

# Segmentacja obrazów histopatologicznych niedrobnokomórkowego raka płuc

Projekt Capstone



**Mentor:** Mateusz Bednarski

Liliana Gmerek  
lek. Marta Krysik  
lek. Jakub Miąskowski

# Karta projektu

## Czas trwania projektu:

11 sierpnia 2024 - 24. Września 2024

## Typ projektu:

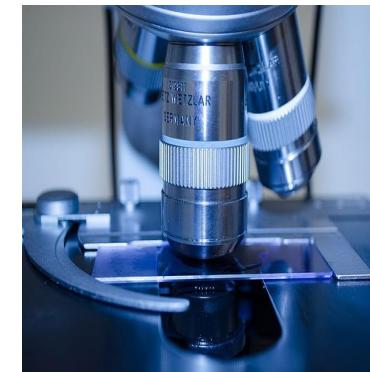
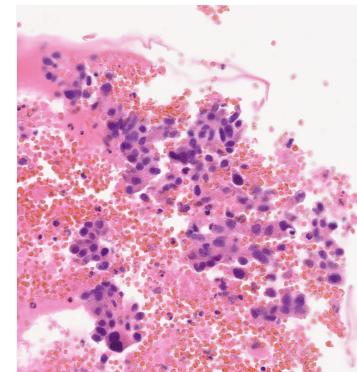
projekt zaliczeniowy 3-semestralnych studiów podyplomowych Data Science w Medycynie na Uniwersytecie Medycznym im. Karola Marcinkiewicza w Poznaniu.

## Podziękowania:

Chcielibyśmy podziękować za udostępnienie i wprowadzenie do danych dr. inż. Adamowi Kozakowi. Ponadto chcielibyśmy wyrazić swoją wdzięczność mentorowi projektu, Mateuszowi Bednarskiemu, za nieocenione wskazówki oraz wsparcie przed rozpoczęciem projektu oraz w czasie jego trwania.

## Typ Problemu:

- Computer Vision
- Deep Learning
- Image Segmentation



# Agenda

I Prezentacja problemu.

II Metodologia pracy nad projektem.

III Prezentacja wyników.

# Agenda

I Prezentacja problemu.

II Metodologia pracy nad projektem.

III Prezentacja wyników.

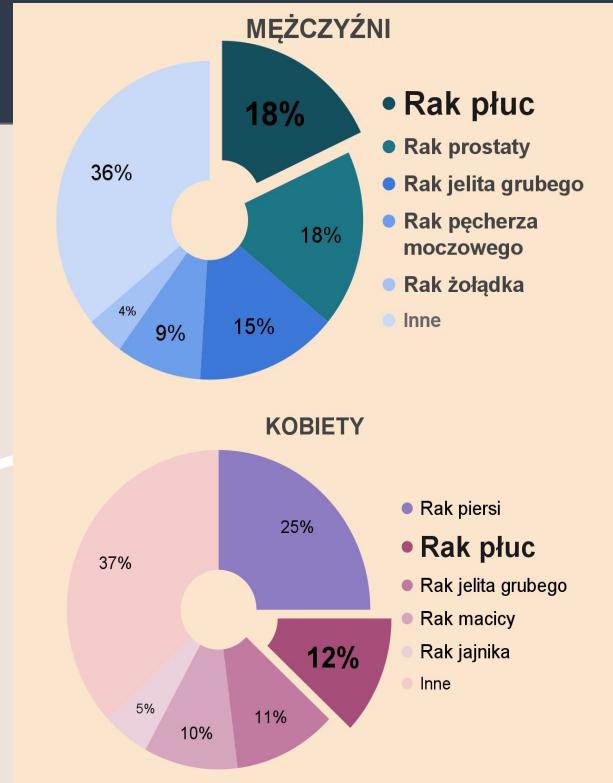
# WPROWADZENIE

*Na wynik badania histopatologicznego, będącego niezbędnym elementem diagnozy niedrobnokomórkowego raka płuc, czeka się nawet kilka tygodni.*

*Nie potrafimy do końca przewidzieć jak Pacjent odpowie na leczenie - cały czas poszukiwane są cechy pozwalające na precyzyjne dobranie skutecznych terapii celowanych.*

# RAK PŁUC

- Rak płuc pod względem zachorowalności jest najczęstszym (obok raka prostaty) nowotworem w Polsce w przypadku mężczyzn oraz drugim co do częstotliwości nowotworem w przypadku kobiet.
- Zachorowalność na raka płuc w Polsce **przekroczyła średnią UE** w 2020 r.<sup>1)</sup>
- Niedrobnokomórkowy rak płuca (Non-Small Cell Lung Cancer, NSCLC) stanowi około 75% wszystkich przypadków raka płuca.
- Z uwagi długiego okresu bezobjawowego większość przypadków NSCLC rozpoznawana jest na **bardzo zaawansowanym etapie** rozwoju choroby.
- Interwencja chirurgiczna w przypadku NSCLC możliwa jest tylko na wczesnym etapie rozwoju choroby (stadium I-II).
- Chemicoterapia ma znaczenie w leczeniu nowotworów rozpoznanych w późniejszych stadiach zaawansowania, a także jako uzupełniająca terapia pooperacyjna.
- NSCLC są **bardzo heterogenną grupą nowotworów**, różniącą się zarówno pod względem utkania histologicznego jak i ekspresji specyficznych biologicznych markerów.



Rozkład zachorowalności na raka wg płci  
w 2020 r. w Polsce  
(źródło: Krajowe profile dotyczące nowotworów 2023)

<sup>1)</sup> Źródło: Europejski System Informacji o Raku (ECIS). <https://ecis.jrc.ec.europa.eu>

# WYZWANIA

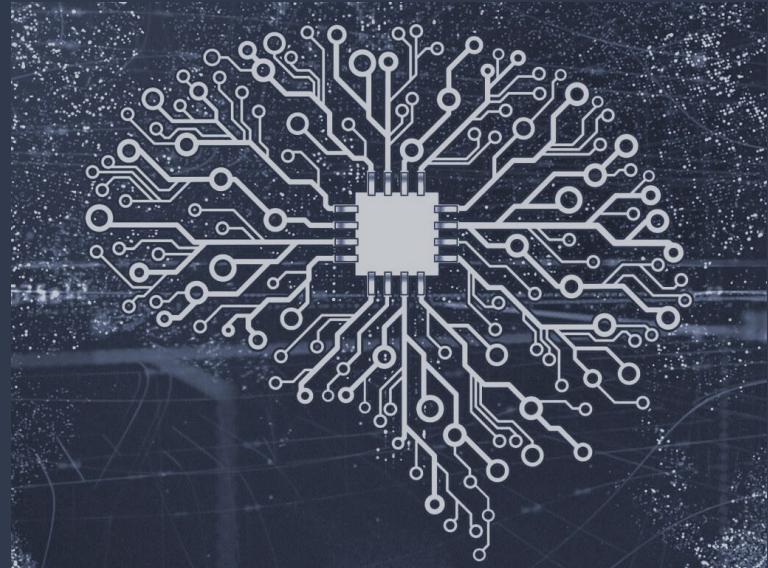
- Aktualnie brakuje wiarygodnych markerów predykcyjnych, które pozwolłyłyby kwalifikować chorych do immunoterapii.
- W celu sprofilowania pacjentów i dopasowania odpowiedniej metody terapii konieczne jest wykonanie zaawansowanych metod diagnostycznych- z zastosowaniem m.in. immunohistochemii oraz badania genetycznego materiału pochodzącego z biopsji.
- Biopsja tkankowa jest procedurą diagnostyczną obarczoną większym ryzykiem dla pacjenta niż pobranie materiału metodą płynnej biopsji.
- Płynna biopsja materiału pochodzącego z popłuczyn oskrzelowo-płucnych jest stosunkowo nieinwazyjną procedurą diagnostyczną, jednak nie zawsze pozwala na określenie typu histologicznego oraz na trafną ocenę zróżnicowania komórek nowotworowych pod względem obecności specyficznych biomarkerów czy mutacji.

