Bioinformatics, 2018, 1–7
doi: 10.1093/bioinformatics/btx838
Advance Access Publication Date: 23 December 2017
Original Paper

Bioimage informatics

A benchmark for comparing precision medicine methods in thyroid cancer diagnosis using tissue microarrays

Ching-Wei Wang[1,2,*], Yu-Ching Lee[2,3], Evelyne Calista[1,2], Fan Zhou[4], Hongtu Zhu[4,5], Ryohei Suzuki[6,7], Daisuke Komura[6], Shumpei Ishikawa[6] and Shih-Ping Cheng[8]

[1]Graduate Institute of Biomedical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, [2]NTUST Center of Computer Vision and Medical Imaging, Taipei, Taiwan, [3]Graduate Institute of Applied Science and Technology, National Taiwan University of Science and Technology, Taipei, Taiwan, [4]Department of Biostatistics, University of North Carolina at Chapel Hill, NC, USA, [5]Department of Biostatistics, University of Texas, MD Anderson Cancer Center, TX, USA, [6]Department of Genomic Pathology, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan, [7]Department of Physics, The University of Tokyo, Tokyo, Japan and [8]Department of Surgery, Mackay Memorial Hospital, Taipei, Taiwan

*To whom correspondence should be addressed.

Associate Editor: Robert Murphy

Received on July 10, 2017; revised on November 22, 2017; editorial decision on December 17, 2017; accepted on December 22, 2017

Abstract

Motivation: The aim of precision medicine is to harness new knowledge and technology to optimize the timing and targeting of interventions for maximal therapeutic benefit. This study explores the possibility of building AI models without precise pixel-level annotation in prediction of the tumor size, extrathyroidal extension, lymph node metastasis, cancer stage and BRAF mutation in thyroid cancer diagnosis, providing the patients' background information, histopathological and immunohistochemical tissue images.

Results: A novel framework for objective evaluation of automatic patient diagnosis algorithms has been established under the auspices of the IEEE International Symposium on Biomedical Imaging 2017— A Grand Challenge for Tissue Microarray Analysis in Thyroid Cancer Diagnosis. Here, we present the datasets, methods and results of the challenge and lay down the principles for future uses of this benchmark. The main contributions of the challenge include the creation of the data repository of tissue microarrays; the creation of the clinical diagnosis classification data repository of thyroid cancer; and the definition of objective quantitative evaluation for comparison and ranking of the algorithms. With this benchmark, three automatic methods for predictions of the five clinical outcomes have been compared, and detailed quantitative evaluation results are presented in this paper. Based on the quantitative evaluation results, we believe automatic patient diagnosis is still a challenging and unsolved problem.

Availability and implementation: The datasets and the evaluation software will be made available to the research community, further encouraging future developments in this field. (http://www-o.ntust.edu.tw/cvmi/ISBI2017/).

Contact: cweiwang@mail.ntust.edu.tw

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

With the recent advent of automated microscopy scanners, dramatic increases in computational power and improvements in image analysis algorithms, digitized tissue histopathology has now become amenable to the application of computerized image analysis and machine learning techniques. The ability to extract pathological features from a patient's tissue sample objectively and consistently can be driven by AI and machine learning. The goal of this study is to investigate the possibility of building automated methods for prediction of clinical diagnosis parameters based on the patients' background information, histopathological and immunohistochemical tissue images. In this study, the thyroid cancer samples are used, which is the most common malignancy of the endocrine system, and the incidence of thyroid cancer is around 12/100 000 people per year in the United states (Kasper *et al.*, 2015).

Tissue microarray (TMA) is an effective tool for high throughput molecular analysis to help identify new diagnostic and prognostic markers and targets in human cancers (Avninder *et al.*, 2008; Chen and Foran, 2016; Jawhar, 2009; Simon *et al.*, 2010; Voduc *et al*, 2008; Wang *et al.*, 2011; Zhang *et al.*, 2009). The technique allows rapid visualization of molecular targets in thousands of tissue specimens at a time and facilitates rapid translation of molecular discoveries to clinical applications. TMAs have applied immunohistochemistry (IHC) for *in-situ* protein expression analysis in tissue samples. IHC is a typical method for studying archived tissues by staining the protein of interest and scoring the staining intensity using visual examination. Quantitative IHC techniques have often yielded clinically important information regarding patient diagnosis, prognosis or both; the scores are used to quantify protein expression, stratifying patients and further identifying effective biomarkers, which may be important for therapeutic purposes.

