Deep learning from multi-expert annotations: need for prior consensus or not?

Christine Decaestecker

FNRS
Laboratory of Image Synthesis and Analysis, EPB, ULB

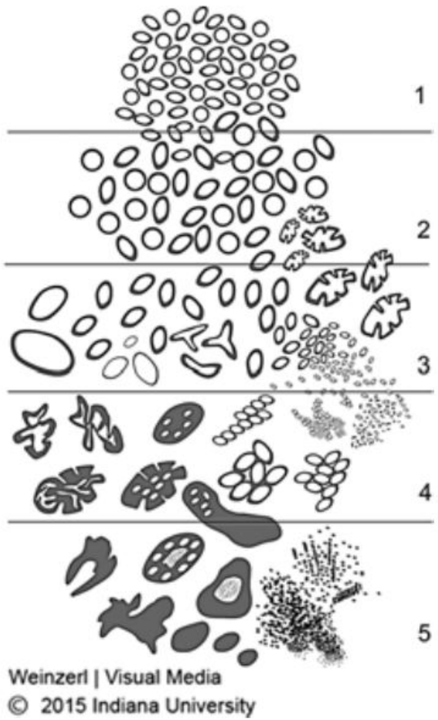
TRAIL doctoral school seminar
UCL – February 2023

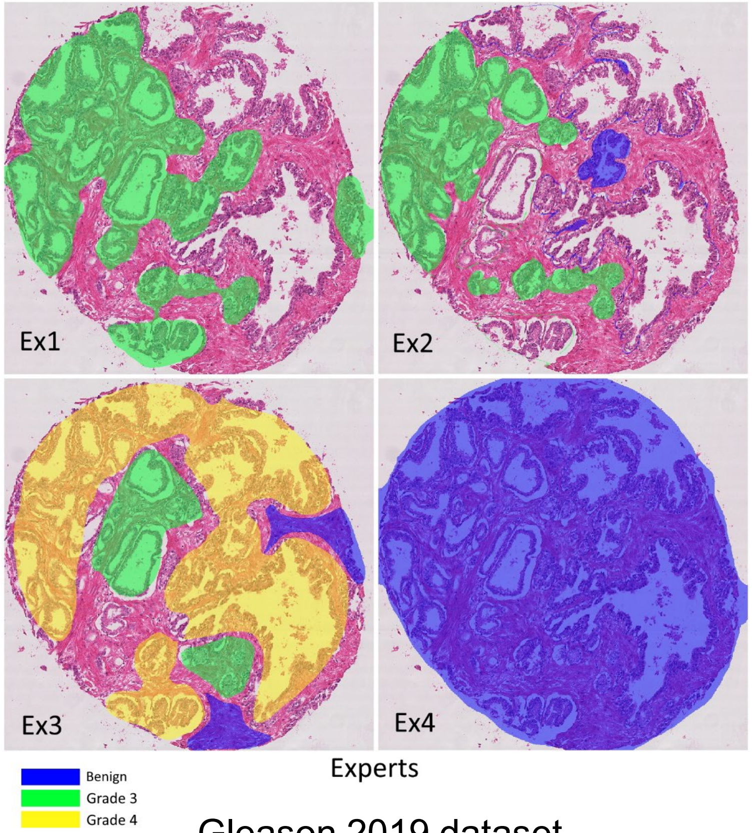
Expert annotations in medical imaging

- Time-consuming
- Subjective
- Expert's experience
- **High inter-expert variability**

