

# Predicting colon cancer in biopsies

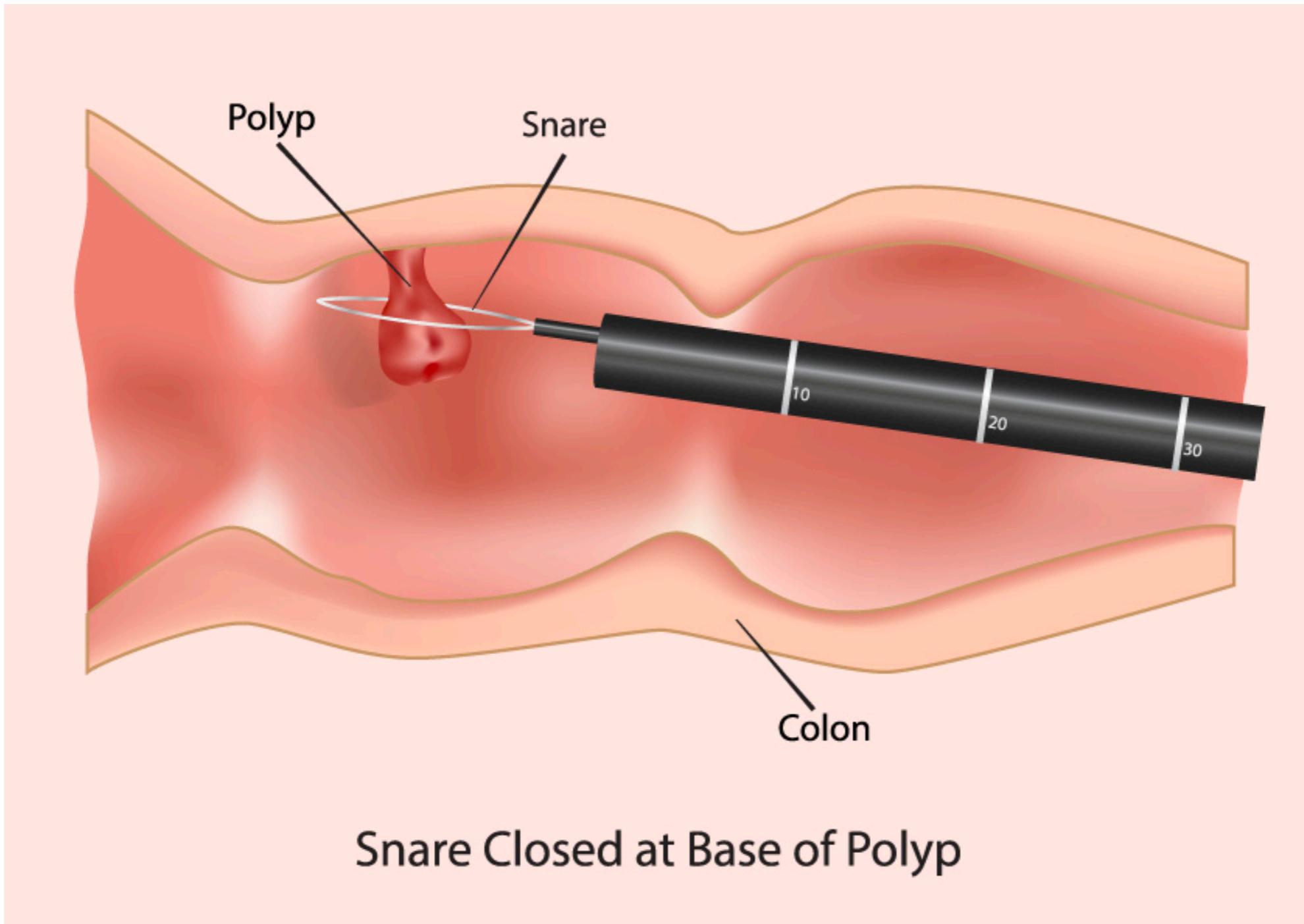
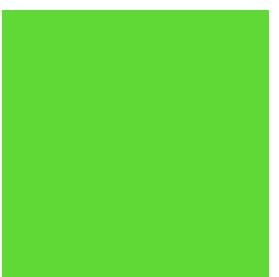
Ignasi Sols

