Uniwersytet Jagielloński w Krakowie

Wydział Fizyki, Astronomii i Informatyki Stosowanej

Łukasz Wójcik

Nr albumu: 1188524

Opracowanie jedno- i wielomodalnych modeli predykcji emocji

Praca magisterska
na kierunku Informatyka Stosowana

Praca wykonana pod kierunkiem:
dr inż. Krzysztofa Kutta
Zakład Technologii Gier

Kraków 2023

Abstract

This in an abstract in English...

Abstrakt

...a to jest abstrakt po polsku.

Spis treści

Wstęp **4**

1 Automatyczna predykcja emocji **5**
1.1 Uogólniony system rozpoznawania emocji 5
1.2 Metody rozpoznawania emocji 6
1.2.1 Wyraz twarzy 7
1.2.2 Postawa ciała i gestykulacja 8
1.2.3 Mowa 9
1.2.4 Sygnały biofizyczne 10
1.3 Reprezentacja emocji w systemie komputerowym 13
1.4 Modalność w modelach predykcji emocji 14

2 Uczenie maszynowe **16**
2.1 Podstawy uczenia maszynowego 16
2.1.1 Rodzaje uczenia maszynowego 16
2.1.2 Przeuczenie i niedouczenie 17
2.2 Przykładowe algorytmy uczenia maszynowego 18
2.2.1 Drzewa decyzyjne i lasy losowe 18
2.2.2 Maszyna wektorów nośnych 20
2.3 Sztuczne sieci neuronowe 21
2.3.1 Głębokie sieci neuronowe 21
2.3.2 Funkcje aktywacji 22
2.3.3 Gradient descent 23

3 Część praktyczna **25**
3.1 Zbiór danych 25
3.2 Przygotowanie danych 26
3.2.1 Oczyszczanie i ekstrakcja cech 26
3.2.2 Grupowanie 27
3.3 Wyniki 28

Podsumowanie **30**

Bibliografia **33**

Spis rysunków **35**

Spis tabel **36**

Wstęp

Emocje są nieodłączną częścią ludzkiego życia. Stanowią bardzo ważny element komunikacji niewerbalnej, wpływają na zachowanie i postrzeganie świata. Możliwość rozpoznawania emocji, reagowanie na nie oraz ich wywoływanie pozwoliłoby na rozwój w wielu dziedzinach. Do najważniejszych należą robotyka, zwłaszcza interakcja człowiek-robot, marketing, szkolnictwo, przemysł rozrywkowy. Doprowadziło to do powstania interdyscyplinarnej nauki pod nazwą informatyka afektywna (ang. *affective computing*), która łączy elementy informatyki, psychologii, neurologii, inżynierii i wielu innych.

Rozwój technologiczny pozwala na podejmowanie prób automatycznego rozpoznawania emocji przy pomocy systemów komputerowych. Algorytmy uczenia maszynowego potrafią przetwarzać bardzo zróżnicowanie źródła informacji o emocjach. Coraz mniejsze rozmiary sensorów i większa dokładność pozwalają na rejestrowanie przeróżnych parametrów: wyraz twarzy, sposób mowy, a nawet sygnały biofizyczne, takie jak EKG lub EEG.

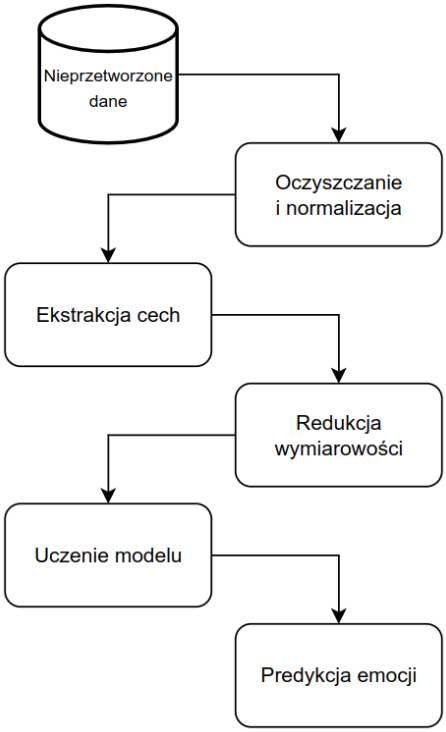
Celem pracy jest wykorzystanie istniejącego zbioru zawierającego zapisy elektrokardiografii oraz reakcji skórno-galwanicznej i stworzenie systemu, który będzie w stanie dokonywać predykcji emocji. W tym celu zaprojektowane zostanie kilka modeli, z wykorzystaniem różnych algorytmów uczenia maszynowego. Następnie wykonane zostanie porównanie osiągniętych wyników.

Rozdział 1

Automatyczna predykcja emocji

1.1 Uogólniony system rozpoznawania emocji

W badaniach nad automatycznym rozpoznawaniem emocji dominują systemy oparte na uczeniu maszynowym lub modelach statystycznych [1, 2]. Powoduje to, że wykonuje się w nich podobne kroki. Dane w takich systemach przechodzą zazwyczaj sekwencyjnie przez tak zwany potok (ang. *pipeline*) [3].



Rysunek 1.1: Ogólny schemat systemu predykcji emocji

Rysunek 1.1 przedstawia uproszczony schemat systemu predykcji emocji. Na początku system otrzymuje nieprzetworzone dane, często nazywane surowymi (ang. *raw*

data). W zależności od źródła mogą to być zdjęcia, filmy, zapisy sygnałów biofizycznych (elektrokardiografia, elektroencefalografia itp.) oraz wiele innych. Są to wartości pochodzące bezpośrednio z sensorów, bazy danych lub publicznie dostępnych zbiorów [1, 4].

Dane wejściowe zazwyczaj nie są wystarczającej jakości dlatego kolejny krok to ich oczyszczanie. Może to być na przykład redukcja szumów w sygnale, odrzucanie skrajnych wartości lub uzupełnianie brakujących. Dodatkowo niektóre modele dają lepsze wyniki po normalizacji danych [3].

Większość modeli nie przyjmuje na wejściu surowych danych, dlatego następnie wykonuje się proces ekstrakcji cech (ang. *feature extraction*). Pozwala on na zmianę danych wejściowych na wartości istotne dla rozpoznawania emocji. Często są to wartości z funkcji i miar statystycznych, takie jak mediana, średnie, odchylenia itp. Przykładowe cechy to geometria twarzy, szybkość mówienia, czas pomiędzy uderzeniami serca [5].

W zależności od metody ilość cech może wynosić ponad 700 [6], dlatego w niektórych systemach kolejnym krokiem jest redukcja wymiarowości. Ma ona na celu zmniejszenie liczby cech wejściowych poprzez łączenie tych silnie skorelowanych, rzutowanie w mniej wymiarowe przestrzenie lub odrzucanie wartości, które nie poprawiają wyników. Powszechnie stosowane podejścia to: analiza głównych składowych (ang. *principal components analysis (PCA)*), grupowanie hierarchiczne (ang. *hierarchical cluster analysis (HCA)*), Gaussian random projection [3].

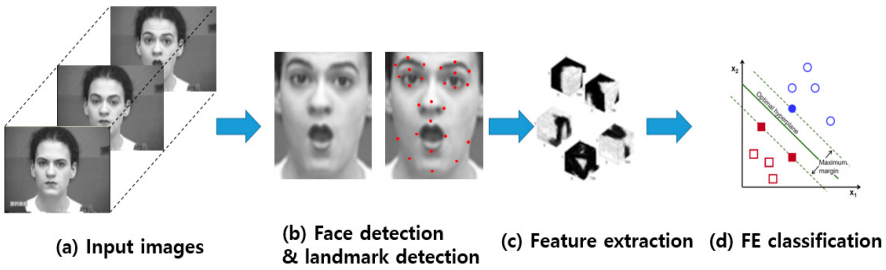
Po uzyskaniu ostatecznych danych następuje trening modelu, zazwyczaj jest to uczenie nadzorowane [4]. Oznaczanie danych dobywa się na dwa sposoby. W pierwszym każdy wektor danych ma przypisaną kategoryczną emocję, na przykład strach lub złość. W drugim emocje opisane są w przestrzeni wielowymiarowej, zatem zamiast jednej kategorii posiadają zazwyczaj dwie lub trzy wartości. Do klasyfikacji stosuje się różnorakie podejścia, między innymi: support vector machine (SVM), random forest classifier (RFC), stochastic gradient descent (SGD), AdaBoost, k-nearest neighbor (k-NN), hidden Markov models (HMM), linear discriminate analysis (LDA), sztuczne sieci neuronowe [1, 2, 7].

1.2 Metody rozpoznawania emocji

Istnieje wiele sposobów, na podstawie których można wnioskować stan emocjonalny człowieka. Pozwala to na wykorzystanie bardzo zróżnicowanych podejść, od oceny wyglądu, przez analizę zachowań, aż po pomiary aktywności elektrycznej w organizmie. Poniżej znajduje się opis najczęściej stosowanych metod [1, 2], ich wady oraz zalety.

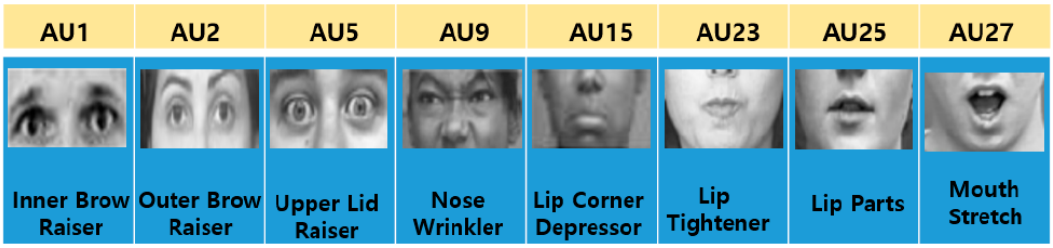
1.2.1 Wyraz twarzy

Ludzka twarz jest bardzo znaczącym źródłem informacji i odgrywa dużą rolę w komunikacji niewerbalnej. Na jej podstawie można oceniać między innymi: płeć, wiek, pochodzenie etniczne, czy stan emocjonalny [5]. Dzięki temu twarz jest bardzo popularnym źródłem w automatycznym rozpoznawaniu emocji, z początkami prac sięgającymi lat 90. XX wieku [5].



