Author(s): Petäinen, Liisa; Väyrynen, Juha P.; Ruusuvuori, Pekka; Pölönen, Ilkka; Äyrämö, Sami; Kuopio, Teijo

Title: Domain-specific transfer learning in the automated scoring of tumor-stroma ratio from histopathological images of colorectal cancer

Year: 2023

Version: Published version

Please cite the original version:

Petäinen, L., Väyrynen, J. P., Ruusuvuori, P., Pölönen, I., Äyrämö, S., & Kuopio, T. (2023). Domain-specific transfer learning in the automated scoring of tumor-stroma ratio from histopathological images of colorectal cancer. PLoS ONE, 18(5), Article e0286270. https://doi.org/10.1371/journal.pone.0286270

PLOS ONE

RESEARCH ARTICLE

Domain-specific transfer learning in the automated scoring of tumor-stroma ratio from histopathological images of colorectal cancer

Liisa Petäinen[1☯*], **Juha P. Väyrynen**[2☯], **Pekka Ruusuvuori**[3,4,5], **Ilkka Pölönen**[1], **Sami Äyrämö**[1‡], **Teijo Kuopio**[6,7,8‡]

1 Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland, **2** Cancer and Translational Medicine Research Unit, Medical Research Center, Oulu University Hospital, and University of Oulu, Oulu, Finland, **3** Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland, **4** Cancer Research Unit, Institute of Biomedicine, University of Turku, Turku, Finland, **5** FICAN West Cancer Centre, Turku University Hospital, Turku, Finland, **6** Department of Education and Research, Hospital Nova of Central Finland, Jyväskylä, Finland, **7** Department of Biological and Environmental Science, University of Jyväskylä, Jyväskylä, Finland, **8** Department of Pathology, Hospital Nova of Central Finland, Jyväskylä, Finland

☯ These authors contributed equally to this work.
‡ SÄ and TK also contributed equally to this work.
* lihesalo@jyu.fi

Citation: Petäinen L, Väyrynen JP, Ruusuvuori P, Pölönen I, Äyrämö S, Kuopio T (2023) Domain-specific transfer learning in the automated scoring of tumor-stroma ratio from histopathological images of colorectal cancer. PLoS ONE 18(5): e0286270. https://doi.org/10.1371/journal.pone.0286270

Editor: Chenchu Xu, Anhui University, CANADA

Received: January 3, 2023

Accepted: May 11, 2023

Published: May 26, 2023

Data Availability Statement: Central Finland Health Care District Ethics Committee has agreed to the publication of the test dataset. It has been published at Mendeley Data (DOI:10.17632/37t2d6xmy2.1). If this article is published in PLOS ONE, a reference will be added to the description of the dataset.

Funding: This study is one part of AI Hub Central Finland project that has received funding from the Council of Tampere Region (https://www.pirkanmaa.fi/en/) (Decision number: A75000) and

Abstract

Tumor-stroma ratio (TSR) is a prognostic factor for many types of solid tumors. In this study, we propose a method for automated estimation of TSR from histopathological images of colorectal cancer. The method is based on convolutional neural networks which were trained to classify colorectal cancer tissue in hematoxylin-eosin stained samples into three classes: *stroma*, *tumor* and *other*. The models were trained using a data set that consists of 1343 whole slide images. Three different training setups were applied with a transfer learning approach using domain-specific data i.e. an external colorectal cancer histopathological data set. The three most accurate models were chosen as a classifier, TSR values were predicted and the results were compared to a visual TSR estimation made by a pathologist. The results suggest that classification accuracy does not improve when domain-specific data are used in the pre-training of the convolutional neural network models in the task at hand. Classification accuracy for *stroma*, *tumor* and *other* reached 96.1% on an independent test set. Among the three classes the best model gained the highest accuracy (99.3%) for class *tumor*. When TSR was predicted with the best model, the correlation between the predicted values and values estimated by an experienced pathologist was 0.57. Further research is needed to study associations between computationally predicted TSR values and other clinicopathological factors of colorectal cancer and the overall survival of the patients.

Leverage from the EU 2014–2020, funded by European Regional Development Fund (ERDF) (https://ec.europa.eu/regional_policy/funding/erdf_en). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Deep learning (DL) has been the state-of-the-art medical image analysis technology for the last decade. It has been applied to various tasks also in digital pathology, e.g., tissue classification between normal and tumor tissues, defining tumor subtype, recognition (e.g. dividing cells) and segmentation (patch- or pixel-level segmentation) [1]. Tasks other than purely morphological have also been carried out, such as training DL models to predict certain genetic changes from hematoxylin-eosin (H&E)-stained histopathological sections without e.g. immunohistochemical staining [2–6]. These are relevant and meaningful efforts because the aforementioned tasks are time-consuming and expensive when carried out using manual laboratory methods [7].

A common challenge for developing artificial intelligence methods lies in the insatiable data hunger of DL algorithms. The lack of annotated data is also one of the most significant challenges for digital pathology [8]. Gaining a sufficient amount of high-quality training and testing data means hours of work for pathologists to annotate regions of interest to digitized images. One way to alleviate the data scarcity problem could be, e.g., transfer learning.

Transfer learning means utilizing a pre-trained neural network that has already learned a machine learning task in some domain that is not necessarily the same as in the target application. Transfer learning can be accomplished, e.g. using pre-trained ImageNet [9] neural network architecture that will provide the initial parameter values for the model. Although ImageNet model has been trained with images representing dogs, planes and houses, that are essentially very dissimilar to histopathological images, initializing weights of the convolutional neural network (CNN) model with ImageNet has been shown to increase prediction accuracy in medical imaging tasks [1, 10]. Sometimes a pre-trained model can also be available from the target domain. Some studies have shown that such a domain-specific pre-trained model may further improve prediction performance in the context of histopathological image analysis when compared to the ImageNet-initialization [10, 11]. In the medical domain, the deeper models have been shown to perform better as a feature extractor compared to shallow and linear models [12]. On the other hand, also lightweight models have performed well when transfer learning is applied, as seen in a study by Zhang et al. [13]. ImageNet-based transfer learning is a common approach in digital pathology in general [9, 14, 15], and it is also the most common transfer learning approach in DL models trained with colorectal cancer (CRC) histopathological images [16].

