

## ARTICLE OPEN

# A principled machine learning framework improves accuracy of stage II colorectal cancer prognosis

Neofytos Dimitriou<sup>1</sup>, Ognjen Arandjelović<sup>1</sup>, David J. Harrison<sup>2</sup> and Peter D. Caie<sup>2</sup>

Accurate prognosis is fundamental in planning an appropriate therapy for cancer patients. Consequent to the heterogeneity of the disease, intra- and inter-pathologist variability, and the inherent limitations of current pathological reporting systems, patient outcome varies considerably within similarly staged patient cohorts. This is particularly true when classifying stage II colorectal cancer patients using the current TNM guidelines. The aim of the present work is to address this problem through the use of machine learning. In particular, we introduce a data driven framework which makes use of a large number of diverse types of features, readily collected from immunofluorescence imagery. Its outstanding performance in predicting mortality in stage II patients (AUROC = 0.94), exceeds that of current clinical guidelines such as pT stage (AUROC = 0.65), and is demonstrated on a cohort of 173 colorectal cancer patients.

*npj Digital Medicine* (2018)1:52; doi:10.1038/s41746-018-0057-x

## INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer worldwide and the leading cause of death among gastrointestinal tumours.<sup>1,2</sup> Annually, there are 1.4 million new cases and more than half a million of deaths worldwide.<sup>1</sup> A typical CRC diagnosis requires the evaluation of histopathological slides from a biopsy or resected specimen by a pathologist.<sup>3,4</sup> Subsequent to a positive diagnosis, prognosis is assessed based on the tumour-node-metastasis (TNM) staging system.<sup>5</sup> The TNM stage is considered by far one of the best predictors of CRC<sup>6</sup> and as a consequence, statistics specific to the stage primarily guide therapy. However, stages that exhibit higher variability in survival, encounter greater uncertainty. Stage II patients do not experience nodal (N) or distant (M) metastasis of their cancer and so only the depth of local invasion (T) is reported under TNM staging. Stage II CRC patients countenance an estimated 20% of 5-year poor prognosis, and 35% of 10 years poor prognosis.<sup>7,8</sup> Nevertheless, there are no definite criteria for selecting which, if any, stage II patients should undergo adjuvant chemotherapy with different trials reaching inconsistent conclusions.<sup>9,10</sup> It is therefore imperative to improve upon the prognosis of stage II CRC patients to better aid clinical guidance, reduce the survivability variance, and consequently, ameliorate treatment research.

Histopathological review of patient tissue sections by a pathologist remains subjective and thus suffers from inherent inter- and intra-observer variability. This affects TNM staging, especially due to the introduction of criteria within the staging guidelines, which are harder to standardize.<sup>11,12</sup> Nevertheless, this has a greater negative impact when reporting features independent of TNM that may aid in determining stage II patients with a higher risk of disease specific death.<sup>11,12</sup> One such feature is histological grading, or equivalently differentiation, currently within the core data set of international reporting guidelines for CRC.<sup>3,13</sup> Despite attempts to maintain consistency in reporting this feature, such as moving from a three-tiered system down to two

tiers, reproducibility issues persist.<sup>3,13</sup> Other promising histopathological features for further stratifying stage II CRC patients include lymphatic vessel invasion and tumour budding.<sup>14–16</sup> However, they are currently listed within non-core data items,<sup>3</sup> despite consistent demonstration of their prognostic significance. This has been attributed to the high observer variability and hence, methodological shortcomings of quantifying these features in a standardized manner.<sup>17,18</sup>

Both medical practice and research are moving towards a more nuanced approach in clinical decision-making. Pathology is now embracing the era of digitization with a multitude of interdisciplinary studies employing techniques from fields such as image analysis, machine learning (ML) and deep learning.<sup>19–22</sup> The use of these techniques markedly increases efficiency and efficacy compared to traditional methods, while removing the subjectivity imposed by the human pathologist.<sup>23–25</sup> Moreover, multiplexed detection of target proteins is becoming more commonplace in pathology research through wider adoption of immunofluorescence (IF). Data collected through IF provide a multi-dimensional representation of the tumour micro-environment with each biomarker co-registered to the same physical coordinates in the tissue. In addition, utilizing specific antibodies to visualize histopathological features overcomes common issues of reporting from H&E stained tissue, such as retraction artefact confounding lymphatic vessel invasion and high density immune infiltrate obscuring tumour buds.<sup>17,18</sup> Therefore, employment of techniques from the aforementioned fields on IF data have the potential to exploit multidimensional data, ranging from morphometric to spatial characteristics of selected histopathological features, and aid in improving prognosis for stage II CRC patients.

The present work builds upon previous efforts in the field,<sup>26</sup> which make use of image analysis for the extraction of histopathological features (such as nuclear grade, tumour budding and lymphatic vessel invasion, cellular shape, size, texture, etc.), a priori known or expected to be salient, and simple statistical

<sup>1</sup>School of Computer Science, University of St Andrews, St Andrews KY16 9SX, UK and <sup>2</sup>School of Medicine, University of St Andrews, St Andrews KY16 9TF, UK  
Correspondence: Neofytos Dimitriou (neofytosd@gmail.com)

Received: 15 March 2018 Revised: 22 August 2018 Accepted: 4 September 2018  
Published online: 02 October 2018

techniques for the subsequent inference. In particular, we describe a principled and data driven framework which uses modern machine learning to predict the survival outcome for a stage II CRC patient from a large number of histopathological features.

## RESULTS

### Full feature set based prognosis

Each baseline classifier's hyperparameter values were learnt by maximizing the corresponding average area under the receiver operating characteristic curve (AUROC) on the validation data corpus. Table 1 summarizes the results. The average AUROC across all classifiers was found to be 0.89 both for 5- and 10-year prognosis. One-way analysis of variance (ANOVA) and Tukey's honest significance difference test (THSD) showed no statistical significance between classifiers for 10-year prognosis. The only statistically significant difference is that between naïve Bayes (NB)

and logistic regression (LR)-based approaches for 5-year prognosis (ANOVA  $p$  value  $< 0.01$ , THSD  $p$  value  $< 0.003$ ).