No actual ground truth for *training* and *assessing* machine learning models

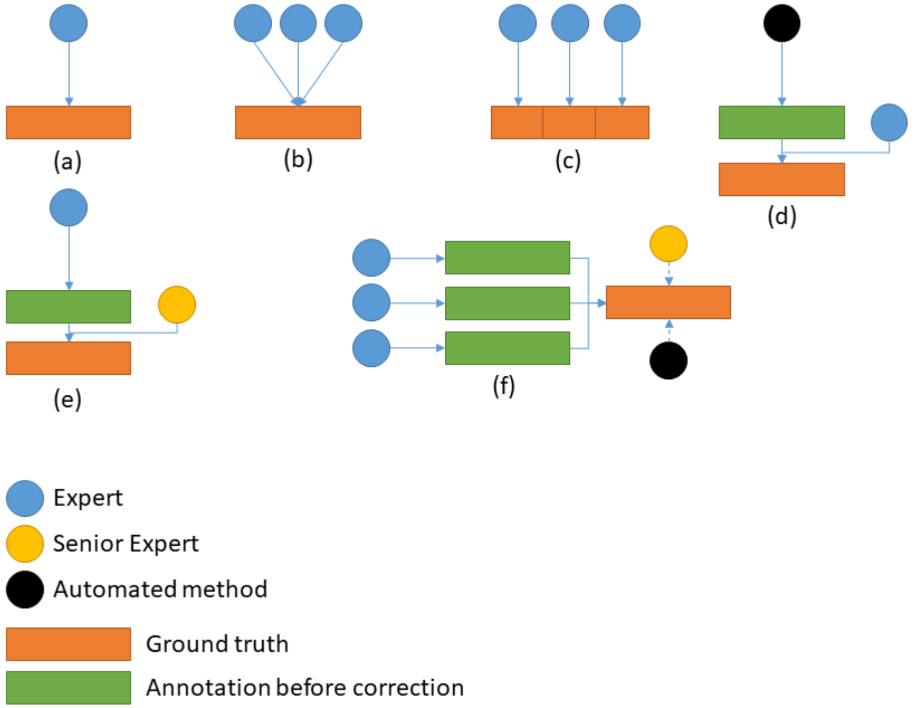
Gleason grade





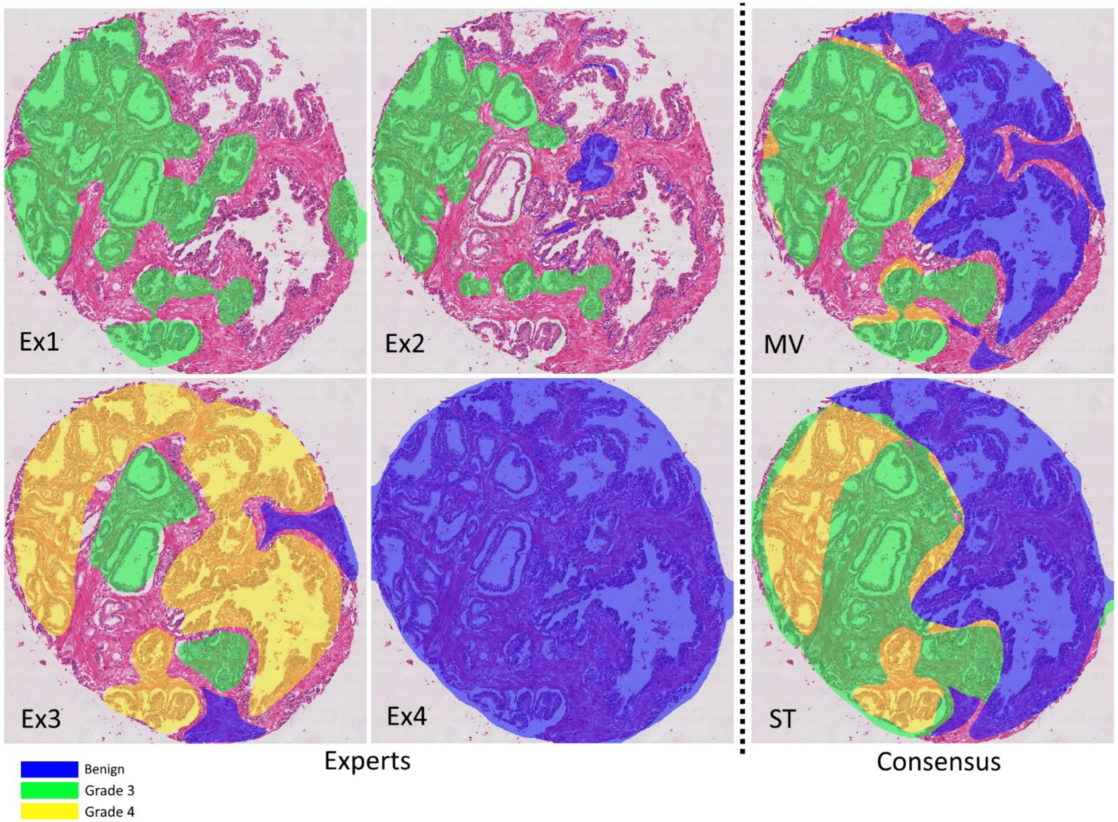
Gleason 2019 dataset

GENERATING "GROUND TRUTH" ANNOTATIONS IN THE MEDICAL FIELD



(a) Single expert
(b) Multiple experts working collegially
(c) Multiple experts working independently on subsets (possibly training and testing sets)
(d) Automated method refined by expert(s)
(e) Expert with senior review
(f) Multiple experts working independently on the same set, with automated consensus (or senior review)

