Zastosowanie algorytmów uczenia maszynowego do wykrywania raka w płynnej biopsji

Franciszek Górski
Piotr Juszczyk
Sebastian Lewalski

opiekun: mgr inż. Sebastian Cygert

Założenia projektu

Projekt jest realizowany we współpracy z Gdańskim Uniwersytetem Medycznym.

Celem projektu jest weryfikacja rezultatów otrzymanych przez zespół z Gumedu, który zastosował własną architekturę sieci konwolucyjnej do klasyfikacji danych osób zdrowych i chorych na nowotwór.

Od zespołu z Gumed otrzymaliśmy zbiór danych, na którym pracowali.

Założenia projektu c.d.

Weryfikację rezultatów zespołu Gumed podzieliliśmy na 3 części:

1. Zastosowanie obecnie dostępnej architektury sieci konwolucyjnej do klasyfikacji raka z otrzymanych danych - wybraliśmy architekturę ResNet
2. Zastosowanie algorytmu drzew decyzyjnych do klasyfikacji raka z otrzymanych danych - wybraliśmy algorytm XGBoost
3. Analiza otrzymanych danych pod kątem optymalizacji obliczeń i maksymalizacji wyników

Wykorzystywane narzędzia

Korzystamy z języka Python i różnych bibliotek implementujących mechanizmy uczenia maszynowego/głębokiego takich jak:

- scikit-learn
- PyTorch
- numpy

PyTorch

NumPy

scikit learn

Dostępne zasoby sprzętowe

- Własne komputery
- GPU udostępniane w ramach Google Colaboratory
- Serwer DGX-Station

Colaboratory

Badanie skuteczności architektury ResNet (Franek) - przebieg prac

- Prace rozpoczęliśmy w drugiej połowie marca od implementacji niezbędnych funkcji do przeprowadzania badań.
- Pierwsze wyniki treningu architektury ResNet 18-warstwowej dawały zaskakująco wysokie wartości dokładności (accuracy), powyżej 90% na zbiorze walidacyjnym.
- Szybko okazało się to jednak niewystarczające z powodu niezbalansowanego zbioru danych - przewaga liczby przypadków negatywnych nad przypadkami pozytywnymi

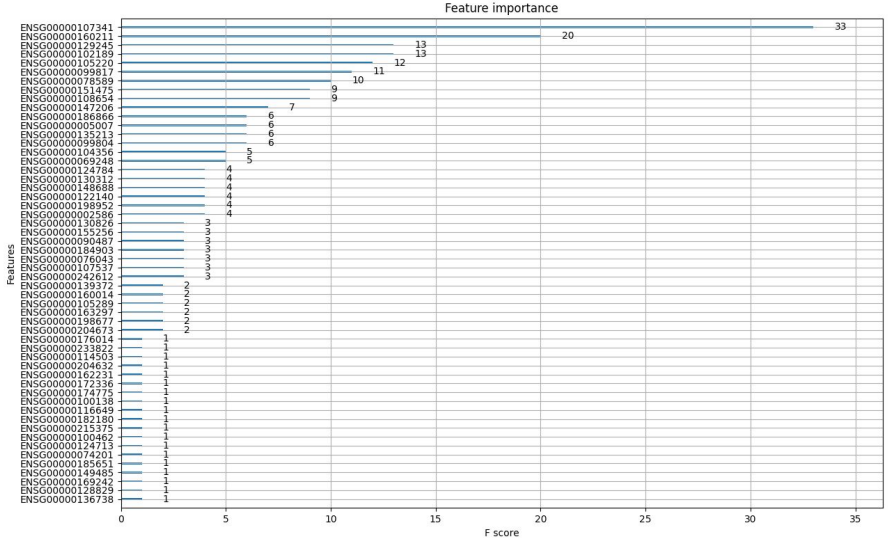
Badanie skuteczności architektury ResNet (Franek) - przebieg prac c.d.

