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

Applied machine learning for liver surgery

The prediction of liver function from routine CT-images with convolutional neural networks.

***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

Abstract

Background and objectives: Both hepatic functional reserve and the underlying histology are important determinants in the preoperative risk evaluation before major hepatectomies. In this project we developed a new approach that implements cutting-edge research in machine learning and nevertheless is cheap and easily applicable in a routine clinical setting is needed.

Methods: After splitting the study population into a training and test set we trained a convolutional neural network to predict the liver function as determined by hepatobiliary mebrofenin scintigraphy and single photon emission computer tomography (SPECT) imaging.

Results: We developed a workflow for predicting liver function from routine CT imaging data using convolutional neural networks. We also evaluated in how far transfer learning and data augmentation can help to solve remaining manual data pre-processing steps and implemented the developed workflow in a clinical routine setting.

Conclusion: We propose a robust semiautomatic end-to-end classification workflow for abdominal CT scans for the prediction of liver function based on a deep convolutional neural network model that shows reliable and accurate results even with limited computational resources.

Keywords: Machine learning, artificial neural network, convolutional neural network, hepatobiliary scintigraphy, liver function

https://doi.org/10.1515/cdbme-2021-1033

1 Introduction

The preoperative liver function is important to estimate the risk of adverse outcomes after major hepatic resections. While laboratory and breath tests have known prognostic value in chronic liver disease, these variables do not reflect the degree of liver fibrosis or the distribution of functional reserve to guide the planning process before major hepatectomies because of portal hypertension, which are often present in chronic liver disease [1]. CT based liver volumetry is often used in the preoperative risk assessment [2]. In chronic liver disease or liver damage, however, CT-volumetry does not accurately predict the function of the remnant liver [3]. Planar dynamic hepatobiliary scintigraphy and single photon emission computer tomography (SPECT) with [99mTc]mebrofenin are established methods in nuclear medicine for this task [4]. Predicting the early postoperative recurrence of hepatocellular carcinoma from standard preoperative CT-imaging has been attempted in the literature [5], but to the best of the authors knowledge, there is no literature that attempts to predict liver function from this imaging modality. combined use of different data sources.

The implementation of an artificial neural network is a directed, acyclic, hierarchical graph. Its nodes, the neurons, take a weighted sum of inputs and transform them to the output by using a so called “activation function”. This output serves as input to nodes of the next network layer [6]. Network training minimizes the so-called objective or loss function which compares the network output to the ground truth. The parameters of this function are the network weights. The back-propagation algorithm [7,8] iteratively updates them starting at the last network layer towards the input typically using stochastic gradient descent.

2 Methods

[99mTc]Mebrofenin dynamic was done according to the protocol given in [9]. Additional diagnostic CT scans with contrast agent (arterial and venous phase) were acquired following the S3 guidelines of the Association of the Scientific Medical Societies (AWMF, No. 032/053OL).

The dataset consists of CT, the liver uptake rate indicating liver function as determined by hepatobiliary scintigraphy and liver tumor histology as diagnosed by a senior pathologist from 35 patients admitted to the University of Freiburg Medical Center with advanced hepatic tumors.

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 within those groups were done.

All tomographic images were converted from DICOM (NM or CT) to a 16-bit PNG format. 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 interaction. To keep costs low, the analysis was performed on a CPU to evaluate whether sufficiently precise results can also be obtained without expensive additional hardware (e.g. GPUs). All analyses were performed using Python 3.6 and the deep learning library Keras with Tensorflow as backend library [10].

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 architectures were used [11,12]. Both architectures achieved a very good classification performance on ImageNet [30]. The dataset was divided randomly into three parts: 80% of the data was used for training (300 images from 15 patients), 10% of this set for cross validation and 75 images from 5 independent patients for testing. The performance of the algorithm was evaluated by using the accuracy and the area under the receiver operating curve (AUC). Image data augmentation was applied to increase the variety of the training data; for training, scaling, zooming and shearing of 20%, 40%, and 60% were applied. The predictive performance of different adaptive gradient descent optimizers with momentum was compared. The maximum number of epochs trained was 100 due to computational constraints. The output of the network was binary (liver clearance rate below or above 5 %/(m²*min)).

