



# Klasifikacija histopatoloških slika karcinoma dojke

Mašinsko učenje – seminarski rad  
Matematički fakultet

Natalija Drašković

# Uvod

- Karcinom dojke jedan je od najčešće dijagnostikovanih karcinoma i najčešći uzrok smrti kod žena širom sveta. Invazivni duktalni karcinom (IDC - Invasive Ductal Carcinoma) je najrasprostranjeniji tip karcinoma dojke sa oko 80% svih dijagnostikovanih slučajeva.
- Tehnike snimanja, poput patologije karcinoma dojke, pomažu u dijagnostici i praćenju bolesti. Međutim, identifikacija malignih ćelija može biti izazov, s obzirom na visoku heterogenost u apsorpciji agenasa za bojenje od strane samog tkiva.
- Kako bi se odredio nivo agresivnosti čitavog uzorka patolozi se obično fokusiraju na regione koji sadrže IDC, pa je jedan od uobičajenih koraka da se uzorak podeli na delove i onda razgraniče tačna područja IDC unutar čitavog uzorka kako bi se odredio stepen agresivnosti, kao i tačno područje zahvaćeno IDC-om.

# Skup podataka

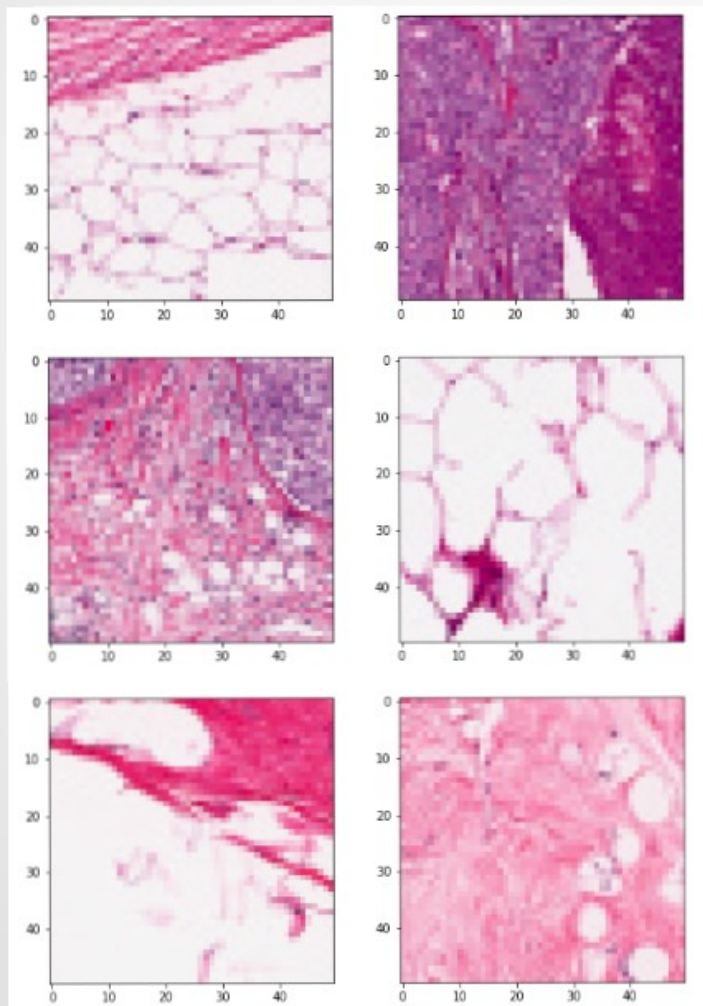
- Originalni skup podataka sastojao se od 162 cele slike uzoraka karcinoma dojke skeniranih pri uvećanju od 40x. Od toga je izdvojeno 277.524 zakrpa veličine 50 x 50 (198.739 IDC negativnih i 78.786 IDC pozitivnih).
- Korišćeni skup podataka je nešto drugačiji od originalnog, te on sadrži podatke za više pacijenata, njih 279, ali ne sadrži čitave slike uzoraka već samo unapred izdvojene zakrpe.
- Takođe, ukoliko pokušamo rekonstrukciju čitave slike uzoraka možemo videti da skup ne sadrži sve zakrpe jednog uzorka, već imamo i nedostajućih podataka.

# Skup podataka

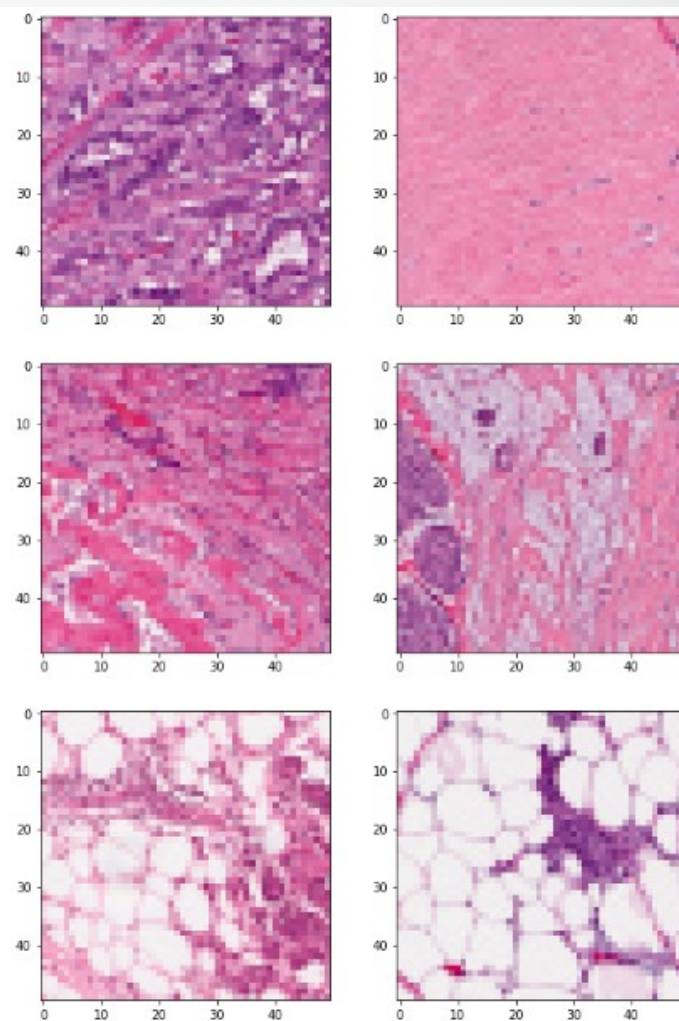
- Skup podataka za svakog pacijena sadrži poseban folder čiji je naziv id pacijenta, unutar koga postoje još dva foldera označena sa 0 i 1, unutar kojih se nalaze slike koje zdravo tkivo i tkivo IDC-a.
- Svaka od fotografija ima format ID\_idx5\_xX\_yY\_classN.png, gde je ID id pacijenta, X vrednost x koordinate zakrpe, Y vrednost y koordinate zakrpe i N klasa kojoj data zakrpa pripada, 0 u slučaju zdravog tkiva i 1 u slučaju IDC-a.
- Primer, 8863\_idx5\_x51\_y1251\_class0.png.

