Segmentation and classification for lung nodule analysis

Gijs Thuis, s4494044
Faculty of Science
Radboud University
Nijmegen, Netherlands
gijs.thuis@ru.nl

Charlotte Janssen, s1069532
Faculty of Social Sciences
Radboud University
Nijmegen, Netherlands
charlotte.janssen@ru.nl

Lisanne van Gelderen, s1108561
Faculty of Social Sciences
Radboud University
Nijmegen, Netherlands
lisanne.vangelderen@ru.nl

Floris Rossel, s1010794
Faculty of Science
Radboud University
Nijmegen, Netherlands
floris.rossel@ru.nl

***Abstract*—This report details our participation in the LUNA23 challenge, hosted at https://luna23-ismi.grand-challenge.org/, which involves analyzing lung nodules in chest CT volumes. Our study addresses three key objectives: malignancy risk estimation, nodule type classification, and nodule segmentation. We built upon a U-net and explored learning rate schedules, optimizers and data augmentation techniques to improve the segmentation task. Ensemble learning was applied to both the segmentation and classification models. Our efforts to improve segmentation with learning procedure optimization did not result in an improvement over the baseline parameters that produced a Dice score of 0.823 on our validation set. For all three tasks combined with ensemble learning applied to every model, we achieved the highest score (0.816) on the Grand Challenge test set. The limited number of samples for non-solid and part-solid nodules in our validation was detrimental for the balanced accuracy scores of nodule type classification, while malignancy estimation showed excellent performance with an AUC of 0.967 on our validation set.**

***Index Terms*—pulmonary lung nodules, image segmentation, image classification, ensemble learning, 3D CNN, U-Net**

I. Introduction

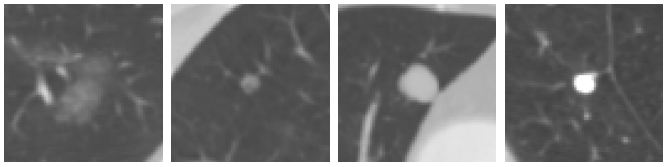
Lung cancer has a high mortality rate and its symptoms are difficult to diagnose in an early stage [1]. Radiologists use chest computed tomography (CT) volumes to identify pulmonary nodules in the lungs, but manual inspection can be very time-consuming.

In recent years, computer-aided diagnosis (CAD) systems have been developed to assist radiologists in the segmentation and classification of these nodules. Accurate segmentation and classification of specific nodule types and malignancy rates are crucial in determining the appropriate treatment for each patient. Compared to manual inspection, CAD systems offer remarkable advantages in terms of speed, significantly accelerating the cancer detection process. Moreover, these systems possess the potential for greater accuracy, resulting in lower rates of misdiagnosis, which is crucial for improving patient survival rates.

In our study, we will investigate the training procedures of two separate models to perform the following three tasks: malignancy risk estimation of the nodule (i), nodule type classification (ii) and nodule segmentation (iii).

II. Lung Nodule Dataset

The proposed methods will be evaluated on the LUNA23-ISMI dataset. Our training set consists of 687 3D CT volumes of nodules (128 x 128 x 64) which will be used for training and validation. The nodule types that are present in this dataset are non-solid (51), part-solid (18), solid (466) and calcified (152). Examples of these nodules are shown in Figure 1.



(a) Non-solid (b) Part-solid (c) Solid (d) Calcified

Fig. 1. Four nodule types

III. Methods

In this study, we utilize two pre-existing models and aim to enhance its performance through various strategies. Our focus revolves around two key approaches. Firstly, we explore different kinds of learning procedures and hyperparameters for the segmentation model specifically. After analyzing the performance of the baseline models and their learning procedures, we observed greater potential for improving the performance of the segmentation task as compared to the two classification tasks. Secondly, we employ ensemble learning techniques for each of the three tasks using a stratified 5-fold split.

We prioritize testing ensemble learning and different learning procedures for the segmentation task. Ensemble learning shows promise for medical image segmentation because of its ability to integrate diverse segmentation models, improving accuracy and robustness [2]. By combining multiple trained models and avaraging out their individual weaknesses, ensemble methods can enhance segmentation performance.

The learning procedures include several different optimizers, a range of learning rates, various learning rate schedules, and different degrees of data augmentation. For the two classification tasks, we investigate ensemble learning with our baseline training procedures.

Our code with the corresponding models and training loop is available on GitHub.[1]

A. Models

1) U-net: For the segmentation task, we use an U-net implemented by Ronneberger et al. [5]. The contraction part of this U-net comprises 5 blocks, each consisting of a convolutional layer, batch normalization, ReLU activation and max pooling. The 4 expansion blocks of the network use transposed convolution to restore the original image size of the pulmonary nodules.

