Histopathology

Histopathology 2018, **72**, 227–238. DOI: 10.1111/his.13333

HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues

Talha Qaiser,[1,*] Abhik Mukherjee,[2,*] Chaitanya Reddy PB,[3] Sai D Munugoti,[3] Vamsi Tallam,[3] Tomi Pitkäaho,[4] Taina Lehtimäki,[4] Thomas Naughton,[4] Matt Berseth,[5] Aníbal Pedraza,[6] Ramakrishnan Mukundan,[7] Matthew Smith,[8] Abhir Bhalerao,[1] Erik Rodner,[9] Marcel Simon,[9] Joachim Denzler,[9] Chao-Hui Huang,[10,11] Gloria Bueno,[6] David Snead,[12] Ian O Ellis,[2] Mohammad Ilyas[2,13] & Nasir Rajpoot[1,12]

[1]*Department of Computer Science, University of Warwick, Coventry, UK,* [2]*Department of Histopathology, Division of Cancer and Stem Cells, School of Medicine, University of Nottingham, Nottingham, UK,* [3]*Department of Electronics and Electrical Engineering, Indian Institute of Technology, Guwahati, India,* [4]*Department of Computer Science, Maynooth University, Maynooth, Ireland,* [5]*NLP Logix LLC, Jacksonville, FL, USA,* [6]*VISILAB, E.T.S.I.I, University of Castilla-La Mancha, Ciudad Real, Spain,* [7]*Department of Computer Science and Software Engineering, University of Canterbury, Canterbury, New Zealand,* [8]*Department of Statistics, University of Warwick, Coventry, UK,* [9]*Computer Vision Group, Friedrich Schiller University of Jena, Jena, Germany,* [10]*MSD International GmbH,* [11]*Singapore Agency for Science, Technology and Research, Singapore,Singapore,* [12]*Department of Pathology, University Hospitals Coventry and Warwickshire, Coventry, UK, and* [13]*Nottingham Molecular Pathology Node, University of Nottingham, Nottingham, UK*

Date of submission 15 May 2017
Accepted for publication 29 July 2017
Published online *Article Accepted* 3 August 2017

Qaiser T, Mukherjee A, Reddy PB C, Munugoti S D, Tallam V, Pitkäaho T, Lehtimäki T, Naughton T, Berseth M, Pedraza A, Mukundan R, Smith M, Bhalerao A, Rodner E, Simon M, Denzler J, Huang C-H, Bueno G, Snead D, Ellis I O, Ilyas M & Rajpoot N

(2018) *Histopathology* **72**, 227–238. https://doi.org/10.1111/his.13333

HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues

Aims: Evaluating expression of the human epidermal growth factor receptor 2 (HER2) by visual examination of immunohistochemistry (IHC) on invasive breast cancer (BCa) is a key part of the diagnostic assessment of BCa due to its recognized importance as a predictive and prognostic marker in clinical practice. However, visual scoring of HER2 is subjective, and consequently prone to interobserver variability. Given the prognostic and therapeutic implications of HER2 scoring, a more objective method is required. In this paper, we report on a recent automated HER2 scoring contest, held in conjunction with the annual PathSoc meeting held in Nottingham in June 2016, aimed at systematically comparing and advancing the state-of-the-art artificial intelligence (AI)-based automated methods for HER2 scoring.

Methods and results: The contest data set comprised digitized whole slide images (WSI) of sections from 86 cases of invasive breast carcinoma stained with both haematoxylin and eosin (H&E) and IHC for HER2. The contesting algorithms predicted scores of the IHC slides automatically for an unseen subset of the data set and the predicted scores were compared with the 'ground truth' (a consensus score from at least two experts). We also report on a simple 'Man versus Machine' contest for the scoring of HER2 and show

Address for correspondence: N Rajpoot and T Qaiser, Department of Computer Science, University of Warwick, UK. e-mails: n.m.rajpoot@warwick.ac.uk; t.qaiser@warwick.ac.uk
*These authors contributed equally to this study.

that the automated methods could beat the pathology experts on this contest data set.
Conclusions: This paper presents a benchmark for comparing the performance of automated algorithms for scoring of HER2. It also demonstrates the enormous potential of automated algorithms in assisting the pathologist with objective IHC scoring.

Keywords: automated HER2 scoring, biomarker quantification, breast cancer, digital pathology, quantitative immunohistochemistry

Introduction

The adoption of image analysis in digital pathology has received significant attention recently due to the availability of digital slide scanners and the increasing importance of tissue-based biomarkers in stratified medicine.[1] Advances in software development and an upwards trend in computational capacity have also caused an upsurge of interest in digital pathology.

Breast cancer (BCa) is the most commonly diagnosed cancer among women, and the second leading cause of death worldwide.[2] According to Cancer Research UK, the risk for women being diagnosed with breast cancer is one in eight in the United Kingdom, and approximately 11 600 women died from breast cancer in 2012.[3] In routine diagnostic practice of BCa, tumour tissue is stained with haematoxylin and eosin (H&E) and then examined under the optical microscope for morphological assessment, including grade. In addition, tissues are stained by immunohistochemistry (IHC) to evaluate biomarker expression for prognostic and predictive purposes. This conventional method of diagnosis by visual examination is considered accurate in most areas, but is known to suffer from inter- and intra-observer variability in some areas, such as diagnosis of atypical hyperplasia and reporting of histological grade.[4–6] Digital pathology offers significant potential for improvement to overcome the subjectivity and improve reproducibility.

The human epidermal growth factor receptor 2 (HER2) gene is amplified in approximately 15–20% of breast cancers.[7] Gene amplification can also be identified through fluorescence *in-situ* hybridization (FISH). Alternatively, as HER2 amplification results in increased protein expression, IHC may be used. Given the technical ease of performing IHC it has become the preferred test, and FISH is usually performed only when the IHC is equivocal. In practice, an expert histopathologist will report a score between 0 and 3+ and cases scoring 0 or 1+ are classified as negative, while cases with a score of 3+ are classed as positive. Cases with score 2+ are classified as equivocal and are assessed further by FISH to test for gene amplification. Examples of the four different HER2 scores (0 to 3+) are shown in Figure 1. A summary of recommended guidelines for HER2 IHC scoring criteria[7] is shown in Table 1.