# OCZEKIWANE KORZYŚCI

- Terapia celowana molekularnie jest nową i obiecującą formą terapii onkologicznej, z którą wiąże się nadzieję w zakresie poprawy skuteczności leczenia.
- Użycie metod AI do klastrowania zdjęć materiału pochodzącego z płynnej biopsji w połączeniu z danymi o obecności poszczególnych biomarkerów może pomóc we wstępnej ocenie jakości preparatów oraz szybszej i bardziej trafnej kwalifikacji pacjentów do terapii celowanej molekularnie.
- Zastosowanie metod AI w procesie diagnostycznym może znacząco poprawić organizację pracy w Pracowniach Patomorfologicznych oraz Genetycznych, co mogłoby przełożyć się na skrócenie czasu od postawienia rozpoznania do rozpoczęcia leczenia.

# ZDEFINIOWANIE KONSEKWENCJI BŁĘDNYCH WYNIKÓW

- Błędne zakwalifikowanie pacjenta do terapii
- Opóźnienie procesu diagnostyki
- Bezpośrednie koszty podjęcia nieskutecznej terapii:
  - Koszt leków
  - Leczenie powikłań
  - Progresja choroby
  - Wydłużenie czasu terapii



# Kluczowe Wskaźniki Efektywności

## - Key Performance Indicators

Średni czas pomiędzy wykonaniem biopsji a wdrożeniem terapii u pacjentów z rozpoznaniami NSCLC

np. ~7 dni

Odsetek pacjentów, którzy odpowiedzieli pozytywnie na leczenie w czasie pierwszej podjętej terapii

$$\frac{\text{suma pacjentów z pozytywną odpowiedzią na leczenie NSCLC}}{\text{suma pacjentów z wdrożonym leczeniem NSCLC}}$$

np.  $160/200 = 80\%$  pacjentów

Suma podjętych terapii lekowych NSCLC, które zostały przerwane z powodu nieskutecznej odpowiedzi na leczenie w ciągu roku

np. 15 terapii

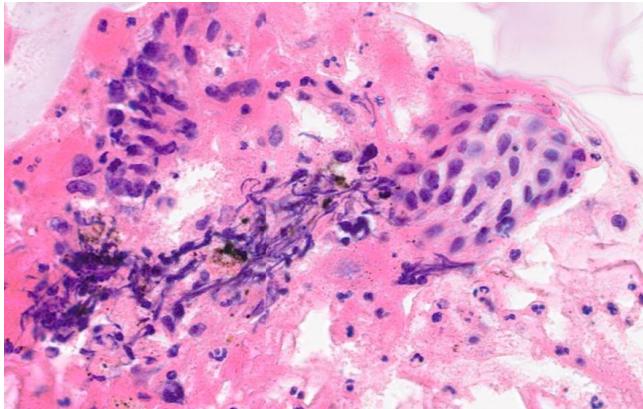
# Agenda

I Prezentacja problemu.

II Metodologia pracy nad  
projektem.

III Prezentacja wyników.

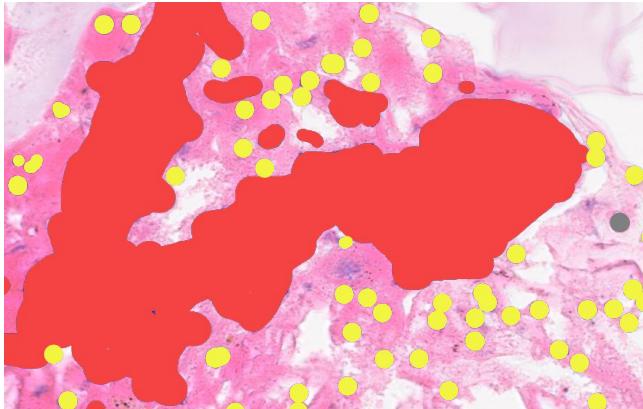
# Dane na których pracowaliśmy



1. Prawie 700 wyselekcjonowanych obszarów skanów mikroskopowych zawierających komórki nowotworowe
2. Dane zostały przeanalizowane i anotowane przez zespół patomorfologów
3. Na zdjęcia zostały nałożone "maski", oznaczające cztery wybrane rodzaje tkanek/komórek:
  - a. NOWOTWOROWE
  - b. ZAPALNE
  - c. TRUDNE DO SKLASYFIKOWANIA
  - d. TKANKĘ ŁĄCZNĄ

ZDJĘCIE PREPARATU

# Dane na których pracowaliśmy



SUMA MASEK

1. Prawie 700 wyselekcjonowanych obszarów skanów mikroskopowych zawierających komórki nowotworowe
2. Dane zostały przeanalizowane i anotowane przez zespół patomorfologów
3. Na zdjęcia zostały nałożone "maski", oznaczające cztery wybrane rodzaje tkanek/komórek:
  - a. NOWOTWOROWE
  - b. ZAPALNE
  - c. TRUDNE DO SKLASYFIKOWANIA
  - d. TKANKĘ ŁĄCZNĄ

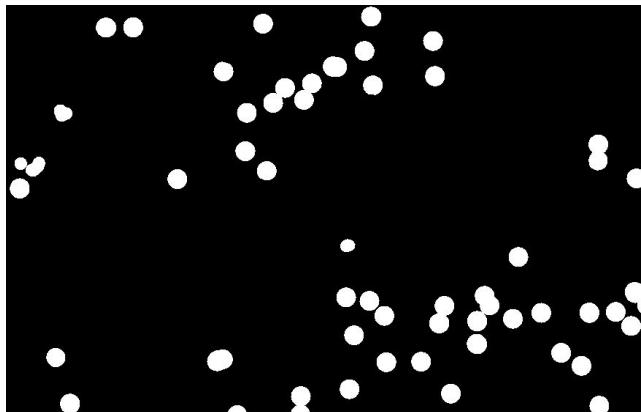
# Dane na których pracowaliśmy



TKANKA NOWOTWOROWA

1. Prawie 700 wyselekcjonowanych obszarów skanów mikroskopowych zawierających komórki nowotworowe
2. Dane zostały przeanalizowane i anotowane przez zespół patomorfologów
3. Na zdjęcia zostały nałożone "maski", oznaczające cztery wybrane rodzaje tkanek/komórek:
  - a. NOWOTWOROWE
  - b. ZAPALNE
  - c. TRUDNE DO SKLASYFIKOWANIA
  - d. TKANKĘ ŁĄCZNĄ

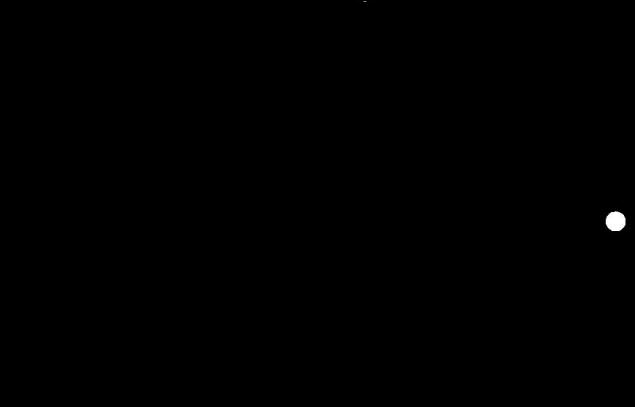
# Dane na których pracowaliśmy



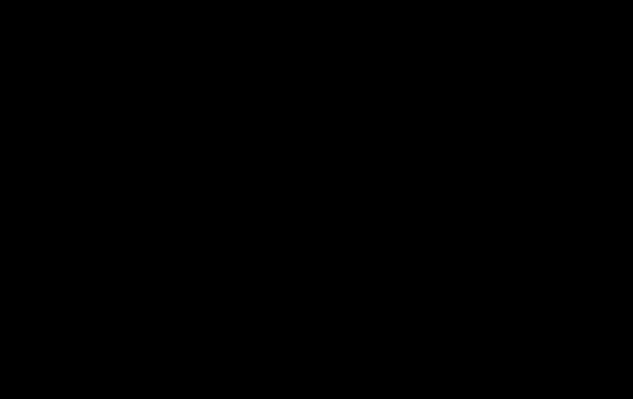
TKANKA ZAPALNA

1. Prawie 700 wyselekcjonowanych obszarów skanów mikroskopowych zawierających komórki nowotworowe
2. Dane zostały przeanalizowane i anotowane przez zespół patomorfologów
3. Na zdjęcia zostały nałożone "maski", oznaczające cztery wybrane rodzaje tkanek/komórek:
  - a. NOWOTWOROWE
  - b. ZAPALNE
  - c. TRUDNE DO SKLASYFIKOWANIA
  - d. TKANKĘ ŁĄCZNĄ