RMSprop is an unpublished, adaptive learning rate method developed by Geoffrey Hinton. It was developed to solve the problem of radically diminishing learning rates of the optimization algorithm Adadelta [13].

The update rule for RMSprop is:

$$\begin{aligned} E[g^2]_t &= 0.9E[g^2]_{t-1} + 0.1g_t^2 \\ \theta_{t+1} &= \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t \end{aligned}$$

where Θ are the parameters, t is the time-step, η is the learning rate, gt is the gradient, Et is a matrix of the sum of the squares of gradients up to time step t, and ε is a smoothing term that avoids division by zero. RMSprop divides the learning rate by an exponentially decaying average of squared gradients. As a default setting, gt is set to 0.9 and the learning rate η was set to 0.001.

All models were trained by optimizing a cross-entropy loss function, with a binary cross-entropy-loss.

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 224x224 pixels. Image data augmentation was applied to increase the variety of the training data. The predictive performance of different adaptive gradient descent optimizers with momentum was compared. The maximum number of epochs trained was 100 due to computational constraints.

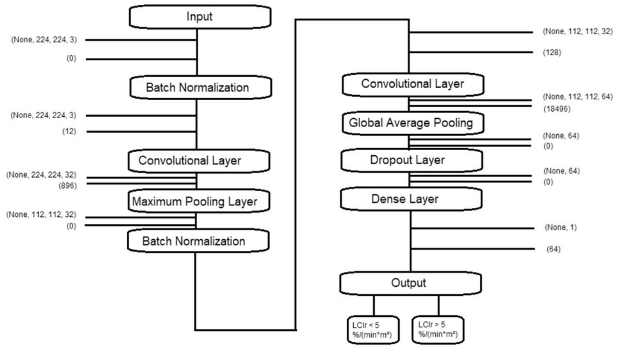

Figure 1: Flowchart for the simple 3-block-model.

3 Results

The mean total liver volume in the study population was 1988 ml (with a standard deviation of +/- 811 ml), the mean total liver function as quantified by the liver clearance rate LClr was 4.6 %/min/m² (with a standard deviation of +/- 1.7 %/min/m²).

The Pearson correlation coefficient between total liver volume and total liver function was 0.31, suggesting a very weak association of the two parameters.

The best performance for the prediction of the hepatic functional reserve as quantified by the liver clearance rate was achieved with the simple 3-block-model, with an accuracy of 81% on the validation dataset which was not used for training and with a ROC-AUC of 0.98, a F1-score of 0.86 and a sensitivity of 75%.

Local functional deficits were more common among large (>1cm) colorectal tumor metastases and hepatocellular carcinomas. Even large cholangiocellular carcinomas - given that there is no cholestasis (6 of 14 cases with cholangiocellular carcinomas, i.e. 40%) - had normal liver clearance rates (> 5%/m²*min) and no qualitative local deficit as shown on SPECT-images.

Three different data augmentation scales were tested (20%, 40% and 60% scaling, zooming and shearing). The best result was obtained using scaling, zooming and shearing transformations up to a strength of 40%, which yielded an accuracy of 62% (with a ROC-AUC of 0.80), suggesting that this may be a good cutoff value.

Finally 5-fold cross-validation was performed to evaluate the results of our simple network, yielding a validation accuracy of 75.0%.

The predictive performance of models based on the VGG16 and ResNet50 architectures was determined with transferred weights from ImageNet. The best result with an accuracy of 80% was achieved with a randomly initialized ResNet50 architecture. To prevent overfitting random dropout was used.

4 Discussion

To summarize our findings, this paper represents three 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 CNN.

One of the main obstacles for the training and deployment of machine learning models in clinical workflows may be the lack of training data due to high acquisition costs. While increasing data variability by data augmentation has been shown to be also beneficial in biomedical image analysis [14], the optimal extent of this approach must still be determined. We included therefore the comparison of several data augmentation schemes in our analysis. The RMSprop optimization algorithm was used to test a novel approach, as it is a popular and powerful algorithm in the machine-learning community but - likely due to its unpublished nature - has not been used for medical applications so far.