# Introduction



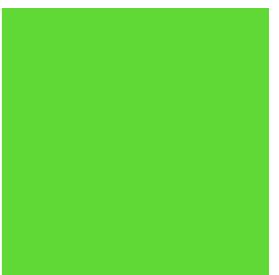
~149,500 americans  
Will be diagnosed of  
colorectal cancer in 2021

# Introduction





**Business impact:** Predict colorectal cancer in data biopsy slices will improve cancer diagnosis.



## Data Science Goal:

Predict colorectal cancer in data biopsy slices

**Dataset:** Kaggle dataset 'Lung and Colon Cancer Histopathological Images',  
<https://www.kaggle.com/andrewmvd/lung-and-colon-cancer-histopathological-images/code>

**10,000 images - already augmented**

**Tools:** Pandas, sklearn, Keras.

## Data Science path:



Build a Deep Learning model that predicts colorectal cancer in biopsy slices.

-The target had two classes:

**Positive** class: adenocarcinoma (TUMOR)

**Negative** class: benign tissue

-The data was **balanced: 5000** Recall: 0.69

**augmented images for each class**

-Images were re-scaled to 128x128

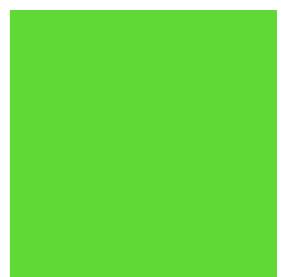
-Trained *Logistic Regression* and  
*Random forest baseline* models.

-Simple CNN.

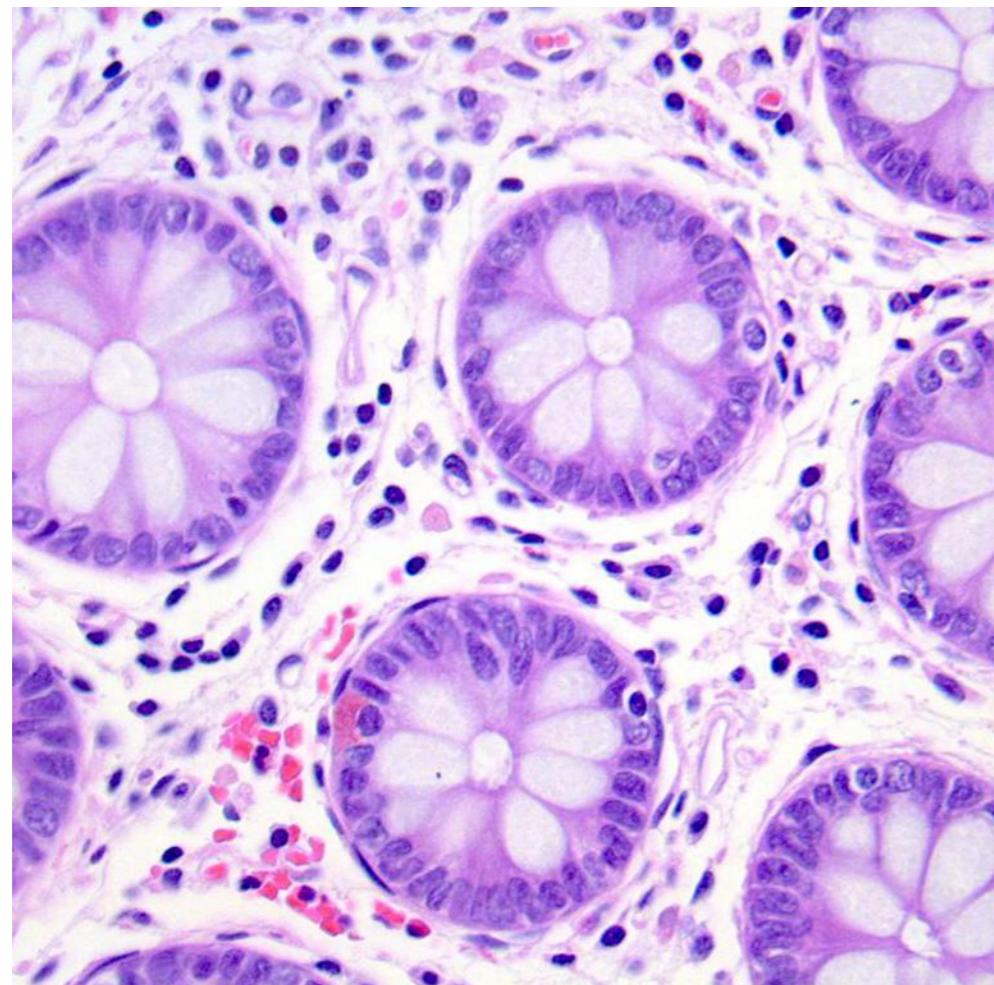
-CNN with transfer learning  
(mobileNet V2)