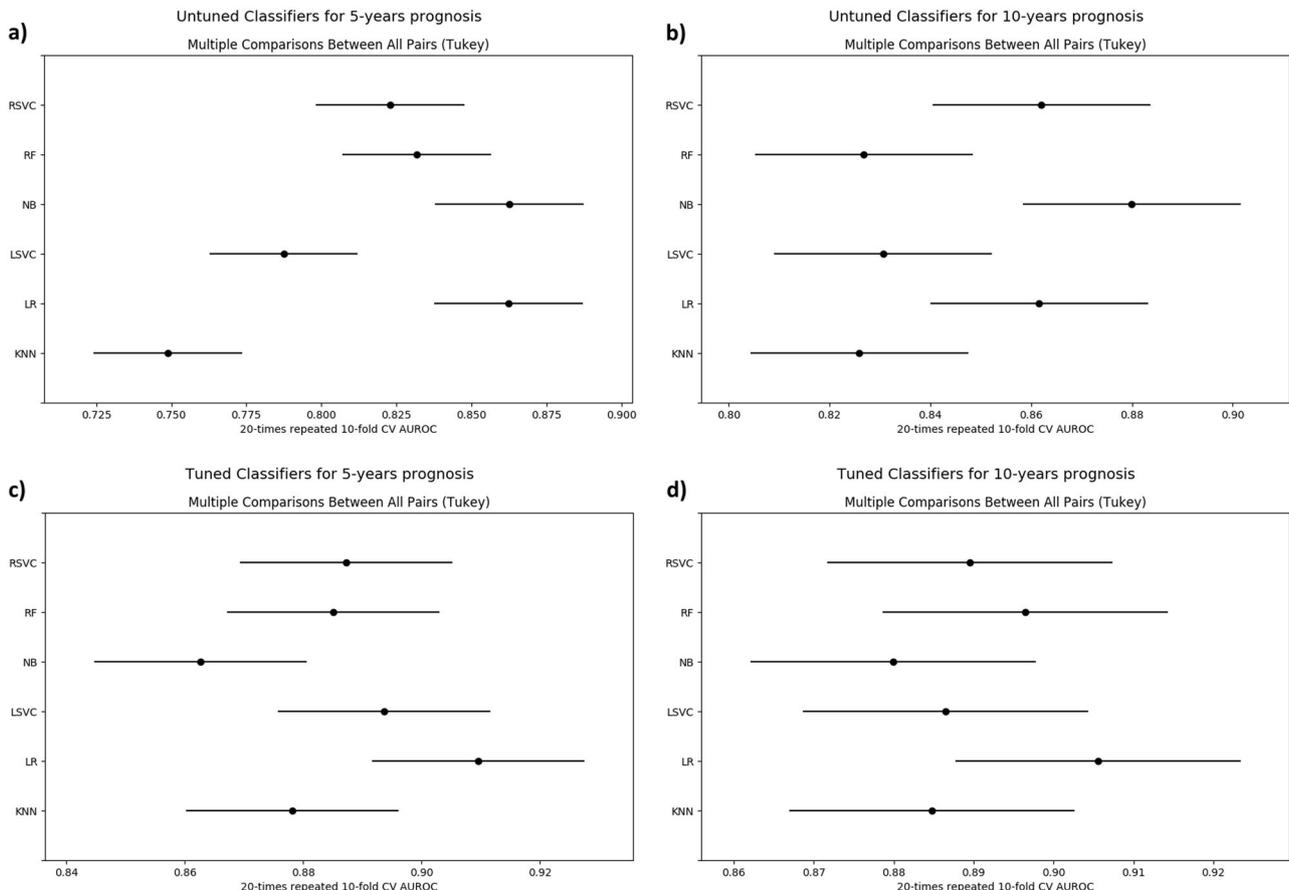
To demonstrate the importance of model selection, we also compared the performance of all classifiers using hyperparameter values, which were learnt as described in the previous section, and with the a priori set hyperparameters values as in the existing literature. As expected, using the latter approach a drop in the average AUROC was observed both for 5- and 10-year prognosis, to respectively 0.82 (approximately 8.0% drop) and 0.85 (approximately 4.5% drop). The results are visualized in Fig. 1.

### Reduced feature sets

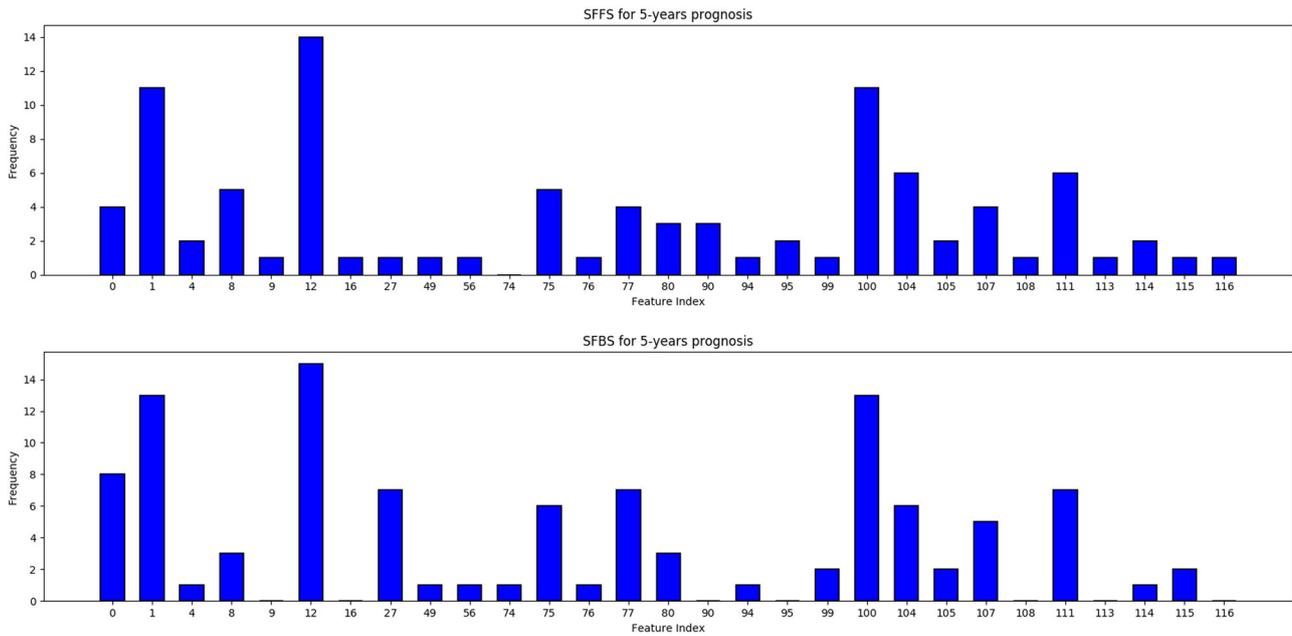
**Feature selection.** The evaluation of each subset of features was performed by tenfold cross-validation on the training data. To reduce outcome variability caused by stochastic effects we adapt the method proposed by Dune et al.<sup>27</sup>. In particular, we performed sequential floating forward search (SFFS) and sequential floating

	LSVM	RSVM	LR	RF	KNN	NB
5 year	$0.89 \pm 0.12$	$0.89 \pm 0.13$	$0.91 \pm 0.12$	$0.89 \pm 0.13$	$0.88 \pm 0.12$	$0.86 \pm 0.14$
10 year	$0.89 \pm 0.13$	$0.89 \pm 0.12$	$0.91 \pm 0.119$	$0.90 \pm 0.13$	$0.89 \pm 0.13$	$0.88 \pm 0.12$

