The role of deep learning in breast screening

By Dr. H Harvey, Dr. E Karpati, Dr. G Khara, Dr. D. Korkinof, M O'Neill, Dr. A Ng, Dr. C Austin, T Rijken& Dr. P Kecskemethy

A bold vision for the future of breast cancer screening is required if programmes are to maintain double-reading standards against the backdrop of a workforce crisis across the UK, much of Europe and even Japan. Traditional Computer-Aided-Detection (CAD) for mammography decision-support could not reach the level of an independent reader. Our deep learning system, Mia (Mammography Intelligent Assistant), is on the cusp of providing consistent, accurate and interpretable mammography reading that can slot into current arbitration workflows. Mia provides the potential for single reading programmes, such as in the US, to reach EU double-reading accuracy, as well as providing new and practical support for adoption of the emerging modality of digital breast tomosynthesis.

The greatest challenge facing breast screening units is not that of accuracy, it is that there is a global radiology workforce crisis taking place against a backdrop of an exponential increase in imaging volume. For instance, there are 80 Breast Screening Units in England. The total number of women invited in 2016/17 in England for Breast Screening rose by 3.7% to 2.96 million. Of those, 2.2 million accepted the invitation and were screened, and 18,402 cancers were detected [1]. A recent UK workforce consensus demonstrated that 25% of units have two or fewer breast radiologists. Furthermore, between now and 2022, for every two breast radiologists that join the NHS, three are predicted to leave [2]. However, the UK is not alone. Japan, for instance, has the lowest proportion of radiologists per population in the G7, and it is estimated that an increase of staff by a factor of 2.5 is required in order to meet international standards [3]. Many other countries are also struggling to recruit and retain the required staff to run screening programmes effectively.

Currently in the EU every mammogram is 'double-read' by two independent radiologists. However, many screening centres struggle to fulfil double reading requirements in a timely manner. Additionally, there is wide variation in recall rates between different centres. In the US, where screening is not invitational and only single read, the accuracy of screening is lower, with significantly higher recall rates than double reading programmes [4]. Blinded double reading reportedly reduces false negative results and the average radiologist can expect an 8%–14% gain in sensitivity with double reading pairing [5]. Double-reading has undoubtedly improved accuracy, but at the cost of increasing human resource requirements.

CURRENT SOLUTIONS – COMPUTER AIDED DETECTION (CAD) SYSTEMS

After the advent of full field digital mammography (FFDM) at the turn of the millenium, medical imaging data became available in a format amenable to computational analysis. The huge volume of screening cases coupled with advances in computing (both from a hardware and software perspective) meant that attempts at automated diagnosis were inevitable. These early systems are known as CAD (Computer Aided Detection), of which iCAD Secondlook, and ImageChecker (by R2/Hologic) are the most widely known. As these products have been around for several years, there is a rich body of research, both positive and negative, into how effective they are, and whether they impact positively on patient outcomes [6-11].

The studies were conducted with different methodologies, sample sizes, population cross-sections and outcome metrics. While a one-to-one comparison is difficult, a number of key trends are easy to identify. Firstly, there is a significant level

The Authors

Hugh Harvey MBBS BSc(Hons) FRCR MD(Res),

Edith Karpati MD Rad.Spec.,

Galvin Khara BSc(Hons) MSc PhD,

Dimitrios Korkinof MEng PhD,

Michael O'Neill BSc(Hons) MSc,

Annie Ng PhD BSc,

Christopher Austin MD MSc, Tobias Rijken BSc(Hons) MSc,

Peter Kecskemethy BSc(Hons) PhD

Corresponding author:

