorzystaniu dwóch profili temperaturowych, w ciągłym cyklu pracy urządzenia. Elektroniczny nos zbiera dane dla 4 kanałów: rezystancji (w Ohmach), temperatury (w °C), ciśnienia (w hPa) oraz wilgotności (w %).

Jako przykład opracowania zebranych danych posłużą falsyfikaty perfum *Euphoria*. Na rysunku poniżej przedstawiono odczyty elektronicznego nosa ze wszystkich 4 kanałów w podziale na dziesięć kroków pomiarowych każdy.



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

Poniżej przedstawiono przebiegi dla poszczególnych kanałów falsyfikatów perfum *Euphoria*.



*Rys. 6. Przebieg badanych zmiennych dla falsyfikatów perfum Euphoria*

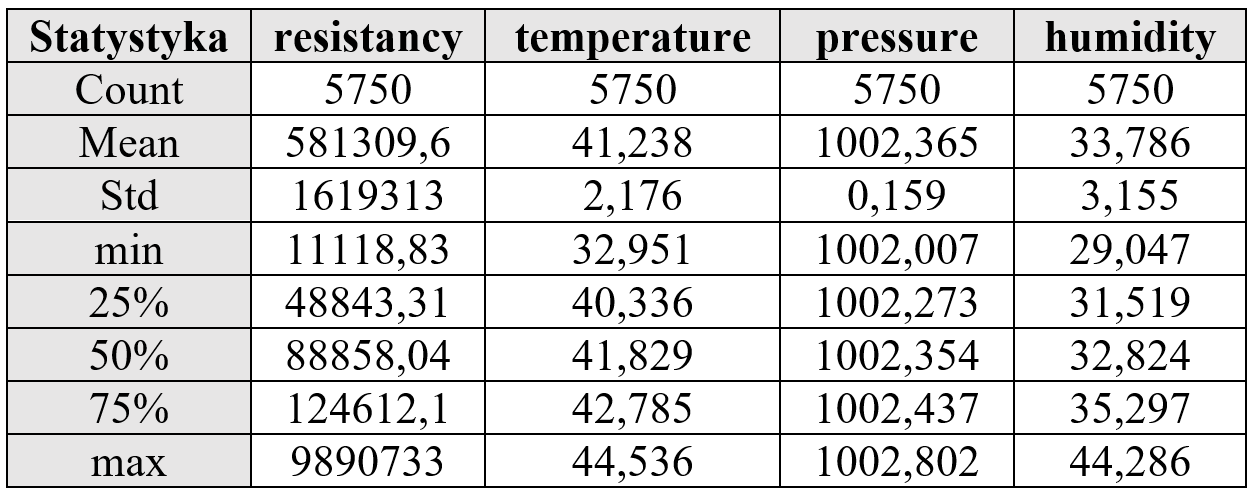
*Źródło: Opracowanie własne*

Na rysunku Rys. 10 widzimy tendencję badanych zmiennych - zwiększa się temperatura otoczenia pod wpływem pracy układu. Zmniejsza się wilgotność powietrza. Wraz z upływem czasu zbierania próbek zwiększają się wartości rezystancji.

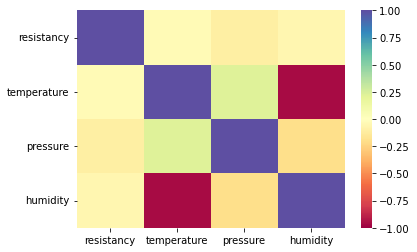
# Eksploracyjna analiza danych

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

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

*  
Źródło: Opracowanie własne.*

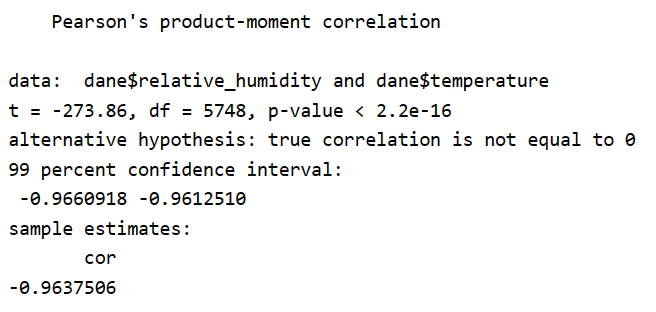
Na podstawie powyższych statystyk opisowych czterech kanałów zebranych przez elektroniczny nos, dostrzec można, że układ scalony nagrzewał się podczas badania - od początkowej temperatury 32°C przeszedł do 44°C. Ciśnienie zmierzone w tym samym czasie pozostawało na względnie stałym poziomie. Wilgotność, podobnie jak i temperatura, ulegała zmianie podczas badania. Warto także zwrócić uwagę na minimalne oraz maksymalne wartości rezystancji – różnią się one od siebie niemal 900 razy. Zebrane dane różnią się znacząco od siebie zakresami zmienności, co skłoniło nas ku decyzji o ich późniejszej normalizacji przed użyciem ich w modelu klasyfikacyjnym.



