Interpretable Deep Neural Network to Predict Estrogen Receptor Status from Haematoxylin-Eosin Images

Philipp Seegerer[1], Alexander Binder[2](✉), René Saitenmacher[1], Michael Bockmayr[3,6], Maximilian Alber[3], Philipp Jurmeister[3], Frederick Klauschen[3], and Klaus-Robert Müller[1,4,5]

[1] Machine Learning Group, Technical University Berlin, Berlin, Germany
[2] Singapore University of Technology and Design (SUTD), Singapore, Singapore
alexander_binder@stud.edu.sg
[3] Institute of Pathology, Charité University Hospital, Berlin, Germany
[4] Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea
[5] Max-Planck-Institute for Informatics, Campus E1 4, Saarbrücken, Germany
[6] Department of Pediatric Hematology and Oncology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Abstract. The eligibility for hormone therapy to treat breast cancer largely depends on the tumor's estrogen receptor (ER) status. Recent studies show that the ER status correlates with morphological features found in Haematoxylin-Eosin (HE) slides. Thus, HE analysis might be sufficient for patients for whom the classifier confidently predicts the ER status and thereby obviate the need for additional examination, such as immunohistochemical (IHC) staining. Several prior works are limited by either the use of engineered features, multi-stage models that use features unspecific to HE images or a lack of explainability. To address these limitations, this work proposes an end-to-end neural network ensemble that shows state-of-the-art performance. We demonstrate that the approach also translates to the prediction of the cancer grade. Moreover, subsets can be selected from the test data for which the model can detect a positive ER status with a precision of 94% while classifying 13% of the patients. To compensate for the reduced interpretability of the model that comes along with end-to-end training, this work applies Layer-wise Relevance Propagation (LRP) to determine the relevant parts of the images a posteriori, commonly visualized as a heatmap overlayed with the input image. We found that nuclear and stromal morphology and lymphocyte infiltration play an important role in the classification of the ER status. This demonstrates that interpretable machine learning can be a vital tool for validating and generating hypotheses about morphological biomarkers.

Keywords: Digital pathology · Deep learning · Explainable AI

A. Holzinger et al. (Eds.): AI & ML for Digital Pathology, LNAI 12090, pp. 16–37, 2020.
https://doi.org/10.1007/978-3-030-50402-1_2

1 Introduction and Motivation

Determining the estrogen receptor (ER) status of a tumor is of high clinical importance for the management of breast cancer patients because it specifies the eligibility for hormone therapy. So far, the gold standard for determining the ER status is immunohistochemical (IHC) staining. However, recent work shows that the ER status also correlates with morphological features found in Haematoxylin-Eosin (HE) slides even though these features are hardly apparent to pathologists [10,32,36]. This would make an additional IHC staining unnecessary if the ER status can already be determined from the HE stain with high confidence and therefore save both time and budget in clinical routine. Furthermore, it can provide a valuable “second opinion” in assessing the ER status, especially if one considers that up to 20% of IHC-based ER and progesterone status assessments might be inaccurate [14]. Nevertheless, it remains unclear what features are used by the learning machine to determine the ER status; analysis was so far limited to low-resolution heatmaps that indicated certain importance of stromal regions [36].

To address this, our work presents a novel end-to-end deep neural network (DNN) ensemble trained on pooled random patches that shows competitive performance compared with prior work [10,32,36] for the prediction of the ER status and pathological cancer grade in terms of area under the ROC curve (AUC). Based on the validation data, a minimum classifier confidence can be chosen below which a sample is rejected from the classification; by doing so, the model reaches a high precision (94%) while still classifying a considerable amount (13%) of patients. For these patients, in principle no additional staining would be needed.

A disadvantage of deep end-to-end learning can be the lack of interpretability of the model. We try to alleviate this issue by analyzing relevance heatmaps of model predictions. This showed that the model mostly relies on stromal texture and nuclear features to identify ER-positive samples and demonstrate how explanation methods can be used to not only verify machine learning models but also to validate and generate biomedical hypotheses. Moreover, by using these explanation methods one can validate whether a trained model relies on unstable features, such as discussed in [26].

2 Glossary

Breast Cancer Grading: Morphological assessment of a breast cancer that measures the degree of differentiation. It consists of the evaluation of the percentage of tubuli formation, the degree of nuclear pleomorphism and the number of mitoses. These three morphological features are assigned a score in a semi-quantitative way and then combined to an aggregated cancer grade score, that can be 1, 2 or 3 [12].

End-to-End Learning: Opposed to the classical pattern recognition pipeline, where features are usually chosen by the developer and only the classifier is learned, end-to-end learning means that everything from the input data to the classifier output, i.e. including the features, is learned from the data. A very successful instance of end-to-end learning are deep neural networks.

Estrogen Receptor Status: This is an important parameter for prognosis and the prediction of therapy response in breast cancer. A lack of ER receptor is correlated with a higher rate of recurrence and shorter survival as well as a decreased probability of responding to endocrine therapy [30].

Immunohistochemical Staining: The different immunohistochemical stains allow the detection and localization of chemical compounds, e.g. proteins, by binding marked antibodies to them. Amongst others, there are stains to detect hormone receptors, such as the estrogen and progesterone receptors.

Model Explanation and Interpretability: Montavon et al. (2017) define interpretation as a "mapping of an abstract concept (e.g. a predicted class) into a domain that the human can make sense of.", such as images or text [29]. An explanation is then "the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g. classification or regression)." A common explanation for image data are heatmaps that visualize which pixels were relevant for the model output. Furthermore, they divide interpretability into two subgroups: post-hoc interpretability that aims at analysing a given trained model and interpretability incorporated into the model (sometimes called ante-hoc interpretability [17]).

Nested Crossvalidation: A method for model selection and performance evaluation of a machine learning model, that makes optimal use of the available data and is therefore suitable for small datasets.

3 State of the Art

3.1 Prediction of ER Status Using DNNs

Recently, several studies applied DNNs to the classification of ER status from HE images. Rawat et al. (2018) constructed spatial maps of engineered features that describe nuclear morphology from HE images [32]. These features are then classified by a DNN into ER+ and ER− with an AUC of 0.72. The approach requires prior segmentation of the nuclei and manual feature engineering. This adds to the complexity of the method but on the other hand the results are better interpretable. Furthermore, the method was developed on a relatively small sample (57 train and 56 test samples).