Historically, up to 20% of the HER2 IHC results may contain inaccuracies[8] due to variations in the technical quality and the subjective nature of scoring. Although adoption of HER2 guidelines and recommendations[7] have served to improve standards in HER2 testing, challenging cases remain, especially with HER2 scores deemed borderline between categories.

Automated IHC scoring of HER2 carries promise to overcome the existing problems in conventional methods. Automated scoring methods are not prone to subjective bias, and can provide precise quantitative analysis which can assist the expert pathologist to reach a reproducible score.

The HER2 scoring contest, documented in this paper, was organized by the University of Warwick, the University of Nottingham and the Academic–Industrial Collaboration for Digital Pathology (AIDPATH) consortium (www.aidpath.eu). It was held in conjunction with the Pathological Society of Great Britain and Ireland meeting in Nottingham (June 2016) to provide a platform for researchers to assess the performance of computer algorithms for automated HER2 scoring on IHC-stained slides. This paper provides an overview of the automated methods for HER2 scoring as presented at the contest and a 'Man versus Machine' comparison of the degree of agreement among histopathologists and the automated methods for HER2 scoring. This may be considered as an initial step towards the development of a reliable computer-assisted diagnosis tool for HER2 scoring of digitized BCa histology slides.

Materials and methods

ETHICS

Ethical approval was by Nottingham Research Ethics Committee 2 (Approval no.: REC 2020313); R&D reference (N) 03HI01.

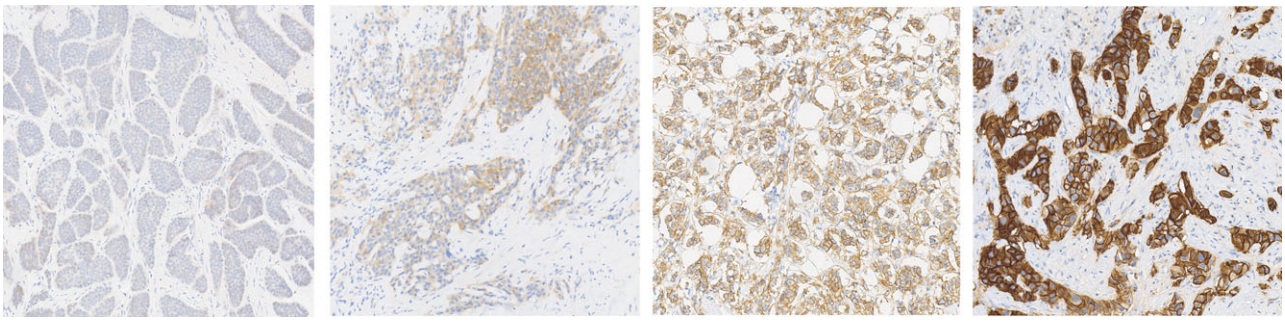


Figure 1. Left to right: examples of regions of interest (800 μm in height and the same in width) from whole slide images (WSIs) scored 0, 1+ (negative), 2+ (equivocal) and 3+ (positive).

Table 1. Recommended automated human epidermal growth factor receptor 2 (HER2) scoring criteria for immunohistochemistry (IHC)-stained breast cancer tissue slides[7]

Score	Cell membrane staining pattern	Staining assessment
0	No membrane staining or incomplete membrane staining in <10% of invasive tumour cells (0+) or faint/barely perceptible or weak incomplete membrane stainaing in 10% of tumour cells (1+)	Negative
1+		Negative
2+	A weak to moderate complete membrane staining is observed in >10% of tumour cells or strong complete membrane staining in ≤10% of tumour cells	Borderline (equivocal)
3+	A strong (intense and uniform) complete membrane staining is observed in >10% of invasive tumour cells	Positive

IMAGE DATA ACQUISITION AND GROUND TRUTH

The histology slides for this contest were scanned on a Hamamatsu NanoZoomer C9600, enabling the image to be viewed from a ×4 to a ×40 magnification, making the process comparable to a clinician's standard microscope. Generally, WSIs are gigapixel images stored in a multiresolution pyramid structure, where the highest resolution is ×40. The contest data set entailed 172 whole slide images (WSI) extracted from 86 cases of invasive breast carcinomas and included both the H&E- and HER2-stained slides. The actual HER2 scoring is normally performed on the IHC-stained slides, while the H&E slides assist the expert pathologist to identify the areas of invasive tumour and discriminate these from areas of *in-situ* disease. Figure 2 shows an example of the two types of WSIs (with a corresponding zoomed-in region of interest) from the contest data set.

The ground truth (GT) was taken from the clinical reports issued on the cases at a tertiary referral centre for breast pathology (Nottingham University Hospitals, NHS Trust). At this centre, each case had been reported or reviewed by at least two specialist consultant histopathologists as part of their routine practice [preliminary reporting and multidisciplinary team (MDT) review]. The centre provides regular internal quality control for HER2 assessment for immunohistochemistry runs and contributes and participates regularly in the UK NEQAS (National External Quality Assessment Scheme) for immunocytochemistry and *in-situ* hybridization (ICC and ISH).

CONTESTANTS

A total of 105 teams from more than 28 countries registered to access the training data set before the end of the registration deadline. By the end of the submission deadline (off-site contest), a total of 18 submissions from 14 teams were received for evaluation. The organizers provided an opportunity to each of the 14 teams for presenting their approach in the contest workshop and six teams chose to present. For the Man versus Machine contest, we received the markings from four pathologists. The contest website was re-opened for new submissions after concluding the workshop. Further details regarding various stages of the contest are described in Data S1 and Table S1.