Hugh Harvey, hugh@kheironmed.com

Kheiron Medical Technologies, London, UK

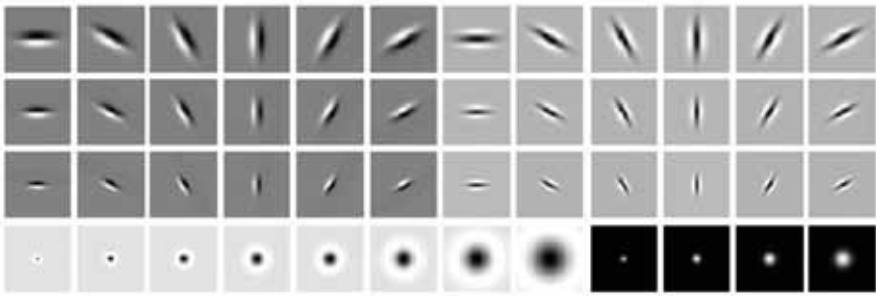

Figure 1. An example of a filter bank used in traditional CAD development. These filters cannot fully encompass the variety of pathologies seen in the real world. Reproduced from [15].

of variance between the conclusions regarding whether the technology is effective. This ambiguity has meant a low, or even lack of CAD adoption in Europe, and decreasing confidence is also illustrated by fewer and fewer US hospitals utilizing it (especially since reimbursement incentives have decreased in recent years).

CAD systems also fall outside the realm of blinded double-reading by their very positioning within the radiologists' workflow. They were designed to overlay Regions of Interest (ROIs) and present them directly to the reading radiologist, in effect giving them more information to interpret, and potentially biasing their decision. The outputs from these CAD systems could not be separated from the radiologist's independent review of the images as any final decision was made in conjunction with the CAD outputs, rather than independently from them. The overall consensus is that these systems output a large number of false positive marks on each case, which can increase reading times [12]. Any of these false positives triggering a recall will have a significant impact on downstream healthcare costs and patient welfare. For example, Elmore et al [13] demonstrated that for every $100 spent on screening with CAD, an additional $33 was spent to evaluate unnecessary false positive results. More importantly, a significant increase in women were undergoing undue stress and emotional worry as they feared for the worst. In a seminal prospective UK trial (CADET II) [14], although CAD was shown to aid single reader detection of cancer, there was a 15% relative increase in recall rates between human/CAD combination and standard double-reading by two humans.

The high proportion of false positive marks given by these early non deep-learning systems rendered them ineffective as a truly independent reader as they were not able to provide meaningful case-wise recall suggestions, hence the only viable output option was providing ROI outlines.

The flaws with traditional CADs stem from the underlying technology used. Traditional machine learning techniques are often referred to as 'expert systems'. In this case, the breast radiologist is the expert who develops hand crafted features and low-level pattern libraries in conjunction with an engineer, usually based on heuristics and pixel-distributions, that can correctly learn to classify objects in images [Figure1]. These are then written into code. In mammography, to achieve an optimal CAD tool, this requires a set of features capable of correctly detecting the diverse range of abnormalities that arise biologically in the breast. This is a demanding task, as microcalcifications are completely different in shape, texture, and size to masses, while the morphology of architectural distortions is even more subtle. It is intuitive and reasonable to assume that the standard low-level feature sets used historically would always be unable to capture the entirety of meaningful information contained in mammograms.

THE PROMISE OF DEEP LEARNING

Deep learning has already revolutionised many image analysis tasks. In 2012, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which challenges entrants to build systems to correctly classify the contents of an image (with 1000 possible classes), was won by AlexNet [16], a deep Convolutional Neural Network (CNN). It reduced error rates from (a then state of the art) 25.7% to 17.0%. This ushered in a paradigm shift for object recognition, with traditional machine learning techniques being universally replaced by deep learning (DL) based systems over the following years. In 2017, the final year of the ImageNet challenge, 29 of the 38 teams competing were able to achieve errors below 5.0% thanks to advances in deep learning (the current state of the art is 2.25%). This rise was facilitated by ever increasing amounts of digital data (as these CNNs require large numbers of images to train), and technological advances in graphical processing units (GPUs), the hardware which allows these models to be trained quickly.