# Skup podataka

- Zdravo tkivo



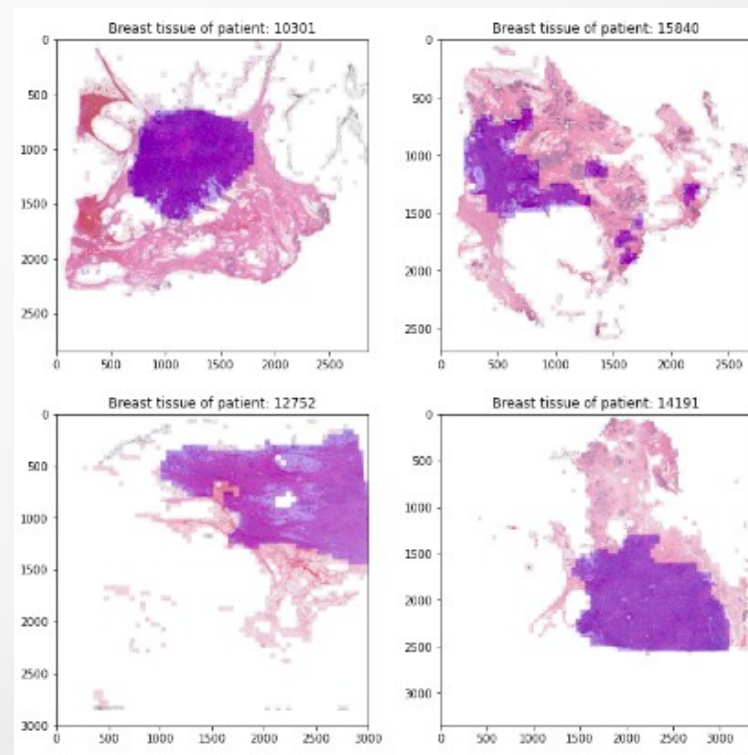
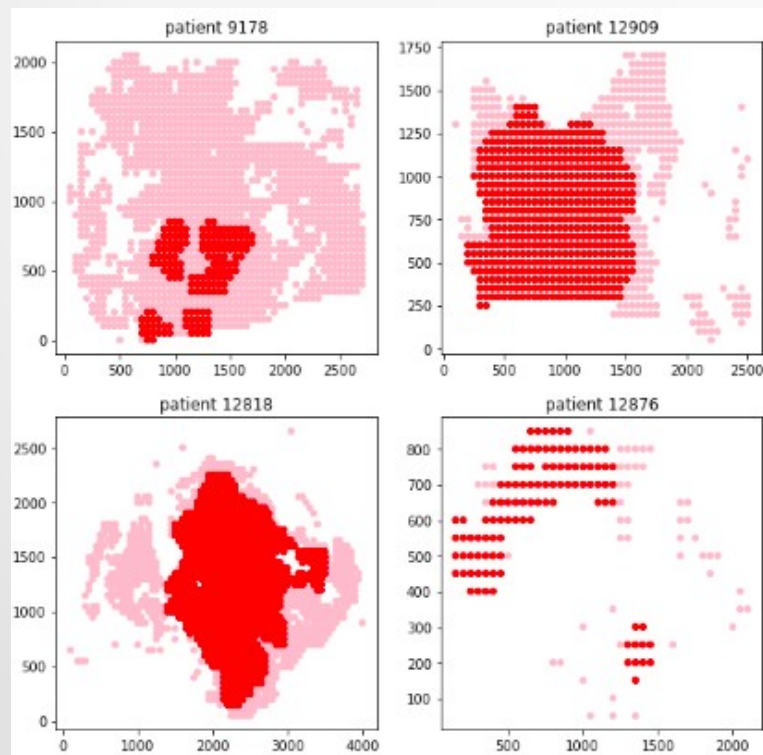
- IDC