## Assumptions/Risks

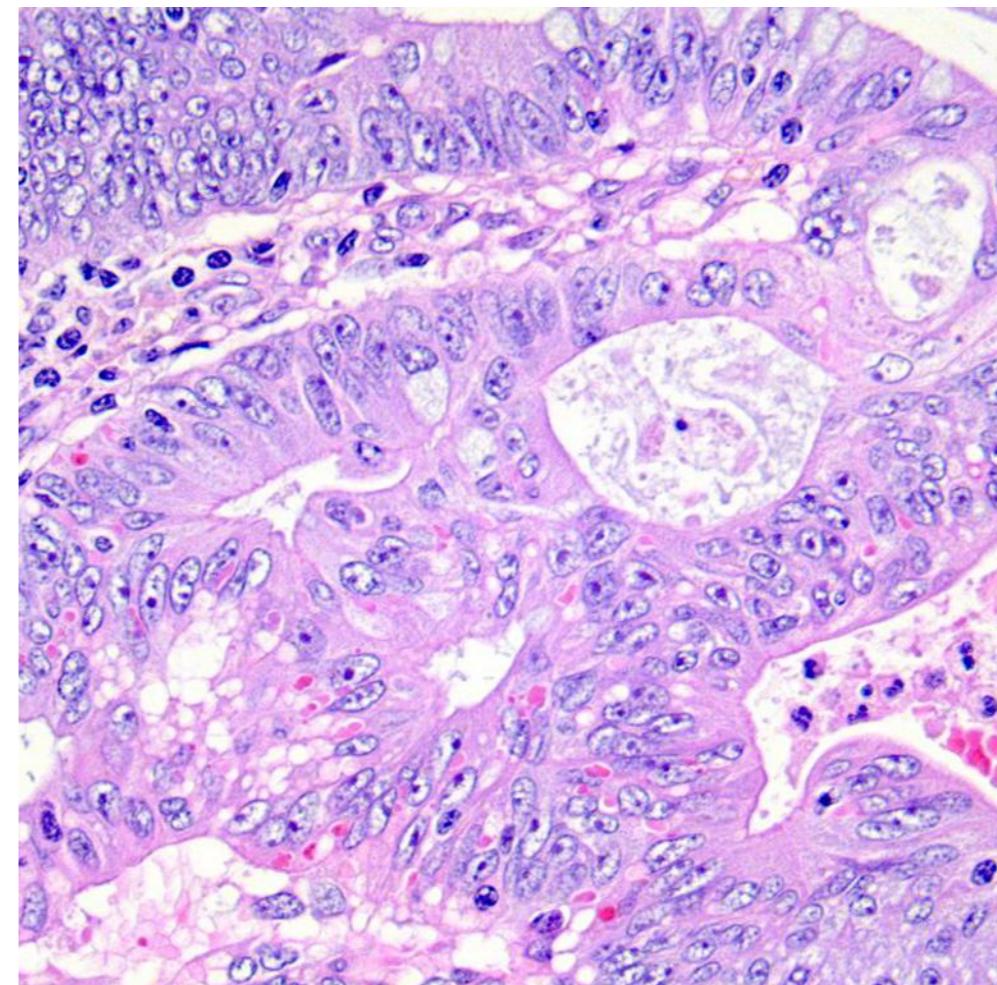
- The dataset images were already augmented and were in a unique folder, with no information regarding to which original images did they belong. Therefore, data bleeding is to be expected.
- The cancer stage of the tumors was not provided, and this might hinder how the model generalizes to data outside this data set.