The success of these models comes down to their flexibility. From a radiologist's perspective, an expert practitioner no longer needs to spend the majority of her/his time developing hand engineered features that capture specific lesion characteristics; instead the CNN is capable of learning the relevant features intuitively from large image datasets and their overall lesion labels. This flexibility comes at the cost of requiring larger datasets and significant computing power.

There has been an increase in research proposing the use of deep learning (DL) for mammography over the past few years, largely divided into two approaches: patch-based and case-wise (whole image).

Patch-based approaches break an image down into smaller regions for analysis, rather than taking an entire image as an input, as CNNs traditionally were developed to accommodate square input image sizes between 250 and 300 pixels in width and are not suitable out-of-the-box for mammography given the large size of images. These systems have traditionally been used to aid in localisation of lesions within an image. However, local decision-based approaches have a number of significant problems. The most notable is that when the many small decisions are re-combined back into the full image the result is typically a large number of false positives or

overall drop of performance in line with the increased complexity of the task.

Dhungeet *et al.*[17], and Ertosunet *et al.*[18] can be accredited with starting off the new wave of deep learning with hybrid approaches, combining traditional machine-learning with deep learning. The former suggested a cascaded CNN-based approach followed by a random forest and classical image post-processing. The latter published a two stage deep learning system where the first classifies whether the image contains a mass, and the second localizes these masses.

In 2016, the DREAM challenge was set up, inviting machine learning researchers to develop systems to detect breast cancers on a proprietary dataset for a monetary grand prize [19]. This was the first public competition to highlight DL's superiority in a mammography screening setting. The input data consisted of around 640,000 images of both breasts and, if available, previous screening exams of the same subject, clinical/demographic information such as race, age and family history of breast cancer. The winning team (Therapixel, France) attained a specificity of 80.8% at a set sensitivity of 80% (AUC 0.87) [20] with their DL system. Ribli *et al.* came second in the challenge with their DL system capable of not only classifying malignancy, but also outlining the regions in an image which supported this decision [21].

Carneiro *et al.* achieved an area under the Receiver Operator Curve (ROC) of 0.9 for malignancy classification on the publicly available DDSM [22] and inBreast datasets [23]. Their model was pre-trained on ImageNet (a tactic regularly employed by practitioners), and yielded significant improvements in mass and microcalcification detection. In 2017, Teare *et al.* [24] proposed a DL system that achieved a malignancy specificity of 80% at a sensitivity of 91% on DDSM, and their own proprietary dataset (which contained an equal number of malignant and non-malignant cases). Geras *et al.* developed a DL system capable of classifying screening cases into BI-RADS 0, 1, or 2 [25].

However the inability to identify malignant lesions limits its ability to perform as an independent reader. Finally, Kim *et al.* developed a system which made a malignancy prediction on an entire mammography case (all 4 views) [26]. The model was trained on malignant (biopsy proven), and normal (with at least 2 years of negative follow up) cases from multiple hardware vendors. They achieved a sensitivity of 75.6% at a specificity of 90.2% (with an overall AUC of 0.903). More recently Rodriguez-Ruiz *et al.* [27] demonstrated that a DL system (Transpara™, Screenpoint Medical, Nijmegen, The Netherlands) could increase the accuracy of radiologists using CAD-style displays of a likelihood of malignancy given a user-selected area.

While all of this research was conducted on various different datasets, with a variety of different outcome metrics, none of the above cited works (including the winners of the DREAM challenge) reached close to the performance of single human reading radiologists as a standalone system. Nevertheless, the progress over the last few years highlights how DL technology is moving beyond the past performance of traditional CAD systems, and closer to the goal of an independent reader.

