From Explainable to Explained AI

Ideas for Falsifying and Quantifying Explanations

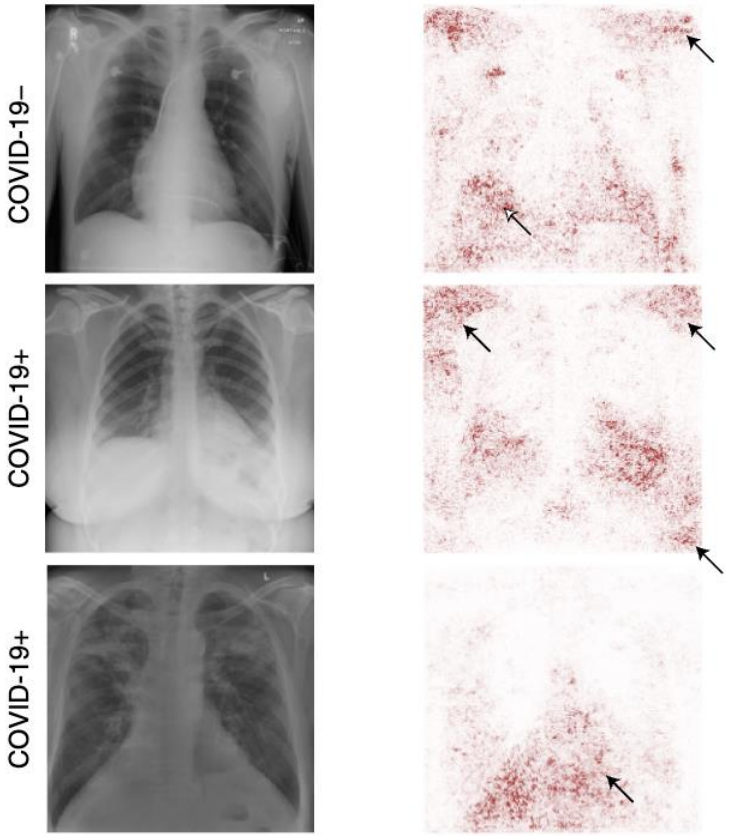
Yoni Schirris, Eric Marcus, Jonas Teuwen, Hugo Horlings, Efstratios Gavves
Netherlands Cancer Institute & University of Amsterdam

https://github.com/nki-ai/x2x

Tylenol leads to autism

Tylenol leads to autism(?)

Deep Learning…



DeGrave, Alex J., Joseph D. Janizek, and Su-In Lee. "AI for radiographic COVID-19 detection selects shortcuts over signal." *Nature Machine Intelligence* 3.7 (2021): 610-619.

Problem : The absence of ideas and methods for the falsification and quantification of explanations for AI models limits the individual researcher to confidently provide good explanations to discover and prevent unwanted actions by the model.

An explanation is…

a guess / conjecture / scientific theory about how something works

A good explanation should be…

Criticizable	Hard to vary	Non-authoritarian
“The sun sets and rises whenever the sun gods sleep and wake, because my mom told me so”	“The sun sets and rises because the sun revolves around the earth”	“The sun sets and rises because the earth rotates around its axis”

Let's test it out in computational pathology

An initial explanation for colorectal cancer prognosis classifier on H&E WSIs

MIL assigns higher scores to patches exhibiting:

(1) Poorly differentiated or highly proliferative tumor regions, often with epithelial–mesenchymal transition.

(2) Invasion into or infiltration of surrounding adipose tissue.

