Artificial Intelligence in Prostate Cancer Diagnostics

Dr. ir. Geert Litjens
Computational Pathology Group
Department of Pathology
Help, the robots are coming!
Most areas lack doctors

The current shortage of U.S. physicians is about 16,000, which affects about 25 million people. As of September 2005, there were 9,594 Health Professional Shortage Areas (HPSA).

- Had a population-group HPSA
- Had a partial-county geographic HPSA
- Was a whole-county geographic HPSA
- Did not have a geographic or population-group HPSA

Not Enough Doctors

Anticipated physician shortage for all specialties

100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0

'14 '15 '16 '17 '18 '19 '20 '21 '22 '23 '24 '25

Source: Association of American Medical Colleges

3% of histopathology departments have enough staff to meet demand.
Hurray, the robots are coming!
Computer-aided diagnosis
A detection/diagnosis/quantification task involving medical images

Performance [a.u.]

human
A detection/diagnosis/quantification task involving medical images

Performance [a.u.]

human
A detection/diagnosis/quantification task involving medical images

- Baseline
- Human performance
- Maximally achievable performance
A detection/diagnosis/quantification task involving medical images

- Baseline
- Early generation CAD product
- Mature CAD product
- Human
- Maximally achievable
A detection/diagnosis/quantification task involving medical images

- Maximum achievable
- Human performance
- Mature CAD product
- Early generation CAD product
- Baseline
Timeline of computer-aided diagnosis

Data → Features → Prediction → Human
Prostate cancer risk models

<table>
<thead>
<tr>
<th>Risk</th>
<th>Stage</th>
<th>Prostate-specific antigen (PSA)</th>
<th>Gleason score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>T1c – T2a</td>
<td>&lt; 10 ng/mL</td>
<td>&lt; 7</td>
</tr>
<tr>
<td>Medium</td>
<td>T2b-c</td>
<td>10 – 20 ng/mL</td>
<td>7</td>
</tr>
<tr>
<td>High</td>
<td>T3</td>
<td>&gt;20 ng/mL</td>
<td>&gt;7</td>
</tr>
</tbody>
</table>
Timeline of computer-aided diagnosis
Lung cancer risk models

Calculator: Solitary pulmonary nodule malignancy risk in adults (Brock University cancer prediction equation)

\[ \text{Logodds} = (0.0287 \times (\text{Age} - 62)) + \text{Sex} + \text{FamilyHistoryLungCa} + \text{Emphysema} - (5.3854 \times ((\text{Nodulesize}/10)^{-0.5} - 1.58113883)) + \text{Noduletype} + \text{NoduleUpperLung} - (0.0824 \times (\text{Nodulecount} - 4)) + \text{Spiculation} - 6.7892 \]

\[ \text{Cancer probability} = 100 \times \left( \frac{e^{\text{Logodds}}}{1 + e^{\text{Logodds}}} \right) \]
Radiomics

Aerts et al. Nature Communications, 2014
Timeline of computer-aided diagnosis

Human

Data

Computer

Features

Prediction
SOON WE WON'T PROGRAM COMPUTERS. WE'LL TRAIN THEM LIKE DOGS

The End of Code —

BEFORE THE INVENTION of the computer, most experimental psychologists thought the brain was an unknowable black box. You could analyze a subject's behavior—ring bell, dog salivates—but thoughts, memories, emotions? That stuff was obscure and inscrutable, beyond the reach of science. So these behaviorists, as they called themselves, confined their work to the study of stimulus and response, feedback and reinforcement, bell to saliva. They waxed virtuoso on the logic of the bell-ringing dog and the 30-second bell-dozing rat, but never pondered mental life.

But in a few decades, the black box opened up. The computer became the black box. It changed the way we think about the brain. And, in some unexpected ways, it changed the way we think about the world.
SOON WE WON’T PROGRAM COMPUTERS. WE’LL TRAIN THEM LIKE DOGS
The breakthrough

The breakthrough

Performance in photo classification

Error in image classification (%)
Imagenet classification with deep convolutional neural networks
A Krizhevsky, I Sutskever, GE Hinton
Advances in neural information processing systems, 1097-1105

Improving neural networks by preventing co-adaptation of feature detectors
GE Hinton, N Srivastava, A Krizhevsky, I Sutskever, RS Salakhutdinov

Learning multiple layers of features from tiny images
A Krizhevsky, G Hinton
Technical report, University of Toronto 1 (4), 7

Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection
S Levine, P Pastor, A Krizhevsky, J Ilgaz, D Gullen
The International Journal of Robotics Research 37 (4-5), 421-430
LEARNING CURVE

Self-taught AI software attains human-level performance in video games

ALL SYSTEMS GO

At last — a computer program that can beat a champion Go player

Radboudumc
Queue the hype...

Journal papers on deep learning in medical imaging

OBJECTIVES: Assess the performance of automated deep learning algorithms at detecting metastases in hematoxylin and eosin-stained tissue sections of lymph nodes of women with breast cancer and compare with pathologist diagnoses in a diagnostic setting.