Rys. 7. Macierz korelacji dla 4 kanałów falsyfikatów perfum Euphoria  
Źródło: Opracowanie własne

Na podstawie powyższej macierzy korelacji, dostrzec można bardzo silną ujemną korelację pomiędzy wilgotnością a temperaturą. Wynosi ona -0,96. Jest to także potwierdzenie zaobserwowanej tendencji dla tych dwóch zmiennych przy okazji przedstawienia przebiegu zmiennych na rysunku 10.

Przeprowadzono także test istotności współczynnika korelacji pomiędzy wspomnianymi zmiennymi na poziomie istotności α=0,01, którego wyniki zaprezentowano poniżej.

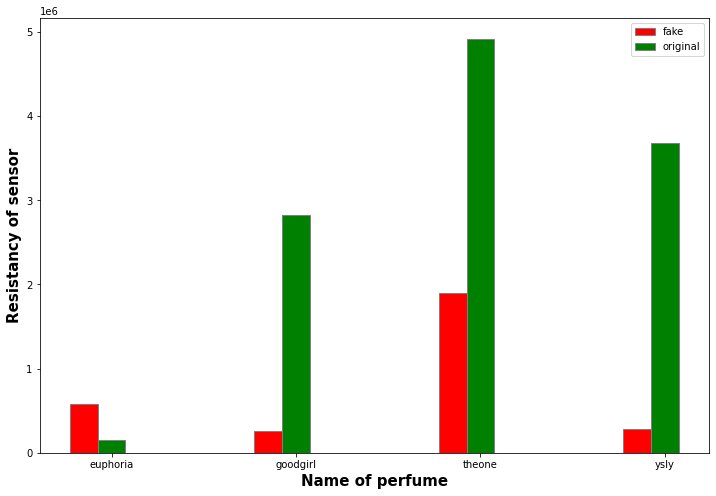


Rys. 8. Wyniki testu istotności współczynnika korelacji dla wilgotności oraz temperatury  
Źródło: Opracowanie własne

Wyniki powyższego testu wskazują na istotność badanego współczynnika korelacji.

Wspomniana, przy okazji statystyk opisowych, potrzeba normalizacji badanych danych została zastosowana w celu umożliwienia ich wzajemnego porównywania. Przebiegło to… ***(dokończyć)***

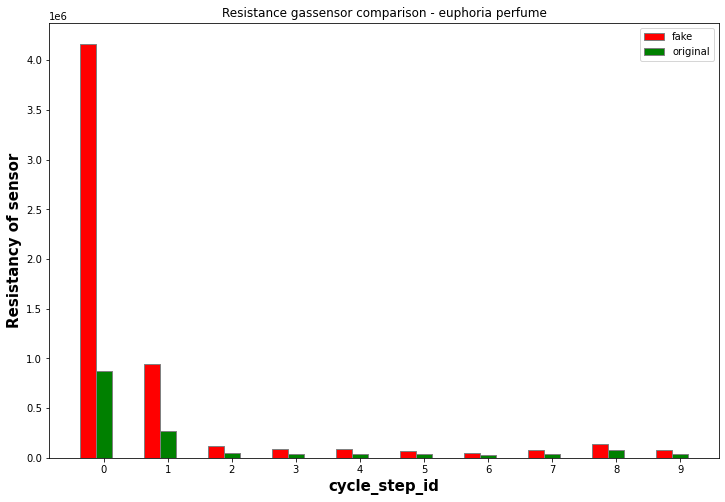
# Porównanie oryginałów oraz falsyfikatów perfum



*Rys. 9. Porównanie odczytów rezystancji dla oryginalnych perfum oraz ich falsyfikatów*

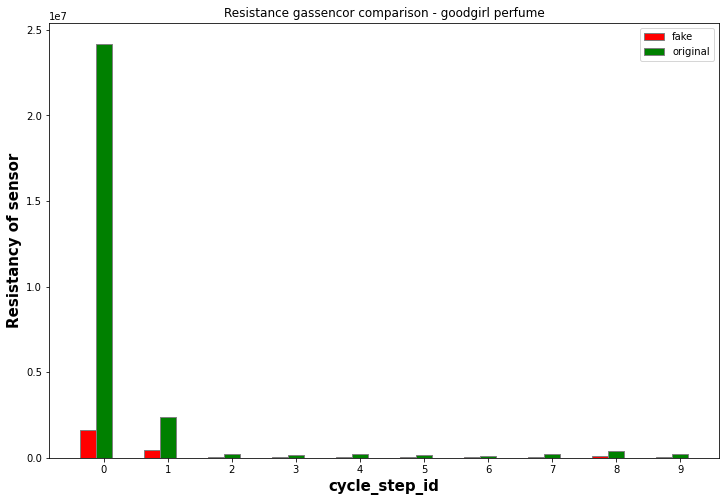
*Źródło: Opracowanie własne*

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.



