Active Learning System for Digital Pathology: A New Tool for Interactive Optimization of Classifier Models

F. Sheikhzadeh[1], J. Larsen[1], H. Fellows[1], J. Martin[1], N. Murari[1], Q. Wong[1], and M. Khojasteh[1]

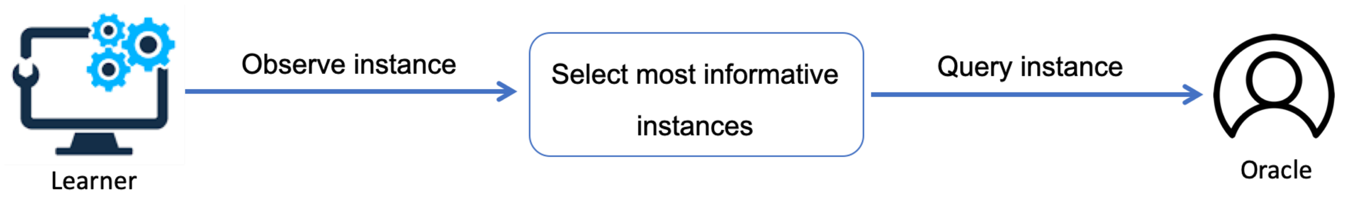
[1] Digital Pathology, Roche Tissue Diagnostics, Santa Clara, CA, USA. **Email: fahime.sheikhzadeh@roche.com**

1 – Background

The goal of **digital pathology** is analyzing whole slide images (WSI) to extract diagnostic and prognostic information.

- Machine learning models are used to detect objects or patterns in WSIs that are related to a specific biological process.
- Training the models requires a large set of manually labeled ground truth, which is tedious and time-consuming to collect.

Active learning is a special case of **machine learning in** which a **learning** algorithm is able to interactively query the user (or some other information source) to obtain the desired outputs at new data points.



Active learning approaches can:

- **Dramatically reduce** the time needed for complete and accurate labeling
- Increase the accuracy for **difficult-to-classify** instances

Two components of any Active Learning approach:

Learning Module
(Machine Learning Model/s)

Sample Selection Module
(Active Learning query strategy to measure how informative an unlabeled instance is)