This study focuses on automating the estimation of TSR from histopathological images of CRC using transfer learning. CRC is the second most death-causing cancer in the world with over 900,000 deaths every year [17]. One prognostic factor for CRC is the proportion of stroma within the tumor site. It has been shown to associate with the survival of the patient in many solid cancer types. The low amount of stroma (TSR $\leq$ 50%) associates with a better prognosis [18–21].

Pathologists determine TSR visually by following a certain scoring protocol [22]. The main problem of this approach is the reproducibility of the TSR scoring. Overview by Van Pelt et al. [22] showed that the Cohen's kappa scores measuring the inter-observer agreement of visual TSR using binary scoring (TSR $>$ 50% = stroma-high and TSR $\leq$ 50% = stroma-low) ranged from 0.60 to 0.89. Automated TSR estimation may improve reproducibility, but it is a challenging task and a new relatively new concept.

Automated estimation of TSR begins by tiling a histopathological whole-slide image (WSI) into smaller image patches. After that TSR can be predicted by classifying the patches and calculating the proportion of tumor and stroma. Another approach is to use a particular spot that is a smaller part of the WSI selected by a pathologist to calculate TSR. Both approaches have been applied in previous studies [23–28].

Sirinukuvattana et al. [23, 24] automated the TSR estimation using a CNN model trained for nuclei detection. They trained a model based on VGG19 [29] architecture to classify nine tissue types and the accuracies for detecting stroma and tumor were 90.4% and 96.0%, respectively. In contrast to other TSR-related studies, their results did not show prognostic value for TSR [25–27].

Zhao et al. [26] proposed a nine-class CNN model using transfer learning for which overall classification accuracies on two test sets were 95.7% and 97.5%. Classification accuracies for tumor and stroma were 92.8% and 70.9% on the test set 1 that was published by Kather et al. [30]. The test set 2 was a random sample from images collected at Yunnan Cancer Hospital. Classification accuracies for tumor and stroma on the test set 2 were 97.2% and 89.1%. Zhao et al. used pathologists annotations as ground truth for TSR estimation on 126 image blocks of size 1 μm^2, the predicted tumor and stroma areas were in high agreement with pathologists' annotations (Pearson r = 0.939, 95% CI 0.914—0.957). When splitting the patient cohort into two categories stroma-low (TSR $<$ 48.8%) and stroma-high (TSR$\geq$48.8%) based on the TSR results of their model, TSR was shown to be an independent prognostic factor in the overall survival of CRC patient.

Instead of using the whole slide as an input for a neural network model, an alternative approach is to calculate TSR from the same circular spot where the visual TSR estimation takes place. The downside of this approach is the manual effort needed to point the spot for the machine learning model. Geessink et al. [28] developed an 11-layer VGG neural network on 129 patients to classify nine tissue types. Using 50% cutoff-value between stroma-high and stroma-low categories there was a considerable disagreement between the model and pathologist (Cohen's kappa $\kappa = 0.239$). Cohen's kappa score was slightly improved ($\kappa = 0.521$) using the median of the TSR values estimated with the model as a cutoff value for stroma-high and stroma-low. Also, stroma-high- and stroma-low grouping showed a strong prognostic value when the median was used as a cutoff value.

In the present study, 12 DL models for estimating TSR for CRC samples were developed and tested using three different transfer learning setups, four different pre-trained CNN architectures, and two distinct CRC data sets. Models were trained to classify image patches into three different tissue classes: *tumor*, *stroma* and *other*. Three best-performing CNN models were chosen to estimate TSR from WSIs in a separate TSR test set. The predicted TSR values were then compared with the pathologist's TSR estimates.

Materials and methods

Data

In this study, two mutually independent CRC data sets were used: a CRC data set from Central Finland Health Care District ("CFHCD-data") and a public CRC data set ("NCT-CRC-HE-100K") [30].

CFHCD-data consists of 1343 patients of primary colorectal cancers (stages I-IV) with one WSI from each patient. The cohort is described in detail by Elomaa et al. [31]. The WSIs were scanned with Hamamatsu NanoZoomer-XR with resolution of 0.5 microns per pixel (MPP).

Automated estimation of the TSR-value for a single WSI is based on calculating the ratio of patches representing stroma and tumor classes. The estimation process consists of two steps: 1) CNN-classification of patches for a WSI 2) Calculation of the ratio of tumor and stroma patches.

For developing and testing the models, CFHCD-data were split into two disjoint sets at random. The first one (Dev-set) consisted of 169 WSIs and was used in developing CNN