We also experimented with adding more convolutional layers to the model. So 6 or 7 contraction blocks and 5 or 6 expansion blocks. This resulted in a similar Dice score as the model with 5 contraction blocks and 4 expansion blocks but required more training epochs. Therefore, we decided not to add additional convolutional layers to the U-net.

2) 3D CNN: For determining the type of lung nodules and their malignancy rate, we use a 3D CNN. This is the baseline network that we were provided with for this challenge. It contains four convolutional layers with 32, 64, 64 and 128 filters respectively. The max pooling layers of this network use a kernel size of 2. We infer the nodule type using a softmax activation, and for malignancy estimation we use a threshold of 0.5 to decide if a nodule is either benign or malignant.

B. Learning procedure and hyperparameter tuning

Before applying ensemble learning on our segmentation model, we test the performance of the model with different learning procedures, optimizers and data augmentation.

1) Learning rate schedule: We explore the effectiveness of incorporating exponential learning rate decay, where a decay rate factor is applied to the learning rate after each epoch. We examine various combinations of decay rates with different initial learning rates. When employing stronger decay rates, we compensate for the diminishing effect on the learning rate later in the training process by utilizing higher initial learning rates.

2) Optimizers: In addition to our baseline optimizer, Adam, we explore two alternative optimizers for training our network: Stochastic Gradient Descent (SGD) and Adaptive Gradient Algorithm (Adagrad). However, given the characteristics of these optimizers, we find it necessary to employ initial learning rates that are three orders of magnitude higher in order to achieve satisfactory outcomes.

3) Data augmentation: In the base line model, we use rotation and translation techniques as augmentation methods for pulmonary nodules. To further enhance the model's robustness, we also incorporate Gaussian noise into each sample. The magnitude of the noise added to each sample is defined by specific values (0.1, 0.3, or 0.5), which correspond to the standard deviation of the Gaussian distribution. For every sample, we generate a noise sample image by drawing random numbers from this distribution, which is subsequently added to the original data sample, which contains brightness values between 0 and 255.

[1]https://github.com/FlorisRX/bodyct-luna23-ismi-trainer

C. Ensemble learning

We use a specific form of ensemble learning called model averaging, where a consensus output is created based on the output of individually trained models. For each task in the LUNA23 challenge, we train the corresponding model once for each fold. The resulting five models are then used for inference on the test set, producing five outputs. These outputs are finally averaged to reach a consensus prediction. Here, we chose to first average the outputs and apply a sigmoid afterwards. The benefit of this approach is that under- and overestimations of the individual models are reduced, preventing predictions that are too extreme.

IV. Results

A. Nodule segmentation

As mentioned in the previous section, we focus on improving the segmentation of the pulmonary nodules. Different configurations of the training schedule have been tested to improve the performance of the model. 19 experiments with different hyperparameter settings and corresponding results are presented in Table I, where the results for the best performing parameters are highlighted in bold. With the baseline parameters, a Dice score of 0.823 is reached, as shown in the first row.

TABLE I
PARAMETER TUNING RESULTS FOR LUNG NODULE SEGMENTATION

Exp.	LR	Scheduler	Decay	Optimizer	Noise	Dice
1	**0.0001**	**none**	**N.A.**	**Adam**	**0**	**0.823**
2	0.0001	none	N.A.	SGD	0	0.035
3	0.0001	none	N.A.	AdaGrad	0	0.095
4	0.0001	none	N.A.	Adam	0.1	0.818
5	**0.0001**	**none**	**N.A.**	**Adam**	**0.3**	**0.828**
6	0.0002	E.D.	0.99	Adam	0	0.803
7	0.0003	E.D.	0.98	Adam	0	0.801
8	0.1	none	N.A.	SGD	0	0.795
9	0.1	none	N.A.	AdaGrad	0	0.746
10	0.0001	none	N.A.	Adam	0.3	0.812
11	0.0001	none	N.A.	Adam	0.5	0.816
12	0.0004	E.D.	0.99	Adam	0	0.790
13	0.0006	E.D.	0.98	Adam	0	0.780
14	0.3	E.D.	0.98	SGD	0.3	0.787
15	0.3	E.D.	0.99	SGD	0.3	0.793
16	0.0002	E.D.	0.99	Adam	0.3	0.805
17	0.0003	E.D.	0.98	Adam	0.3	0.788
18	0.0004	E.D.	0.985	Adam	0.3	0.798
19	0.0006	E.D.	0.975	Adam	0.3	0.794

Of all the configurations we tested, the only one that scored higher than the baseline was when Gaussian noise was added to the data with a standard deviation of 0.3, giving a score of 0.828. However, a second run with the same parameters resulted in a Dice score of 0.812, which was lower than the baseline. The AdaGrad and SGD optimizers resulted in worse performance compared to the Adam optimizer with the parameters we tested. This could be explained by Adam's use of the first and second moment, where Adagrad only

uses second moment information and basic SGD does not use any momentum. We were unsuccessful in achieving better performance using different learning rates (0.0002, 0.0003, 0.0004 and 0.0006) in combination with exponential decay as our learning rate schedule, giving Dice scores of 0.803, 0.801, 0.790 and 0.780, respectively. This suggests that a learning rate of 0.0001 is optimal in the currently explored landscape of parameters for segmenting pulmonary nodules.

