TumorEncode - Deep Convolutional Autoencoder for Computed Tomography Tumor Treatment Assessment

Alexander Katzmann*†, Alexander Mühlberg*, Michael Sühling*,
Dominik Nörenberg‡, Julian Walter Holch§, and Horst-Michael Groß†

*Siemens Healthcare GmbH
Department CT R&D Image Analytics
91301 Forchheim, Germany

†University of Technology Ilmenau
Neuroinformatics and Cognitive Robotics Lab
98693 Ilmenau, Germany

University Hospital Großhadern, Ludwig-Maximilians-University Munich
§Department of Internal Medicine III, Comprehensive Cancer Center
‡Department of Radiology
81377 Munich, Germany

E-Mail: alexander.katzmann@siemens-healthineers.com

***Abstract*—In tumor therapy, estimating tumor growth is crucial to get an early information regarding tumor therapy response and, if neccessary, adapt therapy. We propose a novel deep learning based algorithm using deep convolutional sparse autoencoders to find a minimal representation of tumor shape and texture for colorectal liver metastases. Furthermore, we provide a prediction of future lesion growth based on single slice CT tumor images which prospectively can be used as a prognosis for physicians. The state of the art in tumor treatment assessment for solid tumors mainly uses tumor diameter in single CT slices as the treatment response criterion (RECIST). However, whereas the correlation between RECIST and final treatment outcome was shown to be significant, its effect size is still limited. With our approach we achieve a Matthews correlation coefficient of 52.0% in predicting tumor treatment response compared to 28.2% with radiologic assessment, as well as an AUC of 0.814 opposed to 0.698.**

I. Introduction

In the last years, Deep Neural Networks have been applied to a variety of medical applications, including tumor [1], multiple sclerosis [2] and whole-organ segmentation [3], [4], vessel tracking [5] and others. Another major field which could highly benefit from semi-automatic treatment assessment is oncology, as a manual assessment requires consideration and evaluation of various variables and, therefore, a high amount of attending oncologist's and radiologist's experience. According to [6], [7], in particular tumor treatment assessment includes, but is not limited to, evaluation of:

- visual appearance (e.g. shape, size, density)
- blood values (e.g. haemoglobin, antibodies, tumor markers, e.g., CA19.9)
- histological assessment
- demographic data (e.g. age, gender, ...)
- patient's medical history

Data acquisition, however, is only the first step as all information has to be merged for giving a final state estimate, and subsequently, to decide on an appropriate treatment plan. Most of these estimates are based on clinical experience and require a high amount of expertise, as they include implicit predictions on future course of disease.

Usually these estimates also include an implicit assumption on future tumor growth, as no growth and tumor shrinkage correlate with higher patient survival times [8]. Acquiring some of the aforementioned parameters, e.g. blood values, is highly complex as it requires laboratory diagnostics and/or additional technical and personnel resources. Having a simple, fast and reliable first assessment could therefore be highly beneficial, as it may enable better therapy planning, deeper insight into tumor growth dynamics, a greater patient turnover, a reduction of costs and waiting times and thus may have the potential to improve overall healthcare.

Current research already has shown correlations of tumor's visual appearance, disease progress and survival times to a

This work has received funding from the German Federal Ministry of Education and Research as part of the PANTHER project under grant agreement no. 13GW0163A. The concepts and information presented in this article are based on research and are not commercially available.

World Congress on Computational Intelligence / Int. Joint Conf. on Neural Networks (IJCNN), Rio de Janeiro, Brazil, pp. 3876-3883, IEEE 2018

certain degree: Starting with the first Radiomics publications [9], [10], semiautomatic tumor treatment assessment by using image analytics has become a highly active field [11]–[15]. Although most algorithmical approaches still concentrate on a combination of more classical image features like descriptive intensity histogram statistics, wavelet features or run-length-matrix descriptions, there is already a minor community of scientist using various neural network approaches for image based assessment. Most of this work, however, is focussed on segmentation tasks, as it is easy to generate high amounts of training data which are usually rare for medical tasks due to data privacy regulations. Amongst various tumor entities, colorectal cancer (**CRC**) is of particular interest, as it is the second leading cause of cancer related deaths in modern societies, being responsible for more than 50,260 deaths in the U.S. in 2017 alone [16]. More than 50% of patients with colorectal cancer develop liver metastases, which may lead to liver organ failure, additional organ compression and thus significantly reduce patient life time [17] [18].
Although late research already has shown that deep convolutional neural networks (**DCNN**) can successfully be trained to do familiar tasks, e.g. lung lesion malignancy classification [19], to the best of our knowledge currently no algorithm exists for colorectal cancer metastases assessment, especially doing a continous assessment in time domain. Therefore:

1) ... we present a novel approach which allows prediction of CRC metastases growth from single slice CT images of two treatment timepoints (before- and within-treatment),
2) ... we show that our approach can be utilized as a pre-treatment assessment using one time-point only,
3) ... we further show that our approach outperforms other approaches based on radiological assessment parameters, i.e. RECIST and volume (see sec. II), only .

