Deep learning approach to predict lymph node metastasis directly from primary tumour histology in prostate cancer

Frederik Wessels1,2, Max Schmitt1, Eva Kriehoff-Henning1, Tanja Jutzi1, Thomas S. Worst2, Frank Waldbilig2, Manuel Neuberger2, Roman C. Maron1, Matthias Steeg3, Timo Gaiser3, Achim Hekler1, Jochen S. Utikal4,5, Christof von Kalle6, Stefan Fröhling7, Maurice S. Michel2, Philipp Nuhn2 and Titus J. Brinker1

1Digital Biomarkers for Oncology Group, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Heidelberg, 2Department of Urology and Urological Surgery, 3Institute of Pathology, Medical Faculty Mannheim of Heidelberg University, University Medical Center Mannheim, Mannheim, 4Skin Cancer Unit, German Cancer Research Center (DKFZ), 5Department of Dermatology, Venereology and Allergology, University Medical Center Mannheim, University of Heidelberg, Heidelberg, 6Department of Clinical-Translational Sciences, Berlin Institute of Health (BiH), Charité University Medicine, Berlin, and 7National Center for Tumor Diseases, German Cancer Research Center, Heidelberg, Germany

F.W. and M.S. contributed equally to this work.

Objective
To develop a new digital biomarker based on the analysis of primary tumour tissue by a convolutional neural network (CNN) to predict lymph node metastasis (LNM) in a cohort matched for already established risk factors.

Patients and Methods
Haematoxylin and eosin (H&E) stained primary tumour slides from 218 patients (102 N+; 116 N0), matched for Gleason score, tumour size, venous invasion, perineural invasion and age, who underwent radical prostatectomy were selected to train a CNN and evaluate its ability to predict LN status.

Results
With 10 models trained with the same data, a mean area under the receiver operating characteristic curve (AUROC) of 0.68 (95% confidence interval [CI] 0.678–0.682) and a mean balanced accuracy of 61.37% (95% CI 60.05–62.69%) was achieved. The mean sensitivity and specificity was 53.09% (95% CI 49.77–56.41%) and 69.65% (95% CI 68.21–71.1%), respectively. These results were confirmed via cross-validation. The probability score for LNM prediction was significantly higher on image sections from N+ samples (mean [SD] N+ probability score 0.58 [0.17] vs 0.47 [0.15] N0 probability score, \( P = 0.002 \)). In multivariable analysis, the probability score of the CNN (odds ratio [OR] 1.04 per percentage probability, 95% CI 1.02–1.08; \( P = 0.04 \)) and lymphovascular invasion (OR 11.73, 95% CI 3.96–35.7; \( P < 0.001 \)) proved to be independent predictors for LNM.

Conclusion
In our present study, CNN-based image analyses showed promising results as a potential novel low-cost method to extract relevant prognostic information directly from H&E histology to predict the LN status of patients with prostate cancer. Our ubiquitously available technique might contribute to an improved LN status prediction.

Keywords
prostatic neoplasms, machine learning, deep learning, artificial intelligence, convolutional neural network, neoplasm metastasis, #ProstateCancer, #PCSM, #uroonc

Introduction
Prostate cancer accounts for 20% of all new diagnosed cancer cases in Europe and 10% of all cancer-related deaths in men [1]. Lymph node metastasis (LNM) is usually associated with incurable disease and higher risk of cancer-related death [2]. To determine the LN status in patients with prostate cancer, a lymphadenectomy is usually performed in parallel with
radical prostatectomy (RP), its extent depending on the perceived risk of LNM [3]. Current methods to predict LN status are usually based on clinical information such as PSA, information from the biopsy, including Gleason grading or percentage of positive cores, and in the most current model also MRI metrics as risk factors for LNM [4,5]. These models are able to achieve an area under the receiver-operating characteristics curve (AUROC) between 0.76 and 0.83 on external validation [6,7]. Preoperative adoption of such models is recommended to determine the need for and required extent of a lymphadenectomy [8]. A more extended lymphadenectomy can lead to a greater blood loss, longer hospital stay and more postoperative complications, especially lymphoceles, but can contribute to an improved LN staging [9]. Thus, any improvement of these models can help to reduce complications from unnecessary lymphadenectomies. Alternatively, a prostate-specific membrane antigen (PSMA) positron emission tomography (PET)/CT scan can be performed preoperatively. However, sensitivity for LNM varies significantly [10] and the high costs, as well as the lack of widespread availability, have to date prevented routine preoperative use.