Couture, Williams et al. (2018) trained an ensemble of calibrated SVMs on intermediate VGG16 features to predict patch-wise scores for the ER status and other quantities such as the tumor grade [10]. These scores were then aggregated into quantiles and classified by another SVM to get the final patient-wise prediction. This yielded 84% accuracy, 88% sensitivity and 76% specificity. A possible limitation of this work is that it is not an end-to-end model, i.e. that the used features are not necessarily well-suited for HE images. Moreover, even though the sample size is significantly larger than in [32] (571 train and 288 test samples), it is still small enough that the estimation of the performance on unseen data might be heavily perturbed if only a single hold-out set is used for evaluation.

Recently, Shamai et al. (2019) presented a ResNet model that was trained to predict the ER status from tissue microarray (TMA) samples [36], similar to our work. The model was trained on two cohorts of 20600 HE TMAs of 5356 patients in total. This resulted in an AUC of 0.84 and 0.88, for the two cohorts respectively, that decreased to 0.73 and 0.84 when only one TMA image per patient was used. Furthermore, they report a balanced accuracy of 77% and 82% for both cohorts, respectively. Even though the authors showed response maps of the classifier indicating relevant parts of the input image by overlaying the final feature maps with the image, these response maps were in a lower spatial resolution as the input image.

3.2 Post-hoc Model Explanation

A variety of methods have been developed to explain the output of neural network classifiers [2], with applications including histopathology [7,13,25]. For image data, many of these methods aim to produce explanation heatmaps that indicate the "importance" of each pixel, such that the user can see which structures the model considered relevant. For instance, Smilkov et al. (2017) developed a method called SmoothGrad that averages gradient maps of several noisy versions of an input image [37]. Selvaraju et al. (2017) presented GradCAM that aims to project the activations of the last feature extraction layer of a DNN back to the image space [35]. Bach, Binder et al. (2015) proposed Layer-wise Relevance Propagation (LRP), that decomposes the classifier output in a layer-by-layer fashion to propagate the output signal back to the input space [4]. In the PatternNet and PatternLRP approaches proposed by [22], explanations are learned from the data by optimizing a quality criterion that is based on the observation that an input signal is composed of an informative signal and uninformative noise.

4 Methods

4.1 Data and Preprocessing

The data are taken from TCGA "BRCA" project [1]. From the whole-slide HE images, a board-certified pathologist selected representative regions of interest (ROIs) of 2000 × 2000 px. The ROIs were preprocessed using the method by [38] to account for the variability in staining. In total, 702 cases had labels for the ER status. The classes were imbalanced with 176 (25%) ER− and 526 (75%) ER+ samples. For grading, we used the annotations by [8] where 469 cases were available from which 53 had grade 1, 219 grade 2 and 197 grade 3. We combined grade 1 and 2 to a class "low grade" resulting in 272 (58%) low-grade and 197 (42%) high-grade samples.

[1] Available at https://portal.gdc.cancer.gov/projects/TCGA-BRCA/.

4.2 Training and Evaluation

We designed an ensemble of DNNs for ER status prediction that is trained end-to-end. The core of the model is a ResNet18 [16] that was truncated after the third (out of four) residual block. This allowed benefitting from transfer learning but at the same time limited the number of parameters to reduce overfitting (see Table 3). After the final residual block, the feature maps were spatially averaged to yield a 256-d feature vector. For each ROI, multiple small patches (size 64×64 px) were processed individually by this feature extraction part of the DNN in order to regularize by limiting the context size. The resulting feature vectors were then fused by averaging over the patches to yield a combined representation of the entire image. This averaged feature was then fed to a dense layer with two output dimensions and softmax activation. Before the processing, each ROI was downsampled to half resolution since this performed consistently better than full resolution (see Table 3). The patches were sampled with uniform probability since more sophisticated, novel sampling strategies did not exhibit a significant performance gain (see Sect. 5). The model was trained using the Adam optimizer [23] with a learning rate of 10^{-4} and a cross entropy loss where each class was weighted proportional to the inverse of its frequency in the training data (weights are normalized such that they sum to 1).

Different strategies of random patch sampling have been explored in this paper. The first one is a simple uniform sampling where all patch locations are sampled with equal probability. Secondly, the complement of the red channel ("1 - red") in RGB space was used to undersample regions with high stromal tissue content. The third sampling strategy is intended to focus on nuclei by focusing on the Haematoxylin channel (computed using the method by [38]). Areas with high Haematoxylin content therefore were sampled more frequently.

Model selection and performance evaluation are performed by stratified nested crossvalidation (CV) [39]. This procedure and the terminology used in this paper are sketched in Fig. 1. In the outer CV loop, the dataset was randomly partitioned into five folds with class stratification, i.e. each of the folds had approximately the same number of positive and negative samples, respectively. Then, 4 of these folds were combined (called "development" fold in this paper) and used for training and validation in the inner CV procedure. The remaining fold (called "test" fold) was used to estimate the generalization performance of the models trained in the inner CV procedure. This procedure was repeated five times such that each fold served as a test fold once. In the inner CV loop, the respective development fold was further split randomly into five folds with stratification. Each fold was once excluded from the training and used for validation and early stopping (called "validation" fold). The remaining 4 folds were combined to a "training" fold and used to learn the weights and biases of the model. The mean validation performance over the five inner splits was used as estimate of the performance for this particular outer split. This quantity was used for model selection, i.e. to rank different hyperparameter settings.

After five-fold inner CV for hyperparameter tuning, the final classifier was obtained by forming an ensemble of the five individual models by averaging their output probabilities. This ensembling not only reduces the variance of the estimator [15] but also does not require resource-intensive retraining and no additional validation data for early stopping. Both outer and inner splits were the same for all experiments to facilitate comparison.

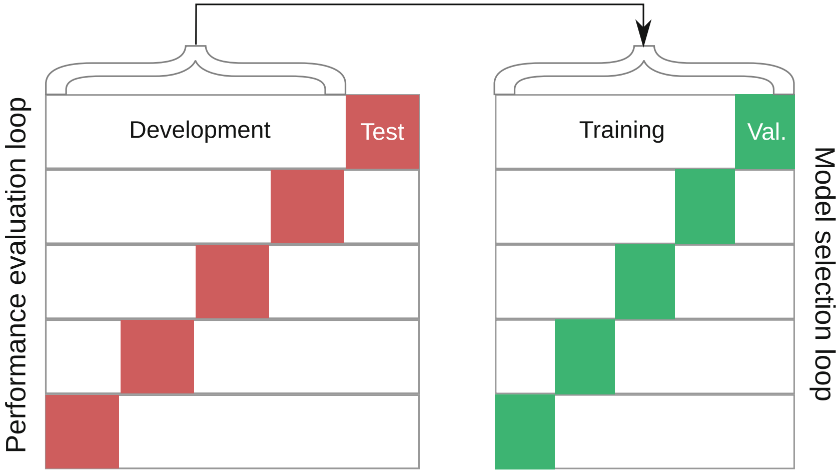


Fig. 1. Illustration of the nested CV procedure for model selection and performance evaluation. *See text for details.*

The performance on unseen data is estimated by averaging the test performances over five outer splits, that are not used for training or model selection. Following [32], we used AUC (averaged over inner folds) as model selection metric because of its applicability to imbalanced data [9]. To evaluate the final performance, we report ROC curves, balanced accuracy, precision and accuracy.