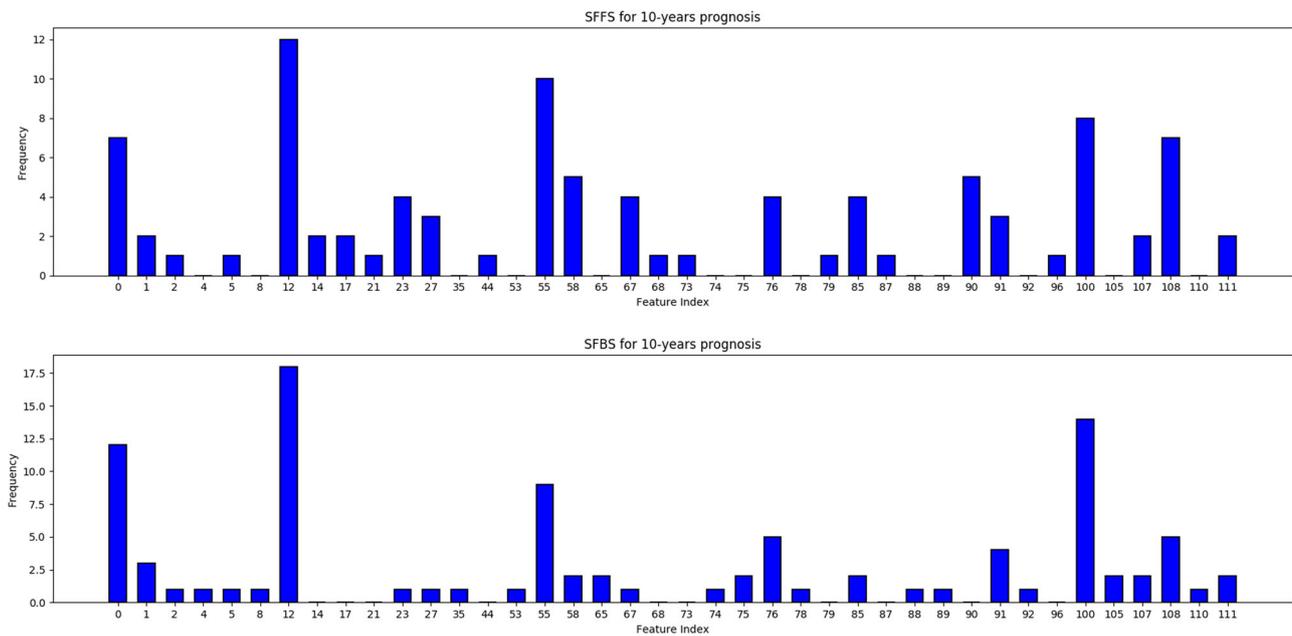
LSVM linear kernel SVM, RSVM radial basis function kernel SVM



**Fig. 1** Tukey's significance difference test. No hyperparameter learning was employed in the experiments corresponding to the plots **a** and **b**, in contrast to **c** and **d**



**Fig. 2** Frequency of occurrence of each feature from the 20 runs of SFFS and SFBS each for 5-year prognosis. Only features with at least one occurrence are shown for clarity



**Fig. 3** Frequency of occurrence of each feature from the 20 runs of SFFS and SFBS each for 10-year prognosis. Only features with at least one occurrence are shown for clarity

backwards search (SFBS) 40 times using different random partitions, each time retaining the feature subset that achieved the best performance. Following aggregation—see Figs. 2 and 3—the subsets from SFFS and SFBS were combined and features ordered based on the frequency of occurrence. Starting with an empty set, features were added in an incremental fashion based on their average AUROC rank, estimated through 20-times repeated tenfold cross-validation. The subset of features that achieved the highest averaged AUROC was selected for each prognostic term, as summarized in Table 2.

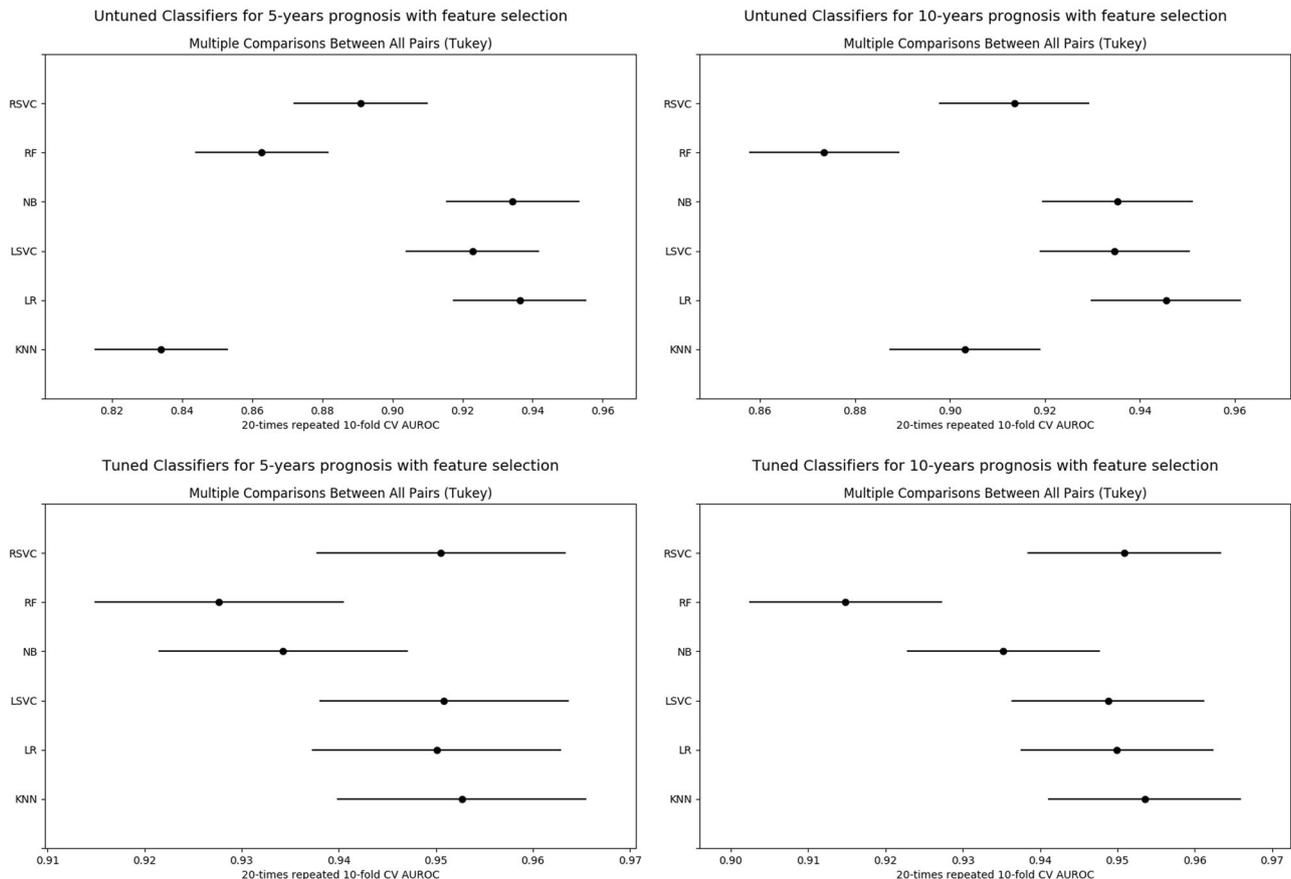
*Experiments.* We followed the same approach to classifier training, model selection, and evaluation as in the previous

section. The sole difference is that instead of the full feature set, for this set of experiments a reduced set of selected features (as described previously) was used.