In recent years, artificial intelligence (AI) has revolutionised oncological research. Multiple studies have proven the capability of AI to classify tissue samples correctly as benign or malignant, especially on haematoxylin and eosin (H&E) stained slides [11–13]. It was shown that AI, in particular image analysis using convolutional neural networks (CNN), can discriminate between malignant and benign tissue in prostate biopsies with an accuracy >90% [13,14]. Also, automated Gleason grading [15] achieved a similar accuracy as Gleason grading by specialised pathologists [16,17]. Recent studies have shown that CNNs are also capable of identifying protein expression alterations and genetic mutations on H&E slides of cancer samples across multiple cancer types, including prostate cancer, breast cancer, and liver cancer [18,19]. These studies suggest that genetic mutations induce changes in cell signalling and cell interaction that can lead to a change in morphology that is detectable by CNNs. AI therefore has the potential to identify oncological relevant patterns by analysing H&E images and to use these patterns to predict oncological outcomes, e.g. risk of metastasis or tumour recurrence [20].

In the present study, we aimed to evaluate whether a CNN could detect relevant patterns on H&E slides to predict LNM in prostate cancer in a Gleason-matched cohort.

**Patients and Methods**

**Study Design and Population**

Eligible patients were retrospectively selected from the prostate cancer database of patients who had undergone open or robot-assisted RP including lymphadenectomy at our institution, a German university hospital. Histologically confirmed LN-positive cases (N+) were selected and matched with LN-negative cases (N0) for age, Gleason score, tumour size (T), perineural invasion (Pn), and venous invasion (V). H&E slides from RP specimens of these patients were evaluated by a pathologist and one representative slide of good quality containing cancerous tissue per patient was chosen for analysis and digitalised. Slides with artefacts or slides of low quality were excluded. This analysis of prostate cancer tissue was approved by the local data protection office and by the local ethics committee (2015-8492-MA), patients had given informed consent to tissue analysis and this work was performed in accordance with the Declaration of Helsinki.

**Pre-processing**

A Zeiss Axio Scan.Z1 (Zeiss, Oberkochen, Germany) was used to digitalise 218 H&E slides with a 20-fold magnification, resulting in whole-slide images (WSIs) with a resolution of 0.22 µm/pixel (px). Using a tissue detection algorithm, WSIs were annotated semi-automatically to exclude the background (non-tissue) and artefacts. These annotated tissue areas were tessellated into square patches of 512 × 512 px resulting in an edge length of 112.64 × 112.64 µm/patch. Tissue detection and tessellation were both implemented in QuPath [21] version 0.1.2 in Groovy (https://qupath.github.io/). Subsequently, blurry patches and patches with too much background were discarded. A patch was classified as blurry if the variation of the Laplacian was <30. If 50% of the pixels were classified as background pixels, the patch was considered to contain too little tissue. A pixel in an image was defined as background if the grey-scale value was >0.74 (190/255). Both steps were implemented in Python version 3.7.7 (Python Software Foundation, Beaverton, OR, USA). To include all possible relevant areas, tumour areas were not annotated and the entire WSI was used as input. This approach is concordant with previous studies that showed that prognostic information on oncological outcome is not limited to tumour areas on WSI [20].

As entire RP tissue sections are very large, 1000 patches per WSI fulfilling the quality criteria were randomly chosen to reduce the duration of training and testing. Each patch was labelled LN-positive or -negative according to the WSI it originated from and its affiliation to the WSI was saved (Fig. 1, i).

**Model Training**

The 218 WSIs were divided into: a training set and a test set. The training set comprised 118 (60 N+, 58 N0) and the hold out test set 100 WSIs (42 N+, 58 N0). To achieve an equal
distribution between both sets, the sets were stratified with respect to class, date of slide creation, Gleason score, and operation technique. We used a pre-trained xse_ResNext34 CNN to predict LNM. During the training of the xse_ResNext34, the training set was split into a training (83 WSIs) and validation (35 WSIs) set. The hyperparameters were tuned using only the validation set before the model was trained a last time with all 118 WSIs and the established hyperparameters. The training of the CNN was divided into two parts. In the first part, the fully connected layers (head) as well as the layers close to the input (body) were trained for 10 epochs, using an initial learning rate of 3.01E-06. In the second part both, head and body, were trained for additional 10 epochs, using a learning rate of 2.08E-6. Training followed Leslie Smith’s one cycle policy [22] and learning rates were selected based on an algorithm that minimises loss for a smaller sample of the training set while maximising learning rate to speed up train time. For data augmentation, a combination of transformations that is suited for a variety of vision datasets was used (Flip \( P = 0.5 \), Warp \([\text{magnitude} = 0.2, \ P = 0.75]\), Rotate \([\text{max deg} = 10, \ P = 0.75]\), Zoom \([\text{min zoom} = 1, \text{max zoom} = 1.1, \ P = 0.75]\), Brightness \([\text{max lighting} = 0.2, \ P = 0.75]\), Contrast \([\text{max lighting} = 0.2, \ P = 0.75]\), according to the method described by Howard et al. [23]. A method called test time augmentation (TTA) was used with standard parameters, as it resulted in the best-balanced accuracy and AUROC on our validation set. TTA is a technique in which for each individual image, the final prediction is an average of the prediction made on both the original image and random transformations of the image [24]. After hyperparameters were established, the model was trained in a final run using all 118 WSIs of the training set. Inference was carried out on all 1000 patches for each WSI (Fig. 1, ii–iv). Every patch was assigned an LNM score by the CNN (probability that the primary tumour has metastasised to LNs). All scores for a given WSI were averaged to a final LNM score for the complete slide. If the final LNM score exceeded a threshold of 0.5, the whole slide image was classified as N+, otherwise as N0. Training, as well as inference, was implemented in Python 3.7.7, using PyTorch 1.6 [25] and fast.ai [23].