4.3 Visual Explanation of ER Status Predictions

Layer-Wise Relevance Propagation. [4] A feedforward DNN processes an input signal $\boldsymbol{x}^0$ by propagating it through a sequence of layers and hence computes a mapping $f(\boldsymbol{x}^0)$. In each layer l, the signal is recombined by a linear projection parametrized by weights W and biases $\boldsymbol{b}$[2] and usually passed through a non-linear activation function σ:

$$\boldsymbol{x}^{l+1} = \sigma(W^l \boldsymbol{x}^l + \boldsymbol{b}^l) = \sigma(\boldsymbol{z}^l) \ , \tag{1}$$

where $\boldsymbol{z}$ denotes the preactivation of the layer. By doing so, the network extracts features from the input signal, that grow more and more abstract and complex after each layer. For a classification task, the feature of the last layer is usually

[2] Note that this formulation includes both fully-connected and convolutional layers.

the output of a fully-connected layer and converted into probabilities via the softmax function.

LRP aims at computing a score for every input dimension, which denotes the contribution of the input dimension to the classification result [4]. The output signal $r^L := f(\boldsymbol{x}^0)$ is distributed backwards through the network, using the same topology as in gradient backpropagation, yet with message functions different from the gradient. During this backward pass, LRP computes for every neuron x_j^{l+1} relevance messages $R_{x_i^l \leftarrow x_j^{l+1}}$ from the neuron x_j^{l+1} to each of its inputs x_i^l. The relevance of a neuron x_i^l is then computed as the sum of all incoming relevance messages.

The key idea of LRP is that in each decomposition step, the signal is conserved, i.e. the total signal $\boldsymbol{r}^{l+1}$ in a layer $l+1$ should be approximately equal to the total signal $\boldsymbol{r}^l$ in the previous layer l:

$$\sum_i r_i^l \approx \sum_j r_j^{l+1} \quad . \tag{2}$$

Once the signal is backpropagated until the input layer $l = 0$, a score is assigned to each input dimension that measures by how much this dimension contributed to the output. Indeed, the relevances are an approximate *decomposition* of the output signal:

$$\sum_i r_i^0 \approx f(\boldsymbol{x}^0) \quad , \tag{3}$$

which directly follows from Eq. 2. This quantity—termed *relevance*—can hence be interpreted as the “importance” of this dimension for the predicted output. For images, the input relevances can be conveniently visualized as a heatmap.

The conservation property is ensured by applying specific decomposition rules, that determine how the relevance of a layer is distributed to the neurons in the previous layer, as described in the following section.

z-rule: The basic intuition is that a neuron i in layer l receives a share of the relevance of each connected neuron j in the following layer $l+1$ and this share should be proportional to its relative contribution to the input of j:

$$r_i^l = \sum_j \frac{z_i^l}{\sum_k z_k^l} r_j^{l+1} \quad . \tag{4}$$

Note that this rule is commonly not directly used in practice but serves as basis for further, improved rules, as described below.

ϵ-rule: Equation 4 is ill-defined for $\sum_k z_k^l = 0$ and therefore, a small positive constant ϵ can be added to the denominator that relaxes the conservation property but enhances numerical stability:

$$r_i^l = \sum_j \frac{z_i^l}{\sum_k z_k^l + \epsilon \ \mathrm{sign}(\sum_k z_k^l)} r_j^{l+1} \quad . \tag{5}$$

$\alpha\beta$-rule: Positive relevance indicates that a neuron increased the output score; negative relevance means decrease. In some applications, positive and negative relevance should be interpreted differently. In order to weight them separately, this rule thus introduces two parameters α and β:

$$r_i^l = \sum_j \left(\alpha \frac{\max(0, z_i^l)}{\sum_k \max(0, z_k^l)} + \beta \frac{\min(0, z_i^l)}{\sum_k \min(0, z_k^l)} \right) r_j^{l+1}. \tag{6}$$

Both parameters are related by $\alpha + \beta = 1$ to ensure relevance conservation. Common choices are $(\alpha = 1, \beta = 0)$ and $(\alpha = 2, \beta = -1)$.

$|z|$-rule: Modern DNNs often apply batch normalization [20]. Hui et Binder (2019) demonstrated that other LRP rules often perform poorly for batch normalization layers and devised a sign-invariant alternative that should be used instead [19]:

$$r_i^l = \sum_j \frac{|w_{ij}^l x_i^l|}{|w_{ij}^l x_i^l| + |b_j^l| + \epsilon} r_j^{l+1}. \tag{7}$$

For further methods, applications and implementation details of LRP, we refer to [1,11,26,28].

Application to ER Status Predictions. In order to elucidate which features were used by the model to classify a tumor sample into ER+ and ER−, we applied LRP to create relevance heatmaps. In particular, we applied the ϵ-rule ($\epsilon = 0.1$) to dense layers and the $|z|$-rule proposed by [19] to Batch Normalization layers. To convolutional layers, the $\alpha\beta$-rule ($\alpha = 1$, $\beta = 0$) was applied. In order to create heatmaps for an ROI, the LRP-decomposition was restricted to the logit of the highest scoring class by clamping the respective other logit to zero. Relevances were summed over color channels.

We chose to overlap patches by striding with half the patch size in order to average out potential border effects and translation dependence of the prediction, as recommended by [29]. To distinguish between positive and negative predictions in the final heatmap tile, the sign of the relevances of predicted negative patches was flipped before averaging. For visualization, relevances were clipped at the 0.995 quantile of the averaged heatmap tile; by doing so, extremely high relevance values did not influence the colormap substantially. This normalization process was performed for negative and positive relevances separately. In this paper, we restrict the heatmap analysis to a single CV fold.

Furthermore, visualizations were computed using SmoothGrad [35]. The magnitude of the gradient was averaged over 15 samples and sign-flipped for ER−predictions for visualization purposes, similar to above. For visualization we clipped the relevances at the 0.98 quantile to achieve a similar appearance as the LRP heatmaps.

5 Experiments

5.1 Hyperparameter Tuning and Performance Evaluation

For the ER status, we tuned the following hyperparameters on the validation sets: learning rate, image resolution, depth of the network and sampling method

Table 1. AUC (in %) for ER status and cancer grade for validation and testing for our end-to-end method and the method by [10] ("SVM") for all outer folds.