As expected, we observed a significant improvement in performance already at the coarsest level of analysis, with the average AUROC across classifiers reaching 0.94, both for 5- and 10-year prognosis. In line with our previous findings, no statistically significant difference was observed between different classifiers, except for the inferiority of random forest (RFs) for 10-year prognosis (ANOVA  $p < 0.0001$ , THSD  $p < 0.01$ ). Just as in the previous set of experiments, our data driven approach to hyperparameter selection was always found to effect a statistically significant improvement over their being set a priori; see Fig. 4.

**Table 2.** Features of significance to both prognosis terms, and those which were specific to a particular term; seven and six features were used for 5 and 10-year prognosis, respectively

# Features	
Unique to 5-year prognosis	4 Nuclei in tumour mean DAPI intensity, number of CK objects with no associated nuclei, sum area of vessels, average DAPI intensity (tumour area)
Unique to 10-year prognosis	3 Nuclei in tumour mean D240 intensity, mean compactness of tumour glands, number of PDCs
Common to both prognoses	3 Nuclei in tumour bud mean DAPI intensity, tumour gland relative area (%), sum area of vessels
CK pancytokeratin, PDCs poorly differentiated clusters	

**Fig. 4** Tukey's significance difference test. No hyperparameter learning was employed in the experiments corresponding to the plots **a** and **b**, in contrast to **c** and **d**

### Final testing

We started by examining training set performance of different classifiers using 20-times repeated tenfold cross-validation. It can be readily seen that classifiers trained on the subset of features selected by SFFS and SFBS performed better, as illustrated in Tables 1 and 3. Though simple, the best performing classifier was found to be KNN-based classifier (with the Minkowski distance metric) both for 5-year ( $k = 36$ ) and 10-year prognosis ( $k = 28$ ).

Kaplan-Meier (KM) survival curves were employed to visualize the difference in survivability between the predicted prognosis groups, and the log-rank test used for objective quantification thereof. For 5-year prognosis, our KNN-based approach achieved the AUROC of 0.77, effecting a good separation patients into high and low-risk ( $p$  value  $< .02$ ). On 10-year prognosis, the classifier achieved the AUROC of 0.94, significantly outperforming the current clinical gold standard of pT stage (AUROC of 0.65), and

even better separation between high- and low-risk patients (log-rank test  $p < .0001$ ). The sensitivity of 42.9%, specificity of 89.2%, and accuracy of 81.8% were achieved for 5-year prognosis, and the sensitivity of 100%, specificity of 84%, and accuracy of 88.9%, for 10-year prognosis. The differentiation (poor/moderate vs. good) and T stage discrimination (T3 vs. T4) results are summarized in Figs. 5, 6 and 7, as well as in Table 4.

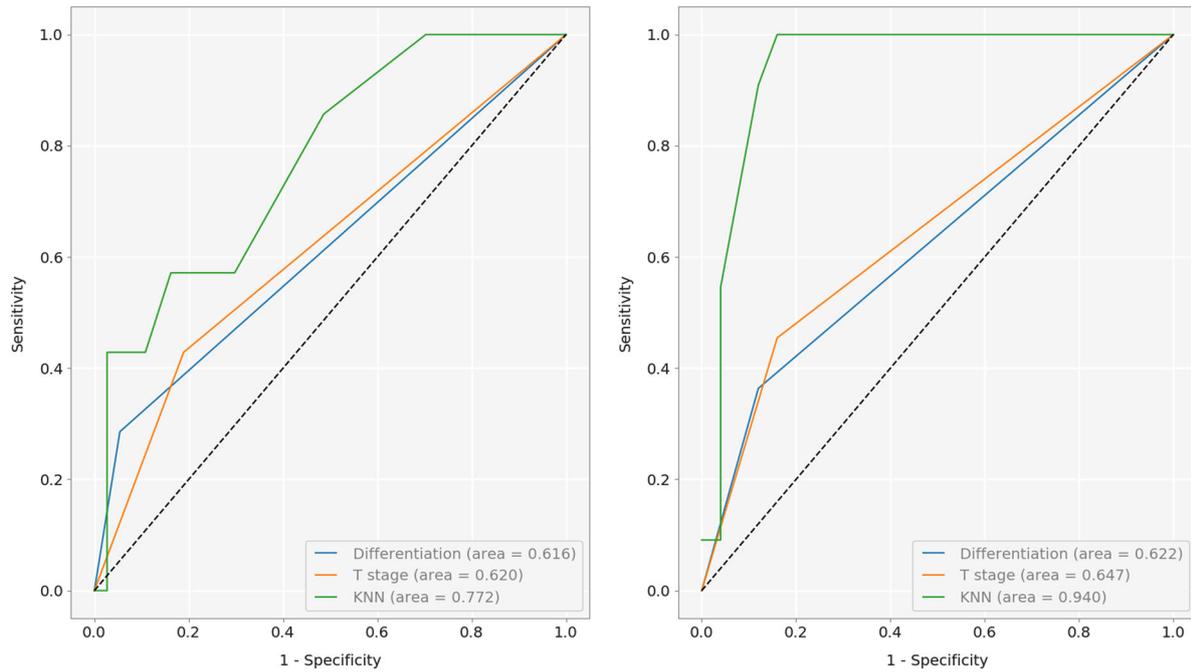
### DISCUSSION

CRC is a highly heterogeneous disease, which limits the prognostic accuracy of the TNM staging system or the reporting based on individual features such as tumour budding,<sup>28</sup> or lymphatic vessel invasion and density.<sup>29</sup> Prior work on the use of automated image analysis and ML applied to other types of cancer has focused on parameters solely from tumour cells.<sup>20,21</sup> However, the evidence

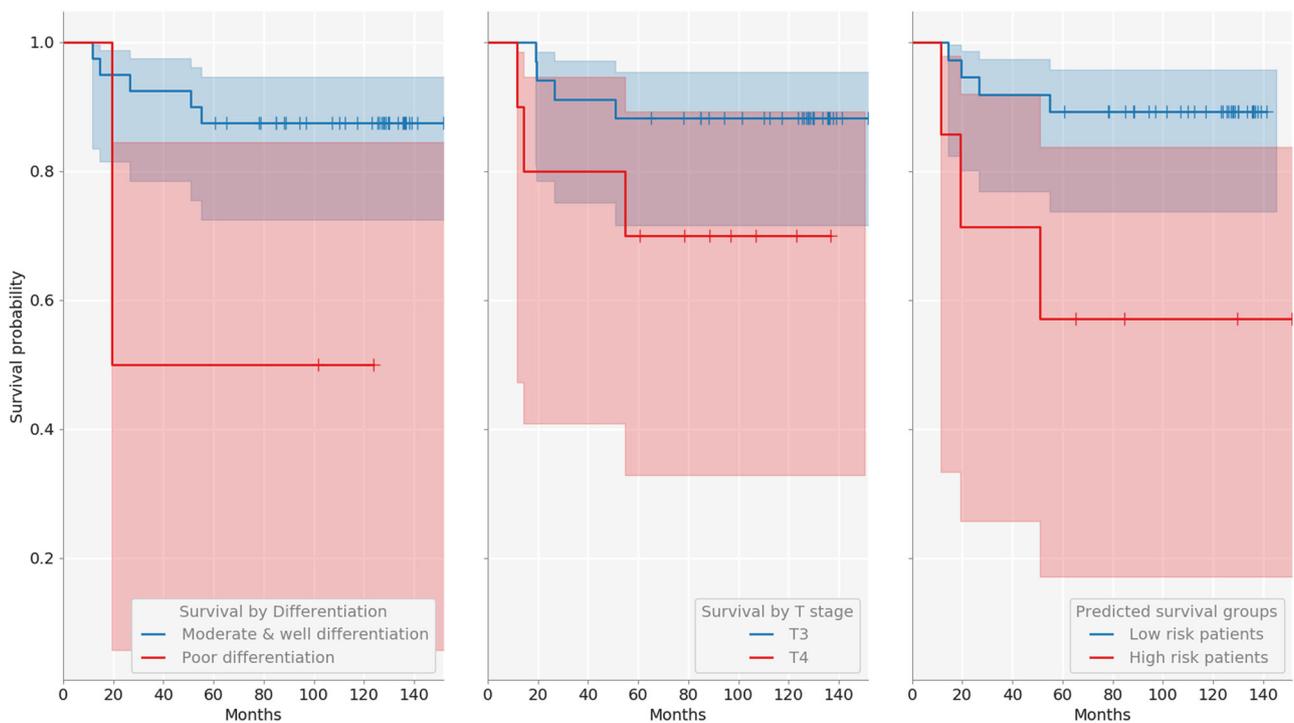
**Table 3.** Average AUROC and standard deviation (for  $n = 200$ ) of each trained classifier using only features selected by SFFS and SFBS