The learning curves for the different experiments shown in Figures 2-5 (see Appendix) indicate that the models do not overfit on the training set, since the Dice scores for the training and validation sets are very similar.

B. Nodule type classification

In Table II and Table III, the confusion matrices for nodule type classification is presented for the training and validation set. The labels 0, 1, 2 and 3 refer to the classes non-solid, part-solid, solid and calcified, respectively. We can see from this table that all nodules are correctly classified during training but, as expected, on the validation set some nodules are being misclassified. The learning curves for the training and validation set in Figure 6 (see Appendix) show that the model severely overfits on the training data, and does not generalize well to validation data. One of the biggest problems here is probably that we have a very limited number of non-solid and part-solid nodules available in our validation set. This makes it difficult to determine how well the model would actually classify these nodule types. In contrast, we can see that the model is able to classify 102 out of 114 solid nodules correctly, which is not really surprising considering that we have way more nodules available for this type.

TABLE II
CONFUSION TABLE FOR NODULE TYPE CLASSIFICATION AT EPOCH 754 ON TRAINING SET

		Predicted			
	Class	0	1	2	3
Actual	0	144	0	0	0
	1	0	129	0	0
	2	0	0	132	0
	3	0	0	0	134

TABLE III
CONFUSION TABLE FOR NODULE TYPE CLASSIFICATION AT EPOCH 754 ON VALIDATION SET

		Predicted			
	Class	0	1	2	3
Actual	0	2	0	1	0
	1	1	1	0	0
	2	3	8	102	1
	3	0	0	3	26

C. Malignancy risk estimation

For malignancy risk estimation, an AUC of 0.967 was reached on the validation set. Table IV and V show the confusion matrices for malignancy risk estimation after 120 epochs for the training and validation set. When training the model on 1000 epochs, the best metric was found at 120 epochs. Overall, the network seems to estimate the malignancy of a pulmonary nodule well on the training set, but could still be improved on the validation set. This could indicate that there is some overfitting. Out of the 97 nodules that were benign, 84 were correctly indicated as such. This implies that we have 13 false positives. From the 51 nodules that were malignant, 47 were indicated to be malignant. This means that 4 malignant nodules were incorrectly indicated as benign, i.e. that there were 4 false negatives. We can also show this with the precision and recall (or sensitivity) for the malignancy estimations. The precision is 0.783 while the recall is considerably higher at 0.922, which highlights again that we have more false positives than false negatives.

TABLE IV
CONFUSION TABLE FOR MALIGNANCY RISK CLASSIFICATION AT EPOCH 120 FOR TRAINING SET

		Predicted	
	Class	Benign	Malignant
Actual	Benign	254	2
	Malignant	6	277

TABLE V
CONFUSION TABLE MALIGNANCY RISK CLASSIFICATION AT EPOCH 120

		Predicted	
	Class	Benign	Malignant
Actual	Benign	84	13
	Malignant	4	47

D. Ensemble learning

Using ensemble learning with the default training configuration, the segmentation model reaches a Dice score of 0.770 on the validation set. We submitted this model also to Grand Challenge to see how the model would perform on unseen data. There, a Dice score of 0.674 was reached, which is slightly higher than the Dice score that was obtained with the baseline model (0.649). The significant drop in performance on unseen data suggests a need for improved generalization capabilities.

For nodule type classification, our ensemble learning approach yielded an accuracy of 0.693 on the validation set, while the accuracy without ensemble learning was 0.724. For malignancy risk estimation, the ensemble learning approach achieved an AUC of 0.958, compared to 0.967 without ensemble learning. It appears that for these two specific tasks, ensemble learning did not result in performance improvements. Considering the three tasks combined, the ensemble learning approach applied to each model yielded an overall validation score of 0.807.

We are most interested in the performance of the models with ensemble learning on unseen data from the Grand Challenge. For this data, the models achieved an overall score of 0.816, surpassing the scores of the other models we submitted to the Grand Challenge, without ensemble learning.