Faster convergence to the optimal solution and higher accuracy was achieved with simple models, especially for the prediction of tumor histology. The initialization scheme had no significant influence on both. The best model was a simple network trained from scratch with random initialization. We demonstrated that smaller networks with few layers and significantly lower computational effort also yield reasonable results. This is in accordance with recent literature [15], 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 pretrained models are based has 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.

The normal range for the liver clearance rate is reported to be 8.5±1.7 (SD) %/min/m². The preoperative cutoff value 5 %/min/m² was chosen to allow for a removal of about half of the liver volume (before or after preconditioning), as 2.7 %/min/m² is seen to be the tolerable minimum value, regardless of the presence of liver disease [16].

A major advantage of our approach is that annotated data can be generated from objective parameters like laboratory tests (in the case of global liver function) or functional imaging like hepatobiliary scintigraphy and SPECT (in the case of local distribution of liver function) without requiring relevant user-interaction. Another important aspect of our study was the use of CT data from clinical routine. These data are typically affected by different noise and variable imaging protocols from different CT vendors, even if they were acquired following consensus guidelines.

This study has two limitations: First, the number of samples in the dataset with hepatobiliary function as ground-truth was very limited; as medical imaging data, especially functional imaging data are very expensive, this may be in general a limitation of this kind of data. Second, 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.

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.

5 Conclusion

Recent advances in the development of CNN architectures and deep learning libraries allow that these algorithms now perform tasks which were previously the exclusive domain of human experts. Moreover, CNN can also be used to predict objective and therefore automatically producible labels based on functional imaging studies. 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. This can only be overcome as soon as models pre-trained on radiological imaging data are available.

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] Hoekstra L. et al. (2013): Physiological and biochemical basis of clinical liver function tests: a review. In: Annals of surgery 257 (1), S. 27–36. DOI: 10.1097/SLA.0b013e31825d5d47.

[2] Hackl C., Schlitt, H. J., Renner P., Lang S. A. (2016): Liver surgery in cirrhosis and portal hypertension. In: World journal of gastroenterology 22 (9), pp. 2725–2735. DOI: 10.3748/wjg.v22.i9.2725.

[3] Truant S. et al. (2015): Liver function following extended hepatectomy can be accurately predicted using remnant liver volume to body weight ratio. In: World journal of surgery 39 (5), pp. 1193–1201. DOI: 10.1007/s00268-014-2929-9.

[4] Kotani K. et al. (2018): Heterogeneous liver uptake of Tc-99m-GSA as quantified through SPECT/CT helps to evaluate the degree of liver fibrosis: A retrospective observational study. In: Medicine 97 (31), e11765. DOI: 10.1097/MD.0000000000011765.

[5] Cieslak K. et al. (2016): Measurement of liver function using hepatobiliary scintigraphy improves risk assessment in patients undergoing major liver resection. In: HPB : the official journal of the International Hepato Pancreato Biliary Association 18 (9), S. 773–780. DOI: 10.1016/j.hpb.2016.06.006.

[6] Zhou Y. et al. CT-based radiomics signature: a potential biomarker for preoperative prediction of early recurrence in hepatocellular carcinoma. Abdom Radiol (NY). 2017 Jun;42(6):1695- 1704. DOI: 10.1007/s00261-017-1072-0.

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

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

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

[10] Ekman M., Fjälling M., Friman S., Carlson S., Volkmann R. (1996). Liver uptake function measured by IODIDA clearance rate in liver transplant patients and healthy volunteers. In: Nuclear medicine communications 17 (3), pp. 235– 242.

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

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

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

[14] Cao C. et al. (2018): Deep Learning and Its Applications in Biomedicine. In: Genomics, proteomics & bioinformatics 16 (1), S. 17–32. DOI: 10.1016/j.gpb.2017.07.003.

[15] Roth H. R., Lee C. T., Shin H. C., et al. Anatomy-specific classification of medical images using deep convolutional nets. IEEE 12th International Symposium on Biomedical Imaging (ISBI); April 2015; Brooklyn Bridge, NY, USA. pp. 101– 104.

[16] Tann M., Woolen S., Swallen A., Nelson A., and Fletcher J. Establishing a normal reference range for mebrofenin clearance rate (MCR) for overall liver function assessment. J Nucl Med 2015 56:49