II. Background

A patient with colorectal cancer usually is scanned with CT every 2-3 months (depending, e.g. on tumor stage), to rule out liver and/or lung metastases, being the most common sites for CRC, as metastases are correlated with significantly lower patient survival (non-metastatic: 53-92% depending on stage; metastatic: 11%) [20]. The first scan is usually called the *baseline* (**BL**), the latter *followup scans* (**FU1**, **FU2**, ...). Following the the gold standard for radiologic solid tumor assessment - Response Evaluation Criteria for Solid Tumors (**RECIST**) - CRC treatment includes measuring up to 5 (RECIST v1.0), respectively 2 (v1.1), target lesions per organ. Measure is taken as the largest lesion diameter within one slice [21].

According to RECIST's single lesion response criteria, tumor progress is described using the following rule table:

- (**PR**) Partial Response - shrinkage in tumor diameter of at least 30% compared to last scan
- (**PD**) Progressive Disease - growth in tumor diameter of at least 20% compared to last scan
- (**SD**) Stable Disease - neither significant growth nor shrinkage
- (**CR**) Complete Response - dissappearance of all target lesions and lymph node reduction to $\emptyset < 10\,\text{mm}$ in short axis

In accordance with the RECIST guideline, we measure tumor diameter as the longest diameter within one slice. We neglect Complete Response (CR) as it can be seen as a special case of PR. We choose computed tomography (**CT**) as the image modality. Although other approaches like positron emmision tomography (PET) inherently provide some advantages like direct correlations between images and cell metabolism, in CRC treatment CT is generally seen as the gold standard, as image acquisition is less costly, better available, easier to use and has higher quantitative interpretability.
As shown in sec. IV-C, single lesion assessment only shows minor correlation, both, with future tumor growth and patient survival time. A more reasonable prediction requires acquisition and integration of blood values, histology and demographic data into a joint model. It may therefore be highly beneficial to have a first, minimal-invasive assessment based on CT images, especially when CT data are standardly acquired in treatment process. Also it is possible, that tumor structure contains additional information which is not contained in histological or blood value data.

As mentioned in sec. I, applying deep learning to medical volume image analysis often requires multiple augmentation techniques, as, compared to most computer vision tasks, medical data is rare and hard to obtain. Data augmentation often can be done efficiently by using the sliding-window approach, potentially combined with multiple other augmentation techniques (e.g. affine image transforms). For segmentation the amount of network parameters also can effectively be reduced by using low-resultion images or stacked networks [4].

When having classification tasks especially image transformations like varying rotation, shear, jittering, etc. can be used. However, even when combined with dropout [22], batch normalization [23], 2D or 3D image augmentation, the degree to which augmentation results in additional performance is limited, as images are still highly correlated.

The limit to data augmentation especially holds true for our task, as the underlying theory claims that phenotypical manifestations (e.g. specific tissue structure, size and shape of central necrosis, etc.) correlate with image structures and/or noise patterns. As these manifestations are currently a field of active research, it can not be said whether larger transformations are realistic in these terms too. This reduces the amount to which image transformations can be done, so augmentation does not fully solve the problem of few data. Therefore, a main goal of the proposed approach was to keep the number of parameters as low as possible. We thus decided to use a 2D approach, as usual slice-wise CT reconstruction has much lower spatial resolution in z-axis ($\sim$ 1-3mm) than in x- and

y ($\sim$ 0.7mm) and so can more easily be ommitted while efficiently reducing the training parameter count.

III. Predicting tumor growth using sparse representation

We present an approach for using 2D deep convolutional autoencoders to train sparse embeddings which efficiently describe tumor variance. These sparse representations can subsequently be used to create an easier-to-train network with very low parameter count to predict future tumor growth.

