# NADWORSKA WERONIKA[[1]](#footnote-1), SZYSZKA PIOTR1, KOZIEŁ MAKSYMILIAN[[2]](#footnote-2), PIŁAT-ROŻEK MAGDALENA[[3]](#footnote-3)

# 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

# Wstęp

Perfumy, będące przedmiotem codziennego użytku są mieszaniną 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 wraz z oprogramowaniem pozwalającym analizować uzyskane dane. Algorytmy uczenie maszynowego stosowane w takim rozwiązaniu pozwalają rozwiązywać wiele problemów. Elektroniczne nosy znajdują zastosowanie w badaniach zanieczyszczeń powietrza [3], klasyfikacji produktów spożywczych [1] czy analizie oddechu pacjentów w celu predykcji występowania nowotworów płuc [4].

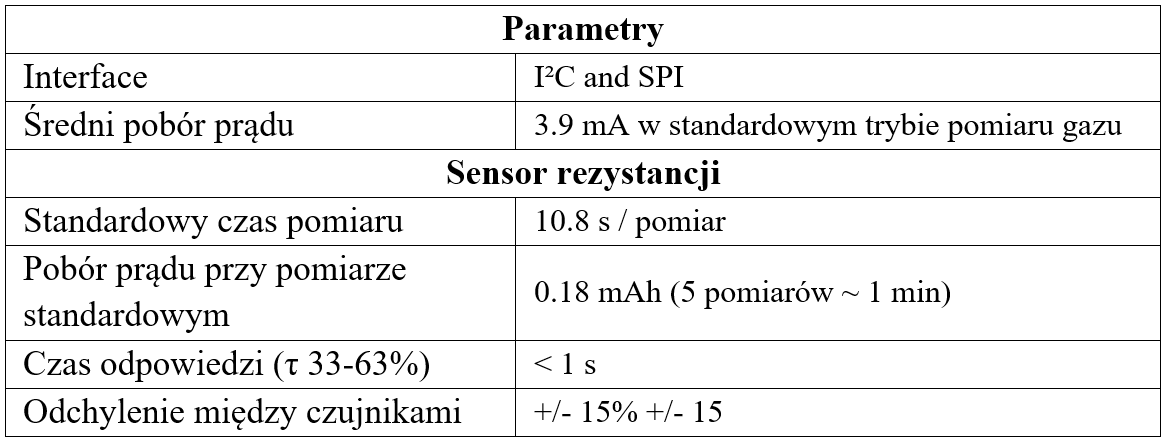
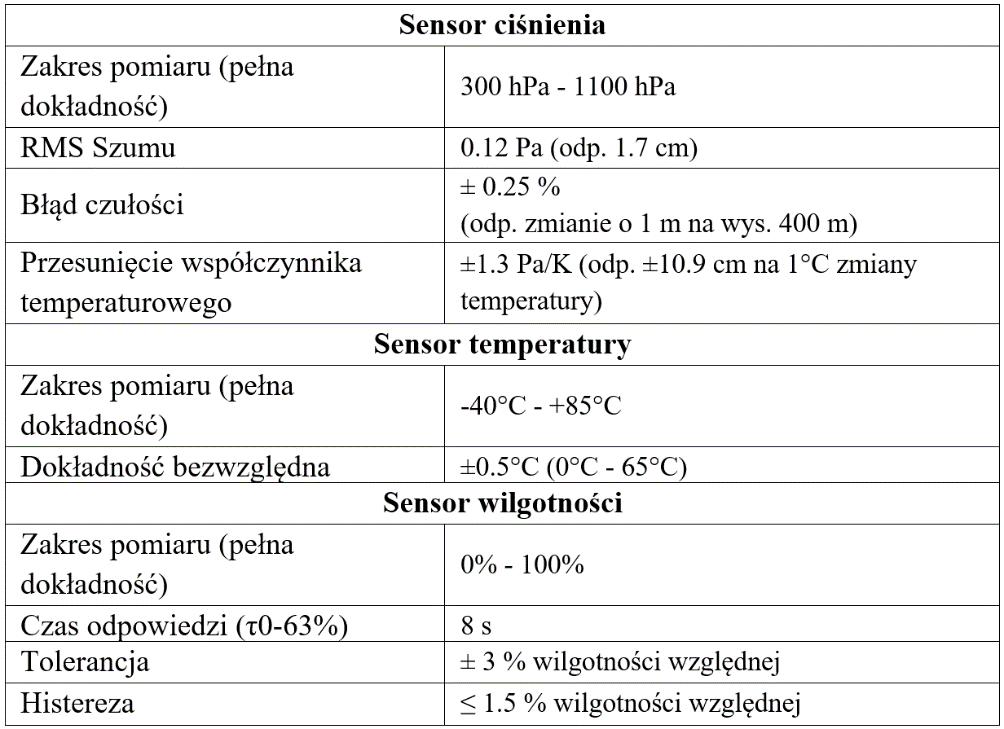
W pracy przedstawiono zastosowanie technik statystycznych oraz uczenia maszynowego w zadaniu eksploracji oraz klasyfikacji danych pochodzących   
z czujników elektronicznego nosa. Takie postępowanie pozwoliło na odkrycie interesujących zależności stojących za zapachami, a w konsekwencji, zbudowanie modeli klasyfikacyjnych, skutecznie odpowiadających na zadane problemy.

