Extreme Scale, Tissue Analytics and AI

Joel Saltz MD, PhD
Chair and Distinguished Professor Department of Biomedical Informatics
Professor Department of Pathology
Cherith Endowed Chair
Stony Brook University

Workshop on Clusters, Clouds, and Data for Scientific Computing
September 2022

Application Challenges

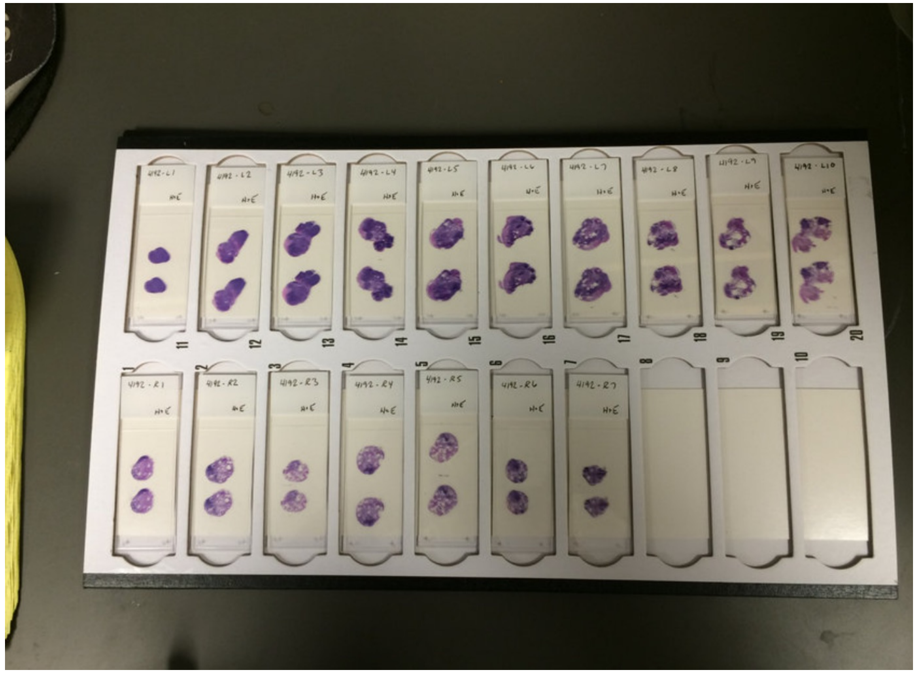
- AI Pathologist: Answer nuanced questions about patients and disease categories
- Big Picture Clinical Goal: Predict outcome and response to treatment using all available image, clinical, molecular data

Answers to Nuanced Questions

- Difficult cases require involved back and forth discussions
- Current AI methods select between a small number of alternatives
- New AI methods should be able to select between hundreds of alternatives and to provide nuanced reasoning with visual and verbal examples

Applications context – Cancer

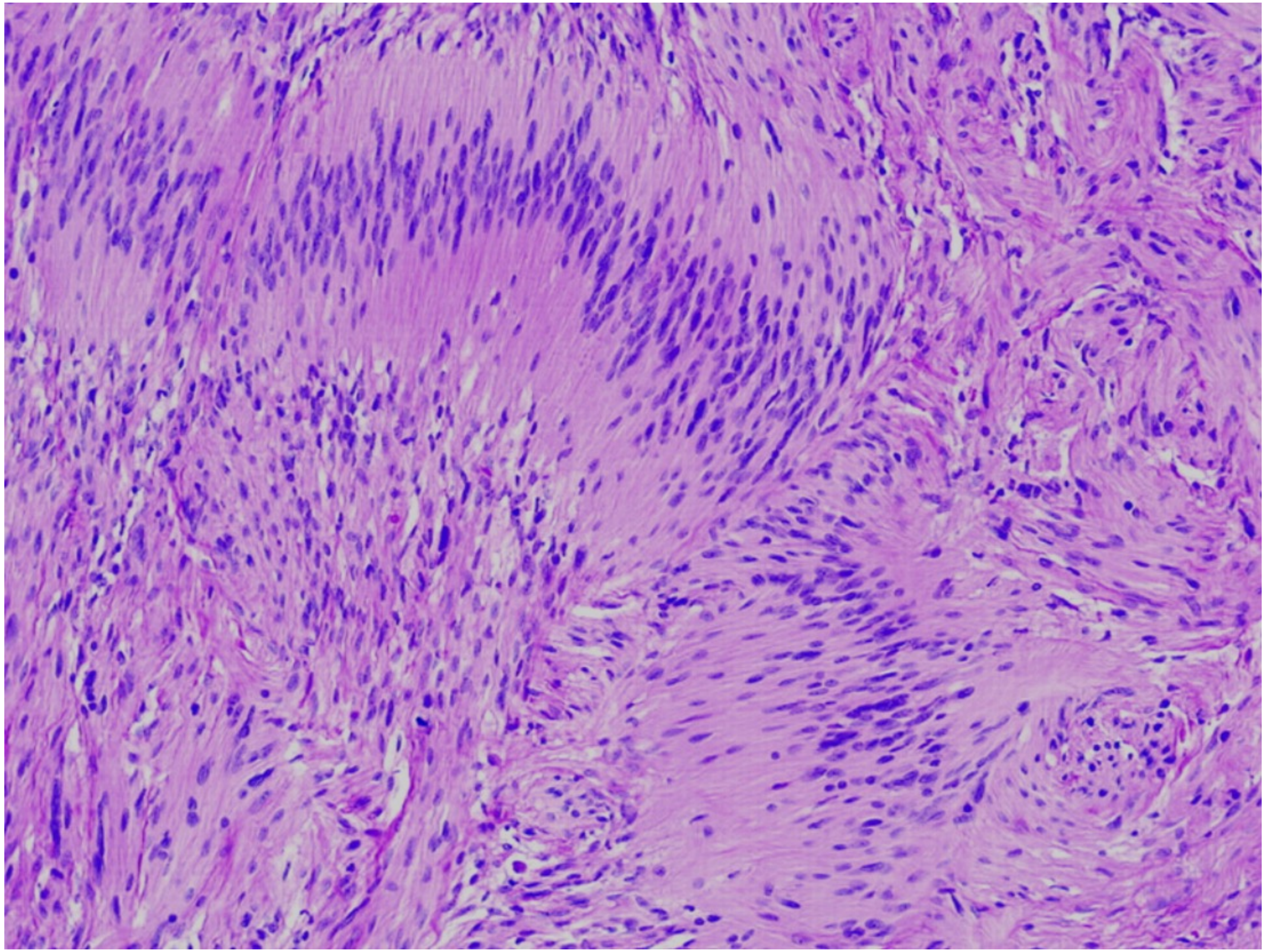
Introducing Aperio GT 450
Leica
Directly load racks from cover slipper to scanner



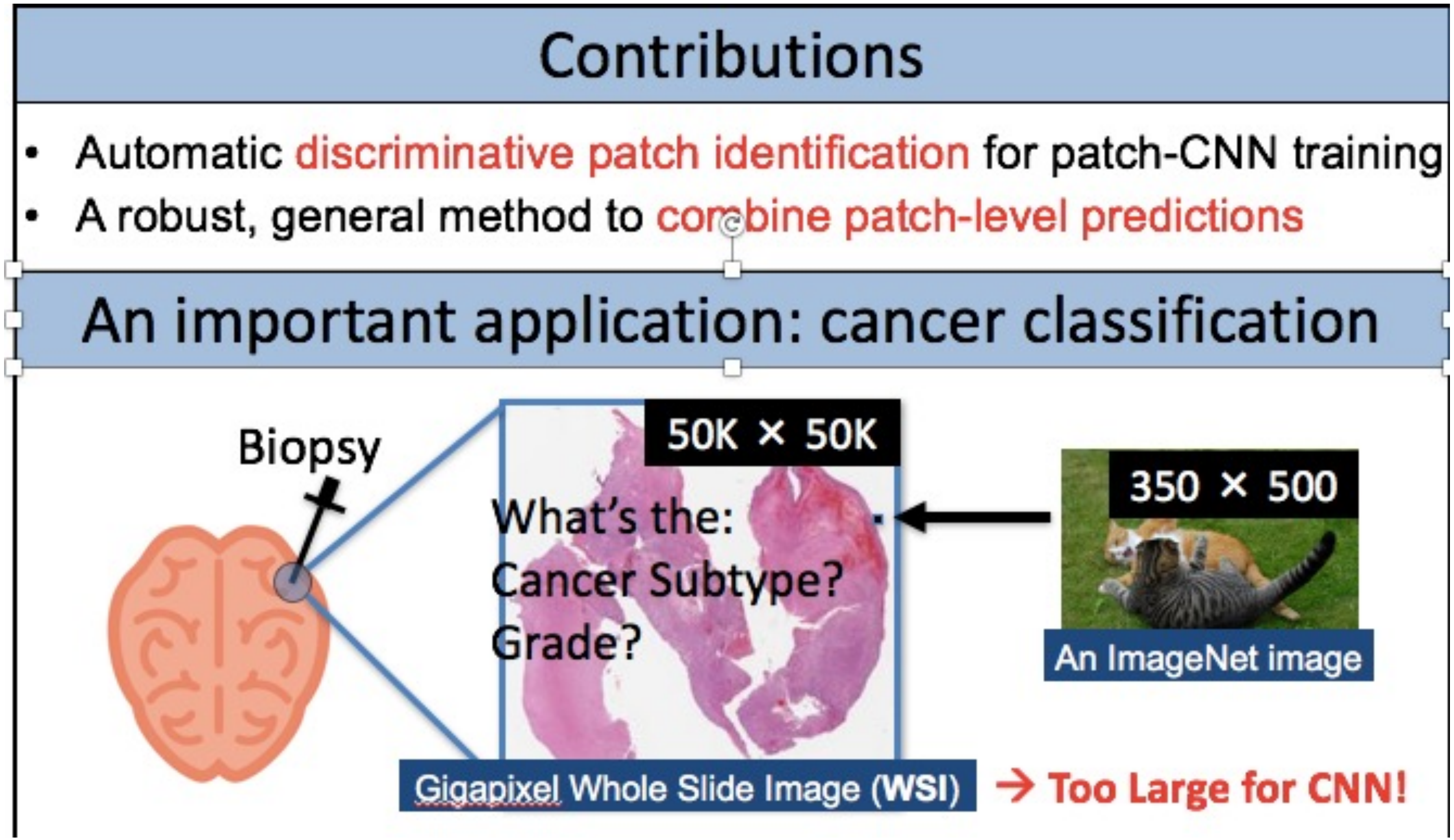
Slide Scanner – GT450 – Roughly 1 minute to acquire a (roughly) 10 Gigapixel image

Tray of glass slides – generally a Pathologist needs to examine many slides for each patient

Example GBM (Brain Tumor) Image



Brain Tumor Classification – CVPR 2016



Contributions

- Automatic discriminative patch identification for patch-CNN training
- A robust, general method to combine patch-level predictions