Rysunek 1.2: Schemat systemu rozpoznającego emocje na podstawie twarzy. *Źródło: [7]*

Rysunek 1.2 przedstawia ogólny schemat systemu rozpoznającego emocje na podstawie wyrazu twarzy. Na wejściu program otrzymuje pojedyncze zdjęcie lub nagranie zawierające twarz. Pierwszym krokiem jest wykrycie twarzy. W dostarczonym źródle może być ich wiele. Następnie przeprowadza się proces ekstrakcji charakterystycznych miejsc. Po uzyskaniu danych o twarzy zostają one przetworzone przez algorytm uczenia nadzorowanego [5].



Rysunek 1.3: Przykłady różnych Action Units w trzech częściach twarzy. *Źródło: [7]*

Jedno z popularnych podejść rozpoznaje emocje na podstawie Facial Action Coding System (FACS) [8]. Zbiór ten zawiera, w zależności od wersji, od 33 do 44 tak zwanych Action Units (AU). Powstały one przy pomocy stymulacji elektrycznej mięśni twarzy, które biorą udział w wyrażaniu emocji. Dzięki temu uzyskano obiektywne ruchy mięśni o różnej intensywności zależnej od napięcia prądu. Sam zbiór nie zawiera ścisłego określenia połączeń AU i odpowiadającym im emocjom, a jedynie hipotezy [5].

Inne podejścia bazują na zbiorach, w których osoby były proszone o wyrażenie danej emocji. Pozyskane w ten sposób dane mają jednak wadę w postaci zbyt intensywnego wyrazu twarzy, w dodatku opartych na stereotypach. Osoba poproszona o to, aby pokazała zdziwienie, zazwyczaj wygląda zupełnie inaczej, niż gdy jest naprawdę zdziwiona. Nawet dobrze wyszkolony aktor nie jest w stanie dokładnie odwzorować naturalnej reakcji [5]. Powoduje to, że modele szkolone na takich zbiorach nie są w stanie rozpoznawać emocji wyrażanych w sposób „normalny". Jedną z możliwości zapobiegania temu zjawisku jest tworzenie zbiorów, w których emocje są wywoływane przez prawdziwe zdarzenia, a nie odgrywane przez aktorów [5].

Główną zaletą rozpoznawania emocji na podstawie twarzy jest stosunkowa prostota, aparaty są tanie i powszechnie dostępne. Dodatkowo zbieranie danych nie wymaga kontaktu fizycznego i nie powoduje dyskomfortu. Same wyrazy twarzy dla wielu emocji są uniwersalne między członkami różnych kultur, płci i niezależne od wieku [5].

Mimo to z tym podejściem wiąże się wiele problemów. Począwszy od trudności wynikających z samego rozpoznawania twarzy, na przykład różne oświetlenie, czy kąt, pod jakim znajduje się twarz. Następnie pojawiają się problemy związane z emocjami: osoba jest w stanie celowo nie wyrażać żadnych emocji lub mogą być one bardzo nikłe [5].

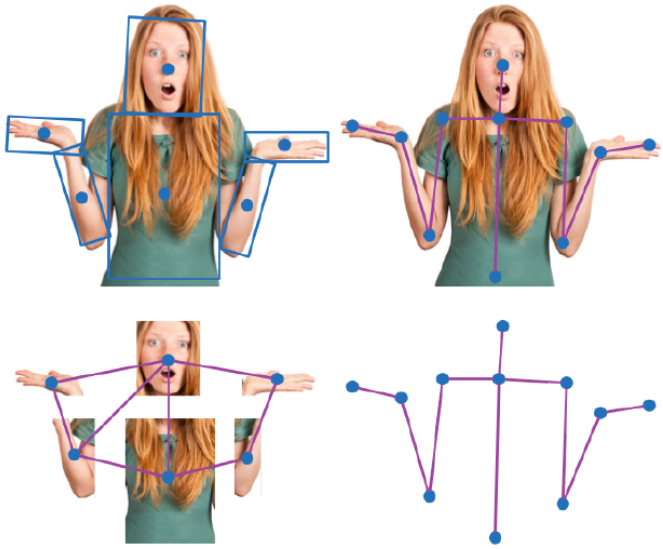
1.2.2 Postawa ciała i gestykulacja

Drugim bardzo ważnym źródłem informacji o emocjach jest postawa ciała człowieka, jego gesty lub ich brak. Stanowią one znaczną część komunikacji niewerbalnej, ruch dłoni jest drugim co do wielkości źródłem, mówiącym o stanie emocjonalnym, więcej informacji pochodzi jedynie z wyrazu twarzy [9]. Co więcej, postawa ciała pomaga w zmaganiu się z aktualnie odczuwanymi emocjami [10].

Jednym z powszechnie stosowanych sposobów śledzenia ruchu ciała są kamery termowizyjne [5]. Pomiary są możliwe dzięki odblaskowym płytkom, które umieszczane są na odzieży. Pozwala to na zapis ruchu w trójwymiarowej przestrzeni. Tego typu podejście wymaga jednak noszenia specjalnego stroju, a dokładność jest zależna od ilości znaczników. Z tego powodu mierzenie ruchu dłoni, a zwłaszcza palców jest problematyczne. Z drugiej strony zbieranie jest mniej danych, a ich przetwarzanie jest łatwiejsze. Dodatkowo zapewniona jest anonimowość badanych [5].

Dzięki rozwojowi widzenia maszynowego możliwe stało się również używanie zwykłych kamer. Takie podejście zapewnia większą swobodę, nie wymaga specjalnego stroju. Co najważniejsze pozwala na dokładniejsze odwzorowanie ruchów, zwłaszcza palców. To podejście również musi zmagać się z problemami typowymi dla rozpoznawania obrazów: oświetlenie, kolor skóry, ubrania mogą negatywnie wpływać na dokładność [5].

Rysunek 1.4 przedstawia dwa sposoby modelowania ludzkiego ciała w systemach komputerowych. Po lewej stronie widnieje model oparty na częściach ciała (ang. *part*



Rysunek 1.4: Sposoby reprezentowania ciała w komputerze: zbiór częsci ciała (lewa strona) oraz reprezentacja szkieletowa (prawa strona). *Źródło: [9]*

based model). Każda część jest rozpoznawana osobno na podstawie wiedzy o budowie ludzkiego ciała. Otrzymywana jest reprezentacja dwuwymiarowa. Po prawej stronie przedstawiono model szkieletowy (ang. *kinematic model*). W tej reprezentacji ciało jest zbiorem wierzchołków połączonych krawędziami, przez co można je reprezentować jako graf [9]. Wierzchołki interpretowane są jako stawy, które posiadają pewne stopnie swobody, odpowiednie dla danej części ciała. Pozwala to na złożoną reprezentację w przestrzeni trójwymiarowej [5].

Po uzyskaniu reprezentacji ciała, w systemie następuje proces rozpoznawania postawy, a następnie oceny emocji. Używa się do tego zarówno statycznych obrazów, jak i nagrań ruchu [9, 10].

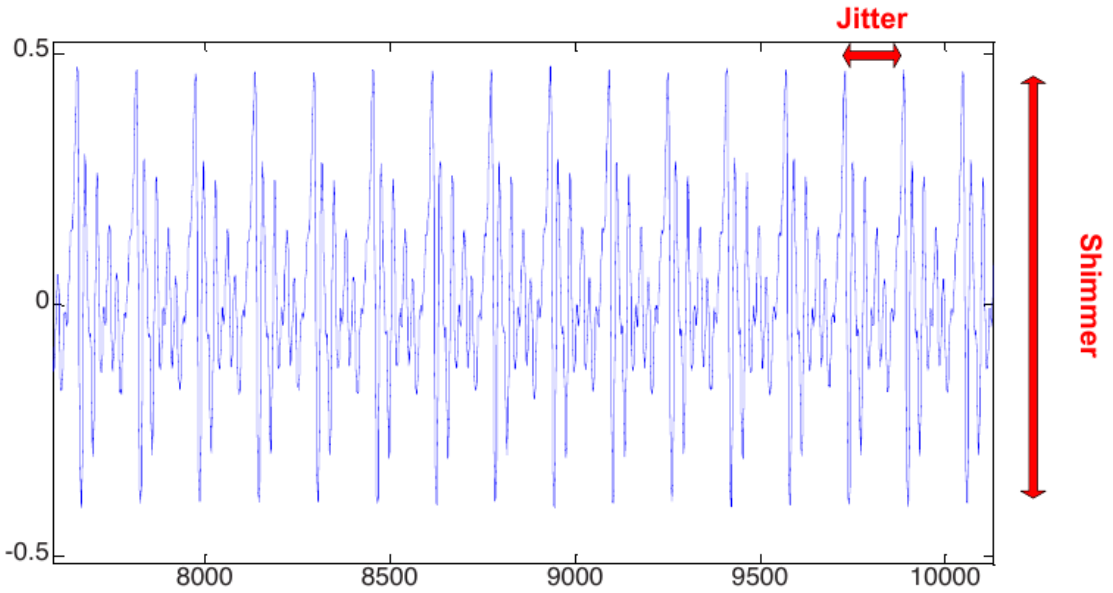
1.2.3 Mowa

Poza niewerbalnymi źródłami, emocje można również rozpoznawać na podstawie mowy. Ludzki głos stanowi bardzo bogate źródło informacji. Pozwala na wnioskowanie o wieku, płci, stanie emocjonalnym, osobowości, dialekcie i pochodzeniu mówcy [11].

W porównaniu do poprzednich źródeł mowa jest o wiele bardziej podatna na zakłócenia, szum, hałasy w tle. Wymaga więc dokładniejszego procesu oczyszczania. Bardzo ważna jest również normalizacja danych. Zakres podstawowej częstotliwości głosu, który wynosi około 50 — 500 Hz, jest o wiele większy niż różnica między wypowiedzią neutralną i w stanie złości, czyli około 68 Hz [5].