Discussion

The main goal of this study was to enhance the segmentation of pulmonary nodules through hyperparameter tuning. Despite our efforts, we were unable to achieve improvements in the model by adjusting the learning rate and experimenting with alternative optimizers such as AdaGrad and SGD. One possible explanation for this outcome is that Adam utilizes both first and second moments, while AdaGrad solely relies on the second moment and SGD lacks momentum altogether. To address this, future research could explore a wider range of learning rates for these optimizers to identify more effective configurations.

For the data augmentation, we made an error in the way noise was added to the data. Our desired approach was to multiply the standard deviation values by the maximum pixel value of 255, but this last step was omitted. This resulted in a tiny, imperceptible amount of noise to be added to the data. This augmentation step led to a slight increase (0.828) of this score in one of the runs, but additional runs showed lower scores, indicating that it does not provide a significant benefit. Future work may use stronger noise for increasing model robustness and generalizability, and combine this with other parameters to investigate synergistic effects. We also recommend employing additional data augmentation techniques, such as brightness and contrast adjustment, warping, scaling and flipping.

Future work could also benefit from adding attention gates to the U-Net, as was done by Oktay et al. [5]. Attention gates have the potential to enhance performance by performing feature selection, such that more attention is allocated to regions of interest while suppressing irrelevant information. We made an attempt at implementing attention gates, but did not manage to get a working model in time.

Both the segmentation task and the two classification tasks can benefit from additional data augmentations and advanced model architectures. The learning curves of the nodule type classification model (see Appendix) showed that this model is severely overfitting the training data, indicating a need for improvement. In particular, incorporating more data augmentations can be a promising strategy to mitigate overfitting in the future.

During our analysis of the classification task, we noticed that the model performs reasonably well in classifying solid nodules. However, it is worse at classifying non-solid and part-solid nodules due to the limited number of samples available in the validation set. To address the issue of class imbalance, one potential strategy would have been to employ an oversampling technique specifically for non-solid, part-solid, and calcified nodules. It is important to note that the omission of oversampling in our study is a notable limitation, and we strongly recommend its implementation in future research endeavors.

For the malignancy risk estimation, the recall is much higher than the precision (0.922 versus 0.783). Having a higher recall is probably more important in cancer screening than high precision, as missing malignant nodules has higher costs than misclassifying benign nodules.

Finally, when using ensemble learning for all three tasks we were able to achieve an overall score of 0.816 on Grand Challenge. What is interesting is that ensemble learning did not always lead to better performance for the individual tasks. For example, the classification model with ensemble learning resulted in a lower performance than without using ensemble learning. We expected an increase rather than a decrease in performance here. Possible reasons for this unexpected decrease could be the stochastic nature of weight initialization and weight updating. Also, we did not apply any noise with ensemble learning right now. Perhaps this could also lead to a slight increase in performance.

Conclusion

This study focused on analyzing lung nodules and addressing three specific tasks: malignancy risk estimation, nodule type classification, and nodule segmentation. Our primary objective was to enhance the performance of the provided baseline models (U-net and 3D CNN). To achieve this, we employed ensemble learning across all three tasks and optimized the training procedure by incorporating data augmentation specifically for the segmentation task. However, our training procedure experiments did not yield a clear performance improvement over the baseline models. Conversely, the introduction of ensemble learning significantly improved the overall performance of our models. In the nodule type classification task, we identified severe overfitting, highlighting the need for further investigation and improvement. Nevertheless, for malignancy risk estimation, we achieved an AUC score of 0.967 on the validation set. Although there is room for improvement in the generalizability of our classification models, our study demonstrated the benefits of using ensemble learning for lung nodule analysis.

References

[1] P. Zhai, Y. Tao, H. Chen, T. Cai, and J. Li (2020). “Multi-task learning for lung nodule classification on chest CT.” IEEE access., vol. 8, 180317-180327.

[2] D. Müller, I. Soto-Rey, and F. Kramer. ”An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks.” Ieee Access., vol. 10, 66467-66480.

[3] W. Chen, Q. Wang, D. Yang, X. Zhang, C. Liu, and Y. Li (2020). “End-to-End multi-task learning for lung nodule segmentation and diagnosis.” In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 6710-6717). IEEE.

[4] B. Wu, Z. Zhou, J. Wang, and Y. Wang (2018, April). “Joint learning for pulmonary nodule segmentation, attributes and malignancy prediction.” In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) (pp. 1109-1113). IEEE.

[5] O. Ronneberger, P. Fischer and T. Brox (2015). ”U-net: Convolutional networks for biomedical image segmentation.” In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 234-241). Springer International Publishing.

[6] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker and D. Rueckert (2018). ”Attention u-net: Learning where to look for the pancreas.” arXiv preprint arXiv:1804.03999.