# Dane na których pracowaliśmy

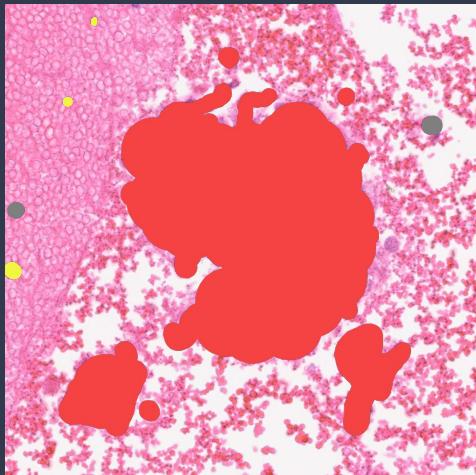
- 
1. Prawie 700 wyselekcjonowanych obszarów skanów mikroskopowych zawierających komórki nowotworowe
  2. Dane zostały przeanalizowane i anotowane przez zespół patomorfologów
  3. Na zdjęcia zostały nałożone "maski", oznaczające cztery wybrane rodzaje tkanek/komórek:
    - a. NOWOTWOROWE
    - b. ZAPALNE
    - c. TRUDNE DO SKLASYFIKOWANIA
    - d. TKANKĘ ŁĄCZNĄ

# Dane na których pracowaliśmy

- 
1. Prawie 700 wyselekcjonowanych obszarów skanów mikroskopowych zawierających komórki nowotworowe
  2. Dane zostały przeanalizowane i anotowane przez zespół patomorfologów
  3. Na zdjęcia zostały nałożone "maski", oznaczające cztery wybrane rodzaje tkanek/komórek:
    - a. NOWOTWOROWE
    - b. ZAPALNE
    - c. TRUDNE DO SKLASYFIKOWANIA
    - d. TKANKĘ ŁĄCZNĄ

TKANKA ŁĄCZNA

# CEL PROJEKTU



**W naszym projekcie skupiliśmy się na trzech rodzajach tkanek:**

- 1) Rakowej (*cancerous*)
- 2) Zapalnej (*inflammatory*)
- 3) Trudnej do sklasyfikowania (*hard to classify*)

**Cel projektu:**

**Wytrenowanie modelu zdolnego do segmentacji wybranych 3 typów tkanek na zdjęciach preparatów histopatologicznych z płynnej biopsji.**

# CHARAKTERYSTYKA ZBIORU DANYCH

Liczba elementów w zbiorze danych:

**673 zdjęcia dla każdej z tkanek (odtworzone maski)**  
+ 673 zdjęcia preparatów z biopsji

**Klasa CANCEROUS:** 99,8% zdjęć zawiera obrysy tkanki

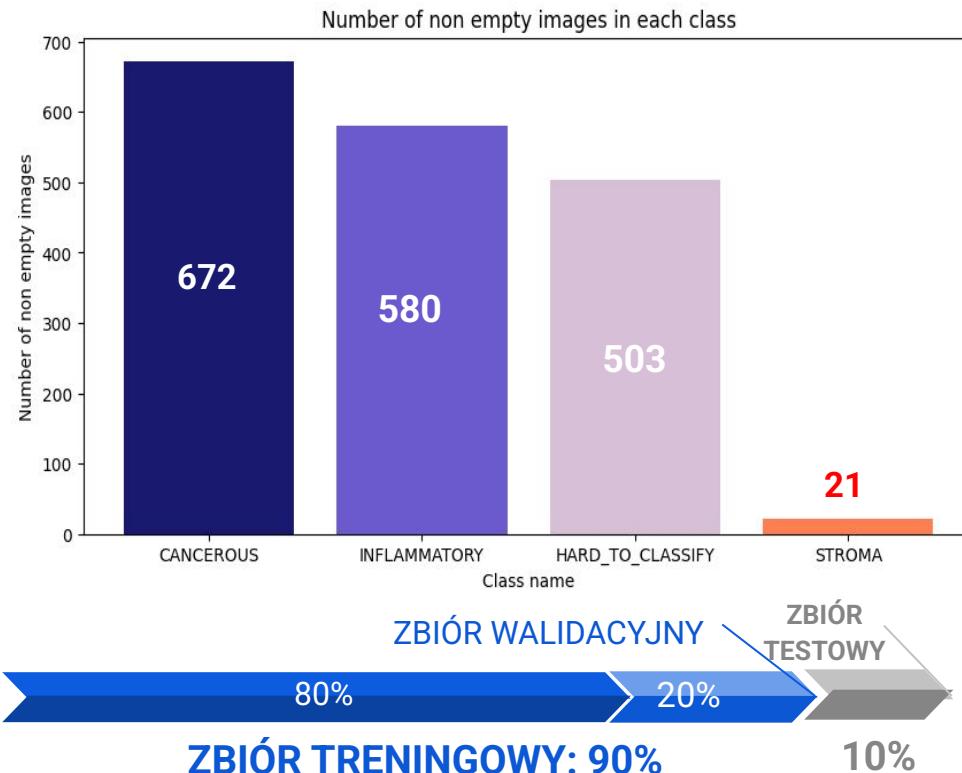
**Klasa INFLAMMATORY:** 86,2% zdjęć zawiera obrysy tkanki

**Klasa HARD TO CLASSIFY:** 74,3% zdjęć zawiera obrysy tkanki

**Klasa STROMA:** **jedynie 3,1% zdjęć** zawiera obrysy tkanki

Uwzględniając klasę STROMA zbiór jest silnie **niezbalansowany** (Imbalance Ratio=32)

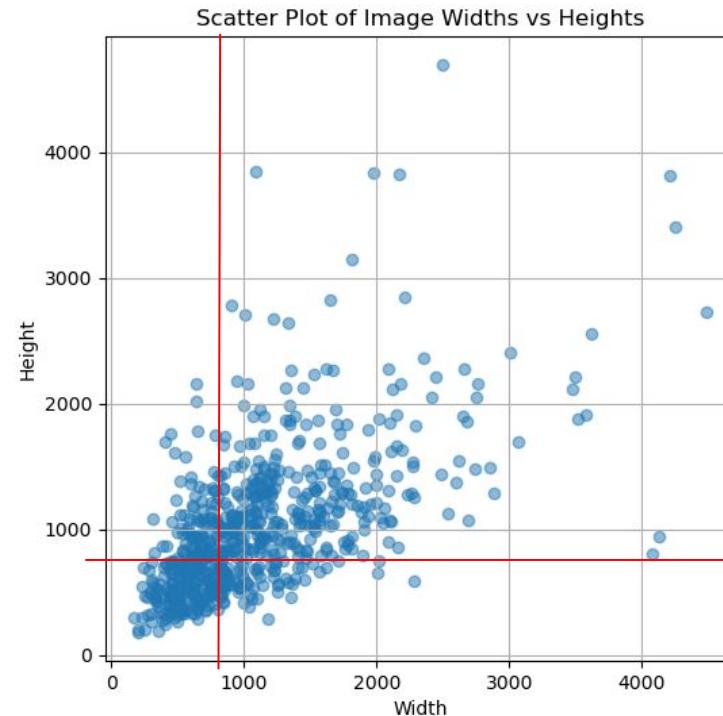
Po wykluszeniu klasy STROMA zbiór jest lepiej zbalansowany (Imbalance Ratio=1.3)



# ROZMIARY OBRAZÓW

Analiza rozmiarów zdjęć:

- Zdjęcia nie mają tego samego rozmiaru
- Zdjęcia nie są kwadratowe
- Mediana szerokości zdjęcia to **931** px., a wysokości **913** px.
- Rozstęp ćwiartkowy dla szerokości (IQR) wynosi **738** pikseli, a dla długości zdjęcia **601** pikseli, co wskazuje na wysoki poziom zróżnicowania rozmiaru oraz potencjalnie dużą liczbę wartości odstających, które można zaobserwować na wykresie punktowym.