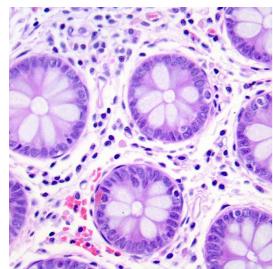
**Benign**



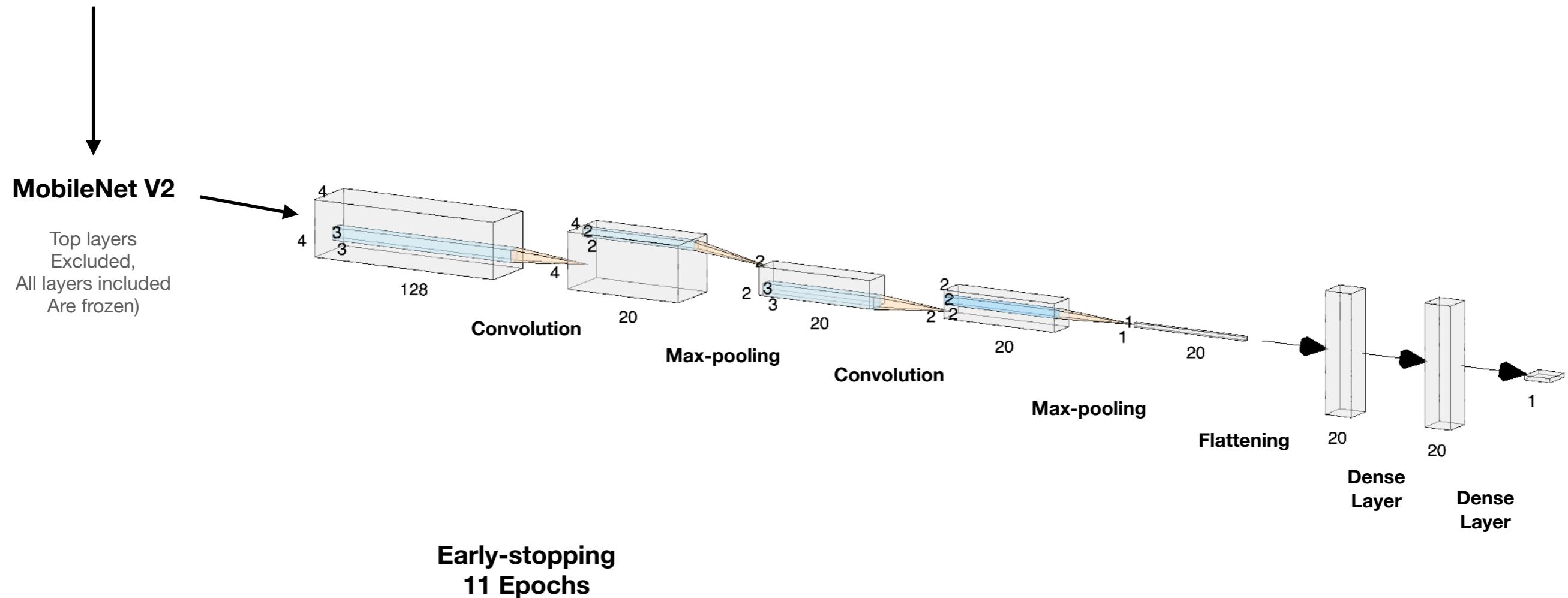
**Tumor**



**Benign**

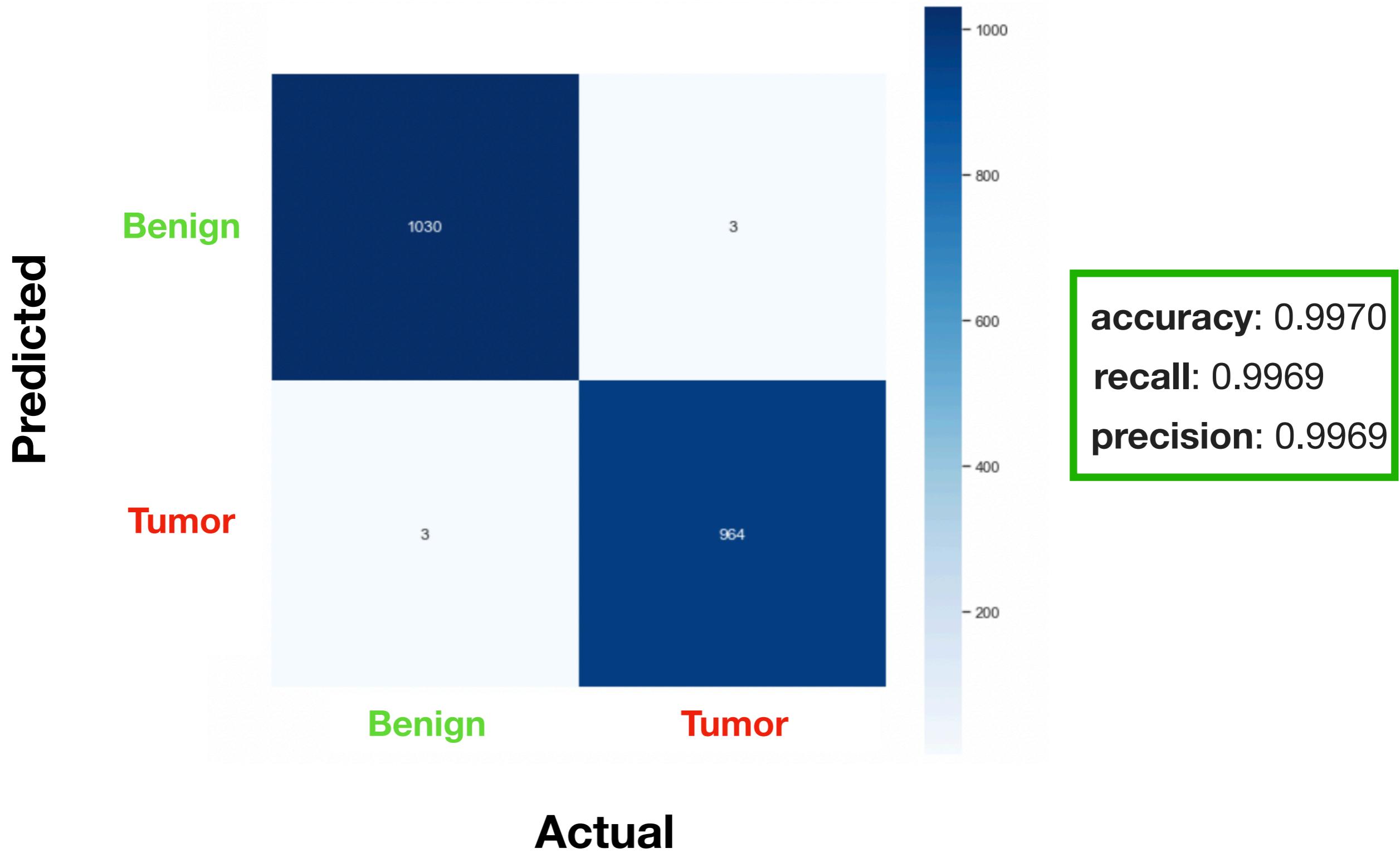


**128 x 128 x 3**



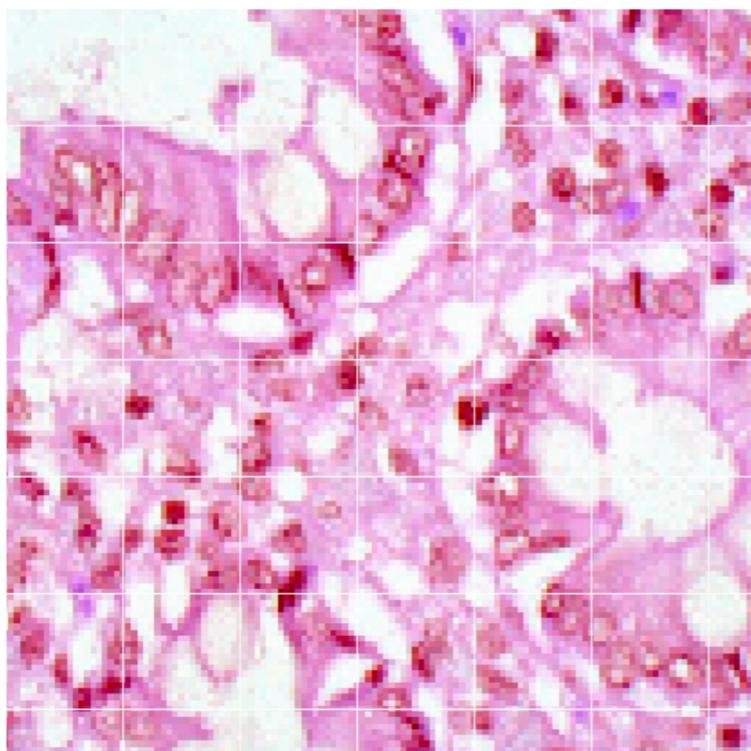