APPENDIX

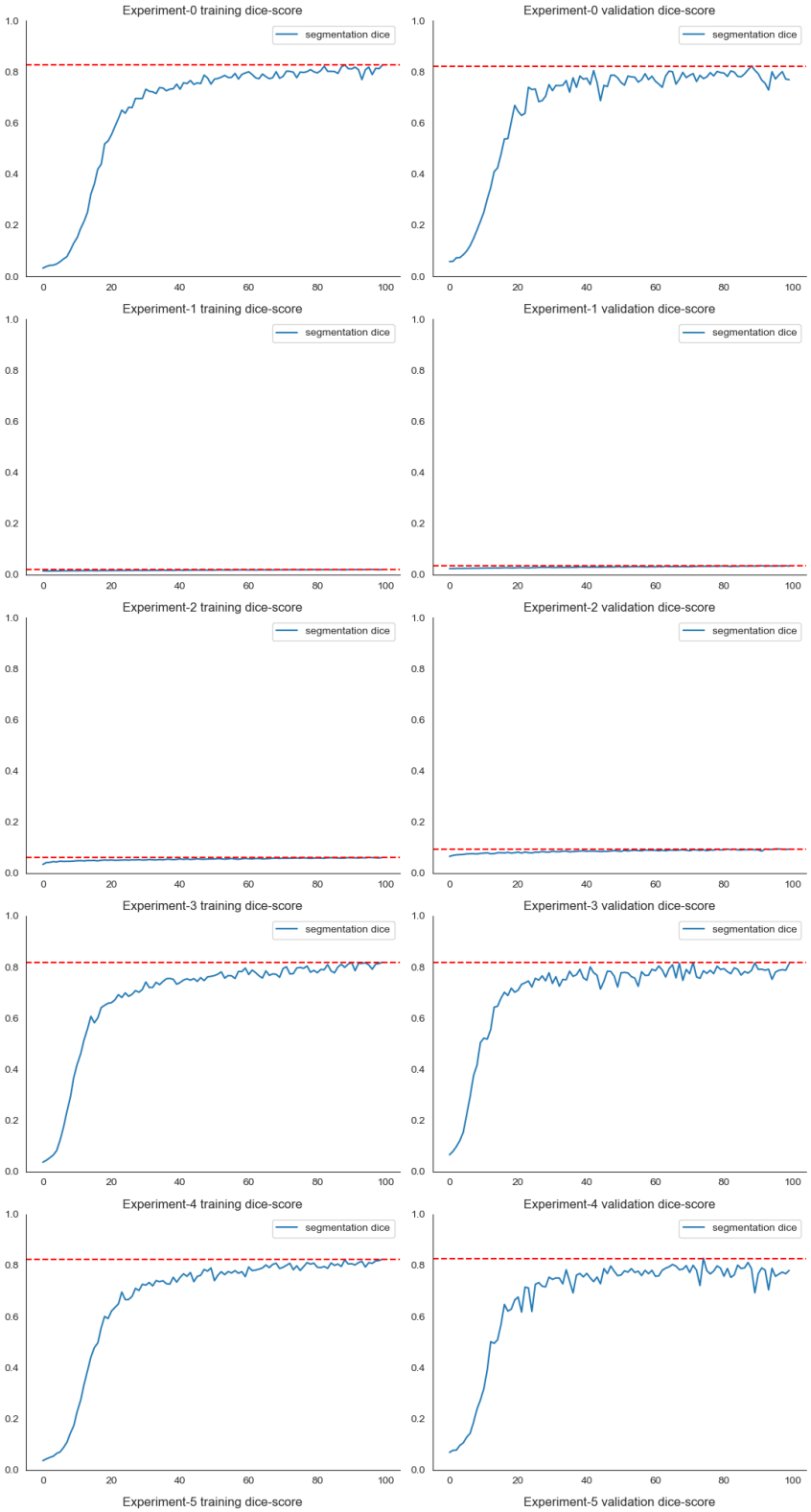


Fig. 2. Learning curves for nodule segmentation experiments 1-5

The learning curves for the training set (left) and validation set (right) from experiments 1-5 in Table I. The curves represent the DICE scores.