**Model stability/robustness**

To further verify the stability of our proposed method independently of the training and test set, we used a cross-
validation approach. We divided the whole set of 218 WSIs into five folds, stratifying it in the same way as described in Model training. Five models were trained so that for each model a different fold was excluded from training and used as a test set for evaluation of the model. Both training and inference were executed using the previously established method with random patches and the established hyperparameters.

Statistics

As primary endpoints, balanced accuracy and AUROC were reported with 95% CIs. The CIs are the result of 10 models trained with the same data and using the established hyperparameters. The LNM probability scores per patient of the N+ and N0 classes were compared using a two-sided Mann–Whitney U-Test with a predefined significance level of \( P = 0.05 \). For the cross-validation approach the variance between the five models of the primary endpoints instead of their 95% CIs was reported. These calculations were conducted in Python 3.7.7 extended with the library SciPy. AUROC of our CNN was compared with the performance of the Memorial Sloan-Kettering Cancer Center (MSKCC) nomogram for prediction of LNM [26] on our test set. Univariable and multivariable logistic regression analyses were performed to identify independent predictors for LNM using JMP version 15.2.1 (SAS Institute, Cary, NC, USA).

Results

Patient Data

Detailed patient characteristics can be found in Table 1. The compared groups showed no significant differences in Gleason score, age, operation technique, tumour category, venous and perineural invasion. The median (interquartile range [IQR]) preoperative PSA level was N+ 14.0 (8–28) vs 10.33 (6.1–17) ng/mL N0 (\( P = 0.004 \)); the median (IQR) number of resected lymph nodes was N+ 14 (10–19) vs 12 (9–16) N0 (\( P = 0.02 \)); and the rate of lymphovascular invasion (LVI) was 75%, (\( n = 91 \)) vs 25% (\( n = 31 \)) (\( P < 0.001 \)), with significant differences between groups.

Performance

On the validation set, balanced accuracy was 70.99% (95% CI 68.52–73.47%), while the AUROC was 0.84 (95% CI 0.83–0.85), sensitivity 56.11% (95% CI 50.35–61.87%), and specificity 85.88% (95% CI 83.70–88.05%).

For the primary endpoints using the test set, a mean AUROC of 0.68 (95% CI 0.678–0.682) and a mean balanced accuracy of 61.37% (95% CI 60.05–62.69%) were achieved over the 10 CNNs that were trained with the same hyperparameters. Furthermore, a mean sensitivity and specificity of 53.09% (95% CI 49.77–56.41%) and 69.65% (95% CI 68.21–71.1%), respectively, could be observed. The model performing best on validation set achieved an AUROC of 0.69 and a balanced accuracy of 64.24% on the test set, while the sensitivity was 59.52% and the specificity was 68.97%. In Fig. 2, mean ROC curve, AUROC including SDs over 10 CNNs trained and tested with the same data (Fig. 2a) and ROC curve and AUROC of the individual CNN performing best on the validation set (Fig. 2b) are shown.