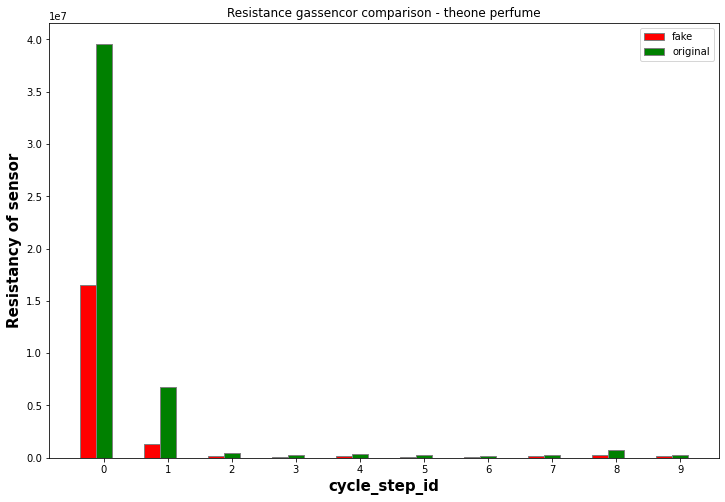
*Rys. 10. Zestawienie rezystancji oryginału oraz falsyfikatu zapachu Euphoria*

*Źródło: Opracowanie własne*



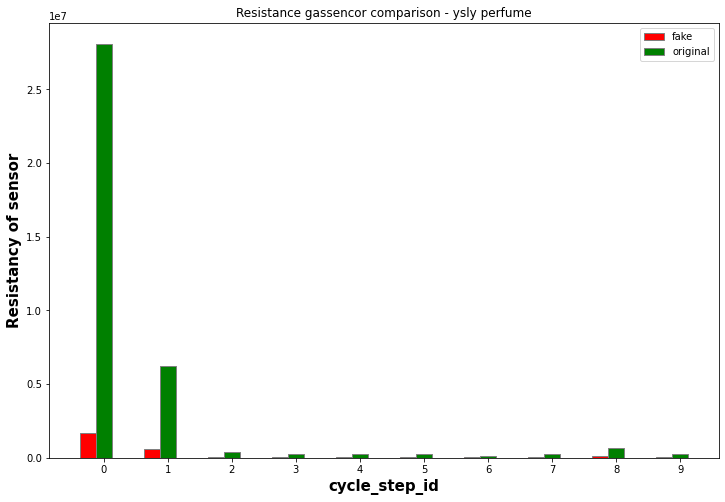
*Rys. 11. Zestawienie rezystancji oryginału oraz falsyfikatu zapachu Good Girl*

