Autoreferat

1 Imię i nazwisko

Neo Christopher Chung

2 Posiadane dyplomy i stopnie naukowe

Stopień doktora (ang. Ph.D.) w dziedzinie Biologii Ilościowej i Obliczeniowej (ang. Quantitative and Computational Biology; Q.C.B.), *Uniwersytet Princeton* (USA) **2014**

Rozprawa: *Statistical Inference of Variables Driving Systematic Variation in High Dimensional Biological Data*
Promotor: Prof. John D. Storey

Magisterium (ang. M.A.) z Biologii Ilościowej i Obliczeniowej (ang. Quantitative and Computational Biology; Q.C.B.), *Uniwersytet Princeton* (USA) **2011**

Licencjat inżyniera (ang. B.S.E.) w Inżynierii Biomedycznej (ang. Biomedical Engineering; B.M.E.), *Uniwersytet Duke'a* (USA) **2009**

Licencjat niewielki (ang. B.A. minor) w studiów wizualnych (ang. Visual Studies), *Uniwersytet Duke'a* (USA) **2009**

3 Zatrudnienie w jednostkach naukowych

Adiunkt, Informatyka **od 10/2018***
Uniwersytet Warszawski, Wydział Matematyki, Informatyki i Mechaniki
*Data rozpoczęcia; staż na adiunkcie naukowym (10.2018 - 8.2019) w ramach grantu SONATA

Adiunkt naukowy, Informatyka **7/2016 – 8/2019**
Uniwersytet Warszawski, Wydział Matematyki, Informatyki i Mechaniki

Wykonawca, Informatyka **2/2016 – 7/2016**
Uniwersytet Warszawski, Wydział Matematyki, Informatyki i Mechaniki

Współpracownik naukowy **1/2020 – 6/2022**
NHLBI Integrated Cardiovascular Data Science Training Program
Zakłady Fizjologii i Medycyny (Kardiologia)
David Geffen School of Medicine, Uniwersytet Kalifornijski w Los Angeles (USA)

Wizytujący naukowiec **1/2018 – 1/2020**
NIH BD2K Center of Excellence for Biomedical Computing
Zakłady Fizjologii i Medycyny (Kardiologia)
David Geffen School of Medicine, Uniwersytet Kalifornijski w Los Angeles (USA))

Profesor wizytujący, **10/2015 – 2/2016**
Grupa Biostatystyczna, Katedra genetyki
Uniwersytet Przyrodniczy we Wrocławiu

NIH UJMT Fogarty Global Health Fellowship **9/2014 – 7/2015**
Centrum Badań nad Chorobami Zakaźnymi w Zambii (CIDRZ)
Instytut Globalnego Zdrowia i Chorób Zakaźnych na Uniwersytecie Karoliny Północnej w Chapel Hill (USA)

4 Opis osiągnięć

Cykl powiązanych tematycznie artykułów naukowych pod tytułem:

Wnioskowanie i wyjaśnialność modelu uczenia maszynowego.

Należy pamiętać, że znak funta (#) oznacza studenta lub naukowca podoktoranckiego pod moim bezpośrednim nadzorem i finansowaniem. Korespondujący autorzy są oznaczeni sztyletami (†). W razie potrzeby, wspólni pierwsi autorzy są oznaczeni gwiazdkami ($\star$).

A: Chung, N.C.†, Błażej M., Startek, M., and Gambin, A. (2019) Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. *BMC Bioinformatics* **20**(15)

B: Chung, N.C. (2020) Statistical significance of cluster membership for unsupervised evaluation of cell identities. *Bioinformatics* **36**(10)

C: Brocki, L.$^{\#}$ and Chung, N.C.† (2020) Concept Saliency Maps to Visualize Relevant Features in Deep Generative Models. *IEEE International Conference On Machine Learning And Applications (ICMLA)*

D: Brocki, L.$^{\#}$; Chung, N.C.† (2023) Fidelity of Interpretability Methods and Perturbation Artifacts in Neural Networks. International Conference on Learning Representations (ICLR) Tiny Paper

E: Brocki, L.$^{\#}$ and Chung, N.C.† (2023) Feature Perturbation Augmentation for Reliable Evaluation of Importance Estimators in Neural Networks. *Pattern Recognition Letters* **176**
*Wcześniejsza wersja podczas *Trustworthy ML Workshop* na *International Conference on Learning Representations (ICLR)* w 2023 r.

F: Brocki, L.$^{\#\star}$, Marchadour, W.*, Maison, J., Badic, B., Papadimitroulas, P., Hatt, M.†, Vermet, F.†, Chung, N.C.† (2022) Evaluation of importance estimators in deep learning classifiers for Computed Tomography. *4th International Workshop on Explainable and Transparent AI and Multi-Agent Systems (EXTRAAMAS)* at the *2022 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*

G: Brocki, L.$^{\#}$, Binda, J.$^{\#}$, and Chung, N.C.† (2024) Class-Discriminative Attention Maps for Vision Transformers. *Transactions on Machine Learning Research (TMLR)*
*Wcześniejsza wersja podczas *XAI Workshop* na *International Joint Conference on Artificial Intelligence (IJCAI)* w 2024 r.

4.1 Wstęp i podsumowanie

Rozwój uczenia maszynowego (ang. machine learning; ML) - od klasycznych metod statystycznych po głębokie sieci neuronowe (ang. deep neural networks; DNN) - koncentruje się na dużej ilości danych generowanych we wszystkich aspektach społeczeństwa. Wraz ze wzrostem ilości danych i zwiększaniem się złożoności systemów, coraz trudniej jest zidentyfikować dokładne sygnały i zrozumieć procesy podejmowanych decyzji. W szczególności potencjał ML do rozwiązywania trudnych problemów w biologii, medycynie i innych dziedzinach nauki jest częściowo ograniczony przez brak rygorystycznych metod, które pomagają w zrozumieniu danych o wysokiej wymiarowości.

Motywowani wyzwaniami związanymi z dużą ilością danych w biologii i medycynie, opracowaliśmy metody na potrzeby inferencji modeli ML oraz ich wyjaśnialności (ang. explainability.). Ogólnie rzecz biorąc w metodach ML, wejście $\mathbf{X}$ skutkuje wyjściem $\mathbf{Y} = f(\mathbf{X})$, gdzie funkcja f może ilościowo określać podobieństwo [A], tworzyć klastery [B], charakteryzować reprezentację ukrytą [B,C,G] lub przewidywać

klasy [D,E,F,G]. Zasadniczo istotne jest uzyskanie odpowiedzi na pytanie: *jak możemy dokładnie wnioskować i wyjaśniać relacje między wejściami i wyjściami?* W bardziej klasycznych problemach inferencji rozważamy, jak ilościowo określić niepewność i kontrolować poziomy błędów. Korzystając z DNN, gdzie konieczne jest przybliżenie, zadaniem jest przekształcenie modelu w taki sposób, aby miał pożądane właściwości.

W [A] badamy klasyczną miarę podobieństwa dla danych binarnych, gdzie stosunek ich przecięcia do ich sumy nazywany jest współczynnikiem Jaccarda/Tanimoto [1, 2]. Dane binarne pochodzą z szerokiego obszaru badań biologicznych, od ekologii po genomikę; a ich znaczne podobieństwo sugeruje współwystępowanie. Pomimo swojej popularności, statystyczna inferencja przy użyciu współczynnika Jaccarda/Tanimoto nie została w pełni rozwinięta. W tym celu przedstawiamy szereg rozwiązań, które były szeroko stosowane w badaniach obejmujących sekwencjonowanie RNA pojedynczych komórek (scRNA-seq), modelowanie chorób, bioróżnorodność i wiele innych.

W [B] skupiamy się na nienadzorowanym klasteryzowaniu (ang. unsupervised clustering), które odkrywa grupy podobnych danych. W przypadku scRNA-seq, które mogą mierzyć profile ekspresji genów pojedynczych komórek na dużą skalę, tożsamości komórek są często szacowane za pomocą klasteryzacji K-średnich [3, 4, 5] i pokrewnych algorytmów. Jednak trudno jest stwierdzić, czy przynależność do klastrów (np. przypisanie komórki do jednego z K klastrów) jest dokładna z powodu problemu podwójnego dopasowania (ang. double fitting problem)[1]. Opracowujemy metody pozwalające na ilościowe określenie tej niepewności, które pomagają w selekcji cech i zapewniają bardziej niezawodne klasteryzowanie.

W [C] używamy nienadzorowanego uczenia z użyciem konwolucyjnych sieci neuronowych (ang. convolutional neural networks; CNNs) do charakteryzowania i wyjaśniania ukrytych pojęć. Szczególnie głębokie modele generatywne, takie jak wariacyjne autoenkodery (ang. variational autoencoders; VAEs), mogą szacować zmienne ukryte, które mogą być użyte do identyfikacji i manipulacji złożonych pojęć (np. okulary przeciwsłoneczne lub warstwy morfologiczne). Aby zwiększyć przejrzystość działania głębokich modeli generatywnych, wprowadzamy metodę oszacowania ważności cech wejściowych (np. pikseli w obrazie)[2] w odniesieniu do ukrytych pojęć. Uzyskujemy zarówno ukryte pojęcie, jak i odpowiadające mu ważne cechy wejściowe.

W [D,E] badamy ocenę metod wyjaśnialnej sztucznej inteligencji (XAI), które szacują ważność cech wejściowych. W zadaniach z zakresu widzenia komputerowego, takich jak przewidywanie biomarkerów przy użyciu obrazów medycznych, istnieje wiele oszacowań ważności, które mogą dawać różne wyjaśnienia. Aby poradzić sobie z brakiem rzeczywistych danych referencyjnych (ang. ground truth), opracowujemy niezależną od modelu metodę umożliwiającą ilościową ocenę wiarygodności takich oszacowań ważności. Jednocześnie, bierzemy pod uwagę, że proponowana przez nas metoda oceny wykorzystuje perturbacje, które wprowadzają istotne zaburzenia w obrazach. Tego rodzaju sztucznie wprowadzone artefakty nieobecne w czasie trenowania modelu, mogą zniekształcić wyniki oceny. Z tego względu rozwijamy również technikę augmentacji danych treningowych, która łagodzi ten problem.

W [F] przyglądamy się bliżej ocenie metod XAI dla obrazów medycznych (tj. tomografii komputerowej; CT) z użyciem klasyfikatorów opartych na CNN. Oprócz wspomnianej metryki wierności, stosujemy także dodatkową ocenę skoncentrowaną na człowieku. Podczas gdy większość modeli głębokich sieci neuronowych dla widzenia komputerowego została opracowana dla obrazów naturalnych, ich potencjał dla obrazów medycznych nie został szeroko zbadany.

W [G] opracowujemy estymator ważności dla modeli o architekturze transformera wizyjnego (ang. vision transformer; ViT)- rodziny modeli widzenia komputerowego. Ze względu na wbudowany w tę architekturę mechanizm atencji (ang. attention mechanism), możliwym jest tworzenie tak zwanych map atencji (ang. attention maps). Tego rodzaju mapy są w stanie zobrazować ogólne cechy jakie ekstrahowane są przez model, ale nie uwzględniają szczegółowej informacji z ostatnich warstw transformera, dokonujących finalnej predykcji, np. klasyfikacji. Z tego powodu, opracowaliśmy dyskryminujące klasę mapy atencji, które mogą być stosowane zarówno w sposób nadzorowany, jak i nienadzorowany. Zainspirowani potrzebą wyjaśnień o dużej szczegółowości i wrażliwości na klasy dla obrazów medycznych, zastosowaliśmy nasze metody na skanach CT płuc.