In the five-fold cross validation, mean (SD) balanced accuracy was 63.26 (4.9%). The mean (SD) AUROC was 0.70 (0.05), therefore showing a comparable performance. Additionally,

### Table 1 Patient data and comparison.

|                     | All  | N+   | N0   | \( P \)  |
|---------------------|------|------|------|----------|
| Patients, n         | 218  | 102  | 116  |          |
| Age, years, median (IQR) | 68 (64–73) | 68 (63–73) | 69 (64–73) | 0.72   |
| PSA level, ng/mL, median (IQR) | 12 (7.3–21) | 14 (8–28) | 10.33 (6.1–17) | 0.004 |
| Open surgery, n (%) | 121  | 59   | 62   | 0.51    |
| Resected LNs, n, median (IQR) | 13 (9–17) | 14 (10–19) | 12 (9–16) | 0.02   |
| Tumour category, n (%) |       |      |      | 0.77    |
| T2b                 | 1    | 0    | 1    |          |
| T2c                 | 19   | 9    | 19   |          |
| T3a                 | 50   | 22   | 28   |          |
| T3b                 | 148  | 71   | 77   |          |
| Venuous invasion, V+, n (%) | 41 (19) | 24 (23) | 17 (15) | 0.09   |
| Perineural invasion, Pn+, n (%) | 217 (100) | 101 (100) | 115 (99) | 0.35   |
| Residual tumour, R+, n (%) | 101 (46) | 48 (47) | 53 (46) | 0.84   |
| LV+, n (%)          | 109  | 78   | 31   | <0.001  |
| Gleason score, n (%) |       |      |      | 0.9     |
| Gleason score 7a    | 20   | 9    | 11   |          |
| Gleason score 7b    | 54   | 25   | 29   |          |
| Gleason score 8     | 29   | 12   | 17   |          |
| Gleason score 9     | 112  | 54   | 58   |          |
| Gleason score 10    | 2    | 1    | 1    |          |

© 2021 The Authors

BJU International published by John Wiley & Sons Ltd on behalf of BJU International
the use of three different tile distributions on the test set also showed comparable results as demonstrated in Table A1.

To assess the difference in the LNM scores between both classes (N+ and N0), a two-sided Mann–Whitney U-test was performed using the above mentioned best performing model, as this test takes the LNM probability score for each individual patient as an input. Results showed a significant difference between N+ and N0 patients (mean [SD] N+ probability score 0.58 [0.17] vs 0.47 [0.15] N0 probability score, \( P = 0.002 \)). Figure 3 exemplarily demonstrates the random distribution of tiles and their individual class prediction based in the probability score by the CNN.

As the groups were not matched for LVI and PSA, we performed a multivariable analysis including PSA, LVI, resected lymph nodes, age and the LNM prediction by the CNN to identify independent predictors for LNM (Table 2).

Here, LNM prediction probability (continuous) by the CNN (odds ratio [OR] 1.04/\% probability, 95% CI 1.02–1.08; \( P = 0.04 \)) and LVI (OR 11.73, 95% CI 3.96–35.70; \( P < 0.001 \)) proved to be independent predictors for LNM.

In addition, a confusion matrix of LVI vs predicted LNM classes indicated that although LVI correlated with LNM to some extent, it was not the main predictor for LNM in our model (Fig. A1).

Comparison with Existing Models and Predictors

To evaluate the predictive value of established nomograms predicting the risk of LNM in our matched high-risk cohort, we exemplary calculated the MSKCCs probability of LNM. This model showed an AUROC of 0.63 for the prediction of LNM on the test set, showing a slightly worse performance than our CNN-based prediction. Additionally, we calculated

![ROC curves](image)

**Fig. 2** ROC curves on patient-level (n = 100 patients): the continuous line in (a) shows the mean ROC curve of 10 identically trained CNNs and their CI in grey. The continuous line in (b) represents the ROC curve of the model performing best on the validation set and the dotted line in both plots indicates a ROC curve resulting from random classification. The 1-specificity (false positive rate) was plotted on the x-axis and the sensitivity (true positive rate) on the y-axis.

**Fig. 3** Exemplary demonstration of the tile distribution and the individual tile classification by the CNN: the figure shows the slide of a patient with a Gleason 4 + 5 tumour. The tiles are colour-coded, green represents a tile rated by the CNN with a probability of >0.5 for the presence of an LNM, yellow shows the rating <0.5. The random distribution of the tiles is clearly visible. The colour coding on this example demonstrates that both tiles from the tumour area and from the non-tumour area were assessed by the CNN for the presence of LNM with a probability of >0.5.

© 2021 The Authors
BJU International published by John Wiley & Sons Ltd on behalf of BJU International
Recently, Kather et al. [18] were able to demonstrate that detection of molecular changes, including tumour protein p53 (TP53), BRCA1, caspase 8 (CASP8) and others, on H&E slides is possible across 14 solid cancer types. Also, AI-aided prediction of outcome was performed successfully. Yamamoto et al. [20] were able to demonstrate that prediction of biochemical recurrence after RP was possible by a CNN-based image analysis using H&E slides of RP specimens.

