Maria Sailer*, Florian Schiller, Thorsten Falk, Andreas Jud, Sven Arke Lang, Juri Ruf, Michael Mix

Prediction of the histopathological tumor type of newly diagnosed liver lesions from standard abdominal computer tomography with a machine-learning classifier based on convolutional neural networks.

Abstract

Background and objectives: Liver lesions are a relatively common incidental finding in computer tomography (CT) of the abdomen. The current gold standard is liver biopsy, which has the downside of respecting only a small part of the total lesion volume. Furthermore, this invasive method carries interventional risks like bleeding or infection. Therefore, an image-based biomarker would be highly desirable. Conventional "radiomics" methods have often been utilized for similar problems, but the results are often not reproducible. This is mainly due to sampling errors and interobserver variability, but also the seemingly complex nature of the problem. We present a new approach that implements cutting-edge research in machine learning which is nevertheless cheap and easily applicable in a routine clinical setting. To achieve this, we use convolutional neural networks (CNN) to predict the histopathological findings from liver lesions from preoperative liver CT.

Methods: After splitting the study population into a training and test set we trained a CNN to predict the histopathological tumor type from CT data.

Results: The developed CNN workflow is able to predict liver tumor histology from routine CT images. We also evaluated in how far transfer learning and data augmentation can help in solving this problem and implemented the developed workflow in a clinical routine setting.

Conclusion: We propose a robust semiautomatic end-to-end classification workflow for the prediction of the histopathological type of tumor lesions based on abdominal CT and a deep convolutional neural network model. In our cohort, the model shows reliable and accurate results even with limited computational resources.

Keywords: Hepatic tumors, artificial neural network, convolutional neural network, machine learning

https://doi.org/10.1515/cdbme-2021-1032

1 Introduction

Deep learning algorithms based on convolutional neural networks (CNNs) are a powerful tool for many image classification tasks. Due to their outstanding performance in other domains, they are a promising method to facilitate advanced computer-aided-diagnosis for routine CT-imaging data. The bottleneck for this method is the annotation of imaging-data; currently this requires experienced experts, which is not only expensive but also largely qualitative and not standardized. Standard radiomic features have not made it into practical clinical use due to their sensitivity on interobserver-variability. The solution we propose for this problem consists of using quantitative outcomes as determined by functional imaging as ground truth and the combined use of different data sources.

***Corresponding author: Maria Sailer:** Department of Nuclear Medicine, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany, maria.sailer@yahoo.de
Florian Schiller, Juri Ruf, Michael Mix: Department of Nuclear Medicine, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany
Thorsten Falk: Department of Computer Science, Core Facility Image Analysis, University of Freiburg, Freiburg, Germany
Andreas Jud, Sven Arke Lang: Department of General and Visceral Surgery, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany

Artificial neural networks are directed, hierarchical, acyclic graphs where the nodes are called “neurons”. Each neuron receives several inputs, takes the weighted sum over its inputs and passes them to a nonlinear function called “activation function”, whose output is forward-propagated to nodes of the next network layer [1]. The last layer of the network is compared to the ground truth with a predefined “loss” or “error” function, resulting in a total error or „loss“ value for the given sample. The loss is subsequently minimized using gradient descent optimization with the so called backpropagation algorithm. More specifically, the parameter gradients are minimized after computing all of them with the backpropagation algorithm [2, 3]. The number of samples for which average loss and gradient are calculated is called batch-size. While simple artificial neural networks are fully connected, convolutional neural networks have the property of local connectivity and shared weights, which also leads to a hierarchical, representation of image features naturally adapted to the local redundancy of images. The latter corresponds to learnable filter kernels. For image processing tasks, convolutional neural network architectures are a more efficient way to process information [4].

2 Methods

The dataset consists diagnostic CT images with contrast agent (arterial and venous phase) of the abdomen from 38 patients admitted to the University of Freiburg Medical Center with advanced hepatic tumors and the histopathological diagnosis written by senior pathologists. CT Images were acquired following S3 guidelines of the Association of the Scientific Medical Societies (AWMF, No. 032/053OL).

For the prediction of the liver function as determined by hepatobiliary scintigraphy all patients in the dataset were included. The sampling into training, validation and test set was done in a stratified manner to ensure equal representation of classes. For random sampling, random permutations of the IDs (identification numbers) within those groups were done.
For each patient five slices in different contrast-enhancement phases were used. Validation was done on a randomly determined set of patients. For validation one slice per patient was used.

All tomographic images were converted from a DICOM (NM or CT) to 16-bit PNG. The transversal slice showing the portal vein bifurcation was defined as reference and the two slices above and below this level were taken. Native CT, venous and arterial phase images were included. No further segmentation was done. This approach requires a minimum of simple user interactions saves computing time and reduces hardware requirements. To keep costs low and the whole workflow affordable no GPU's were used. Instead, the analysis was performed on a CPU to evaluate in whether sufficiently precise results can also be obtained without expensive additional hardware. All analyses were performed in Python 3.6 using the deep learning library Keras [5] based on Google’s Tensorflow library [6].