DESIGN, SETTING, AND PARTICIPANTS: Researchers challenged competition in 2014 with 135 teams from 23 centers in the Netherlands to develop automated solutions for detecting breast node metastases (November 2013–November 2014). A test series of whole-slide images from 2 centers in the Netherlands with 176 slides with and without metastases was chosen to challenge participants to build algorithms. Algorithm performance was evaluated in an independent test set of 129 whole-slide images (116 with and 13 without metastases). The same test set of corresponding glass slides was also evaluated by seven pathologists with time constraints (November 2015) from the Netherlands to ascertain pathologist-related variability across four years, simulating real-world pathology workflow, and too pathological workflow time constraints (WCT).

EXPOSURE: Deep learning algorithms submitted as part of a challenge competition or pathologist interpretation.

OUTCOMES AND MEASURES: The performance of specific metastatic features and the absence of lymph node metastases in whole images using receive operating characteristic curve analysis. The four algorithms participating in the competition achieved diagnostic accuracy as defined by areas under the receiver operating curve (AUC) for the algorithms and the best of the algorithms achieved AUCs of 0.86, 0.86, and 0.87, which were comparable with the pathologist interpretation in the absence of time constraints (mean AUC, 0.850; range, 0.831–0.863) for the top 10 algorithms. A U.S. 0.80 (0.06) for the pathologist interpretation.

CONCLUSIONS AND RELEVANCE: The setting of a challenge competition, where deep learning algorithms achieved better diagnostic performance than panel of pathologists, may have encouraged researchers to develop improved algorithms. The algorithm performance was comparable with an expert pathologist interpreting whole-slide images without time constraints. Whether this approach has clinical utility will require evaluation in a clinical setting.
A detection/diagnosis/quantification task involving medical images

- Baseline
- Early generation CAD product
- Mature CAD product
- Human
- Maximally achievable

Performance [a.u.]
Prostate Cancer Diagnosis
Histopathology data not digital
Whole-slide imaging
Digital acquisition of an entire histopathology slide
Whole-slide imaging
Detectie van metastasen in lymfeklieren
Cancer detection

Best AUC of 0.99!
But only 270 patients...
Tumor
1) Training of IHC network

Specimens are stained with CK8/18 and P63 to mark epithelial tissue and basal cell layer.

- Color deconvolution is applied to each slide. Only the channel representing the epithelial tissue is used, the rest is discarded.
- Artifacts are introduced due to imperfections in the staining and color deconvolution method (Example: top left corner).
- Artifacts are removed manually in selected regions. Training data is sampled from these regions.

A 5-layer deep U-Net is trained on the corrected IHC regions. Areas with artifacts are sampled more.

The IHC network produces precise segmentation masks given an IHC slide, independent of the color deconvolution.

Input data: 25 IHC WSIs (20 training, 5 validation)

2) Training of H&E network

- Slide pairs are registered on cell-level due to the use of restained slides and non-linear patch based registration.
- The trained IHC network is applied to each IHC slide. The network output is used as the training mask for the H&E network. No additional post processing or manual annotations are used.

A 6-layer deep U-Net is trained on H&E and the masks generated by the IHC network.

The trained H&E network segments epithelial tissue on H&E.

Input data: 62 restained and registered IHC/H&E pairs (50 training, 12 validation)
Gleason grading

1. Semi-automatic data labeling

1. 5834 prostate biopsies are used to develop the system. The training set is labeled semi-automatically.

2. First, a rough tumor outline is generated by a tumor detection system.

3. Non-epithelial tissue is then removed by an epithelium segmentation system.

4. The Gleason pattern from the pathologist’s report is assigned to the detected tumor area.

5. A system is trained on pure biopsies only (3+3, 4+4, 5+5). After training, the system can segment patterns individually.

Gleason grading

2. Refinement & training

6. The full training set is labeled using the network trained on pure biopsies. Reports are used to further refine the labels.

7. Using the new labels the final system is trained.

3. Grade group prediction

8. To evaluate, the final system is applied to the test set. Each gland is labeled with Gleason 3, 4, 5 or benign. The normalized percentages are used to compute the Gleason grade group.
Tumor grade:

Gleason Grade

Group 5

Gleason 5
Gleason 4
Gleason 3
Benign

20% benign
15% Gleason 4
65% Gleason 5

Artificial neural network
Reference standard

- 3 expert uropathologists
- 550 prostate biopsies
- Gleason growth patterns, tumor volumes & grade groups
Case distribution

Full test set:
- 550 biopsies
- 3 experts

Subset
- 100 biopsies (selected from the 550)
- 15 external pathologists
International comparison

15 pathologists
10 countries
Gleason grading

Observer experiment on 100 cores

- < 8 years experience
- 8-15 years experience
- 15+ years experience

All pathologists

> 15 years experience

8-15 years experience

< 8 years experience

Pathologists (median + iqr)
Classification – ROC

Internal test set:
Benign vs Tumor

Observer set:
Benign & grade group 1-2 versus grade group >=3

N=535

N=100

Mean ROC (AUC = 0.99 ± 0.00)

Mean ROC (AUC = 0.90 ± 0.03)
Usefulness of AI in practice

Before AI

After AI

Mean (before AI)

Mean (after AI)
Future directions

- Algorithm available via grand-challenge.org
- Decision thresholds for grade groups
- Correlation with survival / recurrence
- Direct prediction of survival / recurrence from morphological patterns