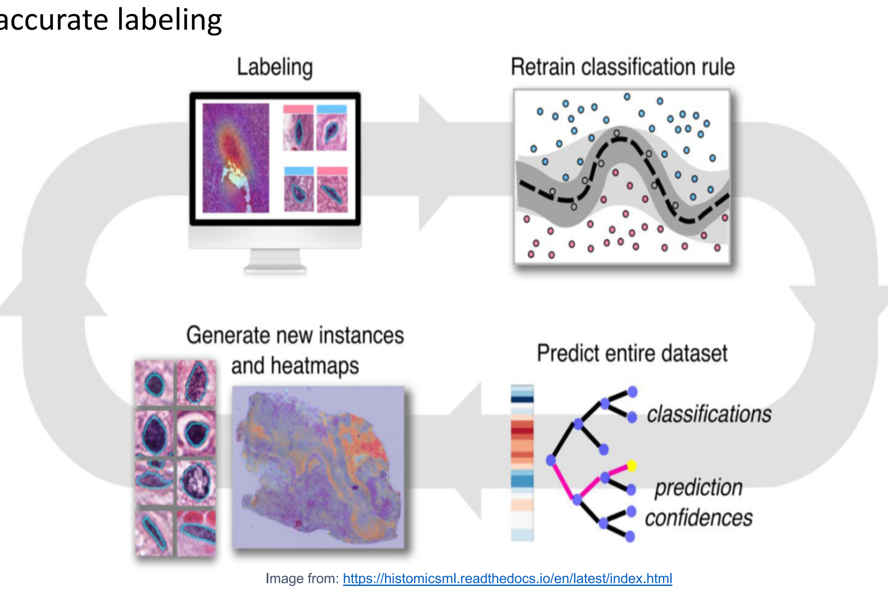


Image from: https://histomicsml.readthedocs.io/en/latest/index.html

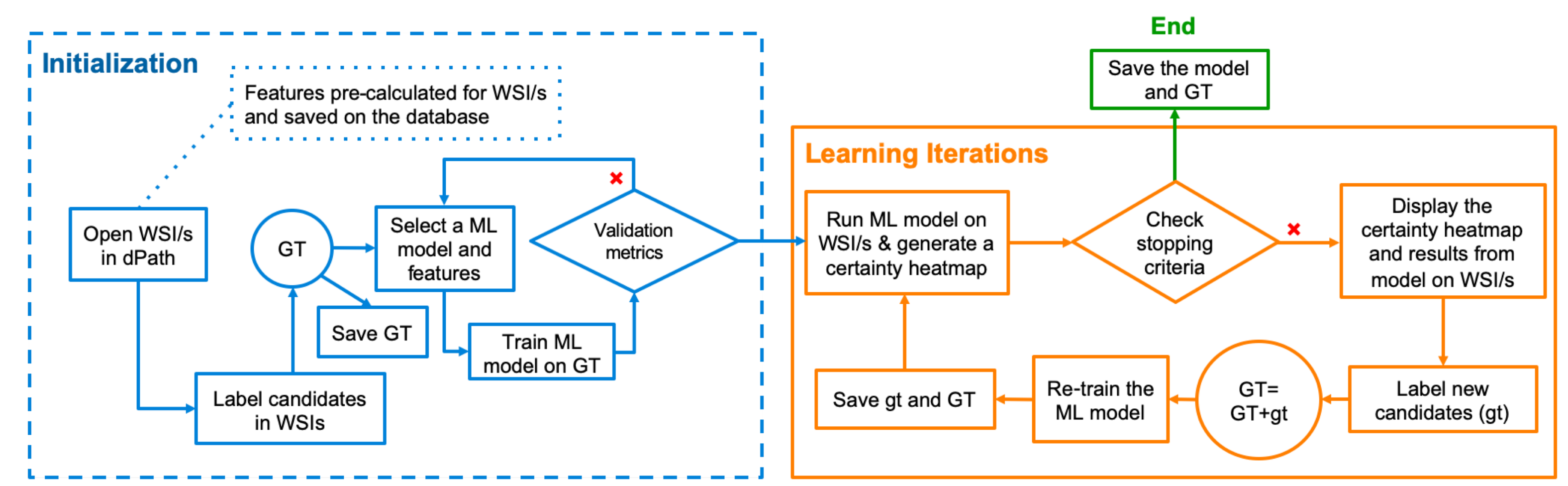
2 – Objective

Developing an **active learning system** that guides the user to focus their labeling effort on those examples that contribute the most to the learning performance of the machine learning model.

By employing our active learning system:

- Machine learning models can be trained using **far less labeled data**
- Users can train or optimize classifier models while **iteratively collecting the ground truth.**

Designed Workflow:



Any user, from pathologists to imaging scientists, can use our system to collect ground truth (annotate objects or regions of interest on WSIs) and train a classifier. The designed system has a great impact on DP applications as it:

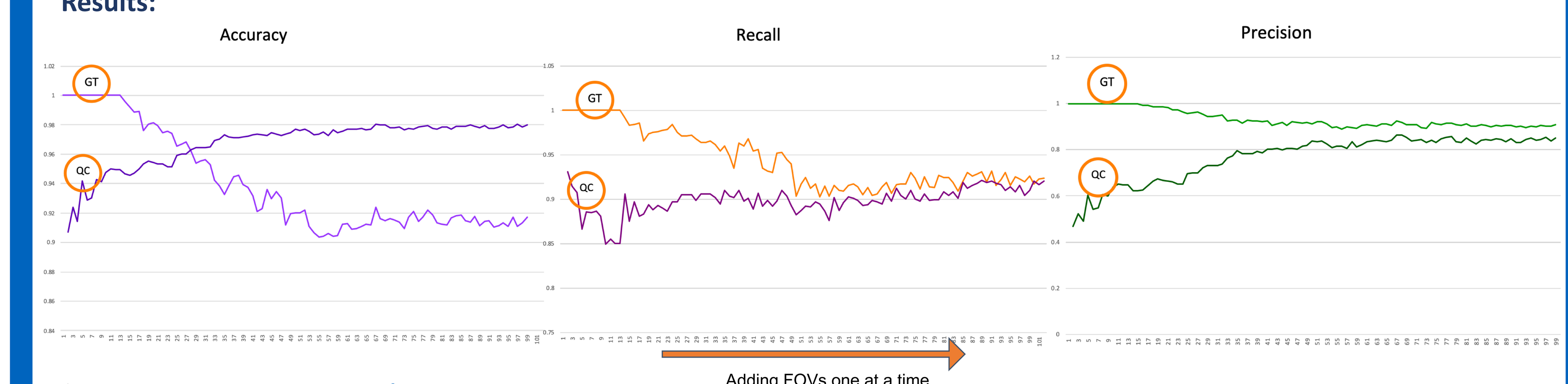
- Provides a wise ground truth collection approach, which is fast and less expensive
- Enables fine tuning of pre-trained machine/deep learning models on new datasets
- Aggregates ground truth annotated on different cohorts or projects in one place
- Saves all the ground truth ever collected, into database for future use by all the users.

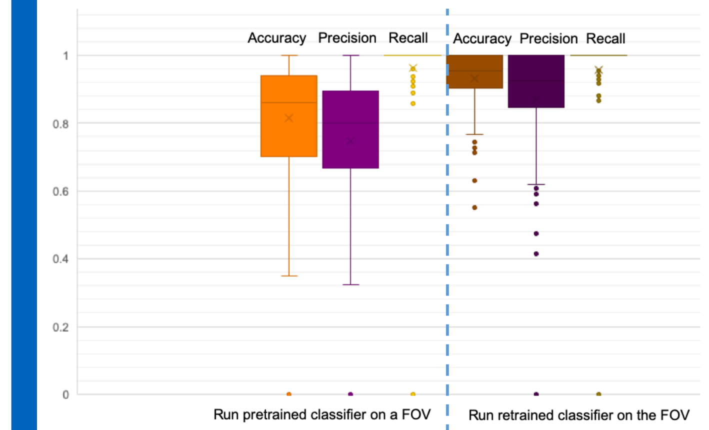
3 – Feasibility Study

Detection of macrophages in multiplex immunofluorescence (IF) stained tissue images

Data: Development dataset: 103 Field of Views (FOV) selected from WSIs (2880 labeled instances: 1492 pos., 1388 neg.)
QC dataset: 30 FOVs (16013 labeled instances: 1324 pos., 14689 neg.)
Ground Truth (GT): Instances that are labeled by the user during active learning process (14 instances from each FOV)

Results:





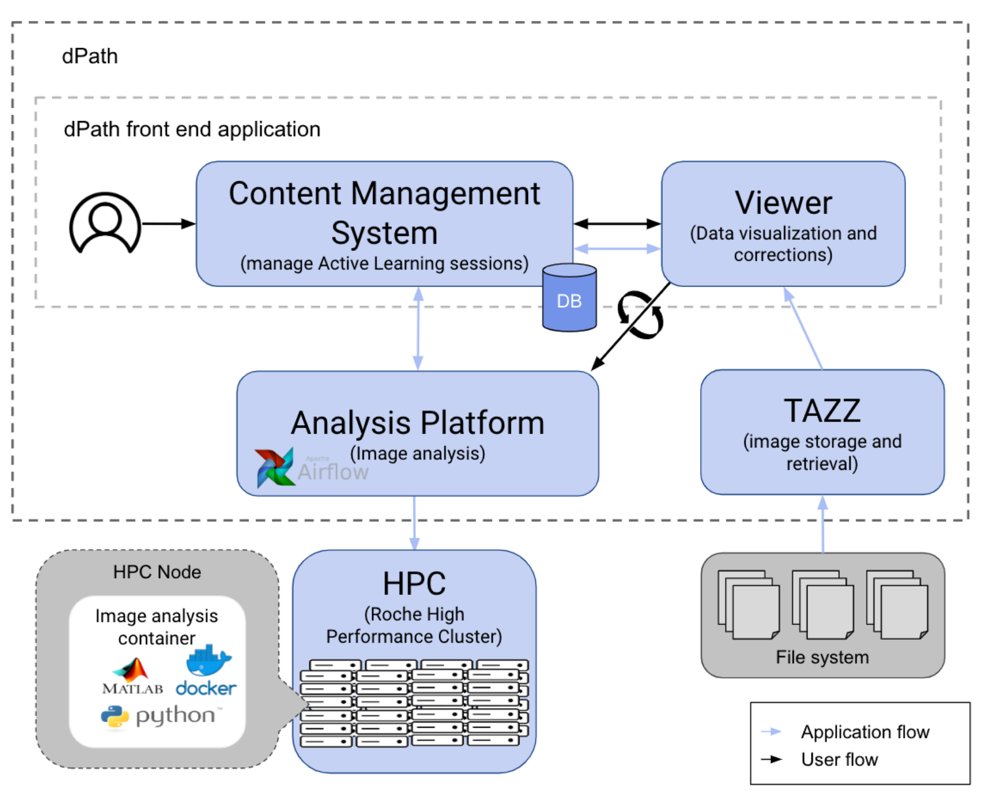
Training on	Performance on	Accuracy	Precision	Recall
All 2880 instances in development dataset	Development dataset	0.93	0.92	0.95
	QC dataset	0.98	0.88	0.90
1440 random instances from development dataset (720 positive, 720 negative)	Random sample	0.94	0.93	0.95
	Development dataset	0.92	0.91	0.94
	QC dataset	0.97	0.83	0.90
1439 instances queried by active learning (709 positive, 730 negative)	Ground truth	0.91	0.91	0.92
	Development dataset	0.93	0.92	0.96
	QC dataset	0.98	0.85	0.92

Observations:

- Enabled **wiser ground truth generating** (active learning approach vs random sample selection)
- Trained the classifier with **half the size of the labeled data** needed in random sample selection approach