An important application: cancer classification

Biopsy

50K × 50K

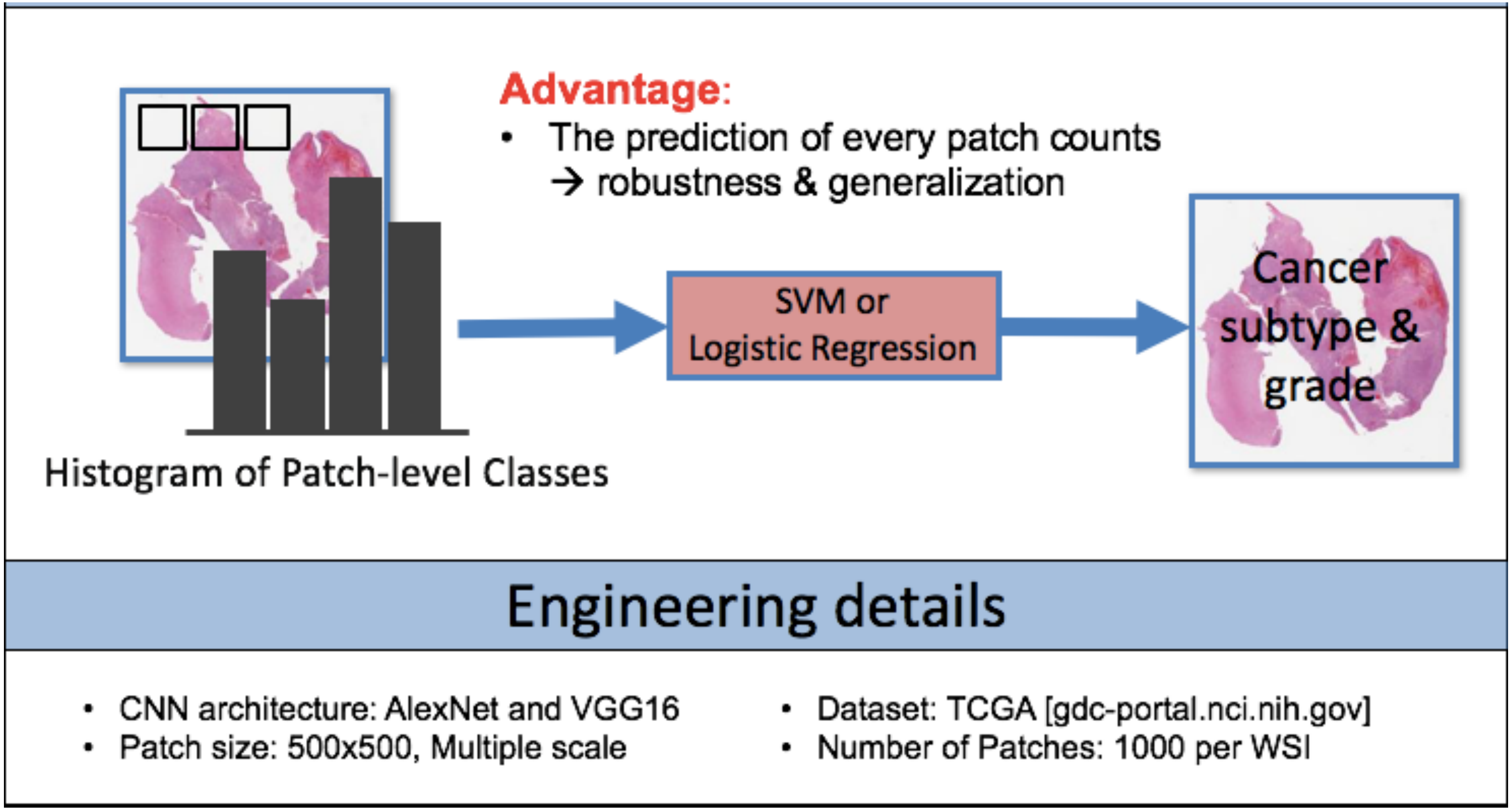
What's the:
Cancer Subtype?
Grade?

350 × 500

An ImageNet image

Gigapixel Whole Slide Image (**WSI**) → **Too Large for CNN!**

Combining Information from Patches

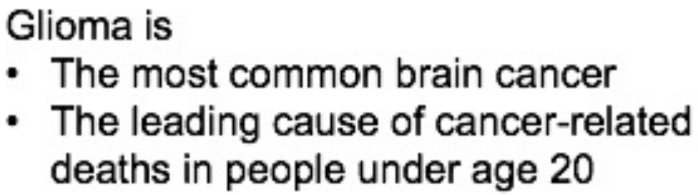


Advantage:
- The prediction of every patch counts
  → robustness & generalization

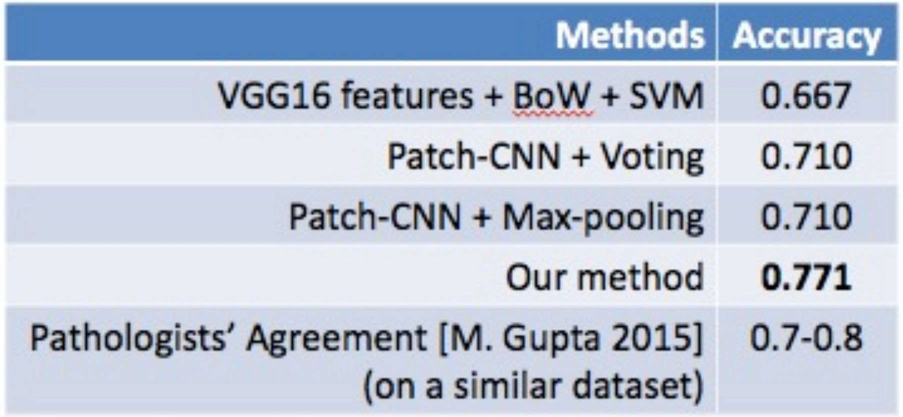
Engineering details

- CNN architecture: AlexNet and VGG16
- Patch size: 500x500, Multiple scale
- Dataset: TCGA [gdc-portal.nci.nih.gov]
- Number of Patches: 1000 per WSI

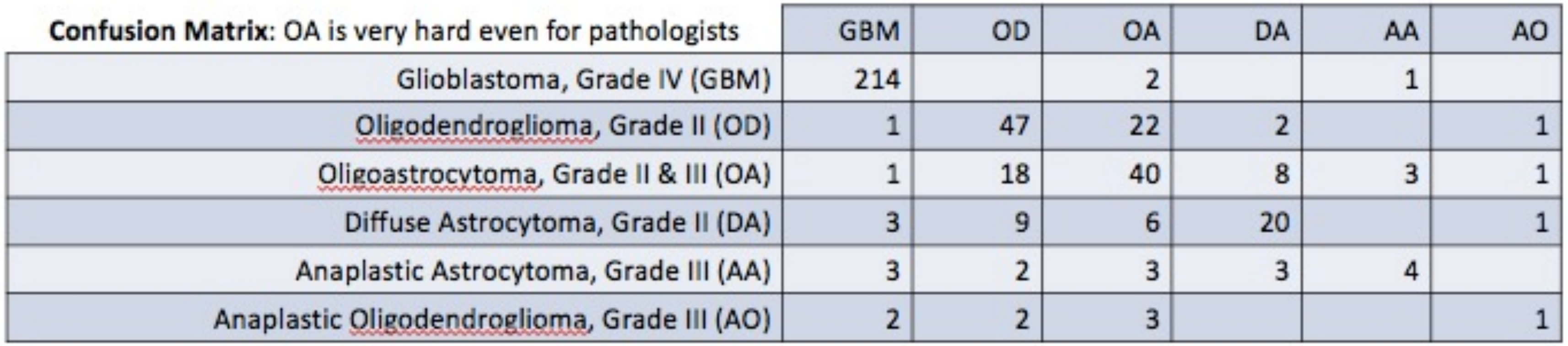
Brain Tumor Classification Results



Glioma is
- The most common brain cancer
- The leading cause of cancer-related deaths in people under age 20



Methods	Accuracy
VGG16 features + BoW + SVM	0.667
Patch-CNN + Voting	0.710
Patch-CNN + Max-pooling	0.710
Our method	**0.771**
Pathologists' Agreement [M. Gupta 2015] (on a similar dataset)	0.7-0.8



Confusion Matrix: OA is very hard even for pathologists	GBM	OD	OA	DA	AA	AO
Glioblastoma, Grade IV (GBM)	214		2		1	
Oligodendroglioma, Grade II (OD)	1	47	22	2		1
Oligoastrocytoma, Grade II & III (OA)	1	18	40	8	3	1
Diffuse Astrocytoma, Grade II (DA)	3	9	6	20		1
Anaplastic Astrocytoma, Grade III (AA)	3	2	3	3	4	
Anaplastic Oligodendroglioma, Grade III (AO)	2	2	3			1

Le Hou, Dimitris Samaras, Tahsin Kurc, Yi Gao, Liz Vanner, James Davis, Joel Saltz

691 Google Scholar citations; PaigeAI FDA cleared clinical application using closely related method

Many Categories of Brain Tumor

Image Based AI Algorithms Currently Only Consider Most Common Options

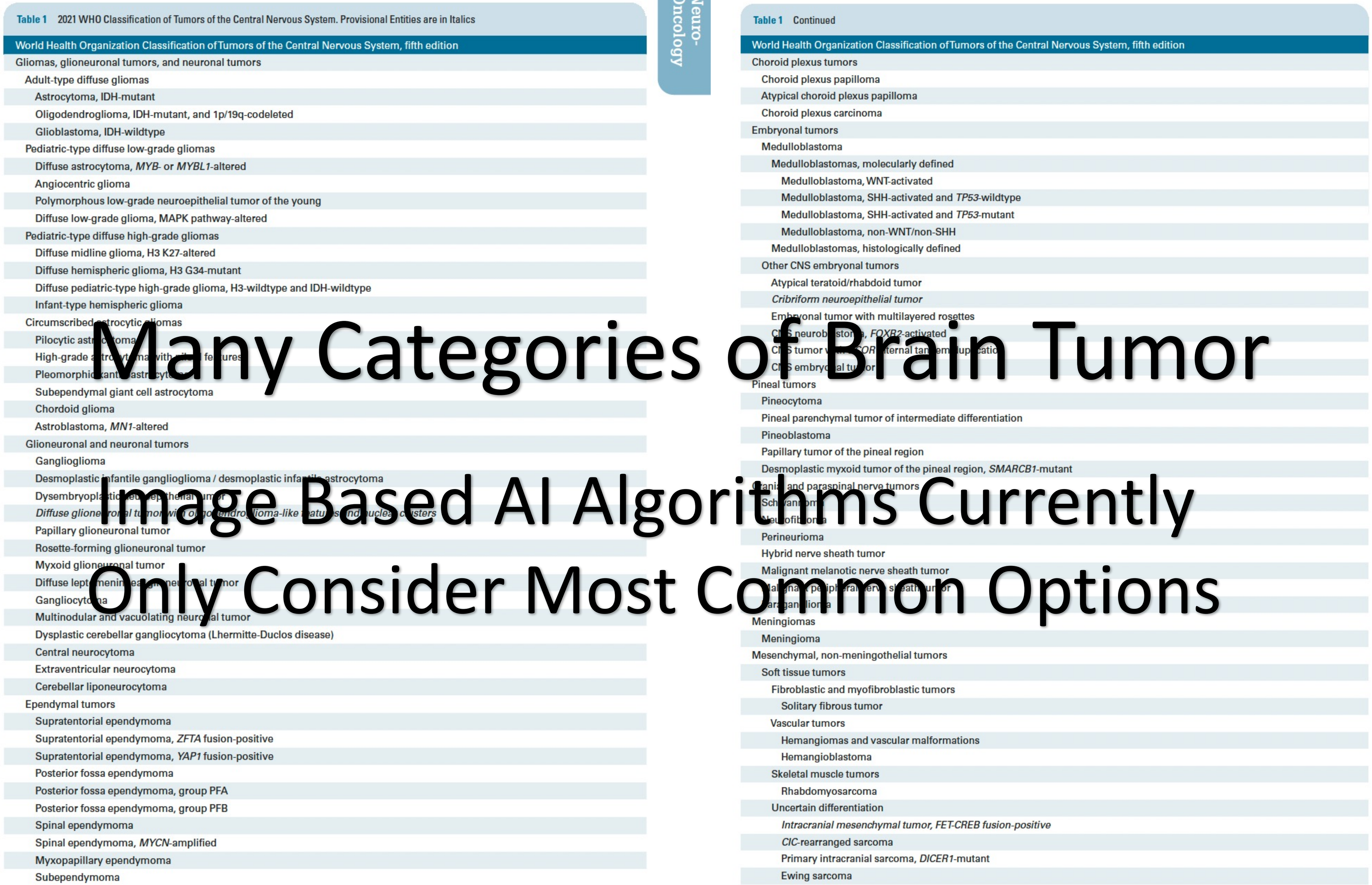

Neuro-Oncology

Table 1 2021 WHO Classification of Tumors of the Central Nervous System. Provisional Entities are in Italics