A. Dataset

Our dataset consists of 321 volumetric CT scans from 135 patients taken between 12/2009 and 02/2017. The data contain 460 unique liver lesions with fully volumetric segmentations at in average 2,92 time points per patient. Totally this results in a dataset of 1344 liver lesion volumes. As the data were acquired retrospectively (and so no unified scan protocol was used), the images suffer from very high heterogeneity, resulting - in terms of images - in various contrast levels, illuminations, noise levels, resolutions, etc..

To unify the dataset, as a first step the volumetric image data were resampled to isotropic voxel size using bicubic interpolation. As described in sec. II, we decided to use 2D images to train our classifier. We chose to extract windows representing 80x80mm at each lesion's middle slice for each time point. The window sizes represent usually expectable values of tumor size in our dataset with quantiles of $\emptyset_{Pr(\emptyset)\leq 0.1}, \emptyset_{Pr(\emptyset)\leq 0.9}$ of 11.3mm, resp. 53.3mm.

The windows are extracted as lesion centered 256x256 pixel bicubic resamplings of the original images. This upsampling takes varying original voxel sizes into account, as small and big lesions should be equally well represented. Also, to reduce the influence of different contrast agent levels (and thus image illumination), histogram equalization was done as an additional step. An example of one baseline followup lesion pair is found in Fig. 1.

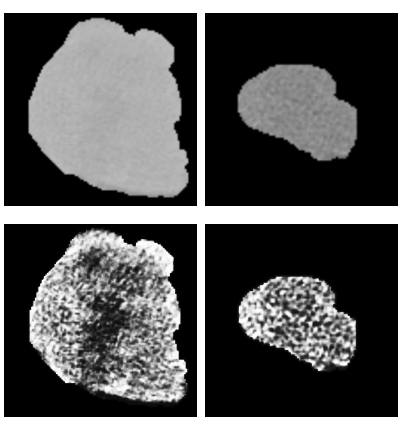

Fig. 1. Lesion example images as given as input for autoencoder and classification network. Top left: baseline image from 12/2015, followup from 04/2016. Bottom: Histogram equalized images.

B. Autoencoder Network Architecture

As already described in sec. II, a major goal was to keep parameter count as low as possible. Nevertheless, we experienced worse performance when using smaller images, as our lesion data highly vary in size and resolution and either lesions were not completely contained when using smaller windows, or could not be fully represented when using lower resolution. We take care of this by using 4x4 max pooling in the early layers to quickly reduce image size. We use batch normalization throughout all layers, as it is generally thought to be similarly regularizing as dropout [23] and in our experiments resulted in better generalization performance as well as lower training times. All layers except for the output layer use leaky rectified linear activation (Leaky ReLUs) with a slope of 0.2. The output layer uses $\tanh$-activation, input and output are scaled to the interval $[-0.5; +0.5]$. Also mean image subtraction was used. The complete architecture can be seen in Fig. 2, details for every layer can be found in Table I.

TABLE I
AUTOENCODER NETWORK ARCHITECTURE

type	filter	stride	reg.	output	# param
in			BN	2×256×256×1	
conv3d	2×1×1	1×1×1	BN	256×256×32	224
conv2d	5×5	1×1	BN	256×256×32	25760
pool	4×4	4×4	—	64×64×32	
conv2d	5×5	1×1	BN	64×64×48	38640
pool	4×4	4×4	—	16×16×48	
conv2d	3×3	1×1	BN	16×16×64	27968
pool	2×2	2×2	—	8×8×64	
conv2d	3×3	1×1	BN	8×8×96	55776
pool	2×2	2×2	—	4×4×96	
conv2d	3×3	1×1	BN	4×4×128	111232
flat				(2048)	
fc	(20)		L1+BN	(20)	41060
dense	(2048)		BN	(2048)	51200
reshape	(2048)		—	4×4×128	
dconv2d	3×3	1×1	BN	4×4×96	111072
up	2×2	1×1	—	8×8×96	
dconv2d	3×3	1×1	BN	8×8×64	55616
up	2×2	1×1	—	16×16×64	
dconv2d	3×3	1×1	BN	16×16×48	27888
up	4×4	1×1	—	64×64×48	
dconv2d	5×5	1×1	BN	64×64×32	38560
up	4×4	1×1	—	256×256×32	
dconv2d	5×5	1×1	BN	256×256×32	25760
dconv3d	2×1×1	1×1×1	BN	2×256×256×1	69

C. Predictor Network Architecture

To reduce the chance of overfitting, we chose to append a very simple, 2-layered network architecture consisting of 8 fully connected leaky ReLUs followed by a two neuron layer with softmax activation. Also we use batch normalization in