The aim of this study is to explore the possibility of building automated methods for prediction of clinical diagnosis results, including the tumor sizes, the extrathyroidal extension status, the lymph node metastasis status, TNM stage and BRAF mutation presence, based on the patients' background information, histopathological and immunohistochemical tissue images. Studies have shown that thyroid cancer with BRAF mutation may be associated with the resistance to radioactive iodine treatment, a higher risk of recurrence and possibly increased cancer-related mortality (Wang *et al.*, 2014 b). The BRAF gene belongs to a class of genes known as oncogenes, and when mutated, oncogenes have the potential to cause normal cells to become cancerous. As BRAF mutation is the most common genetic alterations found in the thyroid cancer, the BRAF IHC images are utilized in this study in addition to the hematoxylin-eosin (H&E) tissue images and patients' background information. Although in recent years, there have been various medical image analysis benchmark frameworks (https://grand-challenge.org/All_Challenges/), most of the benchmarks focus on image segmentation, and only a few deal with classification of cancer metastasesor subtypes, such as CAMELYON17 https://camelyon17.grand-challenge.org/and MICCAI 2017 Computational Precision Medicine http://miccai.cloudapp.net/competitions/56. To the authors' best knowledge, this is the first benchmark for investigation of AI with respect to precision medicine directly linked to the five clinically diagnosis parameters, including tumor size, extrathyroidal extension status, metastasis status, cancer stage and BRAF mutation.

This paper presents the evaluation and comparison of a representative selection of current methods presented during the Grand Challenge for Tissue Microarray Analysis in Thyroid Cancer Diagnosis held in conjunction and with the support by IEEE ISBI (2017). The outline of the paper is organized as follows. In Section 2, the challenge aims, participants, datasets and evaluation approaches are described. Quantitative evaluation results are presented in Section 4. Discussions on the advantages and drawbacks of individual methods are given in Section 5. Finally, conclusions are given in Section 6.

2 Materials and Evaluation Methods: A grand challenge for tissue microarray analysis in thyroid cancer diagnosis

2.1 Organization

The challenge was open to teams from academia and industries and held in conjunction and with the support of IEEE ISBI 2017. The definitions of the five clinical outcomes to predict are described as follows:

i. Size: the greatest dimension in *cm* of the primary tumor.
ii. Extrathyroidal extension E: the extent of the primary tumor. $E = 0$: thyroid cancer may be confined to the thyroid gland; $E = 1$: the patient has minimal extra-thyroid extension (e.g. extension to sternothyroid muscle or perithyroid soft tissues); $E = 2$: the cancer presents as an advanced disease, extending beyond the thyroid capsule to invade subcutaneous soft tissues, larynx, trachea, esophagus, or recurrent laryngeal nerve. A higher value implies more aggressive disease with local invasion.
iii. Lymph node metastasis N: the regional lymph node status. $N = 0$: there is no regional lymph node involvement (no cancer found in the lymph nodes); $N = 1.1$: when cancer is found in the lymph nodes taken from the level VI central compartment (pretracheal, paratracheal, and prelaryngeal/Delphian lymph nodes) of the neck, it is regarded as an N1a disease; $N = 1.2$: when cancer is found in the lymph nodes within the lateral cervical basins (levels I, II, III, IV or V) or retropharyngeal or superior mediastinal lymph nodes (level VII), it is regarded as an N1b disease. It is more difficult to control cancer which has spread from the primary tumor to the nearby lymph nodes. A higher value suggests the tumor spreads farther.
iv. TNM stage S: the tumor-node-metastasis (TNM) cancer staging developed and maintained by the American Joint Committee on Cancer (AJCC) and the Union for International Cancer Control (Edge *et al.*, 2010). The TNM stage plays an important role both in treatment planning and prognosis implications. $S = 1, 2, 3, 4$ where a higher number denotes more advanced disease. Specifically, patients with a more advanced disease have a higher possibility of disease recurrence after treatment and have a higher risk of mortality (died of the disease).
v. BRAF: the presence or absence of BRAF mutation. In this study, the BRAF mutation status is determined by the genome sequencing.

2.2 Participants in the challenge

A total of 54 teams from 18 countries registered for the 2017 IEEE ISBI grand challenge. Participating teams are from the academia and industry, and many groups are from deep learning communities. In evaluation, the test data are available to the participants, but each team is limited to one test run. However, as these are challenging tasks, involving tissue image pattern analysis, quantitative biomarker analysis, big data analysis and machine learning, only four teams successfully provided prediction results on all five parameters, and one team decided not to publish their results. It is understandable that this challenge contains five difficult prediction tasks, and teams with lower performance are reluctant to provide full submission results or be included in the challenge paper. Detailed information of the three remaining approaches is described in the Supplementary Material . The first two are deep learning techniques, and the third one is based on ensemble machine learning approaches. It is interested that three methods are all in a format of committees of machine learning models.

- Zhou and Zhu, TMA-D^{2}LM: Tissue Microarray Analysis via A Deep Dictionary Learning Method (USA).
- Suzuki *et al.*, Hybrid Prediction Model for Thyroid Cancer Diagnosis (Japan).
- Wang *et al.* Ensemble Machine Learning Based Approaches for Thyroid Cancer Diagnosis (Taiwan).