Fold	Val. (ER)		Test (ER)		Val. (Grade)		Test (Grade)	
	SVM	Our	SVM	Our	SVM	Our	SVM	Our
0	79.4	82.6	81.1	73.0	74.7	78.0	81.8	84.8
1	79.7	80.1	72.5	80.8	77.0	80.3	71.9	74.2
2	78.5	80.7	76.1	79.6	76.9	80.6	72.6	71.5
3	78.5	80.8	75.4	75.9	71.6	77.4	78.6	86.5
4	73.5	78.0	86.7	92.0	76.1	81.4	70.0	76.2
Avg.	77.9	80.5	78.4	80.3	75.2	79.5	75.0	78.6

Table 2. Comparison of the mean AUCs (in %) of the individual models (each trained on one inner fold) to the AUC (in %) of an ensemble of them for all outer folds.

Fold	Mean AUC (individual)	Ensemble AUC
0	72.5	73.0
1	78.6	80.8
2	77.4	79.6
3	73.3	75.9
4	89.2	92.0

(see Table 3 for results from hyperparameter tuning following the procedure described in Sect. 4.2). The validation and test results for our end-to-end method and a state-of-the-art method [10] for both ER status and cancer grade are summarized in Table 1.

No significant difference between the different sampling methods (uniform, "1 - red" and Haematoxylin oversampling) could be observed (see Table 3). Hence, we chose to remove this hyperparameter from the model selection and constantly used uniform sampling instead because it has the fewest assumptions and is therefore not prone to introducing any bias.

After determining hyperparameters during model selection, the final performance was estimated by averaging the test performance over all outer splits. As expected, the ensemble performed better than the individual models on average (see Table 2). This yielded an AUC of 0.80 (see Table 1), balanced accuracy of 73%, precision of 72% and accuracy of 79%. The AUC is thus substantially higher than in [32].

In order to test whether the benefit of end-to-end learning translates to other clinically relevant variables, we trained the same architecture to predict the pathological cancer grade. As before, we applied stratified nested CV. Without additional hyperparameter optimization, i.e. using the same configuration as for the ER status, the model achieved an average test AUC of 0.79 (see Table 1) and a balanced accuracy of 72%. This could be further increased to an AUC 0.81 and balanced accuracy of 73% by hyperparameter optimization (see Table 3).

5.2 Comparison to a State-of-the-Art Method

We applied the method by [10] ("SVM") to our dataset with the following minor modifications: Instead of 800 × 800 px patches with 400 px overlap, we used 600 × 600 px with 300 px overlap since the downscaled images only had a size of 1000 × 1000 px. Furthermore, we used standard Platt scaling [31] instead of isotropic regression for the calibration of the first SVM. We found that balancing the classes during training deteriorates the validation performance and thus, we did not apply any reweighting (see Table 3). Apart from that, the method was applied as described in [10].

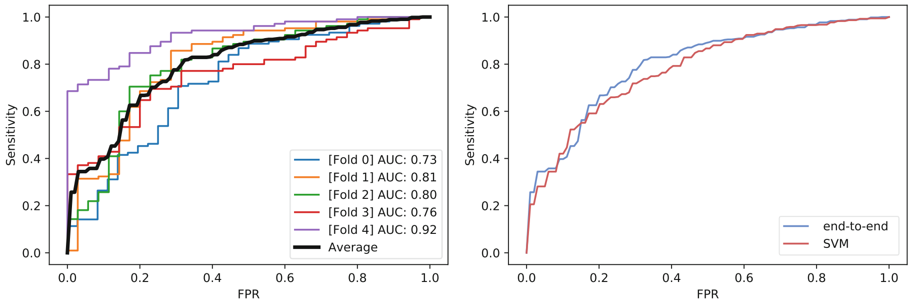
A comparison showed an improvement by our method compared to the SVM method in all folds during validation (see Table 1). Similarly to our method, half resolution yielded a higher AUC than full resolution. Furthermore, the SVM method performed better with VGG16 features as in [10] compared to the truncated ResNet features (see Table 3). Our method performed better not only in the validation but also in the test phase, where the SVM method scored on average an AUC of 0.78 (see Table 1), balanced accuracy of 70%, precision of 70% and accuracy of 78%. The ROC curves of the two methods are compared in Fig. 2.

We compared different resolutions for the prediction of the grade, for each of which the end-to-end method was superior in terms of AUC, even by a larger margin than for the ER status prediction (see Table 3). As for the ER status, half resolution performed best. Using the same hyperparameters as for the ER status, the SVM method scored an AUC of 0.75 (see Table 1) and a balanced accuracy of 70% on the test set which is outperformed by our end-to-end approach.

5.3 Variance over the Splits

We observed that the test AUC varied significantly between the different outer CV splits (see Table 1). Thus, the test performance computed on a single outer split is an unreliable estimate of the true generalization performance. Even more concerning is the fact that this variance can be substantial: In our experiments, the test AUC ranged from 0.73 to 0.92 (see Table 1). Hence, if the evaluation relied on a single train-test split only, the real-world performance could have been severely over- or underestimated. Figure 2 shows the ROC curves for the different outer splits to demonstrate this variance.

Interestingly, the outer fold with the best mean validation performance (fold 0) has the worst test performance. This means that the test set of fold 0 is particularly hard (i.e. harder than average) to classify; vice versa, its train and validation sets are particularly easy since they are disjoint from the test set. This leads to an overly optimistic validation performance in fold 0 (and to a overly pessimistic validation performance in all the other folds) whereas the



(a) Variance of test performance (b) Average test performance

Fig. 2. (a): ROC curves for all outer test sets for our method. (b): Average test ROC curves comparing our end-to-end method to the SVM method by [10].

test performance in fold 0 is overly pessimistic. Similarly, the fold with the worst validation performance (fold 4) shows the best test performance.