Wspólnym motywem przedstawionych projektów badawczych jest: jak wnioskować i wyjaśniać sygnały ze złożonych danych o wysokiej wymiarowości, które mogą być łatwiej rozumiane przez naukowców i profesjonalistów medycznych. Podane rozwinięcia metodologiczne są silnie zainspirowane rzeczywistymi problemami w biologii i medycynie, gdzie rosnąca ilość danych wymaga nowych podejść i modeli ML,

[1] Znane też jako i blisko związane z: double dipping, circular analysis i post-selection inference.

[2] Generowanie wyjaśnień dla klasyfikacji lub charakteryzacji przez DNN i inne modele ML jest często nazywane wyjaśnialną sztuczną inteligencją (XAI). Podczas gdy istnieje kilka podejść do XAI, skupiamy się na wyjaśnieniach post hoc.

aby wyciągnąć użyteczne wnioski. Nasze metody zostały opublikowane jako otwarte pakiety źródłowe w językach R i Python.

4.2 Wnioskowanie współwystępowania w danych o obecności-absencji [A]

Podobieństwo wśród danych binarnych, takich jak obecność (1) i absencja (0) w danych biologicznych, jest często kwantyfikowane za pomocą współczynnika Jaccarda/Tanimoto [1, 2]. W szczególności analiza współwystępowania gatunków pomaga nam zrozumieć relacje ekologiczne i biologiczne między gatunkami. Dla dwóch wektorów binarnych $\mathbf{y}_i$ i $\mathbf{y}_j$ długości m, współczynnik podobieństwa Jaccarda/Tanimoto jest stosunkiem ich przecięcia do ich sumy, $T(\mathbf{y}_i, \mathbf{y}_j) = \mathbf{y}_i \cap \mathbf{y}_j / \mathbf{y}_i \cup \mathbf{y}_j$ [1, 2]. To kwantyfikowanie nakładań pozwala określić współistnienie gatunków [6, 7, 8, 9]. Jednakże współczynnik Jaccarda/Tanimoto nie posiada interpretacji probabilistycznej ani kontroli błędów statystycznych. W **[A]**zaprezentowaliśmy ramy wnioskowania do testowania podobieństwa (np. współwystępowania) w danych binarnych.

Pod modelem zerowym niezależności, zakłada się, że $\mathbf{y}_i$ i $\mathbf{y}_j$ są niezależne i o jednakowych rozkładach (tzw. i.i.d.). Są one modelowane przez rozkład Bernoulliego, z odpowiadającymi im prawdopodobieństwami wystąpienia (czyli sukcesu) p_i i $p_j \in [0, 1]$. W szczególności, dla $k = 1, \ldots, m$, $\mathbf{y}_{i,k} \sim_{i.i.d.}$ Bernoulli(p_i) oraz $\mathbf{y}_{j,k} \sim_{i.i.d.}$ Bernoulli(p_j). Ponieważ ta konwencjonalna definicja jest niezdefiniowana, jeśli oba wektory binarne zawierają tylko zera, dla których $\mathbf{y}_i \cup \mathbf{y}_j = 0$, redefiniujemy współczynnik Jaccarda/Tanimoto

$$T(\mathbf{y}_i, \mathbf{y}_j) = \begin{cases} \frac{\mathbf{y}_i \cap \mathbf{y}_j}{\mathbf{y}_i \cup \mathbf{y}_j} & \text{jeśli } \mathbf{y}_i \cup \mathbf{y}_j \neq 0 \\ \frac{p_i p_j}{p_i + p_j - p_i p_j} & \text{w przeciwnym przypadku.} \end{cases} \tag{1}$$

Postępując za definicją współczynnika podobieństwa Jaccarda/Tanimoto w równaniu (1), wyprowadzamy jego wartość oczekiwaną $\mathbb{E}[T(\mathbf{y}_i, \mathbf{y}_j)] = \frac{p_i p_j}{p_i + p_j - p_i p_j}$. To pozwala nam zdefiniować zcentrowany współczynnik Jaccarda/Tanimoto jako

$$T^c(\mathbf{y}_i, \mathbf{y}_j) = T(\mathbf{y}_i, \mathbf{y}_j) - \mathbb{E}[T(\mathbf{y}_i, \mathbf{y}_j)] \tag{2}$$

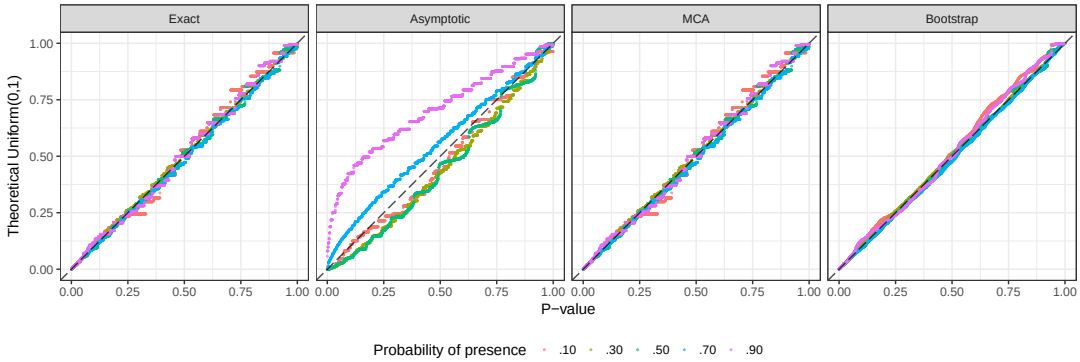
To uwzględnia wartości oczekiwane, naturalnie rozróżniając negatywne i pozytywne powiązania. Aby ocenić, czy $\mathbf{y}_i$ i $\mathbf{y}_j$ są niezależne, przeprowadzany jest następujący test statystyczny hipotez:

$$H_0 : T^c(\mathbf{y}_i, \mathbf{y}_j) = 0 \quad \text{vs.} \quad H_1 : T^c(\mathbf{y}_i, \mathbf{y}_j) \neq 0. \tag{3}$$

Warto zauważyć, że jest to równoważne z tym, iż konwencjonalny (niezcentrowany) współczynnik Jaccarda/Tanimoto równa się wartości oczekiwanej pod warunkiem niezależności. Znaczne odchylenie od wartości oczekiwanej oznacza podobieństwo. Aby uzyskać jego wartość p, wyprowadziliśmy rozkład współczynnika Jaccarda/Tanimoto pod hipotezą zerową. Sformułowane są rozwiązania dokładne i asymptotyczne. Wprowadziliśmy dwie metody estymacji do efektywnego obliczania wartości p dla współwystępowania przy użyciu danych binarnych. Po pierwsze, użyto algorytmu koncentracji miary (MCA) [10] do oszacowania krytycznego obszaru, wykorzystując obserwację, że rozkład wielomianowy koncentruje się wokół swojej dominanty. Po drugie, opracowaliśmy algorytm bootstrap w celu uzyskania istotności statystycznej przy porównywaniu podobieństwa ze współczynnikiem Jaccarda/Tanimoto.

W badaniach symulacyjnych, gdzie pewna proporcja (p) danych binarnych faktycznie współwystępuje, wykazaliśmy, jak dobrze kontrolowane są wskaźniki błędów przy użyciu proponowanych metod estymacji MCA i bootstrap w porównaniu z dokładnymi rozwiązaniami. Wykresy QQ empirycznych zerowych wartości p (nasze metody) vs. teoretyczny rozkład Uniform(0,1) są pokazane w Fig. 1). Ostatecznie, w **[A]**, przedstawiliśmy jego zastosowanie w ocenie współwystępowania ptaków na 28 wyspach Vanuatu oraz gatunków ryb w 3347 siedliskach słodkowodnych we Francji. Proponowane metody są zaimplementowane w pakiecie R o nazwie jaccard (`https://cran.r-project.org/package=jaccard`).

Poza współwystępowaniem gatunków, współczynnik Jaccarda/Tanimoto jest używany w różnych dziedzinach nauk biologicznych, gdzie dane binarne są obserwowane i porównywane. Gdy molekuły i reakcje są reprezentowane jako haszowane daktylogramy (ang. hashed fingerprints), jest on używany do ilościowych porównań i klasyfikacji [11, 12, 13]. Podobieństwo między reakcjami biochemicznymi można testować, stosując nasze metody do ich odpowiadających daktylogramów. W genomice, standardowe narzędzia, takie jak BEDTools [14], oceniają interwały genomowe za pomocą współczynników Jaccarda/Tanimoto. Dla danych interwałów genomowych z dwóch próbek lub grup, można testować, czy ich nakładanie jest statystycznie istotne, dostarczając dowodów na współdzielone zmiany genomowe. Ze względu na popularność współczynników Jaccarda/Tanimoto, proponowany zestaw metod byłby przydatny w szerokim zakresie aplikacji naukowych.

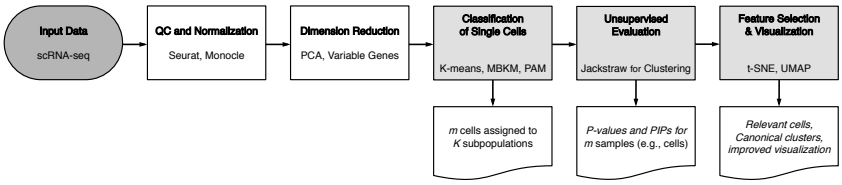


Rysunek 1: Wartości p podobieństwa pomiędzy niezależnymi wektorami obecności-absencji dla $m = 100$, z szerokim zakresem prawdopodobieństw $p = .1, .3, .5, .7, .9$. W każdym scenariuszu, 2000 niezależnych zmiennych jest symulowanych i testowanych za pomocą czterech proponowanych metod. Linie diagonalne wskazują teoretycznie poprawny rozkład Uniform(0,1).

4.3 Wnioskowanie na temat przynależności w nienadzorowanym klasteryzowaniu [B]

Klasteryzacja jest jedną z najpopularniejszych, nienadzorowanych metod uczenia maszynowego do charakteryzowania systematycznych wzorców w danych. W genomice stosuje się ją do identyfikacji współregulowanych podzbiorów genów [15, 16, 17] oraz subpopulacji wśród próbek [18, 19, 20]. Sekwencjonowanie RNA pojedynczych komórek (single-cell RNA sequencing; scRNA-seq) umożliwiło jednoczesny pomiar ekspresji genów w tysiącach pojedynczych komórek [21, 22, 23]. Tożsamości pojedynczych komórek zazwyczaj nie są znane *a priori* i często są charakteryzowane poprzez nienadzorowane klasteryzowanie. Klasteryzowanie m komórek do K subpopulacji zapewnia obliczeniowo zdefiniowane tożsamości m komórek. Te tożsamości komórkowe, oparte na klasteryzacji, są przedmiotem dużego zainteresowania, gdyż złożone fenotypy i choroby mogą wykazywać nieznane dotąd sygnatury molekularne. Biorąc pod uwagę, że tożsamości komórek są określane w sposób nienadzorowany, konieczna jest ocena, czy są poprawnie przypisane. Opracowaliśmy nowe metody szacowania istotności statystycznych i prawdopodobieństw włączenia *a posteriori* (PIP) przypisywania tożsamości komórkowych do oszacowanych subpopulacji. Poprzez uczenie się niepewności stosowania klasteryzacji do danych scRNA-seq, proponowane metody umożliwiają nienadzorowaną ocenę przynależności w klastrach, takich jak tożsamości komórkowe.