2.3 Description of datasets

Patients who had surgery for thyroid cancer at Mackay Memorial Hospital between January 2001 and May 2012 were de-identified and randomly selected for TMA construction. The following patients were excluded: those with medullary thyroid cancer, those aged < 20 years, those with a tumor size of < 8 mm, and those with fewer than three tumor-containing paraffin blocks available. The construction of thyroid cancer TMAs was approved by the institutional review board of Mackay Memorial Hospital (12MMHIS149) and has been described in our previous study (Wang *et al.*, 2014b). The diagnosis of thyroid cancer was reviewed and confirmed by an experienced endocrine pathologist. Thyroid carcinomas were classified according to the last World Health Organization classification of endocrine tumors (DeLellis *et al.*, 2004). The primary tumor was staged according to the seventh edition of the AJCC staging system (Edge *et al.*, 2010). A total of 154 lesions consisting of 128 papillary carcinomas, 17 follicular carcinomas and 9 poorly differentiated or anaplastic carcinomas were initially included.

For each patient, three separate cores were obtained from the viable tumor region and additional one control core from the adjacent normal thyroid tissue. One set of TMA slides were subjected to standard H&E staining. The other set of TMA slides were submitted to immunohistochemical analysis of BRAF V600E mutation. For IHC staining, 5-μm sections were subject to deparaffinization, rehydration and antigen retrieval. The slides were incubated with BRAF V600E-specific clone VE1 antibody (Spring Bioscience, Pleasanton, CA) at 4°C overnight. Positive and negative control slides (without primary or secondary antibodies) were included in each procedure. Molecular validation of BRAF c.1799 T >A mutation was performed, and DNA was isolated from formalin-fixed paraffin embedded sections and analyzed using Sanger sequencing. Initially, there were 14 H&E TMAs and 14 BRAF V600E IHC TMAs collected from 154 patients. Figure 1 shows a H&E TMA and a IHC TMA constructed in this study. Due to poor image quality caused by the production of TMAs, data of some patients were excluded. In total, 14 H&E TMAs with 550 H&E tissue cores from 153 patients and 13 BRAF IHC TMAs with 480 IHC tissue cores from 140 patients were utilized.

As the super high resolution TMA image is very large with the resolution 25412 × 69180 pixels and file size 1.68 GB per TMA image to simpllify the challenge task, instead of processing gigapixel/terapixel TMA images directly, high resolution images of individual tissue cores were extracted from the gigapixel/terapixel TMA slide for analysis using TissueFAXS Viewer 4.2 (TissueGnostics®Gmbh, 2014); the image resolution of individual tissue core ranges from 9755 × 7808 pixels to 11730 × 11712 pixels (around 70 MB to 100 MB per image) and prepared the clinical data information for each patient. In data production process, some data artifacts such as image stitching artifacts and imaging focus artifacts produced by the digital scanner and software as shown in Figure 2 may occur. Although these artifacts do not influence human diagnosis, the methods will need to deal with these artifacts either by manually produced masks or by automated segmentation approaches.