AUTOMATED CONSENSUS



Pixel-wise **Majority Vote**

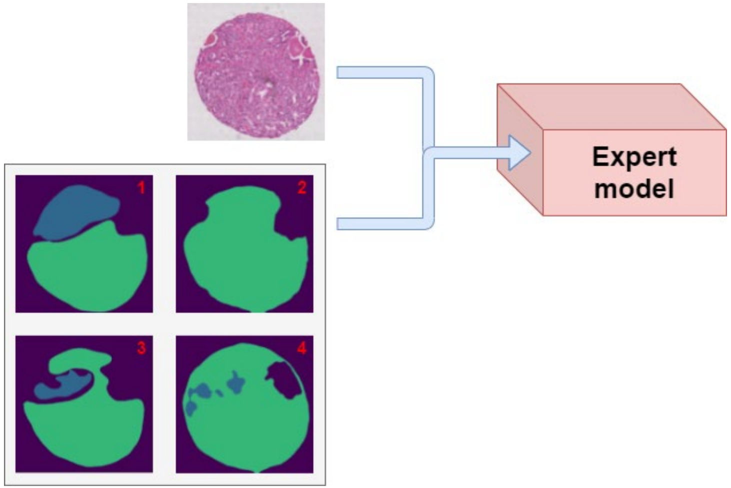
STAPLE (*Simultaneous Truth and Performance Level Estimation*):

- available in Python library SimpleITK
- expectation-maximization (EM) algorithm
- estimates simultaneously the "ground truth" and the confusion matrix characterizing each expert
- with a spatial homogeneity constraint (via additional embedded iteration process)

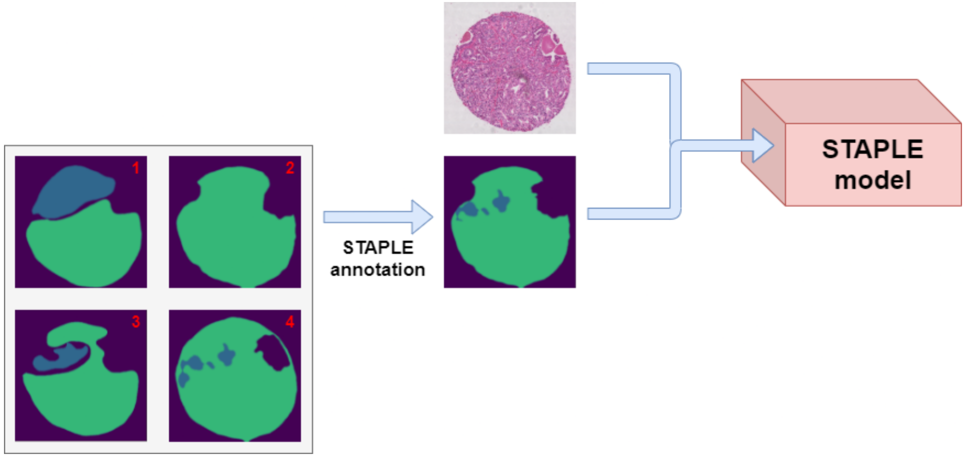
⇒ rather heavy computation

Gleason 2019 challenge dataset

Using either **multi-expert annotations** or a **single consensus annotation** to **train a deep neural network** for the purpose of **automating prostate cancer grading (Epstein scoring)**



Multi-expert annotations



Single consensus annotation

- **Epstein score : based on the 2 most prevalent Gleason grades**

  (simplified : Grade 5 merged with Grade 4 = Grade 4+, due to very few examples)

Simplified Epstein score	Most prevalent grade	Second most prevalent grade
Epstein 0	None	None
Epstein 1	Grade 3	None
Epstein 2	Grade 3	Grade 4+
Epstein 3	Grade 4+	Grade 3
Epstein 4	Grade 4+	None