First, simple convolutional neural networks were trained from scratch in different experiments corresponding to different hyperparameter configurations.
Second, the Keras implementations of ResNet50 and VGG16 CNN architecture were used [7, 8]. Both architectures achieved very good classification performances on ImageNet [9]. The network we used was based on those architectures, pre-trained on approximately 1.2 million images from ImageNet and re-trained on the CT-images collected in our study. The fully-connected layers at the end of the network for ImageNet were removed and replaced by a different fully connected network.

To prevent overfitting random dropout was used. The dataset consist 1710 images (n = 38, 15 images per patient, 3 contrast phases per image) was divided randomly into three parts:
- Training: The radiological data of 25 patients was used to train the network,
- Validation: The image data of 5 patients was used for validation and
- Testing: The records of 8 independent patients whose image data were acquired a few months later served to test the algorithm.

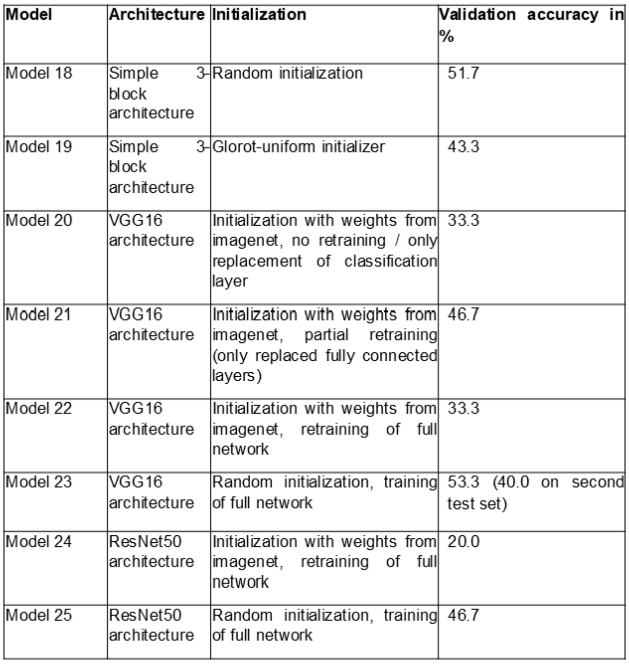
The performance of the algorithm was evaluated by using the accuracy and the area under the receiver operating curve (AUC) by plotting sensitivity versus 1 - specificity in the testing set. Training tiles were automatically resized to 224 x 224 pixels (the fixed input dimension of the VGG16 and ResNet50 CNNs). Image data augmentation was applied to increase the variety of the training data. For the transfer learning task the ResNet50 and VGG16 architecture were compared for both the prediction of functional reserve and tumor histology. For histology prediction random initialization and initialization with a Glorot- uniform method were used as baseline for comparison. The output of the network was categorical with three classes, constituting the most common types of hepatic tumors, that is hepatocellular carcinomas, cholangiocellular carcinomas and colorectal liver metastases. Patients with other, rare tumor types were not included to avoid overfitting to nonrepresentative samples. All models were trained by optimizing a categorical cross-entropy-loss function.

3 Results

The median total liver volume in the study population was 1817ml (with a mean of 1988 ml and a standard deviation of +/- 811 ml).

To predict class membership we used a three-class supervised model with different initialization schemes to test whether transfer learning was superior to random parameter settings. Both random initialization, full or partial weight transfer and a Glorot-uniform-initializer were used to determine a set of initial weights. The latter draws random values from a normal distribution with a mean of zero and a variance that is the multiplicative invert of the number of incoming neurons. The results of all models are shown in Table 3.

Table 3: Initialization and performance of all models that were tested in our study.

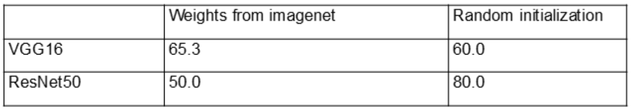


Model	Architecture	Initialization	Validation accuracy in %
Model 18	Simple 3-block architecture	Random initialization	51.7
Model 19	Simple 3-block architecture	Glorot-uniform initializer	43.3
Model 20	VGG16 architecture	Initialization with weights from imagenet, no retraining / only replacement of classification layer	33.3
Model 21	VGG16 architecture	Initialization with weights from imagenet, partial retraining (only replaced fully connected layers)	46.7
Model 22	VGG16 architecture	Initialization with weights from imagenet, retraining of full network	33.3
Model 23	VGG16 architecture	Random initialization, training of full network	53.3 (40.0 on second test set)
Model 24	ResNet50 architecture	Initialization with weights from imagenet, retraining of full network	20.0
Model 25	ResNet50 architecture	Random initialization, training of full network	46.7

The best model for the prediction of tumor histology was a model based on the VGG1 architecture and random initialization with an accuracy of 40% on the independent test dataset.