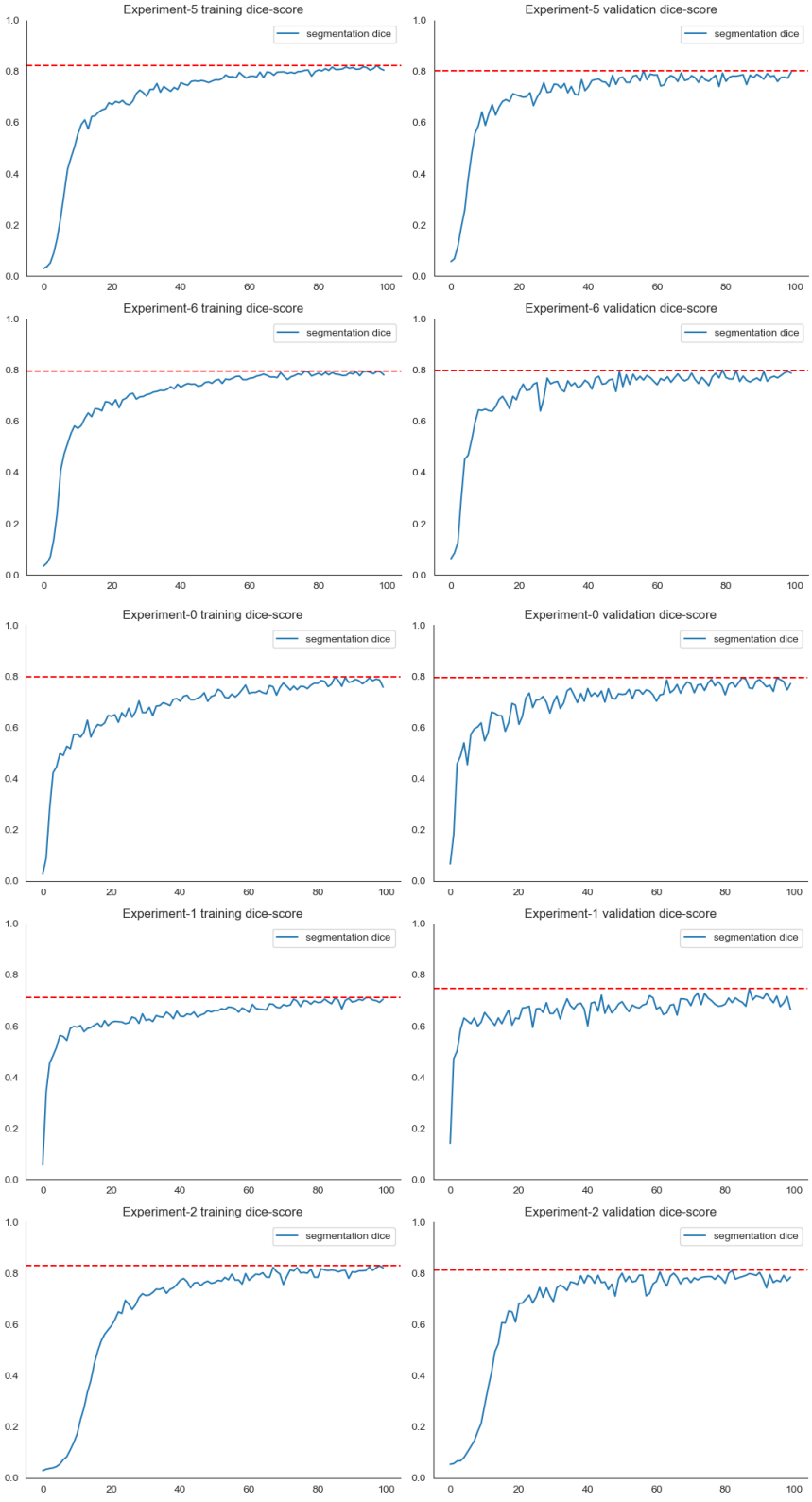


Fig. 3. Learning curves for nodule segmentation experiments 6-10

The learning curves for the training set (left) and validation set (right) from experiments 6-10 in Table I. The curves represent the DICE scores.