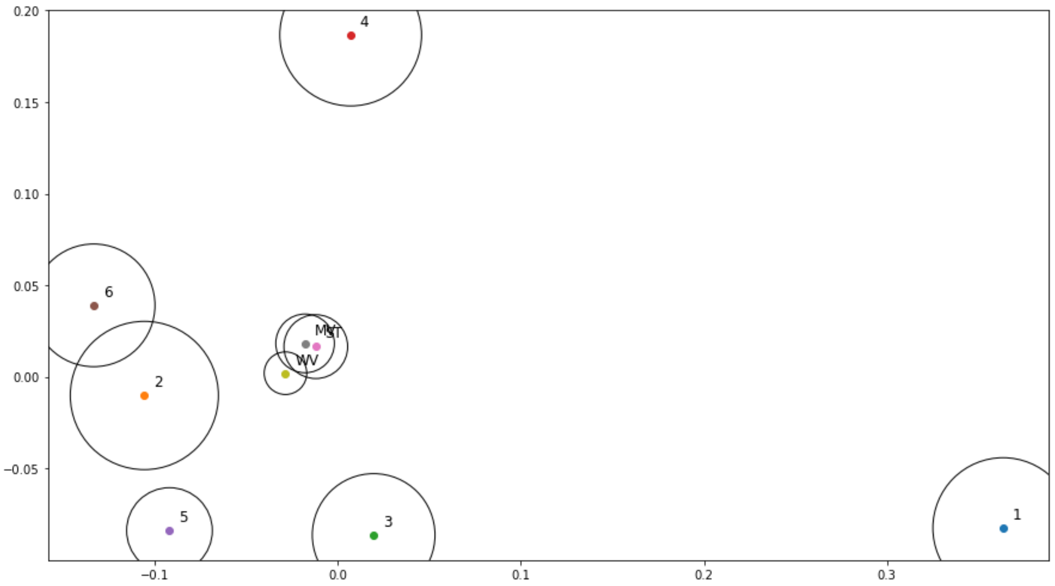
(Complete classification :
Epstein 3 = Grade 4 & Grade 3
Epstein 4 = Grade 4 (all)
Epstein 5 if includes a Grade 5 lesion)

AUTOMATED CONSENSUS FOR EPSTEIN'S SCORING

Disagreement / Dissimilarity matrix
$(1 - \kappa_Q)$

	E1	E2	E3	E4	E5	E6	ST	MV	WV
E1	0	0,52	0,43	0,47	0,44	0,47	0,35	0,36	0,38
E2	0,52	0	0,14	0,33	0,08	0	0,04	0,06	0,06
E3	0,43	0,14	0	0,23	0,16	0,26	0,11	0,11	0,1
E4	0,47	0,33	0,23	0	0,25	0,24	0,15	0,15	0,17
E5	0,44	0,08	0,16	0,25	0	0,18	0,13	0,12	0,1
E6	0,47	0	0,26	0,24	0,18	0	0,12	0,11	0,11
ST	0,35	0,04	0,11	0,15	0,13	0,12	0	0,02	0,03
MV	0,36	0,06	0,11	0,15	0,12	0,11	0,02	0	0,02
WV	0,38	0,06	0,1	0,17	0,1	0,11	0,03	0,02	0