EVALUATION

The performance of each submitted algorithm was evaluated based on three criteria: (1) agreement points, (2) weighted confidence and (3) combined

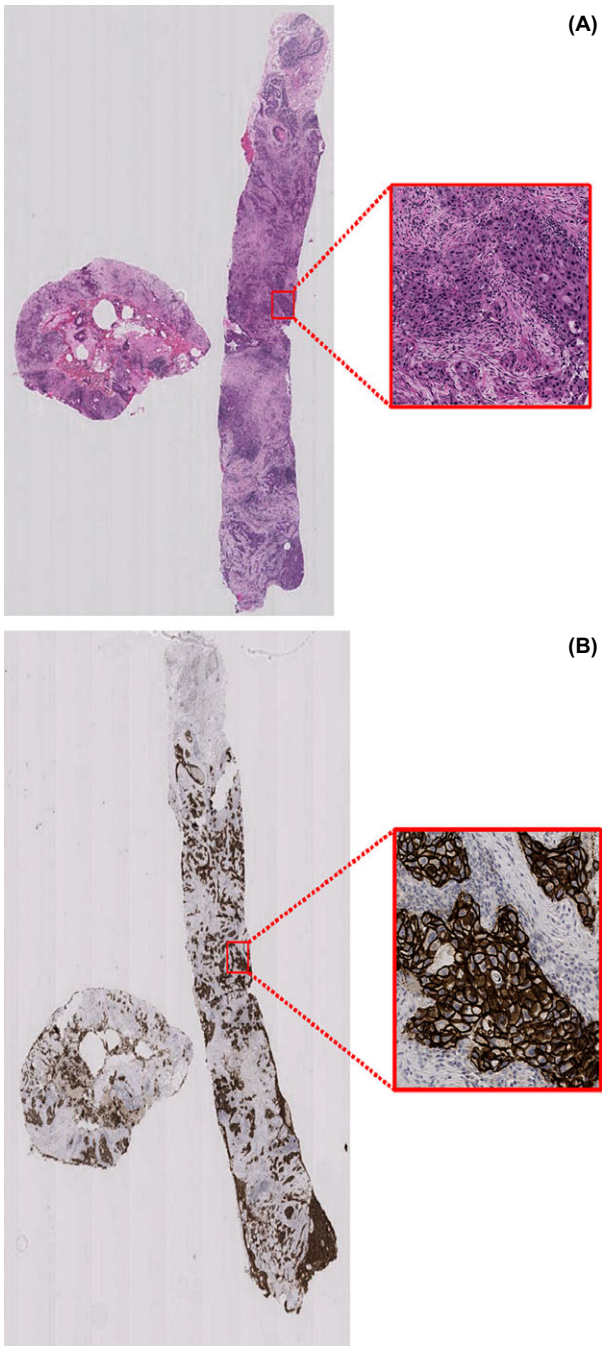


Figure 2. An example whole slide image (WSI) along with a zoomed-in cross-sectional area showing the tumour region (**A**) haematoxylin and eosin (H&E)-stained slide; (**B**) immunohistochemistry (IHC)-stained slide.

points. Each assessment criterion has a separate leader-board.

The evaluation criteria were rationalized according to the clinical significance and implications of HER2 IHC scoring as follows: in everyday clinical practice, for a score of 0 and 1+ no herceptin is offered to the patient; for a 3+ score, herceptin is offered. For an IHC 2+ score, a FISH test is performed; if positive (i.e.) there is evidence of gene amplification and herceptin is offered, while for a negative result it is not offered. The evaluation considers the impact of erroneous classification. For example, a score of 0/1+ being interpreted as 3+ or vice versa is a serious error while a 2+ scored as 0/1+ denies a few patients of valid treatment; a score of 3+ for a 2+ case bypasses the FISH test and may treat a few cases erroneously (which would have been FISH-negative) with toxic drugs while a score of actual 3+ downgraded to a 2+ calls for additional expense of FISH testing, but the end result will probably be the same and hence this should not be regarded as a serious error. These have been summarized in Table 2.

Table 2. (A) Agreement points for predicted calls of ground truth (GT), (B) bonus point criteria, when percentage of cells with complete cell membrane staining (PCMS) lies in certain range of the GT value of the PCMS

	Points for predicted score				
(A)					
Ground truth	Score	0	1+	2+	3+
	0	15	15	10	0
	1+	15	15	10	0
	2+	2.5	2.5	15	5
	3+	0	0	10	15

Ground truth score	Percentage of cells with complete cell membrane staining (PCMS)	
(B)		
0	0	0
1+	1 (PCMS < 3%)	3 (PCMS ± 2)
2+	5 (PCMS ± 5)	2.5 (PCMS ± 10)
3+	5 (PCMS ± 5)	2.5 (PCMS ± 10)

For agreement points, a penalty method was employed whereby each erroneous prediction is penalized with respect to its deviation from the GT, as shown in Table 2A. It can be envisaged that the agreement points may end in a tie, where the accumulative points of two or more teams may be the same. To resolve the tie, a bonus criterion was devised as shown in Table 2B, where the decision was made on the percentage of cells with complete cell membrane staining (PCMS) regardless of the

intensity. The bonus points were <3% introduced for scores 2+ and 3+ as they attain more clinical significance. For the IHC score 1+, 1 bonus point was awarded if there was an accurate prediction of the IHC score and PCMS <3%, while 3 bonus points were awarded if there was an accurate prediction of the IHC score and PCMS >3% but the predicted PCMS value deviated only ±2% from the GT. For the IHC scores 2+ and 3+, 5 bonus points were awarded if there was an accurate prediction of the IHC score and PCMS deviated only ±5% from the GT. Similarly, 2.5 bonus points were awarded for scores 2+ and 3+ if there was an accurate predication of IHC score and PCMS deviated only ±10% from the GT.

The weighted confidence was devised to measure the credence of the predicted score by the submitted algorithm. The criteria to measure the weighted confidence w_c were distinct for both truly and wrongly classified cases. In cases where the predicted HER2 score p_S matched with the GT with higher confidence c, the weighted confidence amplified the confidence value for true prediction, whereas wrong predictions with high confidence were penalized accordingly, as given in equation (1). This type of assessment is important for the development of an interactive diagnostic module. The confidence value may indicate those cases or regions where further examination by the experts may be required before concluding the final HER2 score.

$$w_c = \begin{cases} \frac{2c-c^2}{2} & \text{if } p_s = \text{GT} \\ \frac{-c^2+1}{2} & \text{otherwise} \end{cases} \quad (1)$$

The third assessment criterion is a combination of both agreement points and weighted confidence-based evaluations. The combined points were calculated by taking the product of two assessment criteria for each case.