# Skup podataka

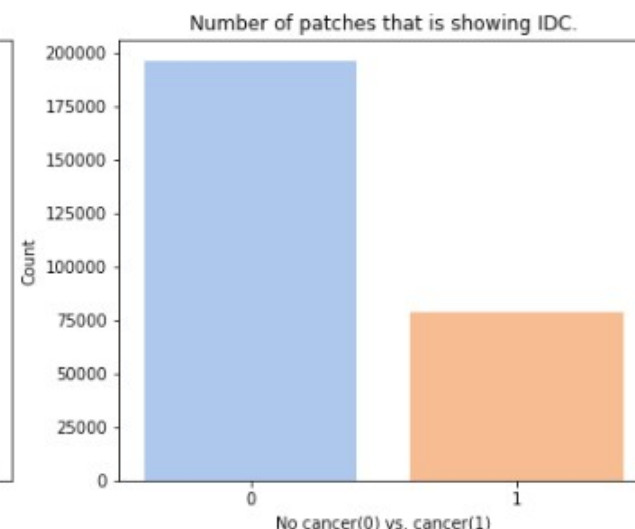
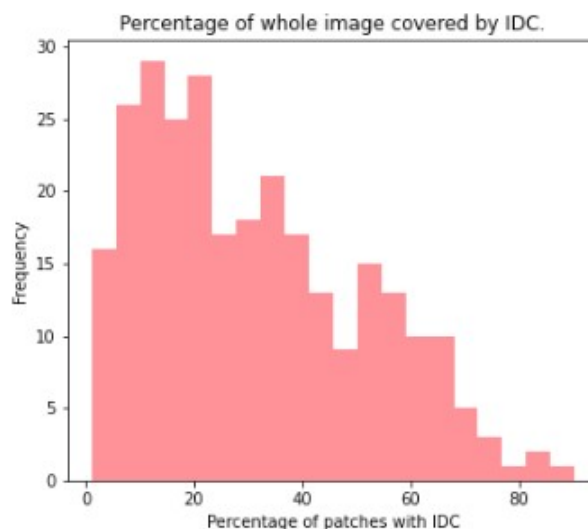
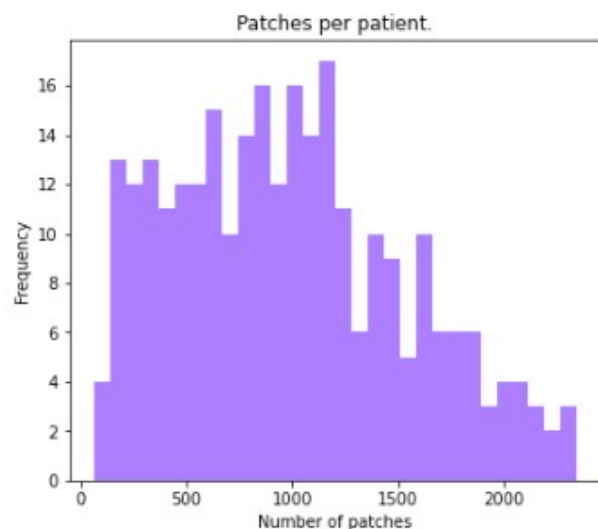
	patient_file_id	x	y	is_cancer	img
0	10253	1001	1001	0	[[[226, 164, 206], [224, 154, 196], [225, 175,...
1	10253	1001	1051	0	[[[219, 150, 197], [217, 158, 201], [228, 173,...
2	10253	1001	1101	0	[[[248, 245, 249], [248, 246, 248], [249, 246,...
3	10253	1001	1151	0	[[[249, 247, 249], [249, 247, 249], [249, 247,...
4	10253	1001	1201	0	[[[241, 231, 237], [246, 242, 245], [234, 222,...



# Skup podataka

Number of patches in Class 0 after filtering: 196454  
Number of patches in Class 1 after filtering: 78768  
Total number of patches after filtering: 275222  
Percentage of negative images: 71.38  
Percentage of positive images: 28.62

is_cancer	0	1
patient_file_id		
10253	0.872495	0.127505
10254	0.910377	0.089623
10255	0.665441	0.334559
10256	0.750000	0.250000
10257	0.672441	0.327559



# Modeli

- Kako nemamo čitave slike uzoraka, zbog čega se ne možemo osloniti na korišćenje pozicije zakrpe i posmatranje njene okoline, deo skupa podataka koji se odnosi na poziciju zakrpe korišćen je samo pri rekonstrukciji rezultata.
- Modeli su trenirani na skupu podataka koji predstavlja slike označene klasama 0 i 1, gde 0 predstavlja zdravo tkivo, a 1 IDC.
- Neko buduće unapređenje moglo bi biti posmatranje čitavog skupa slika koje pripadaju pacijentu i korišćenje lokacije zakrpe kao dodatan parametar modela. Na osnovu rekonstruisanih slika vidimo da je IDC obično lokalizovan, što podržava činjenicu da bi takav pristup imao potencijalno tačnije rezultate.



# Modeli

- Model 1

Model: "sequential\_2"

Layer (type)	Output Shape	Param #
conv2d_8 (Conv2D)	(None, 50, 50, 32)	896
max_pooling2d_8 (MaxPooling2D)	(None, 25, 25, 32)	0
dropout_10 (Dropout)	(None, 25, 25, 32)	0
conv2d_9 (Conv2D)	(None, 25, 25, 64)	18496
max_pooling2d_9 (MaxPooling2D)	(None, 12, 12, 64)	0
dropout_11 (Dropout)	(None, 12, 12, 64)	0
conv2d_10 (Conv2D)	(None, 12, 12, 128)	73856
max_pooling2d_10 (MaxPooling2D)	(None, 6, 6, 128)	0
dropout_12 (Dropout)	(None, 6, 6, 128)	0
conv2d_11 (Conv2D)	(None, 6, 6, 128)	147584
max_pooling2d_11 (MaxPooling2D)	(None, 3, 3, 128)	0
dropout_13 (Dropout)	(None, 3, 3, 128)	0
flatten_2 (Flatten)	(None, 1152)	0
dense_4 (Dense)	(None, 128)	147584
dropout_14 (Dropout)	(None, 128)	0
dense_5 (Dense)	(None, 2)	258