Multi-Dimensional Scaling projection

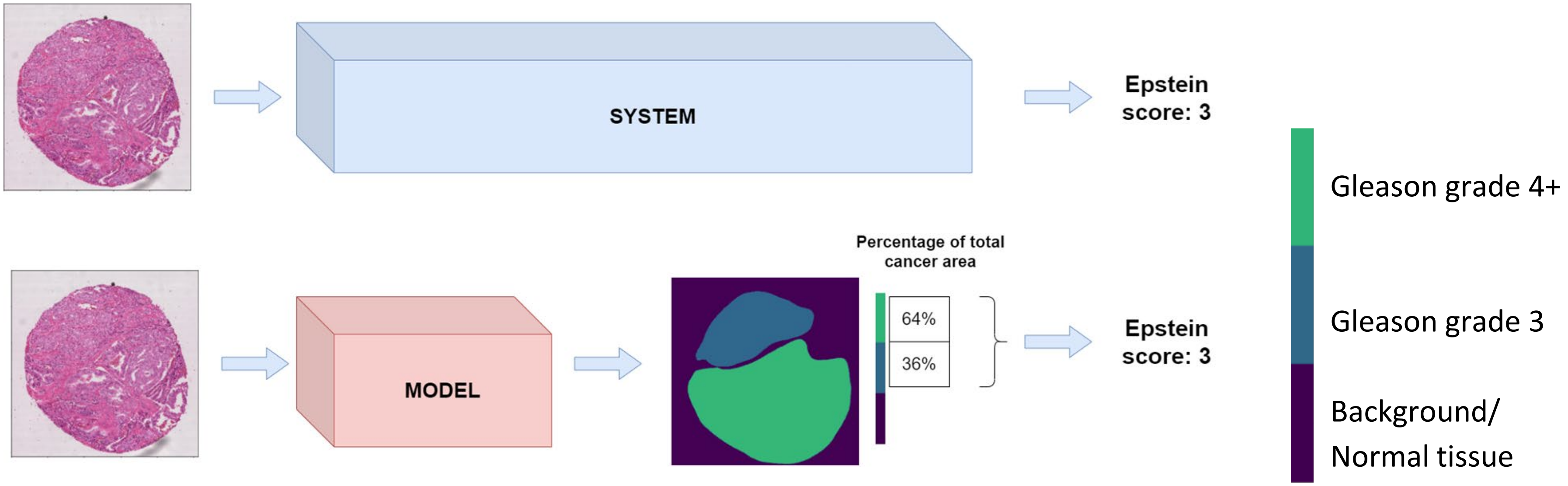


Gleason 2019 dataset

WV: weights each expert by its average head-to-head agreement with the other experts (computed with the unweighted kappa).

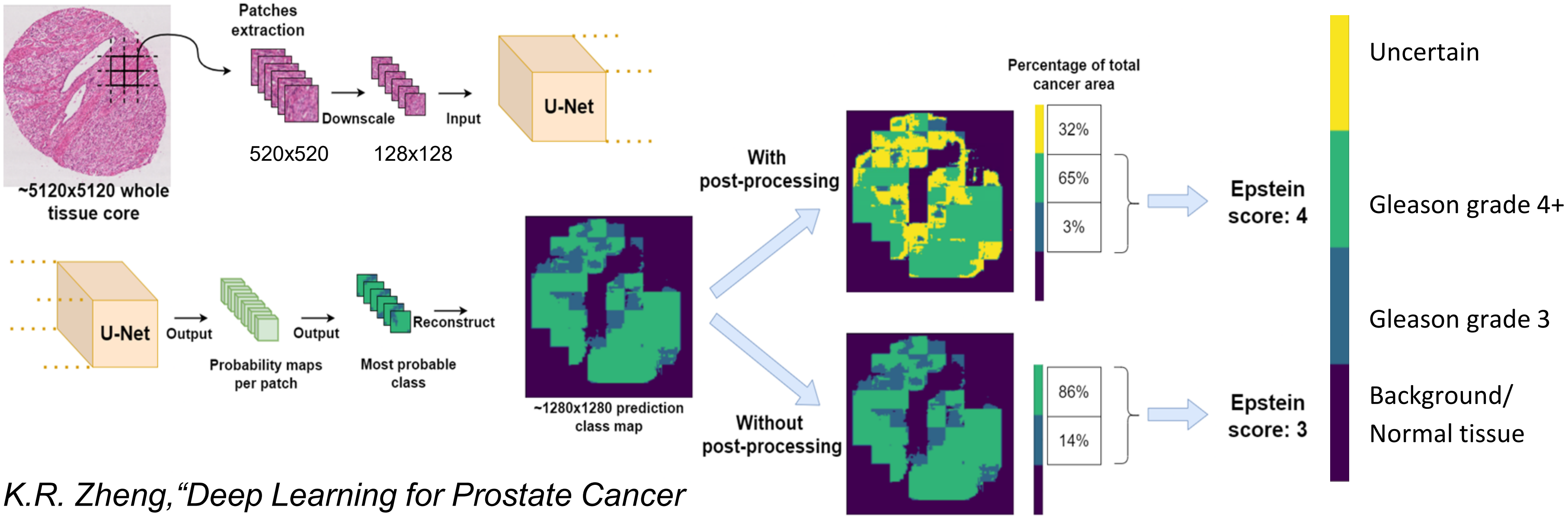
EXPLORING DIRECT MODEL TRAINING FOR PROSTATE CANCER GRADING

- Input: Tissue core
- **System output: Epstein score**
- Intermediate: Class map (explainability)