World Health Organization Classification of Tumors of the Central Nervous System, fifth edition
Gliomas, glioneuronal tumors, and neuronal tumors
Adult-type diffuse gliomas
Astrocytoma, IDH-mutant
Oligodendroglioma, IDH-mutant, and 1p/19q-codeleted
Glioblastoma, IDH-wildtype
Pediatric-type diffuse low-grade gliomas
Diffuse astrocytoma, *MYB*- or *MYBL1*-altered
Angiocentric glioma
Polymorphous low-grade neuroepithelial tumor of the young
Diffuse low-grade glioma, MAPK pathway-altered
Pediatric-type diffuse high-grade gliomas
Diffuse midline glioma, H3 K27-altered
Diffuse hemispheric glioma, H3 G34-mutant
Diffuse pediatric-type high-grade glioma, H3-wildtype and IDH-wildtype
Infant-type hemispheric glioma
Circumscribed astrocytic gliomas
Pilocytic astrocytoma
High-grade astrocytoma with piloid features
Pleomorphic xanthoastrocytoma
Subependymal giant cell astrocytoma
Chordoid glioma
Astroblastoma, *MN1*-altered
Glioneuronal and neuronal tumors
Ganglioglioma
Desmoplastic infantile ganglioglioma / desmoplastic infantile astrocytoma
Dysembryoplastic neuroepithelial tumor
Diffuse glioneuronal tumor with oligodendroglioma-like features and nuclear clusters
Papillary glioneuronal tumor
Rosette-forming glioneuronal tumor
Myxoid glioneuronal tumor
Diffuse leptomeningeal glioneuronal tumor
Gangliocytoma
Multinodular and vacuolating neuronal tumor
Dysplastic cerebellar gangliocytoma (Lhermitte-Duclos disease)
Central neurocytoma
Extraventricular neurocytoma
Cerebellar liponeurocytoma
Ependymal tumors
Supratentorial ependymoma
Supratentorial ependymoma, *ZFTA* fusion-positive
Supratentorial ependymoma, *YAP1* fusion-positive
Posterior fossa ependymoma
Posterior fossa ependymoma, group PFA
Posterior fossa ependymoma, group PFB
Spinal ependymoma
Spinal ependymoma, *MYCN*-amplified
Myxopapillary ependymoma
Subependymoma

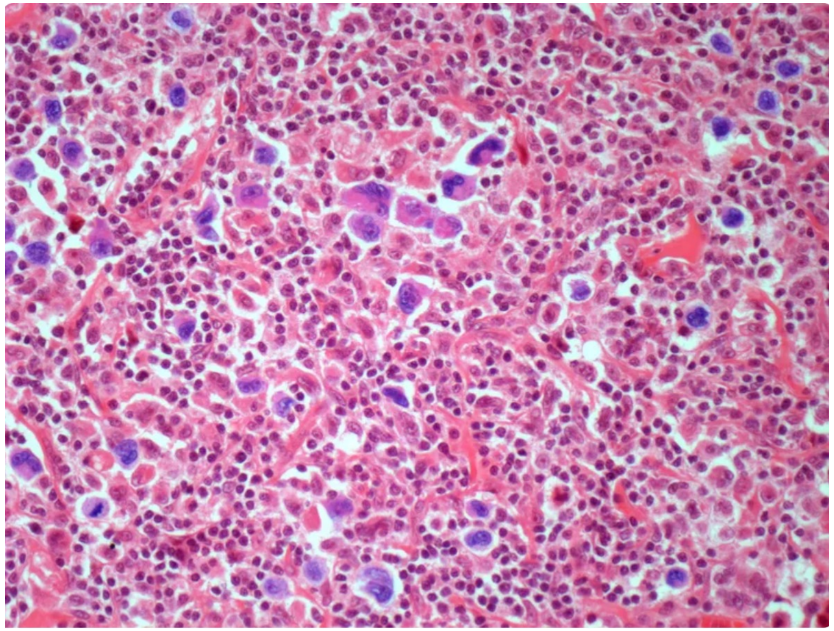
Table 1 Continued

World Health Organization Classification of Tumors of the Central Nervous System, fifth edition
Choroid plexus tumors
Choroid plexus papilloma
Atypical choroid plexus papilloma
Choroid plexus carcinoma
Embryonal tumors
Medulloblastoma
Medulloblastomas, molecularly defined
Medulloblastoma, WNT-activated
Medulloblastoma, SHH-activated and *TP53*-wildtype
Medulloblastoma, SHH-activated and *TP53*-mutant
Medulloblastoma, non-WNT/non-SHH
Medulloblastomas, histologically defined
Other CNS embryonal tumors
Atypical teratoid/rhabdoid tumor
Cribriform neuroepithelial tumor
Embryonal tumor with multilayered rosettes
CNS neuroblastoma, *FOXR2*-activated
CNS tumor with *BCOR* internal tandem duplication
CNS embryonal tumor
Pineal tumors
Pineocytoma
Pineal parenchymal tumor of intermediate differentiation
Pineoblastoma
Papillary tumor of the pineal region
Desmoplastic myxoid tumor of the pineal region, *SMARCB1*-mutant
Cranial and paraspinal nerve tumors
Schwannoma
Neurofibroma
Perineurioma
Hybrid nerve sheath tumor
Malignant melanotic nerve sheath tumor
Malignant peripheral nerve sheath tumor
Paraganglioma
Meningiomas
Meningioma
Mesenchymal, non-meningothelial tumors
Soft tissue tumors
Fibroblastic and myofibroblastic tumors
Solitary fibrous tumor
Vascular tumors
Hemangiomas and vascular malformations
Hemangioblastoma
Skeletal muscle tumors
Rhabdomyosarcoma
Uncertain differentiation
Intracranial mesenchymal tumor, FET-CREB fusion-positive
CIC-rearranged sarcoma
Primary intracranial sarcoma, *DICER1*-mutant
Ewing sarcoma

The Promise of Self Supervised Learning

- Large language models are trained on huge quantities of text using methods that include filling in masked words, predicting next sentence
- These models are now able to generate sophisticated responses to questions
- Many open issues but see next example of nuanced question involving lymphomas

what lymphomas resemble hodgkins disease but do not have reed sternberg cells



- There are several types of lymphomas that can resemble Hodgkin's disease but do not have Reed-Sternberg cells. Reed-Sternberg cells are large, abnormal lymphocytes that are the hallmark of classical Hodgkin's lymphoma.
- Some of the lymphomas that can mimic Hodgkin's disease but lack Reed-Sternberg cells include:
- Nodular lymphocyte-predominant Hodgkin's lymphoma: This is a rare type of Hodgkin's lymphoma that accounts for about 5% of all cases. It is characterized by the presence of large, popcorn-shaped lymphocytes called "L&H cells" that resemble normal B cells.
- T-cell/histiocyte-rich large B-cell lymphoma: This is a subtype of diffuse large B-cell lymphoma that can mimic classical Hodgkin's lymphoma. It is characterized by the presence of scattered large B cells, surrounded by a background of small T cells and histiocytes.
- Anaplastic large cell lymphoma: This is a type of non-Hodgkin's lymphoma that can mimic Hodgkin's disease. It is characterized by the presence of large, anaplastic lymphocytes that express CD30, a marker also seen in Hodgkin's disease.
- Mediastinal large B-cell lymphoma: This is a subtype of diffuse large B-cell lymphoma that often presents as a large mass in the mediastinum (the area between the lungs). It can mimic Hodgkin's disease due to the presence of fibrosis and the presence of some large, atypical lymphoid cells.

ChatGPT

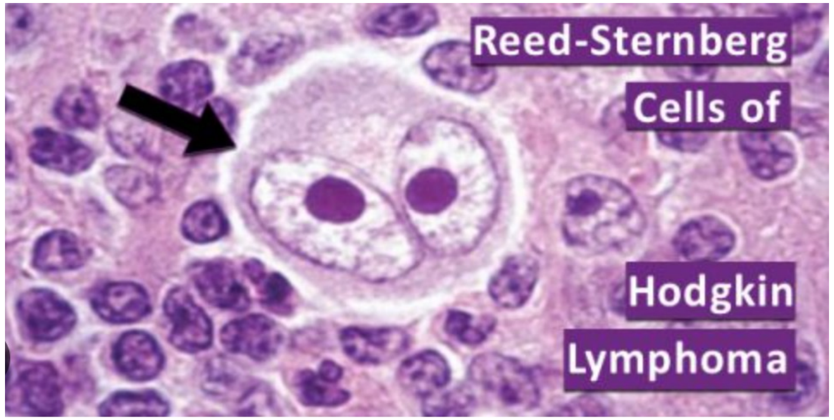
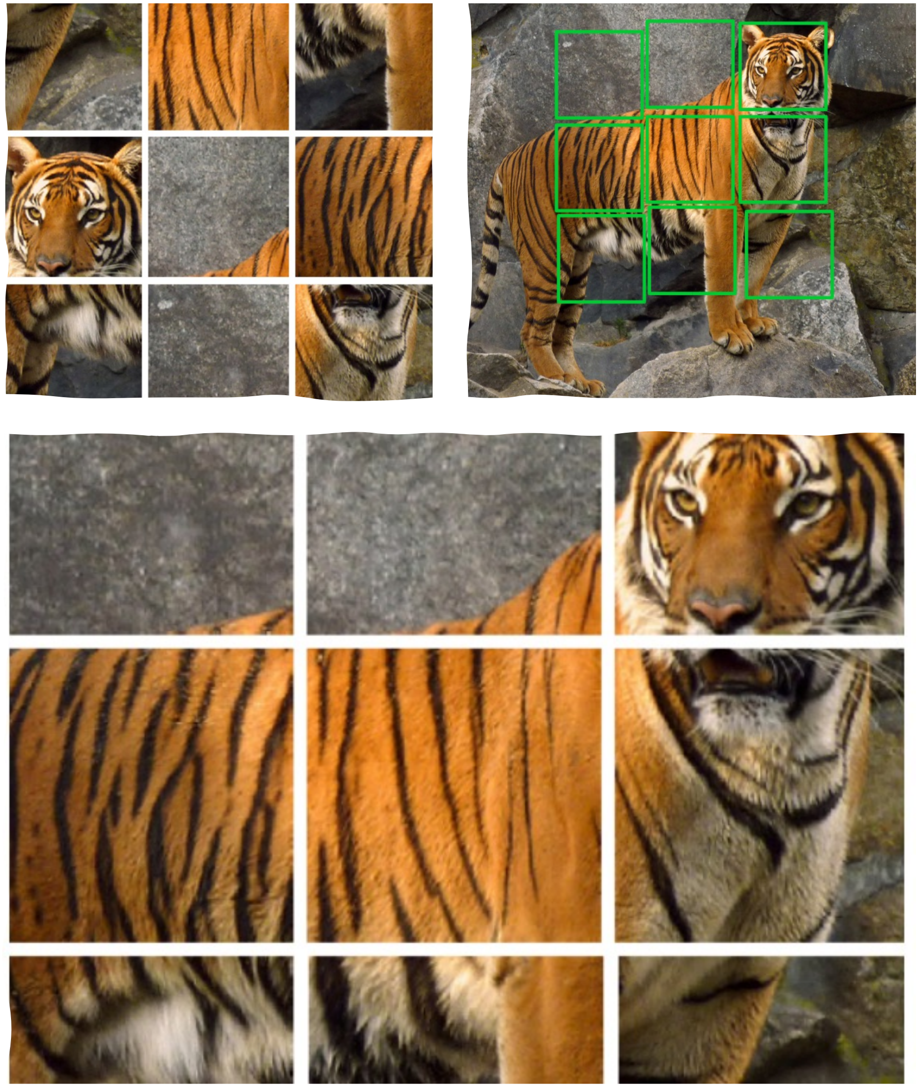
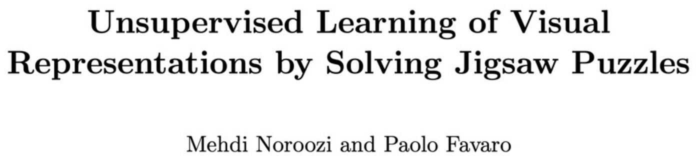