Po oczyszczeniu i normalizacji następuje proces ekstrakcji cech niskiego poziomu (ang. *low-level descriptors (LLD)*). Są to wartości oparte o częstotliwość głosu oraz

o zmiany w sposobie wypowiedzi (na przykład szybkość mówienia lub poziom głośności). Sama ilość cech niskiego poziomu nie jest z góry określona i może być różna w zależności od podejścia. Do najpopularniejszych LLD należą: fundamental frequency (F0), Mel-frequency cepstral coefficients (MFCCs), jitter, shimmer, harmonic-to-noise ratio oraz wartości z widma akustycznego [5, 12].



Rysunek 1.5: Przykładowy sygnał mowy z zaznaczonymi jitter i shimmer.
Źródło: [13]

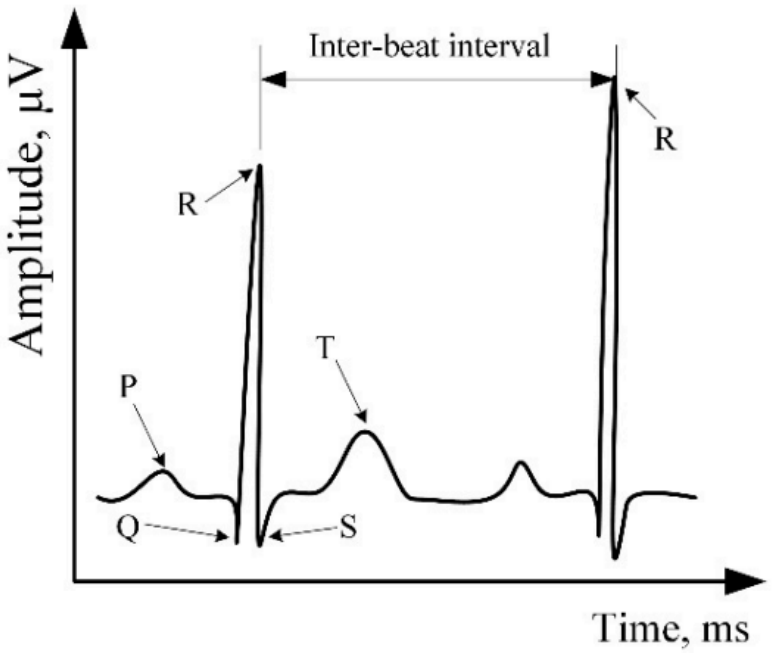
Po uzyskaniu cech niskiego poziomu można zastosować funkcje i miary statystyczne, takie jak średnie i odchylenia, aby otrzymać tak zwane cechy wysokiego poziomu (ang. *high-level descriptors (HLD)*) [5].

Podobnie jak wyrazy twarzy, mowa jest zależna od kultury i pochodzenia osoby. Dodatkowo wyszkolona osoba jest w stanie kontrolować wymowę w taki sposób, aby ukrywać odczuwane emocje lub udawać inne [5].

1.2.4 Sygnały biofizyczne

Emocje wywołują również zmiany, których nie da się zaobserwować za pomocą wzroku lub słuchu. Różne stany emocjonalne wpływają między innymi na szybkość bicia serca, wydzielanie potu, oddech, temperaturę ciała. Są to parametry, które można zmierzyć i wnioskować na ich podstawie odczuwane emocje [5].

Jednym z najpopularniejszych sposobów jest elektrokardiografia (EKG), czyli mierzenie aktywności elektrycznej serca. Do pomiarów używa się elektrod umieszczonych na skórze, najczęściej jest ich 3 lub 12. Analiza sygnału odbywa się na podstawie załamków P, Q, R, S, T [5].



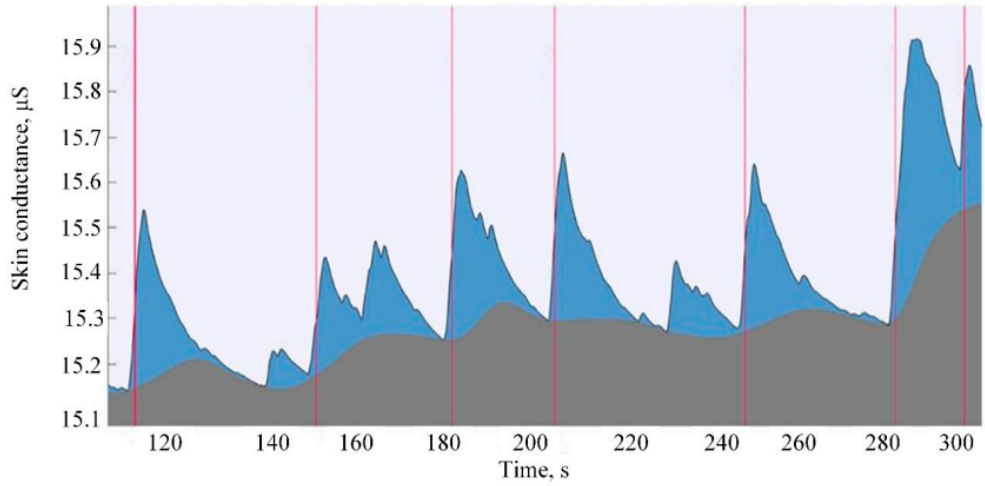
Rysunek 1.6: Przykładowy sygnał EKG z zaznaczonymi załamkami. *Źródło: [2]*

Rysunek 1.6 przedstawia przykładowy sygnał EKG. Jako pierwszy występuje załamek P, który oznacza depolaryzację mięśnia przedsionków. Potem następuje zespół QRS opisujący depolaryzację mięśnia komór. Po nim pojawia się załamek T odpowiadający repolaryzacji komór [2].

W automatycznym rozpoznawaniu emocji najczęściej bierze się pod uwagę zespół QRS oraz odległości między załamkami R (ang. *R-R interval / inter-beat interval*), które wykorzystuje się w analizie zmienności rytmu zatokowego (ang. *heart rate variability (HRV)*) [5].

Drugim często używanym sygnałem biofizycznym jest reakcja skórno-galwaniczna, powszechnie używa się dwóch skrótów: GSR (ang. *galvanic skin response*) lub EDA (ang. *electrodermal activity*). Opisuje ona zmiany w przewodnictwie skóry spowodowane aktywnością gruczołów potowych. Prowadzi to do różnic w wilgotności i w następstwie do zmiany oporu elektrycznego [2]. Pomiary wykonuje się za pomocą elektrod, które mogą być umieszczone w dowolnym miejscu na skórze. Zazwyczaj wykorzystuje się miejsca najbardziej czułe na zmiany emocjonalne: dłonie oraz podeszwy stóp [5].

Rysunek 1.7 przedstawia przykładowy sygnał GSR, który składa się z dwóch głównych komponentów. Szarym kolorem zaznaczono tonic component, który zmienia się powoli i zależy głównie od reakcji na czynniki środowiska (temperatura, wilgotność powietrza itp.). Na niebiesko oznaczono phasic component, przejawiający się jako krótkie piki w odpowiedzi na stan emocjonalny [2].



Rysunek 1.7: Przykładowy sygnał GSR. Czerwone linie oznaczają momenty pojawiania się stymulantu. *Źródło: [2]*

Poza elektrokardiografią oraz reakcją skórno-galwaniczną stosuje się również wiele innych podejść.

Fotopletyzmografia (ang. *photoplethysmography (PPG)*) jest alternatywnym sposobem mierzenia aktywności serca. Do pomiarów używa się światła, które reaguje na zmiany w ilości krwi w tkankach. Różnice w odbijanym lub przepuszczanym świetle odpowiadają uderzeniom serca [2].

Elektroencefalografia (EEG) jest używana do badania aktywności mózgu na podstawie fal $\delta, \theta, \alpha, \beta, \gamma$. Pomiary odbywają się za pomocą elektrod umieszczonych na głowie. Zazwyczaj używa się 8, 16 lub 32 pary [2].

Elektromiografia (EMG) służy do pomiaru aktywności elektrycznej mięśni. Podczas skurczu mięśni pojawia się napięcie, które można zmierzyć na powierzchni skóry przy pomocy elektrod. EMG jest zazwyczaj stosowane dla mięśni twarzy [2].

Oddychanie jest również sygnałem biofizycznym. Pomiary wykonuje się zazwyczaj za pomocą opaski wokół klatki piersiowej, która mierzy jej ruch wywołany wdechami i wydechami [2].

Dużym problemem sygnałów biofizycznych są zakłócenia związane z aktywnością człowieka. Ruch ma duży wpływ na pracę serca, która nie zmienia się liniowo w stosunku do wysiłku. Kichnięcie powoduje w organizmie reakcję podobną do odczuwania strachu, mimo że osoba kichająca raczej nie jest przestraszona [5].

1.3 Reprezentacja emocji w systemie komputerowym

Po uzyskaniu danych należy je przypisać do odczuwanych emocji. Najprostszym sposobem jest przydzielenie im pewnej kategorii, na przykład: strach, złość, radość, smutek itp. W automatycznym rozpoznawaniu emocji ich liczba jest zazwyczaj niewielka i wynosi od 4 do 8 [2]. Czasem są to również klasyfikatory binarne, które przewidują jedynie czy dane należą do danej klasy, czy nie.