Praca składa się z czterech części. W pierwszej opisano narzędzie   
i metodologię zbierania danych. Druga część zawiera analizę eksploracyjną   
i przygotowanie danych do modelu predykcyjnego. W trzeciej części omówione zostały zastosowane metody, zaś w ostatniej części zaprezentowano wyniki   
i wnioski z badania.

# Sprzęt

Na potrzeby naszych badań wykorzystano zespół Evaluation Kit Board BME688 [8], będący połączeniem płytki deweloperskiej Adafruit HUZZAH32 [11] z wbudowanym modułem Wi-Fi ESP32, oraz płytki deweloperskiej BME688 zawierającej osiem sensorów BOSCH BME688 [9]. 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. specyfikacja Evaluation Kit Board BME688



Źródło: Opracowanie własne na podstawie specyfikacji producenta [9]

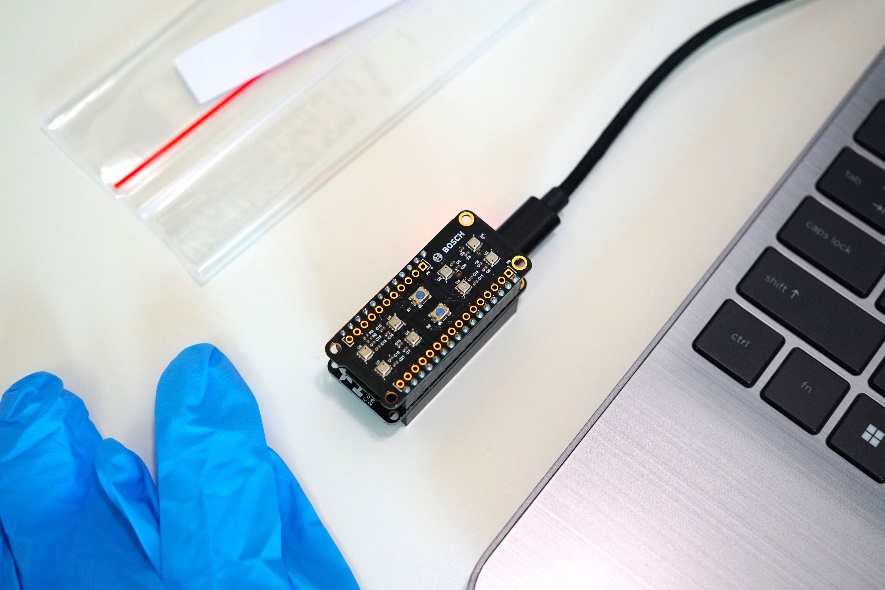
# Metodologia

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, postanowiono aplikować zapachy na paski z gęstego, niezadrukowanego papieru, pochodzącego z tego samego arkusza 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 zadbano, by pomiary próbek oryginału i imitacji danego zapachu zawsze były oddzielone pomiarem wody perfumowanej innego producenta. W ten sposób ograniczono 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 momencie dokonywania pomiarów 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