*Źródło: Opracowanie własne*



*Rys. 12. Zestawienie rezystancji oryginału oraz falsyfikatu zapachu The One*

*Źródło: Opracowanie własne*

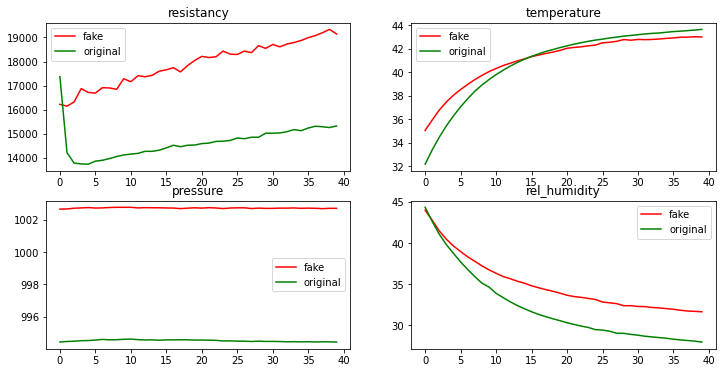


*Rys. 13. Zestawienie rezystancji oryginału oraz falsyfikatu zapachu Y*

*Źródło: Opracowanie własne*

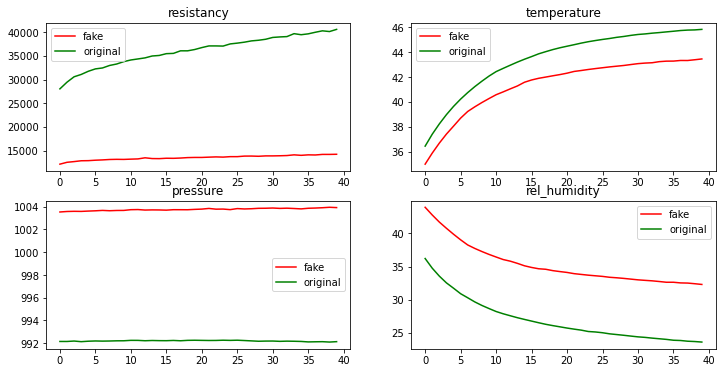
Na podstawie powyższych rysunków 10-13, zauważyć można znaczące różnice w odczytach rezystancji pomiędzy perfumami oryginalnymi a ich fałszywymi odpowiednikami.

Rysunki 14-17 prezentują przebiegi czterech par badanych perfum w podziale na cztery kanały, dla których zostały one zebrane.



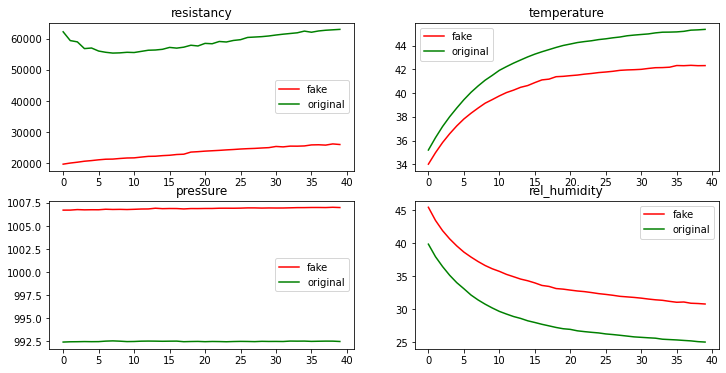
*Rys. 14. Porównanie odczytów z czterech kanałów oryginału oraz falsyfikatu zapachu Euphoria*

*Źródło: Opracowanie własne*



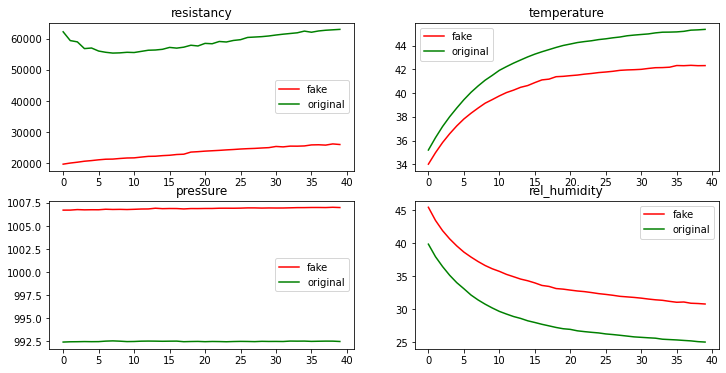
*Rys. 15. Porównanie odczytów z czterech kanałów oryginału oraz falsyfikatu zapachu Good Girl*

*Źródło: Opracowanie własne*



*Rys. 16. Porównanie odczytów z czterech kanałów oryginału oraz falsyfikatu zapachu The One*

*Źródło: Opracowanie własne*



*Rys. 17. Porównanie odczytów z czterech kanałów oryginału oraz falsyfikatu zapachu Y*

*Źródło: Opracowanie własne*

Możemy zaobserwować klarowne tendencje wykazane na powyższej serii wykresów porównujących oryginały z ich imitacjami. Pierwowzory zapachów *The One*, *Y* oraz *Good Girl* odznaczają się zauważalnie wyższymi wartościami rezystancji od swoich odpowiedników. Natomiast w przypadku zapachu *Euphoria* to imitacja przeważa nad oryginałem pod względem odczytanych wartości rezystancji gazowej. Znaczące różnice w odczytach ciśnienia dotyczą za to każdej pary perfum. Temperatura oraz wilgotność utrzymują się na zbliżonym do siebie poziomie dla każdej z zebranych próbek. ***(tu koniec poprawek MK 05.05)***