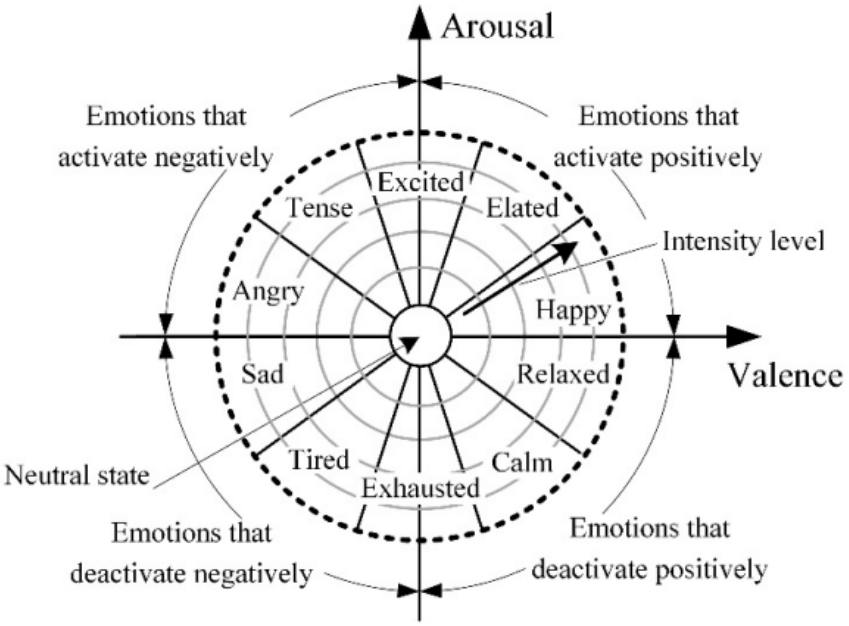
Rysunek 1.8: Przykładowa baza danych zawierająca zdjęcia twarzy przedstawiające podstawowe emocje. *Źródło: [14]*

Najczęściej wykorzystuje się kategorie należące do tak zwanych podstawowych emocji. Zostały one zaproponowane między innymi przez Paula Ekmana w 1971 roku [15]. Należą do nich: radość, smutek, złość, zdziwienie, strach oraz wstręt.

Inne popularne podejścia reprezentują emocje za pomocą dwóch lub trzech ciągłych wartości liczbowych. Stany emocjonalne są przedstawione w przestrzeni, dzięki czemu można reprezentować o wiele więcej kategorii, w sposób bardziej płynny i dokładny [5].

Wiele podejść opiera się na dwuwymiarowym modelu zaproponowanym przez Jamesa Russella [16]. Emocje w tym modelu reprezentowane są za pomocą wartości opisujących przyjemność odczuwanej emocji (ang. *valence*) oraz pobudzenie jakie wywołuje (ang. *arousal*). Sam model jest zazwyczaj w kształcie koła, które może być podzielone na wycinki przedstawiające emocje. Punkt leżący w danym wycinku przedstawia odpowiednią emocję [2].

Dwuwymiarowy model nie jest jednak w stanie wystarczająco rozróżniać niektóre emocje. Przykładowo strach oraz złość są reprezentowane jednakowo: przez wysokie pobudzenie i niską przyjemność. Aby poprawić rozpoznawanie stanów emocjonalnych, inne podejścia dodają trzeci wymiar utożsamiany z dominacją, jaką wywiera dana emocja [5].



Rysunek 1.9: Kołowy model oparty o torię Russella. *Źródło: [2]*

Samo przypisywanie emocji do danych odbywa się za pomocą dwóch podejść [5]. Pierwsze z nich opiera się na wyszkolonych obserwatorach, którzy oceniają stan emocjonalny badanej osoby. Nie zawsze jest to jednak możliwe, dlatego drugim powszechnie stosowanym sposobem jest samoocena. Metoda ta jest prostsza i pozwala na klasyfikacje emocji w sygnałach biofizycznych. Jest jednak bardziej zawodna, ponieważ osoba może źle sklasyfikować odczuwaną emocje, lub niedokładnie ocenić moment, w którym do niej doszło [5].

1.4 Modalność w modelach predykcji emocji

Początkowo modele automatycznej predykcji emocji opierały się wyłącznie na jednym źródle informacji, były to zatem systemy jednomodalne. Ten trend był tym bardziej wzmacniany przez skupienie się na rozpoznawaniu emocji na podstawie wyrazu twarzy. Takie podejście ma jednak jedną główną wadę, system nie jest w stanie rozpoznawać emocji, gdy brakuje danych wejściowych. Zdarza się, że twarz jest zakryta, osoba nic nie mówi lub stoi nieruchomo. Aby zapobiec temu problemowi oraz przez chęć uzyskania lepszych wyników rozpoczęto prace nad systemami wykorzystującymi więcej niż jedno źródło informacji [5].

Model wielomodalny to taki, który wykorzystuje co najmniej dwa różne źródła informacji. Może to być twarz oraz mowa, gestykulacja i sygnały biofizyczne, lub wszystkie na raz. Tego typu systemy o wiele rzadziej napotykają problem braku danych oraz zapewniają lepsze wyniki [17].

Systemy wielomodalne zmagają się jednak z innymi problemami. Największy to łączenie ze sobą danych, które wymagają różnych okienek czasowych do analizy. Film może być analizowany na podstawie pojedynczych klatek, jednak sygnały biofizyczne lub mowa wymagają zazwyczaj dłuższych pomiarów, aby dać wartościowe dane [5].

Rozdział 2

Uczenie maszynowe

2.1 Podstawy uczenia maszynowego

Klasyczne programy komputerowe składają się z szeregu instrukcji, w których zawarte są wszystkie możliwe akcje podejmowane przez użytkownika. Istnieją jednak problemy o tak dużej złożoności, że niemożliwe staje się opisanie wszystkich sytuacji. Przykładem może być autonomiczne sterowanie pojazdami. Przewidzenie wszystkich możliwych zjawisk na drodze jest bardzo trudne, jeśli nie niemożliwe. Klasyczny program musiałby być bardzo rozbudowany i złożony, a co za tym idzie również niełatwy do zrozumienia, implementacji i rozwoju. Chcemy zatem, aby taki pojazd potrafił sam podejmować decyzje nawet dla nieznanych sytuacji i uczył się na ich przykładzie.

O uczeniu maszynowym mówimy, gdy system komputerowy jest w stanie stworzyć model, który na podstawie obserwacji danych pozwala mu na tworzenie hipotez i podejmowanie decyzji [4]. Model to w rzeczywistości funkcja h, przybliżająca rzeczywistą funkcję f, która opisuje dane wejściowe. Dane te to wektory zawierające pewne cechy, na przykład: piksele w zdjęciu, częstotliwości dźwięku, ilość dni z opadami deszczu itp. W zależności od typu wartości, jakie są przewidywane, mówi się zazwyczaj o dwóch rodzajach problemów [4]:

- klasyfikacji (ang. *classification*), gdy dane wyjściowe zawierają się w skończonym zbiorze wartości,
- regresji (ang. *regression*), gdy dane wyjściowe są wartością liczbową.

2.1.1 Rodzaje uczenia maszynowego

Model uczy się dzięki zmienianiu tak zwanych parametrów. Są to wartości, które kontrolują działanie systemu, a ich ilość różni się w zależności od zastosowanego algorytmu [18]. Ponadto model uczy się na podstawie pewnej informacji zwrotnej. Rozróżnia się trzy główne podejścia [4]:

- W uczeniu nadzorowanym (ang. *supervised learning*) system uczy się na podstawie par składających się z wektora danych wejściowych i wartości, którą chcemy przewidzieć (ang. *label*). Celem uczenia jest przewidzenie wartości na podstawie danych wejściowych.

- W uczeniu nienadzorowanym (ang. *unsupervised learning*) system otrzymuje surowe dane wejściowe, w których ma wykryć pewne zależności lub wzorce. Program nie otrzymuje informacji zwrotnej, ponieważ nie istnieje pewien ściśle określony rezultat, który chcemy otrzymać.

- W uczeniu przez wzmacnianie (ang. *reinforcement learning*) uczenie odbywa się na podstawie systemu nagród i kar. Gdy system komputerowy robi to, co chemy osiągnąć otrzymuje za to pewną nagrodę, w przeciwnym wypadku jest karany. Celem systemu jest zatem podejmowanie akcji, które prowadzą do jak największej ilości nagród.

Algorytmy uczenia maszynowego posiadają zazwyczaj również szereg tak zwanych hiperparametrów (ang. *hyperparameters*), które można traktować jako dodatkowe opcje. Nie są one zmieniane w trakcie uczenia, a wymagają ustawienia ich przed uczeniem. Mają one znaczący wpływ na osiągane wyniki, dlatego zazwyczaj przeprowadza się osobny proces mający wybrać jak najlepsze wartości [18].

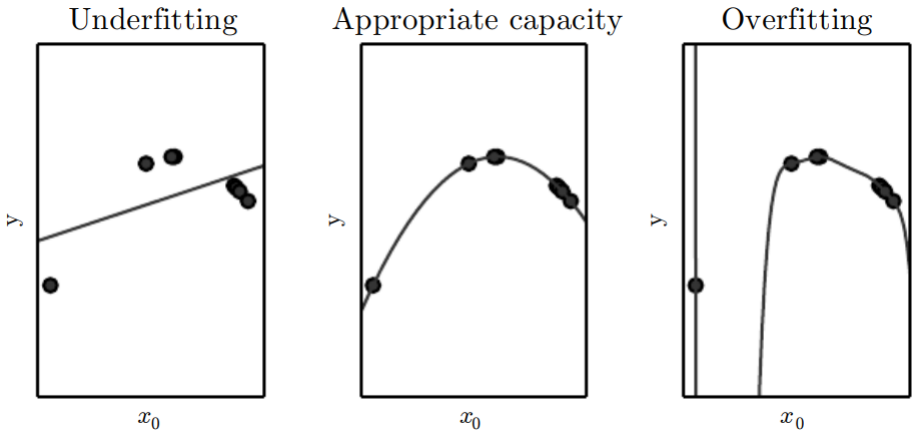
2.1.2 Przeuczenie i niedouczenie

Wystarczająco skomplikowany model jest w stanie osiągnąć bardzo dobre wyniki dla danych, na których był trenowany. Jednak głównym celem uczenia maszynowego jest stworzenie systemu, który będzie radził sobie z niezaistniałymi wcześniej wejściami. Jest to proces tak zwanej generalizacji (ang. *generalization*) [18].