The predictive performance of models based on the VGG16 and ResNet50 architectures was determined with transferred weights from ImageNet. The accuracy of the corresponding models is summarized in Table 4. The best result with an accuracy of 80% was achieved with a randomly initialized ResNet50 architecture.

Table 4: Accuracy [%] of all models with complex architectures and weight transfer.



	Weights from imagenet	Random initialization
VGG16	65.3	60.0
ResNet50	50.0	80.0

4 Discussion

Faster convergence to the optimal solution and higher accuracy for the prediction of tumor histology was achieved with simple models. The initialization scheme had no significant influence on both. The best model was based on the VGG16 architecture trained from scratch with random initialization. The second-best model was a simple network trained from scratch with random initialization, suggesting that transfer learning from standard image domains does not yield relevant advantages over training from scratch with random initialization. Furthermore, small networks with few layers and significantly lower computational effort also yield reasonable results. This is in accordance with recent literature [10], where it has also been confirmed that small networks give higher predictive performance than standard machine learning approaches with conventional feature engineering and feature selection. One reason may be, that the data that pre-trained models are based on, have little similarity to biomedical image data. This may lead to a bad initialization - in some cases near local minima of the gradient function - which in some cases may not even be surmountable by adaptive learning rates.

We showed that simpler networks have a better computational cost/performance tradeoff and that good performance can also be achieved with only minimal preprocessing and without much cost. If a complex architecture like VGG16 or ResNet50 architecture is chosen, there is no relevant difference between these two options. Furthermore, as histology is a very complex endpoint much more data may be required to build better predictive models. Although transfer learning has been praised as the solution to sparse data in other domains, for the question we tried to solve transfer learning from domains unsimilar to medical images does not bring any benefit. Therefore, there is an urgent need to generate pre-trained models from a large amount of biomedical imaging data.

This study has the following limitations: First, the number of samples in the dataset was very limited; as medical imaging data is very expensive this may be called a general limitation of this kind of data, though. Second, due to the small number of samples even after data augmentation only hold-out validation could be implemented to compare the trained models, as in k-fold-crossvalidation the dataset is split into k equally sized folds for which the amount of data was not sufficient. Third, only a limited number of architectures and hyperparameters could be tested over comparatively few epochs due do computational resource constraints, as all computations were required to terminate within a reasonable time frame on a CPU.

To summarize our findings, this paper represents three major contributions to the biomedical image analysis literature. First, to the best of our knowledge, it presents the first study on the use of one imaging modality as a ground truth for building prediction models from another imaging modality. Second, we offer a framework for an affordable, easily implementable prediction model which is based on state-of-the art computer vision algorithms sin the preoperative setting for advanced hepatic tumor surgery. Third, we identified good hyperparameter configurations and data augmentation schemes for the predictive analysis of abdominal CT-images using convolutional neural networks.

5Conclusion

Recent advances in the development of convolutional neural network architectures and deep learning libraries allow that these algorithms now perform tasks which were previously the exclusive domain of human experts. We showed this in case, that simple models yield comparative results to deep models initialized with pre-trained models from Imagenet, where random initialization may be the best choice.

Author Statement

Research funding: The author state no funding involved. Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent has been obtained from all individuals included in this study. Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

References

[1] Goodfellow I, Bengio Y, Courville A (Eds.), Deep learning, MIT Press, Cambridge (2015), pp. 162-481

[2] Rumelhart D, Hinton G, Williams R: Learning representations by back-propagating errors. In: Nature. Band 323, 1986, pp. 533–536.

[3] Cui, Nan (2018). Applying Gradient Descent in Convolutional Neural Networks. In: J. Phys.: Conf. Ser. 1004, pp.12027. DOI: 10.1088/1742-6596/1004/1/012027.

[4] Krizhevsky, A, Sutskever, I, Hinton G. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[5] Abadi M et al., TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org

[6] Chollet, F (2015) keras, GitHub. http: // github . com / fchollet / keras

[7] He, K; Zhang, X; Ren, Shaoqing; S(2015): Deep Residual Learning for Image Recognition http: // arxiv. org / pdf / 1512.03385v1.

[8] Simonyan K; Zisserman A (2014): Very Deep Convolutional Networks for Large-Scale Image Recognition.http: // arxiv. org / pdf / 1409.1556v6.

[9] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision. 2015;115:211–252.

[10] Du H, Ghassemi M, Mengling F (2016): The effects of deep network topology on mortality prediction. In: Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference 2016, pp. 2602–2605. DOI: 10.1109/EMBC.2016.7591263.