EXPLORING DIRECT MODEL TRAINING FOR PROSTATE CANCER GRADING

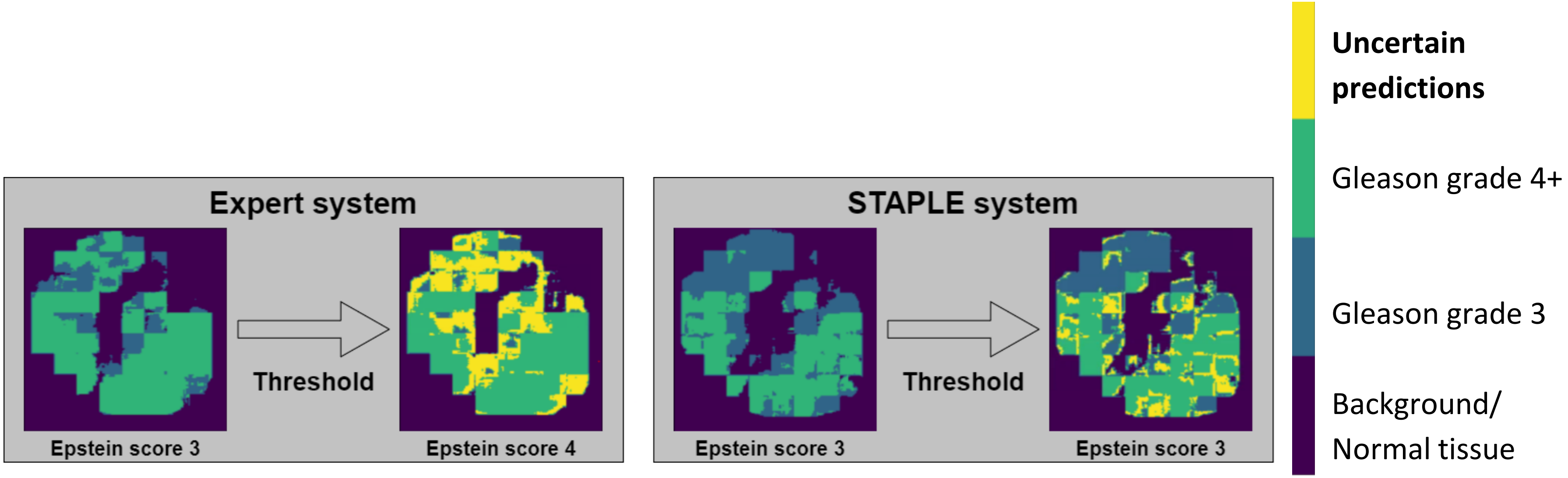
- Tissue patches
- Output = maps of probabilities to belong to each grade class
- **Post-Processing**: Identifying **uncertain** predictions ($P_{max} \leq 2*P_2$)



A.K.R. Zheng,"Deep Learning for Prostate Cancer Grading," Master Thesis, ULB, 2021.

EXPLORING DIRECT MODEL TRAINING FOR PROSTATE CANCER GRADING

1. The « Expert system » returns more uncertain predictions
2. The proportion of uncertain pixels can be used to highlight difficult cases, requiring more advice from experienced pathologists
3. Removing them leads to better Epstein predictions



EXPLORING DIRECT MODEL TRAINING FOR PROSTATE CANCER GRADING

- McNemar test to compare predictions of the 4 systems (on an independent test set):

		Post-Processing	
		Without Threshold	With Threshold
Network training supervision	STAPLE	STAPLE w/o Th	STAPLE Th
	Expert	Expert w/o Th	Expert Th

- **Need of a ground truth for evaluating performance metrics**

  ⇒ High inter-pathologist variability

  ⇒ STAPLE annotations to estimate “ground truth” Epstein score

EXPLORING DIRECT MODEL TRAINING FOR PROSTATE CANCER GRADING

- **Post-processing significantly improves the expert system that outperforms the STAPLE system**

		Post-Processing		
		Without Threshold		With Threshold
Network training supervision	STAPLE	Predicted 23/77	**0.227 <->**	Predicted 28/77
		| 0.263		**| 0.007**
	EXPERT	Predicted 29/77	**<-> 0.001**	Predicted 44/77

McNemar test: P-values are in bold
Number of agreements with STAPLE-based Epstein scores

EXPLORING DIRECT MODEL TRAINING FOR PROSTATE CANCER GRADING

- Inter-pathologists agreements (R_K*) range: 0.37 to 0.55 (excluding Pa. 2 and Pa. 6 with too few annotations)
- Post-processed expert system has the best agreement levels with the pathologists