	LSVM	RSVM	LR	RF	KNN	NB
5 years	$0.95 \pm 0.08$	$0.95 \pm 0.08$	$0.95 \pm 0.08$	$0.93 \pm 0.11$	$0.95 \pm 0.08$	$0.93 \pm 0.10$
10 years	$0.95 \pm 0.08$	$0.95 \pm 0.08$	$0.95 \pm 0.08$	$0.92 \pm 0.10$	$0.95 \pm 0.07$	$0.94 \pm 0.09$

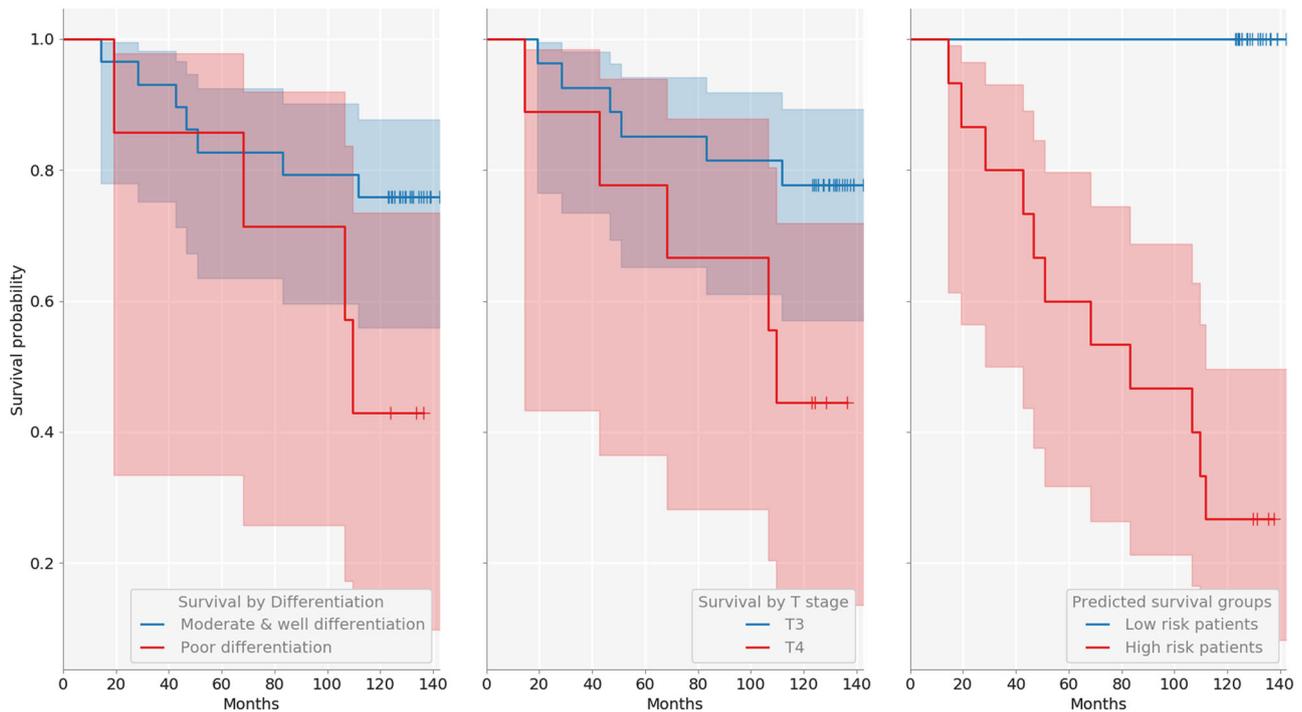
The experiments were performed by 20 times repeating tenfold cross-validation on training data.



**Fig. 5** ROC curves for the two prognostic terms of interest



**Fig. 6** KM curves for 5-year prognosis



**Fig. 7** KM curves for 10-year prognosis

	Differentiation (5/10 year)	T stage (5/10 year)	KNN (5/10 year)
Specificity	0.95/0.88	0.82/0.84	0.89/0.84
Sensitivity	0.39/0.36	0.43/0.46	0.43/1.00
Accuracy	0.84/0.72	0.75/0.72	0.82/0.89
AUROC	0.62/0.62	0.62/0.65	0.77/0.94

from an increasing number of studies suggests that the tumour micro-environment is just as informative,<sup>30–32</sup> which motivated us to use information not only from tumour nuclei but also from numerous hierarchical features such as texture, morphology, fluorescence intensity, and spatial relationships across the micro-environment of the invasive margin. Hence, we introduced a carefully crafted ML based framework capable of nuanced prediction of survival for stage II CRC patients. Our methodology was shown to outperform significantly the current gold standard in the form of pT staging. Specifically, our method achieved AUROC of over 77 and 94% for 5 and 10-year prognoses respectively, compared to pT stage, which stratifies patients with the AUROC of approximately 62% both for 5- and 10-year prognosis, and the differentiation, which achieves the corresponding AUROC of approximately 62 and 65%, respectively. Moreover, we demonstrated high interpretability of the proposed approach, allowing clinicians to gain new insight by identifying prognostically the most salient features.