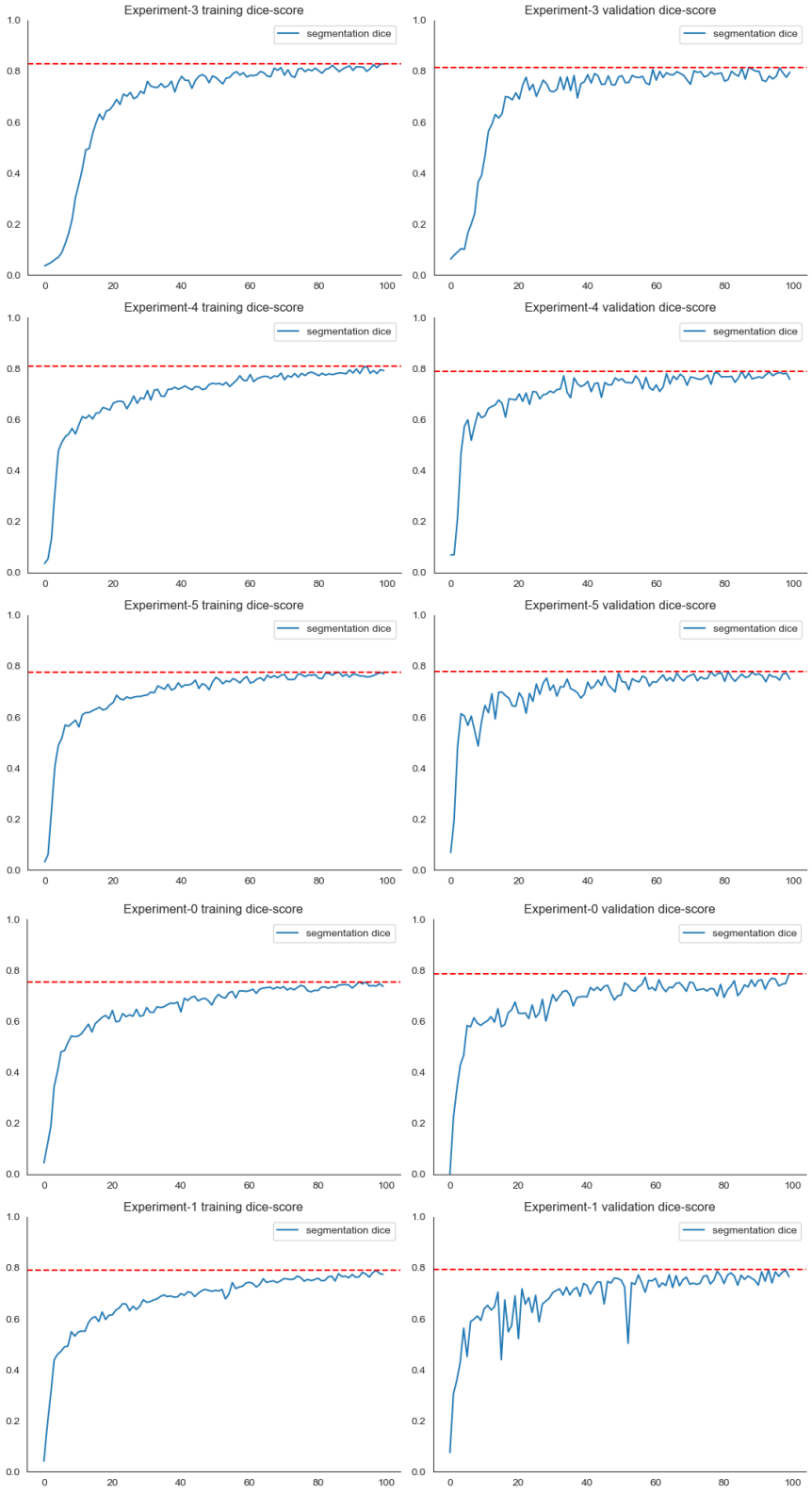


Fig. 4. Learning curves for nodule segmentation experiments 11-15

The learning curves for the trianing set (left) and validation set (right) from experiments 11-15 in Table I. The curves represent the DICE scores.