Results

CONTEST LEADERBOARDS

Comprehensive results comprising all the submissions for automated methods are shown in Table 3. The teams in were ranked with respect to the combined point-based assessment with bonus points. For the off-site contest, the total possible points were 420 (28 cases with a maximum of 15 points each), whereas for weighted confidence the maximum points were 28, 1 for each case. The top three-ranked teams with respect to point based assessments were Team Indus, MUCS-1 and MUCS-2, whereas according to weighted confidence assessment the top-ranked teams were VISILAB, FSUJena and MTB NLP. The combined results rank the top three teams in the following order: VISILAB, FSUJena and Huangch. The performance of top-ranked teams including bonus points and the trend for total points (without the bonus points) can be seen in Figure 3. MUCS-1, MUCS-3, CS_UCCGIP and MTB NLP achieved equal points, but MUCS-1 secured more bonus points, as their PCMS was more accurate compared to remaining counterparts. Similarly, Team VISILAB and Rumrocks resulted in a tie where both teams attained equal points, but the VISILAB method was more precise in predicting PCMS. Comprehensive tables for all three leaderboards are available for download from the contest website.

SUMMARY OF PROPOSED AUTOMATED METHODS

Most of the automated methods (described in Data S2 and Figure S1) applied a supervised patch-based classification approach to solve this problem. The most common pipeline was based on three main components: (1) pre-processing including the methods to identify the regions of interest for patch generation, (2) classification based on handcrafted or neural network learned features and (3) post-processing techniques to aggregate the HER2 score at WSI level and to estimate the PCMS. Deep learning, especially convolutional neural network (CNN)-based approaches, dominated as eight of the top 10 methods were based on CNN. The majority of the CNN architectures [Team Indus, MUCS-(1–3), MTB NLP, VISILAB, RumRocks, FSUJena] were inspired from the state-of-the-art deep neural networks.[9,10]

In pre-processing and patch extraction stage, most of the teams followed the conventional thresholding techniques with a combination of morphological operators. These techniques are computationally less expensive and generally work well, as background regions lack any texture contents in contrast with other tissue components. The MUCS-(1–3), MTB NLP, VISILAB and FSUJena probe the regions of interest manually through calibration or customized methodologies. These methods aimed to pick the best possible regions for training their algorithm, generally without affecting the testing phase. To segment tissue regions, the RumRocks team implemented a deconvolutional neural network (DCNN) and a two-dimensional CNN for selection of patches based on their texture. The Huangch team performed mean filtering and stain normalization using the control tissue

Table 3. A summary of results of all three assessment criteria for the automated human epidermal growth factor receptor 2 (HER2) scoring contest, ordered by the combined points criterion

Team	Affiliation	Points	Points + bonus	Weighted confidence	Combined
VISILAB	Universidad de Castilla-La Mancha	382.5	404.5	23.552	348.041
FSUJena	Computer Vision Group, University of Jena	370	392	23	345
HUANGCH	Bioinformatics Institute, Singapore	377.5	391.5	22.622	335.77
MTB NLP	NLP Logix, LLC	390	405.5	22.937	335.737
VISILAB (density)	Universidad de Castilla-La Mancha	377.5	391	21.878	322.067
Team Indus	Indian Institute of Technology Guwahati	402.5	425	18.451	321.414
UC-CSSE-CGIP group	University of Canterbury, New Zealand	390	395	21.07	316.05
MUCS-3	Computer Science, Maynooth University	390	411	20.434	300.813
HERcules	University of Oxford	360	380	20.572	295.633
MUCS – 2	Computer Science, Maynooth University	385	413	19.51	290.171
Rumrocks	Department of Statistics, University of Warwick	382.5	395	19.649	277.705
TissueGnostics	TissueGnostics GmbH, Austria	365	366	17.78	266.41
Team Indus (Stainsep)	Indian Institute of Technology Guwahati	332.5	345.5	18.451	250.715
MUCS – 1	Computer Science, Maynooth University	390	416	16.765	248.876
HersRockers	Indian Institute of Technology Guwahati	320	330	17.318	223.007
VIP-UGR	University of Granada	305	322.5	15.41	211.748
TartanSight	Computational Biology, CMU	230	230	15.148	159.745
Cancer_Detector	Indian Institute of Technology Kanpur	255	260	12.994	138.962

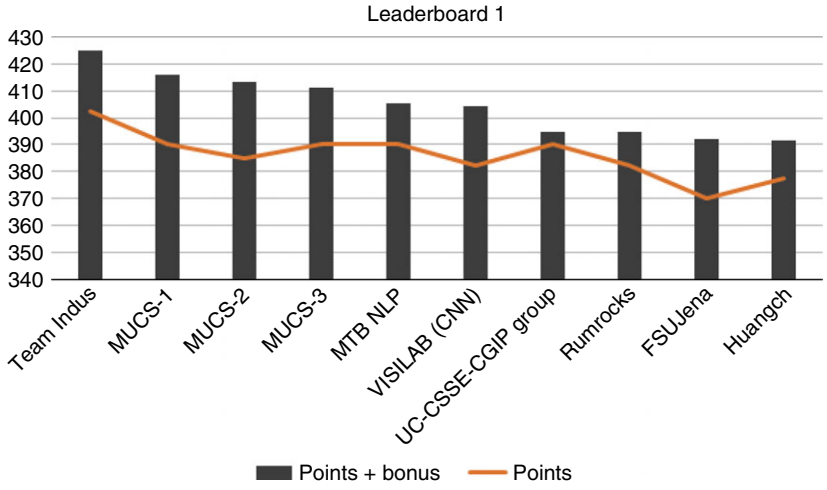


Figure 3. Combined results for top-ranked teams with respect to agreement and bonus points. The trend shows the significance of predicting correctly the percentage of cell membrane.

intensity values to calibrate the stain colour intensity as a pre-processing step.