In our CNN-based prediction of LNM, we were able to achieve a mean AUROC of 0.68 with a mean balanced accuracy of 61.37%. To put these results into perspective, the following considerations on the study design have to be considered. We specifically intended to evaluate whether a CNN can identify other than the already established oncologically relevant image features such as Gleason grade on H&E slides and whether additional information gain can result from such analyses. Thus, we matched our present cohort for Gleason score, as well as age, tumour size and extent, venous and perineural invasion. Especially Gleason score and tumour size have been proven to be significant risk factors for LNM [5]. To account for the fact that the groups were not matched for LVI and PSA level, possible predictors of LNM, we performed a multivariable analysis that showed that the CNNs predictions were an independent predictor for LNM in our test set. This suggests that the CNN can indeed provide an additional benefit on LN prediction and that LVI was very likely not the main predictor for the CNN. Due to the exclusion of known risk factors, a comparison to published performances of models predicting LNM that incorporate these risk factors, e.g. the Briganti nomogram or the MSKCC nomogram with a reported AUROC of >0.8 [5], is only possible to a limited extent. We exemplarily evaluated the MSKCC’s nomogram performance on our test set. Results showed a low performance in our present cohort, which we expected as the cohort was matched for some of the factors considered. We specifically intended to evaluate whether a CNN can identify other than the already established oncologically relevant image features such as Gleason grade on H&E slides and whether additional information gain can result from such analyses. Thus, we matched our present cohort for Gleason score, as well as age, tumour size and extent, venous and perineural invasion. Especially Gleason score and tumour size have been proven to be significant risk factors for LNM [5]. To account for the fact that the groups were not matched for LVI and PSA level, possible predictors of LNM, we performed a multivariable analysis that showed that the CNNs predictions were an independent predictor for LNM in our test set. This suggests that the CNN can indeed provide an additional benefit on LN prediction and that LVI was very likely not the main predictor for the CNN. Due to the exclusion of known risk factors, a comparison to published performances of models predicting LNM that incorporate these risk factors, e.g. the Briganti nomogram or the MSKCC nomogram with a reported AUROC of >0.8 [5], is only possible to a limited extent. We exemplarily evaluated the MSKCC’s nomogram performance on our test set. Results showed a low performance in our present cohort, which we expected as the cohort was matched for some of the factors included in the model. In comparison with the CNNs performance, the MSKCCs AUROC of 0.63 was slightly worse, indicating a possible benefit of combining our method with established classifiers.

Interestingly, when combining LVI and our CNNs prediction an AUROC of 0.83 resulted, significantly higher than the individual predictions alone. Apparently, these were the main remaining predictors for LNM in our high-risk cohort. This also emphasises the fact that LVI is a significant risk factor for a more aggressive prostate cancer type [35]. However, this risk factor is not found in the current nomograms. This could be due to the fact that LVI can rarely be diagnosed on the

### Table 2 Multivariable logistic regression analysis Predictors of lymph node positivity on test set (n = 100).

| Predictor                     | AUROC univariate | Wald X    | OR (95% CI) | P     |
|-------------------------------|------------------|-----------|-------------|-------|
| LVI (L1 vs L0)                | 0.74             | 24.26     | 11.73 (3.96–35.70) | <0.001 |
| LNM prediction probability by the CNN (continuous, %) | 0.68 | 7.45 | 1.04 (1.01–1.08) | 0.009 |
| PSA (continuous, ng/mL)       | 0.61             | 4.55      | 1.01 (0.99–1.03) | 0.12  |
| Age (continuous, years)       | 0.51             | 1.73      | 1.058 (0.97–1.15) | 0.18  |
| Resected LNs (≥15 vs <15)    | 0.57             | 1         | 1.74 (0.58–5.15) | 0.32  |
prostate core biopsy. Therefore, the routine reporting of LVI is usually recommended, but not mandatory according to the current reporting guides on prostate biopsy [36]. Still, this result should be evaluated further. Interestingly, adding PSA to the model only led to a minor further improvement and PSA was no independent predictor of LNM in the multivariable analysis. However, in a larger non-matched consecutive cohort, PSA is expected to also significantly contribute to LNM prediction before [2,4].

Another point to consider is the risk classification of the patients in our present cohort. In our cohort, more than half of the patients had a Gleason score of 9 or 10 due to the described process of matching. Therefore, also the patients that were diagnosed as N0 had a significant risk of LNM. Although extended lymphadenectomy was standard procedure in these cases, positive LNs might have been missed during lymphadenectomy, especially considering the reported difference in resected LNs between groups. In addition, micrometastases could have been missed in routine pathology [37], both possibly leading to false negative labels for training and testing of our CNN. These labels might have lowered AUROC and accuracy of our method.