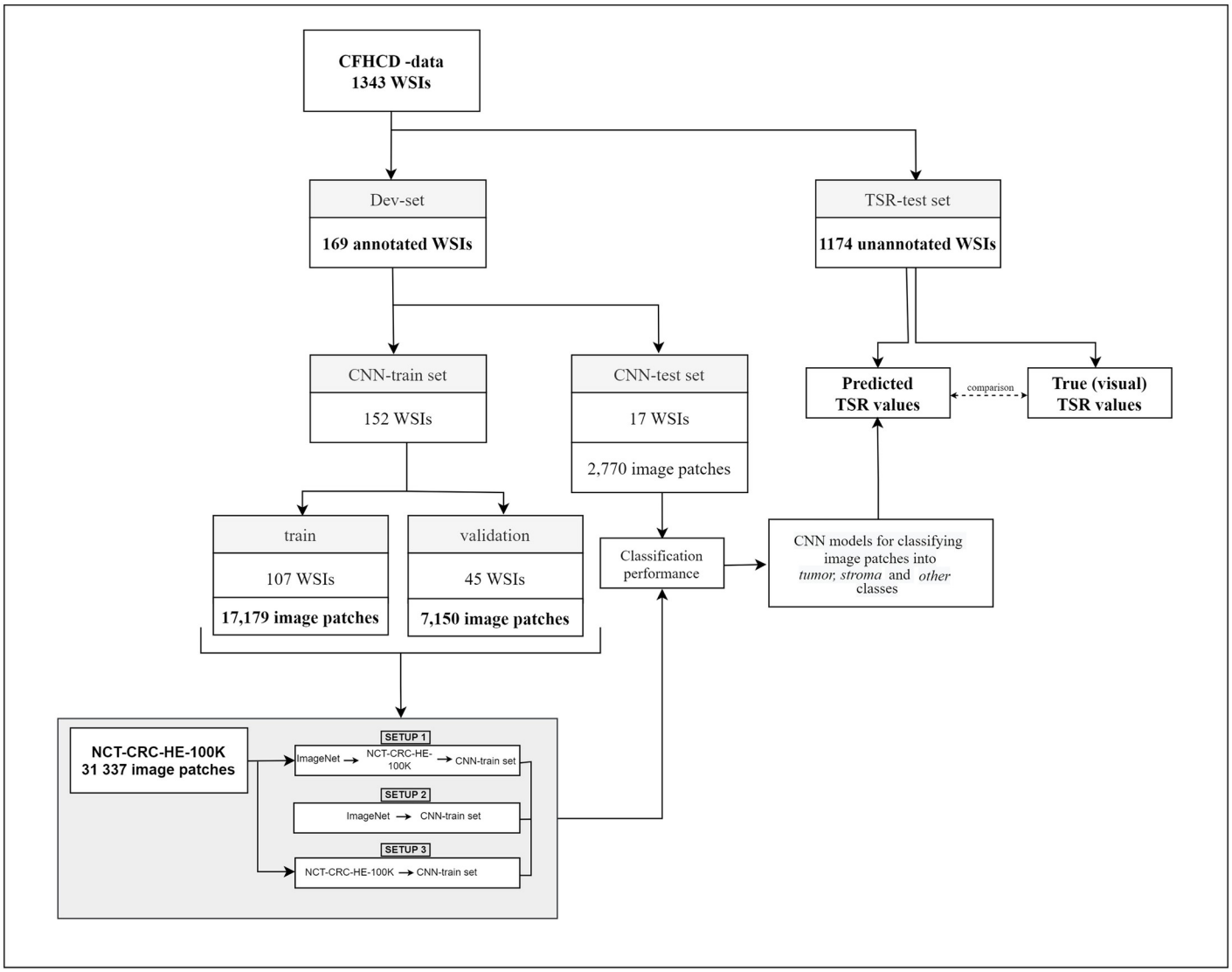

Fig 1. Diagram of the studyflow. Patches for training were tiled from 152 annotated WSIs (CNN-train), the classifier test set of 17 WSIs (CNN-test) was excluded from the training data. TSR-values were predicted for the remaining 1174 WSIs (TSR-test) and compared with the TSR-values determined visually by a pathologist.

https://doi.org/10.1371/journal.pone.0286270.g001

classification models and the second one (TSR-test set) consisted of 1174 WSIs and it was applied to test CNN-based estimation of the TSR-value. The overall study flow is presented in Fig 1.

Before tiling and preprocessing, Dev-set was annotated by an experienced pathologist into three different tissue categories, *tumor*, *stroma* and *other*. The annotations were accomplished with QuPath image analysis tool [32]. The class *other* includes debris, lymphocytes, mucus, normal epithelium and smooth muscle (See Fig 2). Ground truth TSR-values (from now on referred to as true TSR-values) were visually assessed by the pathologist for the TSR-test set following the protocol by Van Pelt et al. [22].

A subsample (n = 31,337) of NCT-CRC-HE-100K data set were used to investigate the effects of domain-specific pre-training of the CNN models. The original data includes 100,000 patches tiled from 86 WSIs (0.5 MPP). The patches are pre-labeled to nine tissue classes:

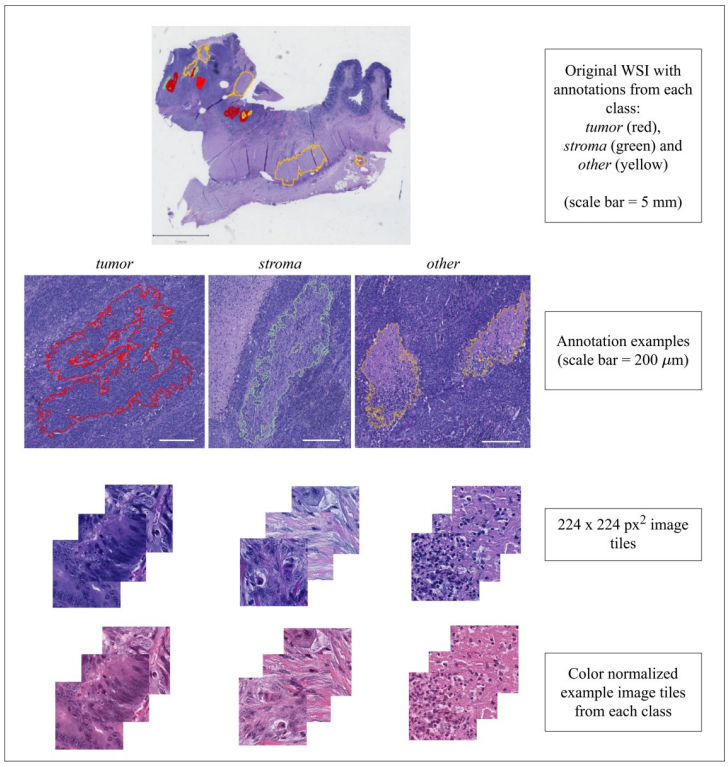

Fig 2. Annotation examples. Examples of annotations from each class. The patches were extracted from annotated areas, and they were color normalized using Macenko's method [33].

https://doi.org/10.1371/journal.pone.0286270.g002

adipose, debris, lymphocytes, mucus, normal, smooth muscle, stroma and tumor. In this study, only patches representing *stroma* (10,446 patches) and *tumor* (10,446 patches) classes were used as such, whereas a random sample of 10,445 patches from debris, lymphocytes, mucus, normal and smooth muscle classes with even distribution was drawn and assigned to a class called *other*.