In the second step, most of the teams (specifically the top 10) employed deep learning approaches, whereas other teams such as CS_UCCGIP and Huangch derived handcrafted characteristic curves and employed standard machine learning approaches. Team Indus used a combination of data-driven and handcrafted features. They incorporated the average control tissue intensity value along with learned

features maps before passing them to the fully connected layers. Some of the top-ranked teams deployed variants of Alexnet[9] and GoogLeNet[10] for predicting the HER2 score. The FSUJena team computed the bilinear features after retrieving activations from convolutional layers of the AlexNet. The derived activations contain the learned feature maps representing a d-dimensional $w \times h$ spatial grid. This approach enables them to perform their analysis on top of the learned features maps from CNN. In combination with standard approaches for data regularization, MTB NLP and RumRocks trained multiple models. The final HER2 score and PCMS was estimated by averaging over all the models. Additionally, a wide range of data augmentation and regularization techniques were employed to overcome the overfitting issues. As in practice, the standard data augmentation techniques such as affine transformations (e.g. rotation, flip, translation), random cropping, blurring and elastic deformations were applied to train the network. MUCS-2, MTB NLP and RumRocks broadly used the data augmentation techniques to assist the network to generalize well on unseen data.

In the final stage of pre-processing and predicting the PCMS, most of the teams employed standard image processing and machine learning approaches on top of the results attained from the last step. A Random Forest classifier was trained by MTB NLP to produce the final class probabilities and to estimate the PCMS. FSUJena simply used the mean tumour cell percentage seen in the training set for a particular class as an estimate. Team Indus used both IHC- and H&E-stained slides to estimate the PCMS by using standard image processing approaches such as contour detection, thresholding and morphological features. All the remaining teams limited their analysis to only IHC-stained images. All the submissions used high-magnification images ($\times 10$ or above), except MUCS and Rumrocks, who used images from low resolution for selection of ROIs.

MAN VERSUS MACHINE EVENT

Organization

One way of evaluating the automated algorithms for IHC (HER2) scoring is to perform comparative analysis of the assessment of expert pathologists and automated methods for a handful of cases compared to the scores for those cases as agreed by at least two consultant breast pathologists (GT). On the day of the contest workshop, we organized an event called Man versus Machine. The main aim of this event was to analyse the performance of automatic methods and to explore the disagreements among conventional and automatic methods. This type of analysis can lead us to a more sophisticated protocol for automatic HER2 scoring and to overcome the inter- and intra-observer agreements that can be found in normal practice.

The analysis between the expert's agreement and the evaluation of the automatic HER2 scoring method was performed with a subset (15 cases) of the off-site test data set. For this event, we set up an online webpage for the pathologists. The webpage enabled the experts to load and navigate (including pan and zoom) through the WSI of those cases. Both IHC-(HER2) and H&E-stained digital images were made available to mimic the conventional scoring environment. On the contest day at Pathological Society meeting 2016 we requested the expert pathologists to score each case by providing the HER2 score, PCMS and a confidence value.

Man versus machine results comparison

Table 4 summarizes the overall evaluation scores achieved by each participant for this event. Each table entry gives the cumulative score for all 15 cases, which indicates the overall performance. The agreement-points-based assessment was used to evaluate the performance for this event. In total, we received four responses from expert pathologists and, as shown in Table 4 we ranked the top six submissions, including the top three automated methods. From submitted responses, three participant pathologists reported themselves as 'consultant pathologist' and one as 'trainee pathologist', and all three marked breast pathology as a subspeciality.

As can be seen in Table 4, one of the automated methods slightly outperformed the top-performing participant pathologist. These results point to the potential significance of automated scoring methods

Table 4. Summary results for the Man versus Machine event. The evaluation was carried out according to the contest criteria as described in the Evaluation section

Rank	Team name	Score	Bonus	Score + bonus
1	Team Indus	220	12.5	232.5
2	Expert 2	210	20.5	230.5
3	VISILAB	212.5	15	227.5
4	MUCS-1	205	20.5	225.5
5	Expert 1	185	10	195
6	Expert 3	180	13	193

and the recent advancements in digital pathology. It is worth mentioning that automated HER2 scoring algorithms submitted in this contest are not ready to deploy in their current form, as they will require extensive validation on a significantly large-scale data set and also a great deal of input from experts to prepare the GT on the larger data set.

Table 5 shows the pooled data for HER2 scoring among the three top-ranked automated methods and the scores from three participant pathologists and comparison with the GT. Table 5 was determined for the 15 cases selected from the off-site contest data set. On the basis of HER2 scores, a 100% agreement with the GT was observed for score 3+ among the participant pathologists and the automated methods. For the scores of 1+ and 2+, there were disparities between the GT and the new scores. In all cases except one, for both man and machine, the error resulted from overcalling the score. Thus, for score 1+, six of nine (67%) were overcalled as 2+ by humans while four of nine (44%) were overcalled by the machine algorithms. For the score of 2+, seven of 15 (46%) were overcalled as 3+ by humans while machines overcalled one of 15 (6%) as 3+ and one of 15 (6%) was undercalled as 1+. Clinically, a score of 2+ is critical, as in routine practice cases of score 2+ are recommended to undergo FISH testing. It is equally important to avoid predicting score 2+ as 1+ or 0 cases, as such erroneous prediction will deny the further assessment of HER2. As can be seen in Table 5, none of the cases with score 2+ was misclassified by the participant pathologists as either 1+ or 0, whereas for one of the cases an automated method wrongly predicted a score of 2+ as 1+.

Most of the incorrect predictions by the participant pathologists were found to be in cases where there was considerable heterogeneity. Two such examples are shown in Figure 4A–D. In tumour cells of HER2 score 2+, a pattern of weak to moderate complete membrane staining is observed whereas for score 3+, an intense (uniform) complete membrane staining is observed. Estimating the complete membrane staining is a difficult and highly subjective process, especially for score 2+ and 3+, as it is extremely difficult to detect subtle differences in the morphological appearance for those cases.

