Improving the detection of breast cancer through artificial intelligence

By Prof. Nico Karssemeijer

The reading by radiologists of the mammograms generated in current breast screening programs is an extremely time-consuming process, a situation which is likely to be exacerbated by the increasing use of digital breast tomosynthesis systems in the future. In addition, even with the currently recommended process of using double readers, there are still a significant number of breast cancers which are not identified in screening mammography. Several years ago, it was thought that the use of Computer-Aided Detection (CAD) systems could address these issues, but in practice such CAD systems failed to live up to expectations.

The advent of extremely powerful algorithms generated by Deep Learning technology and massive training processes is now revolutionizing the field.

This article describes the performance of a new, commercially available system developed using Deep Learning and designed to provide focussed decision support to breast radiologists. Initial evaluation data are presented which demonstrate the potential of the approach, for example in the use of such Artificial Intelligence systems as "second readers".

Breast cancer screening programs have adopted international quality assurance guidelines with the aim of maximising the chances of achieving the main objective of screening mammography, namely the early detection of breast cancer. These guidelines cover all aspects that have an impact on screening outcomes, and have resulted in excellent technical quality control of imaging equipment, which is one of the fundamental and critical factors in screening. However, it is well established that, even when working with high performance equipment under optimal conditions, radiologists can still fail to detect breast cancer. The quality of human interpretation of mammograms appears to be one of the most difficult factors to control —indeed it can be argued that this is the single most important factor in the whole screening process. Extensive audits of breast cancer screening programs reveal that over 50% of cancers in screened populations were already visible on previous mammograms when it is known where to look. This holds for both screen-detected and interval cancers detected in between screening rounds. This figure has not changed over the years, not even after the introduction of digital mammography. The crucial question is whether these missed cancers could have been detected earlier, without leading to an unacceptable increase in false positives. Some decades ago researchers thought the answer to this question was simply yes: Computer Aided Detection (CAD) systems were designed to identify potential abnormalities in mammograms. The idea was that if radiologists were to carefully inspect locations marked by CAD they would not overlook cancers marked by the system. However, this turned out to be an illusion. Radiologists still miss cancers, even when they are marked by a CAD system. The assumption that radiological errors in screening are due to radiologists not looking in the right place was wrong. Therefore, it is not surprising that despite widespread use of CAD in practice, there is mounting evidence that current CAD technology is not fulfilling its promise [1].

To design a better system to support radiologists in screening, we, with radiologists, investigated in detailed experiments the causes of screening errors. We found that the biggest difficulty is not the actual detection of suspicious regions, but rather their assessment. Since many regions in mammograms can look somewhat suspicious, the hard part for the radiologists is the decision on which ones they should act and how. This led to the conclusion that the key to improve the reading of screening mammograms is to help radiologists with decision-making. The potential benefit of this approach is great, as was demonstrated in a study where we assessed the best achievable performance in detecting malignant soft tissue lesions in mammograms. We used a large series of screening mammograms in which breast cancer had been missed, mixed in with normal exams. By using a panel of radiologists independently interpreting the mammograms we found that assessment by the panel was much better than that of any of the individual radiologists. The sensitivity of the individual radiologists ranged from 15% to 48%, at a recall rate of 5%. With a panel of 8 radiologists, sensitivity at the same recall rate increased to 61% and the results suggest that with more readers this could increase even further. The hope is that by using smart Artificial Intelligence techniques this 'best achievable performance' can become within reach of screening programs, for the benefit of all women participating in screening.

"... even when working with high performance equipment under optimal conditions, radiologists can still fail to detect breast cancer..."