Considering the above-mentioned points, we are confident that our present results reflect the capability of our CNN to extract new relevant information from H&E images to predict LN status in patients with prostate cancer. The present work is one in a series of first works to show the potential to predict clinical outcome in a prostate cancer cohort using a CNN and specifically to investigate the capability of CNNs to predict LN status directly from histology in patients with prostate cancer.

Although this was not the main aim of our present study, the analysis also revealed that in this high-risk cohort, LVI was the best individual predictor of LNM. Although it is rarely diagnosed on prostate core biopsies, our results suggest that LVI should be integrated as a risk factor in future models for predicting the LN status in patients with prostate cancer and that the presence of LVI in a prostate core biopsy alone might justify an extended lymphadenectomy.

Limitations

Our present study has several limitations. First, retrospective analysis of slides has to be mentioned. It should be noted here that before prospective studies can be started, well-established algorithms from retrospective analyses need to be developed. Second, results of our study have to be confirmed in further studies. As our results were obtained with RP specimens, our algorithm needs to be tested on core biopsies of the prostate before translation into clinical practice is feasible. As shown in previous studies, in the process of validation, adaption and retraining of the CNN usually is necessary to account for subtle differences in staining and image quality, as well as different type of scanner when applying our method to external validation sets.

Clinical Implementation

The described approach is a cost-effective technique that has the potential to provide improved predictions on LN status. Before clinical implementation can be attempted, further research on the digital biomarkers and image patterns that contribute most to the CNNs classification and adapting this method to prostate cancer biopsies is needed. First, external validation of this method is necessary to emphasise robustness and generalisation. Second, translation of this method to prostate core biopsies is desirable as an application during the ongoing operation on RP specimen is likely to be far more complicated to implement. To achieve a high accuracy LN status prediction, the existing CNN could be combined with/integrated into existing or future nomograms via logistic multivariable regression analysis. Alternatively, a new CNN-based classifier could be trained on unmatched cohorts. That way, the resulting CNN would be able to incorporate known risk factors such as tumour stage or Gleason grading into its decision-making process and weight these factors automatically. Thereby, it might achieve a high predictive accuracy as a stand-alone diagnostic tool that could be compared directly with existing nomograms. However, to conduct such studies, due to the low incidence of LNM, a large and preferably multicentre cohort is necessary. Ultimately, such models might have the capability to predict LNM with accuracy high enough to avoid unnecessary lymphadenectomies and thus reduce the morbidity of RP.

Conclusion

A CNN-based image analysis has the potential to improve the preoperative risk-assessment of LNM in prostate cancer. However, its true value can only be defined after further research. Additionally, LVI appears to be a significant risk factor for LNM, which is not yet reflected in current predictive models.

Disclosure of Interests

Titus J. Brinker would like to disclose that he is the owner of Smart Health Heidelberg GmbH (Handschuhsheimer Landstr. 9/1, 69120 Heidelberg, Germany) which develops mobile apps, outside of the submitted work.

Author contributions

Frederik Wessels: conceptualisation, data curation, investigation, project administration, methodology, formal analysis, writing – original draft; writing – review and editing; Max Schmitt: conceptualisation, investigation, formal analysis,
Deep learning in prostate cancer

software, methodology, writing – original draft; writing – review and editing; Eva Kriehoff-Hennig: conceptualisation, funding acquisition, methodology, supervision; writing – original draft; writing – review and editing; Tanja Jutzi: conceptualisation, funding acquisition, methodology, supervision; writing – review and editing; Thomas S. Worst: conceptualisation, methodology data curation, writing – review and editing; Frank Waldbilling: conceptualisation, writing – review and editing; Manuel Neuberger: data curation, writing – review and editing; Roman C. Maron: software, writing – review and editing; Matthias Steeg: data curation, writing – review and editing; Timo Gaiser: conceptualisation, supervision, writing – review and editing; Achim Hekler: software, supervision, writing – review and editing; Jochen S. Utikal: writing – review and editing; Christof von Kalle: writing – review and editing; Stefan Fröhling: resources, writing – review and editing; Maurice S. Michel: resources, writing – review and editing; Philipp Nuhn: conceptualisation, funding acquisition, supervision, resources, writing – review and editing; Titus J. Brinker: conceptualisation, methodology, resources, supervision, project administration, funding acquisition, writing – original draft, writing – review and editing.