Tiling and preprocessing of WSIs

Following the annotation of the WSIs in Dev-set they were tiled into smaller WSI patches (224 × 224 pixels / 101 μm × 101 μm). Tiling was performed with a sliding window procedure with 64 pixels overlapping. If less than 75% of the patch area included annotated pixels, the patch was discarded from further analyses. All patches were color normalized by Macenko's method [33].

Since the classifiers were not trained to detect the image background and adipose tissue, the WSIs in the TSR-test set were tiled in the following way: First, to remove image background and adipose tissue, a binary mask was applied using Otsu's algorithm for choosing the optimal threshold value [34]. After this the image patches were tiled using a sliding window procedure with no overlapping. If the masked area was less than 75%, the patch was discarded from further analyses. Macenko's method was applied for color normalization of the patches [33].

Training and evaluation of classifiers

For training and selecting the best CNN models Dev-set was split at random into two distinct WSI sets of CNN-train (n = 152) and CNN-test (n = 17) with the constraints that after the

tiling process the maximum difference of eighty patches in the training distribution of the target classes was allowed and the minimum number of patches was nine hundred in each test class. The constraints were applied to ensure approximately balanced class distribution in CNN-train and that the size of CNN-test is at least 10% of Dev-set.

This resulted in 24,329 and 2,770 input patches for training and testing of the classifiers, respectively. The final numbers of patches in the CNN-train/CNN-test classes were 8139/900, 8060/900, and 8130/900 for *tumor*, *stroma* and *other*, respectively. Examples of image patches from each class are shown in Fig 3.

For predicting TSR in WSIs, four convolutional neural network (CNN) architectures, AlexNet [35], GoogleNet [36], ResNet50 [37], VGG19 [29], were trained on CNN-train to classify the patches into three tissue types: *tumor*, *stroma* and *other*. For all the CNN models, the input layer was set according to the WSI patch dimensions 224 × 224, and the output layer was softmax function (dimension = 3).

Three different pre-training strategies for CNN models were applied (SETUP-1, SETUP-2 and SETUP-3). In SETUP-1 the CNN model was first initialized with ImageNet [9] weights and thereafter pre-trained with the domain-specific subsample of NCT-CRC-HE-100K-data before finalizing the training with CNN-train. In SETUP-2 training on CNN-train started directly from ImageNet weights. In SETUP-3 the CNNs were first initialized with random weights using PyTorch default parameters and then subsequently pre-trained with the subsample of NCT-CRC-HE-100K-data and finalized with CNN-train.

For both pretraining the CNN models on the NCT-CRC-HE-100K patches as well as fine-tuning the models on CFHCD-data (CNN-train and CNN-test), the most suitable values for hyperparameters were selected using 5-fold cross-validation (C-V) CNN-train. In 5-fold C-V data are split at random into the five subsets. Then each subset acts once as a test fold (set) while the four other are used for training. The error estimate for a model is the average error over the test folds.

After finding the best hyperparameter values for the neural networks, the optimal number of epochs was determined with an early stopping model selection strategy on CNN-train using approximately 1/3 of the training patches for validation. The final model was then trained once more on the full CNN-train until the optimal number of epochs was reached.

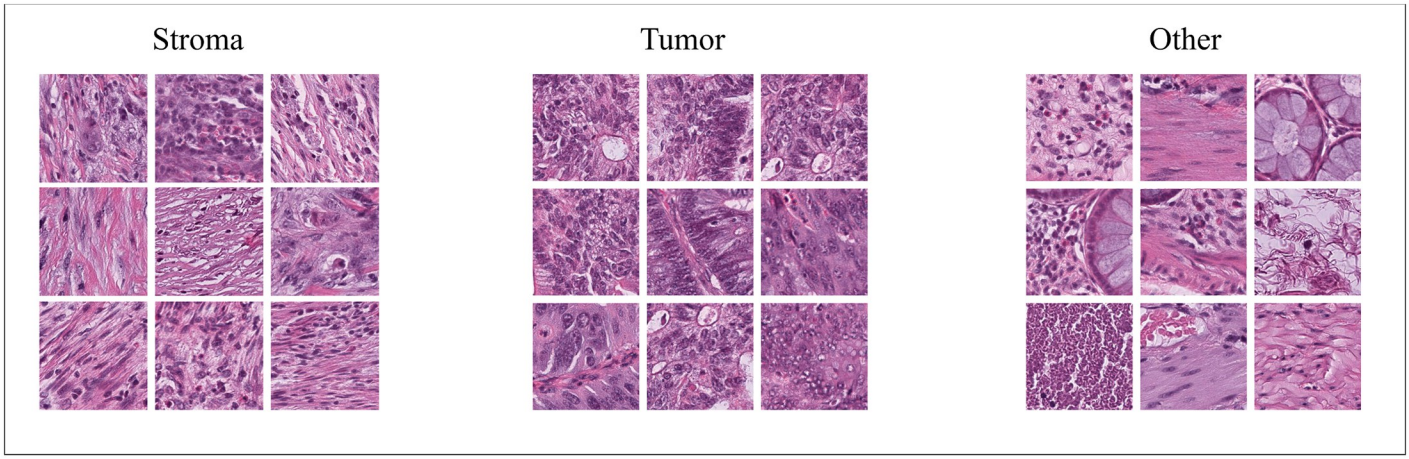

Fig 3. Example patches from each class. The image patches were taken from annotated areas and the number of patches was balanced between the groups. The *other* class includes, e.g., debris, normal epithelium and smooth muscle.

https://doi.org/10.1371/journal.pone.0286270.g003

The generalization performance of each CNN model was then assessed by computing the classification accuracy, precision, recall and F_1-score of the models on CNN-test (2,770 CFHCD-patches).