Obserwowane dane $\mathbf{Y}_{(m,n)}$ zawierają m wierszy i n kolumn. W danych scRNA-seq zakładamy, że próbki pojedynczych komórek są ułożone w wierszach, a kolumny to zmienne genetyczne (np. geny). Różnorodne narzędzia [24, 25, 26, 27] są używane do kontroli jakości i normalizacji (Fig. 2). Co więcej, redukcja wymiarów może być zastosowana na zmiennych genetycznych, aby podkreślić niektóre aspekty zmienności lub biomarkerów. Dlatego też n kolumn może obejmować wszystkie dostępne geny, geny o wysokiej zmienności, główne komponenty lub inne. Niemniej jednak, kiedy kontekst jest jasno określony, należy odnieść się do n kolumn jako genów.



Rysunek 2: Schemat analizy danych scRNA-seq dla wyjaśnienia transkrypcyjnej heterogeniczności. Bez znajomości tożsamości komórek można uzyskać profile ekspresji genów pojedynczych komórek. Proponowane metody umożliwiają statystycznie rygorystyczną ocenę tożsamości komórkowych, poprawiając nienadzorowane klasteryzowanie i wybór cech.

Załóżmy, że m komórek tworzy K subpopulacji, wykazujących różne systematyczne wzorce zmienności. Dla $k = 1, ..., K$, wzajemnie wykluczający się podzbiór komórek (m_k z m) jest przypisany do

k-tego klastra. Wtedy, $\sum_{k=1}^{K} m_k = m$. Próbki w obrębie k-tego klastra wykazują systematyczną zmienność, którą można podsumować za pomocą ich środka, centroida, medoidu lub innego reprezentatywnego $\mathbf{c}_k(\mathbf{Y})$ dla $k = 1, ...K$. W klasteryzacji metodą K-means, środek jest definiowany jako średnia euklidesowa; najbliższe środki (odległość L_2) są następnie używane do klasyfikacji próbek pojedynczych komórek [3, 4, 5]. Zakładamy, że istnieją nieobserwowane środki $\mathbf{l}_k$ oraz współczynniki $\mathbf{b}_k$ dla $k = 1, ..., K$. Dane są wtedy modelowane jako:

$$\mathbf{Y}_{(m,n)} = \mathbf{B}_{(m,K)}\mathbf{L}_{(K,n)} + \mathbf{E}_{(m,n)}, \tag{4}$$

gdzie $\mathbf{E}$ to szum będący zbiorem niezależnych zmiennych losowych o identycznym rozkładzie. W odniesieniu do klastra, $\mathbf{b}_k$ składa się z masy punktowej w zerze i ciągłego rozkładu wartości współczynników. Ten model typu "spike and slab"wprowadza ukryte zero-jedynkowe zmienne $\boldsymbol{\gamma}_k$ z początkowymi prawdopodobieństwami włączenia [28, 29]. Jeśli konkretna i-ta próbka jest naprawdę skojarzona z $\mathbf{l}_k$, wtedy $\boldsymbol{\gamma}_{i,k}$ wynosi 1. W przeciwnym razie, 0. $\mathbf{b}_k = \boldsymbol{\gamma}_k\mathbf{b}_k$, gdzie $\mathbf{b}_k$ może przyjąć ciągły rozkład, co ilościowo ujmuje związek pomiędzy $\mathbf{L}$ a $\mathbf{Y}$. Pozwala to na biologiczną, w tym międzykomórkową, zmienność wewnątrz klastra. Średnie w wierszach można łatwo obsłużyć poprzez centrowanie danych.

Używamy statystyki F do powiązania próbek pojedynczych komórek $\mathbf{Y}$ i środków klastrów $\mathbf{c}_k(\mathbf{Y})$ dla $k = 1, ..., K$. Profile ekspresji genów danej próbki pojedynczej komórki $\mathbf{y}_i$ są modelowane z użyciem środków klastrów $\mathbf{c}_k(\mathbf{Y})$ oraz innych zmiennych $\mathbf{X}_i$, co prowadzi do pełnego modelu $\mathbf{y}_i \sim f_{\text{full}}(\mathbf{c}_k(\mathbf{Y}), \mathbf{X}_i)$. Alternatywnie, model zerowy to $\mathbf{y}_i \sim f_{\text{null}}(\mathbf{X}_i)$. Następnie, nieskorygowana reszta sum kwadratów mierzy niespójność między $\mathbf{y}_i$ a dwoma konkurującymi modelami:

$$\text{RSS}_{\text{full,i}} = \sum(\mathbf{y}_i - f_{\text{full}}(\mathbf{c}_k(\mathbf{Y}), \mathbf{X}_i))^2 \qquad \text{RSS}_{\text{null,i}} = \sum(\mathbf{y}_i - f_{\text{null}}(\mathbf{X}_i))^2 \tag{5}$$

Nieprzystosowana statystyka F dla i-tej pojedynczej komórki jest zdefiniowana jako $F_i = \frac{\text{RSS}_{\text{null,i}} - \text{RSS}_{\text{full,i}}}{\text{RSS}_{\text{full,i}}/(n - p_{\text{full,i}})}$, gdzie p_{full} oznacza liczbę parametrów w pełnym modelu. Jednak, jako że $\mathbf{c}_k(\mathbf{Y})$ jest oszacowane z $\mathbf{Y}$, występuje zależność kołowa prowadząca do sztucznie zwiększonej istotności. Używanie etykiet zależnych od danych, takich jak subpopulacje komórkowe pochodzące z klasteryzacji, nie kontroluje poziomu błędów. Dlatego opracowaliśmy jackstraw do klasteryzacji — podejście oparte na próbkowaniu z zamianą, aby oszacować empiryczny rozkład statystyk F pod modelem hipotezy zerowej, które koryguje tę zależność cykliczną. Oprócz klasteryzacji metodą K-średnich, wdrożyliśmy również jackstraw dla skalowalnej wersji mini batch metody K-średnich (MBKM), Partitioning Around Medoids (PAM) i innych.

Na podstawie jackstraw do klasteryzacji opracowaliśmy podejście empirycznego Bayesa do obliczania prawdopodobieństw włączenia *a posteriori* (posterior inclusion probabilities; PIPs) dla m komórek poprawnie przypisanych do swoich klastrów. Rozważmy, że m wartości p jackstraw $\mathbf{p} = p_1, \ldots, p_m$ są uzyskane dla m próbek pojedynczych komórek, które zostały zgrupowane w K subpopulacji. Szacujemy prawdopodobieństwo *a posteriori*, że $b_i \neq 0$, ponieważ niezerowe współczynniki oznaczają ich wiarygodne włączenie do klastrów:

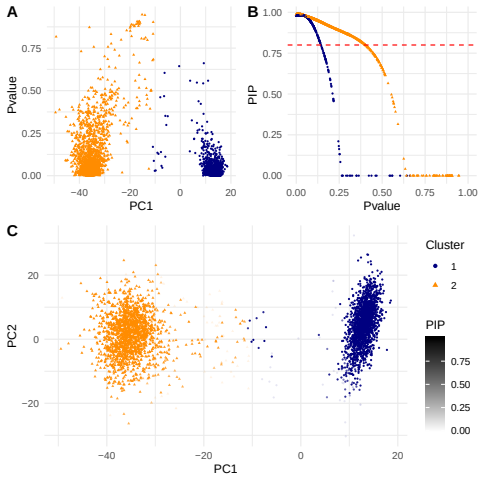
$$\rho_i = \Pr(b_i \neq 0|\mathbf{p}_m) = 1 - \Pr(b_i = 0|\mathbf{p}_m). \tag{6}$$

Proponowane m PIPs mogą być elastycznie łączone do dalszych analiz, w celu usprawnienia wyboru cech i redukcji wymiarów. Przy stosowaniu proponowanych metod do oceny identyfikacji komórek w danych scRNA-seq, PIPs są używane do twardego i miękkiego progowania próbek pojedynczych komórek. Po pierwsze, w twardym progowaniu, komórki z niskimi PIPs byłyby usuwane lub maskowane dla pewnych dalszych analiz, osiągając selekcję cech. Po drugie, w miękkim progowaniu, PIPs mogą być użyte jako wagi dla pojedynczych komórek do dalszych analiz. Nasze analizy pojedynczych komórek w **[B]** pokazują te zastosowania.

Przeprowadziliśmy zestaw kompleksowych badań symulacyjnych, które umożliwiły krytyczną ocenę wartości p przy użyciu danych symulowanych. Po pierwsze, wygenerowaliśmy dane z modelu (Eq. (4)) zmieniając ilość szumu (σ^2), liczbę komórek (m) oraz liczbę genów (n). Po drugie, rozważaliśmy strukturę klastrów pochodzącą z profili ekspresji genów 2700 jednojądrzastych komórek krwi obwodowej (PBMC). Osiem klastrów z różną ilością sygnałów użyto do symulacji danych. Po trzecie, przeprowadziliśmy eksperyment Splatter [30] używając ludzkich indukowanych pluripotencjalnych komórek macierzystych (iPSC), w których wszystkie komórki pochodzą z $K = 3$ subpopulacji. W tych trzech badaniach symulacyjnych jackstraw dla klasteryzacji wykazał zgodność z kryterium hipotezy zerowej, demonstrując swoją dokładność [31].

Przykładowo, jego zastosowanie w mieszaninie linii komórkowych Jurkat i 293T (50:50) które były sekwencjonowane przy użyciu GemCode przez 10X Genomics [23] jest pokazane w Fig. 3. Linie komórkowe Jurkat i 293T są wysoce zróżnicowane, pochodząc odpowiednio od osobników męskich i żeńskich. [23]

zastosował klasteryzację K-średnich, która dzieli $m = 3381$ komórek na $K = 2$ subpopulacje. Wartości p odzwierciedlają odchylenie od dwóch centrów, wzdłuż osi pierwszego PC (Fig. 3(a)). Dla PIP < 0.80 (równoważnego 20% lokalnym FDR), 5.97% spośród 3381 pojedynczych komórek jest zidentyfikowanych jako niejednoznaczne i usuniętych z odpowiednich klastrów (Fig. 3(b)). Wizualizowaliśmy PIPs jako poziomy transparentności w wykresie rozrzutu dwóch głównych składowych PC (Fig. 3(c)).



Rysunek 3: Tożsamości komórek w danych mieszaninach komórek Jurkat:293T z [23]. Dwie odrębne linie komórkowe tworzą $K = 2$ subpopulacje komórkowe. Proponowana metoda jackstraw jest stosowana na 10 najważniejszych składowych (PCs) UMIs. *(a)* Wartości P z proponowanych metod są przedstawione w stosunku do pierwszej głównej składowej (PC). Dwie kolorowe grupy punktów odpowiadają dwóm klastrom. *(b)* Przy PIP < 0.80, 3.4% z 3381 pojedynczych komórek zostałoby usuniętych. *(c)* PIP kontroluje poziomy przezroczystości na wykresie rozpraszania PC. Kiedy PIP=0, punkt danych jest całkowicie przezroczysty.