Breast density assessment is another area of research interest for the deep learning community. In the US in particular, where density assessments are mandated in several states, DL systems can provide an automated BI-RADS density category. Such a system has

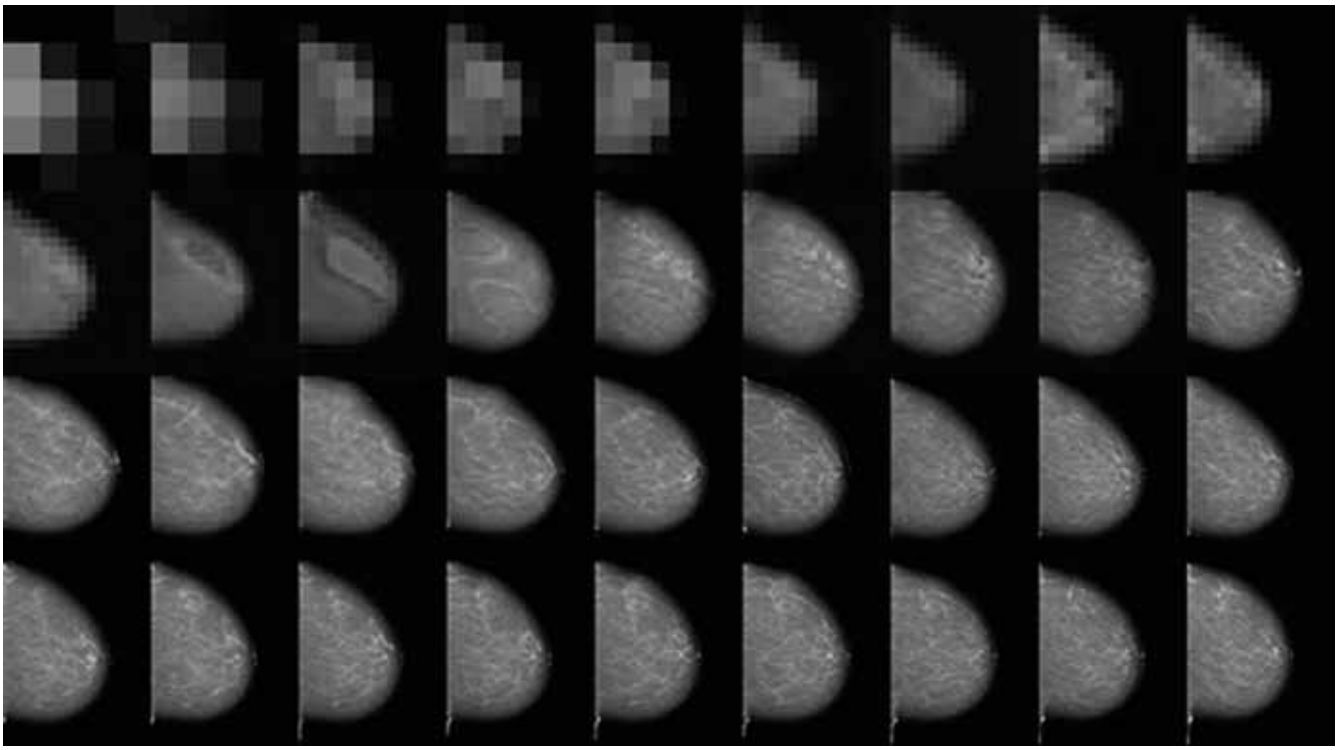

Figure 2. Progressive generation of synthetic mammograms from low to high pixel resolution. Reproduced from [30].

already been implemented in clinical practice at Massachusetts General Hospital [28].

Despite the data and computational resource requirements of DL, it has the potential to drive a revolution in medical image analysis, improving accuracy due to improved false positive and negative rates, increasing consistency and speed, automating analyses and speeding up assessment time. All of these elements provide useful practical support for screening programmes and services. The ultimate goal is to detect malignancies at a level that undoubtedly supports radiologists and breast units, which is at or beyond the level of an experienced single reader's average performance. However, current available mammography datasets are largely unrepresentative of the range of true screening cases in clinical practice [29], so further work is required in order to provide developers with the necessary data to train and validate clinically applicable models.