Discussion

A major aim of organizing this contest was to provide a platform for computer scientists and researchers to contribute and to evaluate the performance of their computer algorithms for automated IHC scoring of HER2 in images from BCa tissue slides. Automated scoring can overcome significantly the subjectivity found, due to varying standards adopted by different diagnostics laboratories. There is a current wealth of literature[11,12] using individual platforms (both freely and commercially available) for digital analysis of HER2 in BCa. This, however, was the first comparison of platforms and algorithms, and provides a pilot for independent comparison of computing algorithms for HER2 assessment on a benchmark data set. The contest highlights the wealth of potential carried by artificial intelligence (AI) techniques for the assessment of IHC slides.

The contest 'training data set' was selected deliberately such that it contained a reasonable number of cases from all HER2 scores, bearing in mind the need for the training algorithms to learn features for each score. For the test data set (both off- and on-site), the GT was withheld at the time of image evaluation. Results showed that the automated analysis performed comparably to histopathologists. Many of the algorithms achieved high accuracy – often close to the maximum. Our main objective was to analyse the performance of algorithms based on clinical relevance, and hence the three particular evaluation criteria described above were chosen. It may be possible that other assessment criteria may influence the ranking of comparative results.

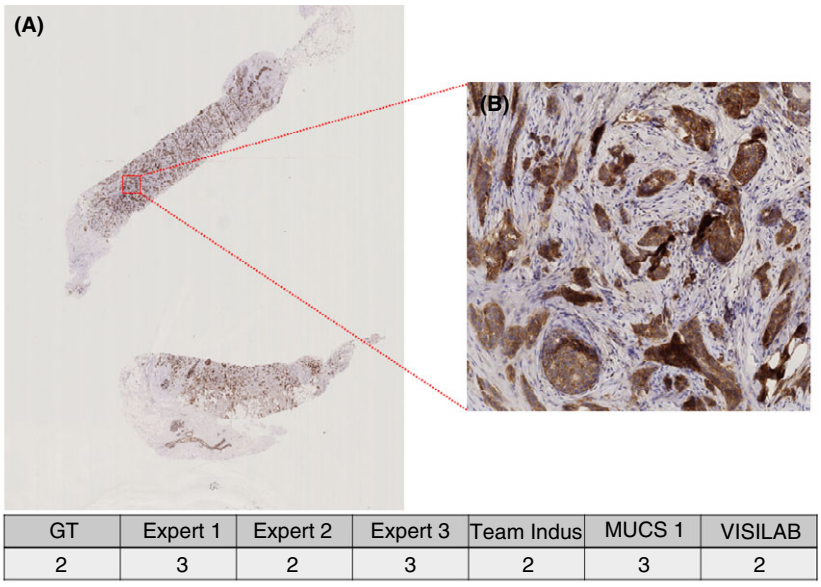
The data from the Man versus Machine comparison showed that, reassuringly, all participants (whether human or computer) identified cases correctly with a GT score of 3+. This means that no one in the category would have been denied treatment. Similarly, for the cases with a score of 0 or 1+, although there was some overcalling, this never exceeded 2+ and thus none would have received treatment without further testing. The most problematic category was, not unexpectedly, cases with a score of 2+ in both human and machine evaluations. If overcalled as 3+, the FISH negative subset would be overtreated. The GT information for the FISH results were not released to the participants, as the contest was aimed only at comparing interpretation of HER2 IHC results. Hence, most of the automated algorithms aimed at predicting the equivocal cases as 2+. Table 5 incorporates the FISH results for all the cases that were marked as 2+ in the test data GT (including the Man versus Machine data set). From Man versus Machine cases (15 in total), a score of 2+ (subsequently FISH negative) was overcalled by the machine as 3+ in only one instance (VISILAB). In contrast, on three

Table 5. Combined matrix for agreement among the three experts and the top three automated methods based on agreement points against the ground truth (GT) scores for 15 cases in the Man versus Machine event. Borderline case 7 was deemed negative and cases 16 and19 were deemed positive for treatment decision (based on the human epidermal growth factor receptor 2:chromosome 17 centromere (HER2:CEP17) amplification ratio for HER2 over-expression: 1.96, 2.1 and 2.07, respectively

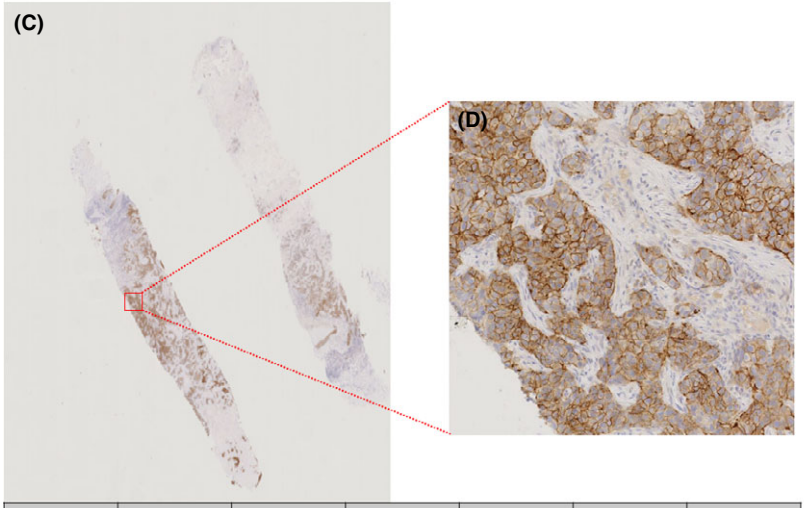
Case	Ground truth	FISH results	Expert 1	Expert 2	Expert 3	Team Indus	Visilab	MUCS-1
1	2+	Negative	3+	2+	2+	2+	2+	2+
2	0	–	0	1+	1+	1+	1+	0
3	3+	–	3+	3+	3+	3+	3+	3+
4	0	–	1+	1+	1+	0	1+	1+
5	1+	–	2+	1+	2+	1+	2+	1+
6	3+	–	3+	3+	3+	3+	3+	3+
7	2+	Borderline amplified	3+	3+	3+	2+	2+	2+
8	2+	Negative	3+	2+	3+	2+	3+	2+
9	3+	–	3+	3+	3+	3+	3+	3+
10	3+	–	3+	3+	3+	3+	3+	3+
11	1+	–	1+	1+	2+	0	1+	1+
12	2+	Positive	2+	2+	3+	2+	2+	2+
13	1+	–	2+	2+	2+	2+	2+	1+
14	2+	Negative	2+	2+	2+	2+	2+	1+
15	0	–	0	1+	0	0	1+	0
16	2+	Borderline amplified	–	–	–	0	1+	2+
17	2+	Negative	–	–	–	2+	2+	2+
18	2+	Positive	–	–	–	2+	1+	2+
19	2+	Borderline amplified	–	–	–	2+	2+	2+
20	1+	–	–	–	–	1+	1+	1+
21	1+	–	–	–	–	1+	1+	2+
22	0	–	–	–	–	1+	0	1+
23	1+	–	–	–	–	0	1+	1+
24	1+	–	–	–	–	0	1+	2+
25	3+	–	–	–	–	3+	3+	3+
26	0	–	–	–	–	1+	0	1+
27	0	–	–	–	–	0	0	1+
28	0	–	–	–	–	0	0	0