# 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 hiper-parametrów użytych modeli.

# Analiza skupień

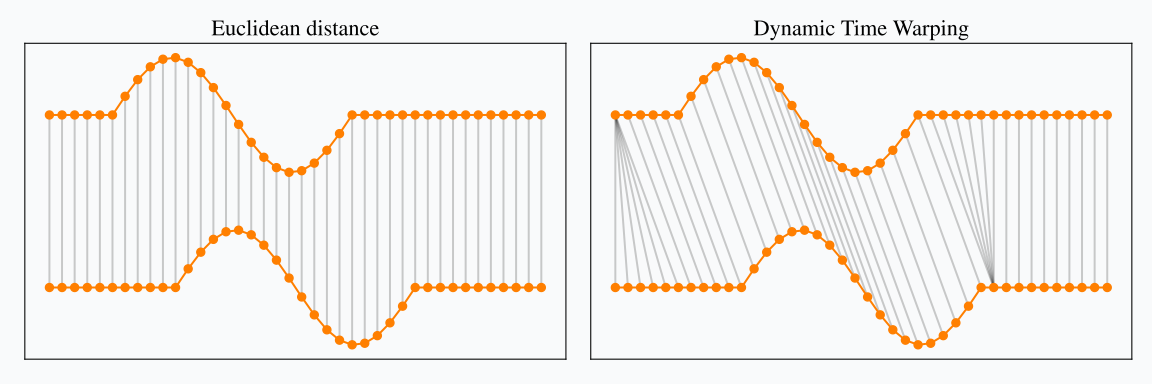
W celu odkrycia wzorców i podobieństw w zbadanych próbkach zapachowych, sięgnęliśmy po algorytmy analizy skupień. Czynnikiem uzasadniającym klastrowanie danych jest to, że choć analiza skupień należy do grupy nienadzorowanych metod uczenia maszynowego, to mając oznaczone próbki, mogliśmy porównać wyniki z rzeczywistymi etykietami.

Algorytmy analizy skupień polegają na grupowaniu najbardziej podobnych do siebie obiektów. Podział na klastry polega na grupowaniu elementów na podstawie dzielącej je odległości. Celem tego procesu jest minimalizacja różnic pomiędzy wektorami wewnątrz klastrów, oraz maksymalizacja zróżnicowania pomiędzy poszczególnymi grupami. Zależy nam na tym, by obiekty wewnątrz jednego klastra były podobne do siebie podczas gdy klastry osobne klastry wykazują znaczące różnice.

Ze względu na sekwencyjny charakter danych, w naszym badaniu nie mogliśmy wykorzystać klasycznych metod klastrowania – nie uwzględniają one w żaden sposób kolejności, a rekordy traktowane są przez nie jak punkty w wielowymiarowej przestrzeni. 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 [cytowanie tslearn]. Wykorzystanie informacji o kolejności i zależnościach czasowych między punktami danych umożliwiło uwzględnienie dynamicznego charakteru sygnału.

W kontekście oceny podobieństwa między próbkami szeregów czasowych wykorzystujemy metrykę *Dynamic Time Warping* (DTW), polegającą na znalezieniu optymalnej ścieżki dopasowania pomiędzy dwoma szeregami, minimalizującej sumaryczną odległość między odpowiadającymi sobie punktami czasowymi. DTW 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. 18.Porównanie dwóch sposobów obliczania odległości szeregów czasowych  
Źródło:

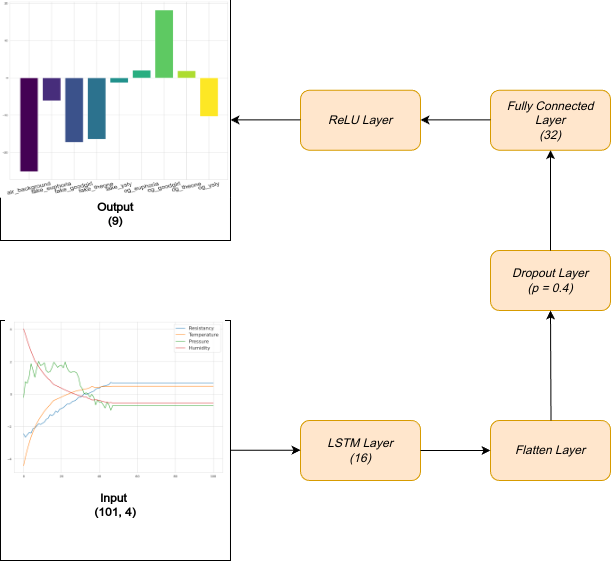
https://rtavenar.github.io/blog/dtw.html#dynamic-time-warping

[napisać dlaczego 16 skupień]

# Klasyfikacja rodzajów perfum

Celem podjęcia problemu klasyfikacji rodzajów perfum było sprawdzenie, czy za pomocą użytego przez nas urządzenia wraz z technikami uczenia maszynowego jest możliwość odróżniać różne typy perfum, włącznie z ich fałszywymi odpowiednikami.