DEEP LEARNING

Remarkable advances in machine learning have resulted in a breakthrough in the field of computer image analysis over the

The Author

Prof Nico Karssemeijer, PhD [1,2]

1. Radboud University Medical Center Nijmegen, The Netherlands
2. ScreenPoint Medical BV, Nijmegen, The Netherlands

email: nico.karssemeijer@screenpointmed.com

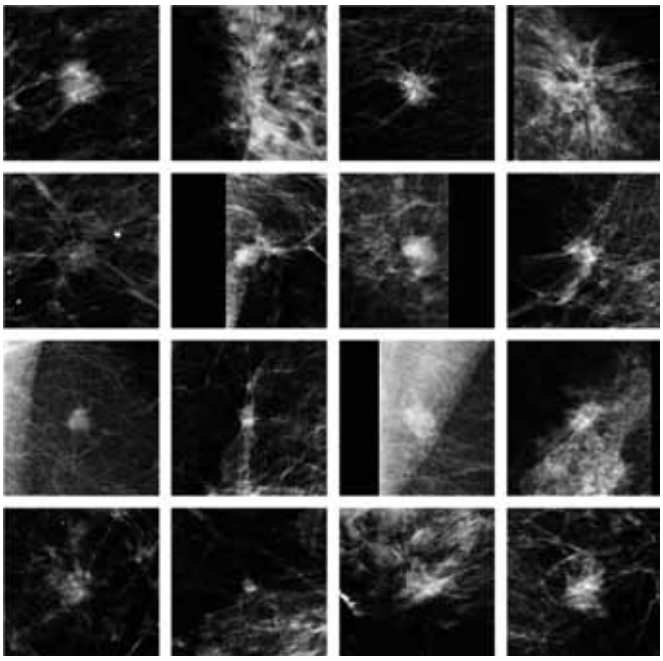


Figure 1. Deep learning systems are trained with small patches containing suspicious regions and learn to distinguish cancers from normal tissue or benign lesions. In one experiment 400 patches were interpreted by four experienced radiologists and by the deep learning system, a Convolutional Neural Network (CNN). The performance of the automated system was better.

last few years. Using the technique known as deep learning, artificial neural networks can be trained to recognize patterns in the same way as human beings. It is only a matter of time before such artificial networks outperform humans for certain specific tasks. The reading of screening mammograms is one task where conditions are ideal for the application of deep learning: the reading of mammograms is a repetitive task for which large amounts of reliable data are available for training. In addition, the imaging procedure itself is highly standardized.

The basic principle behind the deep learning approach is that the computer will learn to directly recognize features in images without the intervention of an expert who has previously "taught" the system what the features of interest are. In the training phase, literally millions of examples are presented to the system. Typically, these examples are sub-images (patches) of images containing a target (e.g. cancer) or a non-target pattern. After training, the system can distinguish the two types of patches. The more examples that are provided in training, the better the system learns the task.

In a recent study, we trained a deep learning system to recognize cancer in patches of 5x5 cm^2 extracted from mammograms [2]. Only soft tissue lesions were included. Typical examples are shown in Figure 1. Over one million patches were used to train the system. The performance of the system was subsequently compared to that of four experienced screening radiologists using a set of 400 patches that had not been seen by the system during training. It turned out that the deep learning system had a higher performance than the radiologists [3]. Even though this study had limitations — in practice the reading of mammograms is more complex than judging small regions of interest — this result nevertheless demonstrates the potential of the whole approach of the new machine learning technology.

DECISION SUPPORT

To assist radiologists with the interpretation of suspicious regions, we designed an interactive approach in which radiologists can select regions in mammograms for a second opinion provided by an AI system. In an experimental evaluation of this approach, we asked nine screening radiologists to read 200 exams with and without the support of a system for automated detection of malignant soft tissue lesions. The series contained 80 mammograms with cancer, 20 false positives, and 100 normal exams. The algorithms used in this study were developed at the Radboud University Medical Centre. Both the "classical" CAD approach and the interactive decision support approach were evaluated in the study. It was found that the use of the interactive decision support was highly effective, while results confirmed that just showing CAD marks — as is done in the classical CAD approach — in fact did not help the readers [4].

Given the rapid development of deep learning techniques in general, especially when coupled with massive training, it is now possible to produce much more powerful algorithms than were possible at the time of the Radboud study cited above. The remarkable progress made in the field over the past few years is shown in Figure 2. Using the same series of test exams as in the above-mentioned reader study, (which of course had not been used for the training process, so the algorithms had not previously "seen" the images), we compared the stand-alone performance of the system used in the earlier reader study with that of Transpara, a new system developed using deep learning by ScreenPoint Medical.