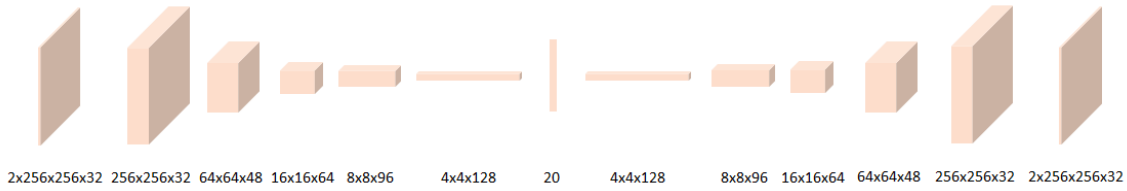

Fig. 2. Autoencoder used for training sparse representation of 20 neurons. One slice of baseline and followup images are given as input, resp. target output. An overview of all layers can be found in Table I.

the former layer. By doing so, the number of trainable parameters is reduced to 218. When training the classifier, we also analyzed more complex multi-layer-architectures with strong regularization but did not experience significant advantages. The final network architecture can be found in Table II.

TABLE II
PREDICTOR NETWORK ARCHITECTURE

type	filter	stride	reg.	output	# param
in				2×256×256×1	
conv3d	2×1×1	1×1×1	BN	256×256×32	224
conv2d pool	5×5 4×4	1×1 4×4	BN —	256×256×32 64×64×32	25760
conv2d pool	5×5 4×4	1×1 4×4	BN —	64×64×48 16×16×48	38640
conv2d pool	3×3 2×2	1×1 2×2	BN —	16×16×64 8×8×64	27968
conv2d pool	3×3 2×2	1×1 2×2	BN —	8×8×96 4×4×96	55776
conv2d flat	3×3	1×1	BN	4×4×128 (2048)	111232
fc	(20)		L1+BN	(20)	41060
dense	(8)		BN	(8)	200
dense	(2)		—	(2)	18

IV. EXPERIMENTS

A. Ground truth

We labeled our data by extracting the RECIST diameter of all lesions to any timepoint and determined the final label by:

$$y_i = \begin{cases} 1 & \text{if } \varnothing_{i,t+1}/\varnothing_{i,t} \geq 1.2 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

As already introduced in sec. II, gold standard radiologic assessment is based on RECIST. Our above definition matches the RECIST single lesion assessment criteria of significant growth. More formally our classifier goal is consistent to RECIST's discrimation criterion for *progressive disease* opposed to *partial response*, *complete response* and *stable disease*. As the RECIST diameter as used above is defined as the largest lesion diameter measurable in one slice of an image, notabely it does not take viewport into account but relies on the (usually accurate) assumption of isotropic lesion growth. Due to the differing target of classifying tumor growth starting from the current time point and not to get overall clinical parameters, opposed to RECIST we define growth or shrinkage relative to current timepoint t where RECIST compares to the best response t_{BR} [21].

B. Performance Measures

Especially for medical treatment choosing the right classifier can be crucial, as deciding whether to use a conservative or optimistic classifier highly depends on the concrete purpose as well as a critical cost-benefit-analysis. As we expect this analysis to be a case-to-case decision, one goal was to choose a measure which is invariant to this decision to the greatest possible extent. When evaluating our classifier, we decided to state various metrics for all trained classifiers (see Table III). We also provide a ROC-based AUC. For the final evaluation

TABLE III
OVERVIEW ON USED MEASURES

measure	equation
true positive rate, sensitivity, recall	$TPR = REC = \frac{tp}{tp+fn}$
true negative rate, specifity, inverse recall	$TNR = \frac{tn}{tn+fp}$
positive predictive value, precision	$PPV = \frac{tp}{tp+fp}$
negative predictive value, inverse precision	$NPV = \frac{tn}{tn+fn}$
F_1 score	$F_1 = \frac{2 \cdot tp}{2 \cdot tp+fp+fn}$
Informedness Youden's J statistic	$IFD = TPR + TNR - 1$
Markedness	$MKD = PPV + NPV - 1$
Matthews correlation coefficient/Φ-Score	$MCC = \sqrt{IFD \cdot MKD}$
Area under curve (ROC)	AUC

we choose Matthews correlation coefficient as the preferred measure, as according to [24] it has the following favorable properties:

1) It is a common measure. Known as Φ-coefficient, MCC also is a common measure in statistics for most scientific domains and therefore easy to understand for most scientists.
2) For guessing, it is zero-centered. This allows it to be directly interpretable as an informed measure of quality over guessing.
3) It can handle unbalanced datasets. This is important as tumors in treatment usually shrink, meaning negative

samples become overrepresented. Also it ensures that results are comparable across various datasets without being influenced by class distribution.