Confirming findings from prior empirical research as well as one of the premises of the present work, our experiments demonstrated that a diverse set of characteristics of the entire micro-environment have a prognostic value. This explains the outstanding performance of our method and the major improvement on the current state of the art which focuses on a single aspect thereof (usually tumour cells). DAPI intensity within the nuclei of tumour buds was consistently found to carry the greatest prognostic weight, which too agrees with previous empirical

findings—cells within more invasive and mesenchymal tumour buds have increased plasticity and gene expression,<sup>15,33</sup> which effects an increase in DAPI intensity. Furthermore, in the present study, this feature was highly correlated with parameters associated with tumour bud nuclei morphometry, whereby features linked to larger and more irregular shaped nuclei (such as found in more aggressive poorly differentiated cancer cells) were associated with poorer prognosis. This phenomenon would further explain an increase in the DAPI intensity within parameters describing other tumour subpopulations and which are reported parameters from the model. Tumour gland nuclear morphometry, also found to be of major prognostic importance, has also been identified in the past.<sup>34,35</sup> Other selected features included known histopathological features such as the number of PDCs,<sup>36</sup> the number and area of lymphatic vessels,<sup>29</sup> and the shape and area of tumour glands.<sup>37,38</sup>

It is interesting to observe and comment on our finding that certain features were specifically associated with a particular prognostic term. Having looked at this in detail, we found high correlation between these features, within specific survival terms, and outcomes, suggesting that the features are not specific to set survival times per se but are rather associated with poorer outcomes. For example, the number of small pan cytokeratin positive objects with no associated nuclei was found to be an important feature for 5-year survival. On the other hand, the number of PDCs was found to be an important prognostic feature for 10-year survival. Nevertheless, both were highly correlated with the number of tumour buds.

Digital pathology is becoming more common in the clinical workflow, with recently, Glasgow and Oxford hospitals committing to a fully digital workflow. The digitization of pathology will allow the embedding of image analysis and AI solutions into a pathologist's routine practice. Fully automated workflow, such as the one presented here, allows results to be reported to the patient in a shorter time frame while freeing up more of a pathologist's large workload. Studies such as these add to the body of work exemplifying proof of concepts, which use image analysis and AI for cancer pathology. In order for automated

**Table 5.** Summary of patient cohort statistics

Number of patients		173
Age (years)	Range	62.5 ± 33.5
	Median	67
Gender	Male	86 (50%)
	Female	87 (50%)
T Stage	TX	1 (1%)
	T1	6 (3%)
	T2	7 (4%)
	T3	122 (71%)
	T4	37 (21%)
N Stage	N0	163 (94%)
	N1	8 (5%)
	N2	1 (1%)
	N3	1 (1%)
M Stage	MX	9 (5%)
	M0	161 (93%)
	M1	3 (2%)
Site	Rectum	56 (32%)
	Colon	117 (68%)
Differentiation	Undetermined	3 (2%)
	Poor	25 (14%)
	Moderate	138 (80%)
	Good	7 (4%)

**Table 6.** The search space of each classifier based on the distributions over its hyperparameters (n.b.  $F$  denotes feature count; for biased categorical distributions, tuples  $(p_s, v)$  designate the sampling probability and the value assigned)

Classifier	Hyperparameter	Distribution	Values
SVM, linear kernel	C	Log-uniform	[ln (1e−5), ln (1e2)]
	Class weight	Categorical	Balanced or none
SVM, RBF kernel	C	Log-uniform	[ln (1e−5), ln (1e2)]
	Gamma	Log-uniform	[ln (1e−3), ln (1e3)]
LR	Class weight	Categorical	Balanced or none
	Type of penalty	Categorical	L1 or L2
RF	C	Log-uniform	[ln (1e−5), ln (1e2)]
	Class weight	Categorical	Balanced or none
	Number of trees	Log-uniform integer	[10, 1000]
	Criterion	Categorical	Gini or entropy
	Maximum features	Biased categorical	(0.2, √F), (0.1, ln F), (0.1, F), (0.6, U(0, F))
KNN	Maximum depth	Biased categorical	(0.1, 2), (0.1, 3), (0.1, 4), (0.7, none)
	Bootstrap	Categorical	True or False
	Class weight	Categorical	Balanced or none
KNN	K	Log-uniform integer	[1, 50]
	Weights	Categorical	Uniform, or Euclidean distance
	Metric	Categorical	Balanced or none
	P	Categorical	Balanced or none

image analysis and AI to be translated into the clinic further regulatory approved validation studies must be applied utilizing large patient cohorts sourced from multiple institutions.

In summary, the present work made several important contributions: (i) a principled framework for data driven ML based precise prognosis of stage II CRC cancer outcomes, (ii) significantly better performance than the current state of the art, (iii) clinical insight into the disease, and (iv) demonstrated the general potential of modern ML in digital pathology and health care more broadly. Following the highly promising results reported herein, our future work will focus on the application of computer vision and ML directly on histopathological tissue slides, so as to avoid the loss of information associated with ‘atomization’ of the process<sup>39</sup> effected by human driven feature extraction and the subsequently applied learning from these. Additionally, in order to increase the potential for clinical adoption of the developed methodologies, it will likely be of interest to consider how the results should be presented to the clinician.<sup>40,41</sup>

## METHODS

Our experimental data were obtained from tissue samples of 180 Scottish patients who had been diagnosed with CRC and who underwent surgical resection, with a minimum follow-up of 11.5 years. Patients that succumbed within 5 days of the surgery were excluded to ensure that surgical complications did not contribute to the cause of death, as were the three patients that received therapy due to potential effects on the

relevant micro-environment and hence survival.<sup>42</sup> Table 5 summarizes the key clinical and demographic characteristics of the cohort.

The use of tissue samples was approved by the East of Scotland Research Ethics Service (13/ES/0126). Further ethical clearance was not required as the acquired data was anonymized. For more detailed patient information please see the previous work of Caie et al.<sup>26</sup>

## Features

The digitization of the tissue samples, and subsequent quantification and extraction of histological features were part of the work completed by Caie et al.<sup>26</sup> Both are briefly described hereunder but interested readers should refer to Caie et al.<sup>26</sup> and the corresponding supplementary document for a more thorough overview.

Tissue samples were prepared for multiplex immunofluorescence with pan cytokeratin and D2-40 antibodies, along with DAPI stain for the detection of epithelial cells, lymphatic vessels, and cell nuclei. The invasive front was manually identified through the pan cytokeratin channel of each whole-slide image captured at 40× magnification. Fifteen evenly spaced high-resolution (200× magnification) images were captured across the invasive front of each sample. Regions of interest (ROIs) (including stroma, tumour glands, invasive tumour subpopulations, lymphatic vasculature, and cell nuclei) were detected and segmented from each imported image using Definiens AG image analysis software package. Each ROI was described by a collection of morphometric, spatial, and fluorescence related characteristics associated with each patient, resulting in 123 histopathological features (independent variables); for further detail see Supplementary Document.

For each patient, pathological and demographic features were collected as well. The former set comprises the level of differentiation, site of primary tumour, and the corresponding disease stage, and the latter the patient’s age, gender, survival status at multiple clinically relevant follow-up intervals, and (where applicable) time until death. Except for the survival

status, which was the dependent variables of interest in the present work, the remaining features were used for the analysis of experimental results, and not for the actual learning and prediction.

### Data preparation

We followed the standard approach to algorithm training and evaluation, by splitting the cohort dynamically into non-overlapping training, validation (or development), and test subsets. In particular, data were first randomly (with stratification) split into two, 70 and 30%, the latter being the test subset. Using tenfold cross-validation, the former, large subset was in each iteration of the process further randomly split into training and validation subsets.

It is worth noting that, given the key aim of the present work, while the evaluation corpus contain only stage II patients, we decided to include differently staged patients in the training corpus. Our hypothesis was that in spite of not being the target population for our prediction, useful pathological patterns could be learnt from this data too, allowing a degree of interpolation to take place. Stratified sampling was employed in order to maintain the prognosis distribution of each cohort as a means of countering the imbalanced nature of our data, and thus avoid class under-representation.<sup>43</sup> Lastly, features were normalized to zero mean and unity variance.