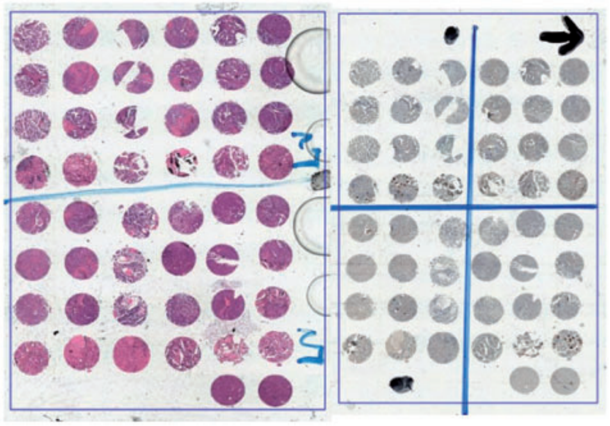

Fig. 1. H&E and IHC TMA images

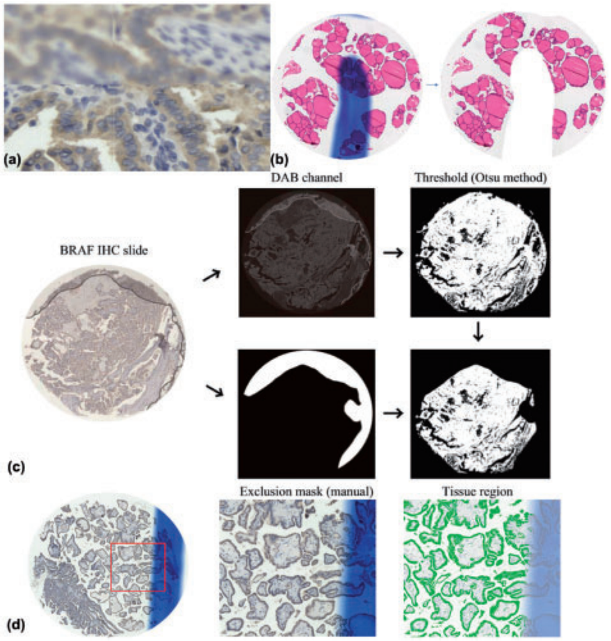

Fig. 2. Image Artifacts. (**a**) Imaging focus artifacts and image stitching artifacts generated by the digital scanner and the associated software; (**b**) blue marking produced by the pathologists. In Zhou and Zhu's method, they manually remove the imaging artifacts in the data preprocessing step; (**c**) Staining artifacts also often occur in data production. In Suzuki *et al.*'s method, artifacts are also removed by manually produced masks. (**d**) In wang *et al.*'s method, an automated segmentation method is built to locate tissues of interests for further processing. The tissues of interests are highlighted in green (Color version of this figure is available at *Bioinformatics online.*)

The patients' background information are provided for training and testing purposes, including the age, sex, hashimoto, body weight, body height, body mass index and the cancer type (see Table 1). The information about the data distributions in training

Table 1. Patient background information

Age	Age
Sex	1 = Male; 2 = Female
Hashimoto	0 = The patient does not have Hashimoto's thyroiditis 1 = The patient has Hashimoto's thyroiditis
BW	Body weight (kg)
BH	Body height (m)
BMI	$\frac{\mathrm{BW(kg)}}{\mathrm{BH^2(m^2)}}$
Cancer subtype	11 = Papillary cancer, classical 12 = Papillary cancer, follicular variant 13 = Papillary cancer, solid variant 25 = Follicular cancer 26 = Follicular cancer, minimally invasive 30 = Poorly differentiated cancer 40 = Anaplastic cancer

and testing with respect to the number of TMAS, patients and tissue cores and the data distributions with respect to the cancer type, sex and hashimoto status are presented in the Supplementary Material.

2.4 Quantitative evaluation approaches

For quantitative evaluation, the tumor size is regarded as numerical data, and the others, including the extension, the lymph node metastasis, TNM stage and BRAF mutation status are treated as categorical data. Both types of data are evaluated using the mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE) and root relative-squared error (RRSE). Given N pairs of two variables y_i and $\widehat{y}_i$ where $i = 1 \ldots N$, y_i represents the actual value from the referenced standard, and $\widehat{y}_i$ represents the prediction outcomes, the four measurements are formulated as follows

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \widehat{y}_i| \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \widehat{y}_i)^2} \tag{2}$$

$$\mathrm{RAE} = \frac{\sum_{i=1}^{N} |y_i - \widehat{y}_i|}{\sum_{i=1}^{N} |y_i - \bar{y}_i|} \tag{3}$$

where $\bar{y}_i = \frac{1}{N}\sum_{i=1}^{N} y_i$.

$$\mathrm{RRSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \widehat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2}} \tag{4}$$

In addition, for prediction of the numerical data, i.e. tumor size, all presented methods are evaluated using the correlation coefficient (CC)

$$\mathrm{CC} = \frac{\sum_{i=1}^{N} (y_i - \bar{y}_i)(\widehat{y}_i - \bar{\widehat{y}}_i)}{\sqrt{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2 \sum_{i=1}^{N} (\widehat{y}_i - \bar{\widehat{y}}_i)^2}} \tag{5}$$

where $\bar{\widehat{y}}_i = \frac{1}{N}\sum_{i=1}^{N} \widehat{y}_i$.

On the other hand, for prediction of the categorical data, four additional measurements are utilized, including Kappa (κ), recall (R), precision (P) and F-measure (F)

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \tag{6}$$

where p_o is the relative observed agreement among raters (identical to accuracy), and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

where TP and FN represent the number of true positive and false negative cases.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{8}$$

where FP stands for the number of the false positive cases.

For multiple classification problems, we adopt the weighted average f-measure using WEKA's definition (Frank *et al.*, 2016)

$$F = \frac{\sum (F_a \times \#a)}{\sum \#a} \tag{9}$$

where F_a represents the F-score for the class a; $\#a$ represents the number of instances of class a; $\sum \#a$ represents the total number of instances.

3 AI methods

This section summarizes the methods that were successfully submitted full prediction results to the challenge. Due to limited number of paper length, detailed information of the methods are described in the Supplementary Material.