Ponadto, jackstraw do klasteryzacji, PIPs i podejście do wyboru cech zostały zintegrowane do platformy CV.Signature.TCP do analizy czasowych sygnatur molekularnych w proteomice, w następującym artykule:

> Chung, N.C.[†], Choi, H., Wang, D., Mirza, B., Pelletier, A.R., Sigdel, D., Wang, W., Ping, P. (2020) Identifying temporal molecular signatures underlying cardiovascular diseases: A data science platform. *Journal of Molecular and Cellular Cardiology*. 145:54-58

Ogólnie rzecz biorąc, proponowana metoda jackstraw dla klasteryzacji pomaga prowadzić wnioskowanie w modelach dyskretnych zmiennych ukrytych, gdzie subpopulacje oszacowane z nienadzorowanej klasteryzacji mogą być używane w analizach późniejszych. Otwiera to nowe możliwości dla wybierania kanonicznych członków klastra, zmniejszania centrów klastrów oraz kierowania wyborem stabilnych klastrów. Ponieważ proponowane metody nie są ograniczone do scRNA-seq, przewidujemy ich adaptację w innych dziedzinach wymagających dużej ilości danych, gdzie trenowanie z danymi nieetykietowanymi staje się standardem.

4.4 Wyjaśnianie ukrytych pojęć w głębokich modelach generatywnych [C]

Algorytmy uczenia się bez nadzoru odgrywają coraz większą rolę, ponieważ wykorzystywanie ogromnej ilości danych nieoznaczonych jest kluczem do współczesnego paradygmatu uczenia maszynowego. Głębokie modele generatywne wykorzystują głębokie sieci neuronowe (DNN) do analizy danych na dużą skalę w sposób nienadzorowany. Szczególnie wariacyjne autoenkodery (VAE) [32, 33] pozwalają na ekstrakcję niskowymiarowych przestrzeni ukrytych, które kodują wysokowymiarowe dane wejściowe i ujawniają ukryte relacje. Ogólna architektura VAE składa się z enkodera i dekodera, które zbudowane są z wielu warstw sieci neuronowych. Enkoder, który zazwyczaj przeprowadza drastyczną redukcję wymiarów, kompresuje obserwowane dane wejściowe do zmiennych ukrytych, podczas gdy dekoder rekonstruuje obserwowane dane z tych zmiennych.

Rozważmy, dane $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, które są generowane przez zmienne ukryte $\mathbf{z}$ z rozkładu *a priori* $p_\theta(\mathbf{z})$. Następnie, $\mathbf{X}$ są uzyskiwane z rozkładu warunkowego $p_\theta(\mathbf{x}|\mathbf{z})$, gdzie θ i $\mathbf{z}$ są nieznane. Staramy się przeprowadzić wnioskowanie w tym modelu tak, aby z podanego $\mathbf{X}$, odnaleźć $\mathbf{z}$, poprzez obliczenie gęstości

a posteriori $p_\theta(\mathbf{z}|\mathbf{x}) = \frac{p_\theta(\mathbf{x}|\mathbf{Z})p_\theta(\mathbf{Z})}{p_\theta(\mathbf{X})}$, co jest nieobliczalne, gdyż musielibyśmy otrzymać rozkład brzegowy $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})d\mathbf{z}$ poprzez marginalizację po zmiennych ukrytych. VAE omija to wyzwanie, dopasowując model przybliżonego wnioskowania $q_\phi(\mathbf{z}|\mathbf{x})$, co jest aproksymacją dla prawdziwego rozkładu $p_\theta(\mathbf{z}|\mathbf{x})$. Iteracyjny proces uczenia jednocześnie optymalizuje parametry θ i ϕ. Probabilistyczny enkoder $q_\phi(\mathbf{z}|\mathbf{x})$ i dekoder $p_\theta(\mathbf{x}|\mathbf{z})$ są skonstruowane z sieci neuronowych o L warstwach. W przypadku obrazów i innych danych przestrzennych często używa się konwolucyjnych sieci neuronowych, aby skorzystać z ich zdolności do przetwarzania danych przestrzennych [34, 35]. Trenowanie sieci neuronowej zachodzi poprzez optymalizację gradientową jej parametrów w celu minimalizacji określonej funkcji kosztu.

Chociaż głębokie modele generatywne potrafią generować nowe obrazy [36, 37, 38] i umożliwiają manipulację atrybutami specyficznymi dla obrazów [39, 40], wciąż stanowi wyzwanie wyjaśnienie i interpretacja ich zachowania [41, 42, 43]. Interesuje nas zrozumienie ukrytej reprezentacji pojęć wysokiego poziomu w modelach generatywnych. Wykorzystując VAE, proponujemy ocenę znaczenia cech wejściowych w odniesieniu do wektorów pojęć i przedstawiamy nową metodę uzyskiwania map istotności pojęć (ang. concept saliency maps; CSM). To zasadniczo rozszerza popularną metodę map istotności (ang. saliency maps)[3] w zadaniach klasyfikacyjnych na uczenie bez nadzoru.

Podczas przewidywania znanych klas przy użyciu DNN, mapy istotności zostały wprowadzone jako naturalne podejście do wyjaśniania działania modeli [44, 45, 46]. Mapa istotności wizualizuje względną ważność pikseli w odniesieniu do klas, dla których została zaprojektowana i wytrenowana sieć neuronowa. Daje ona wgląd w zachowanie modelu, które doprowadziło do określonej predykcji. Mapy istotności uzyskuje się poprzez obliczenie gradientu wyniku dla klasy S_{class} względem pikseli wejściowych p_{ij}, gdzie S_{class} jest zwykle aktywacją neuronu w warstwie wyjściowej kodującej klasę, którą chcemy przeanalizować.

Aby osiągnąć zastosowalność dla VAE i innych modeli generatywnych, które naturalnie nie mają znanych klas, proponujemy użycie wektorów pojęć do obliczania wyników S_{concept}, które mogą być rozumiane jako zamiennik dla S_{class}. Wektor pojęcia to ukryta reprezentacja pojęcia wysokiego poziomu, co może obejmować znane atrybuty [39, 47], członkostwa w klastrach [48] lub inne. Gdy model generatywny jest wytrenowany na $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$, wektor pojęcia $\mathbf{z}_c$ jest pozyskiwany poprzez uśrednienie reprezentacji ukrytych próbek zawierających interesujący atrybut (oznaczonych jako $\mathbf{Z}^+$) i odejmowanie średniej próbek, które go nie mają ($\mathbf{Z}^-$):

$$\mathbf{z}_c = \frac{1}{n^+}\sum \mathbf{Z}^+ - \frac{1}{n^-}\sum \mathbf{Z}^-$$

Wynik S_{concept} jest następnie uzyskiwany poprzez mierzenie podobieństwa ukrytej reprezentacji obrazu wejściowego $\mathbf{z}_i = q_\phi(\mathbf{x}_i)$ oraz wektora pojęcia odpowiadającego danemu atrybutowi $\mathbf{z}_c$, gdzie $q_\phi(\mathbf{x}_i)$ to enkoder VAE. W **[C]** zastosowaliśmy iloczyn skalarny jako $S_{\text{concept}} = \mathbf{z}_c \cdot \mathbf{z}_i$.

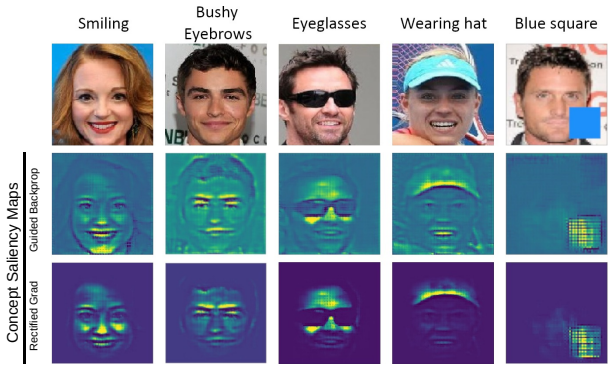
Na koniec obliczany jest gradient S_{concept} względem pikseli wejściowych p_{ij}, aby wygenerować mapę istotności pojęcia M_{ij}

$$M_{ij} = \frac{\partial S_{\text{concept}}}{\partial p_{ij}} \tag{7}$$

Ten ostatni krok może wykorzystywać różne metody do obliczenia (zmodyfikowanych) gradientów. Różnią się one tym, jak traktują propagację wsteczną gradientu poprzez funkcję aktywacji (ReLU), definiowaną jako $f(x) = \max(x, 0)$. Załóżmy, że mamy sieć gęstą o L warstwach i oznaczmy wejście neuronu i w warstwie l jako x_l^i oraz j-tą wagę neuronu jako w_{ij}. Większym n oznaczamy głębsze warstwy, tak że $n = 0$ to warstwa wejściowa, a $l = L$ to warstwa wyjściowa. Warstwa wyjściowa w zadaniu klasyfikacji zawiera wynik klasy, natomiast w modelu nienadzorowanym, używamy zaproponowanego wyniku pojęcia S_{concept}. Gradient $R_l^i = \frac{\partial S_{\text{concept}}}{\partial x_l^i}$ pokazuje, jak wynik klasy lub wynik pojęcia zmienia się względem x_l^i, gdzie x_l^i mogą być wartościami pikseli wejściowych. Używając zasady łańcuchowej, znajdujemy następującą relację dla propagacji wstecznej gradientów $R_l^k = \sum_j w_{kj}\mathbf{1}(x_l^k > 0) \cdot R_{l+1}^j$, gdzie $\mathbf{1}$ to funkcja wskaźnikowa. Jednak mapy istotności tworzone za pomocą tej propagacji wstecznej są znane jako bardzo szumne i dlatego zaproponowano ulepszone metody; mianowicie *Guided Backpropagation* [49] i *Rectified Gradient* [50], które są używane w kolejnych zastosowaniach map istotności.

Zastosowaliśmy zaproponowaną metodę do dwóch zbiorów danych: duża baza danych twarzy CelebA [51] i przestrzenny zbiór danych transkryptomicznych (ST) opuszki węchowej myszy [52]. CelebA to duża baza danych obrazów twarzy z ustalonymi atrybutami [51], składająca się z 202593 obrazów i 40 binarnych adnotacji cech twarzy dla każdego obrazu. Po wstępnym przetworzeniu obrazów, wytrenowaliśmy VAE z warstwami konwolucyjnymi, ReLU, normalizacją warstw oraz z wymiarowością przestrzeni ukrytej równą 400. CSM pokazane na Fig. 4 uzyskano, używając 2000 obrazów dla każdego atrybutu, gdzie adnotacje

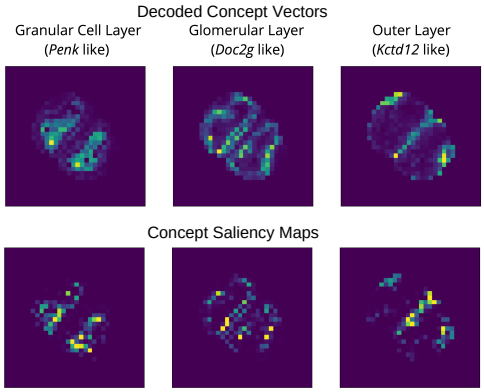
[3] Także nazywane mapami ważności, atrakcyjności lub uwagi. Twierdzimy, że *wyniki* tworzące taką mapę są obliczane przez estymatory istotności.



Rysunek 4: Mapy istotności pojęć (CSM) dla wybranych atrybutów w CelebA. Wektory pojęć i wyniki pojęć są budowane z próbek z wybranymi atrybutami, które nie są bezpośrednio używane przy obliczaniu CSM. Jasność wskazuje na znaczenie dla atrybutu.

zestawu danych CelebA dla cech twarzy, takich jak „uśmiech" lub „zakładanie kapelusza", zostały użyte. CSM często odpowiadają intuicji. Przykładowo podkreślają brwi, kapelusz i niebieski kwadrat w sposób wyraźny.