(3) Morphological features indicative of an aggressive tumor-stroma interface. Conversely, lower scores are associated with better differentiation and organized

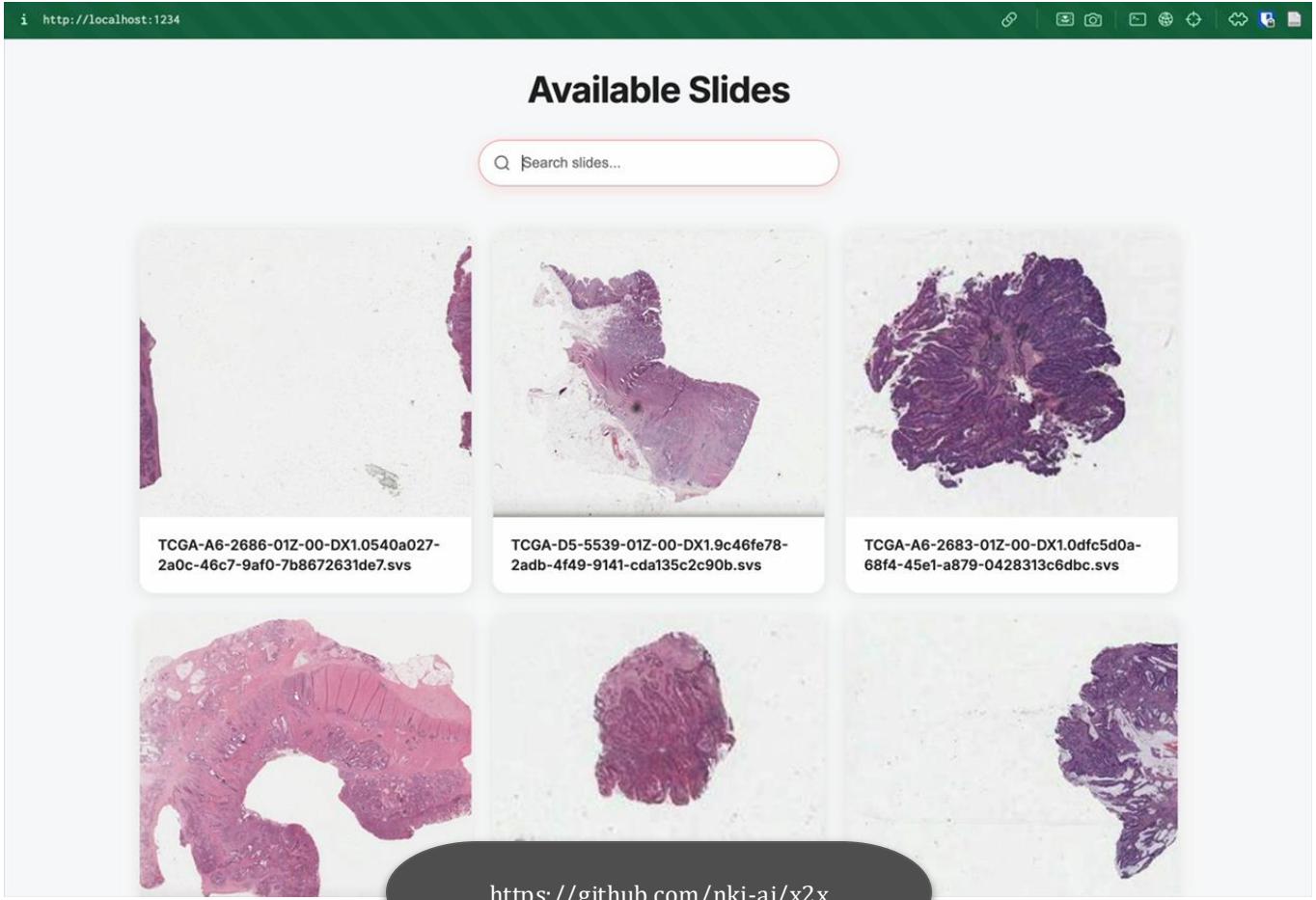
✅ Criticizable

✅ Hard to vary

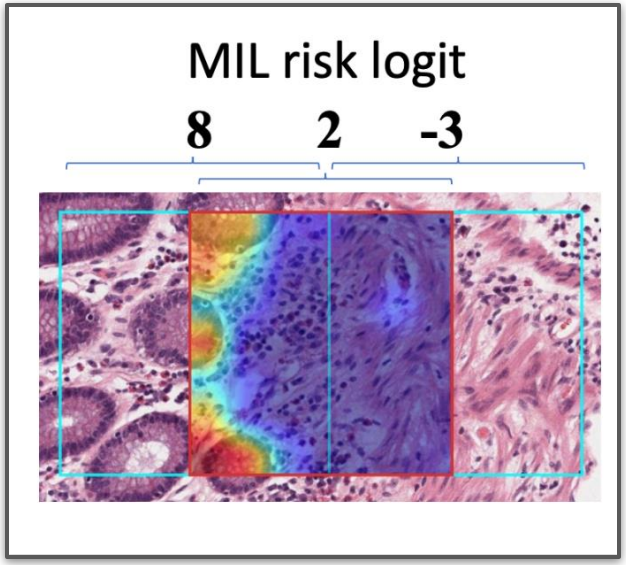
✅ Non-authoritarian

Falsification

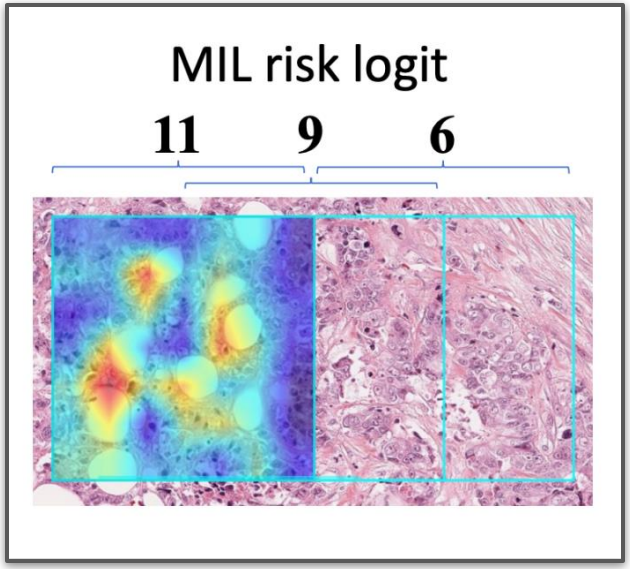
From Explainable to Explained AI: Ideas for Falsifying and Quantifying Explanations, by Yoni Schirris. https://arxiv.org/pdf/2508.09205. Gh: nki-ai/x2x



From Explainable to Explained AI: Ideas for Falsifying and Quantifying Explanations, by Yoni Schirris. https://arxiv.org/pdf/2508.09205. Gh: nki-ai/x2x



https://github.com/nki-ai/x2x



https://github.com/nki-ai/x2x

An initial explanation for colorectal cancer prognosis classifier on H&E WSIs

MIL assigns higher scores to patches exhibiting:

(1) Poorly differentiated or highly proliferative tumor regions, often with epithelial–mesenchymal transition.

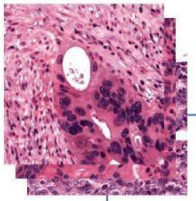
(2) Invasion into or infiltration of surrounding adipose tissue.

(3) Morphological features indicative of an aggressive tumor-stroma interface. ~~Conversely, lower scores are associated with better differentiation~~ and organized immune infiltration

(4) Healthy glandular formation??

Jiang, X., Hoffmeister, M., Brenner, H., Muti, H.S., Yuan, T., Foersch, S., West, N.P., Brobeil, A., Jonnagaddala, J., Hawkins, N., et al.: End-to-end prognostication in colorectal cancer by deep learning: a retrospective, multicentre study. The Lancet Digital Health 6(1), e33–e43 (2024)

Quantification



Black-box model

Risk score = 6.5

AUROC

GPT-4o

"Describe the image"

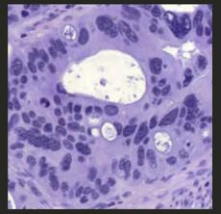
"The image shows a dense population of cells with hyperchromatic, round to oval nuclei."