3.1 Zhou and Zhu (TMA-D^{2}LM): tissue microarray analysis via a deep dictionary learning method

An TMA analysis model (TMA-D^{2}LM) is built, consisting of three steps: (i) pre-processing and segmentation, (ii) feature extraction and (iii) predictive model building. The pre-processing step is used to remove the outlying tissues with staining ingredient, dust or cracked glass. Next, a deep dictionary learning model consisting of five projection layers, which will deeply train the dictionary learning model and learn internal factorization compared to the single-layer model, is built to perform the feature extraction. The deep dictionary learning model is based on a multi-layer operation framework (Trigeorgis *et al.*, 2017) to generate a low dimensional representation of each individual image. The purpose of this framework is to find the most discriminative patch from each image and build a low dimensional representation of the selected patch to denote the features of the 'mother' images. For prediction, Zhou and Zhu use XGBoost (Chen and Guestrin, 2016), in which the Gradient Boosting produces a prediction model in the form of an ensemble of weak prediction models. The extracted image features along with the seven demographic covariates are used as predictors to predict the five clinical parameters of interest. The prediction framework is built with four layers. Specifically, the bottom layer contains the image features extracted from BRAF and H&E TMA, and then they are combined with the demographic covariates to predict BRAF and Extension, respectively. Then, they use the demographic covariates, BRAF and extension to predict size and N, respectively. Finally, size, N, BRAF and extension are combined to predict the cancer stage. The reason to do the 'indirect prediction' is that after computing the partial correlation between all these five clinical parameters, it is found that BRAF mutation and extension is highly correlated to N, S and tumor size. On the other hand, BRAF mutation and extension are proved to be key factors indicating the cancer level. Even though BRAF mutation and extension are not perfectly predicted, according to the experiment, they generally have a linear relationship with the true annotations. Therefore, including them as predictors may still help increase the prediction accuracy of tumor size, N and especially cancer stage.

3.2 Suzuki *et al.*: hybrid prediction model for thyroid cancer diagnosis

Suzuki *et al.* built a hybrid prediction model for thyroid cancer diagnosis. They separate the prediction model into two sub-modules. The first module is dedicated to predict BRAF mutation status using only IHC slide as input, and the second module predicts all the other clinical diagnoses using the rest of input data, namely H&E slide image, clinical features such as age and sex, as well as the BRAF mutation prediction from the first module. The two modules employ different machine learning approaches for building the individual prediction models reflecting the nature of tasks. Suzuki *et al.* use a deep convolutional network (convnet)-based approach for building the mini patch-level discriminative model for BRAF mutation status, respecting the promising performance of convnets in the recent literature in image recognition and competitions in the field of medical imaging (Szegedy *et al.*, 2015; Ronneberger *et al.*, 2015; Wang *et al.*, 2016). A novel network architecture is built to take a set of overlapping image patches with different magnification levels as the input for capturing the image features of cancer slides in diverse biological scales from individual cells to tissue structures. In addition, they employ additional ad-hoc techniques reflecting pathologists' observations for preparing training datasets and deriving the final decision. Moreover, Suzuki *et al.* hypothesized that the nuclear features including size, shape and texture of HE stained slides of thyroid cancer could be useful for the prediction, as some studies demonstrated that aneuploidy correlates to aggressiveness in papillary thyroid carcinoma (Sturgis *et al.*, 1999) and aneuploidy could affect size and texture of nuclei where abnormal quantities of DNA are contained. In order to build predictive models for other clinical diagnoses, Suzuki *et al.* use the image features of nucleus in HE TMA image, clinical features and BRAF mutation. Since these features include categorical, discrete and continuous variables, Suzuki *et al.* use gradient boosting trees (Chen and Guestrin, 2016) for the prediction, which are known to be effective and powerful in such situation. Hyperparameter optimization of the prediction model was efficiently performed using Bayesian optimization technique.

3.3 Wang *et al.*: ensemble machine learning based approaches for thyroid cancer diagnosis

Using IHC, proteins can be directly visualized by antibodies in their natural cellular localization, and the ability of IHC to quantify a potential biomarker provides the opportunity to study the relationship between the biomarker and chemosensitivity in tumour subgroups, enabling hypothesis generation for additional translational research (Wang, 2013). As BRAF mutation is the most common genetic alterations found in the thyroid cancer, and in the previous study, validation using Sanger sequencing supports the use of IHC for the detection of BRAF V600E protein with reliable accuracy, sensitivity, and specificity as compared with the PCR-based methods (Cheng *et al.*, 2014), therefore in Wang *et al.*'s method, BRAF expression patterns are extracted from IHC images both for segmentation of tissue of interests (as potential tumor regions) and for building machine learning models in prediction of clinical diagnosis parameters. An automated segmentation and quantification method is built and applied to the IHC images not only for measuring the BRAF expression levels but also for localization of tissues of interests. Next, various ensemble machine learning models are trained based on the IHC quantification scores and patient's background information. For segmentation of tissue of interests for further machine learning, color deconvolution (Ruifrok and Johnston, 2001) is first applied to extract independent haematoxylin and DAB/BRAF stain contributions, which has been demonstrated to be effective in tissue image analysis in various studies (Wang, 2013; Wang and Chen, 2013; Wang *et al.*, 2014a). Next, tissue of interests (ROI) are obtained by a clustering process using Otsu's thresholding method (Otsu, 1979), and the background cluster and foreground stain cluster are automatically separated by selecting an optimal local threshold t with the overlap of the background distribution and foreground stain distribution minimized.