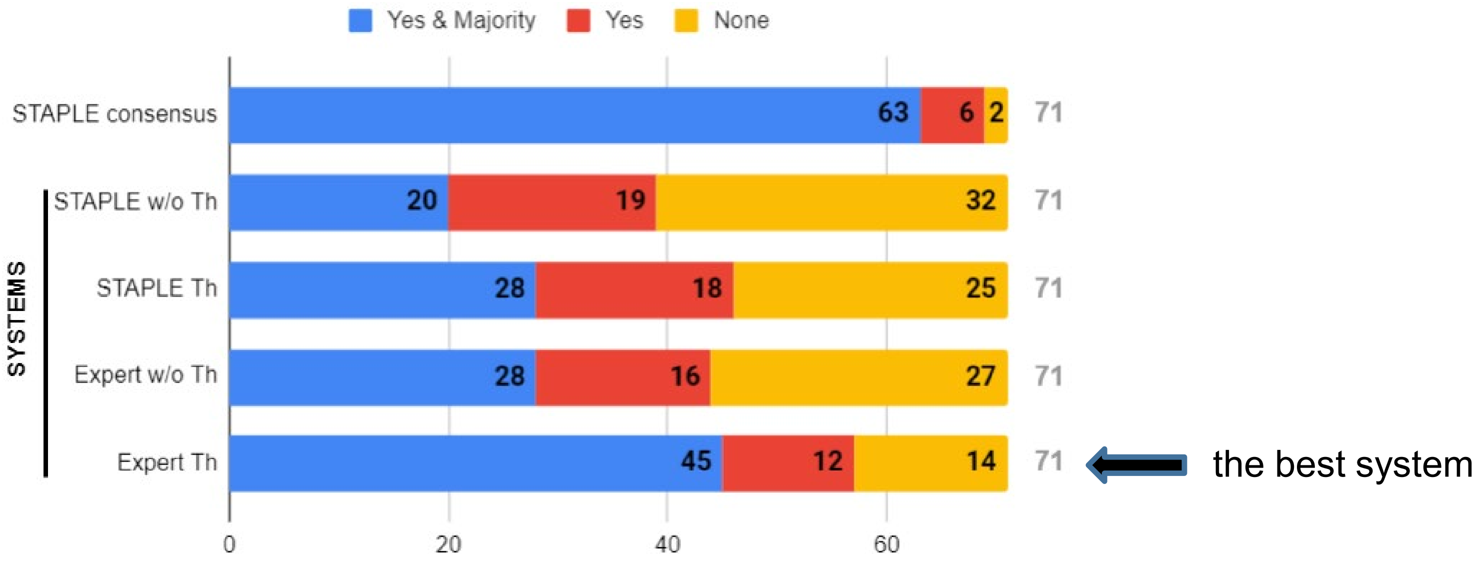
		SYSTEMS			
	STAPLE	STAPLE w/o Th	STAPLE Th	Expert w/o Th	Expert Th
Pa. 1	0.47	0.14	0.22	0.23	0.40
Pa. 2	0.84	0.46	0.66	0.56	0.68
Pa. 3	0.66	0.35	0.45	0.39	0.47
Pa. 4	0.73	0.24	0.37	0.40	0.50
Pa. 5	0.67	0.33	0.37	0.52	0.57
Pa. 6	0.71	0.20	0.31	0.32	0.52

Purple values = agreements with high number of annotated cores (tissue samples)

* Multiclass Matthews correlation coefficient (more reliable than kappa on unbalanced data sets, Delgado & Tibau (2019) PLoS ONE 14(9): e0222916)

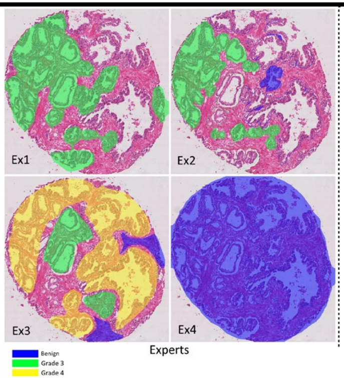
EXPLORING DIRECT MODEL TRAINING FOR PROSTATE CANCER GRADING

- Performance evaluated using STAPLE ⇒ possibly biased
- Alternative: number of Epstein scores among the pathologists' scores and among the majority agreed score



How to train a (classification) model with multi-expert annotations?

Karimi, et al. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis, Medical Image Analysis, 2020.