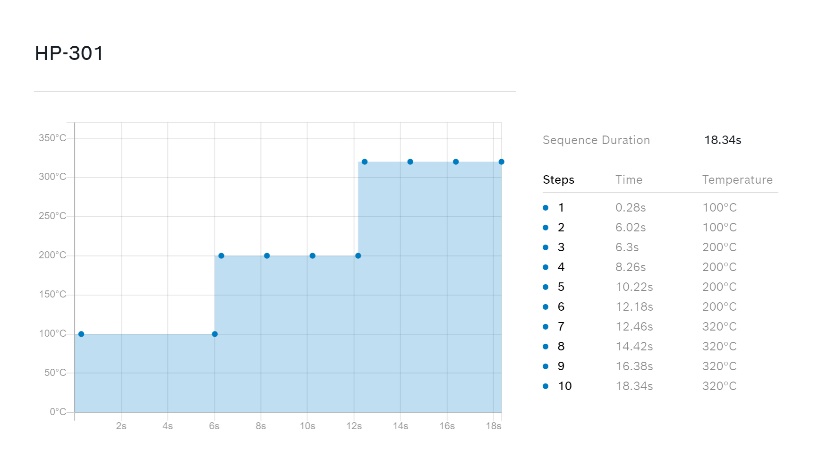
# Program

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

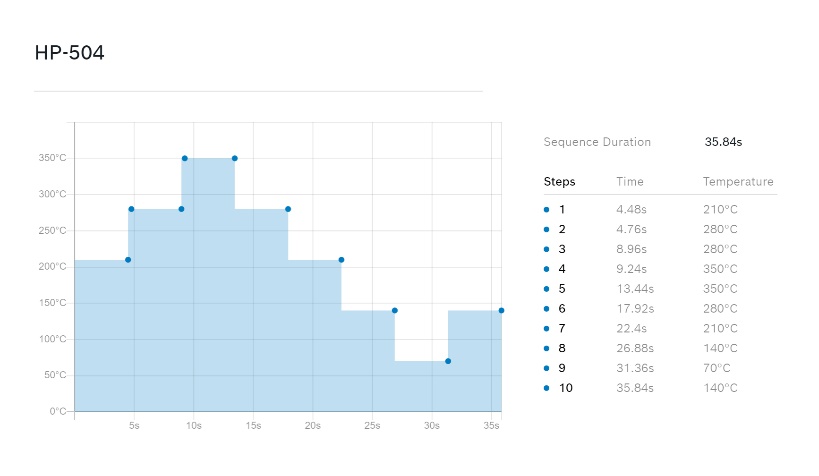
Ustalając szczegółowy program, poszukiwano kompromisu pomiędzy różnorodnością i szerokością zakresu pomiarowego, a powtarzalnością   
i niezawodnością. Połączenie tych cech zapewniła kombinacja dwóch   
profili: HP-301 oraz HP-504, rozłożonych równomiernie na 4 pary sensorów.

Decydując o cyklu pracy układu wybrano 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 na Rysunkach 2-4.



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



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

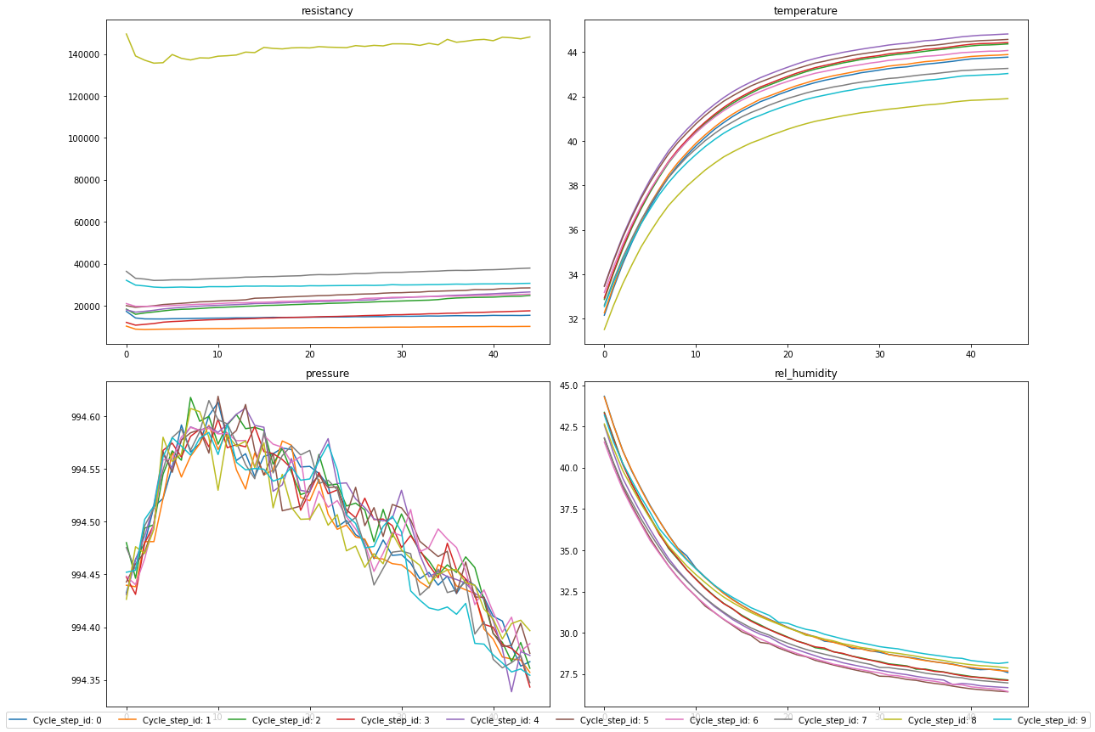


Rys. 4. Układ sensorów BME688  
Źródło: Program BME AI-Studio

# Charakterystyka zebranych danych

W badaniu uwzględniono cztery 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 Omach), 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