4 – Designed Active Learning System

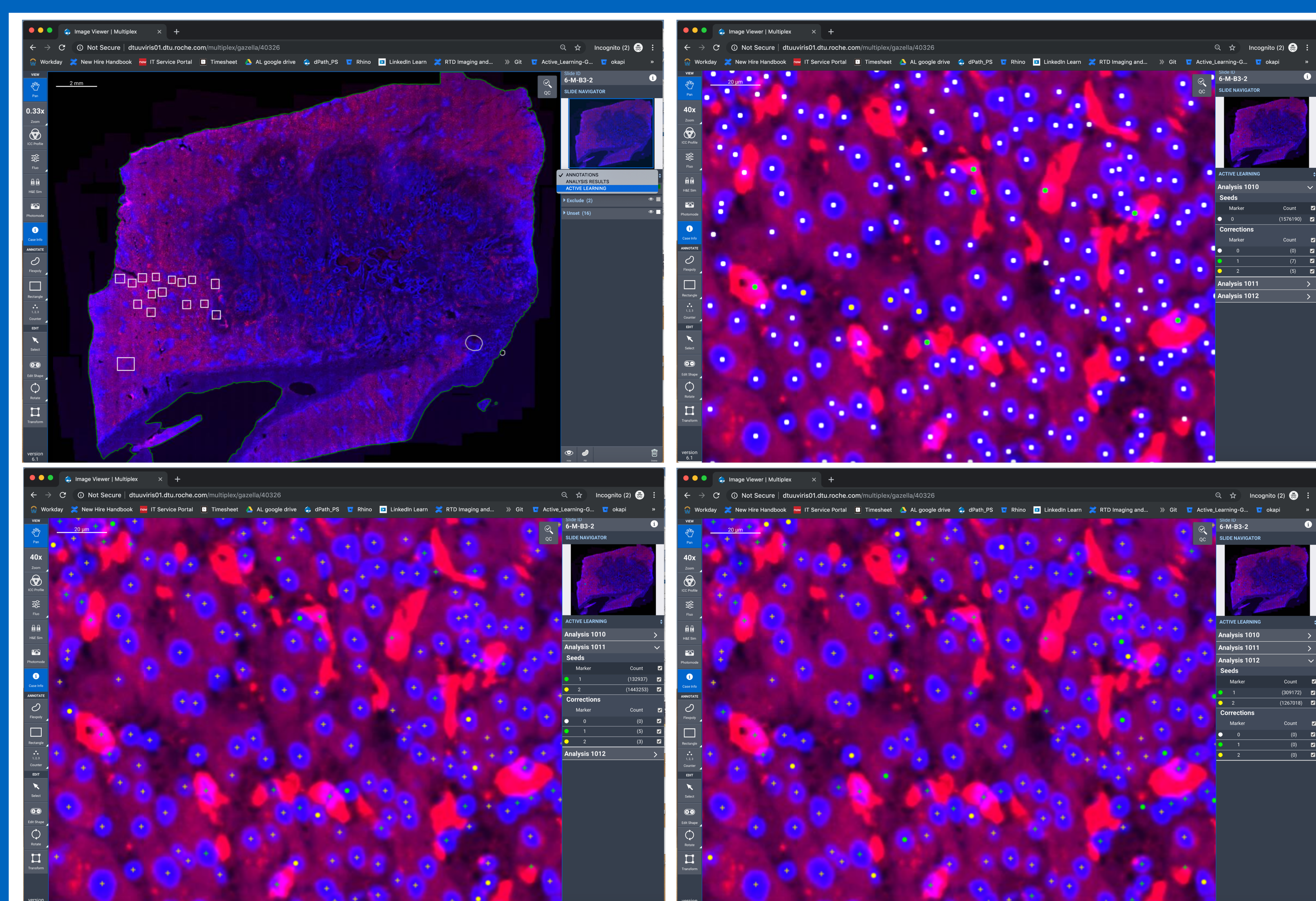
The designed system can be employed by **pathologists**, or **imaging scientists** to collect the ground truth and train classifier models in an iterative manner. Both conventional machine learning and deep learning models can be trained using this system.



Implementing the active learning algorithm requires several different components including:

- high-resolution, high-speed image storage and retrieval,
- Image and data visualization,
- interactive data correction,
- and image analysis, including model training and inferencing.

This framework is available within dPath, a Roche Tissue Diagnostics Computational Pathology Research Platform. A high performance image server handles the requests for images from the front-end application. Using Roche's High-Performance Cluster, results from training and inferencing the model are generated and stored at scale in a database. APIs enable the interaction between the machine learning engine and the database. The dPath platform integrates all of these components and allows the user to train models within a web browser.