Image based self supervised learning

- Generate and solve various visual puzzles
- Reassemble jigsaw puzzles
- Fill in masked regions
- Recognize warped, distorted, noisy images





Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles

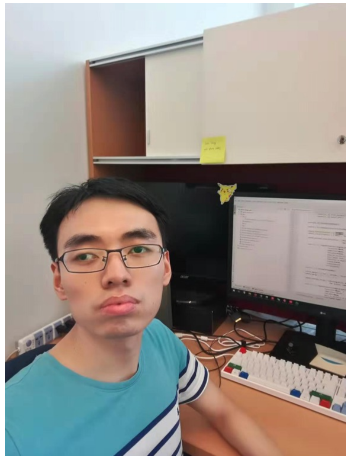
Mehdi Noroozi and Paolo Favaro

https://arxiv.org/pdf/1603.09246.pdf

Self Supervision in Pathology

- Gigapixel Images
- Images can have millions of cells, many tens of thousands + of glands, crypts ducts
- Models need to learn syntax and semantics of tissue
- We are doing this, starting out with generating a model using self supervised training using whole slide images and cell density
- Fine tuned codes have above SOTA performance on classification and segmentation tasks

Saarthak Kapse
Stony Brook

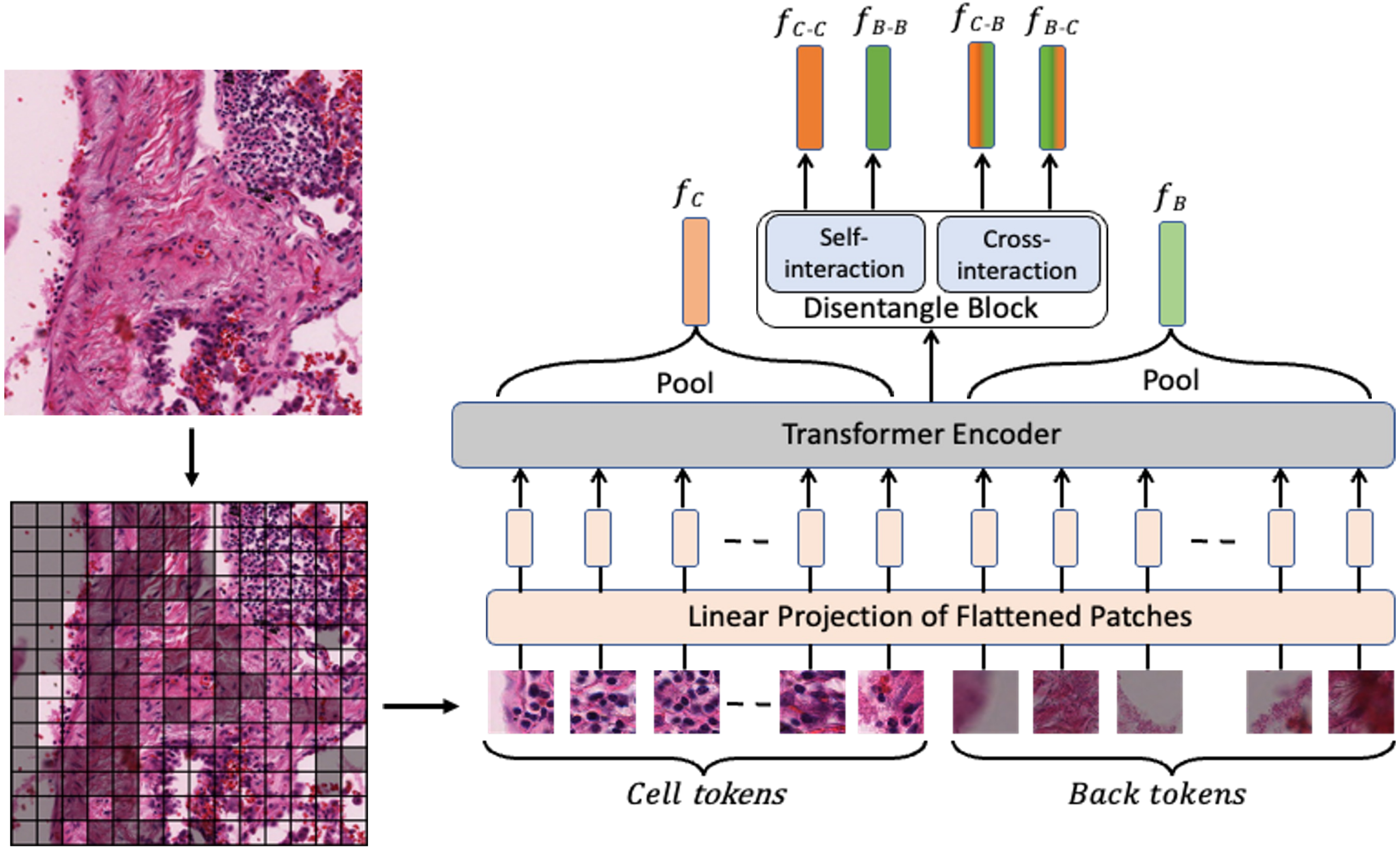


Jingwei Zhang
Stony Brook

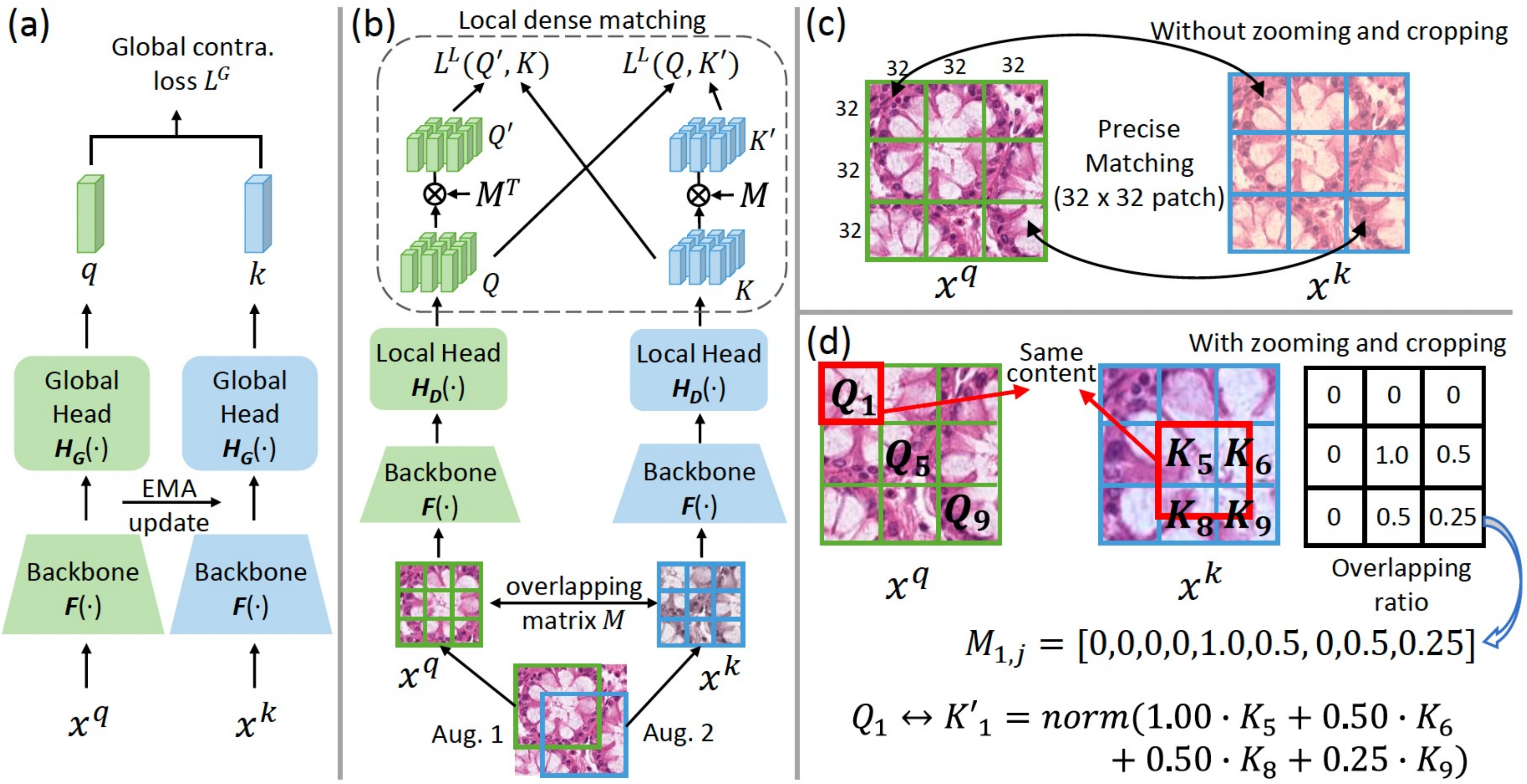
IPMI 2023
2023 International Conference on
Image Processing and Machine Intelligence
February 25-27, 2023 | Singapore

Precise Location Matching Improves Dense Contrastive Learning in Digital Pathology

Jingwei Zhang[1*], Saarthak Kapse[1*], Ke Ma[2], Prateek Prasanna[1],
Maria Vakalopoulou[3], Joel Saltz[1], and Dimitris Samaras[1]



Self Supervised Learning in Pathology



- Deep learning based pipelines adapted to carry out scientific tasks
- Characterize detailed composition and structure of tissue – Radiology, Pathology and molecular composition
  - Predict outcome, treatment response, steer treatment
  - ”Real World” large population research studies
  - Scientific studies involving disease mechanism

Applications of AI based Tumor Infiltrating Lymphocyte Analysis Methods (TILS)

Lymphocytes are immune cells

Immune therapy has become ubiquitous

Many clinical studies involving many types of cancer

Spatial patterns of distribution of TILs maps to ascertain the functional immune status of the tumor microenvironment