To w jakim stopniu model może generalizować, zależy od jego pojemności (ang. *capacity*). Jest to miara opisująca umiejętność dopasowania modelu do różnorodnych funkcji. Model o zbyt małej pojemności nie jest w stanie dopasować się do zbioru treningowego. Z drugiej strony zbyt duża pojemność doprowadza do nauki zależności, które źle wpływają na wyniki dla nowych danych [18].

W procesie uczenia maszynowego celem jest minimalizacja błędu na zbiorze treningowym (ang. *training error*). Takie podejście może jednak doprowadzić do stworzenia modelu, który będzie wykazywał nadmierne dopasowanie (ang. *overfitting*), zwane również przeuczeniem. Powoduje to, że system nie jest w stanie poradzić sobie z danymi spoza zbioru treningowego i daje złe wyniki [18].

Chcemy zatem stworzyć model, którego celem nie będzie tylko minimalizacja błędu treningowego. Aby tego dokonać, zazwyczaj wydziela się pewien fragment ze zbioru wartości wejściowych, który zostaje użyty do oceny modelu, jest to tak zwany zbiór testowy. Nowym celem procesu uczenia jest minimalizacja błędu treningowego, ale



Rysunek 2.1: Wizualna reprezentacja wyników modelu niedouczonego (lewa strona), modelu posiadającego dobry stopień generalizacji (środek) oraz modelu przuczonego (prawa strona). *Źródło: [18]*

z zachowaniem jak najmniejszej różnicy między błędem treningowym a błędem testowym (ang. *test error*), który jest obliczany dla zbioru testowego [18].

Sytuację przeciwną, w której model nie jest wystarczająco skomplikowany i niemożliwe jest uzyskanie niskiego błędu treningowego, nazywa się niedouczeniem (ang. *underfitting*) [18].

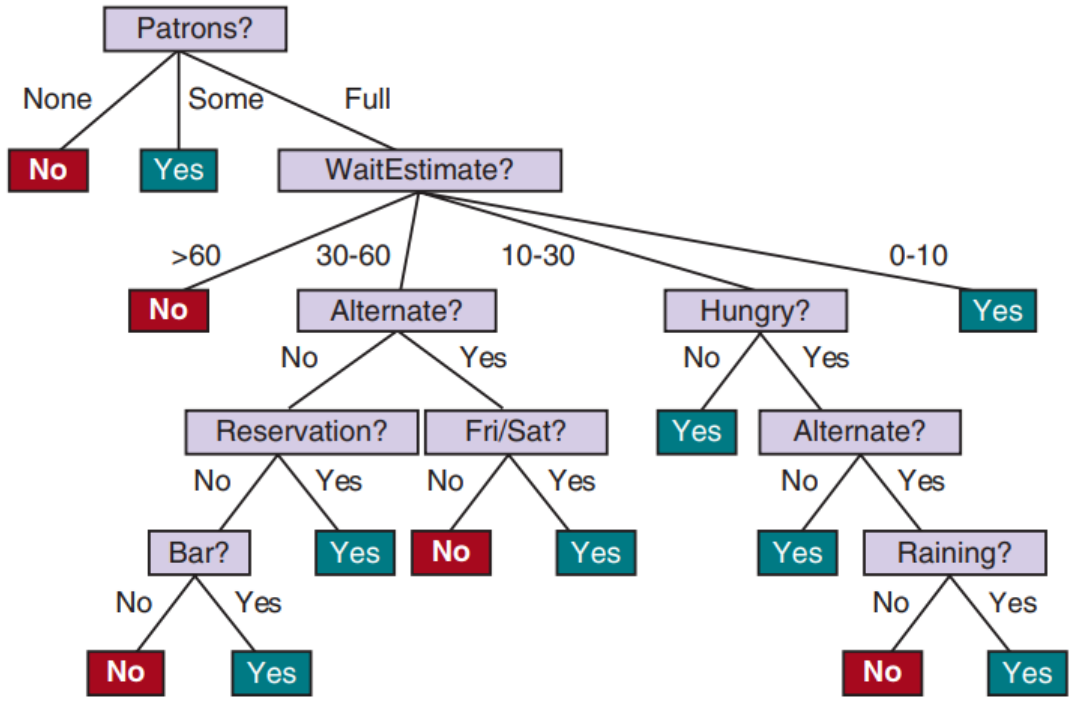
2.2 Przykładowe algorytmy uczenia maszynowego

2.2.1 Drzewa decyzyjne i lasy losowe

Jedną z najprostszych reprezentacji procesu podejmowania decyzji są drzewa decyzyjne (ang. *decision tree*). Przedstawiają one drzewa, czyli spójny, nieskierowany, acykliczny graf, którego wierzchołki przedstawiają pewien test, a liście są decyzjami. Decyzja jest zatem uzyskiwana dzięki wykonaniu szeregu testów na danych wejściowych, aż do momentu dojścia do liścia [4].

W procesie uczenia algorytm stosuje zazwyczaj zachłannie metodę dziel i zwyciężaj. Jako pierwsze do testów używa się wartości, które mają największy wpływ na klasyfikację, czyli dają największy przyrost informacji [4]. Można to obliczyć przy pomocy entropi Shannona lub Ginni index. Najpowszechniej stosowane algorytmy tworzące drzewa decyzyjne to ID3, C4.5 oraz CART [4, 19].

Drzewa decyzyjne są jednak bardzo skłonne do nadmiernego dopasowania. W procesie uczenia drzewo może stworzyć tyle wierzchołków, że będzie w stanie przypisać każdą wartość w zbiorze treningowym do osobnego liścia. Taki model nie będzie w stanie generalizować, co doprowadzi do złych wyników dla nowych danych. Rozwiązaniem tego problemu jest proces zwany przycinaniem drzewa (ang. *pruning*). Działa



Rysunek 2.2: Przykładowe binarne (posiada tylko dwie możliwe decyzje) drzewo decyzyjne. Wierzchołki, oznaczone kolorem fioletowym, reprezentują testy. Przy krawędziach zaznaczono wynik testu, który powoduje jej wybranie. Liście, czyli decyzje, są oznacone kolorem niebieskim - Tak (Yes) oraz czerwonym - Nie (No). *Źródło: [4]*

on na zasadzie usuwania wierzchołków, których dziećmi są tylko liście, i które nie są statystycznie znaczące [4].

Dużą zaletą drzew decyzyjnych jest prostota ich interpretacji, bardzo łatwo jest je przedstawić wizualnie. Można je stosować zarówno do regresji, jak i do klasyfikacji, potrafią nawet przypisywać kilka klas dla jednego wektora wejściowego. Dodatkowo są one w stanie dopasować się do dużych zbiorów, nie potrzebują normalizacji oraz są stosunkowo szybkie. Główną wadą drzew decyzyjnych jest jednak ich niestabilność. Małe różnice w danych wejściowych mogą prowadzić do dużych zmian w finalnej strukturze drzewa oraz wynikach, które dają [19].

Jednym ze sposobów zmniejszenia wariancji jest stworzenie wielu modeli i podejmowanie decyzji na podstawie ich odpowiedzi, jest to tak zwane ensemble learning. Wyniki są zazwyczaj uśredniane lub przeprowadza się głosowanie, w którym ostateczną odpowiedzią jest ta, którą zwraca największa liczba drzew [4]. W przypadku drzew decyzyjnych takie podejścia nazywa się lasami losowymi (ang. *random forest*). Każde z drzew w procesie wyboru nie bierze pod uwagę wszystkie n atrybutów a jedynie pewną ich część, zazwyczaj $\sqrt{n}$. Nadal wybierany jest ten atrybut, który daje jak największy przyrost informacji [4].

Popularne są również extremely randomized trees (ExtraTrees), które dodają kolejny element losowości. Zamiast szukać wartości progowej, która daje największy przyrost informacji, jest ona wybierana z rozkładu jednostajnego danego atrybutu [4].

2.2.2 Maszyna wektorów nośnych

Maszyny wektorów nośnych, w skrócie SVM (ang. *Support Vector Machine*), to rodzina algorytmów uczenia maszynowego służąca do klasyfikacji oraz regresji. W procesie uczenia algorytm tworzy granice (ang. *decision boundary*), będące hiperpłaszczyznami, w taki sposób, aby ich odległość od punktów była jak największa. Następnie na ich podstawie zwracany jest wynik [4].

Podstawowa wersja SVM opiera się na regresji liniowej. Można ją opisać wzorem:

$$y = w^\top x + b,$$

gdzie $\top$ oznacza transpozycję, b to bias, który zawsze wynosi 1, x jest wektorem danych wejściowych, a $w^\top$ jest wektorem wag, które są zmieniane w trakcie nauki [18].

Maszyny wektorów nośnych pozwalają również na tworzenie modeli nieliniowych dzięki tak zwanemu kernel trick. Wynika on z możliwości algorytmów uczenia maszynowego, które mogą być napisanie wyłącznie jako iloczyny skalarne między przykładami. Dzięki temu powyższa liniowa funkcja może być zapisana jako:

$$w^\top x + b = b + \sum_{i=1}^{m} \alpha_i x^\top x^{(i)},$$

gdzie α jest wektorem wag, a $x^{(i)}$ przykładem treningowym. Następnie x może być zamienione na wynik funkcji $\phi(x)$, a iloczyn skalarny na funkcję $k(x, x^{(i)}) = \phi(x)^\top \phi(x^{(i)})$, nazywaną kernelem. Daje to ostateczną funkcję [18]:

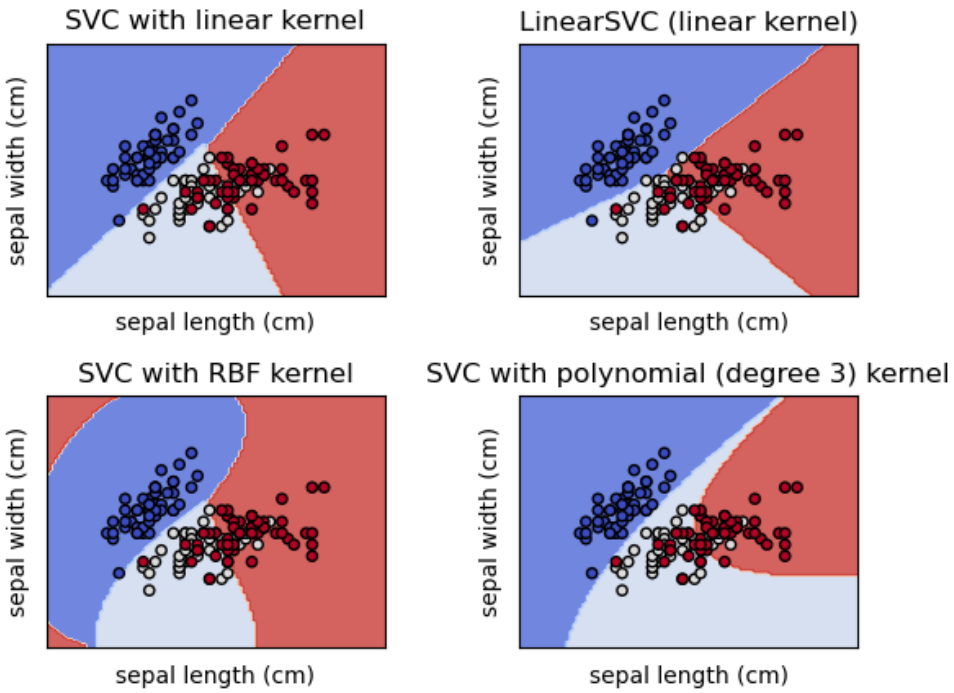
$$f(x) = b + \sum_i \alpha_i k(x, x^{(i)}).$$

Istnieje wiele kerneli, jednym z najpopularniejszych jest radial basis function (RBF). Ma on następujący wzór:

$$k(u, v) = \mathcal{N}(u - v; 0, \sigma^2 I)$$

gdzie $\mathcal{N}(x; \mu, \Sigma)$ oznacza funkcję gęstości prawdopodobieństwa rozkładu normalnego. Wartość funkcji maleje w przestrzeni v wraz z oddalaniem się od punktu u [18].

Jedną z głównych zalet SVM jest to, że wagi α wynoszą 0 z wyjątkiem wektorów wzmacniających (ang. *support vectors*), które są punktami najbliższymi obliczonym hiperpłaszczyznom. Pozwala to na przyśpieszenie obliczeń i ograniczenie zużycia pamięci. Mimo to maszyny wektorów nośnych znacząco tracą na wydajności w przypadku dużych zbiorów danych [4, 18].



Rysunek 2.3: Wizualna reprezentacja granic wygenerowanych przez SVM wykorzystujące różne kernele. *Źródło: https://scikit-learn.org/stable/modules/svm.html (dostęp 26.05.2023).*

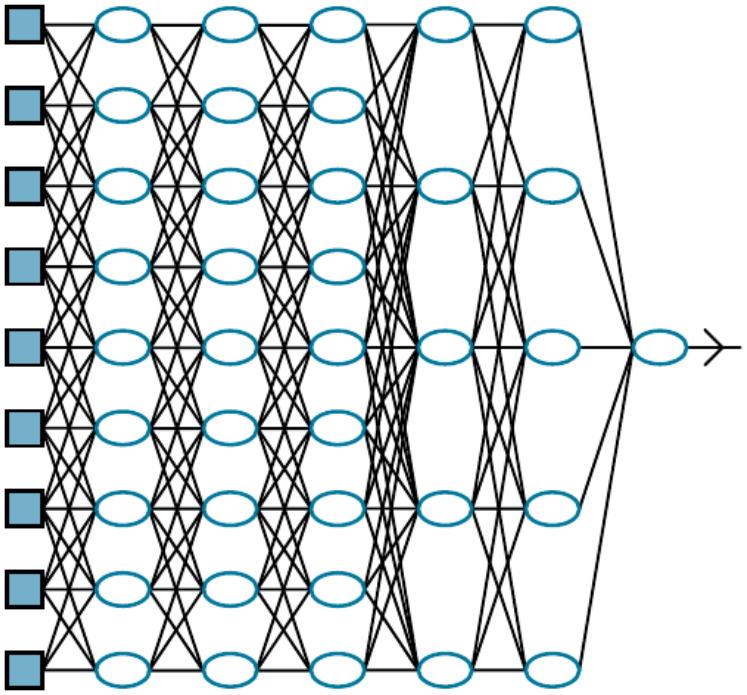
2.3 Sztuczne sieci neuronowe

Budowa ludzkiego mózgu oraz sposób działania komórek nerwowych stały się źródłem inspiracji do stworzenia systemów, które potrafiłyby się uczyć. Doprowadziło to do stworzenia modeli zwanych dzisiaj sieciami neuronowymi [4].

2.3.1 Głębokie sieci neuronowe

Współcześnie najpopularniejszym rodzajem sieci neuronowych są tak zwane sieci głębokie, których nazwa pochodzi od wykorzystywania wielu warstw w modelu. Składają się one z neuronów, które imitują działanie biologicznych komórek nerwowych. Każdy neuron otrzymuje sygnały od wielu innych neuronów oraz sam generuje sygnał po przekroczeniu pewnej wartości. Neurony są grupowane w warstwy, które można traktować jako osobne funkcje. Liczbę neuronów w warstwie nazywa się szerokością modelu, a ilość warstw w modelu głębokością. Sieć jest zatem złożeniem funkcji: $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$, gdzie $f^{(1)}$ oznacza pierwszą warstwę, $f^{(2)}$ drugą itd. Pierwsza warstwa $f^{(1)}$ jest nazywana warstwą wejściową, ostatnia warstwa $f^{(n-1)}$ to warstwa wyjściowa, pozostałe nazywane są warstwami ukrytymi [18].

Sieci zazwyczaj są jednokierunkowe (ang. *feedforward neural networks*). Przepływ sygnałów odbywa się tylko od warstwy wejściowej, poprzez warstwy ukryte, aż do warstwy wyjściowej. Sieci, w których sygnały (końcowy wynik lub wyniki z warstw ukrytych) przesyłane są w obie strony nazywa się sieciami rekurencyjnymi [18].



Rysunek 2.4: Przykładowa jednokierunkowa sieć składająca się z siedmiu warstw. Kwadraty po lewej stronie oznaczają neurony warstwy wejściowej, wyjściem jest pojedyńczy neuron po prawej stronie. *Źródło: [4].*

2.3.2 Funkcje aktywacji

Wartość, jaką neuron przekazuje do kolejnej warstwy, można wyrazić następującym wzorem:

$$a_j = g_j(\Sigma_i w_{i,j} a_i),$$

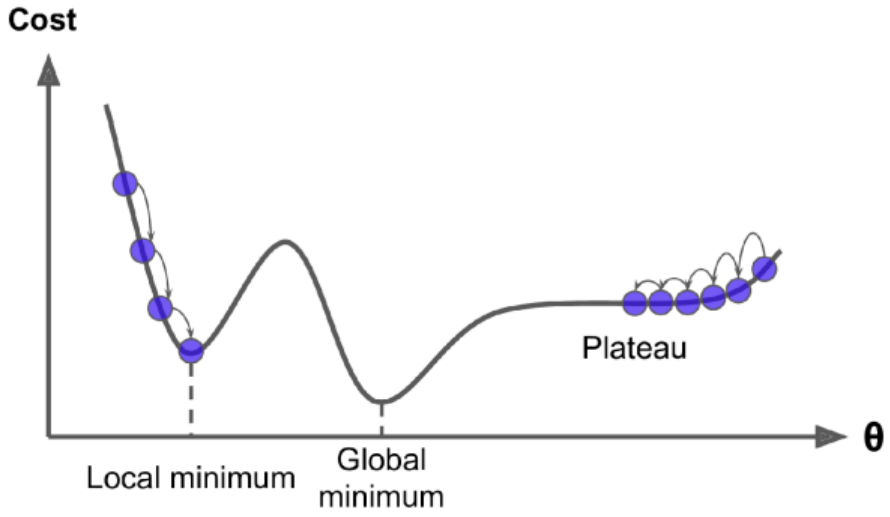
gdzie a_j oznacza neuron j, $w_{i,j}$ wagę połączenia między neuronem i oraz j, natomiast g_j to nieliniowa funkcja aktywacji. Nieliniowość pozwala na odwzorowanie dowolnej funkcji przez wystarczająco skomplikowaną sieć [4].

Do najpopularniejszych funkcji aktywacji należą [4]:

- ReLU (rectified linear unit): $ReLU(x) = \max(0, x)$,
- sigmoid (logistic function): $\sigma(x) = 1/(1 + e^{-x})$,
- tangens hiperboliczny: $\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}$.

2.3.3 Gradient descent

Za uczenie sieci neuronowych odpowiadają algorytmy oparte na metodzie gradientu prostego (ang. *gradient descent*). Metoda ta pozwala na znalezienie lokalnego minimum w przestrzeni dzięki wykonywaniu małych kroków w jego kierunku. Na początku wybiera się losowo punkt należący do danej przestrzeni, następnie oblicza się gradienty i przesuwa w kierunku największego spadku, aż do momentu dojścia do punktu z minimalną wartością funkcji straty (ang. *loss function* lub *cost function*). Odległość o jaką przesuwany jest punkt nazywa się learning rate [4].



Rysunek 2.5: Wizualizacja przykładowego działania metody gradient descent. Po lewej stronie algortym znajduje minimum lokalne, po prawej stronie algorytm natrafia na bardzo małe spadki i może zostać zatrzymany przed znalezieniem globalnego minimum. *Źródło: [3].*

Pojedynczy krok algorytmu można zatem zapisać następująco:

$$w \leftarrow w - \alpha \nabla_w L(w),$$

gdzie w to parametry sieci, α to learning rate, L to funkcja straty [4].

W klasyfikacji jako funkcja straty często stosowana jest entropia krzyżowa (ang. *cross-entropy loss*). Jej ogólny wzór wygląda następująco:

$$H(P, Q) = \int P(x) \log Q(x) dx,$$

gdzie P oznacza prawdziwe wartości zbioru testowego $P^*(x, y)$, a Q wartości przewidziane przez model $P_w(y|x)$. Celem uczenia jest zmiana w tak, aby zminimalizować $H(P^*(x, y), P_w(y|x))$ [4].