C. Baseline

As there is currently no real baseline for predicting future tumor growth of colorectal cancer liver metastases, we decided to train classifiers from radiologic assessment parameters. We extracted tumor volume and longest diameter in one slice (RECIST). Comparing against volume is important since lately tumor volume is assumed to be a more accurate predictor than RECIST [25] [26]. Based on these measures m_r we defined two new input sets X_r with $x_{r,i,t} \in X_r$:

$$x_{r,i,t} = \begin{pmatrix} m_{r,i,t} \\ m_{r,i,t-1} \\ m_{r,i,t} - m_{r,i,t-1} \\ \frac{m_{r,i,t}}{m_{r,i,t-1}} \end{pmatrix} \tag{2}$$

with $m_{r,i,t}$ being the measure $r \in \{\text{RECIST}, \text{Volume}\}$ for sample i at timepoint t. For each X_r we trained one classifier. Hyperparameter optimization was done using 100 iterations of randomized search cross validation with nested 10-fold grouped cross validation for inner validation. Train-test split is done with outer 10-fold grouped cross validation, where in both cases patient name was the grouping parameter. We also compared against uninformed guess (coin-flip-model), informed guess (stratified by label distribution), and most-frequent-class-guess (abbreviated as MFCG). Results can be found in Table IV.

TABLE IV
PERFORMANCE OF PREDICTORS FROM RADIOLOGIC DATA AND GUESSING

Measure	RECIST	Volume	Infd.	Uninfd.	MFCG
TPR	62.9%	**65.2%**	23.5%	50.0%	0.0%
TNR	63.4%	60.7%	76.3%	49.7%	**100.0%**
PPV	35.9%	**40.6%**	23.3%	23.6%	0.0%
NPV	85.2%	**85.6%**	76.2%	76.2%	76.3%
F_1	**44.4%**	44.0%	23.0%	31.8%	0.0%
IFD	25.0%	**26.7%**	0.0%	0.0%	0.0%
MKD	**22.4%**	21.2%	0.0%	0.0%	-23.7%
MCC	**28.2%**	27.1%	0.0%	0.0%	0.0%
AUC	**69.8%**	68.3%	(50.0%)	(50.0%)	(50.0%)

The results show high correlation between RECIST- and volume-based prediction, as well as minor correlation between RECIST, volume and future tumor growth. Both classifiers were significantly superior to guessing regarding $F1$, MCC and AUC with $p < .001$ each. The 95 % confidence intervals were $[.402, .486]$, $[.235, .328]$, $[.671, .726]$ for RECIST, and $[.398, .482]$, $[.220, .323]$, $[.652, .714]$ for volume-based prediction. Significance was tested using 5,000 iterations of bootstrapping. As Pearson correlation for dichotomous variables reduces to MCC (or Φ-coefficient), the results for RECIST- and volume-based prediction can be seen as an effect size of $r = .282/.271$. Also the results show that, as expected, some metrics are more error prone to unbalanced data than others. Expectedly, F_1, IFD, MKD, MCC and AUC seem to be more robust indicators for unbalanced sets.

D. Training process

All training was done on a NVIDIA DGX-1 using Keras with Tensorflow backend [27], [28]. As in IV-C, we divided our dataset into two sets with no shared lesions, scans or patients. Also we ensured similar label distribution amongst both sets.
Training took place in two steps. First, we trained the autoencoder to create sparse representations, second, we trained a classifier appended to the autoencoders sparse representation layer. For autoencoder training binary crossentropy loss was used. All optimization was done using Adam with Nesterov momentum as described in [29]. Although not generally seen as neccessary, we experienced better results when additionally limiting optimizer's parameter update by exponentially annealing the learning rate η_i as a function of the current epoch i:

$$\eta_i = \eta_0 \cdot \left(\frac{\eta_{n-1}}{\eta_0}^{\frac{i}{n-1}^{\gamma}} \right) \tag{3}$$

with start learning rate $\eta_0 = 3 \cdot 10^{-4}$, final learning rate $\eta_{n-1} = 1 \cdot 10^{-7}$, the number of epochs n and $\gamma = 1.2$ being the learning rate exponent. For both, autoencoder and classifier training, we used adversarial training as it was shown to regularize training in a way similar to dropout [30]. As our training dataset is highly unbalanced, we also employed stratified sampling in training, assigning a sampling probability p_i for each sample i with m classes as follows:

$$p_i = \frac{\frac{1}{Pr(y=y_i)}}{\sum_{k=0}^{m} \frac{1}{Pr(y=y_k)}} \tag{4}$$

We combined this with a modified version of the exact importance sampling from [31], multiplying the sampling probability with the norm of the error gradient for each sample, as we encountered huge benefits in training and test set performance.