Generative Adversarial Networks (GANS) are a promising type of machine learning that can be used to synthesise medical images using features learnt from the latent space of a real dataset. Our group recently published a pre-print describing our success at creating high-resolution mammograms using GANs [Figure 2], [30], and postulated that there may be potential for synthetic data to help augment sparse training datasets and assist in domain transfer tasks.

INDEPENDENT READING

To truly make a positive impact in breast cancer screening, the largest pain point must be addressed urgently - that of a radiologist workforce in crisis. In order to achieve this ambitious goal, an automated system is needed that is able to make the same decision that a consultant radiologist makes when reviewing mammogram cases, with at least the same accuracy and consistency. This decision is a binary one - to callback a woman for further investigations or not - which case-wise deep learning systems are now able to achieve. We developed a DL system, known as Mia (Mammography Intelligent Assessment), trained on over 1 million real-world screening mammography images gathered to explicitly achieve this goal, and are the first group to receive regulatory approval for a deep-learning system to act as a second (or third) reader that provides case-wise callback decisions. Our initial retrospective evaluation of this system (under consideration for peer reviewed publication at the time of writing) on an unseen validation set of a screening cohort of 3860 patients (with outcomes proven by biopsy or at least 3 years negative follow-up) indicated that it compares favourably to established performance benchmarks for modern screening digital mammography [31], the criteria for identifying radiologists with acceptable screening mammography interpretive performance [32], and the minimally acceptable interpretive performance criteria for screening mammography [33]. Further studies will of course be needed to make direct comparisons with radiologists in a real-world setting, which is why we are working in partnership with the East Midlands Radiology Consortium (EMRAD), an established collaboration of seven NHS Acute Trusts in the East Midlands, UK, as well as several other NHS, EU and US sites to further validate Mia on a large cohort of screening cases.

The current European standard of double reading followed by arbitration, either via a third experienced radiologist or multi-disciplinary team, could be feasibly maintained by combining human and software decision-making while ensuring true blinding between readers. Just as there is arbitration now when two radiologists disagree on a callback decision, the same process could be applied when a human and machine disagree, and only then would the system's interpretation of the case be queried. In this manner, radiologists would be entirely uninterrupted to assess cases to the best of their training and expertise, reducing any potential for bias introduced from CAD outputs. Furthermore, an independent DL reader does not provide any extra distractions, or require additional clicks within the radiologist's workflow.

Gaining radiologist's acceptance that an automated independent reader is performing an analysis at or above human performance, which can be checked if necessary, will be an important step that has to be a high-priority focus area of work in the field. Indeed, there are already non-radiologist staff taking on the role of an independent reader, with many sites across the UK training and employing consultant mammographers in order to bolster their workforce [34], as such, there is already a move, driven by necessity, towards non-radiologist interpretation of images.

Some may argue that effectively replacing one of two humans within a clinical setting will alter the training of breast radiologists and also affect their required reporting numbers to maintain their qualifications. According to Woodard et al. [35] radiologists become more specific in their assessment of mammograms over their careers, but less sensitive. A system that performs as well as experienced radiologists from day one, with consistency, would help mitigate against this recognised learning curve, and has the potential to help in the training of the future workforce, all-the-while ensuring patients care is maintained, or even improved.

To incorporate an independent DL reader within double reading programmes would not require significant re-organisation of how these programmes are run. A system that integrates with current reporting methodologies into a workflow with arbitration is in theory relatively simple to deploy. However, in the US, and other single-reading nations, double-reading workflows are not the norm, and therefore the infrastructure of double-reading would need to be created, enabling the benefits of European double reading standards to be applied to these single reading programmes.