Często stosowany jest szybszy wariant algorytmu gradientu prostego zwany stochastic gradient descent, w skrócie SGD. W odróżnieniu od zwykłego algorytmu,

w każdej iteracji losowo wybierana jest niewielka liczba wartości, zamiast całego zbioru treningowego. Pozwala to na znaczne przyśpieszenie obliczeń [4].

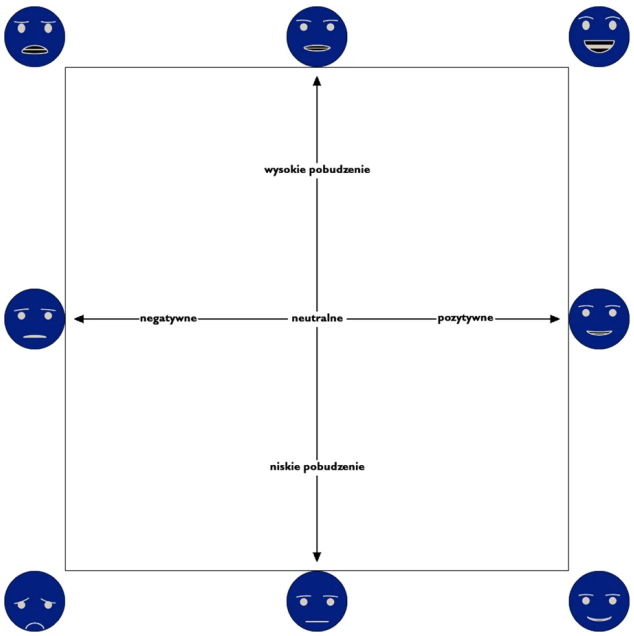
Innym popularnym algorytmem jest Adam, którego nazwa pochodzi od wyrażenia adaptive moments. W trakcie działania Adam dynamicznie zmienia learning rate oraz momentum. Momentum sprawia, że punkt dodatkowo przesuwa się w kierunku opartym na średniej ruchomej poprzednich przesunięć [18].

Rozdział 3

Część praktyczna

3.1 Zbiór danych

W pracy wykorzystano gotowy zbiór danych o nazwie BIRAFFE2 [20]. Zawiera on zapisy elektrokardiografii (EKG), reakcji skórno-galwanicznej (EDA), wyrazów twarzy i ruchu dłoni, które zostały nagrane podczas prób wywołania emocji przez stymulanty audiowizualne i specjalnie przygotowane gry komputerowe. Dodatkowo w zbiorze zawarto subiektywną ocenę stymulantów w dwuwymiarowej przestrzeni przyjemności i pobudzenia (ang. *valence, arousal*), wyniki testu osobowości opartego o tak zwaną wielką piątkę (ang. *big five*) oraz ankiety o doświadczeniu z grami komputerowymi. Dane pochodzą od 102 osób w wieku od 18 do 26 lat, z czego 33% badanych to kobiety [20].



Rysunek 3.1: Widżet użyty do subiektywnej oceny stymulantów. *Źródło: [20]*

Stymulanty audiowizualne prezentowane były w dwóch turach, z sesją gry komputerowej pomiędzy turami. Każdy stymulant prezentowany był przez 6 sekund, po czym badany miał 6 sekund na ocenę wywołanych emocji i następowało kolejne 6 sekund przerwy [20]. Wizualne stymulanty wybrano ze zbioru IAPS [21], a dźwiękowe ze zbioru IADS [22].

W niniejszej pracy wykorzystano jedynie zapisy EKG oraz EDA z obu tur prezentacji stymulantów.

3.2 Przygotowanie danych

3.2.1 Oczyszczanie i ekstrakcja cech

Cały system automatycznej predykcji emocji, opisany w tej pracy, został napisany w języku Python. Aby oczyścić dane i dokonać ekstrakcji cech (ang. *feature extraction*) użyto biblioteki NeuroKit[1] [23]. Zawiera ona wiele funkcji i narzędzi pozwalających na pracę z sygnałami biofizycznymi.

Dane zostały podzielone na okienka o długości 18 sekund, co odpowiada pojawieniu się pojedynczego stymulanta audiowizualnego, czasu na subiektywną ocenę emocji oraz przerwie przed kolejnym stymulantem. Wartości odpowiadające treningowi nie były brane pod uwagę.

Sygnały EKG były poddawane oczyszczaniu funkcją `ecg_clean()` z wykorzystaniem metody zaproponowanej przez Pana i Tompkinsa [24]. Następnie znajdowano załamki R w zespole QRS, wykorzystując metodę zaproponowaną w tym samym artykule oraz funkcję `ecg_peaks()`. Na ich podstawie obliczano średnią częstotliwość występowania załamków funkcją `ecg_rate()` oraz wartości związane ze zmiennością rytmu zatokowego (ang. *heart rate variability, HRV*) stosując `hrv_time()` oraz `hrv_frequency()`.

Podobnie jak EKG, sygnał EDA był na początku oczyszczany i wydzielono z niego tonic component, użyto do tego funkcję `eda_process()`. Następnie obliczono ilość wystąpień reakcji oraz ich średnią amplitudę wykorzystując funkcję `eda_intervalrelated()`. Kolejnym krokiem było obliczenie standardowego odchylenia dla tonic component. Następnie wykorzystano `eda_sympathetic()` aby uzyskać wartości związane z sympathetic component, czyli wartościami w zakresie 0,0045 - 0,25 Hz [25]. Na koniec obliczono autokorelację sygnału stosując `eda_autocorr()`.

W kolejnym kroku zastosowano powyższe metody dla sygnałów z przedziału od pierwszego do ostatniego stymulanta, które potraktowano jako średnią wartość, unikalną dla każdego badanego. Następnie odejmowano wartości uzyskane w każdym z okienek od średniej danej osoby. Miało to na celu uzyskanie danych o zmianie stanu badanego podczas oglądania stymulanta względem normy.

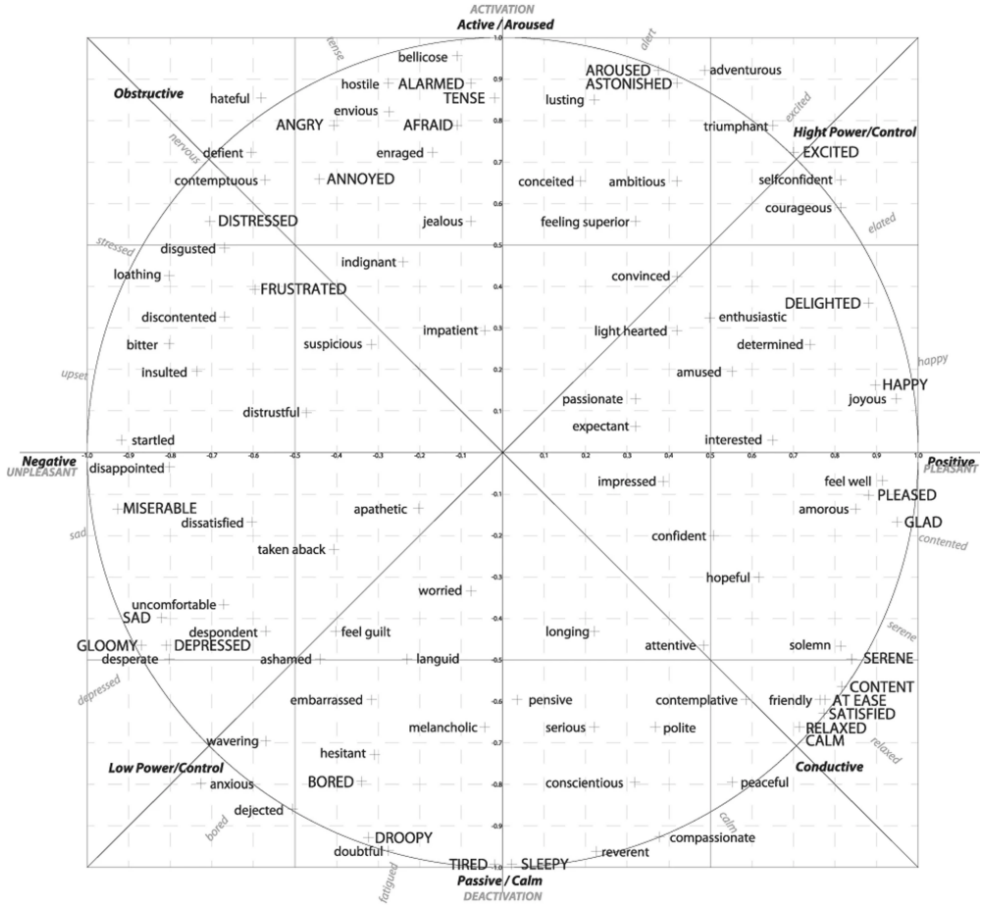
[1] https://neuropsychology.github.io/NeuroKit/

Dla każdego okienka przypisano odpowiadające mu dwie wartości uzyskane przez subiektywną ocenę. Były to przyjemność emocji (ang. *valence*) oraz pobudzenie (ang. *arousal*) jakie wywołały.

3.2.2 Grupowanie

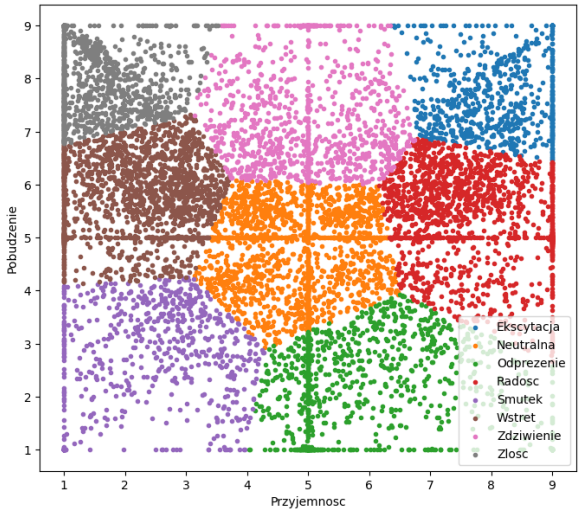
Po uzyskaniu cech przeprowadzono proces grupowania (ang. *clustering*), w celu zmiany problemu z regresji do klasyfikacji dla kilku klas. Wykorzystano do tego algorytm K-Means, który jest przykładem uczenia nienadzorowanego i został zaproponowany przez Lloyda [26]. Sama użyta funkcja `KMeans()` pochodzi z biblioteki scikit-learn[2] [27].