For more information about chosen hyperparameters, see S1 Table.

Calculating tumor-stroma ratio

TSR was calculated for each WSI in TSR-test as the ratio of patches classified by a CNN model as *stroma* and *tumor*:

$$TSR = \frac{n_{stroma}}{n_{tumor} + n_{stroma}}, \quad (1)$$

where n_{stroma} and n_{tumor} are the number of *stroma* and *tumor* patches, respectively.

Equipment and software

All the neural network models were trained on Linux GPU server Tesla P100, x 86_64 with Python-version 3.8.5. using PyTorch- and TorchVision-libraries, versions 1.9.0 and 0.10.0, respectively. OpenCV 4.5.2 was utilized when masking the WSIs. Color normalization was applied with an open-source library StainTools, available for download at GitHub: https://github.com/Peter554/StainTools. Performance metrics were calculated with Scikit-learn 0.24.2 metrics-module.

Results

Validation and test results of the final models

Classification accuracy metrics on the CNN-test set are shown for all the final CNN models in Table 1. The three highest values were obtained by training the models directly from Imagenet-pretrained weights. The differences between the three best CNN architectures are negligible. Only Alexnet showed slightly poorer performance in terms of classification accuracy. To support the interpretation of the results also validation accuracies are reported in Table 2.

A more detailed comparison between true and predicted classifications on CNN-test can be seen for the most accurate top-3 models in Fig 4. All the models performed well in detecting the *tumor* class but had slight difficulties distinguishing between classes *stroma* and *other*.

Precision, recall and F_1-score are shown in Table 3. The numbers show that all the top-3 models attained excellent hit rate (recall > 0.9) for all classes as well as they are accurate in predicting positive cases (precision > 0.9). Consequently all the models attained high F1-score (> 0.9). The observed differences between the top-3 models are so small that they can most likely be explained by random variation.

Table 1. Test accuracies.

	SETUP 1 accuracy	SETUP 2 accuracy	SETUP 3 accuracy
Alexnet	91.95%	92.42%	91.79%
Googlenet	94.94%	**95.40%**	95.37%
ResNet50	94.94%	**96.09%**	92.57%
VGG19	95.19%	**95.65%**	92.64%

Test accuracies of all final models on the CNN-test set, top-3 models are shown bold.

https://doi.org/10.1371/journal.pone.0286270.t001

Table 2. Validation accuracies.

	SETUP 1 accuracy	SETUP 2 accuracy	SETUP 3 accuracy
Alexnet	90.19%	92.90%	92.30%
Googlenet	93.35%	92.19%	**93.69%**
ResNet50	**93.39%**	92.96%	91.49%
VGG19	**94.17%**	93.29%	91.11%

Validation accuracies for all final CNN classification models, top-3 models are shown bold.

https://doi.org/10.1371/journal.pone.0286270.t002

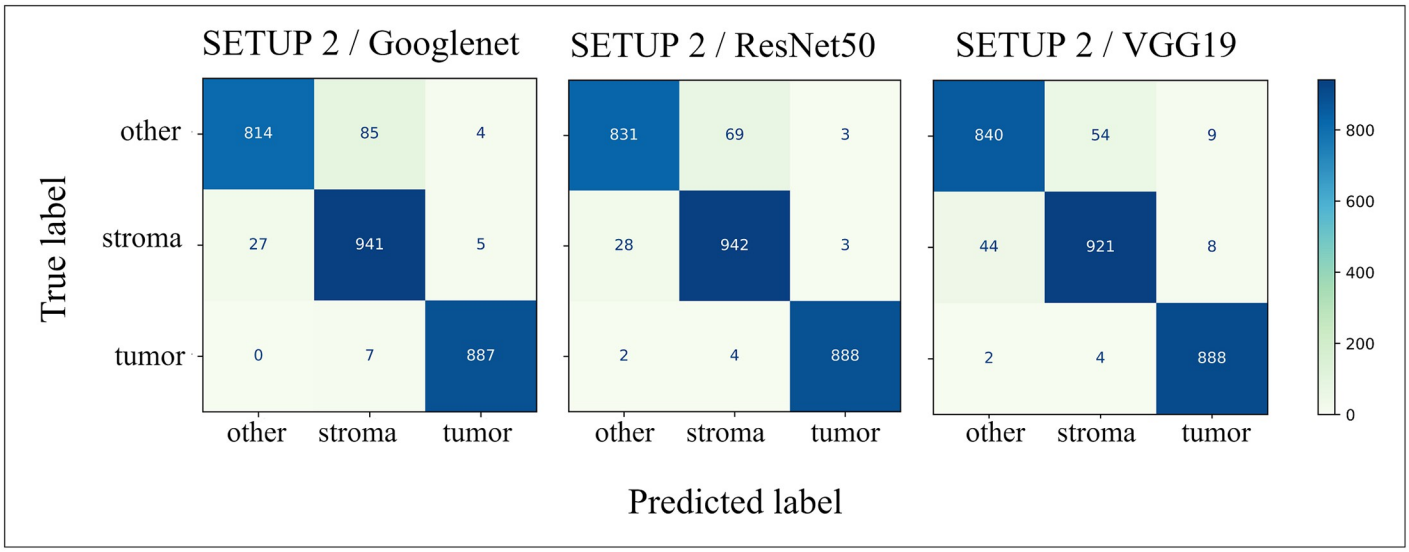

Fig 4. Confusion matrices. Classification results of top-3 models on the CNN-test set.

https://doi.org/10.1371/journal.pone.0286270.g004

Table 3. Precision, recall and F_1-score.