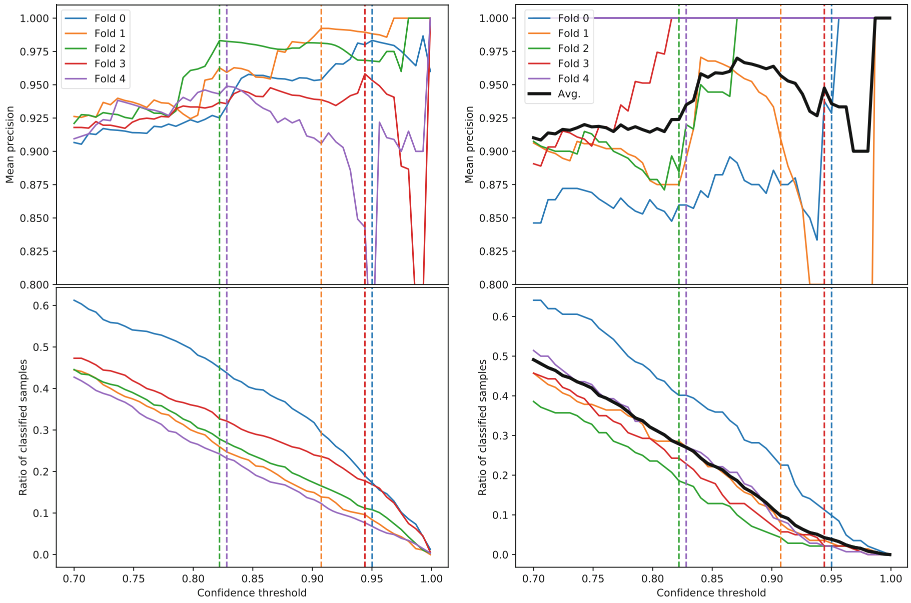
5.4 Rejection Option

The clinical relevance of HE-based determination of the ER status largely depends on the amount of patients that are classified correctly with high confidence[3] because these patients would not require additional examination e.g. by IHC stains. On the other hand, cases that cannot be classified confidently should be rejected by the classifier and undergo additional examination, such as IHC staining.

Therefore, we investigated whether a threshold on the classifier output ("confidence") can be chosen on the validation data such that a reasonable amount of test data can be classified with high confidence, similar to [10] and [21]. To this end, the ER+ precision of the model, i.e. the ratio of true ER+ patients of all patients classified as ER+, as well as the ratio of non-rejected samples, i.e. patients classified as ER+, were plotted against the confidence threshold (see Fig. 3). For the validation data, we averaged over the inner splits.

The confidence threshold was chosen based on the validation data for each outer split individually such that the precision was maximized while classifying more than 10% of the cases. These thresholds are shown as vertical lines in Fig. 3. Introducing this rejection option resulted in a precision of 97% and 94% for validation and test data, respectively, and a ratio of non-rejected samples of 20% and 13%, respectively (averaged over the outer folds).

[3] The output of a DNN typically does not represent the true confidence of the model in terms of probability since it is not properly calibrated. However, here, we stick with the term "confidence" to denote the output probability of the highest scoring class as it is commonly used that way in the literature.



(a) Validation (b) Test

Fig. 3. Precision and ratio of classified (not rejected) samples when introducing a rejection option. Confidence thresholds (depicted as vertical lines) are chosen on the validation data for each fold individually.

5.5 Visual Explanation

The resulting LRP heatmaps for classifying the ER status were inspected visually. From this, the following findings are deduced:

- Stroma is an important indicator for the model for ER+. This is in line with previous findings that suggest that stromal morphology correlates with survival [5] and the ER status [36]. Nevertheless, we consider this a potential issue of the learned predictor (see Sect. 6).
- Lymphocyte infiltration is an indicator for ER−. This is plausible because it has been reported in the literature that increased lymphocytic infiltration correlates with decreased ER activity [24,27].
- A high nuclear grade (prominent nucleoli, nuclear pleomorphism), i.e. poor differentiation, and mitoses are an indicator for ER−, in line with [27].
- The model responds to cells arranged in an “indian file”-like pattern. This pattern is typically observed in infiltrating lobular carcinoma, a histological subtype of breast cancer that correlates with a positive ER status [3].

Exemplary heatmaps are shown in Fig. 4, where red indicates evidence for ER+, blue for ER− and green is neutral (for more examples, see Fig. 6). In the following, we discuss these example cases in more detail.

TCGA-A2-A0CR (see Fig. 4(a)): Cancer cells aligned in an “indian file”-like pattern are evidence for ER+. This pattern is typical for infiltrating lobular carcinoma that more frequently is ER+. Additionally, stroma and the low-grade nuclear morphology are evidence for ER+. Lymphocyte infiltration is evidence for ER−. It is correctly predicted ER+ with confidence of 0.989.

TCGA-E2-A1LL (see Fig. 4(b)): This tumor features a bad differentiation and a high nuclear grade and thus, nuclear morphology is an indicator for ER−. Similarly, mitoses are correlated with uncontrolled tumor growth and are hence evidence for ER−. The small ratio of stromal tissue is however evidence for ER+. It is correctly predicted ER− with confidence 0.923.

TCGA-E2-A158 (see Fig. 4(c)): The large portion of stroma in this tumor is evidence for ER+; lymphocytes are evidence for ER−. The model falsely predicts ER+, albeit with a relatively low confidence of 0.834.

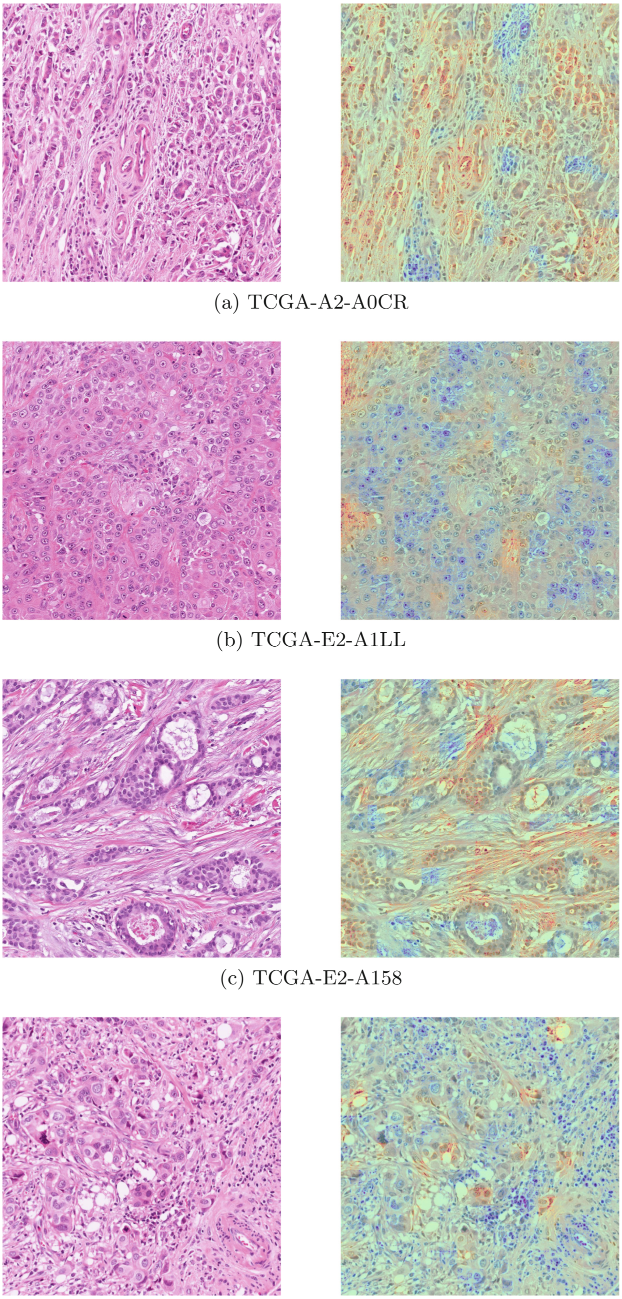
TCGA-C8-A12P (see Fig. 4(d)): This is a tumor with a medium degree of differentiation. The large number of lymphocytes is evidence for ER−. Even though the stroma indicates ER+, the model correctly predicts ER− with confidence 0.804.