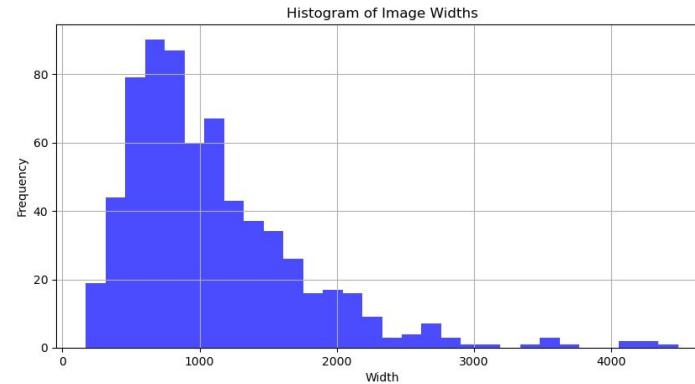
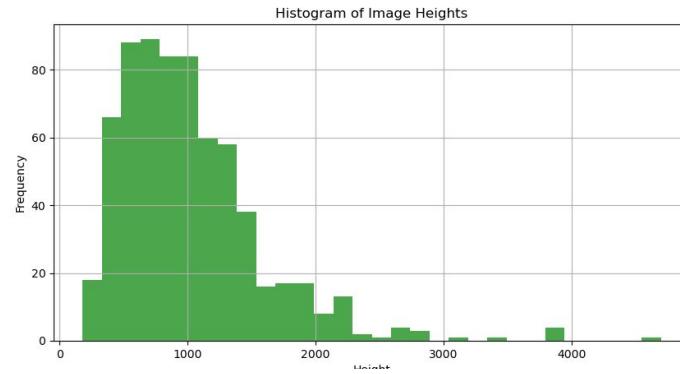


# ROZMIARY OBRAZÓW

Histogramy szerokości oraz wysokości zdjęć wskazują na lewoskońską dystrybucję.

W rezultacie:

1. W modelu zrezygnowano z opcji zmiany rozmiaru zdjęcia, która mogłaby wpływać na niewłaściwe uczenie się modelu (*resizing*), na rzecz **selektywnego wycinania fragmentów zdjęcia (*cropping*)**.
2. Minimalny rozmiar zdjęć do eksperymentów określono na 256x256 pikseli, a ostatecznie wybrany rozmiar zdjęć **512x512** był uzasadniony lepszymi rezultatami oraz ograniczeniami sprzętowymi.



# KONFIGURATOR

```
1 v class CFG:  
2     experiment_id      = 'JM53'  
3     model_name         = 'UNet'  
4     train_continuation = False  
5     pretrained_model_path = 'type_your_model_directory_here'  
6     early_stopping     = True  
7     es_patience        = 3  
8     es_delta            = 0.002  
9     train_bs            = 4  
10    valid_bs            = 4  
11    crop_size           = (512 , 512)  
12    num_crops           = 8  
13    epochs              = 100  
14    lr                  = 0.0001  
15    data_train_test_split = 0.9  
16    data_train_val_split = 0.8  
17    device               = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")  
18    mask_train           = False  
19    cancerous_train      = True  
20    hard_to_classify_train = False  
21    inflammatory_train   = False  
22    stroma_train          = False
```

# KONFIGURATOR

```
1 v class CFG:  
2     experiment_id          = 'JM53'  
3     model_name              = 'UNet'  
4     train_continuation      = False  
5     pretrained_model_path  = 'type_your_model_directory_here'  
6     early_stopping          = True  
7     es_patience              = 3  
8     es_delta                 = 0.002  
9     train_bs                  = 4  
10    valid_bs                  = 4  
11    crop_size                = (512 , 512)  
12    num_crops                  = 8  
13    epochs                     = 100  
14    lr                         = 0.0001  
15    data_train_test_split     = 0.9  
16    data_train_val_split      = 0.8  
17    device                     = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")  
18    mask_train                  = False  
19    cancerous_train            = True  
20    hard_to_classify_train     = False  
21    inflammatory_train         = False  
22    stroma_train                = False
```

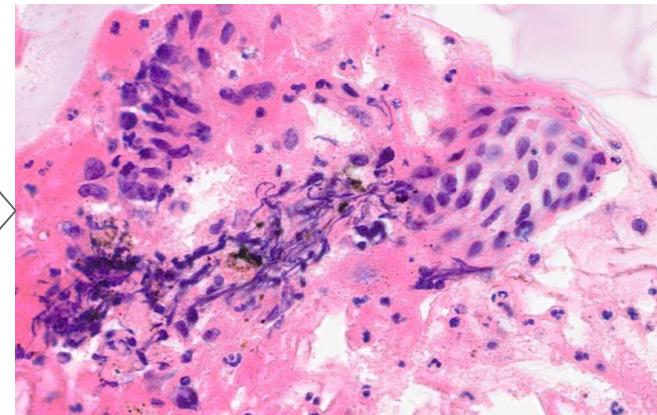
# KONFIGURATOR

```
1 v class CFG:  
2     experiment_id          = 'JM53'  
3     model_name              = 'UNet'  
4     train_continuation      = False  
5     pretrained_model_path  = 'type_your_model_directory_here'  
6     early_stopping          = True  
7     es_patience              = 3  
8     es_delta                 = 0.002  
9     train_bs                 = 4  
10    num_crops                = 4  
11    crop_size                = (512 , 512)  
12    num_crops                = 8  
13    epochs                   = 100  
14    lr                        = 0.0001  
15    data_train_test_split    = 0.9  
16    data_train_val_split     = 0.8  
17    device                    = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")  
18    mask_train                = False  
19    cancerous_train           = True  
20    hard_to_classify_train   = False  
21    inflammatory_train       = False  
22    stroma_train              = False
```