1) Autoencoder: The autoencoder was trained for 1,000 epochs, finding the best model after epoch 713 with respect to train and test loss. As we chose a very sparse representation we did not encounter a significant difference between bias and variance. However, reconstruction quality is also very limited, which implies that the autoencoder network is not able to fully represent the learned dataset. Again, as our training data were very limited for a deep learning task, this may be preferable. An example baseline followup pair and its deconvolved encoding can be seen in Fig. 3.

2) Classification: For analyzing whether actual appearance, e.g. shape and texture, itself holds information on future prognosis, after training the autoencoder we trained two different models:

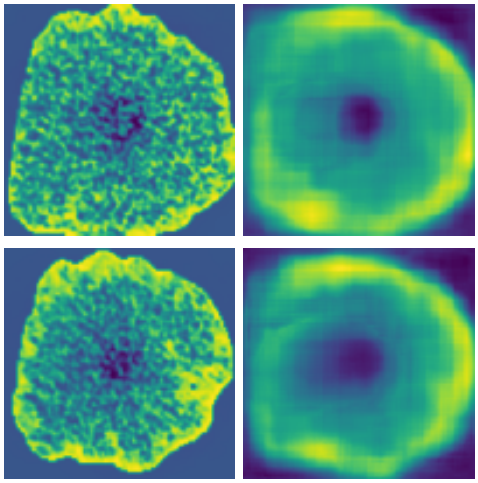

Fig. 3. Baseline (top) and followup images (bottom) in false colors; left: original image, right: autoencoder reconstruction

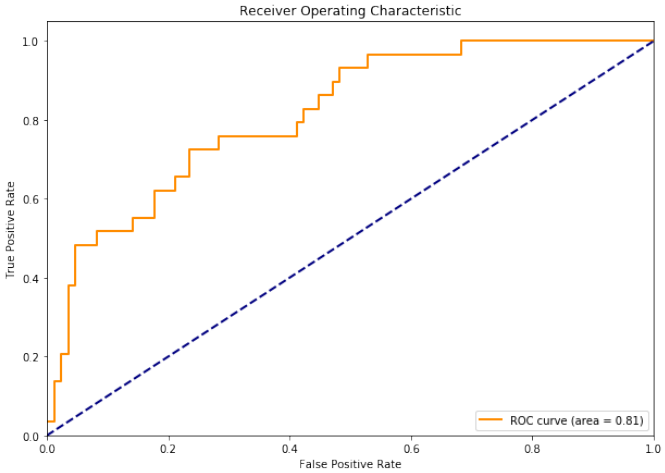

Fig. 4. Receiver operating characteristic for test set using sparse encoding based predictor with baseline + followup.

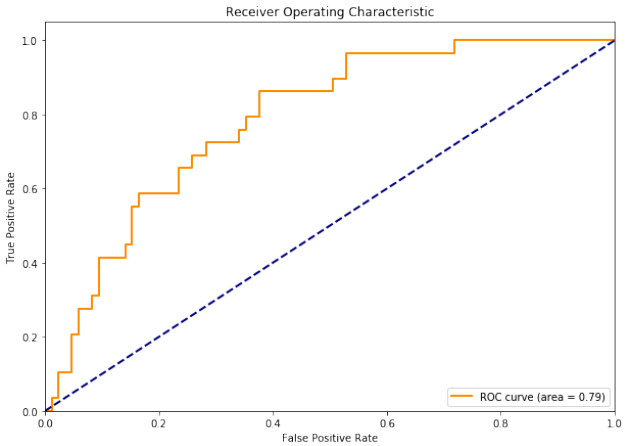

Fig. 5. Receiver operating characteristic for test set using sparse encoding based predictor using one timepoint only.