model	class	precision	recall	F_1-score
SETUP 2 / Googlenet	other	0.97	0.90	0.93
	stroma	0.91	0.97	0.94
	tumor	0.99	0.99	0.99
SETUP 2 / ResNet50	other	0.97	0.92	0.94
	stroma	0.93	0.97	0.95
	tumor	0.99	0.99	0.99
SETUP 2 / VGG19	other	0.95	0.93	0.94
	stroma	0.94	0.95	0.94
	tumor	0.98	0.99	0.99

Precision, recall and F_1-score on the CNN-test set of top-3 models. The CNN-test set included 2770 image tiles.

https://doi.org/10.1371/journal.pone.0286270.t003

Tumor-stroma ratio predictions

Distributions of the predicted TSR values on the TSR-test set are shown for the top-3 models in Fig 5. Examples of the most accurate and least accurate TSR predictions on WSIs shown in S1 Fig.

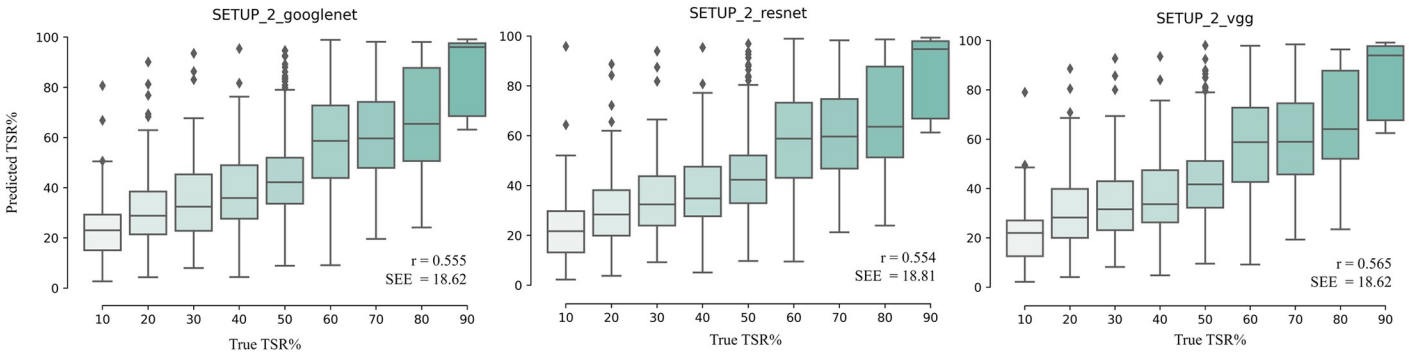

Fig 5. Results from TSR prediction: Boxplots. Boxplots showing the distributions of the predicted TSR values and Pearson correlation coefficient (r) of predicted and true TSR values for the top-3 models on the TSR-test set. The results are shown for each visually estimated TSR value category on the discrete scale {10%, 20%, . . ., 90%}. The standard error of the estimate (SEE) is the overall SEE of all categories.

https://doi.org/10.1371/journal.pone.0286270.g005

Mean, median, standard error of the estimate (SEE), standard deviation (std) of predicted TSR values and the number of WSIs in each category are shown in Table 4. The results show that all the models overestimate the TSR value in the three lowest target categories (10%, 20% and 30%) and underestimate in the remaining target categories. All the models perform best in the category 60% (the mean differences between the predicted and true TSR values 1.4) whereas the poorest performance is observed in the categories 10% and 80% (the mean differences between the predicted and true TSR values 14.4 and 14.2 respectively). The smallest

Table 4. TSR results: Statistics.

SETUP 2 / Googlenet										
True TSR		**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
Predicted TSR	mean	24.6	32.0	35.4	38.4	44.1	58.5	60.5	65.8	86.4
	median	23.0	28.9	32.4	35.9	42.2	58.6	59.6	65.4	96.0
	SEE	20.7	20.5	16.8	15.1	17.2	19.3	21.3	25.0	14.7
	std	14.9	16.6	16.0	15.1	16.1	19.3	19.1	20.8	14.8
	n	53	95	105	200	284	229	143	51	13
SETUP 2 / ResNet50										
True TSR		**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
Predicted TSR	mean	24.7	31.3	35.5	38.5	44.4	58.6	60.6	66.0	86.0
	median	21.7	28.3	32.4	34.8	42.3	58.8	59.6	63.6	94.7
	SEE	21.9	20.2	17.0	15.4	17.4	19.4	21.4	24.8	15.5
	std	16.3	16.8	16.2	15.3	16.5	19.4	19.4	20.7	15.6
	n	53	95	105	200	284	229	143	51	13
SETUP 2 / VGG19										
True TSR		**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
Predicted TSR	mean	23.3	31.0	34.4	37.1	43.1	57.7	60.1	65.9	85.7
	median	22.0	28.3	31.6	33.6	41.7	58.8	58.9	64.1	94.0
	SEE	19.0	19.6	16.5	15.1	17.4	19.4	21.7	25.1	14.8
	std	13.7	16.3	16.0	14.9	16.0	19.3	19.4	21.0	14.8
	n	53	95	105	200	284	229	143	51	13

Main statistics from top-3 models on the TSR-test set. The results are shown based on the visually estimated TSR-value on the discrete scale {10%, 20%, . . ., 90%}. The overall SEE of all categories combined shown in Fig 5.

https://doi.org/10.1371/journal.pone.0286270.t004

SEEs are observed in categories 30%, 40% and 90%. When grouping the TSR values into stroma-high (TSR $>$ 50%) and stroma-low (TSR $\leq$ 50%), the Cohen's kappa scores between the true and predicted TSR values were 0.32 (SETUP 2 / Googlenet), 0.33 (SETUP 2 / ResNet50) and 0.33 (SETUP 2 / VGG19).