Combine diagnostic criteria and TILs to stratify patients, guide clinical management, and select therapy (e.g. immunotherapy)

TILS | BREAST CANCER

INTERNATIONAL IMMUNO-ONCOLOGY WORKING GROUP

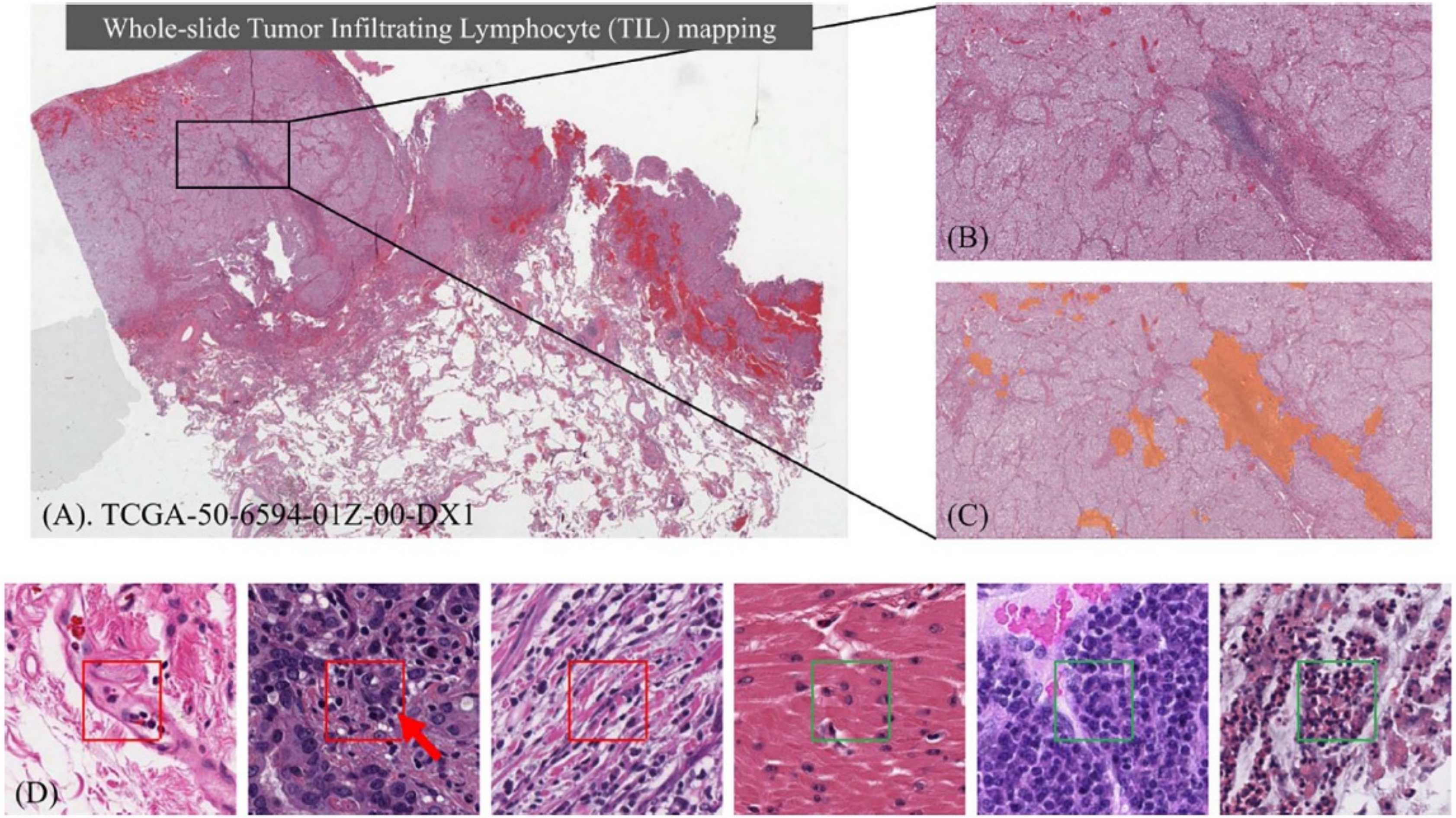
TiGER

Tumor **I**nfiltratin**G** lymphocytes in breast canc**ER**

Teaching Algorithms to Recognize Immune Cells in Confusing Contexts

Deep Learning-Based Mapping of Tumor Infiltrating Lymphocytes in Whole Slide Images of 23 Types of Cancer

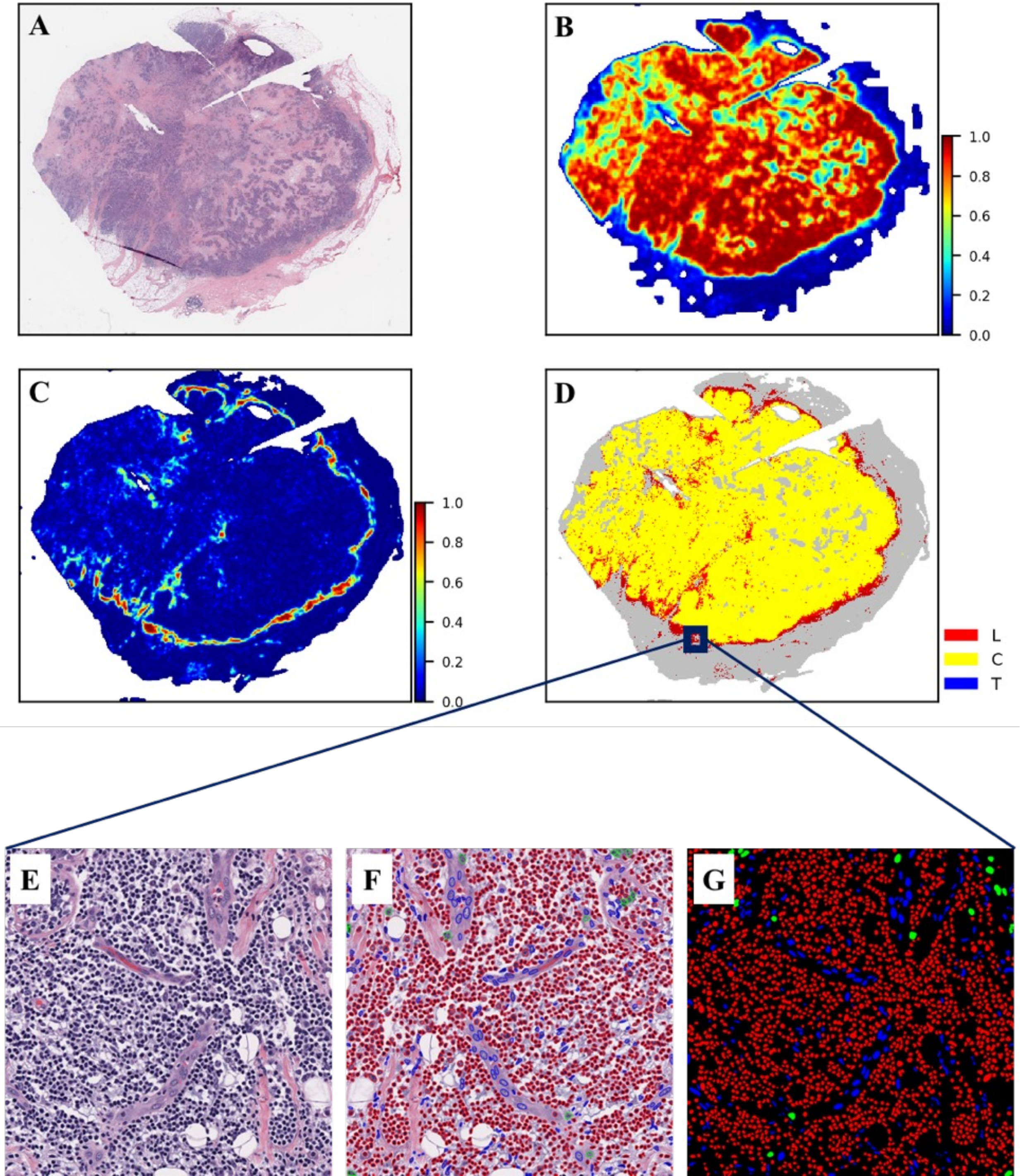
Shahira Abousamra[1], Rajarsi Gupta[2], Le Hou[1], Rebecca Batiste[3], Tianhao Zhao[3], Anand Shankar[4], Arvind Rao[4], Chao Chen[2], Dimitris Samaras[1], Tahsin Kurc[2] and Joel Saltz[2]*


Whole-slide Tumor Infiltrating Lymphocyte (TIL) mapping
(A). TCGA-50-6594-01Z-00-DX1
(B)
(C)
(D)

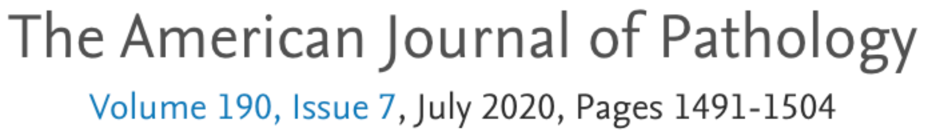
TIL positive – Red
TIL confounders - Green

Tumor TIL Analyses

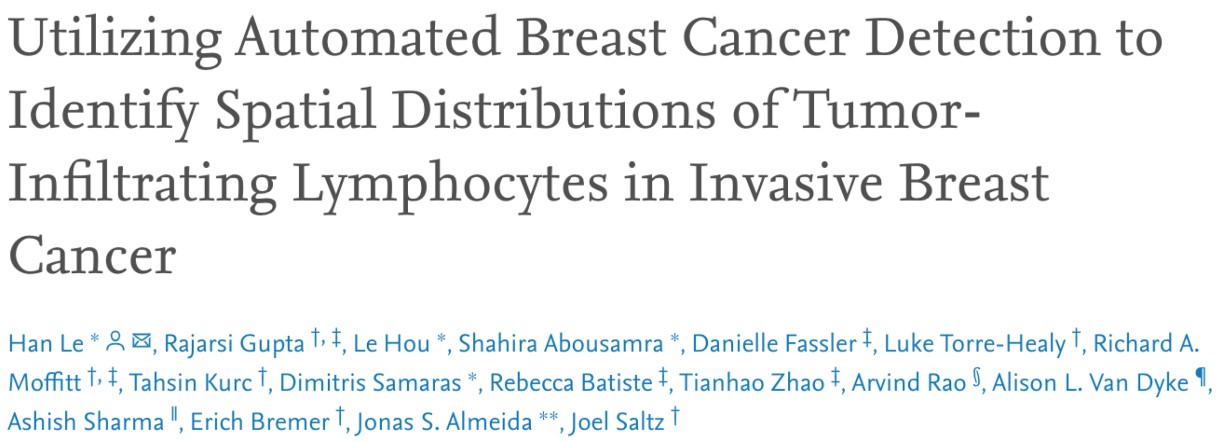
High-resolution detection and classification of tumor cells, lymphocytes, and stromal cells in the entirety of whole slide images