Table 2. Evaluation on predictions of BRAF, stage, extension, N and size

BRAF	Zhou and Zhu	Suzuki *et al.*	Wang *et al.*
κ statistic	0.26	0.56	**0.86**
Precision	0.65	0.80	**0.93**
Recall	0.66	0.80	**0.93**
F-measure	0.65	0.79	**0.93**
Stage	Zhou and Zhu	Suzuki *et al.*	Wang *et al.*
κ statistic	**0.56**	0.41	0.44
Precision	**0.83**	0.69	0.76
Recall	**0.76**	0.69	0.71
F-measure	**0.78**	0.67	0.72
Extension	Zhou and Zhu	Suzuki *et al.*	Wang *et al.*
κ statistic	0.13	**0.13**	0
Precision	0.53	**0.57**	0.35
Recall	0.5	**0.62**	0.60
F-measure	0.51	**0.55**	0.44
N	Zhou and Zhu	Suzuki *et al.*	Wang *et al.*
κ statistic	**0.35**	0.19	0.23
Precision	**0.69**	0.48	0.53
Recall	**0.64**	0.57	0.60
F-measure	**0.62**	0.52	0.55
Size	Zhou and Zhu	Suzuki *et al.*	Wang *et al.*
Correlation coefficient	−0.01	0.13	**0.58**
Mean absolute error	1.16	0.98	**0.89**
Root mean square error	1.55	1.25	**1.06**
Relative absolute error	1.16	0.99	**0.90**
Root relative squared error	1.18	0.99	**0.86**

The bold values represents the results of the best performed method among the three approaches in comparison for individual prediction tasks.

4 Experimental results

All the proposed methods are evaluated against the ground truth on the 47 separate testing patients' samples, which have both H&E and IHC images. The quantitative evaluation results for BRAF, stage, extension, N and size are presented in Table 2, and box-whisker plots of error distributions are displayed in Figure 3. Information about the computation time and hardware/software specifications are provided in the Supplementary Material. For BRAF prediction,

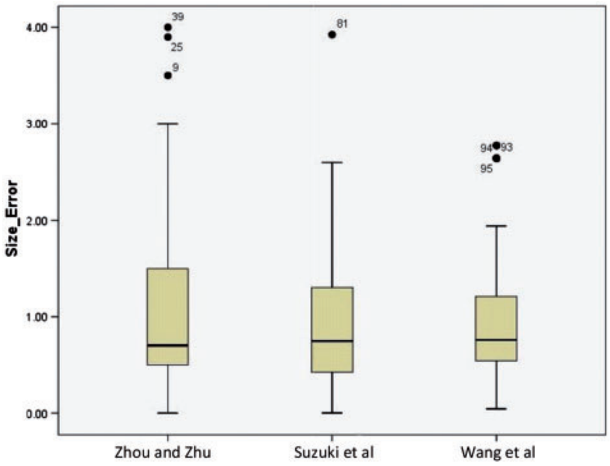

Fig. 3. The box-whisker plots of error distributions for size

it is observed that the best results by Wang *et al.*'s method with the highest Kappa, precision, recall and F-Measure (0.86, 0.93, 0.93, 0.93). The high precision and recall results indicate that the automated method is demonstrated to be promising in prediction of the BRAF mutation status based on the tissue images and patients' background information. In testing the significance of the patient background information in Wang *et al.*'s model, which builds machine learning models based on five automated quantification scores $Q_s|_{s=1...5}$ (see details in the Supplementary Material) and seven patient background features (see Table 1), we test the model using all five automated quantification scores and patient background information and gradually remove the patient background features one by one to evaluate the model performance. The results show deteriorated performance of Wang *et al.'s* method without patients' information (see Table 3).

In cancer stage prediction, Zhou and Zhu's method obtains the best results with the highest Kappa, precision, recall and F-measure (0.56, 0.83, 0.76, 0.78). In prediction of the extension and *N*, the best results on Kappa, precision, recall and F-measure are by Suzuki's method (0.13, 0.57, 0.62, 0.55) and by Zhou and Zhu's method (0.35, 0.69, 0.64, 0.62), respectively. For the size parameter results, Wang *et al.*'s method obtains the highest correlation coefficient and the lowest MSE, RMSE, RAE, RRSE (0.58, 0.89, 1.06, 0.90, 0.86), which is an interested finding as to the authors' best knowledge, this is the first study, which indicates the tumor size of thyroid cancer can be predicted using the patient background information (see Table 1) without directly measuring the tumor size.