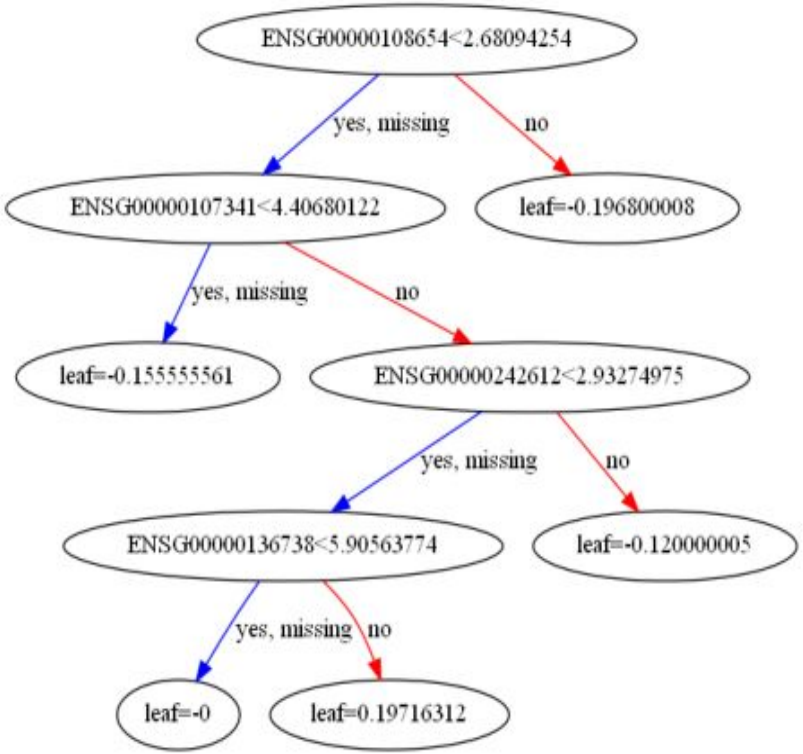
- Przeprowadzenie szeregu eksperymentów takich jak: trening na zredukowanych danych, trening typu transfer learning, trening z wykorzystaniem augmentacji typu MixUp
- Każdy z eksperymentów przeprowadzany był 2-etapowo:
  a. najpierw wykonywane było przeszukiwanie hiperparametrów takich jak learning rate, weight decay i dropout metodą zachłanną
  b. następnie wybierano nastawy z najwyższą wartością metryki balanced accuracy na zbiorze walidacyjnym i powtarzano to 3 krotnie

Badanie skuteczności algorytmu XGBoost

- Badanie wartości metryki *balanced accuracy* (oraz *F1-score weighted*) w zależności od doboru hiperparametrów algorytmu (*learning rate*, *eval_metric*, *n_estiamtors*, *max_depth*)
- Modyfikacja początkowych wag (liczebność przypadków pozytywnych/liczebność przypadków negatywnych)
- Badanie wpływu wykorzystania zredukowanych zbiorów na wynik predykcji
- Wybór cech o największej wartości parametru *feature importance*
- Badanie wpływu normalizacji *feature-wise*
- Badanie skuteczności algorytmu na ograniczonych drzewach