<u>Legend</u>
red = lymphocytes
blue = stromal cells
green = tumor cells

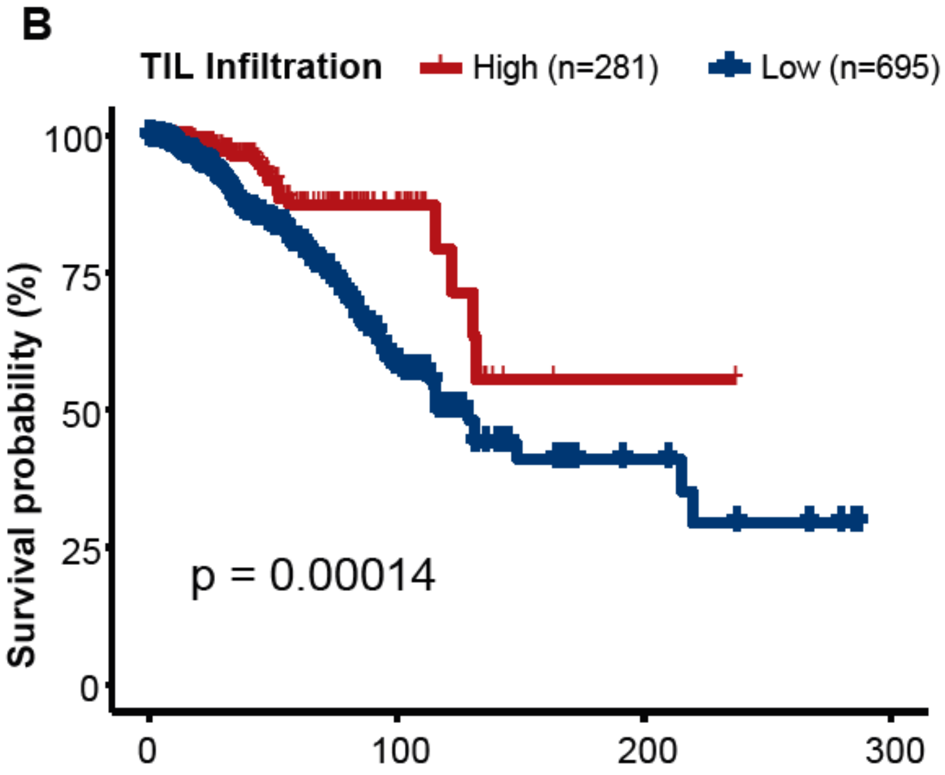


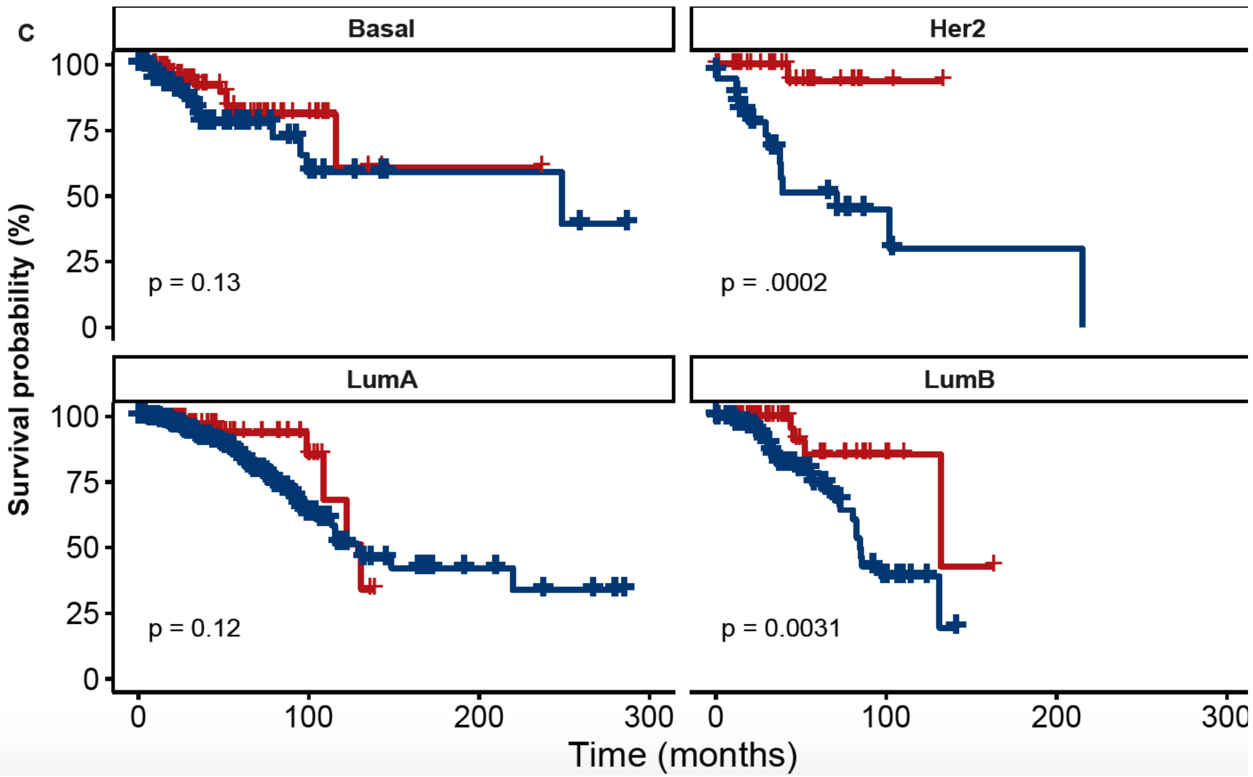
The American Journal of Pathology
Volume 190, Issue 7, July 2020, Pages 1491-1504



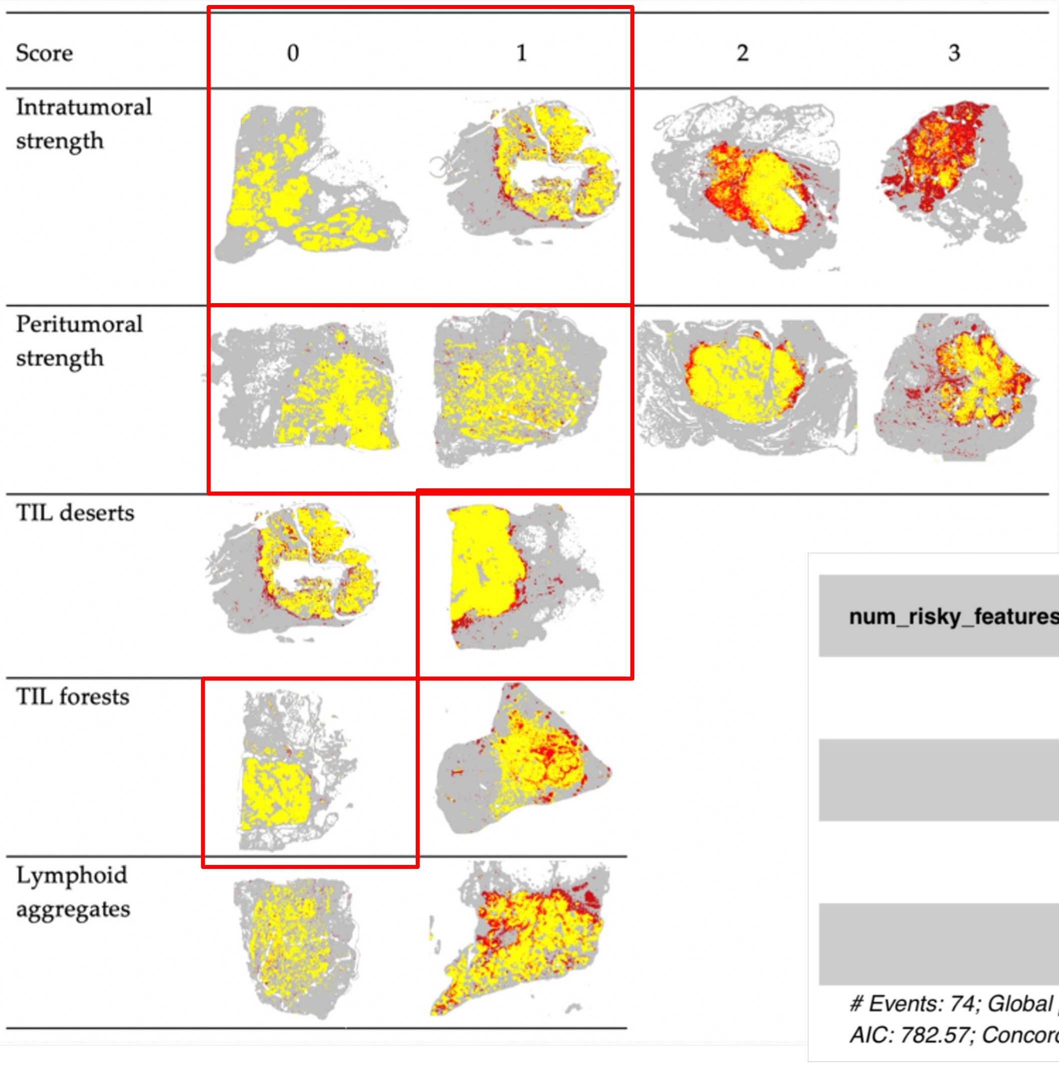
Utilizing Automated Breast Cancer Detection to Identify Spatial Distributions of Tumor-Infiltrating Lymphocytes in Invasive Breast Cancer

Han Le [*], Rajarsi Gupta [†, ‡], Le Hou [*], Shahira Abousamra [*], Danielle Fassler [‡], Luke Torre-Healy [†], Richard A. Moffitt [†, ‡], Tahsin Kurc [†], Dimitris Samaras [*], Rebecca Batiste [‡], Tianhao Zhao [‡], Arvind Rao [§], Alison L. Van Dyke [¶], Ashish Sharma [‖], Erich Bremer [†], Jonas S. Almeida [**], Joel Saltz [†]

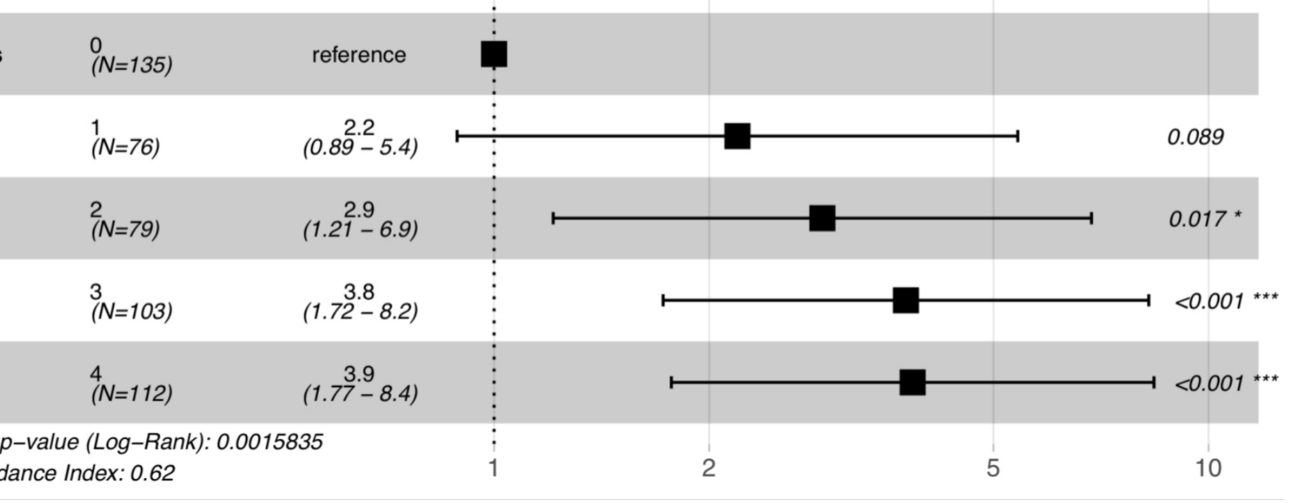




Qualitative evaluation of lymphocyte distribution enables interpretable risk identification