Przestrzenna transkryptomika (ST) mierzy profile ekspresji genów próbki przy zachowaniu informacji przestrzennej. Używamy tej mapy przestrzennej wartości ekspresji genów rozmieszczonych w siatce, aby zademonstrować, jak tworzyć mapy istotności przy ograniczonej wiedzy o próbkach wejściowych. Interesuje nas zrozumienie, jak pojęcia wysokiego poziomu, takie jak struktury morfologiczne, są obecne w przestrzennych ekspresjach genów, które są mierzone poprzez utrwalanie i barwienie pociętej próbki tkanki. Dokładniej mówiąc, dane ST z opuszki węchowej myszy zawierają całogenomowe ekspresje genów z pociętej tkanki opuszki węchowej myszy [52]. Gdy są umieszczone na mikroukładzie, istnieje 267 miejsc w siatce, które mierzą aktywności ekspresji do 16573 genów w tym obszarze.



Rysunek 5: Mapy istotności pojęć (CSM) związane z warstwami morfologicznymi, reprezentowane przez geny kanoniczne.

Oszacowaliśmy wektory koncepcji dla trzech warstw morfologicznych, obliczając korelacje Pearsona w przestrzeni ukrytej między trzema biomarkerami a wszystkimi innymi genami. Wektory pojęć związane z tymi warstwami morfologicznymi zostały przybliżone poprzez średnią z 50 genów o najwyższych statystykach korelacyjnych (Fig. 5). Zastosowaliśmy zaproponowane metody, używając tych wektorów pojęć do dostępnych genów w danych ST. Wydaje się, że mapy istotności rzeczywiście podkreślają miejsca macierzy liczby genów, które odpowiadają odpowiednim warstwom morfologicznym.

Celem tej pracy jest uogólnienie metody map istotności do nienadzorowanego uczenia głębokiego. Wykazano, że VAE uczą się sensownej reprezentacji ukrytej, która pozwala na manipulację atrybutami poprzez przemieszczanie się w przestrzeni ukrytej [39, 40]. CSM mogą pomóc lepiej zrozumieć, jak głębokie modele generatywne mogą być używane do rozdzielania przestrzeni ukrytej obrazów, takich jak obrazy naturalne i przestrzenna ekspresja genów.

Jednak praca nad mapami istotności ujawniła, że ocena wyjaśnień post hoc[4] jest podstawowym

[4]Chociaż mogą istnieć różne definicje, rozważamy tylko najpopularniejsze podejście – oszacowanie ważności funkcji wejściowych względem wyniku modelu [53, 54, 55]

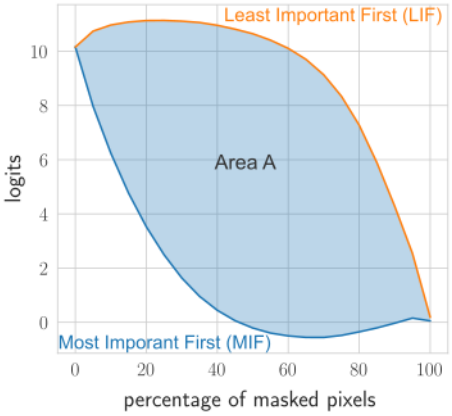
wyzwaniem z powodu braku referencyjnych wyjaśnień. Dlatego większość metod wyjaśnianej AI (XAI), w tym **[C]**, odwołuje się do cech wizualnych i oceny jakościowej.

4.5 Ewaluacja metod wyjaśniania sztucznej inteligencji (ang. explainable AI; XAI) [D, E, F]

Badaliśmy, jak ewaluować wierność wyjaśnień post-hoc, które są niezależne od modelu DNN i metody XAI. W [D, E] wprowadzamy metrykę oceny opartą na zaburzeniach, zwaną wiernością (ang. fidelity) oraz augmentację symulującą wprowadzane w czasie ewaluacji zaburzenia (ang. feature perturbation augmentation; FPA), która przeciwdziała problemom związanym z ewaluacją opartą na wprowadzaniu zaburzeń.

Obiecującym podejściem do porównywania estymatorów ważności, pomimo wspomnianego braku prawdziwych danych, jest zaburzenie cech wejściowych [56, 57, 58]. Dla oceny wierności tworzymy krzywe zaburzeń, które pokazują zmiany w wynikach modelu. W szczególności obserwujemy zmiany w wartościach predykcji przed funkcją softmax. Wartości predykcji przed finalną funkcją aktywacji softmax określane są też logitami (ang. logits). Dany estymator ważności dla każdego piksela wejściowego oblicza wagi, które pokazują, jak bardzo dany piksel przyczynił się do ostatecznej predykcji. Następnie, piksele wejściowe są zaburzane w kolejności od najważniejszych (ang. most important first; MIF) lub najmniej ważnych (ang. least important feature; LIF). Jeżeli estymator ważności estymuje kierunek ważności, estymowane wagi uwzględniają również znak. Wówczas piksele z pozytywnymi wagami dostarczają dowodów przewidywanej klasie, a piksele z ujemnymi wagami mogą dostarczać przeciwdowodów wobec przewidywanej klasy. W przypadku gdy wagi przyjmują wartości dodatnie i ujemne, ranking przebiega od najwyższych wartości dodatnich do najniższych ujemnych (MIF) lub na odwrót (LIF). Najniższe, ujemne wartości wag, które są pierwsze w LIF, mogą wskazywać na silne kontrargumenty/przeciwdowody wobec przewidywanej klasy.

Kiedy estymator ważności uznaje daną cechę za istotną, idealnie jej usunięcie powinno silnie zmniejszyć powiązany logit. Większy spadek wartości logitu wskazuje, że wybrana cecha jest ważniejsza dla predykcji modelu. Odwrotnie, jeśli cecha ma niską rangę ważności, jej usunięcie powinno prowadzić do minimalnego spadku predykcji (estymatory bez znaku) lub potencjalnie wzrostu predykcji poprzez usuwanie pikseli będących przeciwdowodami (estymatory ze znakiem). W praktyce, aby uzyskać końcowe krzywe zaburzeń, uśredniamy znormalizowane logity na dużej liczbie próbek.



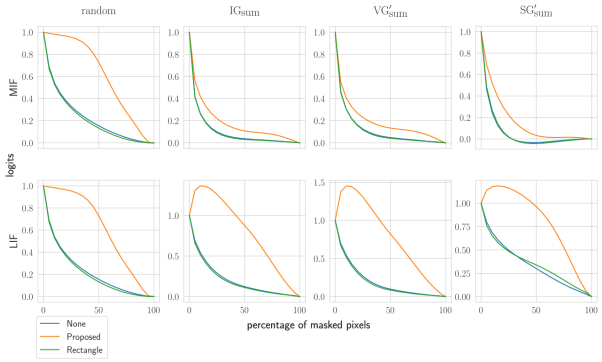
Rysunek 6: Metryka wierności A jest zdefiniowana jako pole między krzywymi LIF (pomarańczowa) i MIF (niebieska). Estymatory ważności z większym A są uznawane za lepiej wyjaśniające model.

Proponujemy użycie pola A między krzywymi MIF a LIF jako miary względnej wierności estymatorów ważności. Małe pole poniżej krzywej MIF wskazuje, że estymator dobrze wykrywa cechy, które są ważnym dowodem dla danej klasy. Duże pole poniżej krzywej LIF oznacza, że estymator może wiarygodnie znaleźć nieistotne cechy lub będące przeciwdowodami. Z A jako metryką wierności, uznajemy estymatory ważności z dużym A za lepsze od tych z mniejszym A.

Jednak taka ocena przez zaburzenia może być problematyczna z powodu ryzyka wprowadzania do zdjęć artefaktów i naruszania kluczowego założenia, że dane treningowe i testowe pochodzą z tego samego rozkładu [59, 60]. Innymi słowy, nawet gdy rzeczywiście nieistotne piksele w zdjęciu są maskowane, predykcja modelu może mimo to znacznie spaść ze względu na sam fakt wprowadzenia zaburzeń do zdjęcia. Poddaje to w wątpliwość wiarygodność opartej na zaburzeniach oceny efektywności. Problem ten jest ściśle związany z przykładami adwersaryjnymi [61].

Proponujemy przezwyciężenie problemu out-of-distribution (OOD) przez augmentację danych w czasie treningu przy użyciu tego samego rodzaju zaburzeń danych, jakie będą używane do późniejszej oceny post-hoc estymatorów ważności. W FPA, mini-batche podczas treningu są wybierane do zaburzeń z prawdopodobieństwem p. W wybranym obrazie iterujemy po pikselach; p^1 odnosi się do prawdopodobieństwa maskowania pojedynczego piksela, a p^2 odnosi się do prawdopodobieństwa tworzenia maski w kształcie kwadratu. Najpierw dla każdego mini-batcha losujemy p^1 z rozkładu Uniform$(0, p^1_{\max})$ i ustawiamy piksele wejściowe na 0 z prawdopodobieństwem p^1. Po drugie, z prawdopodobieństwem p^2 dla każdego

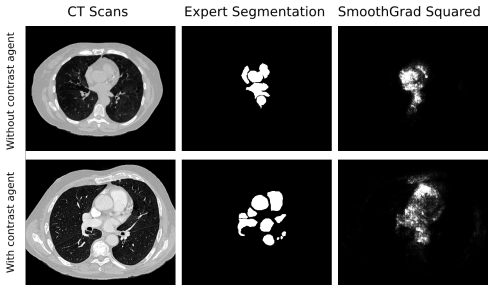
wybranego piksela tworzymy maskę w kształcie kwadratu wypełnionego zerami z losowo wybraną długością boku z przedziału $[1, s_{\max}]$.



Rysunek 7: Krzywe zaburzeń dla ResNet-50 trenowanego na danych Food101 z wagami ważności. „Rectangle" odnosi się do augmentacji danych proponowanej w [62]. „Random" (skrajnie lewe) oznacza, że wagi ważności są przypisane losowo. Początkowe logity zostały znormalizowane do jedynki.

Przeprowadziliśmy eksperymenty na dwóch popularnych architekturach sieci neuronowych i trzech zestawach danych, mianowicie architekturze ResNet-50 [63] trenowanej na ImageNet [64] i Food101 [65], oraz ResNet-18 trenowanej na CIFAR-10 [66]. Porównujemy następujące cztery estymatory ważności: vanilla gradient (VG) [53], Integrated Gradients (IG) [55], SmoothGrad (SG) [54] i SmoothGrad-Squared (SQ-SG) [59]. W [E] dostarczamy ocenę tych metod przy użyciu proponowanej metryki wierności, z i bez FPA. W efekcie, nasza ewaluacja pokazała, że wszystkie rozważane metody: VG, SG i IG przewyższają losowe mapy ważności we wszystkich trzech rozważanych zestawach danych.

Porównując nasze rankingi estymatorów ważności do tych w literaturze, stwierdzamy, że w przeciwieństwie do naszych wyników, ROAR (ang. Remove and Retrain) [59] donosi, że VG, SG i IG działają gorzej niż losowo uzyskane wagi referencyjne. Kilka innych metod oceny jednak informuje, że IG działa lepiej niż losowa referencja [67, 68, 69]. W metodzie ROAR używając wytrenowanego modelu oblicza się dla każdego zdjęcia z danych treningowych wagi ważności. Następnie z danych treningowych usuwa się piksele uznane za ważne i model trenowany jest *de novo* na zmodyfikowanym zestawie danych. Finalnie, badacze mierzą zmianę w dokładności (ang. accuracy) modelu wytrenowanego na zmodyfikowanych danych treningowych. Przypuszczamy, że odmienne wyniki ROAR wynikają z tego, że poprzez mierzenie dokładności ponownie trenowanych modeli ocenia się, ile informacji zostało usuniętych i nie było dostępnych do ponownego trenowania modelu. Różni się to istotnie od oceny, ile informacji zostało usuniętych, na których pierwotnie polegał model. Nasza proponowana metoda unika tego błędu, trenując model tylko raz z FPA, a następnie używając tego samego modelu do wszystkich kolejnych ewaluacji.



