Prediction of Acute Pulmonary Toxicity Events with 3D Convolutional Neural Networks from Radiotherapy Dose Maps

Pedro Juan Soto Vega, Vincent Bourbonne, Wistan Marchadour, Gustavo Andrade-Miranda, Francois Lucia, Martin Rehn, Ulrike Schick, Dimitris Visvikis, Franck Vermet, Mathieu Hatt

Prediction of Acute Pulmonary Toxicity Events with 3D Convolutional Neural Networks from Radiotherapy Dose Maps

Pedro J Soto Vega[1], Vincent Bourbonne[1,2], Wistan Marchadour[1], Gustavo Andrade-Miranda[1], Francois Lucia[1,2], Martin Rehn[2], Ulrike Schick[1,2], Dimitris Visvikis[1], Franck Vermet[3,1], and Mathieu Hatt[1]

[1]Laboratory of medical information processing (LaTIM), INSERM UMR 1101, Univ Brest, Brest, France
[2]Radiation Oncology Department, University Hospital of Brest, Brest, France
[3]Laboratory of Mathematics of Atlantic Britanny (LMBA), CNRS UMR 6205, Univ Brest, Brest, France

Abstract Predicting toxicity events in radiation therapy (RT) is highly beneficial for managing patients effectively. Identifying patients who are at a high risk of experiencing toxicity early on during their treatments can help in taking measures to reduce the risk of adverse events produced by this undesirable effect. Recent works in a related application, namely, acute pulmonary toxicity (APT) in lung cancer patients treated by RT, have demonstrated high accuracy in predicting such an event using dose maps features processed by a multilayer perception network. Thus, motivated by the success of convolutional neural networks (CNN) in learning semantically rich representation directly from images, this work investigates the suitability of CNN architectures in predicting APT directly from dose maps. Our results demonstrate the ability of some CNN models to predict APT from planning dose maps with an accuracy of up to 81% in terms of receiver operative characteristic's area under the curve. However, most of the architectures and configurations under evaluation led to non-satisfactory accuracy, as only shallower architectures using resized dose maps as inputs were able to train models with good accuracy in the testing set.

1 Introduction

Prediction of toxicity events related to radiotherapy (RT) could be very useful for better managing patients and personalizing treatment and follow-up. Indeed, identifying patients at risk of suffering from these toxicity events early in the course of patients monitoring (ideally, before initiating RT), could allow, in the best case scenario, modifying treatment (optimization of treatment planning, de-escalation of dose to sub-volumes of the organs-at-risk (OAR) most responsible for the toxicity, etc.) to reduce the risk, or at least, identifying patients that could benefit from intensified monitoring after RT to better prevent and treat toxicity symptoms. Recently, we have shown that analysis of dose maps can provide predictive markers of toxicity events in cervical [1] and lung cancer [2]. We used radiomics engineered features (i.e., intensity or textural metrics) extracted from the delineated OAR in the dose maps to train multiparametric models that demonstrated higher predictive value than the usual dose-volume histogram (DVH) approach. In another work, we developed an alternative approach where all dose maps were co-registered to a common spatial reference, and a specific region in the lung most correlated with the acute pulmonary toxicity (APT) event was identified through statistical analysis on a voxel-by-voxel level in the dose maps. Dose map features (e.g., mean dose) from this specific area, combined with clinical variables through a multilayer perceptron (MLP), allowed good accuracy (AUC 81%) in predicting patients with APT[3, 4]. The main advantage of this approach was the ability to both predict which patients are most likely to suffer from APT and to identify a spatial area in the lung that was most responsible. In the present work, we investigated the feasibility of achieving similar predictive power by relying on deep learning (DL) convolutional neural network (CNN) architectures trained using the dose maps as inputs, with or without the help of clinical variables. Our long term goals are i) to achieve similar (if not better) predictive power compared to the previous approach, ii) to provide some explainability of the prediction made by the network and thereby identify the area in the lung most responsible for the APT, on a patient-by-patient basis. Objective ii) will be investigated using interpretability tools such as saliency maps of the trained CNN model. Preliminary results will be reported in the present paper, focusing on reporting methods and results of objective i), while reporting on first results using saliency maps for objective ii).