Total params: 388,674  
Trainable params: 388,674  
Non-trainable params: 0

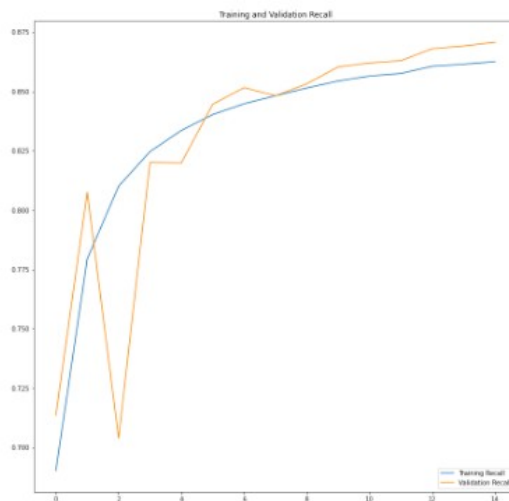
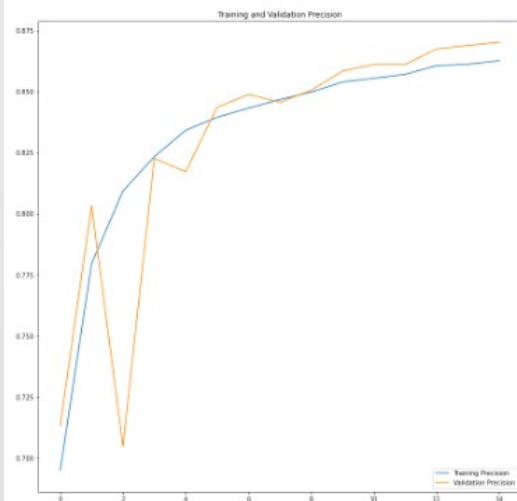
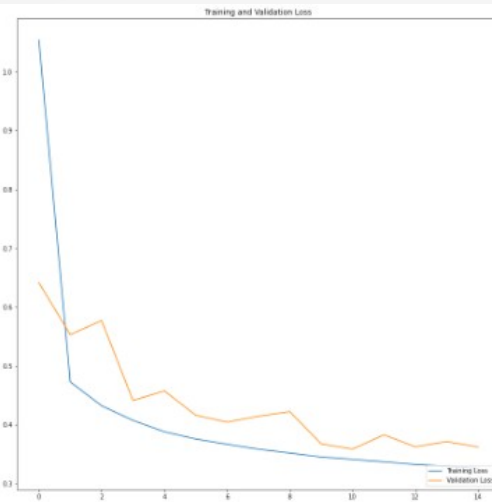
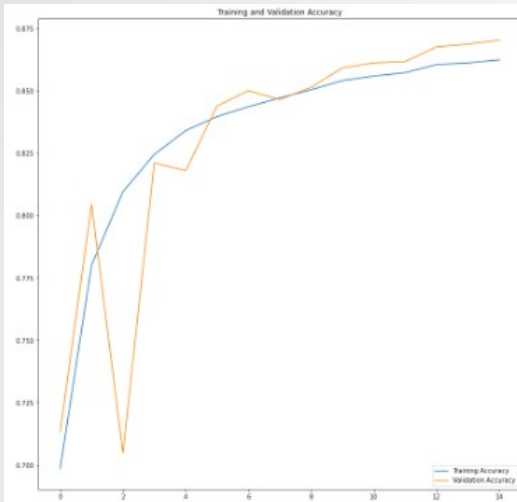
- Model 2

Model: "sequential\_16"

Layer (type)	Output Shape	Param #
conv2d_69 (Conv2D)	(None, 48, 48, 100)	2800
batch_normalization_21 (Batch Normalization)	(None, 48, 48, 100)	400
max_pooling2d_60 (MaxPooling2D)	(None, 24, 24, 100)	0
conv2d_70 (Conv2D)	(None, 12, 12, 256)	102656
batch_normalization_22 (Batch Normalization)	(None, 12, 12, 256)	1024
max_pooling2d_61 (MaxPooling2D)	(None, 5, 5, 256)	0
conv2d_71 (Conv2D)	(None, 5, 5, 384)	393600
batch_normalization_23 (Batch Normalization)	(None, 5, 5, 384)	1536
conv2d_72 (Conv2D)	(None, 5, 5, 384)	147840
batch_normalization_24 (Batch Normalization)	(None, 5, 5, 384)	1536
conv2d_73 (Conv2D)	(None, 5, 5, 256)	98560
batch_normalization_25 (Batch Normalization)	(None, 5, 5, 256)	1024
max_pooling2d_62 (MaxPooling2D)	(None, 2, 2, 256)	0
flatten_16 (Flatten)	(None, 1024)	0
dense_40 (Dense)	(None, 9216)	9446400
dense_41 (Dense)	(None, 4096)	37752832
dropout_68 (Dropout)	(None, 4096)	0
dense_42 (Dense)	(None, 4096)	16781312
dropout_69 (Dropout)	(None, 4096)	0
dense_43 (Dense)	(None, 2)	8194

Total params: 64,739,714  
Trainable params: 64,736,954  
Non-trainable params: 2,760