Rysunek 8: *Górny rząd:* Próbka przekroju tomografii komputerowej klatki piersiowej bez środka kontrastowego, maska segmentacyjna zdefiniowana przez eksperta oraz mapa wyrazistości uzyskana przy użyciu SmoothGrad Squared (od lewej do prawej). *Dolny rząd:* Te same grafiki co w górnym rzędzie, tutaj dla próbki z kontrastem. Warto zwrócić uwagę, jak obszary takie jak aorta i regiony serca są w tym przypadku wyróżnione.

W [F] oceniano siedem metod XAI w odniesieniu do metody bazowej w kontekście obrazów medycznych. Sieć neuronowa o architekturze ResNet-50 była trenowana do klasyfikacji skanów tomografii komputerowej (CT) płuc pozyskanych z i bez środków kontrastowych, w których klinicznie istotne obszary anatomiczne zostały określone ręcznie jako maski segmentacyjne w obrazach. Zestaw danych użyty

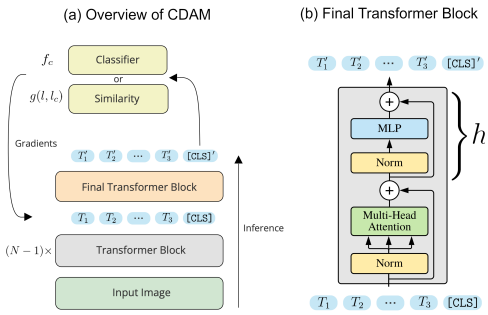
do treningu i testów jest połączeniem kolekcji obrazów udostępnionych przez The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [70] i Bazy Danych Guzków Płucnych (LNDb) [71]. Łącznie dostępne są skany 1312 pacjentów, zawierające ponad 320 tysięcy przekrojów. Model był trenowany przy użyciu 2074 przekrojów uzyskanych przez wybór 2 przekrojów w każdym z 1037 skanów 3D pacjentów.

Trzy metryki oceny były używane do ewaluacji różnych aspektów interpretowalności. Po pierwsze, wspomniana wcześniej wierność mierzy spadek wartości predykcji modelu, gdy pewne piksele wejścia są zaburzane. Po drugie, zgodność między wagami ważności a maskami segmentacyjnymi określonymi przez eksperta jest mierzona na poziomie piksela za pomocą krzywej ROC (ang. receiver operating characteristic curve). Po trzecie, mierzymy zgodność regionu między mapą opartą na XRAI (eXplanation with Ranked Area Integrals; [67]) a maską segmentacyjną za pomocą współczynników podobieństwa Dice'a (DSC). Ogólnie rzecz biorąc, SmoothGrad-Squared (pokazany w Fig. 8) i SmoothGrad [54] znalazły się na szczycie rankingu wierności i ROC, podczas gdy Integrated Gradients [55] oraz SmoothGrad przodowały w ocenie DSC. Oczekiwania i intuicja ekspertów zakodowane w mapach segmentacyjnych nie zawsze zgadzają się z tym, jak model doszedł do swojej predykcji. Zrozumienie tej różnicy może pomóc w wykorzystaniu potencjału głębokiego uczenia maszynowego w medycynie.

4.6 Wyjaśnianie transformerów wizyjnych [G]

Transformery wizyjne (ViT) to rodzina modeli widzenia komputerowego opartych na architekturze transformera [72], które wykazują się porównywalną lub nawet lepszą wydajnością niż konwolucyjne sieci neuronowe [73, 74]. Ważną cechą transformerów jest to, że mechanizm atencji potencjalnie zapewnia wgląd w wewnętrzne działanie modelu [75, 76, 77, 78]. Chociaż mapy atencji zapewniają intuicyjną interpretowalność, nie są wystarczająco różnicujące w stosunku do docelowej klasy. Rozpoznając potencjał map atencji w wyjaśnianiu działania ViT, opracowaliśmy dyskryminujące klasę mapy atencji (ang. class-discriminative attention maps; CDAM), które są rozszerzeniem klasycznych map atencji. Proponowana przez nas metoda wykorzystuje gradienty obliczane dla przewidywanej klasy, co czyni ją wrażliwą na sygnały specyficzne dla klasy lub ukrytego pojęcia.

Pierwotnie architektura transformera została zaprojektowana pod kątem danych tekstowych, jednakże później została zaadoptowana do obszaru danych obrazowych. W obszarze modelowania języka naturalnego, tekst dzielony jest na fragmenty, które następnie kodowane są w reprezentacje wektorowe, nazywane też tokenami. W przypadku obrazów, tokeny tworzone są z małych fragmentów obrazu wejściowego [73]. Tak uzyskane tokeny tworzą sekwencję i analizowane są przez transformera jako sekwencja. Dodatkowo na początku sekwencji dodawany jest token klasyfikacyjny [CLS], który jest wykorzystywany w dalszych zadaniach takich jak klasyfikacja. Mapy atencji są dwuwymiarową wizualizacją wag atencji w ostatniej warstwie atencji z użyciem `[CLS]` jako wektora zapytania (ang. query vector), gdzie wartości atencji są uśredniane dla wszystkich głów atencji (ang. attention heads). W naszej notacji, `[CLS]` odnosi się do tokenu klasyfikacyjnego przed i `[CLS]'` po ostatnim bloku transformera.



Rysunek 9: Schemat graficzny *dyskryminujących klasę map atencji* (CDAM). Zaadoptowano z [73].

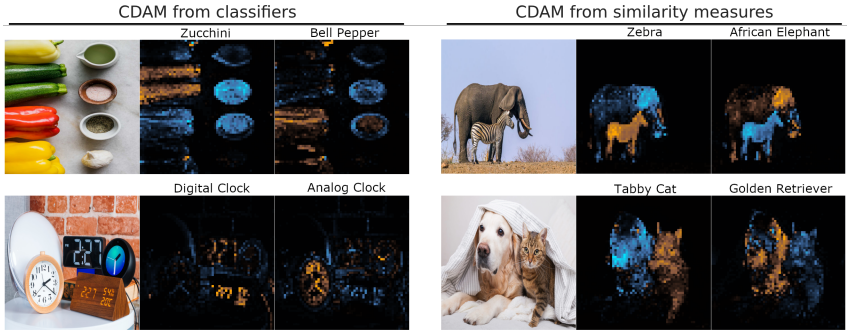
Klasyczne mapy atencji nie są rozróżniają klas, ponieważ nie uwzględniają żadnego sygnału z klasyfikatora pracującego na reprezentacjach z enkodera transformera. Opracowaliśmy metody nadzorowane i nienadzorowane do uzyskiwania CDAM, oparte odpowiednio na znanej klasie i ukrytym pojęciu. Aby oszacować istotność tokenu, CDAM oblicza gradienty logitu klasy lub oceny podobieństwa pojęcia, uzyskane z klasyfikatora używającego tokenu `[CLS]'` w odniesieniu do aktywacji tokenów w końcowym bloku transformera (Fig. 9).

W pierwszym przypadku, w podejściu nadzorowanym, logit klasy jest uzyskiwany z odpowiedniej aktywacji w wektorze przewidywania klasyfikatora trenowanego w sposób nadzorowany na reprezentacjach ViT. W drugim scenariuszu, w podejściu nienadzorowanym (związanym z **[C]**), definiujemy pojęcie poprzez mały zestaw przykładów (np. dziesięć obrazów psa) i uzyskujemy zanurzenie pojęcia (ang. concept embedding) poprzez uśrednianie ich ukrytych reprezentacji. Następnie, bez treningu klasyfikatora, obliczana jest miara podobieństwa między zanurzeniem pojęcia a ukrytą reprezentacją docelowego obrazu testowego. Obliczone podobieństwo używane jest jako cel do obliczania gradientów. Pokazaliśmy, że

CDAM skalują wartości atencji $A_i = \text{softmax}\left(\frac{Q_{\texttt{[CLS]}} K_i^\top}{\sqrt{d_k}}\right)$, pod kątem istotności odpowiednich tokenów T_i dla predykcji klasy lub podobieństwa. Tutaj $Q_{\texttt{[CLS]}} = T_{\texttt{[CLS]}} W^Q$ z $W^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$ oraz $K_i = T_i W^K$ z $W^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$. W tej notacji Q i K są macierzami zapytania i klucza (ang. Queries and Keys Matrices), które łączą wektory zapytań i kluczy [72], a indeksy takie jak K_i wybierają pojedyncze wektory zapytań lub kluczy.

Metody uczenia samonadzorowanego (ang. self-supervised learning; SSL) takie jak DINO (ang. self-DIstillation with No Labels) mogą trenować modele ViT, których mapy atencji są wysokiej jakości, pomimo braku wykorzystania etykiet. Tak uzyskane mapy przypominają wyniki segmentacji semantycznej, czyli dobrze segmentują obiekty obecne na zdjęciach. [79]. Dlatego w naszych badaniach wykorzystujemy model ViT [73] wstępnie wytrenowany za pomocą DINO jako bazowy dla naszych głównych eksperymentów i zastosowań. Ponadto, demonstrujemy ogólną przydatność CDAM poprzez użycie innych modeli ViT, mianowicie wstępnie wytrenowanych z pomocą metody SWAG (ang. Supervised Weakly through Hashtags) [80], DeIT (ang. Data-efficient Image Transformers) [81] oraz DINOv2 [82].

Dodatkowo, inspirowani metodami SmoothGrad [54] oraz Integrated Gradients [55], zbadaliśmy proces uśredniania lub wygładzania podczas obliczania gradientów predykcji klasy lub podobieństwa pojęcia w odniesieniu do aktywacji tokenów w końcowym bloku transformera. Konkretnie, wprowadzamy dwie dodatkowe odmiany CDAM; (1) Smooth CDAM uśrednia wiele gradientów, które są obliczane z dodatkowym szumem dla tokenów oraz (2) Integrated CDAM bierze sumę z gradientów na ścieżce od podstawy do tokenu w końcowym bloku transformera.



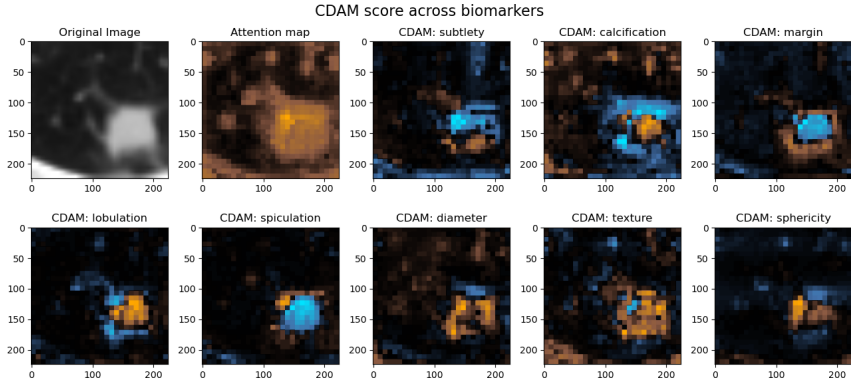
Rysunek 10: Gdy wybieramy różne klasy (po lewej) lub pojęcia (po prawej), uzyskane CDAM są różne i odpowiadają odpowiednim obiektom. Warto zauważyć, że zielone części warzyw są wskazywane jako aktywujące dla klasy *cukinia* (po lewej). Zmarszczki na trąbie słonia wydają się być związane z pojęciem *zebra*, semantycznie mylone przez model jako paski zebry.