Metric	Level	TCGA BRCA Hazard Ratio
Intratumoral strength	1	0.62
	2	0.27
	3	0.19
TIL deserts	Present	2.0 **
TIL forests	Present	0.56 *
Peritumoral strength	1	2.07
	2	1.14
	3	0.61

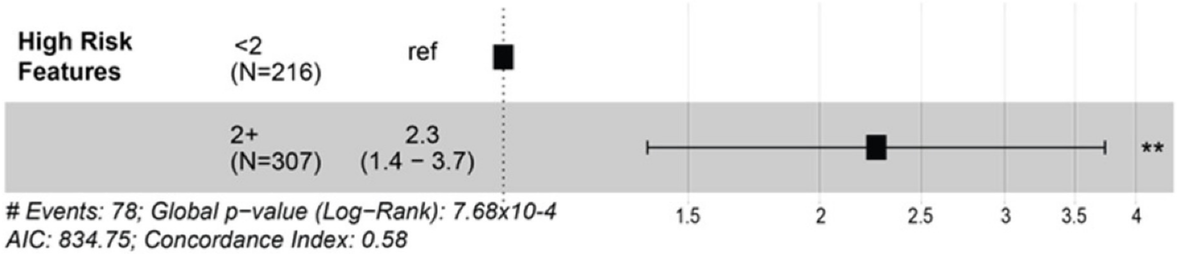


num_risky_features			
	0 (N=135)	reference	
	1 (N=76)	2.2 (0.89 – 5.4)	0.089
	2 (N=79)	2.9 (1.21 – 6.9)	0.017 *
	3 (N=103)	3.8 (1.72 – 8.2)	<0.001 ***
	4 (N=112)	3.9 (1.77 – 8.4)	<0.001 ***

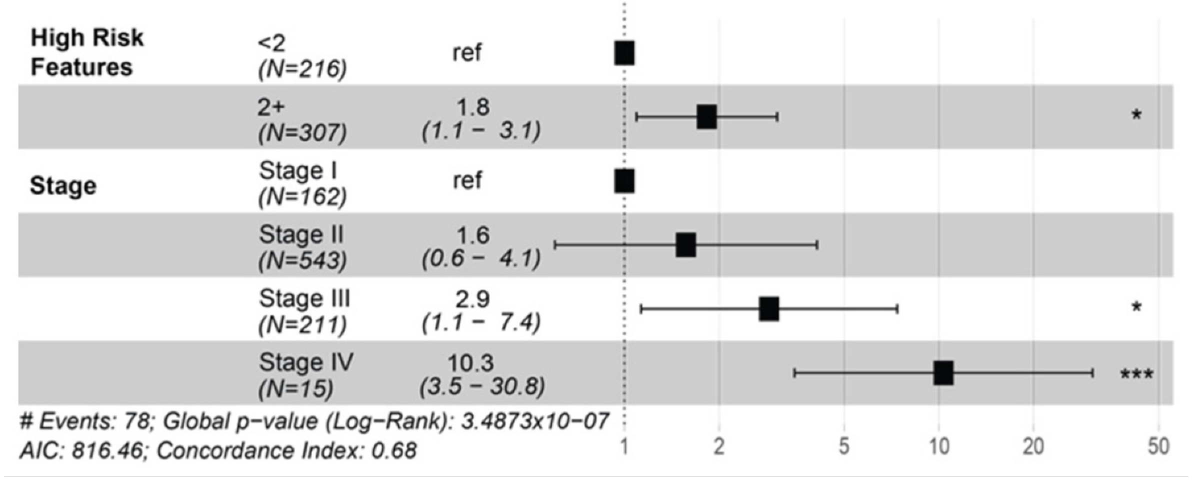
# Events: 74; Global p–value (Log–Rank): 0.0015835
AIC: 782.57; Concordance Index: 0.62

Presence of multiple risky features stably correlates with risk in new data set

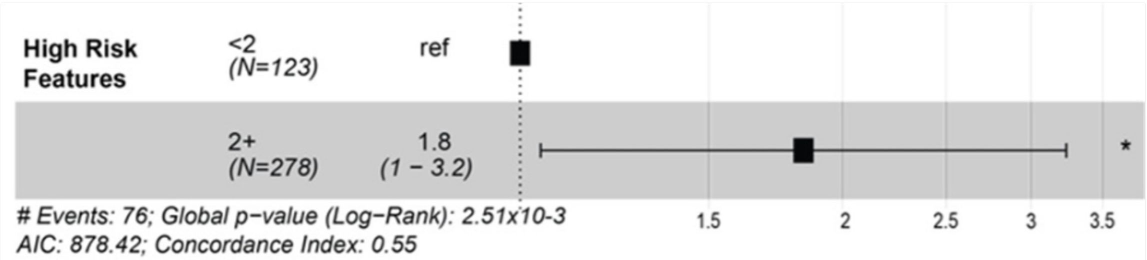
TCGA-BRCA - Original dataset



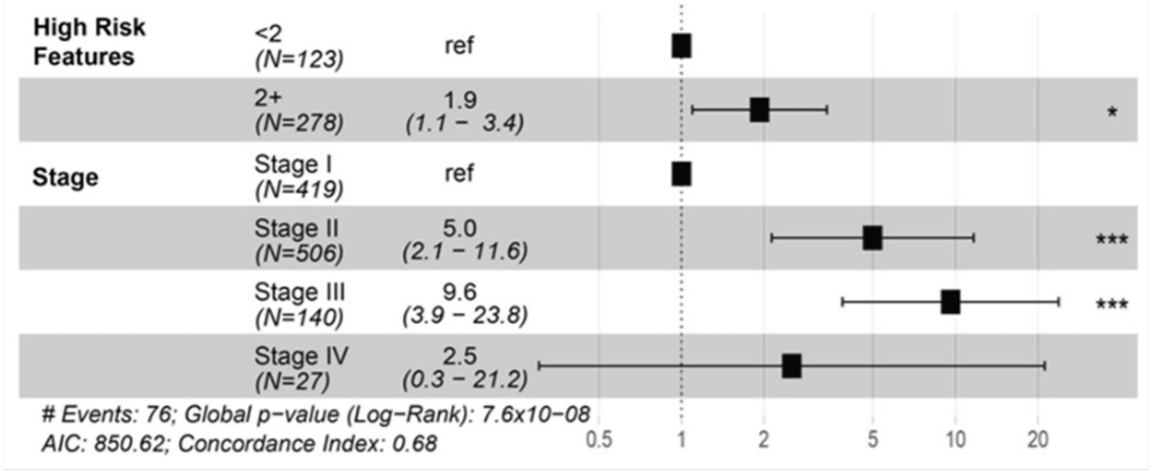
(a)



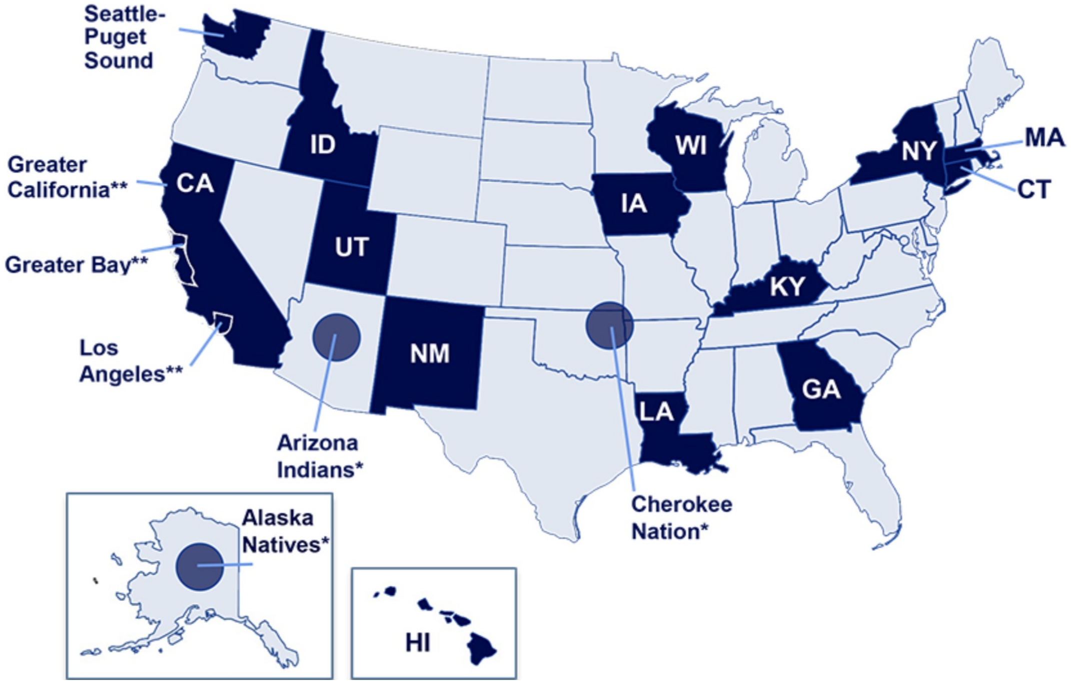
UNC-CBCS - Validation dataset



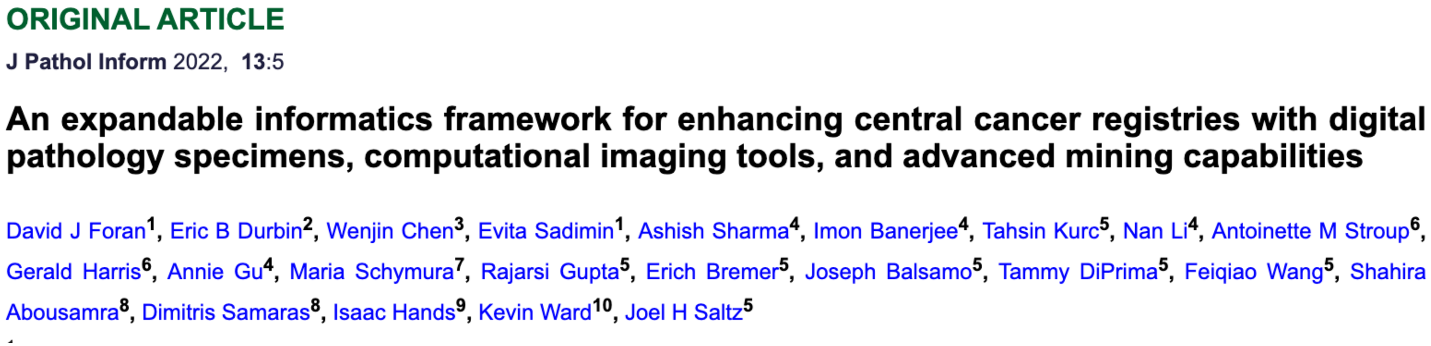
(b)



Collaboration with SEER Registries to Bring AI Pathology to Surveillance and to Create Real World Clinical Research Datasets