Total parameters: 2,492,465. Trainable parameters: 234,481

# Results

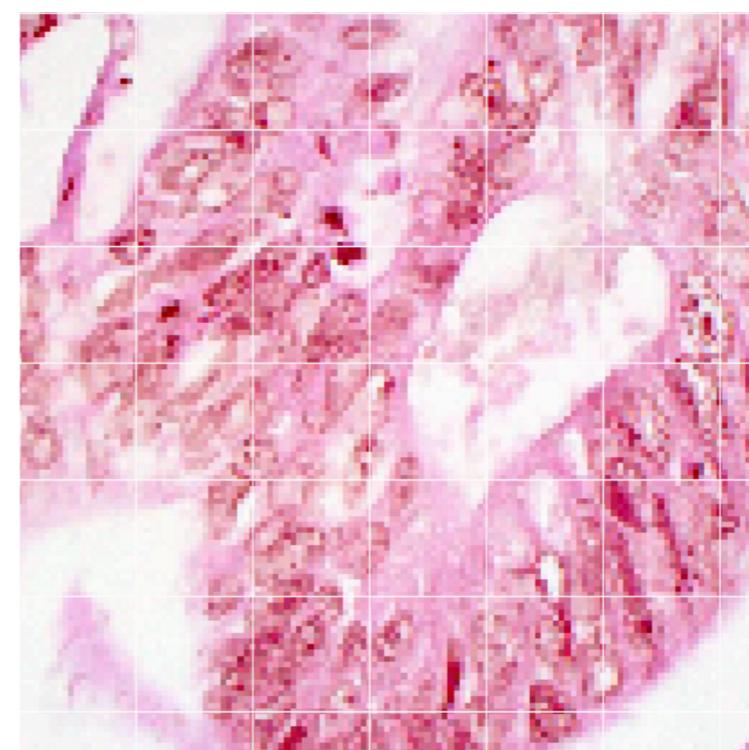


# Misclassifications

**False Positive**



**False Negative**



## **Future directions:**

- Use another dataset that allows to avoid data bleeding (e.g., NCT-CRC-HE-100K).
- Consider transforming it into a multi-class problem (NCT-CRC-HE-100K has different classes for non-cancerous tissue).