### Baseline classification and performance assessment

The problem at hand was formalized as a binary, supervised classification task, whereby the prediction was that of a good or bad prognosis, i.e. survived or not, respectively. We adopted several well-understood baseline classifiers, with different underlying assumptions (explicit or implicit) and mathematical underpinnings. In particular, we compared classifiers based on support vector machines,<sup>44</sup> RFs,<sup>45</sup> *k*-nearest neighbours (KNN),<sup>46</sup> NB,<sup>47</sup> and LR.<sup>48</sup> In an effort to capture performance adequately on a highly imbalanced data set, the AUROC<sup>49</sup> is adopted as the primary performance measure. In addition, for the sake of consistency with related work and ease of comparative analysis, we also report specificity and sensitivity, and accuracy.

### Model selection

The capability of a model to represent information, as well the efficiency its learning is governed by a number of parameters. These parameters, referred to as hyperparameters, need to be set prior to training. However, finding the optimal or close to optimal set of hyperparameter values is challenging. The commonly used and probably the simplest approach, in the form of a grid search has limited applicability due to its intractability for complex models. A random search over predefined ranges of hyperparameters often produces better results while being computationally less demanding.<sup>50</sup> However, both techniques are naïve as they do not take into account historical patterns.

Sequential model based global optimization (SMBO) techniques adopt a more sophisticated approach, approximating the possibly computationally expensive fitness function with a simpler surrogate.<sup>51</sup> Different SMBO approaches optimize different criteria which then guide the surrogate of the fitness function. The one adopted herein is tree-structured Parzen estimator (TPE), which optimizes the so-called 'expected improvement'. Conceptually, TPE initially behaves like a random search, subsequently refining the search so that hyperparameter values associated with poor performance are not re-visited.<sup>51,52</sup> This process is guided probabilistically, using suitable densities or distributions associated with the type of hyperparameter. Those used in the present work are summarized in Table 6. Finally, as the loss function we used the negated AUROC resulting from tenfold cross-validation, averaged over 20 independent runs and using 500 iterations.

### Feature selection

In order to address potential problems associated with the so-called curse of dimensionality, which becomes of increasing concern with a large number of features, we examined the use of dimensionality reduction in the context of the problem at hand.<sup>53,54</sup> In particular, motivated by their successful use in the existing literature<sup>55</sup> we employed SFFS and SFBS,<sup>55–57</sup> which respectively perform recursive removal or addition of features in an attempt to improve a specific metric, until the desired reduction in the feature number is attained.

### Code availability

Full code is available from the authors upon request.

### DATA AVAILABILITY

The data used in this work is available from the authors upon reasonable request.

### ACKNOWLEDGEMENTS

We would like to thank and acknowledge NHS Lothian, and in particular Mrs Frances Rae, for providing the tissue, clinical data, and ethical clearances associated with the present work.

### AUTHOR CONTRIBUTIONS

N.D., O.A., and P.C. conceived of the presented idea, and contributed to the development of the technical approach and experimental analysis. D.H. supervised the project and facilitated access to data. All authors discussed the results and contributed to the final manuscript.

### ADDITIONAL INFORMATION

**Supplementary information** accompanies the paper on the *npj Digital Medicine* website (<https://doi.org/10.1038/s41746-018-0057-x>).

**Competing interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### REFERENCES

1. Ferlay, J. et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *Int. J. Cancer* **136**, E359–E386 (2015).
2. Langner, C. & Schneider, N. Prognostic stratification of colorectal cancer patients: current perspectives. *Cancer Manag Res* 291 (2014). <https://doi.org/10.2147/cmr.s38827>.
3. Loughrey, M. B., Quirke, P. & Shepherd, N. A. Dataset for colorectal cancer histopathology reports. *The Royal College of Pathologists* **343**, 1–47 (2014).
4. Brenner, H., Kloor, M. & Pox, C. P. Colorectal cancer. *Lancet* **383**, 1490–1502 (2013).
5. Edge, S. B. & Compton, C. C. The American Joint Committee on cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann. Surg. Oncol.* **17**, 1471–1474 (2010).
6. Fleming, M., Ravula, S., Tatishchev, S. F. & Wang, H. L. Colorectal carcinoma: pathologic aspects. *J. Gastrointest. Oncol.* **3**, 153–173 (2012).
7. Compton, C. C. Optimal pathologic staging: defining stage II disease. *Clin. Cancer Res.* **13**, 6862s–6870s (2007).
8. Nauta, R., Stablein, D. M. & Holyoke, D. Survival of patients with stage b2 colon carcinoma. *Arch. Surg.* **124**, 180 (1989).
9. Barone, C. Adjuvant chemotherapy of colon cancer current strategies. *Eur. J. Cancer Suppl.* **6**, 60–63 (2008).
10. Lombardi, L. et al. Adjuvant colon cancer chemotherapy: where we are and where well go. *Cancer Treat. Rev.* **36**, S34–S41 (2010).
11. Lea, D., Häland, S., Hagland, H. R. & Søreide, K. Accuracy of TNM staging in colorectal cancer: a review of current culprits, the modern role of morphology and stepping-stones for improvements in the molecular era. *Scand. J. Gastroenterol.* **49**, 1153–1163 (2014).
12. Maguire, A. Controversies in the pathological assessment of colorectal cancer. *World J. Gastroenterol.* **20**, 9850 (2014).
13. von Karsa, L. et al. European guidelines for quality assurance in colorectal cancer screening and diagnosis: overview and introduction to the full supplement publication. *Endoscopy* **45**, 51–59 (2012).
14. Lai, Y.-H. et al. Tumour budding is a reproducible index for risk stratification of patients with stage II colon cancer. *Colorectal Dis.* **16**, 259–264 (2014).
15. Lugli, A., Karamitopoulou, E. & Zlobec, I. Tumour budding: a promising parameter in colorectal cancer. *Br. J. Cancer* **106**, 1713–1717 (2012).
16. Lin, M. et al. Intratumoral as well as peritumoral lymphatic vessel invasion correlates with lymph node metastasis and unfavourable outcome in colorectal cancer. *Clin. & Exp. Metastasis*. **27**, 123–132 (2010).
17. Kojima, M. et al. Pathological diagnostic criterion of blood and lymphatic vessel invasion in colorectal cancer: a framework for developing an objective pathological diagnostic system using the Delphi method, from the Pathology Working