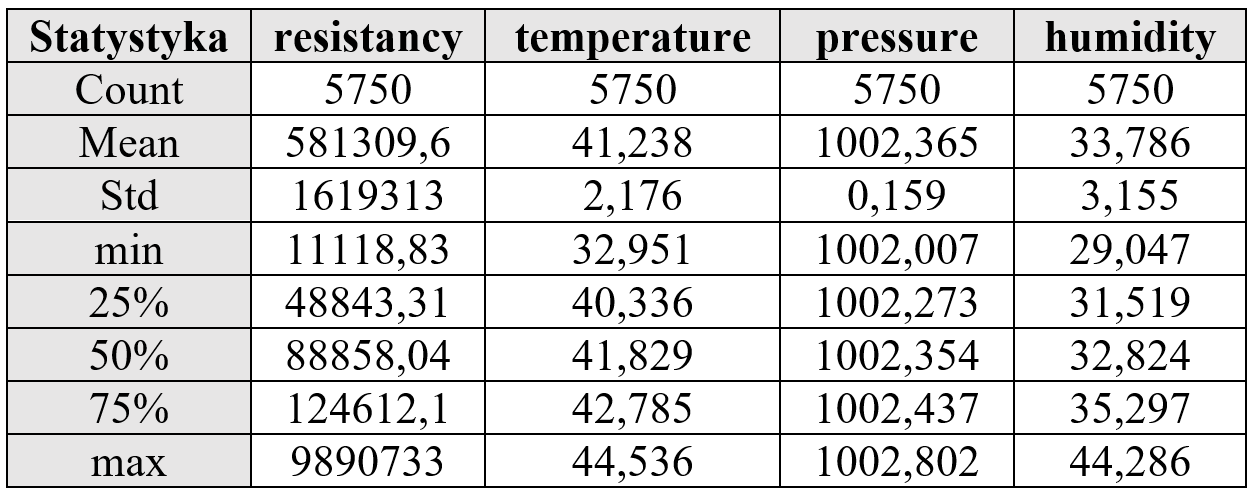
Nawykresach na Rysunku 5 widać tendencję badanych zmiennych dla przedstawionych perfum – zwiększa się temperatura otoczenia pod wpływem pracy układu, zmniejsza się wilgotność powietrza. Wraz z upływem czasu zbierania próbek, zwiększają się wartości rezystancji.

# Eksploracyjna analiza danych

W Tabeli 2 zaprezentowano statystyki opisowe danych odczytanych   
z czujnika gazów na przykładzie zapachu *Euphoria*.

*Tabela 2. Statystyki opisowe czterech zmiennych dla perfum Euphoria*

*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 formuły

gdzie jest średnią arytmetyczną, a odchyleniem standardowym .

Warto także zwrócić uwagę na minimalne oraz maksymalne wartości rezystancji – wartość największa jest niemal 900-krotnie większa od najmniejszej. Wynika to z faktu, że przebieg rezystancji jest skalowany w zależności od kroku pomiarowego, co można zaobserwować na Rysunku 5.

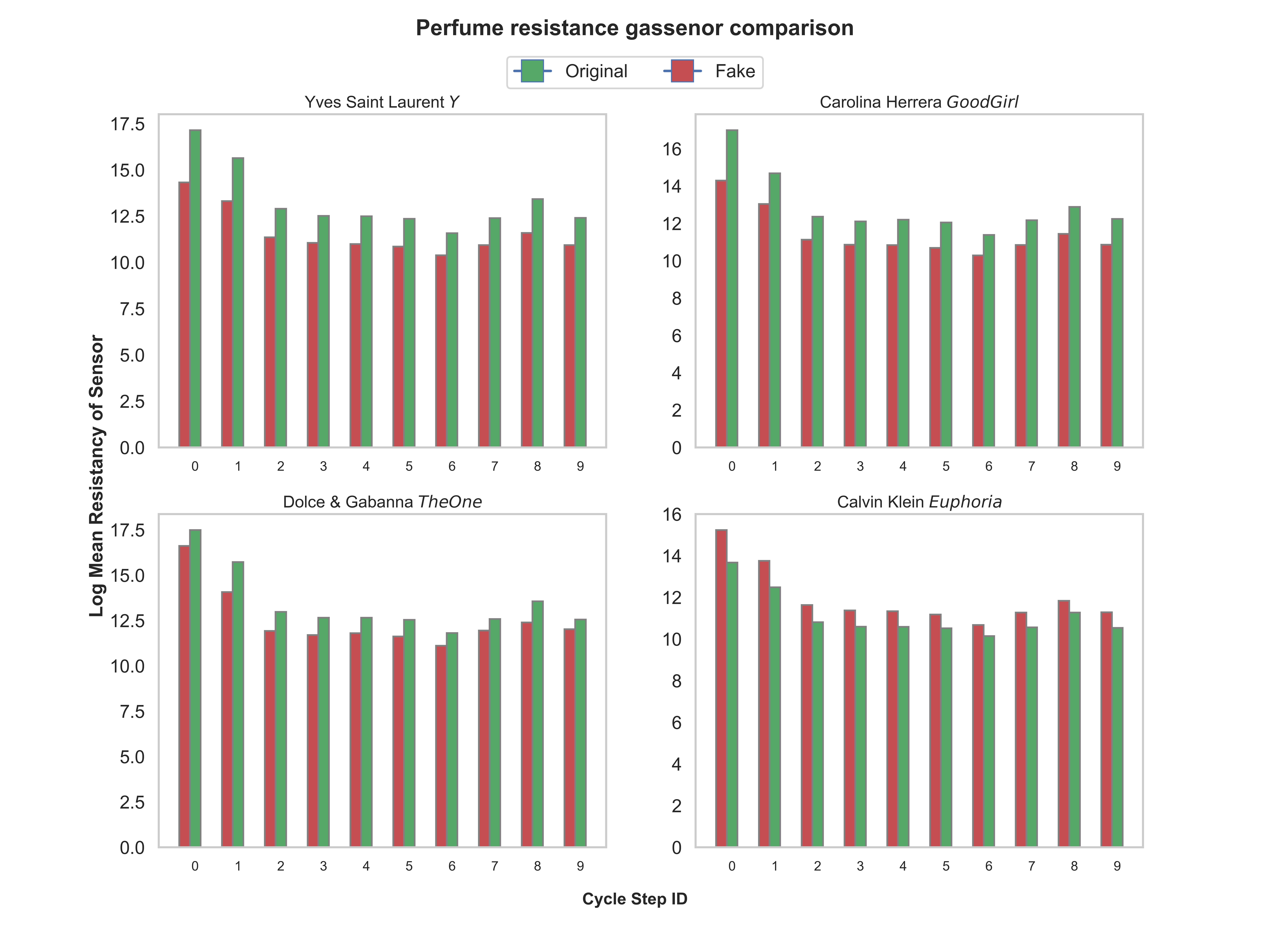
# Porównanie oryginałów oraz falsyfikatów perfum