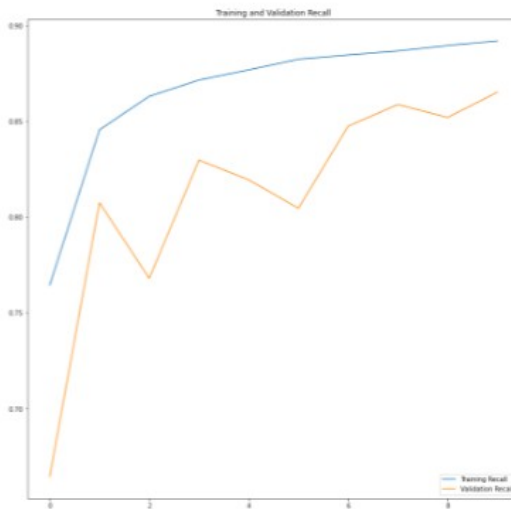
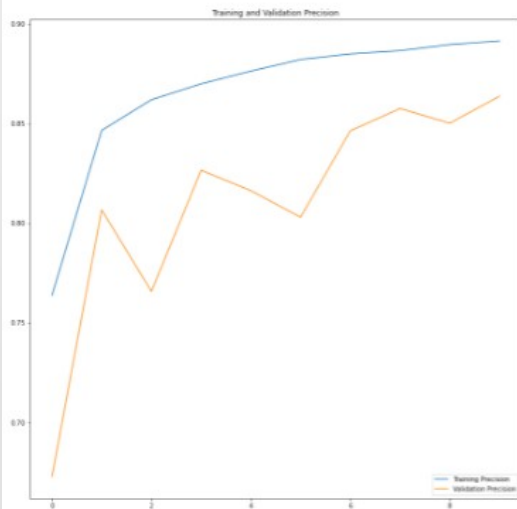
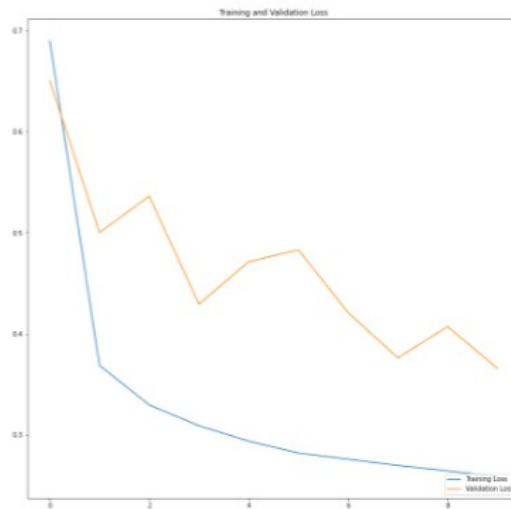
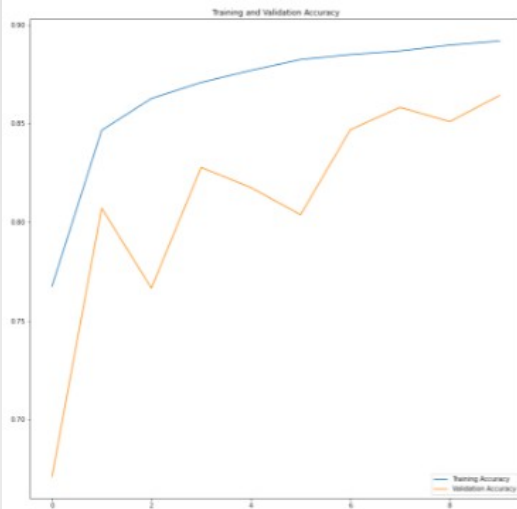
# Model 1



	precision	recall	f1-score	support
0	0.91	0.92	0.92	39009
1	0.80	0.77	0.78	15663
accuracy			0.88	54672
macro avg	0.86	0.85	0.85	54672
weighted avg	0.88	0.88	0.88	54672

```
[[36075 2934]  
 [ 3654 12009]]
```