Participating SEER Registries: New Jersey, Kentucky, Georgia, New York


ORIGINAL ARTICLE

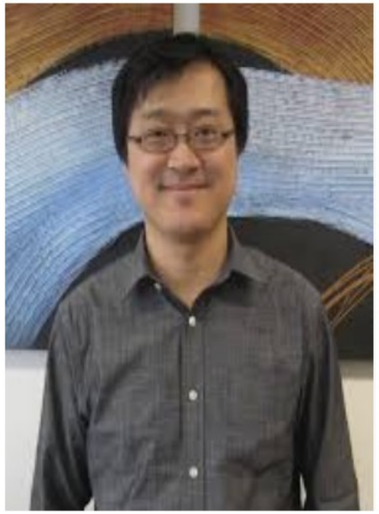
J Pathol Inform 2022, **13**:5

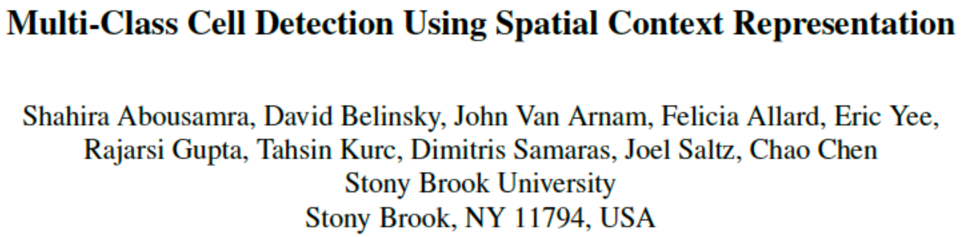
An expandable informatics framework for enhancing central cancer registries with digital pathology specimens, computational imaging tools, and advanced mining capabilities

David J Foran[1], Eric B Durbin[2], Wenjin Chen[3], Evita Sadimin[1], Ashish Sharma[4], Imon Banerjee[4], Tahsin Kurc[5], Nan Li[4], Antoinette M Stroup[6], Gerald Harris[6], Annie Gu[4], Maria Schymura[7], Rajarsi Gupta[5], Erich Bremer[5], Joseph Balsamo[5], Tammy DiPrima[5], Feiqiao Wang[5], Shahira Abousamra[8], Dimitris Samaras[8], Isaac Hands[9], Kevin Ward[10], Joel H Saltz[5]

Complex AI Pipelines
Spatial Contexts -- Cell Detection and Classification

- Classification accuracy is frequently context sensitive
- Training on new tissue types and new cell categories is time consuming
- Method detects and classifies nuclei
- Training requires “dotting” nuclei



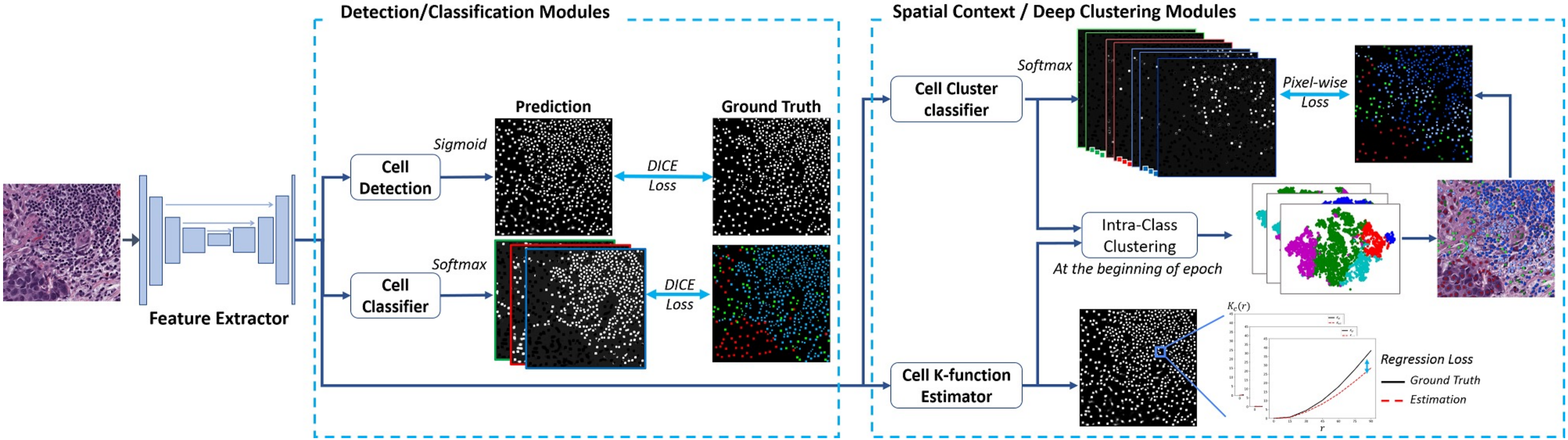


Multi-Class Cell Detection Using Spatial Context Representation

Shahira Abousamra, David Belinsky, John Van Arnam, Felicia Allard, Eric Yee,
Rajarsi Gupta, Tahsin Kurc, Dimitris Samaras, Joel Saltz, Chao Chen
Stony Brook University
Stony Brook, NY 11794, USA

2021 ICCV OCTOBER 11-17 VIRTUAL

Pipeline encompasses cell detection, cell classifier and learning category specific spatial statistics

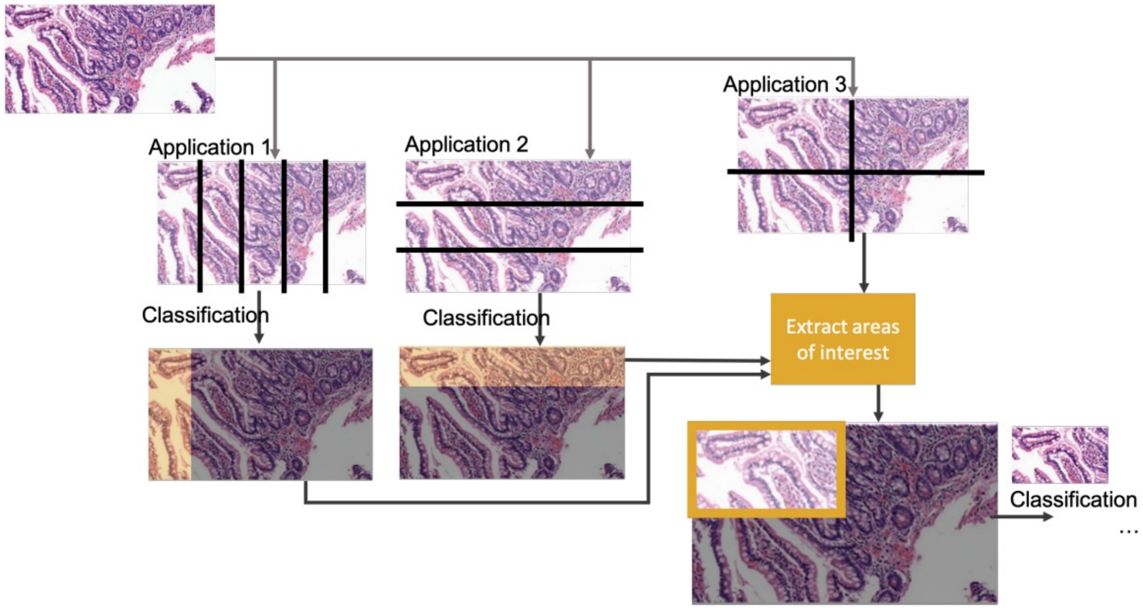


Extreme Scale Deployments

Understanding and leveraging the I/O patterns of emerging machine learning analytics

Ana Gainaru[1], Dmitry Ganyushin[1], Bing Xie[1], Tahsin Kurc[2], Joel Saltz[2], Sarp Oral[1], Norbert Podhorszki[1], Franz Pöschel[4], Axel Huebl[3], and Scott Klasky[1]

[1] Oak Ridge National Laboratory, Oak Ridge, USA,
[2] Stony Brook University, New York, USA,
[3] Lawrence Berkeley National Laboratory (LBNL), Berkeley, USA
[4] Center for Advanced Systems Understanding, Görlitz, Germany



- Middleware for managing image and model data to optimize memory hierarchy performance in complex training and prediction tasks
- Collaboration with ORNL to explore impact using ADIOS and tumor infiltrating lymphocyte tasks

Observations

- Algorithms designed to do specific things -> generalized trainable algorithms
- Self supervision, multi-modal friendly methods such as transformers
- Multi-modal/Multi-task learning
- Combination of very large data, need for self-supervision, multi-modal integration creates a perfect storm for exascale (and beyond) computing requirements
- Analyses typically carried out with <100,000 whole slide images (usually much smaller) – large institutions now have 10M+ whole slide image datasets
- Need for extreme scale system software to control complexity, validate pipelines and to optimize performance

Stony Brook Multi-Modal Deep Learning Faculty

Dimitis Samaras, Chao Chen, Tahsin Kurc, Prateek Prasanna, Raj Gupta

SEER UG3 Team

- Stony Brook
  - Joel Saltz MD, PhD
  - Tahsin Kurc PhD
  - Dimitris Samara PhD
  - Erich Bremer
  - Le Hou
  - Shahira Abousamra
  - Raj Gupta
  - Han Le
  - Bridge Wang
- Emory
  - Ashish Sharma PhD
  - Ryan Birmingham
  - Nan Li
- Georgia SEER Registry
  - Kevin Ward MPH, PhD
- Rutgers
  - David J. Foran PhD
  - Evita Sadimin, MD
  - Wenjin Chen, PhD
  - Doreen Loh M.S.
  - Jian Ren, Ph.D
  - Christine Minerowicz, MD
- Rutgers SEER/ NJ State Cancer Registry
  - Antoinette Stroup, PhD
  - Adrian Botchway, CTR
  - Gerald Harris, PhD
- University Kentucky; Kentucky SEER Registry
  - Eric B. Durbin, DrPH, MS
  - Isaac Hands, MPH
  - John Williams, MA
  - Justin Levens

QuIP ITCR Team

Stony Brook University
Joel Saltz
Tahsin Kurc
Raj Gupta
Dimitris Samaras
Erich Bremer
Fusheng Wang
Tammy DiPrima
Le Hou

NCI / DCEG
Jonas S Almeida

Emory University
Ashish Sharma
Ryan Birmingham
Nan Li

University of Tennessee Knoxville
Jeremy Logan
Scott Klasky

Harvard University
Rick Cummings

ITCR/IMAT Supplement
Richard Levinson
Raj Gupta
Tahsin Kurc
Joel Saltz
Han Le
Maozheng Zhao
Dimitris Samaras

The PRISM Team

UAMS
University of Arkansas for Medical Sciences

- **Fred Prior, PhD**
- Jonathan Bona, PhD
- Kirk Smith
- Lawrence Tarbox, PhD
- Mathias Brochhausen, PhD
- Roosevelt Dobbins
- Tracy Nolan
- William Bennett

EMORY
UNIVERSITY

- **Ashish Sharma, PhD**
- Annie Gu
- Mohanapriya Narapareddy
- Monjoy Saha, PhD
- Pradeeban Kathiravelu, PhD

QARC
QUALITY ASSURANCE
REVIEW CENTER

- TJ Fitzgerald, MD
- Fran Laurie

Stony Brook
University

- **Joel Saltz, MD, PhD**
- Erich Bremer
- Rajrisi Gupta MD
- Tahsin Kurc, PhD
- Tammy DiPrima

Thanks - Funding

- U24CA180924, U24CA215109, NCIP/Leidos 14X138 and HHSN261200800001E, UG3CA225021-01 from the NCI; R01LM011119-01 and R01LM009239 from the NLM, Award 2123920 from the National Science Foundation Bob Beals and Betsy Barton, Stony Brook Mount Sinai Seed Funding, Pancreatic Cancer Action Network

- This research used resources provided by the National Science Foundation XSEDE Science Gateways program and the Pittsburgh Supercomputer Center BRIDGES-AI facility