# 

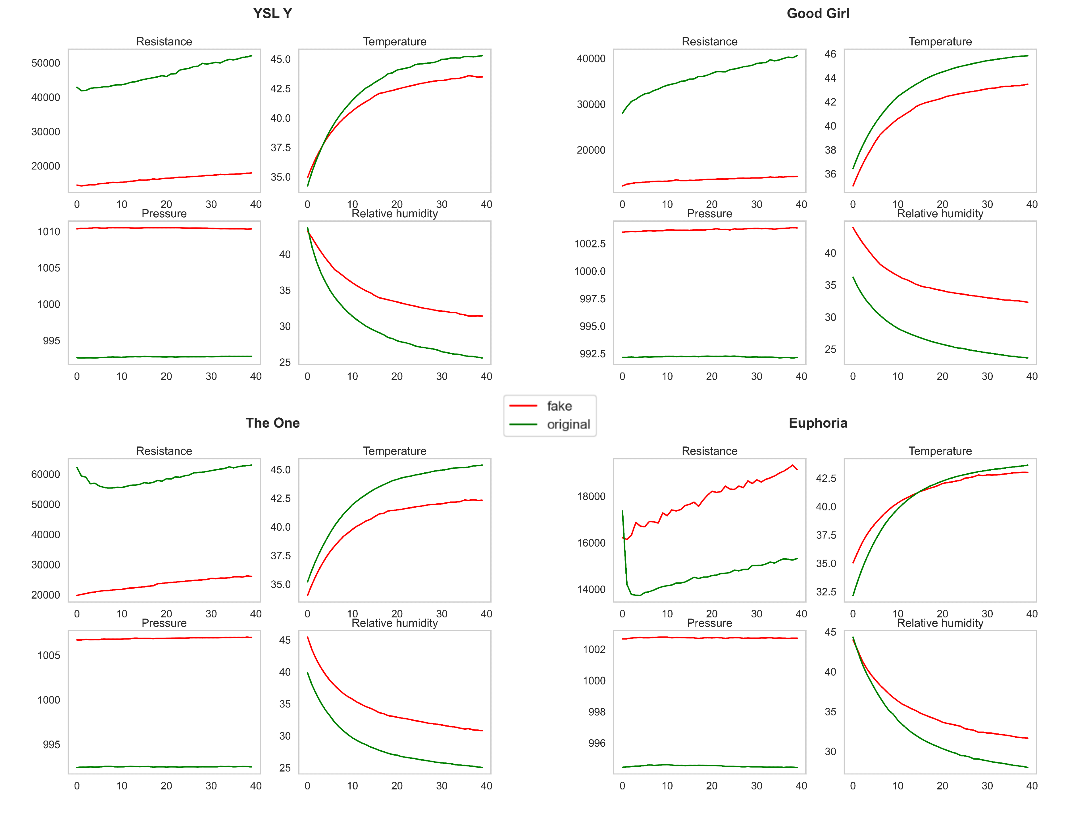
*Rys. 6. Porównanie odczytów rezystancji dla oryginalnych perfum oraz falsyfikatów*

*Źródło: Opracowanie własne*

Na wykresie na Rysunku 6 można zauważyć, 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 falsyfikatów dla każdego kroku cyklu  
Źródło: Opracowanie własne

Na Rysunku7, dla wszystkich zapachów oprócz *Euphoria* (*CK*)*,* na wykresach obserwowane są wyższe wartości odczytów rezystancji dla oryginalnych perfum, niż dla ich falsyfikatów.

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

*Źródło: Opracowanie własne*

Na Rysunku 8 zauważyć można 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.