A comparison of LRP heatmaps to SmoothGrad heatmaps showed that these findings are consistent with a differently motivated explanation method (see Fig. 5) and not a special finding of LRP. However, note that the SmoothGrad heatmaps come along with a significantly higher computational effort compared to LRP because they require an average over multiple gradient maps.

6 Discussion

It is common knowledge in machine learning that training models end-to-end is beneficial because not only the classifier but also the features are adapted specifically to the task at hand [6]. We demonstrated that this also holds for the classification of ER status and cancer grade from HE images even though our sample was relatively small: The end-to-end method compared favorably with recent prior DNN-based methods.

The high variance of the test performance between different outer folds shows that it can be misleading to rely on the performance on a single outer fold only to estimate generalization performance because it can be biased either too pessimistically (e.g. as in fold 0) or too optimistically (e.g. fold 4) and this bias can be substantial. This is due to the high variance with which the generalization performance is estimated that arises from the relatively modest sample size. This dilemma can be solved by averaging the test performances over the different outer splits which gives a more reliable estimate because the averaging reduces the variance of the estimator [15]. We argue that, while this might be irrelevant in general computer vision where usually large labeled datasets are available,



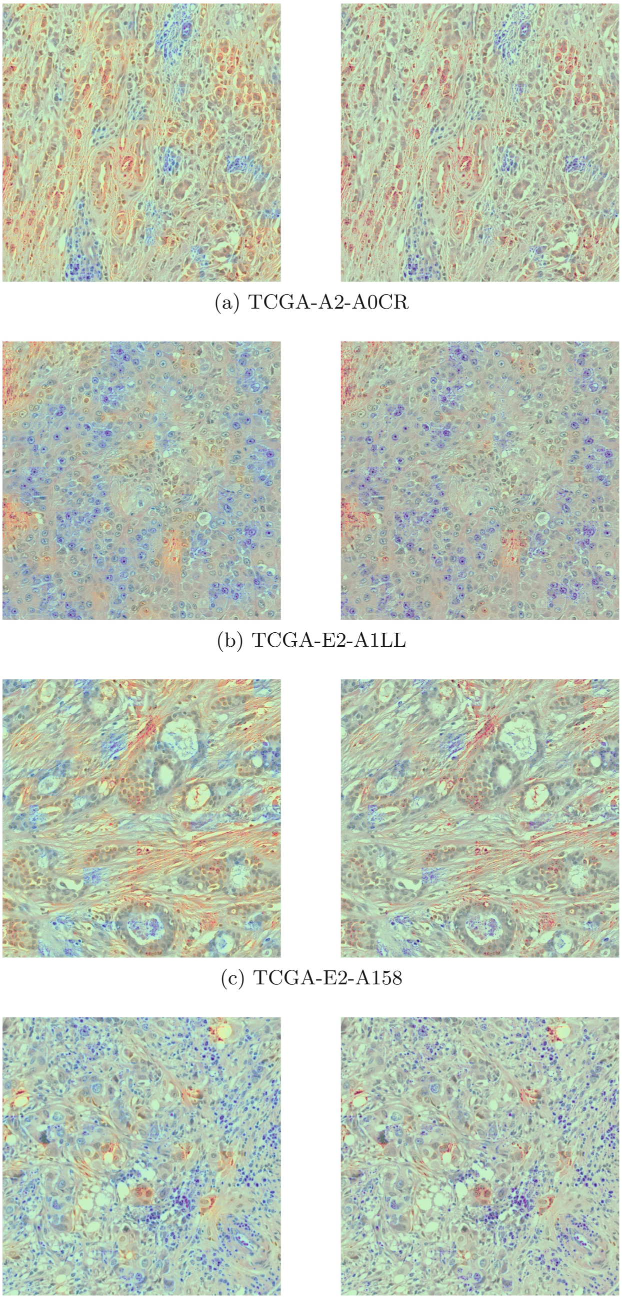
(a) TCGA-A2-A0CR

(b) TCGA-E2-A1LL

(c) TCGA-E2-A158

(d) TCGA-C8-A12P

Fig. 4. Examplary images and LRP-heatmaps (overlayed on image) for the detection of ER status. (a) Cell patterns typical for lobular breast carcinoma indicate ER+ (red). Lymphocyte infiltration indicates ER− (blue). (b) High nuclear and mitotic grade and lymphocyte infiltration indicate ER−. (c) Stromal patterns indicate ER+. (d) Lymphocyte infiltration indicates ER−. (Color figure online)



(a) TCGA-A2-A0CR

(b) TCGA-E2-A1LL

(c) TCGA-E2-A158

(d) TCGA-C8-A12P

Fig. 5. Comparison of LRP - (left) and SmoothGrad (right) heatmaps (overlayed on image), analogous to Fig. 4. This highlights similar features for both methods. (Color figure online)

in many medical image analysis tasks this variance is a serious problem because the datasets are much smaller. Furthermore, especially in the medical domain it is highly relevant to reliably estimate the performance of a machine learning model. Thus, we hope to focus the attention of the field to methodically sound model tuning and evaluation procedures, such as nested CV, even if this comes along with a considerably higher computational burden. We claim that in the context of machine learning applied to medicine, a reliable performance estimation should always have priority over finding an optimal hyperparameter setting because otherwise, trust of medical professionals and patients in learning machines can easily be undermined.

As shown in Table 1, the predictor has a competitive prediction performance. This is even more remarkable as the total sample size is 702 samples only, whereas a larger study [36] which reported AUCs of 0.73 and 0.84 was able to employ a sample size of 5356 patients. Visual explanation revealed several morphological features that the model considered relevant for classification: stromal features, lymphocyte infiltration, high nuclear grade and cell arrangement correlated with a specific histological subtype.