2 Materials and Methods

2.1 Workflow

The APT classification task was conducted following the procedure presented in Figure 1, where the prediction model takes two inputs: the RT dose maps and clinical features (e.g., age, gender, tumor location...). The dose maps, in the figure represented by X_1 are forwarded through a CNN backbone, e.g., ResNet or DenseNet, from which a feature vector of size 512 is obtained and later concatenated with the clinical features X_2 to serve as input to a set of fully connected layers which delivers the final prediction outcome.

2.2 Dataset and pre-processing

The dose maps as images were previously generated and co-registered to a common spatial reference, namely a thoracic phantom, using a segmentation-based elastic registration via MIM Maestro (MIM v7.0.0, Cleveland, OH, USA) [3]. The segmentation used for registration was the volume combining the lungs and the heart. The dataset contains dose maps of dimension $512 \times 512 \times 237$ from 207 patients, out

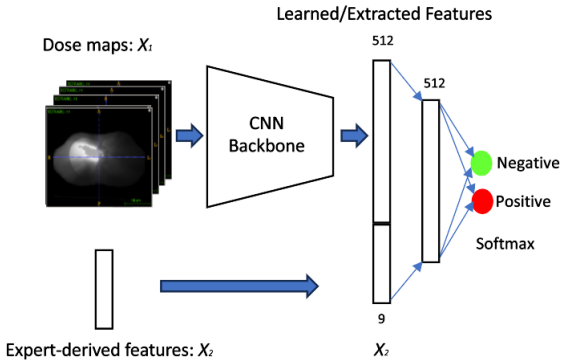

Figure 1: Proposed framework overview.

of which 45 experienced APT whereas 164 did not. Additionally, the dataset includes 37 clinical features for each patient, as described in [3]. Expert-derived knowledge considered as additional input consisted of features previously identified as predictive, i.e., Volume of the heart receiving at least 40Gy ($V40_{Heart}$), Volume of the homolateral lung receiving at least 10Gy ($V10_{LungH}$), Mean dose received by the two lungs ($DMean_{2Lungs}$), American Joint Committee of Cancer (*AJCC Stage*), Chronic Obstructive Pulmonary Disease (*COPD*), Mean Expiratory Volume/Second (*MEVS*), and smoking status [3]. The training set included the same 165 patients used in [3], while the remaining 45 patients were included in the test set. For validation purposes during training, we isolated 20% of the training samples.

2.3 CNN architectures

When using dose maps as input, 3D ResNet [5] and DenseNet-based [6] architectures were used, specifically, ResNet-10, 18, 34, 50, and DenseNet-121. We note that diverse other image classification architectures have been proposed, some with outstanding performance on different applications. To the best of the authors' knowledge, no specific architecture has been designed so far for APT classification. Therefore, we opted to evaluate seminal, general-purpose models, extensively used and well-documented, which we believe would facilitate the reproducibility of our experiments and results.

Succinctly, Residual Networks (ResNet), introduced by [5], aimed at improving convergence issues while training very deep network architectures. ResNet addressed the problem of vanishing gradients, which hinders the optimization process, and the degradation problem, i.e., adding more layers to a deep model leads to higher training error. Such problems, including residual learning blocks among the network layers, were dealt with. The degradation problem suggested that when the network's full capacity was underused for solving a particular task, the optimization process would have difficulty approximating nonlinear layers into identity mappings, which could automatically adjust network depth. Then, instead of hoping that every few stacked layers directly fit a desired underlying mapping, e.g., identity, ResNet explicitly lets those layers fit a residual mapping, which is easier to optimize. DenseNet, on the other hand, introduced by [6], similar to ResNet [5], addresses the vanishing gradient problem by adding densely connected blocks, which consists of linking each layer to any other previous layer. This allows the network to learn more efficiently by reusing features and reducing the number of trainable parameters. The principle behind this proposal states that concatenating the feature maps learned in previous layers allows each layer to access all features of all preceding levels.