# Analiza skupień

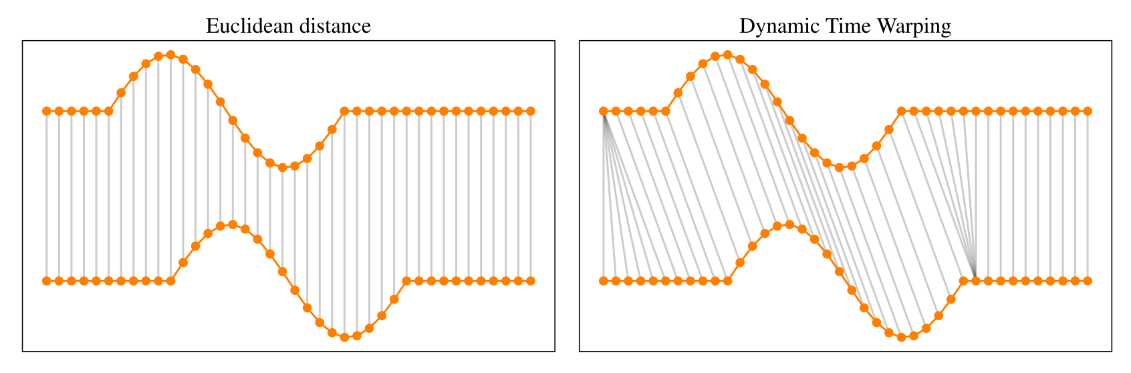
W celu odkrycia wzorców i podobieństw w zbadanych próbkach zapachowych, zastosowano algorytmy analizy skupień. Mimo, że analiza skupień jest jedną z metod nienadzorowanego uczenia maszynowego, w przypadku, gdy próbki były oznaczone, możliwe było porównanie wyników  
z rzeczywistymi etykietami.

Ze względu na sekwencyjny charakter danych, w badaniu nie należało 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. Wykorzystano do tego adaptację algorytmu *k*-średnich dla szeregów czasowych [7]. 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 wykorzystano metrykę *Dynamic Time Warping* *(DTW)* [6], polegającą na znalezieniu optymalnej ścieżki dopasowania pomiędzy dwoma szeregami, minimalizującej sumaryczną odległość między odpowiadającymi sobie punktami czasowymi. Odległość w metryce *DTW* danej wzorem (2), jest szczególnie użyteczna 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.

gdzie to zbiór wszystkich możliwych ścieżek między sekwencjami , to odległość pomiędzy , zaś to parametr metryki Minkowskiego.

 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***,* zdecydowano się na wybór szesnastu 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 wykorzystano płytką sieć neuronową, składającą się   
z jednej warstwy LSTM, jednej warstwy gęstej oraz warstwy wyjściowej, w skład której wchodzi dziewięć 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 standaryzowane sygnały pochodzące ze wszystkich kanałów pomiarowych – tensor rozmiaru (4, 101).

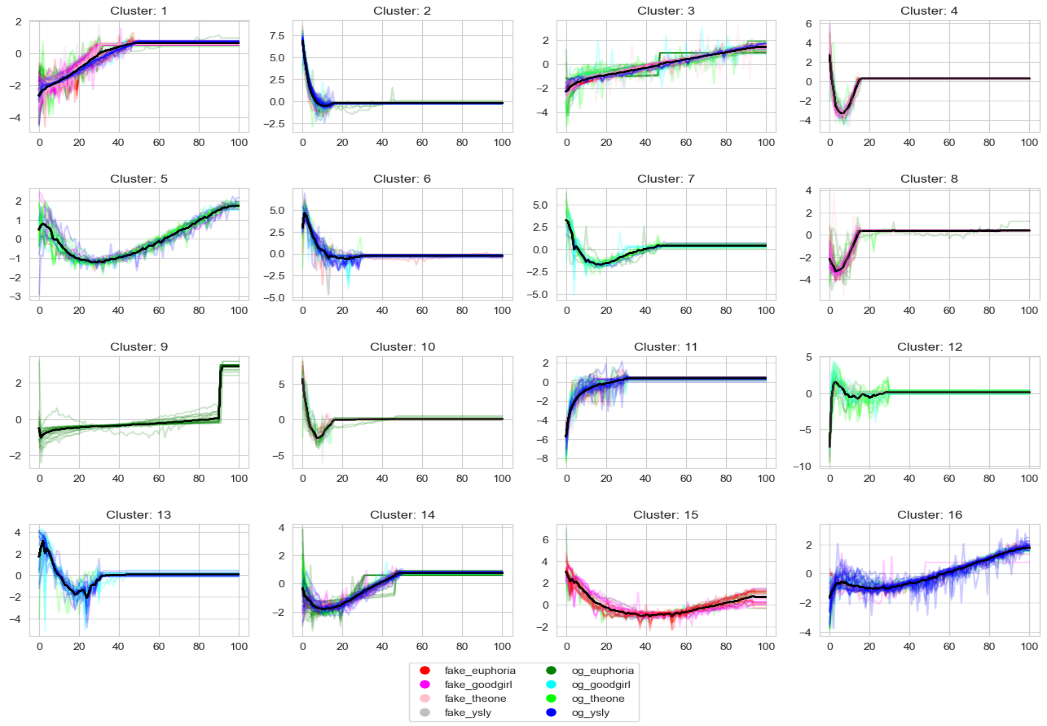
Dane podzielone zostały na trzy podzbiory: treningowy, walidacyjny oraz testowy. Zbiór walidacyjny posłużył monitorowaniu, czy nie wystąpiło zjawisko przeuczenia, a zbiór testowy finalnej ewaluacji dopasowania modelu.