TOMOSYNTHESIS

While 2-dimensional FFDM is the the current standard for breast cancer screening, increasing amounts of research into 3-dimensional digital breast tomosynthesis (DBT) are gaining traction. Even though DBT has the potential to further increase the accuracy of cancer detection, it comes

at the cost of requiring more time to interpret, given the large amount of images produced, and a slight increase in radiation dose [36]. Mostly for these reasons, DBT has not yet seen widespread adoption for population screening where there are large volumes of cases, and instead is mainly used for symptomatic or difficult cases, such as dense breasts [37]. In an attempt to mitigate against the additional time-cost of interpreting these cases, vendors offer 2D synthetic images created from 3D tomosynthesis datasets. Machine learning techniques have been successfully applied to these synthetic images, providing an increase in radiologist accuracy for those that used CAD-enhanced synthetic mammograms when compared to standard 2D FFDM alone [38].

Due to the increased number of image slices in DBT, the labelling requirements are significantly increased, and large enough training and validation DBT datasets for deep learning on screening populations are not yet available. However, machine learning techniques such as domain transfer mean that DL systems trained on 2D mammograms are poised to be applied to DBT. This is achieved by leveraging the useful transferable features from a model trained for 2D mammography, so that a tomosynthesis model does not need to be redeveloped completely from scratch. Nevertheless, as DBT and DL use increases in practice, it is inevitable that independent DL systems will start to be used within this modality also.

CONCLUSION

Traditional machine learning approaches are sufficient to provide simple decision-support such as malignancy detection and density assessment, but our deep learning system, Mia, has the potential to shift the paradigm from simple CAD clinical-decision support to being a truly independent reader. Mia could enable single reading programmes to achieve the low recall rates and accuracy of double read programmes. By incorporating human-level, or superior, automated mammography and tomosynthesis analysis into an arbitrated workflow, the workforce crisis in breast screening could be somewhat mitigated. Reaching this goal requires collaboration between data scientists and clinicians, with consideration towards the requirements of large-scale data sharing and computational resource costs, as well as integration within current practice.

REFERENCES

1. Breast Screening Programme, England - 2016-17 [PAS] https://digital.nhs.uk/data-and-information/publications/statistical/breast-screening-programme/breast-screening-programme-england---2016-17 - accessed October 2018

2. Royal College of Radiologists. The breast imaging and diagnostic workforce in the United Kingdom. Reference: BFCR(16)2. 2016.

3. Nakajima, Y et al. Radiologist supply and workload: international comparison--Working Group of Japanese College of Radiology. Radiation medicine. 2008; 26: 455.

4. Beam C A, et al. Effect of Human Variability on Independent Double Reading in Screening Mammography. Acad Radiol, 1996; 3: 891

5. Domingo L et al. Cross-national comparison of screening mammography accuracy measures in U.S., Norway, and Spain. Eur. Radiol 2016; 26: 2520

6. Philpotts LE. Can computer-aided detection be detrimental to mammographic interpretation? Radiology, 2009; 253: 17.

7. Gilbert FJ et al. Single Reading with Computer-Aided Detection for Screening Mammography. New Eng J Medicine 2008; 359: 1675

8. Taylor P & Potts HWW. Computer aids and human second reading as interventions in screening mammography: Two systematic reviews to compare effects on cancer detection and recall rate. Eur J Cancer, 2008; 44: 798

9. Karssemeijer N et al. Breast Cancer Screening Results 5 Years after Introduction of Digital Mammography in a Population-based Screening Program. Radiology, 2009; 253: 353.

10. Bargalló Xet al. Single reading with computer-aided detection performed by selected radiologists in a breast cancer screening program. Eur J Radiology 2014; 83: 2019.

11. Lehman CD et al. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Int Medicine 2015; 175: 1828.

12. Kohli A and Jha S. Why CAD Failed in Mammography. J Am Coll Radiology : JACR, 2018; 15: 535.

13. Elmore, J G et al. Ten-Year Risk of False Positive Screening Mammograms and Clinical Breast Examinations. New Engl J Med, 1998; 338: 1089

14. Gilbert FJ et al. CADET II: A prospective trial of computer-aided detection (CAD) in the UK Breast Screening Programme. J of Clin Oncol 26(15_suppl): 508–508, 2008.