**Thanks!**

## Appendix:

### Logistic Regression

**accuracy:** 0.74  
**recall:** 0.69  
**F beta score (beta = 2) :** 0.70  
**Precision :** 0.74

### Random Forests (BASELINE)

**accuracy:** 0.80  
**recall:** 0.77  
**F beta score (beta = 2) :** 0.78  
**Precision :** 0.81

### CNN - MobileNet V2

**accuracy:** 0.9970  
**recall:** 0.9969  
**precision:** 0.9969

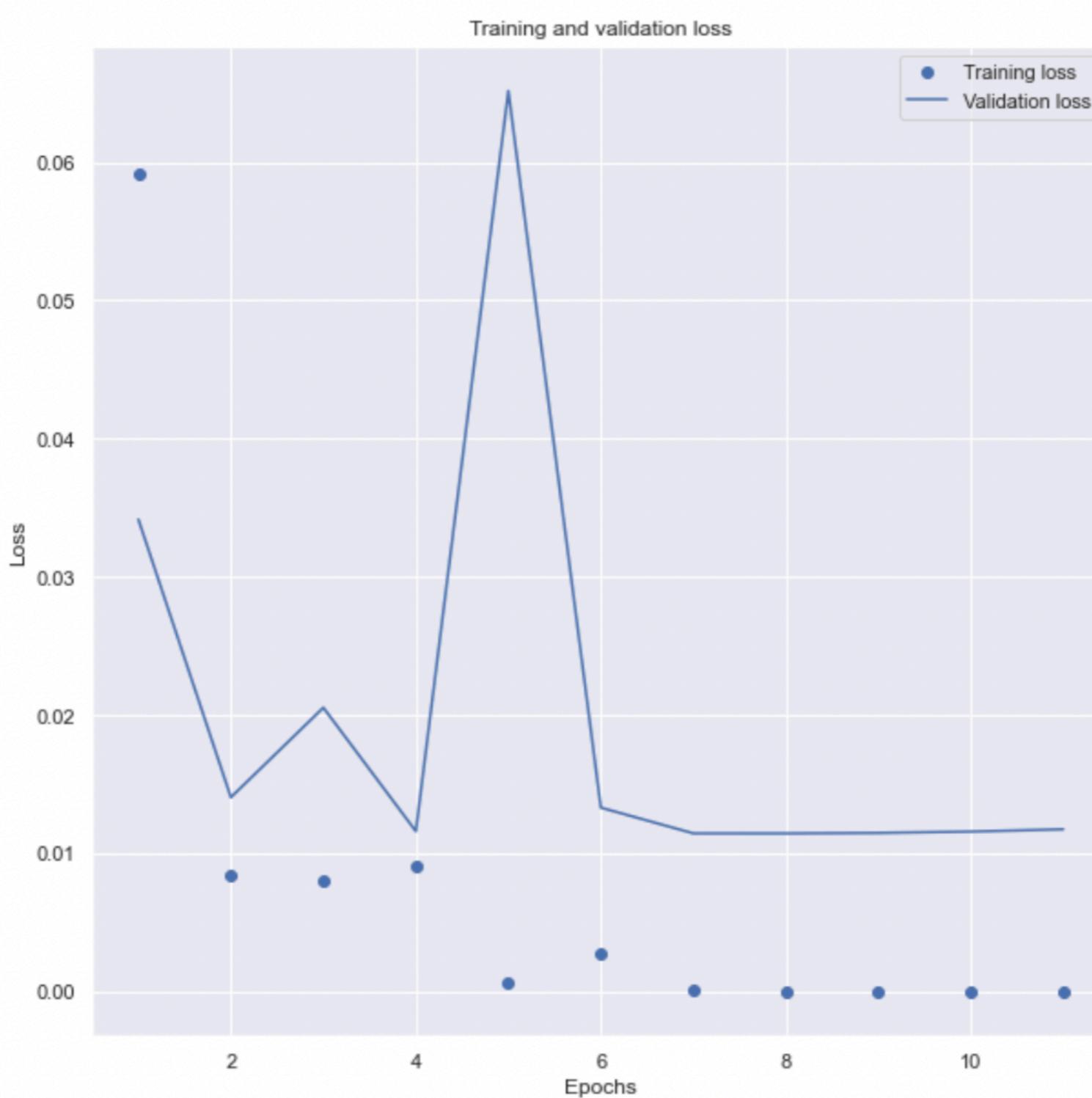
### CNN

**accuracy:** 0.92  
**recall:** 0.99  
**Precision:** 0.87



# Appendix:

## CNN - with Transfer Learning (mobileNet V2)



# Appendix:

## Baseline (PCA: 2 principal components)