Aby zwiększyć liczbę danych treningowych, tym samym poprawiając zdolność modelu do generalizacji, zastosowano technikę zaszumienia próbek treningowych poprzez dodanie do nich szumu gaussowskiego [2] o średniej wartości 0 oraz odchyleniu standardowym równym 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 [10] poprzez optymalizację funkcji straty kategorycznej entropii krzyżowej (ang. *Category Cross Entropy Loss Function*) [5].

# Wyniki klastrowania

Rysunek11 przedstawia wyniki grupowania sygnałów rezystancji zebranych próbek ze względu na kształt przebiegu sygnału w czasie.

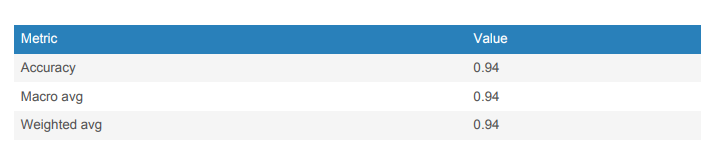
Zauważyć można, że uzyskane skupienia nie są całkowicie jednorodne ze względu na dany zapach. Oznacza to, że kształty sygnałów dla różnych perfum nie wykazują pełnej unikalności każdego z badanych wariantów. Wyjątek stanowią tutaj oryginały zapachów *Euphoria* (*CK*) oraz *The One* (*D&G*)–odpowiednio klastry 9. i 12. Pomimo to, wyniki wskazują, że zapachy mają tendencję do grupowania się w 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 3.

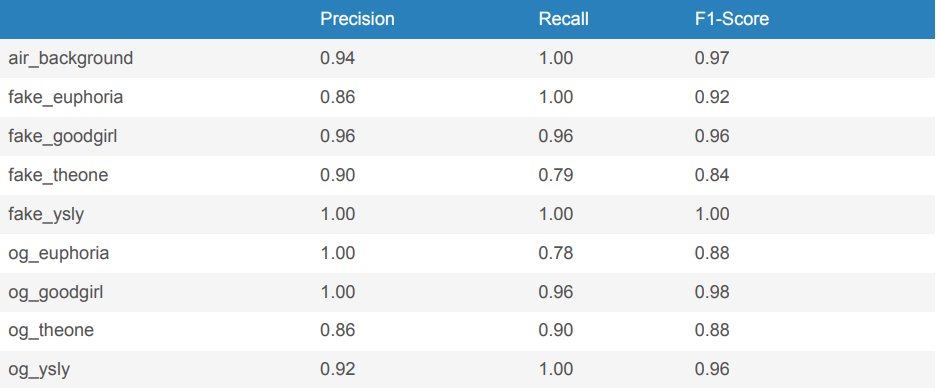
Tabela 3. Wartości ogólnych miar dopasowania na zbiorze testowym