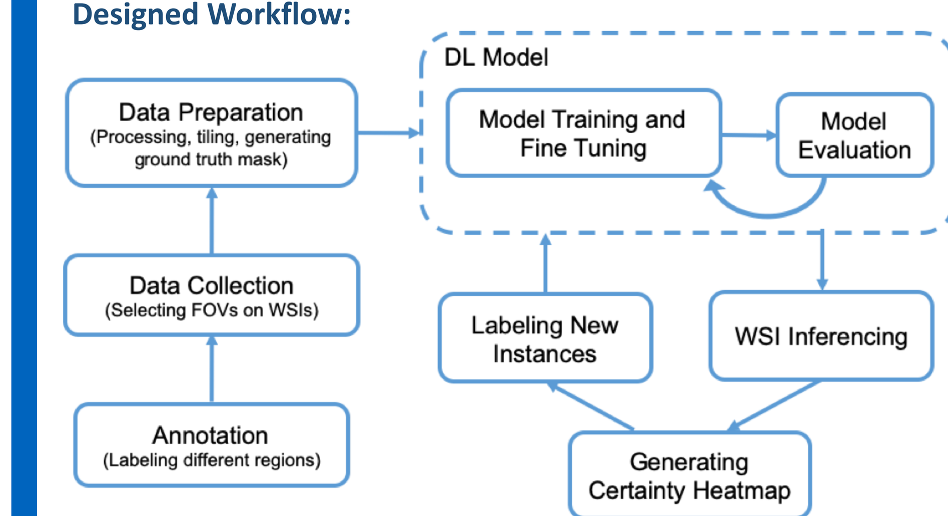


(**Top**)The user starts with labeling cells or regions of the tissue in the images. The labeled examples are used to train a model or optimize a pre-trained model. (**Bottom**) The classification results and corresponding certainty heatmap are visualized. The user then labels more training samples from the most uncertain regions and retrains the classifier. This iteration continues until the model reaches the desired performance and can be deployed (cross: high certainty, circle: low certainty instances).

5 – Active Deep Learning

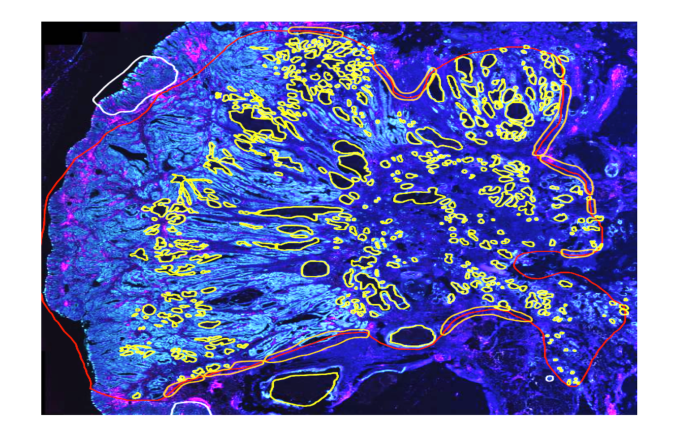
A framework for training or fine tuning Deep Learning (DL) models

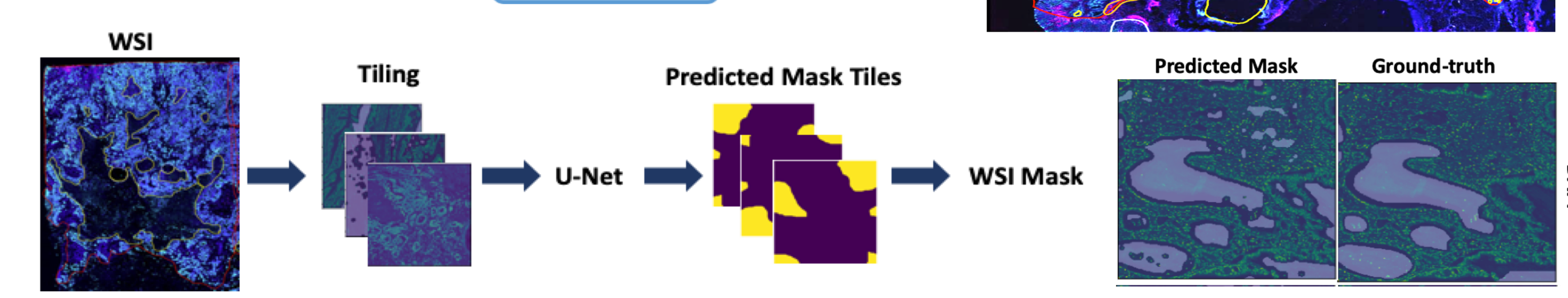
Designed Workflow:



Feasibility Study:

Automated identification of necrotic regions in multiplex IF stained tissue images using deep learning





6 – Conclusion

The digital pathology active learning system enables the end user to **create and optimize a classifier** in an **efficient** and **interactive** manner. The user can train and optimize both conventional machine learning and deep learning models using this system.