# Machine Learning for Faster, More Accurate Tumor Detection

Phil McNamara[1], Carolyn Brewster[1], Anna Lundberg[1], David Degnan[1], Jaclyn Smith[2], Kevin Matlock[2], Gans Srinivasa[2]

[1]Bioinformatics & Genomics Masters Program, University of Oregon [2]Omics Data Automation, Portland, OR

## INTRODUCTION

Over one third of Americans will be diagnosed with cancer in their lifetime.[1] Treatment is expensive, and the lengthy diagnostic process contributes significantly to the cost. Tumor cells function differently from normal cells and those differences are reflected in their appearance. Currently, a highly trained physician visually examines patient tissue samples to identify cancer. Recent advances in computer vision that power technologies such as facial recognition and self-driving cars can potentially be applied to perform this task. Integrating automation into cancer diagnosis will reduce both cost and delays for better outcomes overall.

## HYPOTHESIS

The visual characteristics of tissue in slide images can be used to train a generalizable machine learning model to detect tumors.

## WHY MACHINE LEARNING?

Programming computers to detect cancer is unrealistic; even pathologists have difficulty explaining how they identify tumors. Instead of teaching a computer a set of rules to identify tumors, machine learning allows the computer to define its own rules, then automatically adjust and apply them more effectively.

## WHY A CONVOLUTIONAL NEURAL NETWORK?

**Convolutional filters** look at subsets of an image to find patterns at different scales based on the size of the convolution. Applying these filters in a "sliding window" fashion preserves spatial features. **Neural networks** contain many layers, ensuring that the network can adapt to very complex tasks like image recognition. Our two-part system uses a core network (Inception v3) that has been designed for visual processing.[2] An additional layer of rules help the network adapt specifically to tumor slide images and enhance its transferability between different tumor types.

## GOALS

- Process images into a machine-learnable format
- Train model on well-characterized (CAMELYON) slide set
- Examine efficacy of model with additional training images
- Evaluate model accuracy on novel slides
- Run model on previously unstudied slides, confirm results with a physician

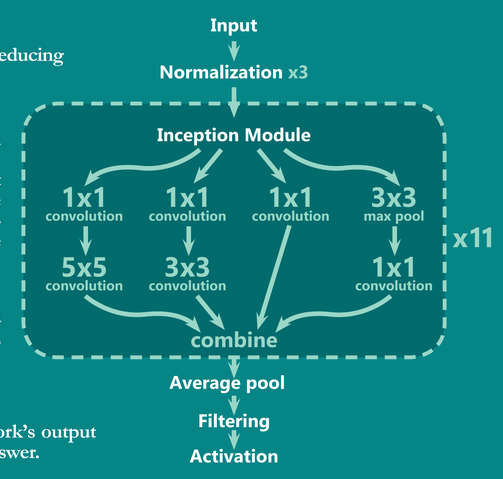
## THE NETWORK

**NOMALIZATION LAYERS**
help "even out" the data, reducing non-meaningful information.

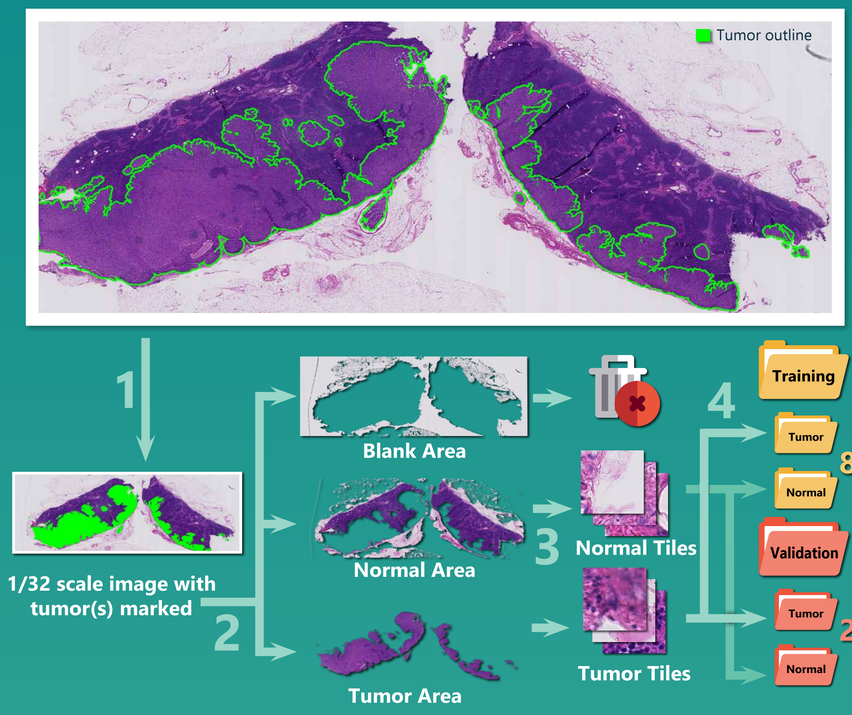
**CONVOLUTION LAYERS**
use a sliding window approach to look at smaller pieces of the image. Larger ones detect large-scale patterns, while smaller ones detect small-scale details and help reduce data size.