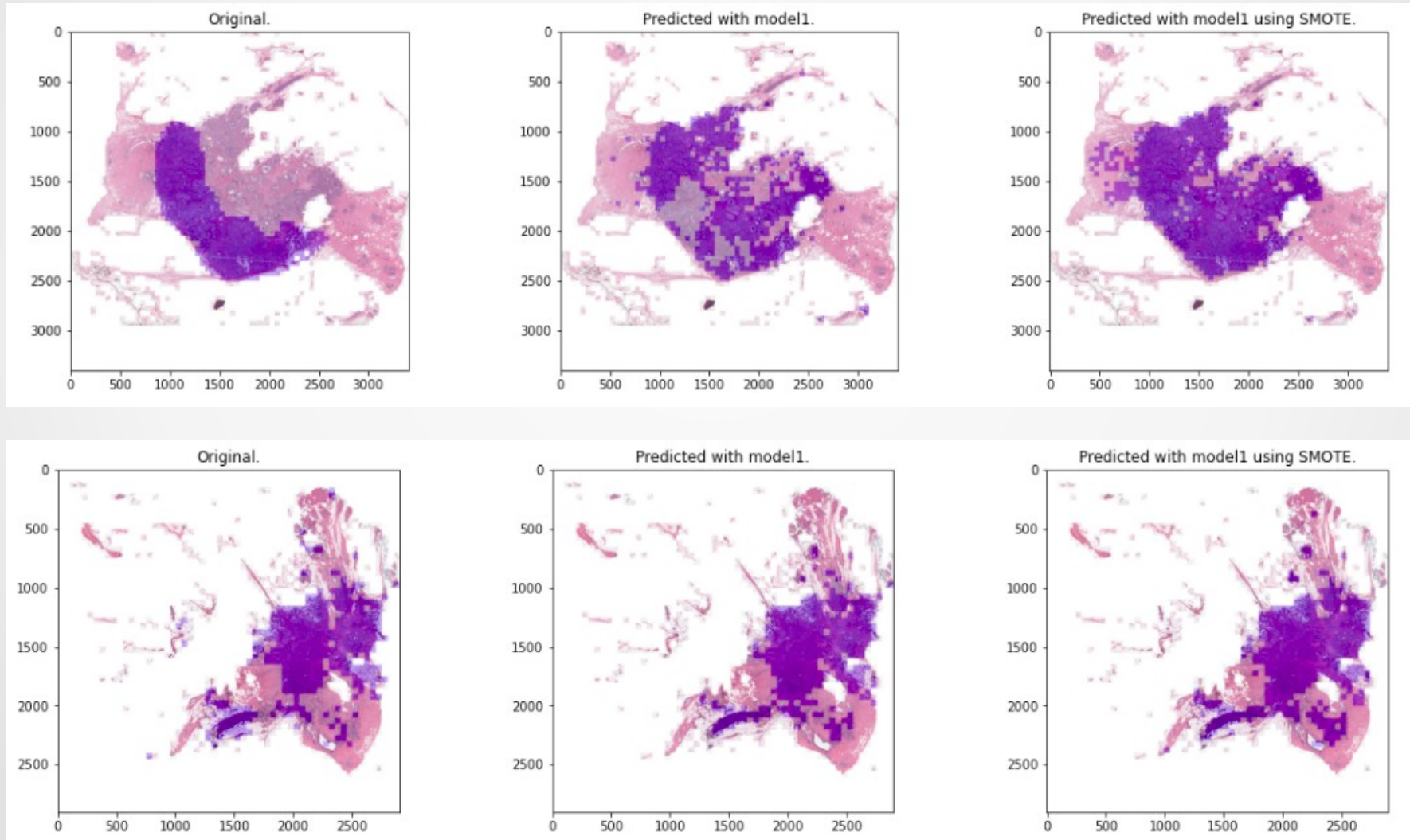
# Model 1 SMOTE



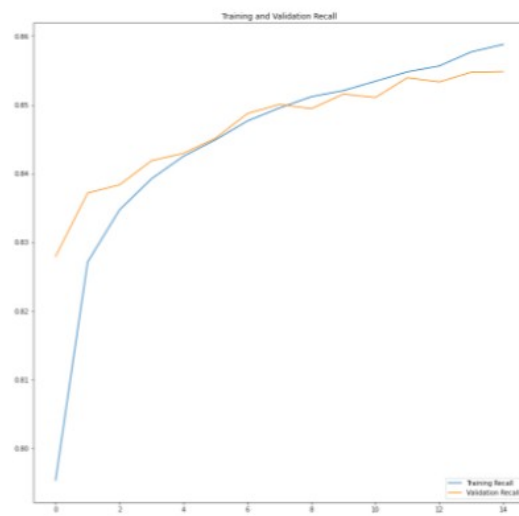
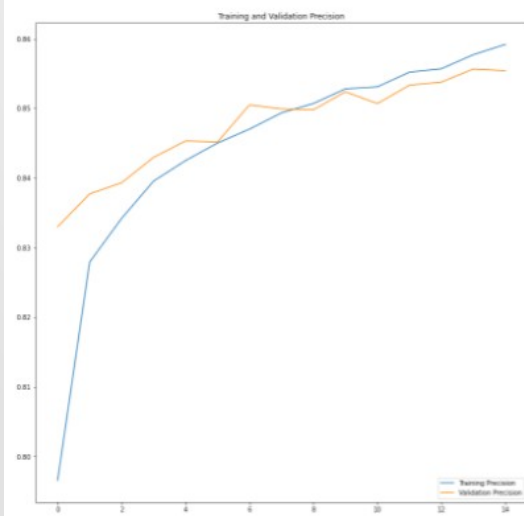
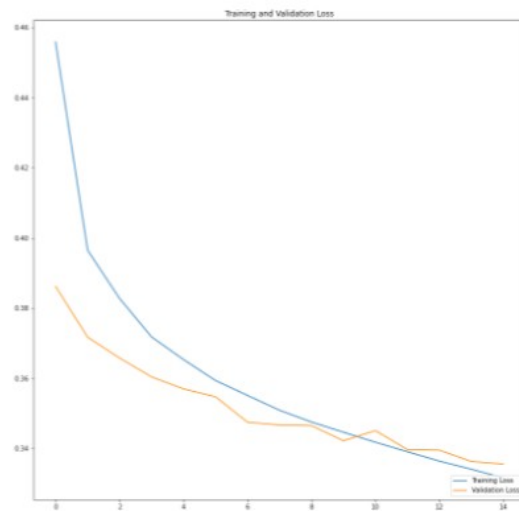
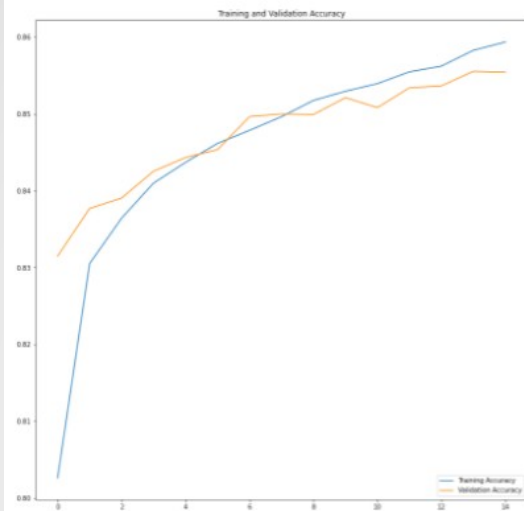
	precision	recall	f1-score	support
0	0.92	0.91	0.91	39009
1	0.78	0.80	0.79	15663
accuracy			0.88	54672
macro avg	0.85	0.85	0.85	54672
weighted avg	0.88	0.88	0.88	54672

```
[[35439 3570]  
 [ 3109 12554]]
```

# Model 1



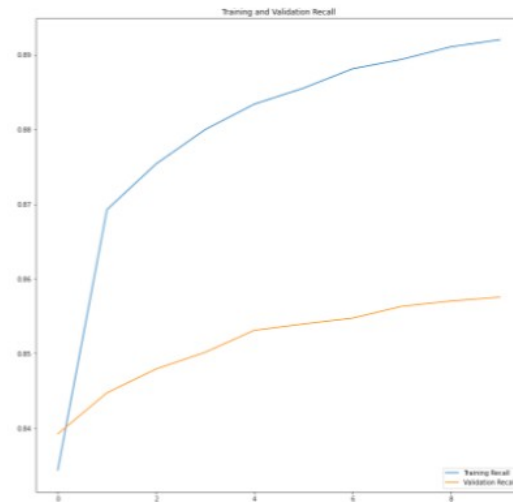
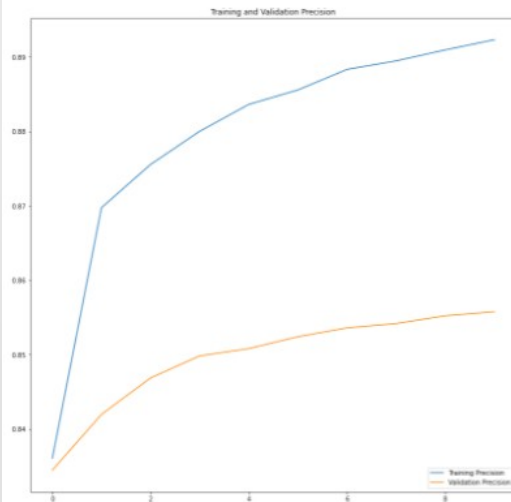
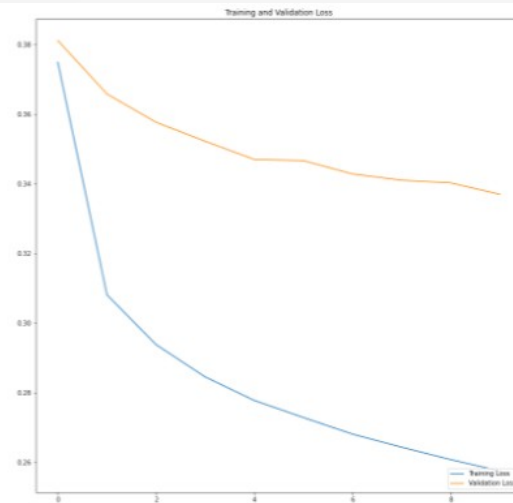
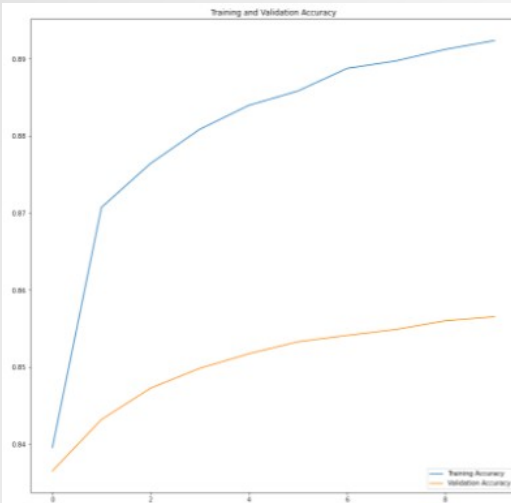
# Model 2