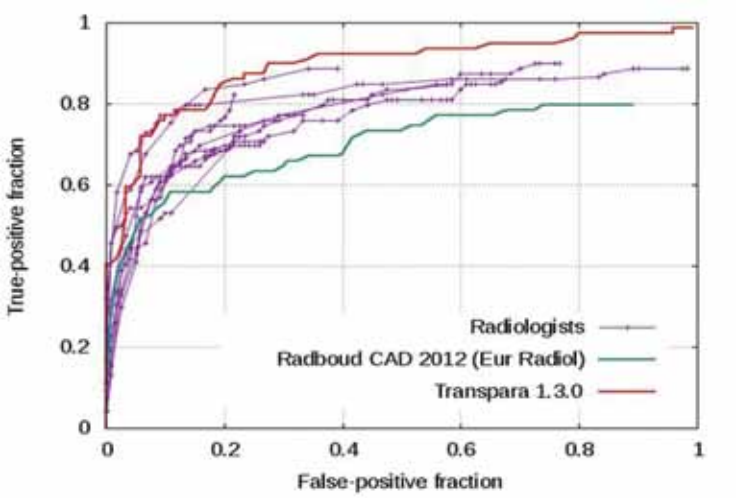
"... the detection performance of the new Transpara system .. for soft tissue lesions appeared to be as good as that of the best reader in this study..."

At the time of the earlier study the AI system developed at Radboud University still performed worse than the radiologists. In contrast, the detection performance of the new Transpara system (V 1.3.0) for soft tissue lesions appeared to be as good as that of the best reader in this study.

INCREASING PRODUCTIVITY

Breast cancer screening is a demanding and time-consuming task for radiologists. With the aim of improving quality, European Guidelines currently recommend independent double reading of screening mammograms by two radiologists. However, with increasing shortages of skilled screening radiologists, in practice such double reading may not be sustainable in the future. In fact, the problem of manpower/resources is likely to become significantly worse when breast screening programs transition to digital breast tomosynthesis, for the simple reason that it takes longer to accurately interpret all slices of a 3D tomosynthesis dataset than it does to read a mammogram. Artificial Intelligence may provide a solution to this problem, since when computers read mammograms or tomosynthesis data as well as radiologists the computers are serving as a second reader, thus potentially removing the need for double human reading.

To investigate the feasibility of this approach, we applied the deep learning system Transpara 1.3.0 to a series of 4,600 screening mammograms acquired using imaging systems from four different



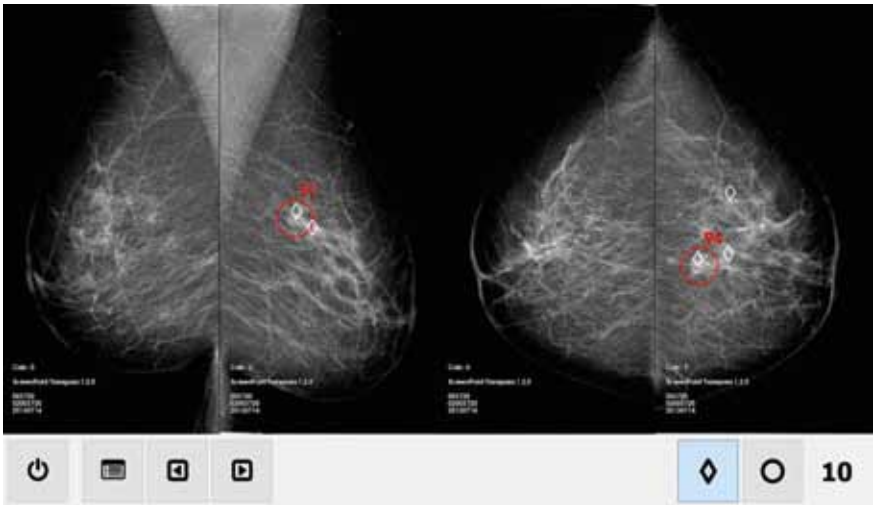


Figure 2: A conventional CAD system (Radboud CAD 2012) for detection of malignant soft tissue lesions was compared with the results of nine radiologists (Bottom Panel). While this conventional CAD system by itself performed less well than the radiologists, it helped them to improve detection [1].
The rapid progress of machine learning can be seen in the results of the Transpara deep learning system from ScreenPoint Medical, Nijmegen. The Top Panel shows the comparative ROC of the radiologists, the conventional CAD system and the Transpara system which can be seen to perform as well as the best radiologist reader in this test.