Badanie skuteczności algorytmu XGBoost

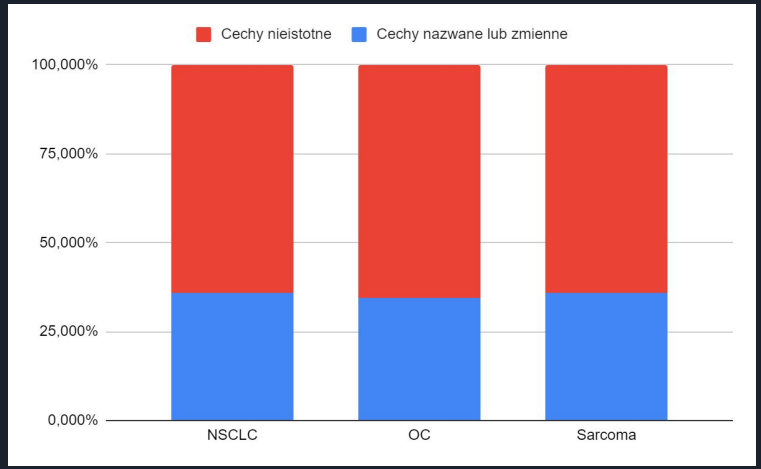




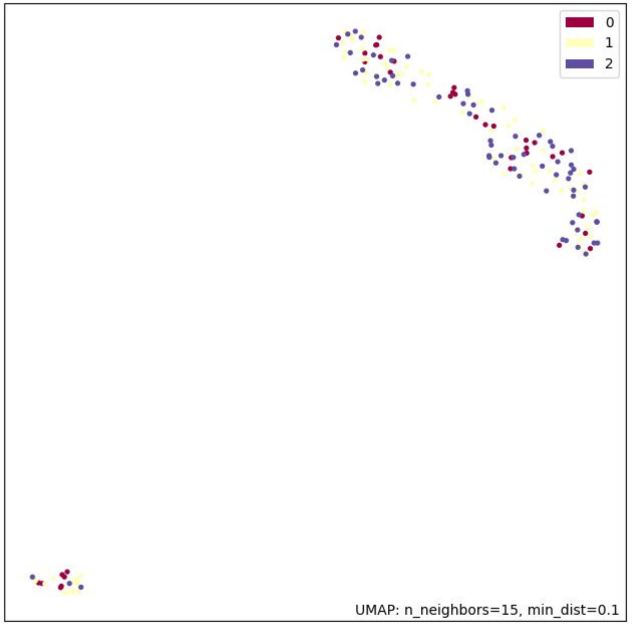
Analiza i przetwarzanie danych (Piotrek)

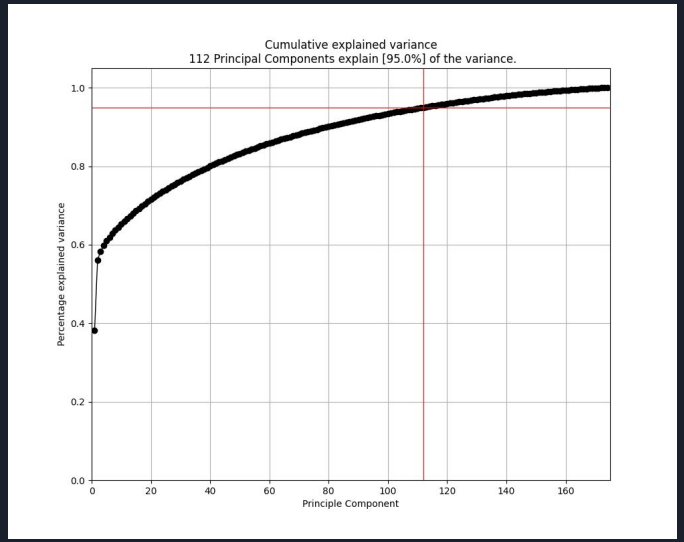
- Niesprawdzony, ograniczony wielkościowo zbiór danych
- Wniosek - trzeba go dokładnie przeanalizować!

W liczbach	NSCLC	OC	Sarcoma
Cechy początkowe	83835	83835	83835
Cechy ważne (TSV)	20242	20242	20242
Cechy nieważne (TSV)	63593	63593	63593
Cechy zmienne	20222	18112	20277
Cechy stałe	63613	65723	63558
Cechy ważne lub zmienne	30063	28940	30097
Cechy ważne zmienne	10401	9414	10422
Cechy ważne stałe	9841	10828	9820
Cechy nieważne zmienne	9821	8698	9855
Cechy nieważne stałe	53772	54895	53738



Analiza i przetwarzanie danych c.d.





Dziękujemy za uwagę