1) One classifier based on sparse representation of BL+FU
2) another classifier based on sparse representation of one time point only

The input of the first classifier consists of baseline-followup-pairs like seen in Fig. 1. For the second classifier we duplicated the image of the current time point and use the same autoencoder trained before. Training for both models was done for 200 epochs. However, training was fully satisfied after 10 iterations. Further epochs did not provide any advantages and lead to overfitting. Training the autoencoder was meant as a pretraining for the sparse representation layers, so we did not fix these in further training. When fixing the layers, AUC results were getting worse for $\sim 5\%$. Not fixing the convolutional layers results in a number of trainable parameters much higher than the one of training samples, so the resulting model is inherently prone to overfitting. The results of our training in comparison to the tested baseline classifiers can be seen in Table V. Fig. 4 and 5 show the ROC curves on the test set.

TABLE V
PERFORMANCE OF OUR APPROACH AND RADIOLOGIC DATA PREDICTION

Measure	Classifier (BL+FU)	Classifier (ONE)	RECIST	Volume
TPR	48.3%	**86.2%**	62.9%	65.2%
TNR	**95.3%**	62.4%	63.4%	60.7%
PPV	**77.8%**	43.9%	35.9%	40.6%
NPV	84.4%	**93.0%**	85.2%	85.6%
F_1	**59.6%**	58.1%	44.4%	44.0%
IFD	43.6%	**48.6%**	25.0%	26.7%
MKD	**62.2%**	36.8%	22.4%	21.2%
MCC	**52.0%**	42.3%	28.2%	27.1%
AUC	**81.4%**	78.7%	69.8%	68.3%

As in section IV-C, significance was tested using 5,000 iterations of bootstrapping. The 95 % confidence intervals for $F1$, MCC and AUC were $[.450, .726]$, $[.356, .694]$, $[.721, .896]$ for the BL+FU-classifier, and $[.455, .693]$, $[.283, .540]$, $[.694, .865]$ when using baseline images only. Thus, both classifiers perform significantly better than classification with RECIST or volume regarding $F1$ and MCC with $p < 0.05$. The classifier using baseline and followup image also reaches significant superiority with respect to AUC.

V. DISCUSSION

The results show that radiologic liver lesion images contain visual information which allows for the prediction of future tumor growth. Assuming most metrics, both approaches perform at least equally well or even better than RECIST or volume based prediction. This holds especially true for balanced or informed metrics, implying that not only tumor size or diameter are important predictors, but structural image information is even more predictive. Both classifiers outperform RECIST or volume based prediction in terms of F1-score, Informedness, Markedness and Matthews correlation coefficient. The results also show that expectedly classifier

metrics like TPR, TNR, PPV and NPV highly depend on the concrete choice of classifier output weighting. As our classifier was chosen to have a high MCC, our BL+FU classifier is much less optimistic than the BL only classifier. Nevertheless, the ROCs implys that both classifiers could be chosen with abritrary prioritization of classes, showing comparable MCC-values.

VI. Conclusion

As already mentioned in sec. I and II, predicting tumor growth is important to get an early assessment whether a patient responds to therapy or not. An algorithm extracting information which is not covered by usual radiologic assessment could be of high clinical value, eventually improve therapy and thus overall patient healthcare. However, for now we did not try to combine image data with clinical parameters, though we expect it to be highly beneficial for the classification goal. While tumor growth is known to correlate with patient lifetime, it should be analyzed whether DCNNs are directly predictive for patient lifetime. Also, the absence of tumor growth does not necessarily mean tumor shrinkage. It may thus be preferrable to separately predict shrinkage in order to distinguish it from non-growth.

One approach which is not covered in this study is the image based volumetric assessment of lesions, utilizing not only one slice but the whole tumor. However, as stated in sec. III-A, the 10% and 90% quantiles for our dataset were 11.3 mm, resp. 53.3 mm. Assuming isotropic growth and expecting usual slice thicknesses of 1-3mm, this results in the necessity of segmenting between 4 and 50 slices for fully volumetric lesion assessment. As our approach only requires segmentation of the lesion's middle slice, it provides high potential for practical use as it allows an early assessment with comparable (RECIST) or lower (volume) costs and significantly better performance than pure radiologic assessment.

References

[1] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.

[2] T. Brosch, Y. Yoo, L. Y. Tang, D. K. Li, A. Traboulsee, and R. Tam, "Deep convolutional encoder networks for multiple sclerosis lesion segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 3–11.

[3] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 556–564.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[5] A. Wu, Z. Xu, M. Gao, M. Buty, and D. J. Mollura, "Deep vessel tracking: A generalized probabilistic approach via deep learning," in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*. IEEE, 2016, pp. 1363–1367.