5 Discussion

This section discusses possible future improvements of the three automated methods. Due to limited length of the paper, more analysis and discussion can be found in the Supplementary Material.

5.1 Zhou and Zhu: TMA-D^{2}LM

There are many potential improvements in the future. First, we will add convolutional layers before running the deep dictionary learning model in order to avoid losing some importance shape features. Second, we will explore other methods to find the center of the 'normal' space, which may give more accurate results. Third, for the continuous responses like 'size', we will develop better regression methods for predicting 'tumor size'.

Table 3. Evaluation on Wang *et al.'s* method by gradually removing the background information

	F-measure				Correlation coefficient
	BRAF	Stage	Extension	*N*	Size
$S_1 = Q \cup B$	0.93	0.75	0.44	0.55	0.58
$S_2 = S_1 \cap \sim b_1$	0.89	0.44	0.44	0.45	0.58
$S_3 = S_2 \cap \sim b_2$	0.86	0.52	0.44	0.45	0.58
$S_4 = S_3 \cap \sim b_3$	0.86	0.50	0.44	0.47	0.58
$S_5 = S_4 \cap \sim b_4$	0.86	0.54	0.44	0.44	0.58
$S_6 = S_5 \cap \sim b_5$	0.91	0.59	0.44	0.48	0.58
$S_7 = S_6 \cap \sim b_6$	0.84	0.54	0.43	0.49	0.58
$S_8 = Q$	0.80	0.59	0.39	0.48	−0.29
$S_9 = B$	0.56	0.70	0.44	0.57	0.58

Note: $Q = \{q_1, \ldots, q_5\}$.
$B = \{b_1, \ldots, b_7\} = \{$ *Age, Sex, Hashimoto, BW, BH, BMI, CancerSubtype* $\}$.

5.2 Suzuki *et al.*: hybrid prediction

Nuclei segmentation and feature extraction from HE stained slides of tumor were performed using CellProfiler version 2.2.0 (Carpenter *et al.*, 2006) onto randomly sampled image patches. The quantitative features include the size, Zernike shape features and pixel intensities. The reasons that Suzuki *et al.*'s approach does not perform well may be that the nuclear image features that Suzuki *et al.*'s method adopts are not enough for pattern analysis, and the image features other than nucleus such as cytoplasm are useful. In addition, inclusion of nuclei of non-tumor cells would affect the performance of the AI model. More sophisticated algorithms for stain normalization (Khan *et al.*, 2014), nuclear segmentation (Irshad *et al.*, 2014), or feature representations using deep learning could lead to more accurate results.

5.3 Wang *et al.*: ensemble

Wang *et al.*'s method is demonstrated to be promising in prediction of BRAF mutation and provide acceptable prediction accuracy in stage and relative high correlation score in estimating tumor size. However, as the method does not utilize the morphological patterns in H&E, the method has limitations in prediction of the clinical outcomes, which relate to tissue morphology, such as the extension, N and size. For future improvements, it is expected that the model may produce better and more reliable predictions outcomes for all five parameters by integration of morphological features extracted from H&E images.

6 Conclusion

Precision medicine is the future, and precision medicine demands personalized pathology. The ability to extract pathological features from a patient's tissue sample objectively and consistently can be driven by AI and machine learning. In this article, we have presented a benchmark for a number of challenging tasks in thyroid cancer diagnosis, including algorithms for prediction of the BRAF mutation status, extrathyroidal extension, lymph node metastastasis, TNM stage and the tumor size, using the patients' background information and H&E and IHC TMA images. The presented results will allow the objective comparison of existing and new developments in the field. All methods were evaluated using a common thyroid cancer dataset repository, ground truth and unified measurements for assessment of the prediction accuracy. Based on the presented results, we can conclude that recent methods achieved acceptable performance in prediction of BRAF mutation and TNM stage even without precise pixel-level annotation. However, the presented results also demonstrate that accurately predicting extrathyroidal extension, lymph node metastastasis and the tumor size remain challenging problems, which are still far from being solved. It is expected that this benchmark will help algorithmic developments, and that more advanced approaches will be built and tested using the provided data repositories and benchmarks.

Funding

Wang, Lee and Calista were supported by Taiwan Ministry of Science and Technology (MOST106-3114-8-011-002, 105-2221-E-011-121-MY2, 106-2622-8-011-001-TE2). Cheng was supported by an inter-institutional grant (MMH-TW-10504) from Mackay Memorial Hospital. Suzuki *et al.* was supported by JSPS Grant-in-Aid for Scientific Research (A), 16H02481 (SI). Zhu's work was partially supported by NIH grants MH086633 and MH092335, NSF grants SES-1357666 and DMS-1407655, and a grant from the Cancer Prevention Research Institute of Texas.