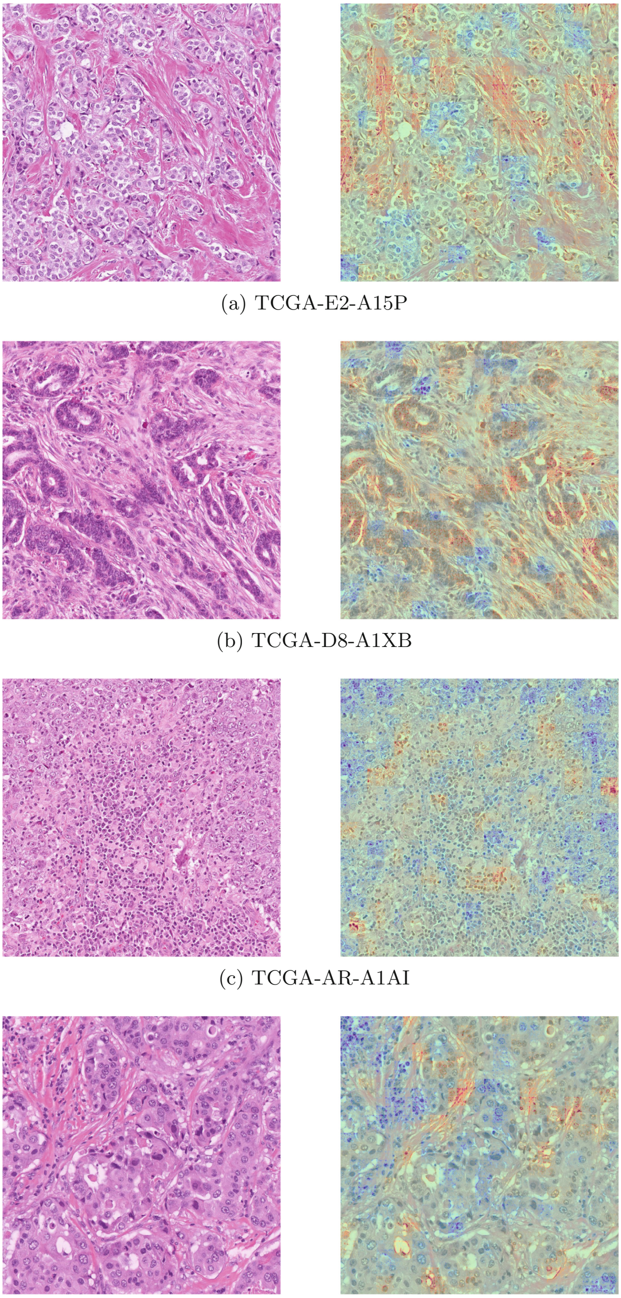
However, we would like to express doubts whether the stroma structure used as evidence for ER+ is a causal feature, despite work supporting this view like [5] and [36], that showed that the ER status could be learned from the stroma to a certain extent. In a larger study, this observation could be validated by inspecting a larger patient dataset, for which we are unable to obtain sufficient samples from the used TCGA dataset. While this hypothesis cannot be ascertained at this stage, this raised concern demonstrates the value of explanation methods for the development of machine learning models, in particular for features such as ER status which are not easily recognized by a human expert from HE stains. This is fundamentally different to many problems in the natural image domain, where the correctness of a prediction can be verified at a glance by non-experts.

An argument for the latter is the reliable selection of high-confidence predictions that results in high precision. Nevertheless, this study shows that explainable AI can be a tool to investigate models for ER status prediction and should be combined with domain knowledge for verification. This is especially important because the prediction of the ER status from HE images has a relatively modest classification performance [10,32,36] and hence could be more prone to a confusion of correlation and causation.

Table 3. Additional validation AUCs (in %) for ER status and cancer grade. (**ER**) SVM method by [10]: full resolution, truncated ResNet18 ("ResNet3L") as feature extractor and class balancing. Our method: full resolution, ResNet18 as model and different sampling methods (Haematoxylin and complement of red channel). (**Grade**) SVM method: quarter and full resolution. Our method: additional hyperparameter tuning. For comparison, the results of the final models are shown (see Table 1).

		SVM				Our				
	Fold	Full Res.	ResNet3L	balanced	Final	Full res.	ResNet18	Haem.	1-red	Final
ER	0	70.6	74.9	76.5	79.4	78.7	79.8	82.1	82.6	82.6
	1	69.1	72.7	74.5	79.7	74.9	78.5	81.3	80.1	80.1
	2	69.1	73.7	73.0	78.5	74.1	80.4	81.5	82.6	80.7
	3	72.7	73.5	73.0	78.5	73.5	79.5	80.1	81.9	80.8
	4	69.1	70.7	69.7	73.5	69.7	75.9	75.9	75.1	78.0
	Avg.	70.1	73.1	73.3	77.9	74.2	78.8	80.2	80.5	80.5

		SVM			Our	
	Fold	Quarter res.	Full res.	Final	Tuned	Final
Grade	0	69.6	68.4	74.7	80.5	78.0
	1	75.4	73.9	77.0	83.0	80.3
	2	76.8	70.6	76.9	82.5	80.6
	3	73.6	69.9	71.6	83.5	77.4
	4	75.1	75.5	76.1	84.7	81.4
	Avg.	74.1	71.7	75.2	82.8	79.5



(a) TCGA-E2-A15P

(b) TCGA-D8-A1XB

(c) TCGA-AR-A1AI

(d) TCGA-BH-A0B7

Fig. 6. Additional images and LRP-heatmaps (overlayed on image). (Color figure online)

7 Conclusion

This work presents a method to learn the prediction of the ER status of breast cancer from HE images, a task that is important for patient stratification but usually not visually apparent to the human observer. In contrast to prior work that relied on engineered or generic representations, this method uses features that are learned from the data at hand. This is achieved by training a DNN ensemble end-to-end on pooled random patches which shows state-of-the-art performance.

We show how CV naturally gives rise to an ensemble to avoid costly retraining after model selection. By introducing a rejection class, a considerable amount of test samples can be classified with high precision. Furthermore, we demonstrate the importance of robust model selection and performance evaluation, in this case nested CV, in the context of deep learning-based medical image analysis. Moreover, the architecture can be applied out-of-the-box to e.g. the prediction of cancer grade.

After training the model, we computed explanation heatmaps to better understand the model intrinsics. This revealed a number of biologically plausible features that are used by the model to predict the ER status: stromal and nuclear morphology, lymphocytic infiltration and disease-specific cell patterns. Explainable AI can thus be an important tool to verify hypotheses about biomarkers, especially in settings with low predictive performance as in the case of ER status prediction.