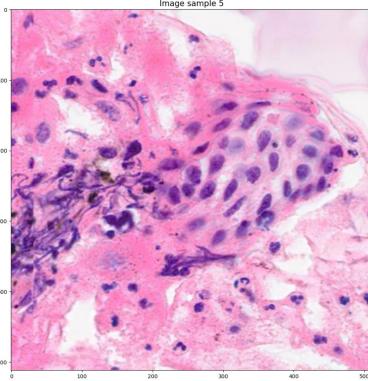
# CROPPING – PRZYCINANIE



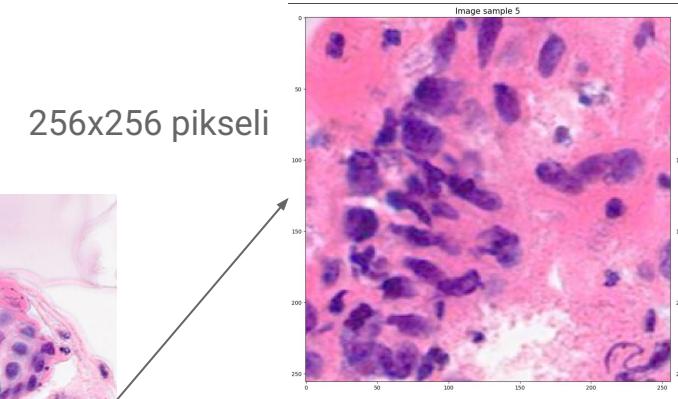
1024 x 1024 pikseli



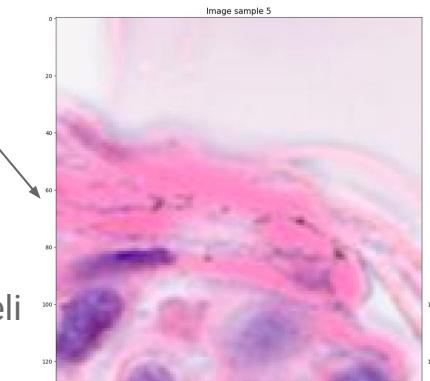
881 X 558 pikseli



512x512 pikseli

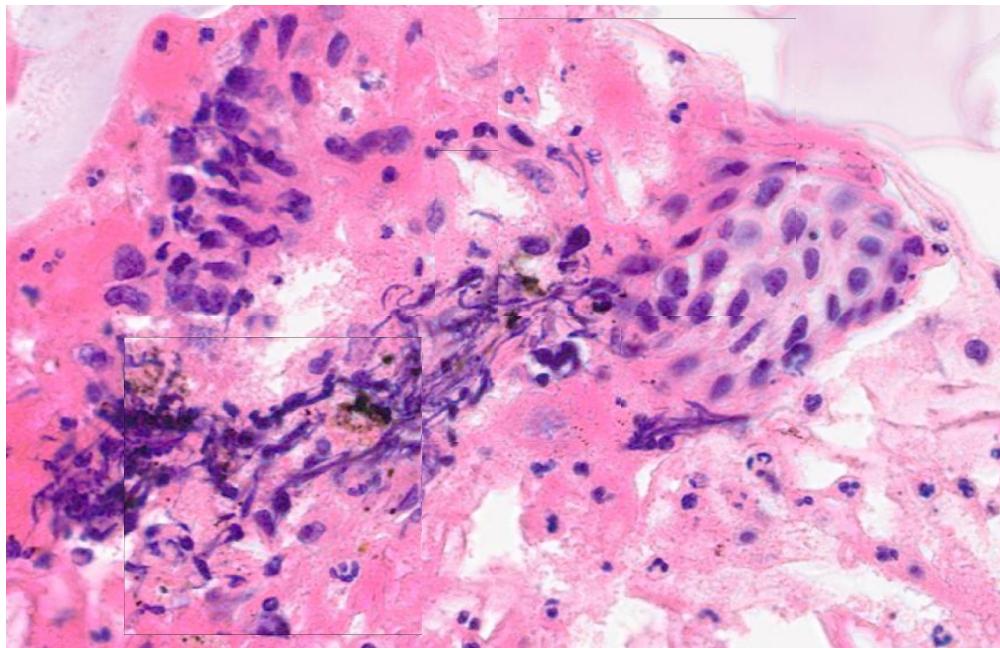


256x256 pikseli

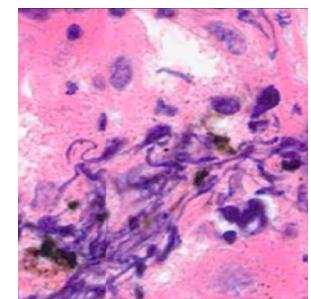
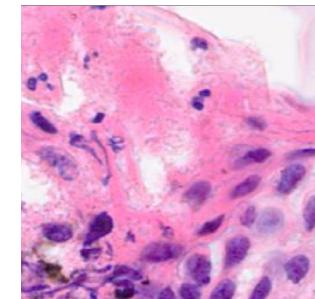
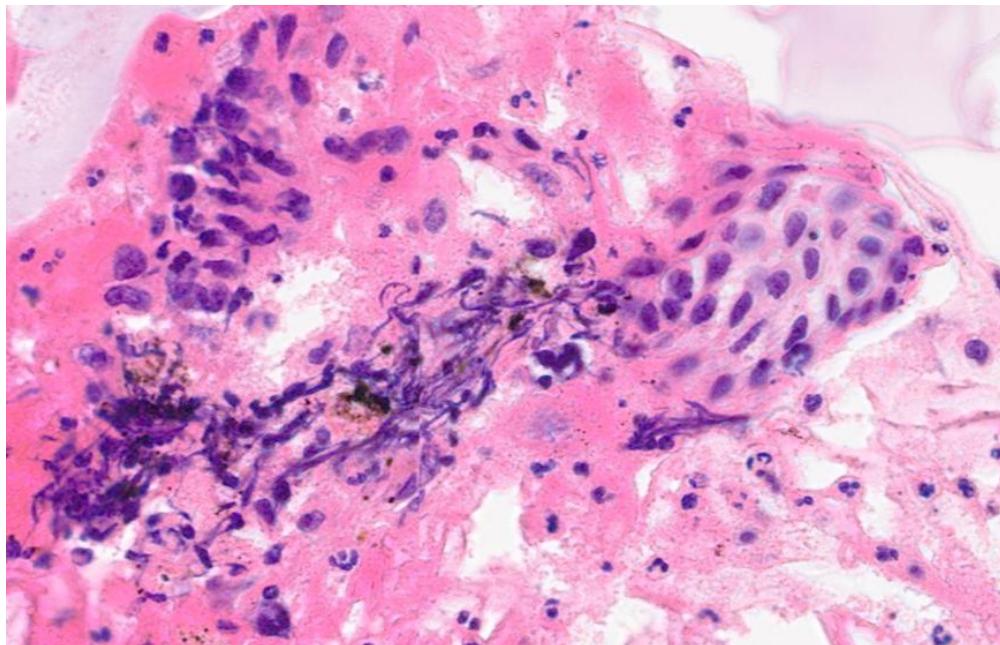
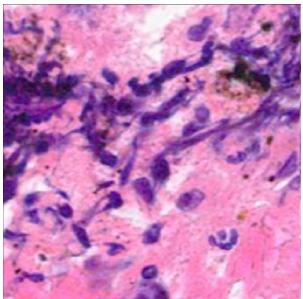
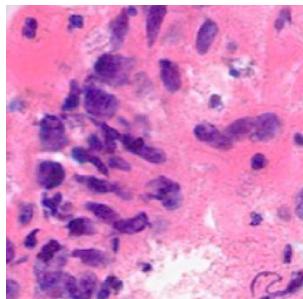


128x128 pikseli

# CROPPING - PRZYCINANIE

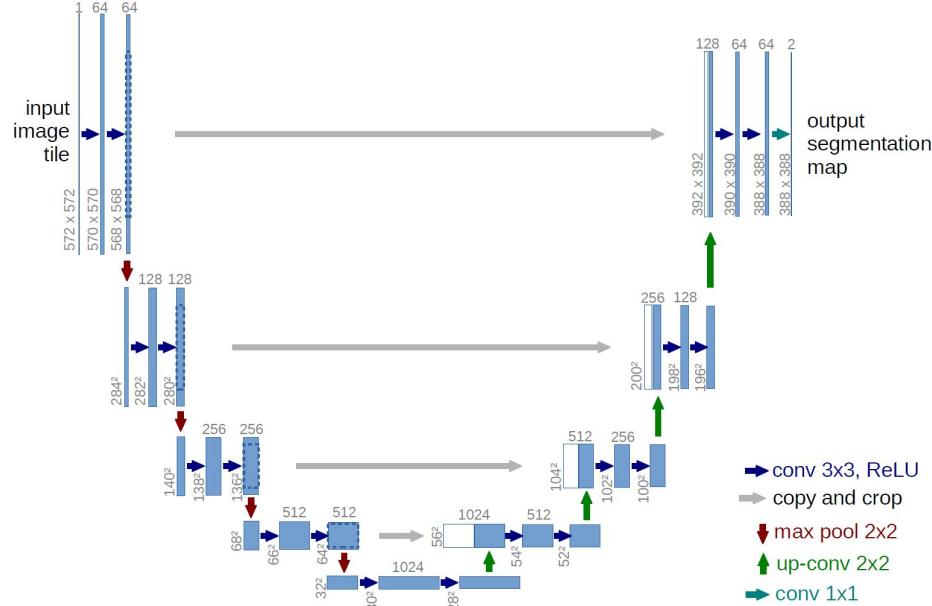


# CROPPING - PRZYCINANIE



# ARCHITEKTURA U-NET

- Sieć neuronowa zaprojektowana specjalnie do segmentacji obrazów
- Opracowana przez zespół z Uniwersytetu we Freiburgu w 2015 roku
- Zdobyła liczne nagrody
- Obraz jest analizowany na różnych poziomach szczegółowości, co poprawia skuteczność uczenia się.