FISH, Fluorescence *in-situ* hybridization.

occasions (subsequently FISH-negative) the participant pathologists overcalled the score 2+ as 3+. Moreover, for the remaining test data set (13 cases), in three instances the score of 2+ (subsequently FISH-positive) were predicted erroneously as either 1+ and 0 by the automated algorithms. Overall, the



GT	Expert 1	Expert 2	Expert 3	Team Indus	MUCS 1	VISILAB
2	3	2	3	2	3	2



GT	Expert 1	Expert 2	Expert 3	Indus	MUCS 1	VISILAB
2	2	2	3	2	2	2

Figure 4. Examples showing immunohistochemistry (IHC)-stained whole slide images (WSIs) (**A**,**C**) and zoomed-in cross-sectional area (**B**,**D**) with corresponding human epidermal growth factor receptor 2 (HER2) ground truth (GT) scores marked by expert pathologists and predictions from the top automated methods.

results indicate that further fine-tuning will be required for 2+ cases with AI. While it is encouraging that automated HER2 scoring algorithms may have sufficient potential as a direct comparison to human diagnosis, it is probably worthwhile to reflect that the number of pathologists actually joining the contest was small (only four) and it would have been better to compare the pathologist's assessment of the slides on a reporting microscope rather than a computer for a fairer comparison to real-life practice.

Conventionally, expert pathologists often switch back and forth between the IHC and H&E slides to map the invasive tumour regions for estimating the percentage of complete membrane staining. With the exception of one of the participants (Team Indus), most of the algorithms reported in this paper have avoided the use of H&E slides, although the use of H&E slide for the automatic detection of ductal carcinoma *in situ* (DCIS) regions cannot be ruled out. In addition, the task of predicting the PCMS is extremely subjective, as the expert has to make an estimation on the basis of the physical appearance of the stained invasive tumour region. The semi-automated methods could provide a comprehensive quantitative analysis on the selected region of interest to assist the experts in estimating the PCMS and HER2 score, especially in borderline cases. As HER2 immunoscoring relies not only on intensity but the completeness of membrane

positivity, automated scoring may be helpful as demonstrated by Brügmann *et al.*,[13] who proposed scoring of HER2 based on an algorithm evaluating the cell membrane connectivity.

This study shows that automated IHC scoring algorithms can provide a quantitative assessment of morphological features that can assist in objective computer-assisted diagnosis and predictive modelling of the outcome and survival.[14] We have demonstrated the potential significance of digital imaging and automated tools in histopathology. In the context of breast histopathology, whereby almost all the invasive tumour cases are considered for HER2 testing, an automated or semi-automated scoring method has potential for deployment in routine practice. Despite all these advances, several challenges remain for the AI algorithms to be optimized and become part of routine diagnosis. It is worth noting that serious optimization will be needed for automated methods while processing a whole-slide image. Some methods required more than 3 h per case which, in the 'real world' of diagnostic service delivery, is not feasible. Another limitation of this contest was that the image data were collected from a single site using a single scanner. A potential extension would be to collect data from multiple pathology laboratories with HER2 scores marked by different experts and images scanned using a variety of different machines. This would also test the differences inherent in staining quality that may affect such procedures. Such enhancements could overcome significantly the overfitting to one particular data set that may occur in the automated scoring methods. In moving across systems other laboratories, for example, have acknowledged the challenges in reaching the optimum Aperio algorithm parameters to provide results that were equivalent to those of the 'automated cellular imaging system' (ACIS) or 'cell analysis system' (CAS 200) quantitation systems,[15] which are fully automated environments for detecting cells based on intensity characteristics and handcrafted features found in IHC-stained images. Therefore, there is a need to learn throughout comparative systems, for which the current study provided a valid starting-point. Also, the study highlights the need for dialogue between histopathologists and informaticians to understand the correct identification of tissue compartments relevant for assessment, correct morphology (normal versus *in-situ* versus invasive) and stromal versus tumour stain. Algorithms will also need to be trained to the natural acceptable variation in staining hues and intensities (intra- and interlaboratories) to work effectively during routine practice.

All cases with score 2+ are recommended routinely for further FISH testing to validate HER2 overexpression at the gene level. It would be an added advantage if the automated methods could be trained with FISH GT to predict the final outcome, and the potential for automated algorithms in calling the actual final HER2 status with reproducible accuracy could be demonstrated. For this, a larger series with 2+ cases alone with FISH data would need to be tested. Indeed, there have been other promising studies that indicate that automated image analysis for HER2 instead of manual assessment may reduce the need for supplementary FISH testing by up to 68%.[16] In a diagnostic setting, this would reduce costs and turnaround time significantly. During the last decade, IHC staining has become ubiquitous in pathology laboratories globally and the role of IHC evaluation in a high-throughput setting becomes key for IHC-based companion diagnostics. Other possible extensions of digital pathology could be to automate the overexpression of the programmed death 1 (PD-1) receptor and its ligand (PD-L1) to evaluate anaplastic lymphoma kinase (ALK) protein and proto-oncogene tyrosine-protein kinase ROS1 in lung cancers.[17] The AI-based algorithms would be more effective if IHC staining and scoring methods were treated as a composite assay.[18,19] The varying staining protocols and scoring parameters may restrain the effectiveness of AI-based automated scoring algorithms, including the HER2 scoring, but with sufficiently variable data from different centres AI algorithms could be trained to overcome that problem.