GPT-4o
"Predict the expected model's output given the explanation"

Expected risk score: "Very high"

Explanation: "Aggressive tumor morphology leads to a high risk score"

Image



Model Output

- Logit: 4.754844665527344
- Attention: 0.09856200218200684

Image Description

Objective description:

- Approximately 40 cells with pleomorphic nuclei (~10-15 px in diameter) concentrated centrally.
- Presence of glandular lumens with partial mucin depletion, occupying about 15% of the area.
- Eosinophilic cytoplasm in most cells.
- Scattered mitotic figures (~5) observed throughout the patch.
- Moderate stromal component, comprising about 25% of the patch area.

Summary: Predominantly pleomorphic tumor cells with partial glandular formation and scattered mitotic figures in a moderate stromal background.

Prediction of Score

Prediction based on image features and current conjecture

The image description highlights pleomorphic nuclei, partial glandular lumens, and scattered mitotic figures amidst a moderate stromal background. According to the conjecture, aggressive tumor behavior is characterized by disorganized cell patterns and irregular nuclei, as seen in the poorly differentiated tumor regions. The presence of scattered mitotic figures may imply subtle signs of proliferative activity, aligning with molecular signals such as dysregulation in cell proliferation pathways (e.g., E2F targets, MYC, G2M checkpoints). Although the description lacks direct mention of fat-infiltrated areas or immune context, the conspicuous pleomorphism and mitotic figures are significant aggressive indicators within the conjecture's framework.

Short summary

Concluding, it is expected this patch is predicted to be VERY HIGH RISK. This is because the presence of pleomorphic nuclei, partial glandular depletion, and mitotic figures indicate poor differentiation and active proliferation, which are key aggressive features per the conjecture.

Final hard conclusion

CONCLUSION=VERY HIGH

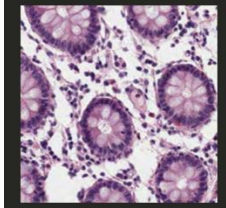
The explanation may predict the output of the model well

https://github.com/nki-ai/x2x

The explanation may predict the output of the model incorrectly: falsification

https://github.com/nki-ai/x2x

Image



Model Output

- Logit: 7.768579959869385
- Attention: -0.1550905704498291

Image Description

Objective description:

- Multiple glandular structures with circular to oval shapes, occupying approximately 60–70% of the visible area.
- Each gland is lined by columnar epithelial cells with basally located nuclei, consistent in size, approximating 5–7 px in diameter.
- Luminal appearance within glandular structures, clear and centrally located.
- Scattered nuclei of smaller diameter (~3–4 px) are visible in the interglandular stromal space, consistent with lymphocytes.
- Thin connective tissue stroma with sparse cell distribution surrounds the glands.

Summary: The image shows regularly arranged glandular structures lined by columnar epithelial cells, with interspersed lymphocytes and connective tissue stroma.

Prediction of Score

Prediction based on image features and current conjecture

According to the conjecture, features indicative of high survival risk involve poorly differentiated tumor regions, fat-infiltration, and signs of proliferative activity. The image description indicates regular glandular architecture with uniformly sized nuclei, suggesting well-differentiated tissue rather than aggressive tumor behavior. The presence of lymphocytes in the interglandular stroma suggests some immune presence, although the conjecture emphasizes that immune context is secondary. There are no mentions of adipose tissue infiltration, irregular nuclei, or highly disorganized patterns, which could indicate a high-risk score. The described features align more with lower-risk indicators rather than the aggressive features specified in the conjecture.

Short summary

Concluding, it is expected this patch is predicted to be VERY LOW RISK. This is because the regular architecture and presence of lymphocytes suggest well-differentiated tissue that does not match the high-risk morphological features described in the conjecture.

Final hard conclusion

CONCLUSION=VERY LOW

Combining them…

Takeaways

Thank you for your attention