Do klasyfikacji wykorzystaliśmy płytką sieć neuronową, składającą się z jednej warstwy LSTM, jednej warstwy gęstej oraz warstwy wyjściowej, w skład której wchodzi 9 neuronów, z których każdy odpowiada jednej z wyjściowych klas, a jeden poświęcony jest dla klasy odpowiadającym jednej z wyjściowych klas, oraz jednemu neuronowi reprezentującemu próbkę czystego powietrza czystego powietrza.



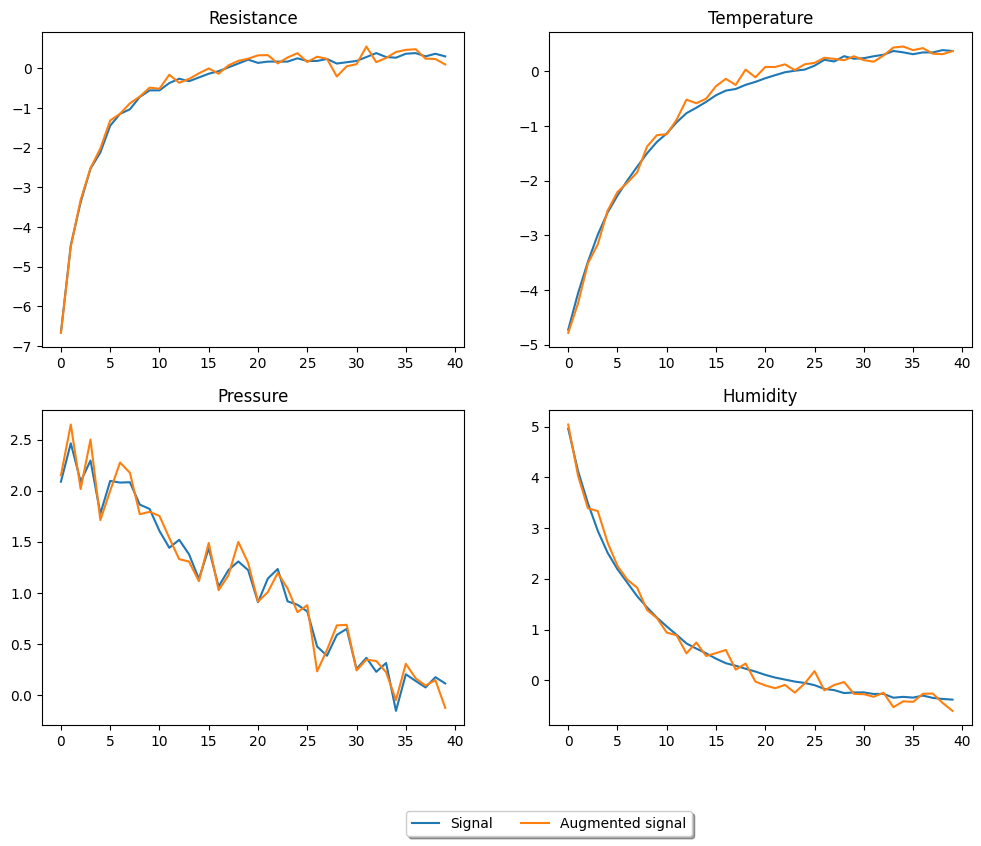
Rysunek 1 Architektura zbudowanego klasyfikatora

Źródło: Opracowanie własne

Dane wejściowe stanowią sygnały pochodzące ze wszystkich kanałów pomiarowych, które rejestrują informacje dotyczące rezystancji, temperatury, ciśnienia oraz wilgotności. To w rezultacie dało na wejściu tensor rozmiaru (4, 101) dla każdej próbki.

Dane podzielone zostały na 3 podzbiory, tj. treningowy, walidacyjny oraz testowy. Zbiór walidacyjny posłużył do monitorowania przebiegu uczenia oraz obserwacji czy nie wystąpiło zjawisko przeuczenia, a zbiór testowy posłużył finalnej ewaluacji dopasowania modelu.

Aby zwiększyć liczbę danych treningowych oraz poprawić zdolność modelu do generalizacji, zastosowaliśmy technikę zaszumienia próbek danych treningowych poprzez dodanie do nich szumu gaussowskiego [cytowanie] o średniej wartości 0 oraz odchyleniu standardowym równym 1/8 odchylenia standardowego oryginalnych danych.