- Group of the Japanese Society for Cancer of the Colon and Rectum. *J. Clin. Pathol.* **66**, 551–558 (2013).
18. Zaorsky, N. G., Patil, D., Freedman, G. M. & Tuluc, M. Differentiating lymphovascular invasion from retraction artifact on histological specimen of breast carcinoma and their implications on prognosis. *J. Breast Cancer* **15**, 478 (2012).
19. Korbar, B. et al. Deep-learning for classification of colorectal polyps on whole-slide images. *Clin Orthop Relat Res.* [abs/1703.01550](https://doi.org/10.1038/s41598-017-13773-7) (2017).
20. Vandenberghe, M. E. et al. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Sci. Rep.* **7**, 45938 (2017).
21. Wang, X. et al. Prediction of recurrence in early stage non-small cell lung cancer using computer extracted nuclear features from digital H&E images. *Sci Rep.* **7** (2017). <https://doi.org/10.1038/s41598-017-13773-7>.
22. Rao, A. et al. Identification of histological correlates of overall survival in lower grade gliomas using a bag-of-words paradigm: a preliminary analysis based on hematoxylin & eosin stained slides from the lower grade glioma cohort of the cancer genome atlas. *J. Pathol. Inform.* **8**, 9 (2017).
23. Galon, J. et al. Towards the introduction of the 'immunoscope' in the classification of malignant tumours. *J. Pathol.* **232**, 199–209 (2013).
24. Thakur, S. S. et al. The use of automated ki67 analysis to predict oncotype DX risk-of-recurrence categories in early-stage breast cancer. *PLoS One* **13**, e0188983 (2018).
25. Bankhead, P. et al. Integrated tumor identification and automated scoring minimizes pathologist involvement and provides new insights to key biomarkers in breast cancer. *Lab. Invest.* **98**, 15–26 (2017).
26. Caie, P. D., Zhou, Y., Turnbull, A. K., Oniscu, A. & Harrison, D. J. Novel histopathologic feature identified through image analysis augments stage II colorectal cancer clinical reporting. *Oncotarget* **7**, 44381–44394 (2016).
27. Dunne, K., Cunningham, P. & Azuaje, F. Solutions to instability problems with sequential wrapper-based approaches to feature selection. *J. Mach. Learn. Res.*, 1–22 (2002).
28. Horcic, M. et al. Tumor budding score based on 10 high-power fields is a promising basis for a standardized prognostic scoring system in stage II colorectal cancer. *Hum. Pathol.* **44**, 697–705 (2013).
29. Cacchi, C. et al. Clinical significance of lymph vessel density in T3 colorectal carcinoma. *Int. J. Colorectal Dis.* **27**, 721–726 (2012).
30. Sugai, T. et al. Vascular invasion and stromal s100a4 expression at the invasive front of colorectal cancer are novel determinants and tumor prognostic markers. *J. Cancer* **8**, 1552–1561 (2017).
31. Heindl, A., Nawaz, S. & Yuan, Y. Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology. *Lab. Invest.* **95**, 377–384 (2015).
32. Isella, C. et al. Stromal contribution to the colorectal cancer transcriptome. *Nat. Genet.* **47**, 312–319 (2015).
33. Bhangu, A. et al. Epithelial mesenchymal transition in colorectal cancer: seminal role in promoting disease progression and resistance to neoadjuvant therapy. *Surg. Oncol.* **21**, 316–323 (2012).
34. Nakashima, Y. et al. Nuclear atypia grading score is a useful prognostic factor in papillary gastric adenocarcinoma. *Histopathology* **59**, 841–849 (2011).
35. Eynard, H. G., Soria, E. A., Cuestas, E., Rovasio, R. A. & Eynard, A. R. Assessment of colorectal cancer prognosis through nuclear morphometry. *J. Surg. Res.* **154**, 345–348 (2009).
36. Barresi, V., Bonetti, L. R., Leni, A., Caruso, R. A. & Tuccari, G. Poorly differentiated clusters: clinical impact in colorectal cancer. *Clin. Colorectal Cancer* **16**, 9–15 (2017).
37. Hynes, S. O. et al. Back to the future: routine morphological assessment of the tumour microenvironment is prognostic in stage II/III colon cancer in a large population-based study. *Histopathology* **71**, 12–26 (2017).
38. Rajaganesan, R. et al. The influence of invasive growth pattern and microvessel density on prognosis in colorectal cancer and colorectal liver metastases. *Br. J. Cancer* **96**, 1112–1117 (2007).
39. Arandjelović, O. A new framework for interpreting the outcomes of imperfectly blinded controlled clinical trials. *PLoS One* **7**, e48984 (2012).
40. Osuala, R. & Arandjelović, O. Visualization of patient specific disease risk. In *Proc. IEEE International Conference on Biomedical and Health Informatics* 241–244, Orlando, Florida, USA (2017).
41. Li, J. & Arandjelović, O. Intuitive and interpretable visual communication of a complex statistical model of disease progression and risk. In *Proc. International Conference of the IEEE Engineering in Medicine and Biology Society* 4199–4202, (2017).
42. O'Neil, M. & Damjanov, I. Histopathology of colorectal cancer after neoadjuvant chemoradiation therapy. *Open Pathol. J.* **3**, 91–98 (2009).
43. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence—Volume 2*, IJCAI'95, 1137–1143 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995).
44. Barracliff, L., Arandjelović, O. & Humphris, G. Can machine learning predict healthcare professionals' responses to patient emotions? In *Proc. International Conference on Bioinformatics and Computational Biology* 101–106 (Honolulu, Hawaii, USA 2017).
45. Karsten, J. & Arandjelović, O. Automatic vertebrae localization from CT scans using volumetric descriptors. In *Proc. International Conference of the IEEE Engineering in Medicine and Biology Society* 576–579, (2017).
46. Nigri, E. & Arandjelović, O. Light curve analysis from Kepler spacecraft collected data. In *Proc. ACM International Conference on Multimedia Retrieval* 93–98, Bucharest, Romania (2017).
47. Beykikhoshk, A., Arandjelović, O., Phung, D., Venkatesh, S. & Caelli, T. Using Twitter to learn about the autism community. *Social. Netw. Anal. Min.* **5**, 5–22 (2015).
48. Birkett, C., Arandjelović, O. & Humphris, G. Towards objective and reproducible study of patient-doctor interaction: automatic text analysis based VR-CoDES annotation of consultation transcripts. In *Proc. International Conference of the IEEE Engineering in Medicine and Biology Society* 2638–2641, (2017).
49. Ling, C. X., Huang, J. & Zhang, H. Auc: A statistically consistent and more discriminating measure than accuracy. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, 519–524 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003).
50. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).
51. Bergstra, J. S., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for hyper-parameter optimization. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. & Weinberger, K. Q. (eds.) *Advances in Neural Information Processing Systems 24*, 2546–2554 (Curran Associates, Inc., 2011).
52. Hutter, F., Hoos, H. H. & Leyton-Brown, K. Sequential model-based optimization for general algorithm configuration. In *Proceedings of the 5th International Conference on Learning and Intelligent Optimization*, 507–523 (Springer-Verlag, 2011). [https://doi.org/10.1007/978-3-642-25566-3\\_40](https://doi.org/10.1007/978-3-642-25566-3_40).
53. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
54. Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Comput. & Electr. Eng.* **40**, 16–28 (2014).
55. Gurcan, M. N. et al. Histopathological image analysis: a review. *Ieee. Rev. Biomed. Eng.* **2**, 147–171 (2009).
56. Pudil, P., Novovičová, J. & Kittler, J. Floating search methods in feature selection. *Pattern Recognit. Lett.* **15**, 1119–1125 (1994).
57. Jain, A. & Zongker, D. Feature selection: evaluation, application, and small sample performance. *IEEE. Trans. Pattern Anal. Mach. Intell.* **19**, 153–158 (1997).



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018