Źródło: Opracowanie własne

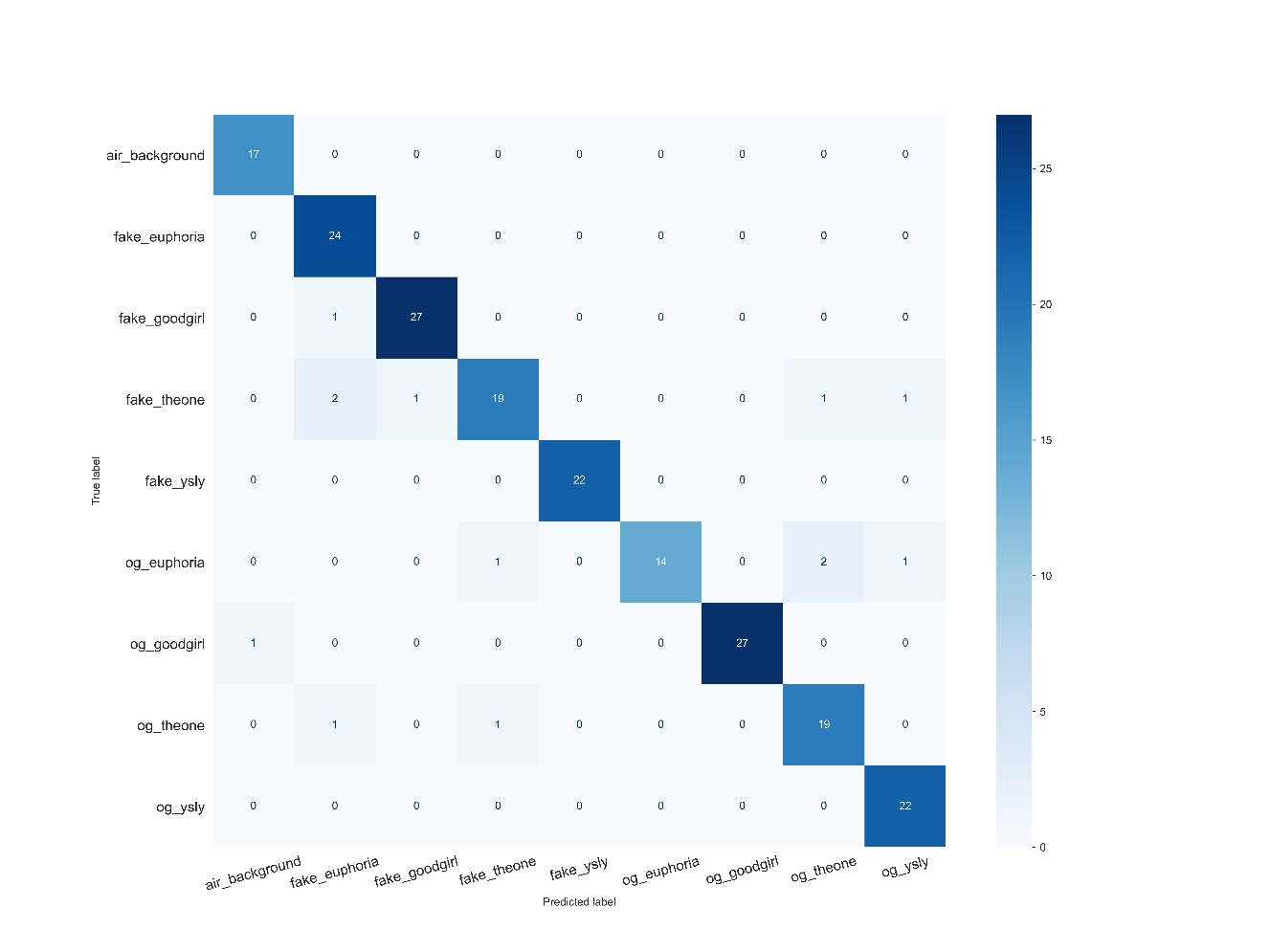
Z metryk zestawionych w Tabeli 4*.* wynika, że najniższe wartości *recall* dotyczą odpowiednika zapachu *The One* (*D&G*)oraz oryginału *Euphoria* (*CK*)*.* To oznacza, że model nie jest w stanie rozpoznać tych klas dla wszystkich przypadków, choć jednocześnie wysokie *precision* świadczy o dobrej jakości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

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



Źródło: Opracowanie własne

Z macierzy kontyngencji przedstawionej na Rysunku 12 odczytać można, że błędne klasyfikacje prawdziwego zapachu jako podróbki stanowią niewielką liczbę przypadków.

Rys. 12. Macierz kontyngencji zbioru testowego dla klasyfikacji rodzajów perfum i ich imitacji  
Źródło: Opracowanie własne

# Podsumowanie

W pracy tej przedstawiono skuteczny modelu klasyfikujący zebrane przez autorów próbki perfum. Dowodzi to zasadności wykorzystania elektronicznego nosa oraz technik analizy danych w rozróżnianiu złożonych lotnych związków organicznych. Wyniki klastrowania wykazały, że większość zapachów grupuje się w jednorodne skupienia, w podziale na oryginały i imitacje, w szczególności klastry zapachów *Euphoria* oraz *The One*.

Badanie potwierdza, że elektroniczny nos BME688 przy użyciu algorytmów uczenia maszynowym jest skutecznym narzędziem do rozróżniania perfum   
i wykrywania falsyfikatów. Otrzymane wyniki niosą perspektywę wykorzystania stworzonego modelu zarówno w branży kosmetycznej, jak i w innych gałęziach przemysłu, w tym związanych z żywieniem czy ochroną zdrowia.

# Literatura

[1] Bhattacharyya N. et al., *Electronic nose for black tea classification and correlation of measurements with ‘tea taster’ marks*, IEEE Transactions on Instrumentation and Measurement, 2008

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

[3] Dang L. et al., *A novel classifier ensemble for recognition of multiple indoor air contaminants by an electronic nose*, Sensors Actuators A Physical, 2014

[4] Di Natale C. et al., *Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors*, Biosensors Bioelectron, 2003

[5] 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

[6] Tavenard R., *An introduction to Dynamic Time Warping*, <https://rtavenar.github.io/blog/dtw.html>, 2021

[7] 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](http://www.jmlr.org/papers/v21/20-091.html), 2020

[8] 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

[9] Gas sensor BME688 |Bosch Sensortec, <https://www.bosch-sensortec.com/products/environmental-sensors/gas-sensors/bme688>, dostęp dnia 15.04.2024 roku

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

[11] Feather, Boards Products Category on Adafruit Industries, <https://www.adafruit.com>, dostęp dnia 15.04.2024 roku

1. Politechnika Lubelska, Wydział Matematyki i Informatyki Technicznej, Studenckie Koło Naukowe Uczenia Maszynowego ATLAS [↑](#footnote-ref-1)
2. Politechnika Lubelska, Wydział Zarządzania, Studenckie Koło Naukowe Uczenia Maszynowego ATLAS [↑](#footnote-ref-2)
3. Politechnika Lubelska, Wydział Matematyki i Informatyki Technicznej, Katedra Matematyki Stosowanej [↑](#footnote-ref-3)