References

Avninder,S. *et al*. (2008) Tissue microarray: a simple technology that has revolutionized research in pathology. *J. Postgrad. Med*., **54**, 158–162.

Carpenter,A.E. *et al*. (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol*., **7**, R100.

Chen,W. and Foran,D.J. (2006) Advances in cancer tissue microarray technology: towards improved understanding and diagnostics. *Anal. Chim. Acta*, **564**, 74–81.

Chen,T. and Guestrin,C. (2016) XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, California, USA, pp. 785–794.

Cheng,S.P. *et al*. (2014) Significance of allelic percentage of BRAF c.1799T > A (V600E) mutation in papillary thyroid carcinoma. *Ann. Surg. Oncol*., **21(Suppl 4)**, S619–S626.

DeLellis,R.A. *et al*. (2004) *World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of Endocrine Organs*. IARC Press, Lyon.

Edge,S.B. *et al*. (2010) *AJCC Cancer Staging Manual*. 7th edn. Springer, New York.

Frank,E. *et al*. (2016) The WEKA Workbench. In: *Data Mining: Practical Machine Learning Tools and Techniques*. 4th edn. Morgan Kaufmann, Burlington, Massachusetts, USA. (Online Appendix) https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf.

IEEE International Symposium on Biomedical Imaging (2017) April 18–21, Melbourne, Australia. http://biomedicalimaging.org/2017/; http://www-o.ntust.edu.tw/~cvmi/ISBI2017/.

Irshad,H. *et al*. (2014) Methods for nuclei detection, segmentation, and classification in digital histopathology: a review-current status and future potential. *IEEE Rev. Biomed. Eng*., **7**, 97–114.

Jawhar,N.M. (2009) Tissue microarray: a rapidly evolving diagnostic and research tool. *Ann. Saudi Med*., **29**, 123–127.

Kasper,D.L. *et al*. (2015) *Harrison's Principal of Internal Medicine*, Mcgraw-Hill Education, New York.

Khan,A.M. *et al*. (2014) A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Trans. Biomed. Eng*., **61**, 1729–1738.

Otsu,N. (1979) A threshold selection method from gray level histograms. *IEEE Trans Syst. Man Cybern*., **9**, 62–66.

Ruifrok,A.C. and Johnston,D.A. (2001) Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol*., **23**, 291–299.

Ronneberger,O. *et al*. (2015) U-Net: convolutional networks for biomedical image Segmentation. *Proc. MICCAI*, **2015**, 234–241.

Simon,R. *et al*. (2010) Immunohistochemical analysis of tissue microarrays. In: Simon, R. (ed.). *Tissue Microarrays Methods and Protocols*. Springer Science+Business Media, New York, Vol. **664**, pp. 113–126.

Sturgis,C.D. *et al*. (1999) Image analysis of papillary thyroid carcinoma fine-needle aspirates: significant association between aneuploidy and death from disease. *Cancer*, **87**, 155–160.

Szegedy,C. *et al*. (2015) Going Deeper With Convolutions. In: *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9.

TissueGnostics®Gmbh (2014) Tissue FAXS Viewer 4.2, Build 6245.1019. Vienna, Austria, http://tissuefaxs-viewer.software.informer.com/download/.

Trigeorgis,G. *et al*. (2017) A deep matrix factorization method for learning attribute representations. *IEEE Trans. Pattern Anal. Mac. Intell*., **39**, 417–429.

Voduc,D. *et al*. (2008) Tissue microarrays in clinical oncology. *Semin. Radiat. Oncol*., **18**, 89–97.

Wang,C.-W. (2013) Fast quantification of immunohistochemistry tissue microarrays in lung carcinoma. *Taylor & Francis*, **16**, 707–716.

Wang,C.-W. and Chen,H.-C. (2013) Improved image alignment method in application to X-ray images and biological images. *Bioinformatics*, **29**, 1879–1887.

Wang,C.-W. *et al*. (2011) Robust automated tumour segmentation on histological and immunohistochemical tissue images. *PLoS One*, **6**, e15818. doi: 10.1371/journal.pone.0015818.

Wang,C-W. *et al*. (2014a) Robust image registration of biological microscopic images. *Sci. Rep*., **4**, 6050.

Wang *et al*. (2014b) Expression of haem oxygenase-1 correlates with tumour aggressiveness and BRAFV600E expression in thyroid cancer. *Histopathology*, **66**, 447–456.

Wang,D. *et al*. (2016) Deep learning for identifying metastatic breast cancer, *arXiv preprint*, arXiv: 1606.05718.

Zhang,D.Y. *et al*. (2009) Proteomics, pathway array and signaling network-based medicine in cancer. *Cell Div*., **4**, 20.