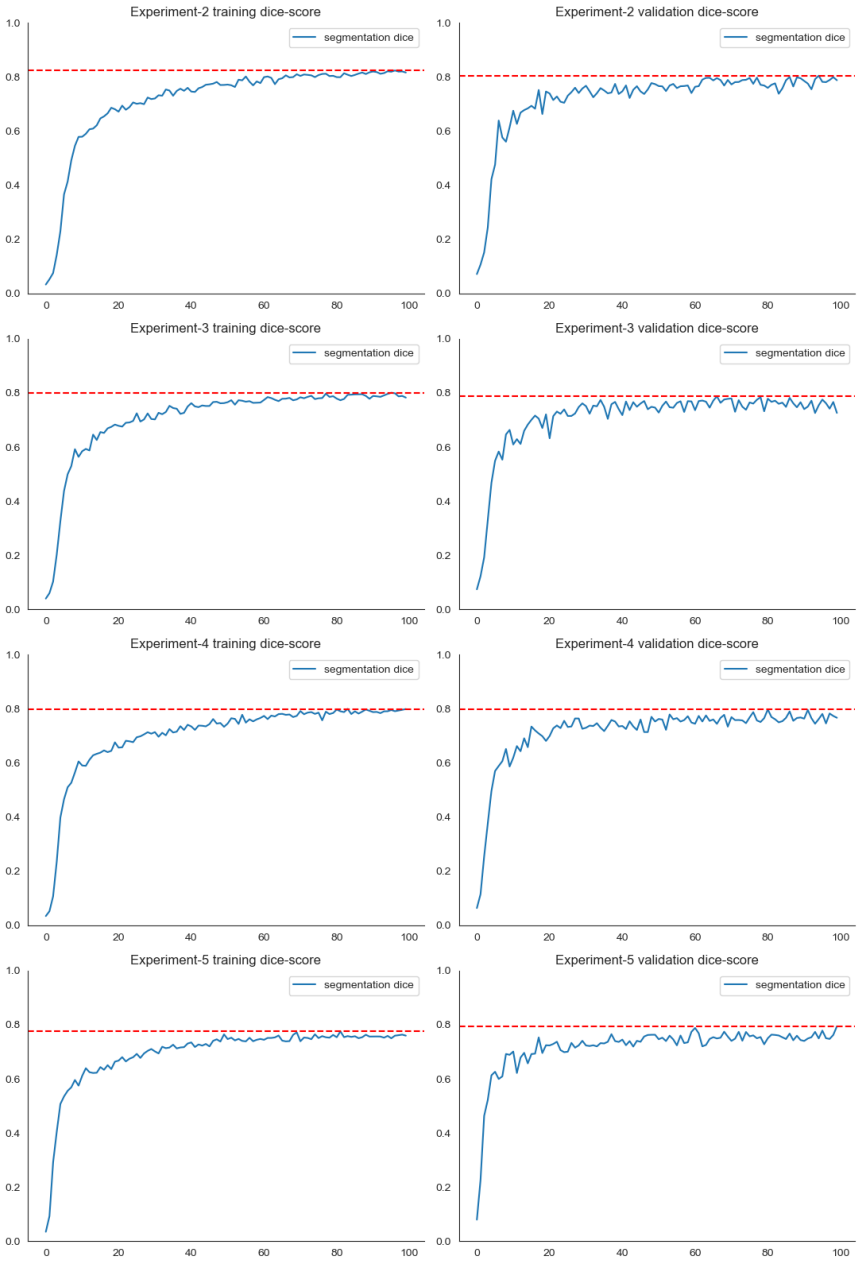


Fig. 5. Learning curves for nodule segmentation experiments 16-19

The learning curves for the trianing set (left) and validation set (right) from experiments 16-19 in Table I. The curves represent the DICE scores.

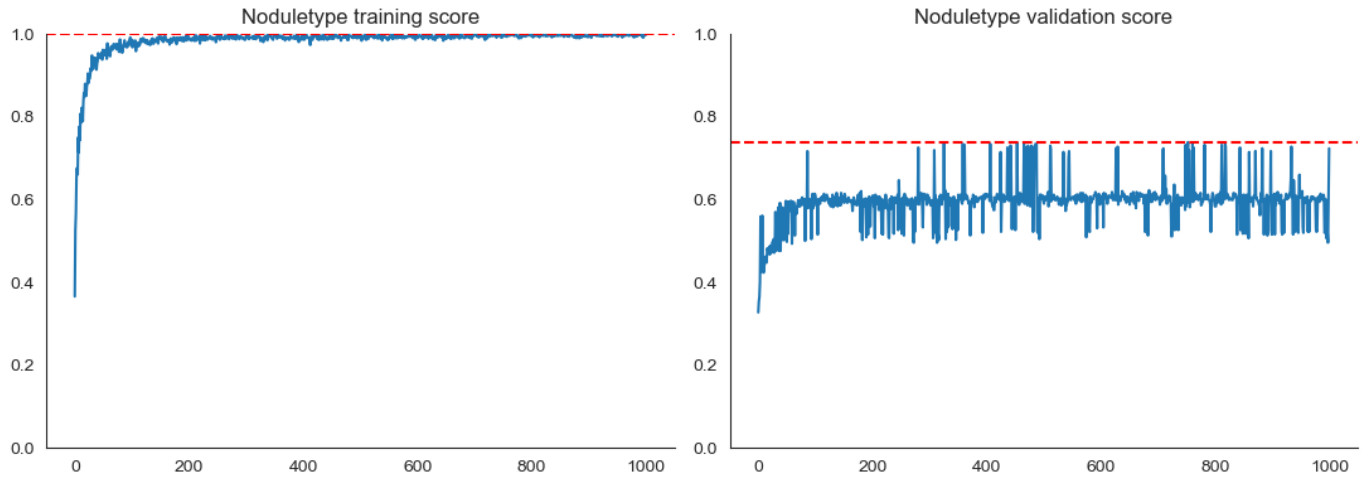


Fig. 6. Learning curves for nodule type classification

The learning curves for the training set (left) and validation set (right) for nodule type classificaton without ensemble learning. The curves represent the balanced accuracy scores. The best scores are 1.0 and 0.739 for the training and validation set respectively, shown by the red lines.