Funding information
This study was funded by the Federal Ministry of Health, Berlin, Germany (grant: Tumorverhalten-Praediktions-Initiative; grant holder: Titus J. Brinker, German Cancer Research Center; grant-number: 2519DAT712).

References
1 Ferlay J, Colombet M, Soerjomataram I et al. Cancer incidence and mortality patterns in Europe: estimates for 40 countries and 25 major cancers in 2018. Eur J Cancer 2018; 103: 356–87
2 Briganti A, Karnes JR, Da Pozzo LF et al. Two positive nodes represent a significant cut-off value for cancer specific survival in patients with node positive prostate cancer. A new proposal based on a two-institution experience on 703 consecutive N+ patients treated with radical prostatectomy, extended pelvic lymph node dissection and adjuvant therapy. Eur Urol 2009; 55: 261–70
3 EAU Guidelines. Edn. presented at the EAU Annual Congress Amsterdam 2020. ISBN 978-94-92671-07-3
4 Gandaglia G, Ploussard G, Valero M et al. A novel nomogram to identify candidates for extended pelvic lymph node dissection among patients with clinically localized prostate cancer diagnosed with magnetic resonance imaging-targeted and systematic biopsies. Eur Urol 2019; 75: 506–14
5 Briganti A, Larcher A, Abdollah F et al. Updated nomogram predicting lymph node invasion in patients with prostate cancer undergoing extended pelvic lymph node dissection: the essential importance of percentage of positive cores. Eur Urol 2012; 61: 480–7
6 Oderda M, Diamand R, Albisininni S et al. Indications for and complications of pelvic lymph node dissection in prostate cancer: accuracy of available nomograms for the prediction of lymph node invasion. BJU Int 2021; 127: 318–25
7 Gandaglia G, Martini A, Ploussard G et al. External validation of the 2019 Briganti Nomogram for the identification of prostate cancer patients who should be considered for an extended pelvic lymph node dissection. Eur Urol 2020; 78: 138–42
8 Mottet N, van den Bergh RC, Briers E et al. EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate cancer 2020 update. Part 1: screening, diagnosis, and local treatment with curative intent. Eur Urol 2021; 79: 243–62
9 Fossati N, Willemsse PM, van den Broeck T et al. The benefits and harms of different extents of lymph node dissection during radical prostatectomy for prostate cancer: a systematic review. Eur Urol 2017; 72: 84–109
10 Abrams-Pompe RS, Fanti S, Schoots IG et al. The role of magnetic resonance imaging and positron emission tomography/computed tomography in the primary staging of newly diagnosed prostate cancer: a systematic review of the literature. Eur Urol Oncol 2020 [Online ahead of print]. https://doi.org/10.1016/j.euo.2020.11.002
11 Brinker TJ, Hekler A, Enk AH et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Cancer 2019; 113: 47–54
12 Hekler A, Utikal JS, Enk AH et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. Eur J Cancer 2019; 118: 91–6
13 Litijens G, Sanchez CI, Timofeeva N et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016; 6: 26286
14 Kott O, Linsley D, Amin A et al. Development of a deep learning algorithm for the histopathologic diagnosis and gleason grading of prostate cancer biopsies: a pilot study. Eur Urol Focus. 2019 [Online ahead of print]. https://doi.org/10.1016/j.euf.2019.11.003
15 Epstein JI. An update of the Gleason grading system. J Urol 2010; 183: 433–40
16 Bulten W, Pinckaers H, van Boven H et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol 2020; 21: 233–41
17 Xu H, Park S, Hwang TH. Computerized classification of prostate cancer Gleason scores from whole slide images. IEEE/ACM Trans Comput Biol Bioinform 2020; 17: 1871–82
18 Kather JN, Heij LR, Grabsch HI et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nature Cancer 2020; 1: 789–99
19 Fu Y, Jung AW, Torne RV et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat Cancer 2020; 1: 800–10
20 Yamamoto Y, Tsuchi T, Akatsuka J et al. Automated acquisition of explainable knowledge from unannotated histopathology images. Nat Commun 2019; 10: 5642
21 Bankhead P, Loughrey MB, Fernandez JA et al. QuPath: open source software for digital pathology image analysis. Sci Rep 2017; 7: 1–7
22 Loshchilov I, Hutter F. SGD: Stochastic gradient descent with warm restarts, 2016
23 Howard J, Gugger S. Fastai: a layered API for deep learning. Information 2020; 11: 108
24 Wang G, Li W, Aertsen M, Deprest J, Ourselin S, Vercauteren TJ. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 2019; 338: 34–45
25 Paszke A, Gross S, Massa F, editors et al. Pytorch: an imperative style, high-performance deep learning library. Advances in neural information processing systems, 2019
26 Memorial Sloan-Kettering Cancer Center. Prediction Tools/Prostate Cancer Nomograms. Available at: https://www.mskcc.org/nomograms/prostate/pre_op. Accessed January 2021
27 Suhovskih AV, Kashuba VI, Klein G, Grigorieva EV. Prostate cancer cells specifically reorganize epithelial cell-fibroblast communication

© 2021 The Authors
BJU International published by John Wiley & Sons Ltd on behalf of BJU International
through proteoglycan and junction pathways. *Cell Adh Migr* 2017; 11: 39–53.