	precision	recall	f1-score	support
0	0.90	0.90	0.90	39009
1	0.76	0.75	0.75	15663
accuracy			0.86	54672
macro avg	0.83	0.83	0.83	54672
weighted avg	0.86	0.86	0.86	54672

```
[[35267 3742]  
 [ 3967 11696]]
```

# Model 2 SMOTE

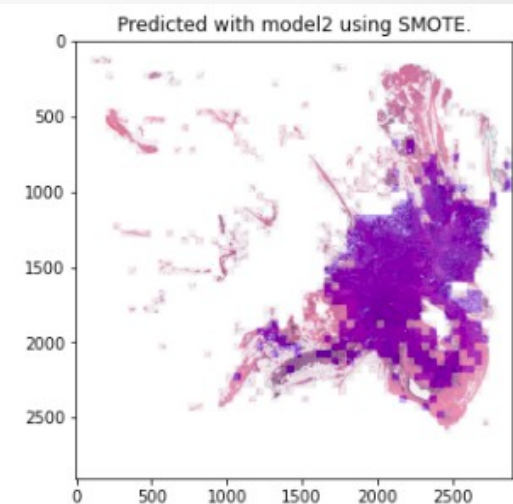
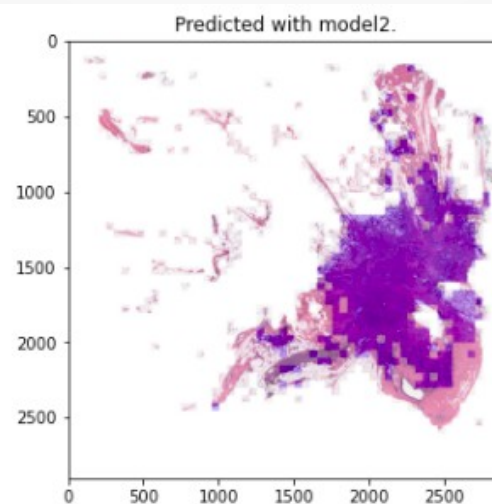
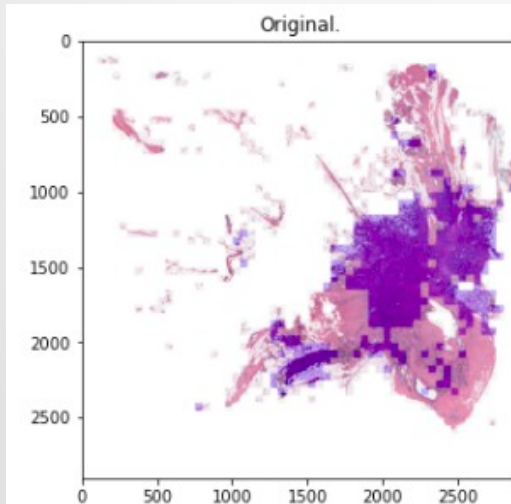
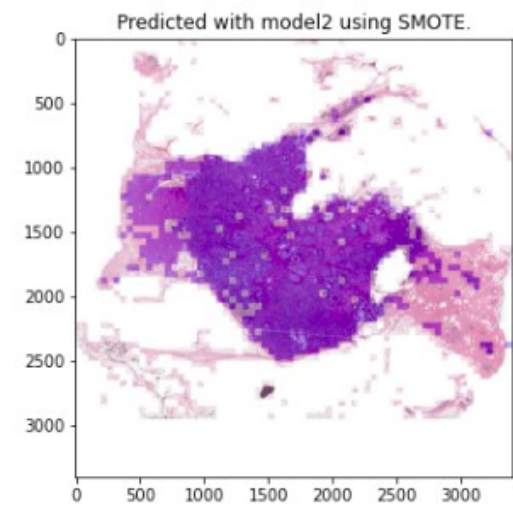
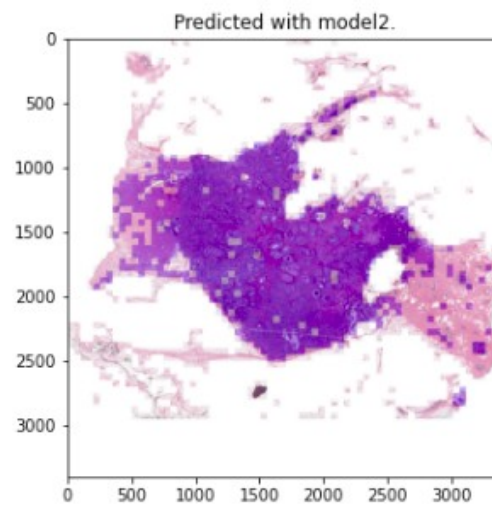
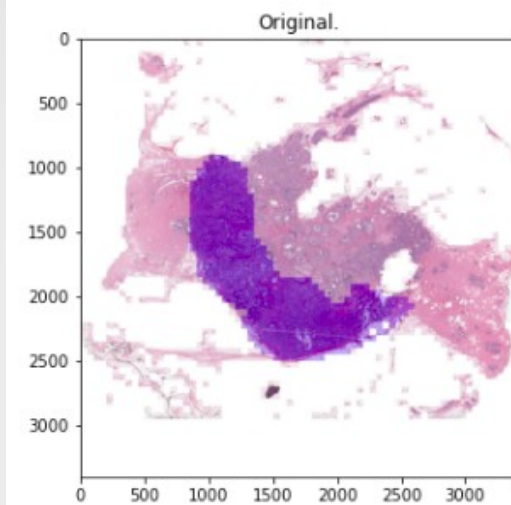