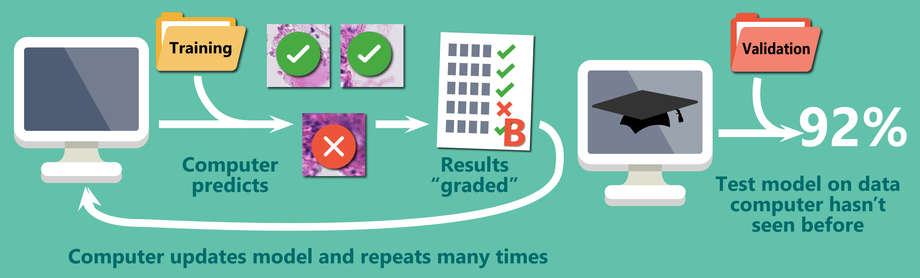
**POOLING LAYERS**
condense data by using a single value (average, max, min) to represent a region.

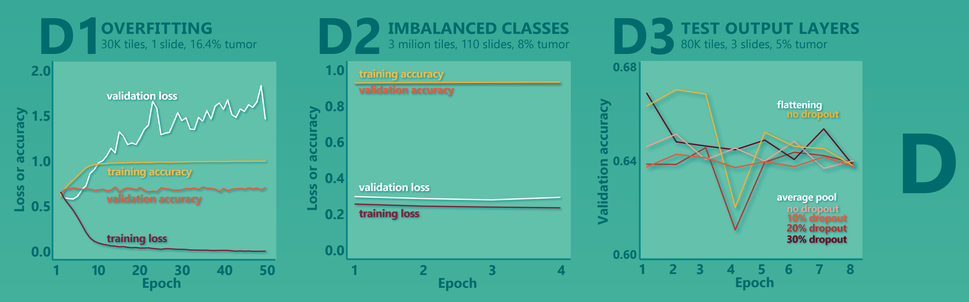
**ACTIVATION LAYERS**
apply a function to the network's output in order to generate a final answer.

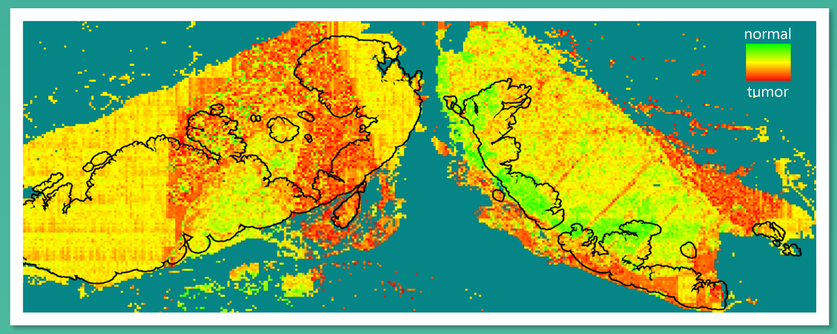



Implemented using TensorFlow with Keras.




## A. SLIDE IMAGES

Tissues from potential tumors are sliced into layers just a few cells thick. Dyes stain the nucleus of each cell blue and the interior red. A powerful microscope images the slides and a histologist looks at the blown-up images to outline any tumors.

## B. IMAGE PROCESSING

High-resolution slide images are far too large for a computer to learn on, so we divide the images into 256x256 pixel **tiles**. Mostly blank tiles are discarded and the rest are labeled as "normal" or "tumor." 80% of each category is used for training and the last 20% is saved for validation.

## C. MACHINE LEARNING

The tiles are converted into a rank 4 **tensor** of 1) number of samples, 2) RGB color values for pixels, 3) x coordinates for the tile, and 4) y coordinates for the tile.

The network applies internal rules known as **weights** to these tensors to predict the probability of tumor in a tile. The **loss function** checks this prediction against the correct answer and tells the computer how far off it was. The computer uses this data to update its weights in a process called **back-propagation**. This cycle is repeated many times using **training** data. Once the model is finalized it is tested on never-before-seen **validation** data to get a final accuracy score.

## D. MODEL TRAINING

One pass through the entire dataset is referred to as an **epoch**. We train the model for many epochs, testing to ensure that is *learning* from the data but not *memorizing* it. Validation accuracy is the true measure of predictive ability.

**D1.** On a low-tumor training set, the model quickly learns to achieve 92% accuracy by always predicting "not tumor."

**D2.** **Overfitting** occurs when a model simply memorizes the training data. Training accuracy quickly approaches 100%, but validation accuracy stagnates at 70%.

**D3.** Output configurations affect results. **Dropout** randomly leaves out a percentage of the weights during training to help reduce overfitting. **Average pooling** uses the mean weight of each layer's output, while **flattening** condenses the tensors without discarding data. Our final network uses average pooling with 20% dropout.

## E. FINDING TUMORS

We ran the trained model on a brand new slide to create the predictive heatmap at left. Green indicates a low predicted probability of tumor and red indicates high probability.

## FUTURE DIRECTIONS

- Improve model by training for more epochs
- Evaluate generalizability to different types of tissue images
- Physicians corroborate the model's predictions on previously uncharacterized tissue images to test prediction reliability and improvement to annotation time
- Apply to slide images linked to comprehensive database of patient and treatment data to improve time to diagnosis and development of an individual treatment plan

## ACKNOWLEDGEMENTS

This work benefited from access to the University of Oregon high performance computer, Talapas, and the knowledge and expertise of the Talapas team, particularly Rob Yelle and Michael Coleman. Jake Searcy provided expert advice on machine learning. We would like to thank Omics Data Automation and UO's Bioinformatics & Genomics Master's Program for their support and guidance.

UNIVERSITY OF OREGON

BGMP@UO Bioinformatics and Genomics Master's Program

OHSU

## REFERENCES

1. American Cancer Society, 2018. *Cancer Facts & Figures 2018.*
2. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition.*

Code, poster, & supplementary materials available at git.io/fpQBU