Źródła:

<https://mb.informatik.uni-freiburg.de/people/ronneber/u-net/>  
<https://medium.com/analytics/how-to-label-data-for-semantic-segmentation-deep-learning-models-907a996f95f7>

# AUTOMATYZACJE

STARTER

## Table of contents

- | DATASET - SPLIT & LOAD
- | ARCHITECTURE U-NET
- | TRENOWANIE MODELU
- | TEST

WYBÓR PLATFORMY

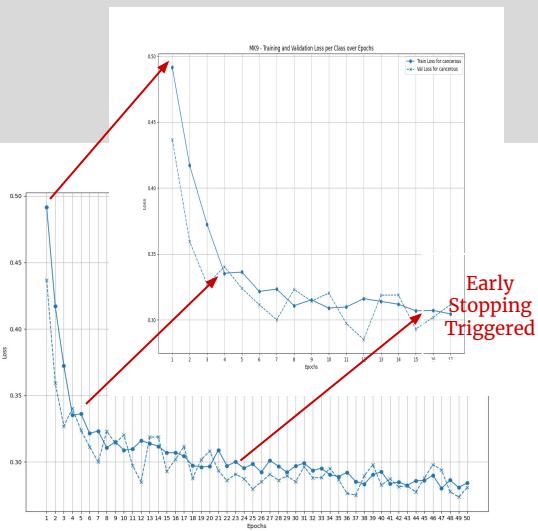


kaggle



# AUTOMATYZACJE

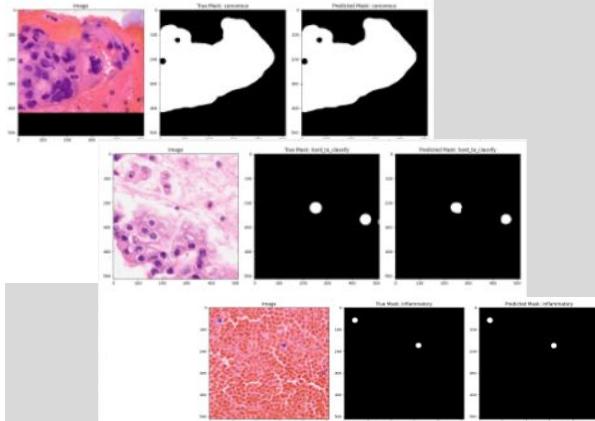
EARLY STOPPING I MOŻLIWOŚĆ  
KONTYNUACJI TRENINGU NA  
WYTRENIOWANYCH WCZEŚNIEJ  
MODELACH



## HYPEROPT



AUTOZAPIS  
WYGENEROWANYCH  
PREDYKCJI

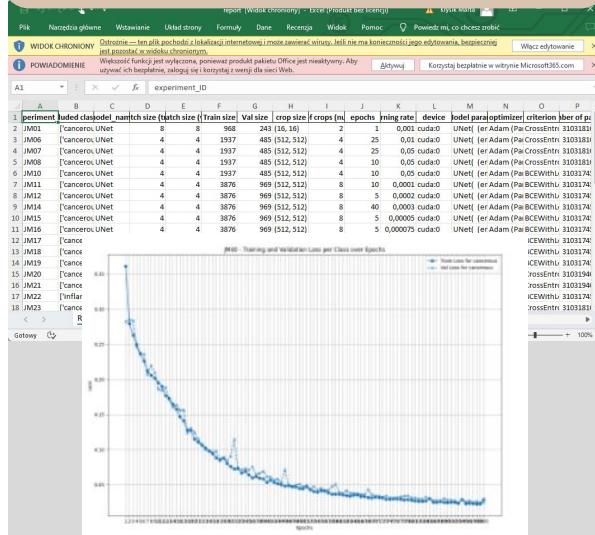


# AUTOMATYZACJE

## EXPERIMENT ID

Pole nazwy	A	B	C	D	E	F
1	experiment_id	included classes	model_name	batch size (train)	batch size (val/Train size)	
2	JM01	['cancerous', 'hard_to_classify']	UNet	8	8	968
3	JM06	['cancerous', 'inflammatory']	UNet	4	4	1937
4	JM07	['cancerous', 'inflammatory']	UNet	4	4	1937
5	JM08	['cancerous']	UNet	4	4	1937
6	JM10	['cancerous']	UNet	4	4	1937
7	JM11	['cancerous']	UNet	4	4	3876
8	JM12	['cancerous']	UNet	4	4	3876
9	JM14	['cancerous']	UNet	4	4	3876
10	JM15	['cancerous']	UNet	4	4	3876
11	JM16	['cancerous']	UNet	4	4	3876
12	JM17	['cancerous']	UNet	4	4	3876
13	JM18	['cancerous']	UNet	4	4	3876
14	JM19	['cancerous']	UNet	4	4	3876
15	JM20	['cancerous', 'hard_to_classify']	UNet	4	4	3876
16	JM21	['cancerous', 'hard_to_classify']	UNet	4	4	3876
17	JM22	['inflammatory']	UNet	4	4	3876
18	JM23	['cancerous', 'inflammatory']	UNet	4	4	3876

## GENEROWANIE RAPORTU Z PARAMETRAMI TRENINGU ORAZ WYKRESAMI PRZEBIEGU UCZENIA

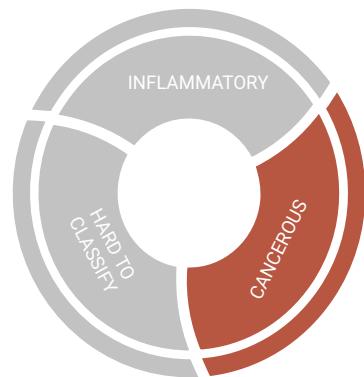
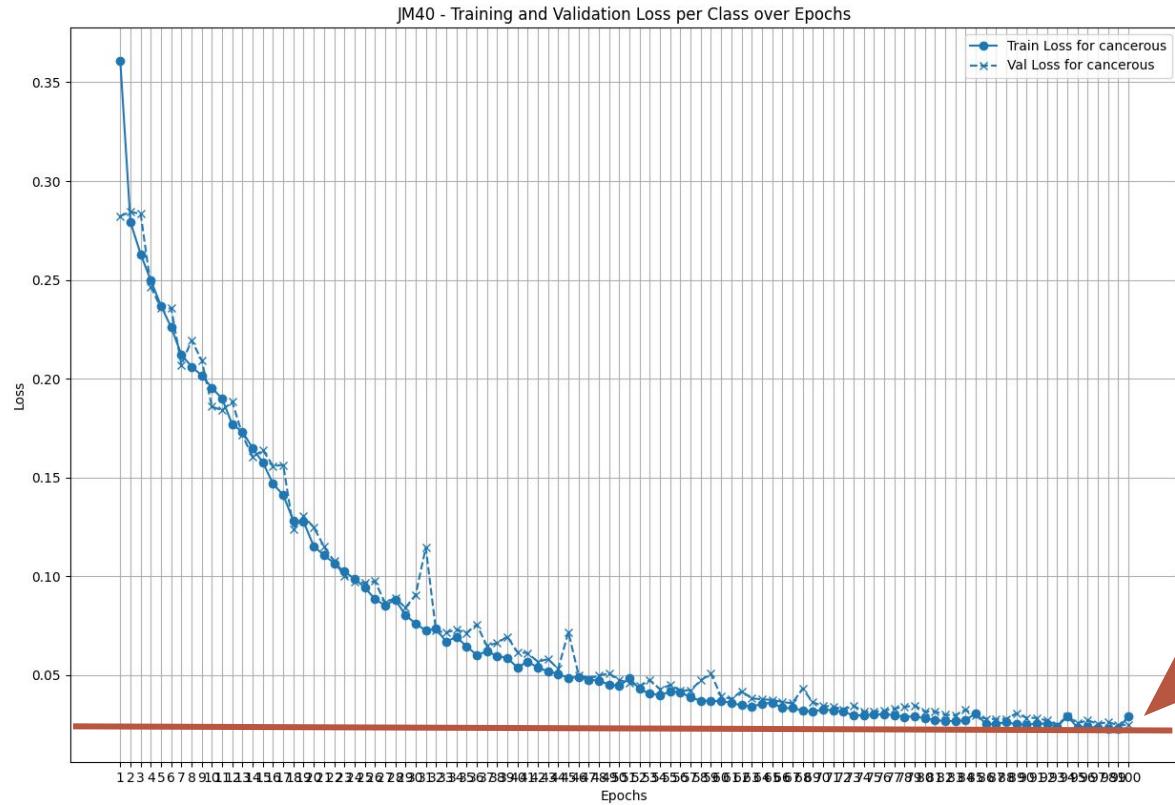