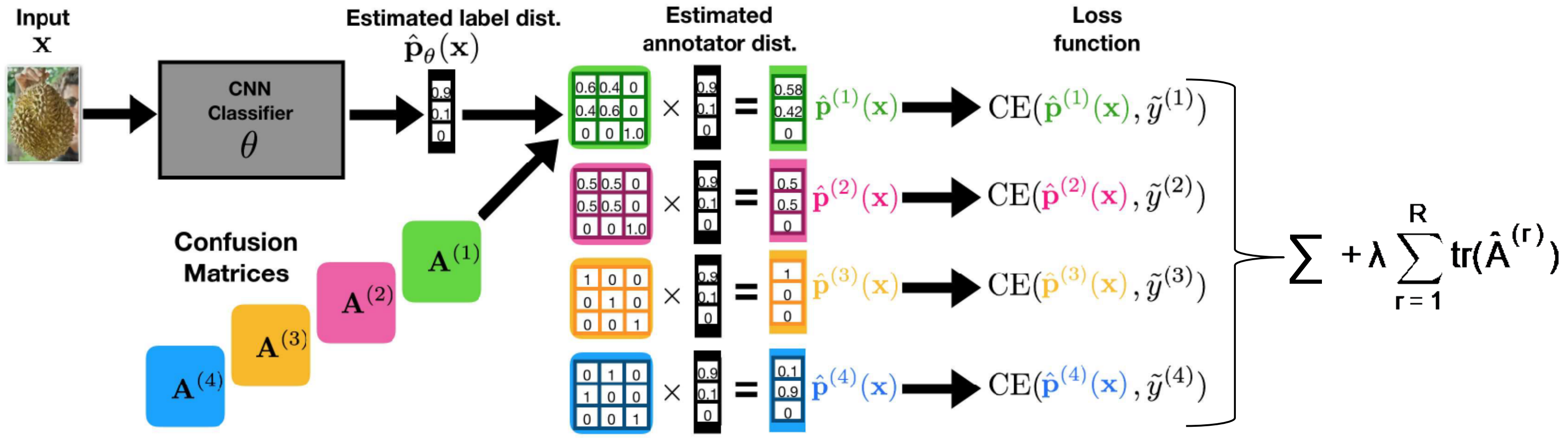


- **Single pathologist**: uses the label provided by one pathologists only (averaged model output)
- **Pixel-wise majority vote**
- **STAPLE**
- **STAPLE + iMAE[1] loss** (reduced the impact of large losses in mean absolute error)
- **Minimum-loss label**: for each training patch, selects the label with the smallest loss for error back-propagation.
- **Annotator confusion estimation[2]** (for image classification): simultaneously learns each individual annotator model (as a confusion matrix) and the underlying true label distribution (like STAPLE process but ”included” into the predictive model), using regularised cross-entropy loss function.

[1] *Wang et al. IMAE for Noise-Robust Learning. arXiv:1903.12141*

[2] *Tanno, et al. Learning from noisy labels by regularized estimation of annotator confusion. Proc. IEEE/CVF conf on computer vision and pattern recognition. 2019.*

Learning From Noisy Labels By Regularized Estimation Of Annotator Confusion (Tanno et al., 2019)



The model parameters $\{\theta, \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)}, \mathbf{A}^{(4)}\}$ are optimized to minimize the sum of four cross-entropy losses between each estimated annotator distribution $p^{(r)}(x)$ and the noisy labels $\tilde{y}^{(r)}$ observed from each annotator, with a regularisation term (= trace of mean $\mathbf{A}^{(r)}$).

Assumptions to be noted : (1) annotators are statistically independent, (2) annotation noise is independent of the input image (does not consider specific instance difficulty).

How to train a (classification) model with multi-expert annotations?

- **Results** (5-fold cross-validation)

    - ***Ground truth labels on the test data are estimated using STAPLE*** (“*given the high inter-observer variability, this would be our best estimate of the ground truth*”)

	Cancerous vs. benign		High-grade (4,5) vs. low-grade		% of large
Method	**accuracy**	**AUC**	**accuracy**	**AUC**	**classif. errors***
Single pathologist	0.8	0.78	0.65	0.61	0.07
Majority vote	0.86	0.87	0.73	0.74	0.03
STAPLE	0.84	0.86	0.73	0.72	0.03
STAPLE + iMAE loss	**0.93**	**0.91**	0.76	0.79	0.03
Minimum-loss label	0.88	0.88	**0.8**	**0.82**	0.03
Annotator confusion estimation	**0.92**	**0.93**	**0.8**	**0.82**	**0.01**
STAPLE (3-3)	0.86	0.86	0.69	0.7	0.02
STAPLE + iMAE loss (3-3)	0.9	0.88	0.75	0.78	0.02
Annotator confusion estimation (3-3)	0.9	0.88	0.73	0.76	0.03

* With a difference of at least 2 (ordered) classes (e.g. “benign” and grade 4 or 5)

Some conclusions

- Different approaches for handling multi-expert annotations, involving or not prior consensus for training

- When the ”ground truth” on the test set is produced using STAPLE consensus, training with consensus annotations

  - is not the most efficient

  - decreases the network ability to “learn” uncertainty

- Promising approaches for image classification (should be adapted for segmentation)

  - Annotator confusion estimation

  - Minimum-loss label

ACKNOWLEDGMENTS

Alain Zheng, Ir

Adrien Foucart, Ir, PhD

Alberto Franzin, Ir, PhD

Laura Galvez Jiménez, Ir