8 Open Challenges and Future Work

More sophisticated sampling strategies that oversample specific tissue structures had a negligible effect in this paper compared to naive uniform sampling. We hypothesize that this is due to the fact that the dataset consists of expert-selected ROIs; uniformly sampling from these ROIs is thus already likely to yield an informative patch. Thus, future work should investigate whether those sampling strategies improve the results on whole-slide images where it is more important to focus the attention to relevant parts of the data than for training on preselected ROIs.

The interpretation of the explanation heatmaps in this study is done by visual inspection. To leverage a more objective evaluation, the overlap of the relevance scores with specific structures in the tissue should be measured quantitatively [13]. However, this would require an expensive pixel-level annotation of e.g. cancer cells, lymphocytes and stroma, and is thus left for future work.

Moreover, the goodness of the explanation heatmaps should ideally be measured quantitatively [33]. This is a challenging task and no consensus on the best evaluation method has been reached yet. However, approximations to this problem exist, e.g. by destroying the information in the image based on the pixel relevances [34] and—after a possible retraining the model [18]—reevaluating the model. These evaluation methods will further help to establish trust in the interpretability methods themselves and can hence be an important ingredient in the path to applying artificial intelligence in a clinical setting.

References

1. Alber, M.: Software and application patterns for explanation methods. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS (LNAI), vol. 11700, pp. 399–433. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6_22
2. Alber, M., et al.: Innvestigate neural networks!. J. Mach. Learn. Res. **20**(93), 1–8 (2019)
3. Arpino, G., Bardou, V.J., Clark, G.M., Elledge, R.M.: Infiltrating lobular carcinoma of the breast: tumor characteristics and clinical outcome. Breast Cancer Res. **6**(3), R149 (2004). https://doi.org/10.1186/bcr767
4. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), e0130140 (2015)
5. Beck, A.H., et al.: Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci. Transl. Med. **3**(108), 108ra113–108ra113 (2011)
6. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
7. Binder, A., et al.: Towards computational fluorescence microscopy: machine learning-based integrated prediction of morphological and molecular tumor profiles. arXiv preprint arXiv:1805.11178 (2018)
8. Budczies, J., et al.: Classical pathology and mutational load of breast cancer-integration of two worlds. J. Pathol. Clin. Res. **1**(4), 225–238 (2015)
9. Cortes, C., Mohri, M.: AUC optimization vs. error rate minimization. In: Advances in Neural Information Processing Systems, pp. 313–320 (2004)
10. Couture, H.D., et al.: Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. NPJ Breast Cancer **4**, 30 (2018)
11. Dombrowski, A.K., Alber, M., Anders, C., Ackermann, M., Müller, K.R., Kessel, P.: Explanations can be manipulated and geometry is to blame. In: Advances in Neural Information Processing Systems, pp. 13567–13578 (2019)
12. Elston, C.W., Ellis, I.O.: Pathological prognostic factors in breast cancer. i. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology **19**(5), 403–410 (1991)
13. Hägele, M., et al.: Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. Sci. Rep. **10**(1), 1–12 (2020)
14. Hammond, M.E.H., et al.: American society of clinical oncology/college of american pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer (unabridged version). Archiv. Pathol. Lab. Med. **134**(7), e48–e72 (2010)
15. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, Heidelberg (2009). https://doi.org/10.1007/978-0-387-84858-7

16. 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, pp. 770–778 (2016)
17. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainabilty of artificial intelligence in medicine. Wiley Interdisc. Rev. Data Min. Knowl. Discov. **9**(4), e1312 (2019)
18. Hooker, S., Erhan, D., Kindermans, P.J., Kim, B.: Evaluating feature importance estimates. arXiv preprint arXiv:1806.10758 (2018)
19. Hui, L.Y.W., Binder, A.: BatchNorm decomposition for deep neural network interpretation. In: Rojas, I., Joya, G., Catala, A. (eds.) IWANN 2019. LNCS, vol. 11507, pp. 280–291. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20518-8_24
20. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
21. Jurmeister, P., et al.: Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. Sci. Transl. Med. **11**(509), eaaw8513 (2019). 11 September 2019, https://doi.org/10.1126/scitranslmed.aaw8513
22. Kindermans, P.J., et al.: Learning how to explain neural networks: PatternNet and PatternAttribution. arXiv preprint arXiv:1705.05598 (2017)
23. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
24. Klauschen, F., et al.: Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning. Semin. Cancer Biol. **52**, 151–157 (2018)
25. Korbar, B., et al.: Looking under the hood: deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR), pp. 821–827 (2017)
26. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. Nat. Commun. **10**(1), 1096 (2019)
27. Millis, R.R.: Correlation of hormone receptors with pathological features in human breast cancer. Cancer **46**(S12), 2869–2871 (1980). https://doi.org/10.1002/1097-0142(19801215)46:12+⟨2869::AID-CNCR2820461426⟩3.0.CO;2-Q
28. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.-R.: Layer-wise relevance propagation: an overview. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS (LNAI), vol. 11700, pp. 193–209. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6_10
29. Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digit. Signal Proc. **73**, 1–15 (2018)
30. Osborne, C.K., Yochmowitz, M.G., Knight III, W.A., McGuire, W.L.: The value of estrogen and progesterone receptors in the treatment of breast cancer. Cancer **46**(S12), 2884–2888 (1980)
31. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv. Large Margin Classif. **10**(3), 61–74 (1999)
32. Rawat, R.R., Ruderman, D., Macklin, P., Rimm, D.L., Agus, D.B.: Correlating nuclear morphometric patterns with estrogen receptor status in breast cancer pathologic specimens. NPJ Breast Cancer **4**, 32 (2018)

33. Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.): Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS (LNAI), vol. 11700. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6
34. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R.: Evaluating the visualization of what a deep neural network has learned. IEEE Trans. Neural Netw. Learn. Syst. **28**(11), 2660–2673 (2016)
35. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
36. Shamai, G., Binenbaum, Y., Slossberg, R., Duek, I., Gil, Z., Kimmel, R.: Artificial intelligence algorithms to assess hormonal status from tissue microarrays in patients with breast cancer. JAMA Netw. Open **2**(7), e197700–e197700 (2019)
37. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: SmoothGrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
38. Vahadane, A., et al.: Structure-preserving color normalization and sparse stain separation for histological images. IEEE Trans. Med. Imaging **35**(8), 1962–1971 (2016)
39. Varma, S., Simon, R.: Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics **7**(1), 91 (2006)