# Agenda

I Prezentacja problemu.

II Metodologia pracy nad projektem.

III Prezentacja wyników.

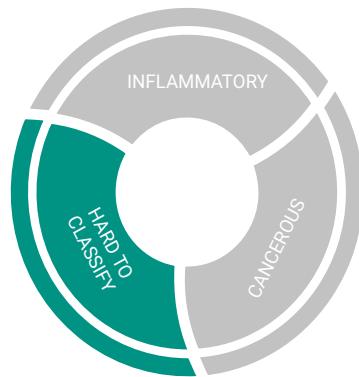
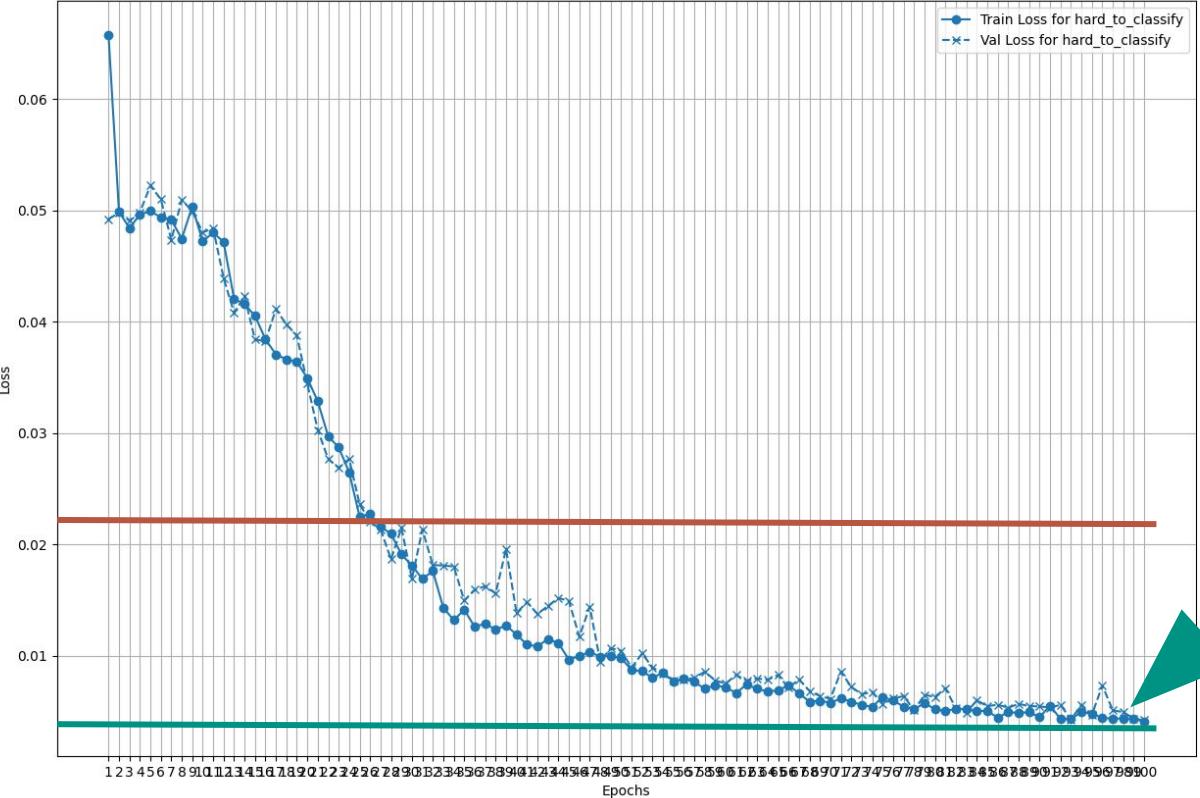


Średnia funkcja straty  
`BCEWithLogitsLoss` wynosi:

Zbiór treningowy: 0,019  
Zbiór walidacyjny: 0,0198  
Zbiór testowy: 0,0225

Rezultat uczenia się modelu - Wykres funkcji straty dla klasy 'cancerous'

JM42 - Training and Validation Loss per Class over Epochs



Środnia funkcja straty

`BCEWithLogitsLoss`

wynosi:

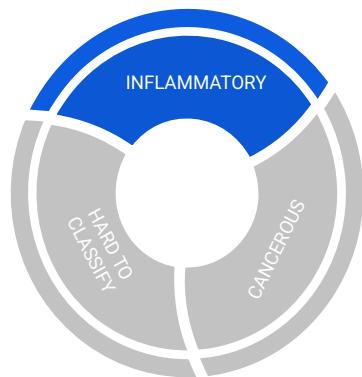
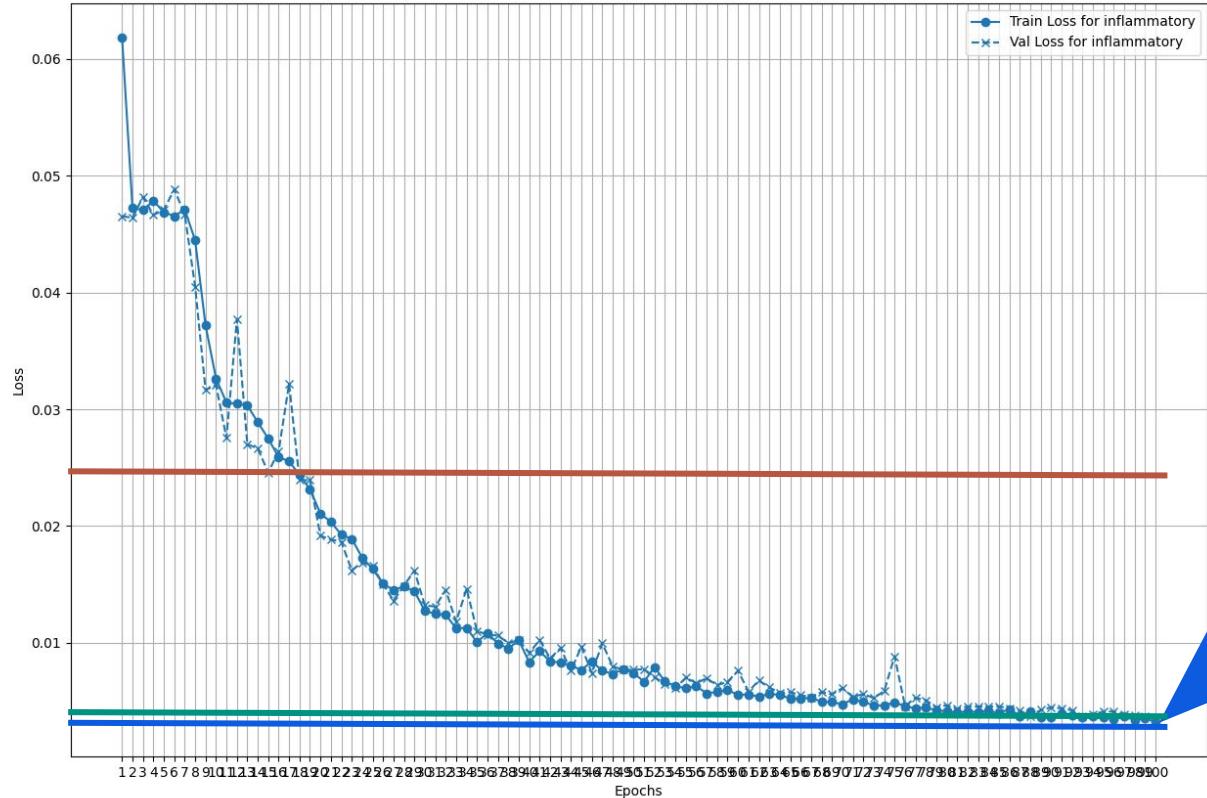
Zbiór treningowy: 0,0042