Discussion

The aim of this study was to investigate how accurately CNN-based machine learning models can predict the ratio of tumor and stroma tissue in WSI samples. For the best model (ResNet50 architecture) the correlation between the true and predicted TSR values was 0.57 (SEE = 18.6). Approximately the same performance was obtained with GoogleNet and VGG19 architectures. Utility of domain-specific pre-training of CNN models was also investigated, but no meaningful differences were observed. The results show that the same or even better performance with comparable computational cost can be achieved in the present task without domain-specific pre-training of CNN.

Even though the automated TSR were predicted from the whole tissue area in contrast to the pathologist who selected a small spot for estimation, their outcomes correlates (r = 0.57) rather well. Cohen's kappa score (κ = 0.33) for stroma-high and stroma-low classification was comparable with results from the previous study by Geessink et al. [28] where Cohen's kappa score for stroma-high and stroma-low with 50% cutoff-value was 0.24. An overview by [22] show that the inter-observer kappa-values variate between 0.60 to 0.89 when pathologists estimate the TSR of CRC binary into stroma-high and stroma-low.

In this study SEE was lowest in TSR categories 30%, 40% and 90%. When comparing the means of the predicted TSR values with the true TSR values, the best performing set of images was the one with TSR 60% and the second largest number of samples. Balancing the data for TSR prediction task might bring more consistency to the results.

The results also indicate that the most difficult part of the classification task is to separate *stroma* and *other*. This can be due to the visual similarity of smooth muscle and fibrotic stroma. This may cause the classifier to make a mistake since the smooth muscle tissue belongs to the *other* class. The smooth muscle and stroma are difficult to separate even by the human eye. Despite the minor weaknesses, the classification accuracy of the best CNN models was comparable to previous studies [25, 26]. The classification accuracy for the *tumor* tissue was over 98% with all top-3 models.

Despite the promising performance of machine learning bringing some aspects of the visual estimation procedure could improve the automated models. For example, only tumor-related stroma tiles could be taken into account. Another option could be mimicking the visual human process by going through the image frame by frame and choosing one spot for making the final TSR prediction. In addition to these, using a smaller tile size might increase classifying accuracy since some tumor areas, as well as stromal areas in between, seem to be quite narrow. This could have a significant effect on the quality of TSR predictions.

Even though accuracy of the proposed automated method for TSR estimation does not fully compare to human visual analysis, the reproducibility of computational model outcomes is a major advantage. Automated machine learning based tools would bring reproducibility to daily practices and the TSR estimation process, in particular. Moreover, TSR estimation with an automated machine learning model can be completed in a fraction of the time compared to the visual method.

This study has some limitations. Firstly, the aim of this research was to develop models on the Finnish population and, therefore, their generalizability to other populations can not be guaranteed without further research. When interpreting the results, it is important to consider

the size of test data. All the test results are, however, produced by trying the models only once on the independent random test set which guarantees that the models were not overfit to the test samples. In order to take the models into clinical use more extensive external tests are needed.

As visually estimated TSR has been shown to be an independent prognostic factor in solid cancer types [18–21], further studies should take place for assessing the correlation of the machine learning predicted TSR values with other clinicopathological factors and the overall survival of patients. If a correlation is found, the model should be optimized to perform better in terms of generalizability. Also a "hotspot"-analysis of TSR, in which a spot where to estimate the TSR would be manually chosen, should be considered as it would ease the computational burden of DL based method and thereby improve the prediction accuracy.

Supporting information

S1 Table. Hyperparameters. Chosen parameters in the final training phase of all models. LR refers to the learning-rate and optimizer is the optimization-function used. Parameters were chosen using 5-fold cross-validation. Loss-function was cross-entropy for all setups.
(TEX)

S1 Fig. TSR prediction examples. Examples of TSR prediction on four WSIs.
(TIF)

Acknowledgments

We thank Central Finland Health Care District and "Suolisyöpä Keski-Suomessa 2000–2015" -project for providing the data set.

Author Contributions

Conceptualization: Liisa Petäinen, Juha P. Väyrynen, Teijo Kuopio.

Data curation: Liisa Petäinen, Juha P. Väyrynen.

Formal analysis: Liisa Petäinen, Pekka Ruusuvuori, Ilkka Pölönen, Sami Äyrämö.

Funding acquisition: Sami Äyrämö.

Investigation: Liisa Petäinen, Juha P. Väyrynen.

Methodology: Liisa Petäinen, Pekka Ruusuvuori, Ilkka Pölönen.

Project administration: Sami Äyrämö.

Resources: Sami Äyrämö, Teijo Kuopio.

Software: Liisa Petäinen.

Supervision: Pekka Ruusuvuori, Ilkka Pölönen, Sami Äyrämö, Teijo Kuopio.

Validation: Liisa Petäinen.

Visualization: Liisa Petäinen.

Writing – original draft: Liisa Petäinen.

Writing – review & editing: Juha P. Väyrynen, Pekka Ruusuvuori, Ilkka Pölönen, Sami Äyrämö, Teijo Kuopio.