28 Russo MV, Esposito S, Tupone MG et al. SOX2 boosts major tumor progression genes in prostate cancer and is a functional biomarker of lymph node metastasis. *Oncotarget* 2016; 7: 12372–85.

29 Wang Q, Diao X, Sun J, Chen Z. Stromal cell-derived factor-1 and vascular endothelial growth factor as biomarkers for lymph node metastasis and poor cancer-specific survival in prostate cancer patients after radical prostatectomy. *Urol Oncol* 2013; 31: 312–7.

30 Castro E, Goh C, Olmos D et al. Germline BRCA mutations are associated with higher risk of nodal involvement, distant metastasis, and poor survival outcomes in prostate cancer. *J Clin Oncol* 2013; 31: 1748–57.

31 Barron DA, Rowley DR. The reactive stroma microenvironment and prostate cancer progression. *Endocr Relat Cancer* 2012; 19: R187–204.

32 Zeng Y, Opeskin K, Horvath LG, Sutherland RL, Williams ED. Lymphatic vessel density and lymph node metastasis in prostate cancer. *Prostate* 2005; 65: 222–30.

33 Singh S, Singh UP, Grizzle WE, Lillard JW Jr. CXCL12-CXCR4 interactions modulate prostate cancer cell migration, metalloproteinase expression and invasion. *Lab Invest* 2018; 8: 17343.

34 Kapil A, Meier A, Zuraw A et al. Deep semi supervised generative learning for automated tumor proportion scoring on NSCLC tissue needle biopsies. *Sci Rep* 2018; 8: 17343.

35 Luo HL, Chiang PH, Chen YT, Cheng YT. Lymphovascular invasion is a pathological feature related to aggressive clinical behavior and predicts early recurrence in prostate cancer. *Kaohsiung J Med Sci* 2012; 28: 327–30.

36 Grignon DJ. Prostate cancer reporting and staging: needle biopsy and radical prostatectomy specimens. *Mod Pathol* 2018; 31(Suppl. 1): S96–109.

37 Heck MM, Retz M, Bandur M et al. Topography of lymph node metastases in prostate cancer patients undergoing radical prostatectomy and extended lymphadenectomy: results of a combined molecular and histopathologic mapping study. *Eur Urol* 2014; 66: 222–30.

Correspondence: Titus J. Brinker, Digital Biomarkers for Oncology Group, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, Heidelberg, 69120 Germany.

e-mail: Titus.brinker@nct-heidelberg.de

Abbreviations: AI, artificial intelligence; AU(ROC), area under (receiver operating characteristic) curve; BRCA, BReast CANcer susceptibility protein; CNN, convolutional neural network; H&E, haematoyxin and eosin; IQR, interquartile range; LN(M), lymph node (metastasis); LVI, lymphovascular invasion; MSKCC, Memorial Sloan-Kettering Cancer Center; N+, LN positive; N0, LN negative; OR, odds ratio; PD-L1, programmed death ligand 1; RP, radical prostatectomy; TTA, test time augmentation; WSI, whole-slide image.

Appendix 1

Table A1 Performance of the CNN* using three different tile distributions (each time 1000 tiles/slide) on the test set.

| Tile Distribution | Bal. accuracy, % | AUROC | Sensitivity, % | Specificity, % |
|-------------------|------------------|-------|----------------|----------------|
| Random tile distribution #1 | 64.24 | 0.690 | 59.52 | 68.97 |
| Random tile distribution #2 | 64.24 | 0.690 | 59.52 | 68.97 |
| Random tile distribution #3 | 61.68 | 0.689 | 54.76 | 68.97 |

*The CNN with the best performance on the validation set was used to evaluate the performance using three different tile distributions on the test set.

Fig. A1 Confusion matrix predicted LNM vs LVI (test set). L-, no lymphatic vessel invasion; L+, lymphatic vessel invasion; p LNM+, predicted positive for lymph node metastasis by the CNN; p LNM-, predicted negative for lymph node metastasis by the CNN.