Zbiór walidacyjny: 0,0044

Zbiór testowy: 0,005

Rezultat uczenia się modelu - Wykres funkcji straty dla klasy 'hard to classify'

JM28 - Training and Validation Loss per Class over Epochs



Środnia funkcja straty

`BCEWithLogitsLoss`

wynosi:

Zbiór treningowy: 0,0033,

Zbiór walidacyjny: 0,0033

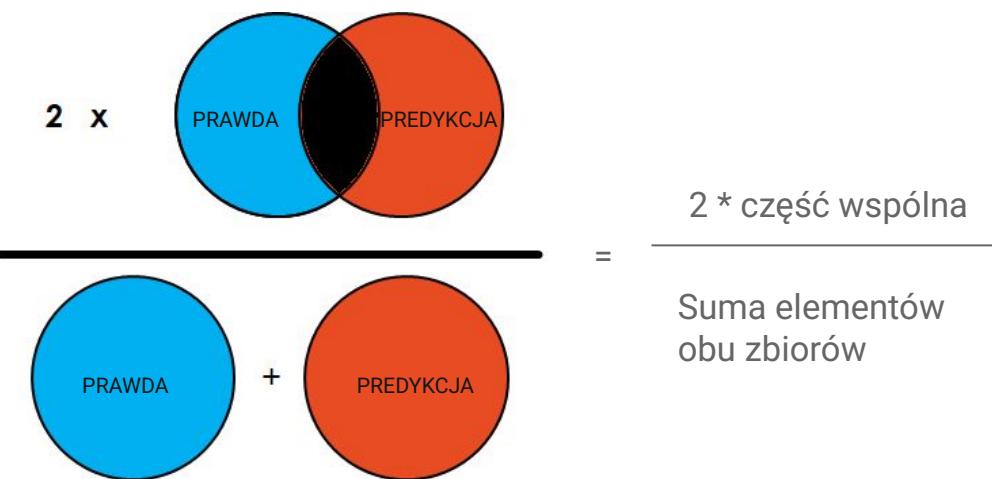
Zbiór testowy: 0,0044

Rezultat uczenia się modelu - Wykres funkcji straty dla klasy 'inflammatory'

# Miara podobieństwa- współczynnik Dice'a

## Współczynnik Dice'a-Sørensenia

$$Dice\ coef = 2 * |A \cap B| / (|A| + |B|)$$



Statystyczna miara podobieństwa dla zbiorów danych binarnych.

- Używana jest w segmentacji obrazów do oceny podobieństwa dwóch obrazów,
- bardziej czuła dzięki traktowaniu zbioru danych jako zbioru pikseli.
- Wysoki współczynnik -> wysokie podobieństwo
- Niski współczynnik, niewielkie podobieństwo

Cancerous - średni współczynnik dice:

Zbiór treningowy: 0,7278

Zbiór testowy: 0,7264

Hard to classify - średni współczynnik dice:

Zbiór treningowy: 0,9954

Zbiór testowy: 0,9956

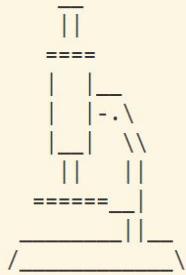
# Command Line Interface

Welcome to the Image Processing Program!

This program uses a UNet model to perform image segmentation on medical images.

You can load images from a specified directory, process them using trained models, and save the output masks to an output directory.

Enjoy the seamless experience!



To easily use the program, paste your images into the input folder using your explorer.

Use the default settings to generate three masks - cancerous, inflammation, and hard to classify.

You can also use an advanced configurator.

"

Would you like to use the default [def] or advanced [adv] settings? (def/adv):

# Command Line Interface

```
Would you like to use the default [def] or advanced [adv] settings? (def/adv): adv
Do you want to change the input directory? (yes/no) [default: C:\Users\jakub\Desktop\Data_Science\CAPSTONE\capstone_group5\src\input]: no
Files in the input folder:
[1] 19.png
[2] 20.png
[3] 21.png
[4] 22.png
[5] 4.png
Would you like to use all files? (yes/no)): 1
Do you want to change the output directory? (yes/no)) [default: C:\Users\jakub\Desktop\Data_Science\CAPSTONE\capstone_group5\src\output]: no
Do you want to change the models directory? (yes/no)) [default: C:\Users\jakub\Desktop\Data_Science\CAPSTONE\capstone_group5\src\models]: no
Available models:
[1] cancerous.pth
[2] hard_to_classify.pth
[3] inflammatory.pth
[4] model_JM 37.pth
[5] model_JM35.pth
Select models to use, separated by commas: 1,2
Processing Model: cancerous: 100% [ time left: 00:00 ]
Processing Model: cancerous: 100% [ time left: 00:00 ]
Processing Model: cancerous: 40% [ time left: 00:20 ]
```

# CZAS TRENINGU

Trening wykonywalny

lokalnie zajął łącznie:

**462009 sekund**

czyli

**128 godzin**

nieprzerwanej maksymalnej pracy GPU

- Długi czas trenowania warunkował ograniczoną ilość eksperymentów
- Znacznym ograniczeniem był czas wykorzystania GPU dostępny na platformie KAGGLE

Trening modeli docelowych (na przykładzie hard\_to\_classify):

- 55880 sekund
- 931 minut
- 15,5 godzin

# PODSUMOWANIE

Co uzyskaliśmy?

Narzędzie które:

- może pomóc w przesiewowej ocenie z uzyskaniem priorytetu oceny przez patomorfologa
- wskazanie podejrzanych obszarów może skrócić czas pracy patomorfologa

Co dalej?

Kolejne potencjalne etapy rozwoju modelu:

- Trening na nowych danych - np. zdrowych Pacjentów - zwiększenie dokładności modelu
- Segmentacja innych tkanek / nowotworów
- przy dostępie do danych klinicznych poszukiwanie kohort Pacjentów o podobnym obrazie mikroskopowym oraz innych cechach klinicznych, które warunkują pozytywną odpowiedź na specyficzne terapie

# WDROŻENIE - POTENCJALNE WYZWANIA

- **Utrudniona walidacja modelu na nowych zdjęciach-** niemożliwy do obliczenia Współczynnik Dice'a z uwagi na brak masek odtworzonych przez specjalistów patomorfologii ze zdjęć preparatów z biopsji (lub konieczność oceny nowych preparatów przez specjalistów- czas, koszty)
- Brak informacji o tym, czy model jest w stanie **rozróżnić pacjentów zdrowych od chorych-** w bazie danych znajdują się wyłącznie zdjęcia materiału biopsycznego pacjentów z postawionym już rozpoznaniem NSCLC.
- Tylko 3 zdjęcia w całym zbiorze danych nie zawierały masek reprezentujących komórki nowotworowe- **trudność z okrešleniem czułości i swoistości predykcji.**
- **Kosztowna moc obliczeniowa** związana z treningiem i testowaniem modelu

# Dziękujemy!