Aby zademonstrować naszą metodę, wytrenowaliśmy liniowy klasyfikator na ImageNet [64], który osiągnął dokładność 77,0%. Przewidywania tego klasyfikatora były następnie użyte do stworzenia map cieplnych pokazanych na lewym panelu Fig. 10. Podobnie, zanurzenia pojęcia l_c zostały uzyskane przez uśrednianie ukrytych reprezentacji 30 obrazów (losowo wybranych z zestawu walidacyjnego ImageNet), po czym obliczono CDAM na prawym panelu Fig. 10. W obu scenariuszach, CDAM wyraźnie wyróżnia semantycznie związane z docelową klasą regiony zainteresowania (ROI) z pozytywnymi wartościami (pomarańczowe) od reszty obrazu. Większość tła i inne obiekty otrzymują albo zero (nieistotne) albo ujemne wartości (przeciwdowody). CDAM dziedziczy segmentacje obiektów wysokiej jakości obecne w mapach atencji.

Na podstawie metryk *fidelity* w **[D,E]**, przeprowadzamy szeroko zakrojone ilościowe oceny skupione na poprawności i wrażliwości na klasy. Poprawność jest mierzona przez zakłócanie rosnącej liczby cech wejściowych i mierzenie wyniku modelu [83, 84]. Korzystając z próbek ImageNet [64] z wieloma obiektami [85] i stosując estymatory ważności dla różnych klas, określamy poziom dyskryminacji klas. Dodatkowo, mierzymy rzadkość i zwartość wyjaśnień, obliczając liczbę wartości ważności bliskich zeru. Oprócz proponowanych wariantów: vanilla, Smooth oraz Integrated CDAM, zastosowaliśmy 7 różnych metod wyjaśnialności do porównania. W sumie, trzy odmiany CDAM przewyższają inne metody w ewaluacji ilościowej.

Dla zastosowań w obrazach medycznych, zastosowaliśmy CDAM na ViT wytrenowanym na kolekcji obrazów LIDC (Lung Image Database Consortium) [86]. Pierwszy model oparty na ViT jest trenowany do przewidywania złośliwości guzków płucnych; CDAM bazujące na tym modelu skupiają się głównie na fragmentach guzków. W drugim modelu dopasowaliśmy model ViT wstępnie wytrenowany za pomocą DINO na 8 biomarkerach (subtelność, zwapnienie, sferyczność, brzeg, lobulacja, spikulacja, tekstura i

średnica) [87]. CDAM zostały uzyskane dla 8 biomarkerów (Fig. 11).



Rysunek 11: Wyjaśnienia dla modelu przewidywania biomarkerów. Osiem biomarkerów dostępnych w zestawie danych LIDC jest używane do wytrenowania modelu regresyjnego opartego na DINO ViT. CDAM dostarczają różne i istotne wyjaśnienia dla różnych biomarkerów.

W kontekście biomarkera "brzeg"(ang. margin), fragmenty obejmujące krawędź guzka otrzymują dodatnie wyniki CDAM, natomiast fragmenty odpowiadające ciału guzka otrzymują wartości ujemne. Dla średnicy, fragmenty odpowiadające ciału guzka otrzymują dodatni wynik, podczas gdy piksele tła wykazują neutralne (bliskie zero) lub lekko ujemne wartości. Jeśli chodzi o sferyczność, fragmenty obejmujące krzywiznę guzka otrzymują dodatni wynik CDAM. Spikulacja odnosi się do obecności igiełkowatych, promienistych struktur na brzegu guzka. Wysoka spikulacja często jest wskaźnikiem złośliwości. W większości przypadków, CDAM podkreśla promieniste wypustki pozytywnymi wartościami, podczas gdy wewnętrzne ciała guzków zazwyczaj otrzymują niskie oceny ważności.

Metody *post hoc* wyjaśniające predykcje głębokich sieci neuronowych mają na celu uczynić proces decyzyjny modelu bardziej przejrzystym poprzez dostarczenie aproksymacji ważności elementów wejścia do modelu. Estymatory ważności powinny charakteryzować się pewnymi właściwościami, takimi jak poprawność, wrażliwość i zwartość. Szczególnie wiele metod wyjaśniających, które szacują ważność cech na wejściu do modelu, nie są zbyt użyteczne, ponieważ dają niemal identyczne wyniki, gdy celują w różne klasy w wyjściu modelu [88, 89]. Nawet losowość w wagach modelu często pokazuje minimalny wpływ na wyjaśnienia (tj. mapy ważności [89]). Stwierdzamy zarówno jakościowo, jak i ilościowo, że CDAM jest wysoce wrażliwy na dany cel, przypisując pozytywne oceny ważności obiektom odpowiadającym docelowej klasie i ujemne innym (semantycznie różniącym się) obiektom w obrazie.

5 Aktywność naukowa

Na Uniwersytecie Warszawskim byłem głównym badaczem w trzech dużych grantach (od 2017 r.) na rozwój nowatorskich metod ML z zastosowaniami w danych biologicznych, obrazach medycznych i innych.

W latach 2014-2015, jako NIH UJMT[5] Fogarty Global Health Fellow, pracowałem w „Centre for Infectious Disease Research in Zambia (CIDRZ)" oraz w „Institute for Global Health and Infectious Diseases" na „University of North Carolina at Chapel Hill (USA)". Badałem dane obserwacyjne na dużą skalę z leczenia HIV/AIDS, aby lepiej zrozumieć, kiedy i jak ludzie żyjący z HIV/AIDS.

W latach 2018-2022 współpracowałem z laboratorium prof. Pinga w „David Geffen School of Medicine, University of California Los Angeles (USA)" w ramach „NIH BD2K Center of Excellence for Biomedical Computing" oraz „NHLBI Integrated Cardiovascular Data Science Training Program". Pracowałem nad rozwojem i zastosowaniem metod ML do analizy danych proteomicznych związanych z chorobami układu krążenia.

Od 2020 roku, dzięki finansowaniu CHIST-ERA, współpracuję z dr Hattem (Laboratorium Przetwarzania Informacji Medycznych, INSERM[6] UMR 1101, Uniwersytet Zachodniej Bretanii, Francja) i dr Papadimitroulasem (BioEmTech, Grecja). Opracowałem i zastosowałem wyjaśnialne metody sztucznej inteligencji do wizji komputerowej, ze szczególnym uwzględnieniem klasyfikacji obrazów medycznych

[5]Konsorcjum Uniwersytetu Karoliny Północnej w Chapel Hill, Uniwersytetu Johnsa Hopkinsa, Szkoły Medycznej Morehouse oraz Uniwersytetu Tulane dla Narodowego Instytutu Zdrowia (NIH) w USA.

[6]Institut national de la santé et de la recherche médicale (Narodowy Instytut Zdrowia i Badań Medycznych)

pod kątem raka. Oprócz prac naukowych, nasza współpraca zaowocowała współorganizacją publicznych warsztatów w Grecji, Polsce i Francji.

Od 2018 roku dr Alex Ochoa (Biostatystyka i Bioinformatyka, Uniwersytet Duke'a, USA) i ja współpracujemy nad genomiką statystyczną. Opracowuję metodę ML do analizy struktur populacji w oparciu o polimorfizmy pojedynczych nukleotydów (SNP) z globalnie zróżnicowanych populacji.

6 Osiągnięcia dydaktyczne, organizacyjne oraz popularyzujące naukę

6.1 Osiągnięcia dydaktyczne

Prowadziłem następujące kursy, z których większość stworzyłem w formie wykładów i materiałów laboratoryjnych:

- *Modelowanie złożonych systemów biologicznych* (kurs magisterski) na Wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego; 2018/2019, 2021/2022, 2022/2023, 2024/2025
- *Sztuczna inteligencja Sztuka* (kurs licencjacki) na Wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego; 2023
- *Analiza danych* (kurs licencjacki) na Katedra Genetyki, Uniwersytet Przyrodniczy we Wrocławiu; 2015/2016
- *Biometria* (kurs licencjacki) na Katedra Genetyki, Uniwersytet Przyrodniczy we Wrocławiu; 2015/2016

6.2 Opieka promotorska

Nadzorowałem jednego post-doca i dwóch studentów studiów magisterskich:

- *Lennart Brocki, Ph.D.* Postdoktorant na Wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego; 2021 – 2024
- *Jakub Binda* Magistrant na Wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego; od 2023
- *Małgorzata Gwiazda* Magistrant na Wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego (przeniesiony na Uniwersytet Techniczny w Monachium na stypendium DAAD); 2024

6.3 Osiągnięcia organizacyjne

Współorganizowałem następujące warsztaty:

- *Workshop on explainability of machine learning models for medical applications* na *International Conference on the use of Computers in Radiation therapy (ICCR)* w Lyonie, Francja, 07/2024
- *Joint workshops on XAI methods, challenges and applications* na *European Conference on Artificial Intelligence (ECAI)* w Krakowie, Polska, 10/2023
- *Advanced Computational Techniques for Oncology* w Atenach, Grecja, 02/2023

6.4 Popularyzacja

- Wywiad na temat niebezpieczeństwa związanego ze sztuczną inteligencją typu blackbox w on *Weekly Chosun* (tygodnik) w Korei Południowej; 2024
- Warsztaty na temat *AI jako nowych mediów* w Polsko-Japońska Akademia Technik Komputerowych; 2022 i 2024
- Telewizyjne wystąpienie na temat przyszłości sztucznej inteligencji dla *subnetTalk*, transmitowane przez *FS1 - Community Television Salzburg*, w Austrii; 2023

- Przyczynił się do ruchu otwartego dostępu i otwartego stypendium poprzez współautorstwo następującego dokumentu:

  Tennant, J., Beamer, J., Bosman, J., Brembs, B., Chung, N.C., ... (2019) Foundations for open scholarship strategy development. `https://open-scholarship-strategy.github.io/site/`

6.5 Inne istotne osiągnięcia

Uzyskałem następujące dotacje[7]:

- *SONATA BIS 2023/50/E/ST6/00694* (2024-2029) z Narodowe Centrum Nauki (NCN)
  "XAICancer: Explainable Artificial Intelligence for Cancer Imaging"
- *CHIST-ERA 2020/02/Y/ST6/00071* z Narodowe Centrum Nauki (NCN) wspierany przez program Horyzont Europa
  "Interpretability of Deep Neural Networks for Radiomics"
- *SONATA 2016/23/D/ST6/03613* (2017-2021) z Narodowe Centrum Nauki (NCN)
  "Statistical and machine learning challenges in classification and integration of molecular and genomic data"
- *UJMT Fogarty Global Health Fellowship* (2014-2015) z Narodowego Instytutu Zdrowia (USA)
  "The HIV/AIDS Treatment Cascade in Zambia and in Southern Africa"

Literatura

[1] Jaccard, P.: The distribution of the flora in the alpine zone. New Phytologist **11**(2), 37–50 (1912). doi:10.1111/j.1469-8137.1912.tb05611.x