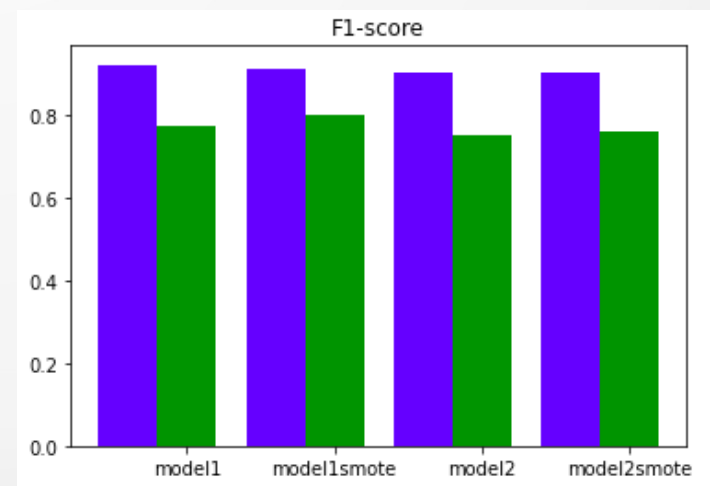
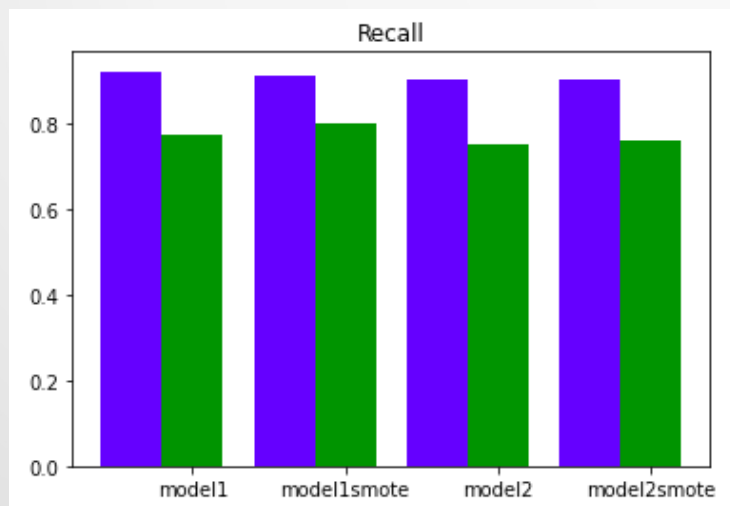
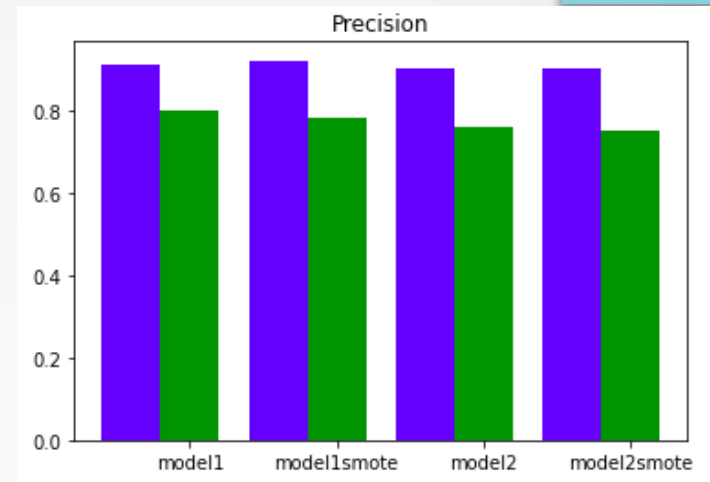
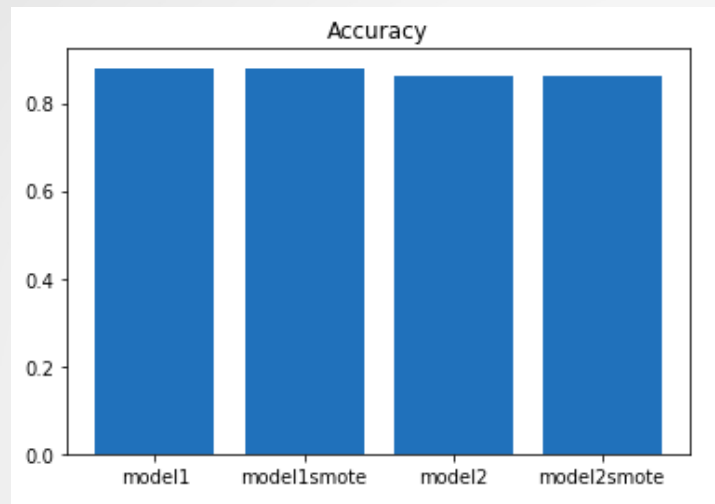
	precision	recall	f1-score	support
0	0.90	0.90	0.90	39009
1	0.75	0.76	0.76	15663
accuracy			0.86	54672
macro avg	0.83	0.83	0.83	54672
weighted avg	0.86	0.86	0.86	54672

```
[[35142 3867]  
 [ 3750 11913]]
```



# Model 2







Hvala na pažnji!

Pitanja?