References

1. Litjens G. A survey on deep learning in medical image analysis. Medical image analysis. 2017; 42:60–88. https://doi.org/10.1016/j.media.2017.07.005 PMID: 28778026
2. Chang P, Grinband J, Weinberg B, Bardis M, Khy M, Cadena G, et al. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. American Journal of Neuroradiology. 2018; 39(7):1201–1207. https://doi.org/10.3174/ajnr.A5667 PMID: 29748206
3. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature medicine. 2018; 24(10):1559–1567. https://doi.org/10.1038/s41591-018-0177-5 PMID: 30224757
4. Schaumberg AJ, Rubin MA, Fuchs TJ. H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer. BioRxiv. 2018; p. 064279.
5. Schmauch B, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nature communications. 2020; 11 (1):1–15. https://doi.org/10.1038/s41467-020-17678-4 PMID: 32747659
6. Wang Y, Kartasalo K, Valkonen M, Larsson C, Ruusuvuori P, Hartman J, et al. Predicting molecular phenotypes from histopathology images: a transcriptome-wide expression-morphology analysis in breast cancer. arXiv preprint arXiv:200908917. 2020;.
7. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. British journal of cancer. 2021; 124(4):686–696. https://doi.org/10.1038/s41416-020-01122-x PMID: 33204028
8. Tizhoosh HR, Pantanowitz L. Artificial intelligence and digital pathology: challenges and opportunities. Journal of pathology informatics. 2018; 9(1):38. https://doi.org/10.4103/jpi.jpi_53_18 PMID: 30607305
9. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015; 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y
10. Mormont R, Geurts P, Marée R. Comparison of deep transfer learning strategies for digital pathology. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2018. p. 2262–2271.
11. Sharmay Y, Ehsany L, Syed S, Brown DE. HistoTransfer: Understanding Transfer Learning for Histopathology. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE; 2021. p. 1–4.
12. Kim HE, Cosa-Linan A, Santhanam N, Jannesari M, Maros ME, Ganslandt T. Transfer learning for medical image classification: A literature review. BMC medical imaging. 2022; 22(1):69. https://doi.org/10.1186/s12880-022-00793-7 PMID: 35418051
13. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 4510–4520.
14. Bayramoglu N, Heikkilä J. Transfer learning for cell nuclei classification in histopathology images. In: European Conference on Computer Vision. Springer; 2016. p. 532–539.
15. Kieffer B, Babaie M, Kalra S, Tizhoosh HR. Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In: 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA). IEEE; 2017. p. 1–6.
16. Davri A, Birbas E, Kanavos T, Ntritsos G, Giannakeas N, Tzallas AT, et al. Deep Learning on Histopathological Images for Colorectal Cancer Diagnosis: A Systematic Review. Diagnostics. 2022; 12(4):837. https://doi.org/10.3390/diagnostics12040837 PMID: 35453885
17. World Health Organization W. Cancer: Key Facts. Cancer. 2020;.
18. West N, Dattani M, McShane P, Hutchins G, Grabsch J, Mueller W, et al. The proportion of tumour cells is an independent predictor for survival in colorectal cancer patients. British journal of cancer. 2010; 102 (10):1519–1523. https://doi.org/10.1038/sj.bjc.6605674 PMID: 20407439
19. Huijbers A. The proportion of tumor-stroma as a strong prognosticator for stage II and III colon cancer patients: validation in the VICTOR trial. Annals of oncology. 2013; 24(1):179–185. https://doi.org/10.1093/annonc/mds246 PMID: 22865778
20. Ma W. Tumor-stroma ratio is an independent predictor for survival in esophageal squamous cell carcinoma. Journal of Thoracic Oncology. 2012; 7(9):1457–1461. https://doi.org/10.1097/JTO.0b013e318260dfe8 PMID: 22843085
21. Mesker WE. The carcinoma–stromal ratio of colon carcinoma is an independent factor for survival compared to lymph node status and tumor stage. Analytical Cellular Pathology. 2007; 29(5):387–398. https://doi.org/10.1155/2007/175276 PMID: 17726261

22. Van Pelt GW. Scoring the tumor-stroma ratio in colon cancer: procedure and recommendations. Virchows Archiv. 2018; 473(4):405–412. https://doi.org/10.1007/s00428-018-2408-z PMID: 30030621

23. Sirinukunwattana K, Raza SEA, Tsang YW, Snead D, Cree I, Rajpoot N. A spatially constrained deep learning framework for detection of epithelial tumor nuclei in cancer histology images. In: International Workshop on Patch-based Techniques in Medical Imaging. Springer; 2015. p. 154–162.

24. Sirinukunwattana K, Snead D, Epstein D, Aftab Z, Mujeeb I, Tsang YW, et al. Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer. Scientific reports. 2018; 8(1):1–13. https://doi.org/10.1038/s41598-018-31799-3 PMID: 30209315

25. Kather JN, Krisam J, Charoentong P, Luedde T, Herpel E, Weis CA, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS medicine. 2019; 16(1):e1002730. https://doi.org/10.1371/journal.pmed.1002730 PMID: 30677016

26. Zhao K. Artificial intelligence quantified tumour-stroma ratio is an independent predictor for overall survival in resectable colorectal cancer. EBioMedicine. 2020; 61:103054. https://doi.org/10.1016/j.ebiom.2020.103054 PMID: 33039706

27. Park JH, Richards CH, DC M. The relationship between tumour stroma percentage, the tumour microenvironment and survival in patients with primary operable colorectal cancer. Ann Oncol. 2014; 25:644–651. https://doi.org/10.1093/annonc/mdt593 PMID: 24458470

28. Geessink OG, Baidoshvili A, Klaase JM, Bejnordi BE, Litjens GJ, van Pelt GW, et al. Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer. Cellular Oncology. 2019; 42(3):331–341. https://doi.org/10.1007/s13402-019-00429-z PMID: 30825182

29. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. 2014;.

30. Kather JN, Halama N, Marx A. 100,000 histological images of human colorectal cancer and healthy tissue; 2018.

31. Elomaa H, Ahtiainen M, Väyrynen SA, Ogino S, Nowak JA, Friman M, et al. Prognostic significance of spatial and density analysis of T lymphocytes in colorectal cancer. British Journal of Cancer. 2022; p. 1–10. https://doi.org/10.1038/s41416-022-01822-6 PMID: 35449453

32. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: Open source software for digital pathology image analysis. Scientific reports. 2017; 7(1):1–7. https://doi.org/10.1038/s41598-017-17204-5 PMID: 29203879

33. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, et al. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE international symposium on biomedical imaging: from nano to macro. IEEE; 2009. p. 1107–1110.

34. Otsu N. A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics. 1979; 9(1):62–66. https://doi.org/10.1109/TSMC.1979.4310076

35. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012; 25:1097–1105.

36. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 1–9.

37. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.