[2] Tanimoto, T.: An elementary mathematical theory of classification and prediction. Technical report, International Business Machines Corporation, New York (1958)

[3] MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability (1967)

[4] Hartigan, J.A., Wong, M.A.: Algorithm as 136: A k-means clustering algorithm. J. Royal Stat. Soc. Series C **28**(1), 100–108 (1979)

[5] Lloyd, S.: Least squares quantization in pcm. IEEE transactions on information theory **28**(2) (1982)

[6] Birks, H.J.B.: Recent methodological developments in quantitative descriptive biogeography. Ann. Zool. Fenn. **24**, 165–178 (1987)

[7] Jackson, D.A., Somers, K.M., Harvey, H.H.: Null models and fish communities: Evidence of nonrandom patterns. The American Naturalist **139**(5), 930–951 (1992)

[8] Real, R., Vargas, J.M.: The probabilistic basis of jaccard's index of similarity. Systematic Biology **45**(3), 380–385 (1996). doi:10.1093/sysbio/45.3.380

[9] Manly, B.F.J.: Randomization, Bootstrap and Monte Carlo Methods in Biology. Chapman & Hall / CRC Press, Boca Raton, FL (2006)

[10] Łącki, M.K., Startek, M., Valkenborg, D., Gambin, A.: IsoSpec: Hyperfast fine structure calculator. Analytical Chemistry **89**(6), 3272–3277 (2017). doi:10.1021/acs.analchem.6b01459

[11] Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., Willett, P.: Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets. J. Chem. Inf. Model. **52**(11), 2884–2901 (2012). doi:10.1021/ci300261r

[12] Rahman, S.A., Cuesta, S.M., Furnham, N., Holliday, G.L., Thornton, J.M.: EC-BLAST: a tool to automatically search and compare enzyme reactions. Nature Methods **11**(2), 171–174 (2014). doi:10.1038/nmeth.2803

[7] Daty zawierają zatwierdzone rozszerzenia

[13] Bajusz, D., Rácz, A., Héberger, K.: Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform **7**(1) (2015). doi:10.1186/s13321-015-0069-3

[14] Quinlan, A.R.: Bedtools: the swiss-army tool for genome feature analysis. Current Protocols in Bioinformatics, 11–12 (2014). doi:10.1002/0471250953.bi1112s47

[15] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell **9**(12), 3273–3297 (1998)

[16] Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America **95**(25), 14863–14868 (1998)

[17] Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., Brown, P.O.: Genomic expression programs in the response of yeast cells to environmental changes. Molecular Biology of the Cell **11**(12), 4241–4257 (2000)

[18] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. PNAS **96**(12), 6745–6750 (1999)

[19] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., *et al.*: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286**(5439), 531–537 (1999)

[20] Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., Van De Rijn, M., Jeffrey, S.S., *et al.*: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS **98**(19), 10869–10874 (2001)

[21] Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., *et al.*: Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. Science **343**(6172), 776–779 (2014)

[22] Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., *et al.*: Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell **161**(5), 1202–1214 (2015)

[23] Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., *et al.*: Massively parallel digital transcriptional profiling of single cells. Nature communications **8**, 14049 (2017)

[24] Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., Regev, A.: Spatial reconstruction of single-cell gene expression data. Nature biotechnology **33**(5), 495–502 (2015)

[25] Guo, M., Wang, H., Potter, S.S., Whitsett, J.A., Xu, Y.: Sincera: A pipeline for single-cell rna-seq profiling analysis. PLoS Comp. Bio. **11**, 1004575 (2015). doi:10.1371/journal.pcbi.1004575

[26] Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., Trapnell, C.: Single-cell mrna quantification and differential analysis with census. Nature methods **14**, 309–315 (2017). doi:10.1038/nmeth.4150

[27] McCarthy, D.J., Campbell, K.R., Lun, A.T.L., Wills, Q.F.: Scater: pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. Bioinformatics **33**, 1179–1186 (2017). doi:10.1093/bioinformatics/btw777

[28] Mitchell, T.J., Beauchamp, J.J.: Bayesian variable selection in linear regression. Journal of the American Statistical Association **83**(404), 1023–1032 (1988)

[29] George, E.I., McCulloch, R.E.: Approaches for bayesian variable selection. Statistica sinica, 339–373 (1997)

[30] Zappia, L., Phipson, B., Oshlack, A.: Splatter: simulation of single-cell RNA sequencing data. Genome Biology **18**(1) (2017). doi:10.1186/s13059-017-1305-0

[31] Leek, J.T., Storey, J.D.: The joint null criterion for multiple hypothesis tests. Statistical Applications in Genetics and Molecular Biology **10**(1), 28 (2011)

[32] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. Proceedings of the International Conference on Learning Representations (2014)

[33] Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. arXiv preprint arXiv:1401.4082 (2014)

[34] Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics **36**(4), 193–202 (1980). doi:10.1007/bf00344251

[35] Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998). doi:10.1109/5.726791

[36] Salimans, T., Kingma, D., Welling, M.: Markov chain monte carlo and variational inference: Bridging the gap. In: International Conference on Machine Learning, pp. 1218–1226 (2015)

[37] Kulkarni, T.D., Whitney, W.F., Kohli, P., Tenenbaum, J.: Deep convolutional inverse graphics network. In: Advances in Neural Information Processing Systems, pp. 2539–2547 (2015)

[38] Mescheder, L., Nowozin, S., Geiger, A.: Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 2391–2400 (2017). JMLR. org

[39] Larsen, A.B.L., Sonderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. Proceedings of The 33rd International Conference on Machine Learning (2016)

[40] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. ICLR **2**(5), 6 (2017)

[41] Lipton, Z.C.: The mythos of model interpretability. arXiv preprint arXiv:1606.03490 (2016)

[42] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). arXiv preprint arXiv:1711.11279 (2017)

[43] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018) (2018). http://arxiv.org/abs/1810.03292v2

[44] Erhan, D., Bengio, Y., Courville, A., Vincent, P.: Visualizing higher-layer features of a deep network. University of Montreal **1341**(3), 1 (2009)

[45] Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., MÃžller, K.-R.: How to explain individual classification decisions. Journal of Machine Learning Research **11**(Jun), 1803–1831 (2010)

[46] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. International Conference on Learning Representations Workshop (2014)

[47] White, T.: Sampling generative networks. arXiv:1609.04468 (2016)

[48] Huang, C., Change Loy, C., Tang, X.: Unsupervised learning of discriminative attributes and visual representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5175–5184 (2016)

[49] Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. International Conference on Learning Representations (2015)

[50] Kim, B., Seo, J., Jeon, S., Koo, J., Choe, J., Jeon, T.: Why are saliency maps noisy? cause of and solution to noisy saliency maps. arXiv preprint arXiv:1902.04893 (2019)

[51] Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (2015)

[52] Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., *et al.*: Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science **353**(6294), 78–82 (2016)

[53] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: In Workshop at International Conference on Learning Representations (2014). Citeseer

[54] Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)

[55] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 3319–3328 (2017). JMLR. org

[56] Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R.: Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems **28**(11), 2660–2673 (2016)

[57] Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. arXiv preprint arXiv:1806.07421 (2018)

[58] Kindermans, P.-J., Schütt, K.T., Alber, M., Müller, K.-R., Erhan, D., Kim, B., Dähne, S.: Learning how to explain neural networks: Patternnet and patternattribution. arXiv preprint arXiv:1705.05598 (2017)

[59] Hooker, S., Erhan, D., Kindermans, P.-J., Kim, B.: A benchmark for interpretability methods in deep neural networks. Advances in neural information processing systems **32** (2019)

[60] Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3429–3437 (2017)

[61] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)

[62] Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 13001–13008 (2020)

[63] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[64] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). Ieee

[65] Bossard, L., Guillaumin, M., Gool, L.V.: Food-101–mining discriminative components with random forests. In: European Conference on Computer Vision, pp. 446–461 (2014). Springer

[66] Krizhevsky, A.: Learning multiple layers of features from tiny images. University of Toronto Technical Report (2009)

[67] Kapishnikov, A., Bolukbasi, T., Viégas, F., Terry, M.: Xrai: Better attributions through regions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4948–4957 (2019)

[68] Rieger, L., Hansen, L.K.: Irof: a low resource evaluation metric for explanation methods. arXiv preprint arXiv:2003.08747 (2020)

[69] Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Towards better understanding of gradient-based attribution methods for deep neural networks. arXiv preprint arXiv:1711.06104 (2017)

[70] Armato, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., Van Beeke, E.J.R., Yankelevitz, D., Biancardi, A.M., Bland, P.H., Brown, M.S., Engelmann, R.M., Laderach, G.E., Max, D., Pais, R.C., Qing, D.P.Y., Roberts, R.Y., Smith, A.R., Starkey, A., Batrah, P., Caligiuri, P., Farooqi, A., Gladish, G.W., Jude, C.M., Munden, R.F., Petkovska, I., Quint, L.E., Schwartz, L.H., Sundaram, B., Dodd, L.E., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Casteele, A.V., Gupte, S., Sallamm, M., Heath, M.D., Kuhn, M.H., Dharaiya, E., Burns, R., Fryd, D.S., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S., Croft, B.Y.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical Physics **38**(2), 915–931 (2011)

[71] Pedrosa, J., Aresta, G., Ferreira, C., Rodrigues, M., Leitão, P., Carvalho, A.S., Rebelo, J., Negrão, E., Ramos, I., Cunha, A., Campilho, A.: LNDb: A lung nodule database on computed tomography. arXiv:1911.08434 [cs, eess] (2019)

[72] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[73] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[74] Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., Vajda, P.: Visual Transformers: Token-based Image Representation and Processing for Computer Vision. arXiv e-prints (2020). doi:10.48550/arXiv.2006.03677. Accessed 2023-08-31

[75] Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., Eisenstein, J.: Explainable prediction of medical codes from clinical text. arXiv preprint arXiv:1802.05695 (2018)

[76] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

[77] Serrano, S., Smith, N.A.: Is attention interpretable? arXiv preprint arXiv:1906.03731 (2019)

[78] Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: Generating token-level explanations for natural language inference. arXiv preprint arXiv:1904.10717 (2019)

[79] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660 (2021)

[80] Singh, M., Gustafson, L., Adcock, A., de Freitas Reis, V., Gedik, B., Kosaraju, R.P., Mahajan, D., Girshick, R., Dollár, P., van der Maaten, L.: Revisiting weakly supervised pre-training of visual perception models. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

[81] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. International Conference on Machine Learning (2021)

[82] Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)

[83] Brocki, L., Chung, N.C.: Fidelity of interpretability methods and perturbation artifacts in neural networks. In: Maughan, K., Liu, R., Burns, T.F. (eds.) The First Tiny Papers Track at ICLR 2023, Tiny Papers @ ICLR 2023, Kigali, Rwanda, May 5, 2023. OpenReview.net, ??? (2023). https://openreview.net/pdf?id=nbqO93YTz-

[84] Brocki, L., Chung, N.C.: Feature perturbation augmentation for reliable evaluation of importance estimators in neural networks. Pattern Recognition Letters **176**, 131–139 (2023)

[85] Beyer, L., Hénaff, O.J., Kolesnikov, A., Zhai, X., Oord, A.v.d.: Are we done with imagenet? arXiv preprint arXiv:2006.07159 (2020)

[86] Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., *et al.*: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (2011)

[87] Brocki, L., Chung, N.C.: Integration of radiomics and tumor biomarkers in interpretable machine learning models. Cancers **15**(9) (2023). doi:10.3390/cancers15092459

[88] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)

[89] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. Advances in neural information processing systems **31** (2018)