Po obliczeniu centroidów oraz uzyskaniu grup ręcznie przypisano im emocje na podstawie modelu kołowego z [28]. W pracy stworzono modele dla 8, 6 i 4 emocji.

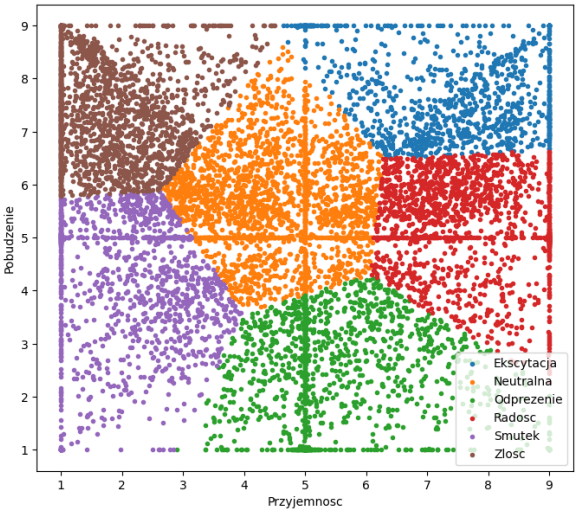


Rysunek 3.2: Model kołowy użyty do przypisania emocj do grup. *Źródło: [28]*

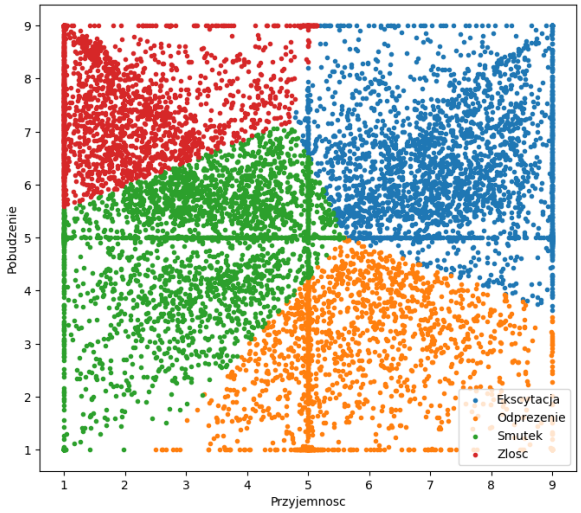
[2]https://scikit-learn.org/



(a) 8 emocji



(b) 6 emocji



(c) 4 emocje

Rysunek 3.3: Uzyskane grupy i przypisane im emocje.

3.3 Wyniki

Porównanie wyników różnych modeli scikit-learn i TensorFlow oraz modalności, kilka tabelek.

Col1	Col2	Col2	Col3
1	6	87837	787
2	7	78	5415
3	545	778	7507
4	545	18744	7560
5	88	788	6344

Tabela 3.1: Table to test captions and labels.

Podsumowanie

Podsumowanie...

Bibliografia

[1] Ashwini Ann Varghese, Jacob P Cherian, and Jubilant J Kizhakkethottam. Overview on emotion recognition system. 2015. doi:10.1109/ICSNS.2015.7292443.

[2] Andrius Dzedzickis, Arturas Kaklauskas, and Vytautas Bučinskas. Human emotion recognition: Review of sensors and methods. *Sensors*, 20:592, 01 2020. doi:10.3390/s20030592.

[3] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc., 2019. ISBN 9781492032649.

[4] Stuart J Russell and Peter Norvig. *Artificial intelligence a modern approach*. Pearson, 2020.

[5] Rafael Calvo, Sidney D'Mello, Jonathan Gratch, and Arvid Kappas. *The Oxford Handbook of Affective Computing*. Oxford University Press, 2015. doi:10.1093/oxfordhb/9780199942237.001.0001.

[6] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljevic, Daniel Wilde, Stephan Garbin, Chirag Raman, Jamie Shotton, Toby Sharp, Ivan Stojiljkovic, Tom Cashman, and Julien Valentin. 3D face reconstruction with dense landmarks. 2022. doi:10.48550/ARXIV.2204.02776.

[7] Byoung Chul Ko. A brief review of facial emotion recognition based on visual information. *Sensors*, 18(2), 2018. doi:10.3390/s18020401.

[8] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. doi:10.1037/t27734-000.

[9] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing*, 12(2):505–523, 2021. doi:10.1109/TAFFC.2018.2874986.

[10] Andrea Kleinsmith and Nadia Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4:15–33, 01 2013. doi:10.1109/T-AFFC.2012.16.

[11] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. A comprehensive review of speech emotion recognition systems. *IEEE Access*, 9:47795–47814, 2021. doi:10.1109/ACCESS.2021.3068045.

[12] Mohammed Abdelwahab and Carlos Busso. Evaluation of syllable rate estimation in expressive speech and its contribution to emotion recognition. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 472–477. IEEE, 2014. doi:10.1109/SLT.2014.7078620.

[13] João Teixeira, Carla Oliveira, and Carla Lopes. Vocal acoustic analysis – jitter, shimmer and HNR parameters. *Procedia Technology*, 9:1112–1122, 12 2013. doi:10.1016/j.protcy.2013.12.124.

[14] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. doi:10.1109/CVPR.2017.277.

[15] Paul Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press, 1971.

[16] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. doi:10.1037/h0077714.

[17] Sidney D'Mello and Jacqueline Kory. Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, page 31–38. Association for Computing Machinery, 2012. doi:10.1145/2388676.2388686.

[18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL `deeplearningbook.org`.

[19] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL `probml.ai`.

[20] Krzysztof Kutt, Dominika Drążyk, Laura Żuchowska, Maciej Szelążek, Szymon Bobek, and Grzegorz J Nalepa. BIRAFFE2, a multimodal dataset for emotion-based personalization in rich affective game environments. *Scientific Data*, 9(1): 274, 2022. doi:https://doi.org/10.1038/s41597-022-01402-6.

[21] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. 2005.

[22] Margaret M Bradley and Peter J Lang. The international affective digitized sounds (IADS-2): Affective ratings of sounds and instruction manual. 2007.

[23] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, feb 2021. doi:10.3758/s13428-020-01516-y.

[24] Jiapu Pan and Willis J Tompkins. A real-time qrs detection algorithm. *IEEE transactions on biomedical engineering*, (3):230–236, 1985.

[25] Hugo F Posada-Quintero, John P Florian, Alvaro D Orjuela-Cañón, Tomas Aljama-Corrales, Sonia Charleston-Villalobos, and Ki H Chon. Power spectral density analysis of electrodermal activity for sympathetic function assessment. *Annals of biomedical engineering*, 44:3124–3135, 2016.

[26] Stuart Lloyd. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[28] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6-7):907–929, 2019.

Spis rysunków

1.1 Ogólny schemat systemu predykcji emocji 5
1.2 Schemat systemu rozpoznającego emocje na podstawie twarzy. *Źródło: [7]* 7
1.3 Przykłady różnych Action Units w trzech częsciach twarzy. *Źródło: [7]* 7
1.4 Sposoby reprezentowania ciała w komputerze: zbiór części ciała (lewa strona) oraz reprezentacja szkieletowa (prawa strona). *Źródło: [9]* . . 9
1.5 Przykładowy sygnał mowy z zaznaczonymi jitter i shimmer. *Źródło: [13]* 10
1.6 Przykładowy sygnał EKG z zaznaczonymi załamkami. *Źródło: [2]* . . 11
1.7 Przykładowy sygnał GSR. Czerwone linie oznaczają momenty pojawiania się stymulantu. *Źródło: [2]* 12
1.8 Przykładowa baza danych zawierająca zdjęcia twarzy przedstawiające podstawowe emocje. *Źródło: [14]* 13
1.9 Kołowy model oparty o torię Russella. *Źródło: [2]* 14

2.1 Wizualna reprezentacja wyników modelu niedouczonego (lewa strona), modelu posiadającego dobry stopień generalizacji (środek) oraz modelu przuczonego (prawa strona). *Źródło: [18]* 18
2.2 Przykładowe binarne (posiada tylko dwie możliwe decyzje) drzewo decyzyjne. Wierzchołki, oznaczone kolorem fioletowym, reprezentują testy. Przy krawędziach zaznaczono wynik testu, który powoduje jej wybranie. Liście, czyli decyzje, są oznacone kolorem niebieskim - Tak (Yes) oraz czerwonym - Nie (No). *Źródło: [4]* 19
2.3 Wizualna reprezentacja granic wygenerowanych przez SVM wykorzystujące różne kernele. *Źródło: https://scikit-learn.org/stable/modules/svm.html (dostęp 26.05.2023).* 21
2.4 Przykładowa jednokierunkowa sieć składająca się z siedmiu warstw. Kwadraty po lewej stronie oznaczają neurony warstwy wejściowej, wyjściem jest pojedyńczy neuron po prawej stronie. *Źródło: [4].* 22
2.5 Wizualizacja przykładowego działania metody gradient descent. Po lewej stronie algortym znajduje minimum lokalne, po prawej stronie algorytm natrafia na bardzo małe spadki i może zostać zatrzymany przed znalezieniem globalnego minimum. *Źródło: [3].* 23

3.1 Widżet użyty do subiektywnej oceny stymulantów. *Źródło: [20]* . . . 25

3.2 Model kołowy użyty do przypisania emocj do grup. *Źródło: [28]* . . . 27
3.3 Uzyskane grupy i przypisane im emocje. 28

Spis tabel

3.1 Table to test captions and labels. 29