Rys. 21. Porównanie danych przed oraz po zaszumieniu  
Źródło: Opracowanie własne

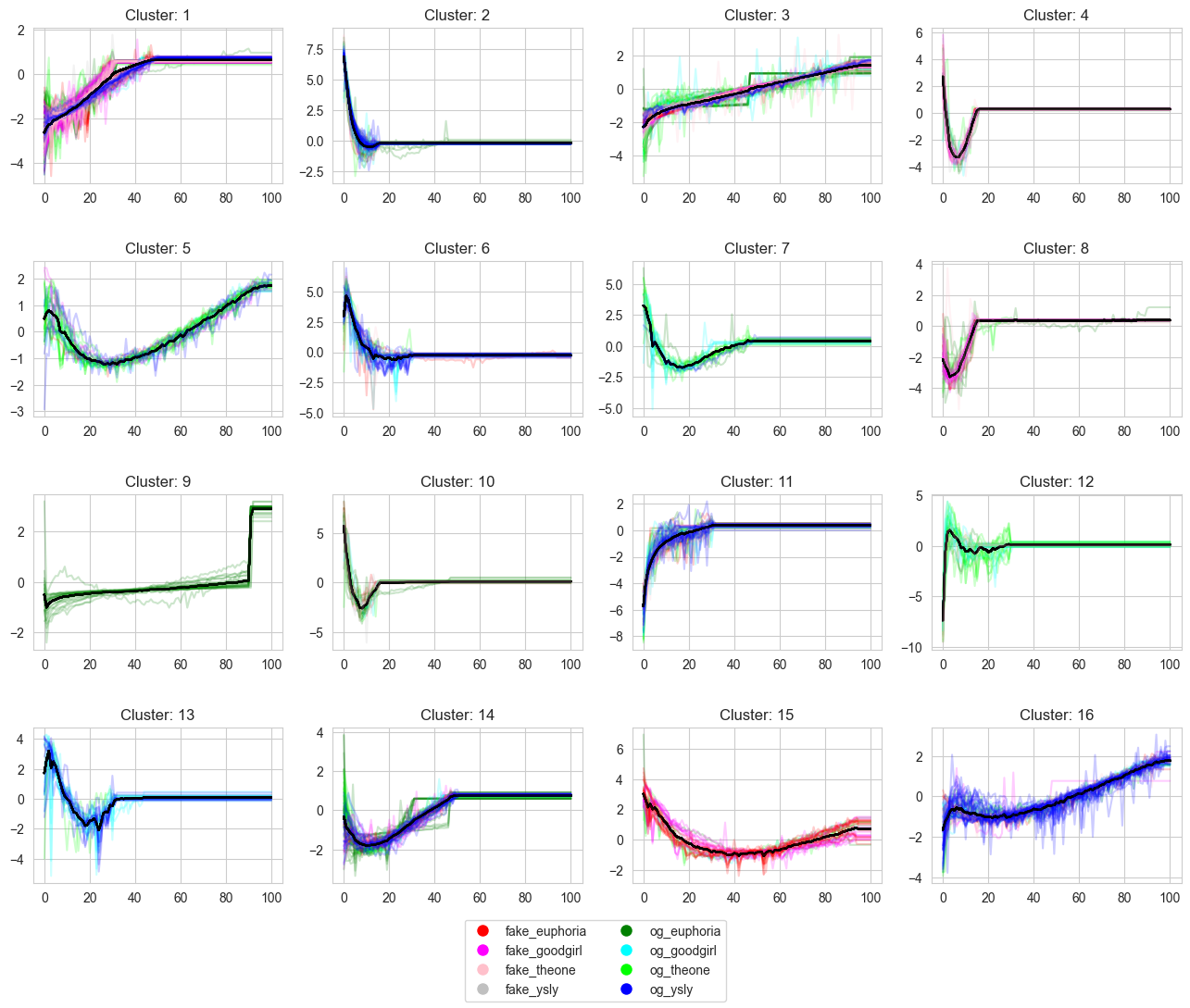
W wyniku tego otrzymano 1426 obserwacji treningowych oraz po 204 obserwacje ze zbioru walidacyjnego oraz testowego.

Proces uczenia modelu trwał przez ponad 150 epok, gdzie każda epoka obejmowała cykl aktualizacji wewnętrznych parametrów modelu poprzez optymalizację funkcji straty kategorycznej entropii krzyżowej zdefiniowanej następująco:

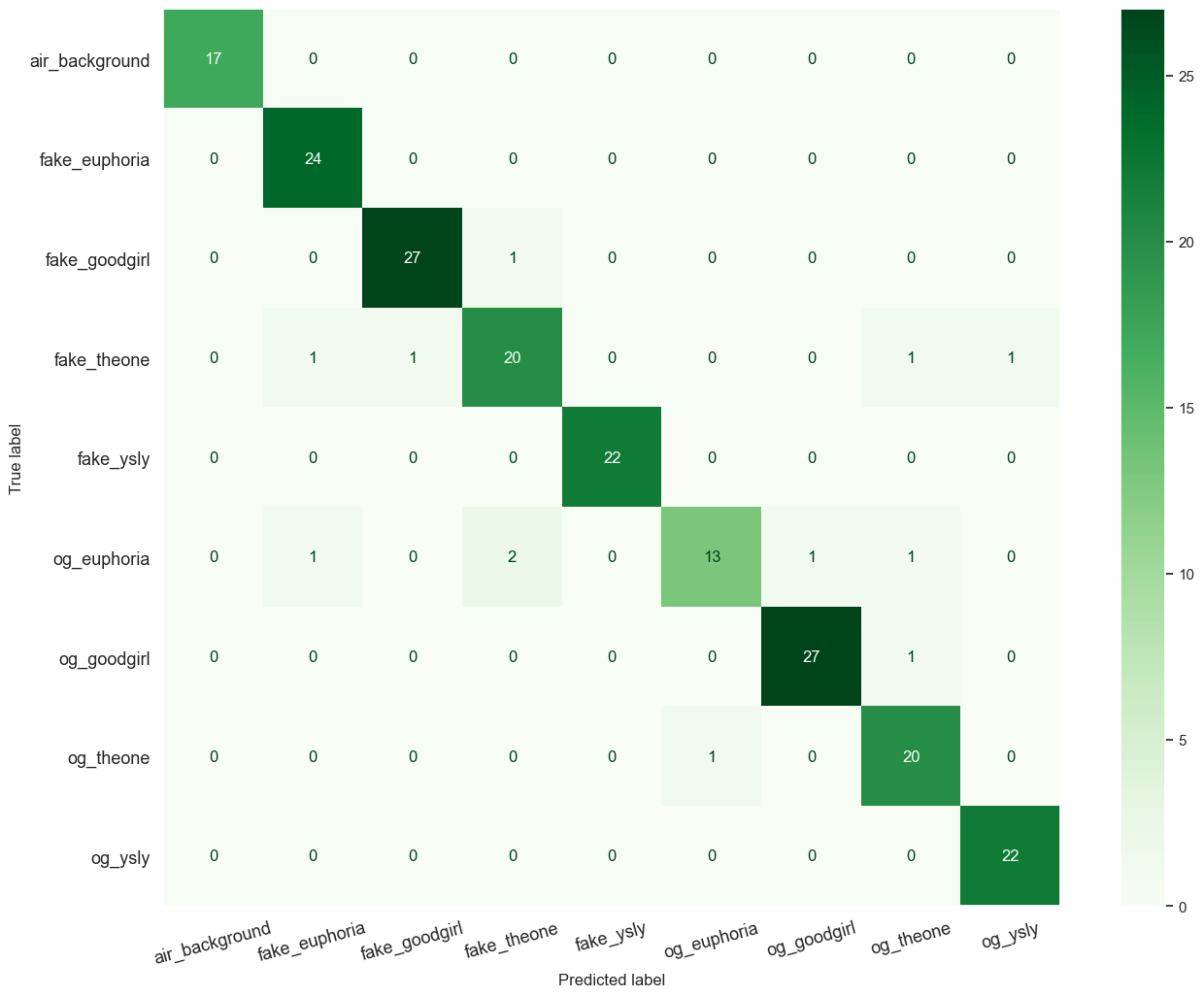
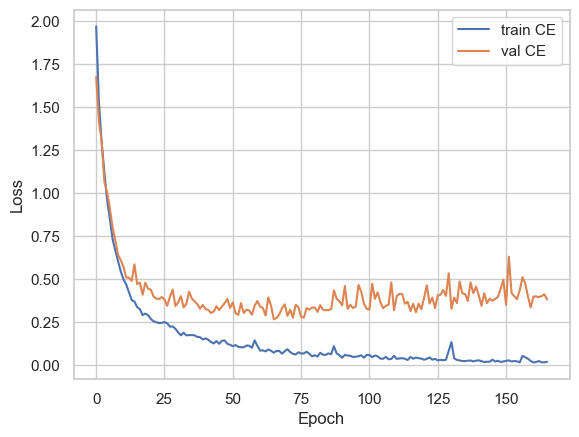
[wzór na CE z PyTorch (bo tam jest log softmax zrobiony)]

# Wyniki oraz wnioski

# Wyniki klastrowania



# Wyniki klasyfikacji



# 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][https://www.reichelt.com/pl/en/bme688-development-kit-bme-688-dev-kit-p310709.html dostęp dnia 15.04.2024](https://www.reichelt.com/pl/en/bme688-development-kit-bme-688-dev-kit-p310709.html%20dostęp%20dnia%2015.04.2024) 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](https://www.bosch-sensortec.com/media/boschsensortec/downloads/datasheets/bst-bme688-ds000.pdf%20dostęp%20dnia%2015.04.2024) roku

[8]

[9]

[10]