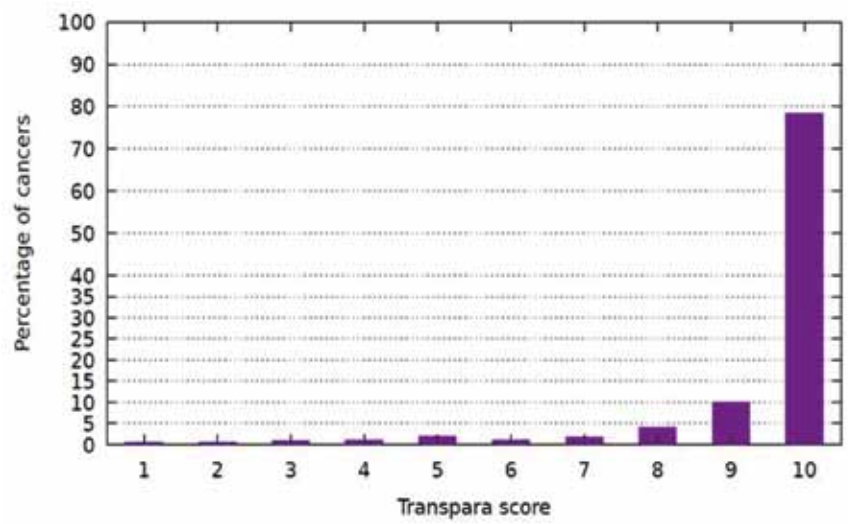


Figure 3. The Transpara deep learning system generates a score indicating the probability that breast cancer is present and detectable in a mammogram. In screening practice, the number of mammograms in each category is approximately equal. The figure shows the expected percentage of cancers in each category: 78% of exams with screen-detected cancers fall in category 10, while very few are in the lowest categories.

vendors, and including a representative series of 600 screen-detected cancers. The Transpara system combines suspicious findings in each screening study and categorizes the study (four mammographic views) according to its overall level of suspicion, attributing a score which reflects this level of suspicion. The scoring system ranges from 1 to 10 and was developed so that approximately 10% of the mammograms fall into each category. Results are shown in Figure 3. In the lower scoring categories very few cancers occur, whereas most exams with cancer fall into the highest scoring category. Thus, a higher score means a higher probability of cancer. It is expected that radiologists will increase their productivity through use of such a system, since it allows them to focus on the most relevant exams. For example, double reading of exams in the lower categories is probably not cost-effective or justifiable, given that the prevalence of cancer in exams in these categories is extremely low.

DISCUSSION

Breast cancer is a leading cause of death in women. In the EU more than 420,000 women are diagnosed with breast cancer each year and 130,000 women die of the disease. While early detection by screening is an effective way to reduce breast cancer mortality, it may not be possible to ensure the continued quality of screening programs in the future, due to the scarcity of skilled radiologists. Even with a sufficient number of experienced radiologists and the use of double reading, the quality of mammographic interpretation remains far from optimal. The fact is that the reading of screening mammograms is hard for humans. It is likely that in the near future computers will outperform radiologists. The use of AI is expected to reduce cost, increase the quality of screening programs, and reduce variability in the detection process.

REFERENCES

1. Lehman, C. D., Wellman, R. D., Buist, D. S., Kerlikowske, K., Tosteson, A. N., & Miglioretti, D. L. . Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA internal medicine. 2015; 175(11):1828-1837.
2. Kooi T, Litjens G, van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, den Heeten A, Karssemeijer N. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal. 2017; 35: 303-312.
3.Kooi, Thijs, *et al.* "A comparison between a deep convolutional neural network and radiologists for classifying regions of interest in mammography." International Workshop on Digital Mammography. Springer International Publishing, 2016.
4. Hupse R, Samulski M, Lobbes, Mann RM, Mus R, den Heeten GJ, Bierock D, Pijnappel RM, Boetes C, Karssemeijer N. Computer-aided detection of masses at mammography: interactive decision support versus prompts. Radiology. 2013; 266(1): 123-9.