[6] B. Glimelius, E. Tiret, A. Cervantes, D. Arnold *et al.*, "Rectal cancer: Esmo clinical practice guidelines for diagnosis, treatment and follow-up," *Ann Oncol*, vol. 24, no. Suppl 6), 2013.

[7] E. Van Cutsem, A. Cervantes, R. Adam, A. Sobrero, J. Van Krieken, D. Aderka, E. Aranda Aguilar, A. Bardelli, A. Benson, G. Bodoky *et al.*, "Esmo consensus guidelines for the management of patients with metastatic colorectal cancer," *Annals of Oncology*, vol. 27, no. 8, pp. 1386–1422, 2016.

[8] L. Claret, F. Mercier, B. E. Houk, P. A. Milligan, and R. Bruno, "Modeling and simulations relating overall survival to tumor growth inhibition in renal cell carcinoma patients," *Cancer chemotherapy and pharmacology*, vol. 76, no. 3, pp. 567–573, 2015.

[9] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker *et al.*, "Radiomics: extracting more information from medical images using advanced feature analysis," *European journal of cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[10] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher *et al.*, "Radiomics: the process and the challenges," *Magnetic resonance imaging*, vol. 30, no. 9, pp. 1234–1248, 2012.

[11] R. T. Leijenaar, S. Carvalho, E. R. Velazquez, W. J. Van Elmpt, C. Parmar, O. S. Hoekstra, C. J. Hoekstra, R. Boellaard, A. L. Dekker, R. J. Gillies *et al.*, "Stability of fdg-pet radiomics features: an integrated analysis of test-retest and inter-observer variability," *Acta oncologica*, vol. 52, no. 7, pp. 1391–1397, 2013.

[12] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Cavalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature communications*, vol. 5, 2014.

[13] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2015.

[14] S. S. Yip and H. J. Aerts, "Applications and limitations of radiomics," *Physics in Medicine & Biology*, vol. 61, no. 13, p. R150, 2016.

[15] M. Bogowicz, O. Riesterer, L. S. Stark, G. Studer, J. Unkelbach, M. Guckenberger, and S. Tanadini-Lang, "Comparison of pet and ct radiomics for prediction of local tumor control in head and neck squamous cell carcinoma," *Acta Oncologica*, vol. 56, no. 11, pp. 1531–1536, 2017.

[16] "American Cancer Society key statistics for colorectal cancer," https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html, accessed: 2017-10-03.

[17] E. P. Misiakos, N. P. Karidis, and G. Kouraklis, "Current treatment for colorectal liver metastases," *World journal of gastroenterology: WJG*, vol. 17, no. 36, p. 4067, 2011.

[18] J. W. Holch, M. Demmer, C. Lamersdorf, M. Michl, C. Schulz, J. C. von Einem, D. P. Modest, and V. Heinemann, "Pattern and dynamics of distant metastases in metastatic colorectal cancer," *Visceral Medicine*, vol. 33, no. 1, pp. 70–75, 2017.

[19] S. Hussein, R. Gillies, K. Cao, Q. Song, and U. Bagci, "Tumornet: Lung nodule characterization using multi-view convolutional neural network with gaussian process," *arXiv preprint arXiv:1703.00645*, 2017.

[20] "American Cancer Society survival rates for colorectal cancer," https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html, accessed: 2017-11-16.

[21] E. Eisenhauer, P. Therasse, J. Bogaerts, L. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney *et al.*, "New response evaluation criteria in solid tumours: revised recist guideline (version 1.1)," *European journal of cancer*, vol. 45, no. 2, pp. 228–247, 2009.

[22] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[24] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[25] S. Hayes, M. Pietanza, D. ODriscoll, J. Zheng, C. Moskowitz, M. Kris, and M. Ginsberg, "Comparison of ct volumetric measurement with recist response in patients with lung cancer," *European journal of radiology*, vol. 85, no. 3, pp. 524–533, 2016.

[26] J. Xiao, Y. Tan, W. Li, J. Gong, Z. Zhou, Y. Huang, J. Zheng, Y. Deng, L. Wang, J. Peng *et al.*, "Tumor volume reduction rate is superior to recist for predicting the pathological response of rectal cancer treated

with neoadjuvant chemoradiation: Results from a prospective study," *Oncology letters*, vol. 9, no. 6, pp. 2680–2686, 2015.

[27] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[28] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[29] T. Dozat, "Incorporating nesterov momentum into adam," 2016.

[30] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[31] A. Katharopoulos and F. Fleuret, "Biased importance sampling for deep neural network training," *arXiv preprint arXiv:1706.00043*, 2017.