This contest provides a baseline for computer science and computational pathology researchers for automated/semi-automated scoring and computer-assisted diagnosis (CAD) tools to assist the pathologists in daily routine analysis. The contest is now over but the registration and the web-portal will remain open for future participants to make novel contributions to automated HER2 scoring.

Acknowledgements

The first author (T.Q.) acknowledges the financial support provided by the University Hospital Coventry Warwickshire (UHCW) and the Department of Computer Science at Warwick. The VISILAB team (A.P. and G.B.) and UNOTT (M.I. and A.M.) acknowledge financial support from the European Project AIDPATH (no.: 612471); http://aidpath.eu/. The MUCS team wishes to acknowledge John McDonald and Ronan Reilly for their valuable contributions to the research,

and acknowledge financial support from Science Foundation Ireland (SFI) under grant no. 13/CDA/2224 and an Irish Research Council (IRC) Post Graduate Scholarship. Co-first author Dr. Mukherjee would also like to thank the NIHR and Pathological Society of Great Britain and Ireland for support. We are also grateful to Dr. Nicholas Trahearn for his input in deriving the weighted confidence evaluation measure.

Conflicts of interest

None.

References

1. Hamilton PW, Bankhead P, Wang Y *et al.* Digital pathology and image analysis in tissue biomarker research. *Methods* 2014; **70**; 59–73.
2. Ma J, Jemal A. Breast cancer statistics. In Ahmed A ed. *Breast cancer metastasis and drug resistance*. New York, NY: Springer New York, 2013; 1–18.
3. Breast Cancer Statistics, Cancer Research UK. Available at: http://www.cancerresearchuk.org/cancer-info/cancerstats/types/breast/ (accessed 12/09/2017).
4. Smits AJJ, Kummer JA, de Bruin PC *et al.* The estimation of tumor cell percentage for molecular testing by pathologists is not accurate. *Mod. Pathol.* 2014; **27**; 168–174.
5. Viray H, Li K, Long TA *et al.* A prospective, multi-institutional diagnostic trial to determine pathologist accuracy in estimation of percentage of malignant cells. *Arch. Pathol. Lab. Med.* 2013; **137**; 1545–1549.
6. Rakha EA, Bennett RL, Coleman D *et al.* Review of the national external quality assessment (EQA) scheme for breast pathology in the UK. *J. Clin. Pathol.* 2017; **70**; 51–57.
7. Rakha EA, Pinder SE, Bartlett JMS *et al.* Updated UK recommendations for HER2 assessment in breast cancer. *J. Clin. Pathol.* 2015; **68**; 93–99.
8. Wolff AC, Hammond MEH, Schwartz JN *et al.* Reply to Vang Nielsen, *et al.* and to Raji. *J. Clin. Oncol.* 2007; **25**; 4021–4023.
9. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In Pereira F, Burges CJC, Bottou L *et al.* eds. *Advances in neural information processing systems 25*. Red Hook, NY: Curran Associates Inc, 2012; 1097–1105.
10. Szegedy C, Liu W, Jia Y *et al.* Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, Massachusetts: 2015, 1–9.
11. Gavrielides MA, Conway C, O'Flaherty N *et al.* Observer performance in the use of digital and optical microscopy for the interpretation of tissue-based biomarkers. *Anal. Cell. Pathol.* 2014; **2014**; 1–10.
12. Tuominen VJ, Tolonen TT, Isola J. ImmunoMembrane: a publicly available web application for digital image analysis of HER2 immunohistochemistry. *Histopathology* 2012; **60**; 758–767.
13. Brügmann A, Eld M, Lelkaitis G *et al.* Digital image analysis of membrane connectivity is a robust measure of HER2 immunostains. *Breast Cancer Res. Treat.* 2012; **132**; 41–49.
14. Chen J-M, Qu A-P, Wang L-W *et al.* New breast cancer prognostic factors identified by computer-aided image analysis of HE stained histopathology images. *Sci. Rep.* 2015; **5**; 10690.
15. Farris AB, Cohen C, Rogers TE *et al.* Whole Slide imaging for analytical anatomic pathology and telepathology: practical applications today, promises, and perils. *Arch. Pathol. Lab. Med.* 2017; **141**; 542–540.
16. Holten-Rossing H, Møller Talman M-L, Kristensson M *et al.* Optimizing HER2 assessment in breast cancer: application of automated image analysis. *Breast Cancer Res. Treat.* 2015; **152**; 367–375.
17. Shtivelman E, Hensing T, Simon GR *et al.* Molecular pathways and therapeutic targets in lung cancer. *Oncotarget* 2014; **5**; 1392.
18. Taylor CR. Predictive biomarkers and companion diagnostics. The future of immunohistochemistry –'in situ proteomics', or just a 'stain'? *Appl. Immunohistochem. Mol. Morphol.* 2014; **22**; 555–561.
19. Ilie M, Hofman V, Dietel M *et al.* Assessment of the PD-L1 status by immunohistochemistry: challenges and perspectives for therapeutic strategies in lung cancer patients. *Virchows Arch. Int. J. Pathol.* 2016; **468**; 511–525.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. Contest format.

Table S1. The ground truth score for 52 cases from the training dataset with percentage of cells with complete membrane staining. The boderline case 63 was deemed negative and the amplification ratio for Her2 over-expression was 1.92.

Data S2. Description of automated methods.

Figure S1. Characteristics curves and the corresponding Her2 score. The *x*-axis denotes range of the saturation value whereas *y*-axis denotes the calculated percentage from saturation limits. The predicted Her2 scores are also shown for each curve.