15. Varma M & Zisserman A 2003 . Statistical Approaches to Material Classification.

16. Krizhevsky A et al. ImageNet classification with deep convolutional neural networks. Communications of the ACM - 05 / 2017

17. Dhungel N et al. Automated Mass Detection from Mammograms using Deep Learning and Random Forest.Int Conf on Dig Image Computing: Techniques and Applications, 2015; pp 1–8.

18. ErtosunMG & Rubin DL. Probabilistic visual search for masses within mammography images using deep learning. IEEE Int Conf on Bioinform and Biomedi 2015; 1310–13155.

19. Sage Bionetworks. The Digital Mammography DREAM Challenge, 2016.

20. DREAM Challenge results. https://www.synapse.org/#!Synapse:syn4224222/wiki/401763 - accessed November 2018.

21. Ribli D Detecting and classifying lesions in mammograms with Deep Learning. Sci Rep, 2018; 8: 4165,

22. Clark K et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. J Dig Imaging. 2013; 2: 1045.

23. Moreira IC et al. INbreast: toward a full-field digital mammographic database. Acad Radiol, 2012; 19: 236

24. Teare P et al. Malignancy Detection on Mammography Using Dual Deep Convolutional Neural Networks and Genetically Discovered False Color Input Enhancement. J Dig Imaging, 2017; 30: 499.

25. Geras K~J, et al. High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks. 2017, arXiv:1703.07047

26. Kim EK et al. Applying Data-Driven Imaging Biomarker in Mammography for Breast Cancer Screening: Preliminary Study Sci Rep, 2018; 8: 2762

27. Rodriguez-Ruiz A et al. Can Radiologists Improve Their Breast Cancer Detection in Mammography When Using a Deep Learning Based Computer System as Decision Support? 14th International Workshop on Breast Imaging (IWBI 2018), June 2018, doi:10.1117/12.2317937.

28. Lehman CD et al. Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation. Radiology, 2018, 180694.

29. Wang X et al. Transfer Deep Learning Mammography Diagnostic Model from Public Datasets to Clinical Practice: a Comparison of Model Performance and Mammography Datasets. 14th IntWorkshop on Breast Imaging (IWBI 2018), June 2018, doi:10.1117/12.2317411.

30. Korkinof D et al. High-Resolution Mammogram Synthesis using Progressive Generative Adversarial Networks. eprint arXiv:1807.03401. 07 2018.

31. Lehman C D et al. National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. Radiology. 2017; 283. 49

32. Miglioretti DL et al. Criteria for Identifying Radiologists With Acceptable Screening Mammography Interpretive Performance on Basis of Multiple Performance Measures. Amer J Roentgenology 2015; 204: 4.

33. Carney PA et al. Identifying Minimally Acceptable Interpretive Performance Criteria for Screening Mammography. Radiology. 2010; 255:

34. Culpan, A.M. Radiographer Involvement in Mammography Image Interpretation: A Survey of United Kingdom Practice. Radiography 2016; 22: 306.

35. Woodard, D. B., et al. "Performance Assessment for Radiologists Interpreting Screening Mammography." Statistics in Medicine, vol. 26, no. 7, 2007, pp. 1532–1551., doi:10.1002/sim.2633.

36. Gisella Gennaro, D. Bernardi, and N. Houssami. Radiation dose with digital breast to-mosynthesis compared to digital mammography: per-view analysis.European Radiology,28(2):573–581, 2 2018.

37.Srinivasan Vedantham, Andrew Karellas, Gopal R. Vijayaraghavan, and Daniel B. Kopans. Digital Breast Tomosynthesis: State of the Art. Radiology, 277(3):663–684, 12 2015.

38. James, J.J. et al. Evaluation of a computer-aided detection (CAD)-enhanced 2D synthetic mammogram: comparison with standard synthetic 2D mammograms and conventional 2D digital mammography. Clinical Radiology , Volume 73 , Issue 10 , 886 - 892.