2.4 Experiments implementation

In the experiments conducted in this work, we considered different input dose map sizes, see Table 1, the use (✓) or not (×) of data augmentation, and to take into account (✓) or not (×) the dataset imbalance. Each dose map and clinical features, except for AJCC Stage, was normalized with min-max normalization between 0 and 1. AJCC Stage feature was converted from categorical to binary variable. When data augmentation was considered, Gaussian noise, smooth, sharpen, and histogram shift were applied randomly. We used the MONAI library[1] to perform these transformations, as well as to set the network architectures. We used the Adam [7] optimizer and learning rate decay during training following [8]. We set the initial learning rate μ_0 and momentum β_1 equal to 0.0001 and 0.9, respectively. The batch size was set to 2, and the early stopping procedure was used to avoid over-fitting after 20 epochs without signals of accuracy improvements in the validation set.

The training/testing procedures were executed three times, each with a different (random) initialization of the trainable parameters. When class imbalance was addressed in experiments, we adopted a *weighted cross-entropy* cost function to compensate for class imbalance to train the networks, by assigning larger weights to the underrepresented ones, i.e., 0.2 for the negative and 0.8 for the positive class. Those weights were set experimentally.

2.5 Evaluation

Following [3]; the performance of models was evaluated using Receiver Operative Characteristic's (ROC) Area Under Curve (AUC). Considering each network architecture was trained and tested three times, the final performance is expressed in terms of the average AUC. It is worth noting that the standard deviation over the three runs was very low and thus not included in the table for space purposes.

3 Results

Table 1 presents the AUC (average of 3 runs) of each model in predicting APT events according to different criteria: using only dose maps, clinical variables, or both.

[1] https://monai.io/

Input	Clinical only				Dose Map only												Dose Map & Clinical											
Input Size	9				64 × 64				128 × 128				256 × 256				64 × 64				128 × 128				256 × 256			
Balance	×		✓		×		✓		×		✓		×		✓		×		✓		×		✓		×		✓	
Data Aug.	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓
MLP	43.7	56.0	46.5	60.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DenseNet121	-	-	-	-	73.7	69.2	67.4	68.5	76.9	73.2	73.9	74.7	72.0	75.3	64.7	73.7	65.2	64.8	70.9	73.0	72.5	73.3	71.7	71.8	74.2	74.8	73.5	74.4
ResNet10	-	-	-	-	**80.2**	61.2	**80.4**	71.3	67.0	78.4	79.9	76.7	25.3	41.0	73.0	74.9	**80.9**	63.3	73.6	61.2	76.0	66.8	78.5	66.9	41.5	71.4	73.0	47.5
ResNet18	-	-	-	-	77.7	60.6	79.9	63.2	74.2	75.8	76.8	72.3	37.3	37.0	61.5	75.4	76.6	65.4	71.9	58.2	77.7	67.9	76.9	76.8	31.1	61.8	72.3	68.7
ResNet34	-	-	-	-	47.1	57.1	76.6	54.2	39.1	71.7	80.0	59.1	36.7	48.3	50.5	66.1	64.9	50	67.9	65.6	79.5	72.5	77.9	53.7	38.1	43.7	50.7	53.4
ResNet50	-	-	-	-	49.8	44.5	40	59.9	25.2	59.0	39.6	22.7	47.0	23.8	52.6	50.9	58.1	42.3	59.7	53.8	22.5	59.6	70.9	59.1	37.8	43.3	41.0	44.0

Table 1: Average AUC (%) for 3 executions of models.

First of all, results obtained by using only expert-derived knowledge as input (i.e., features previously identified in [3]) to a MLP, the highest performance of 60.2 % of AUC was obtained with balancing and data augmentation. On the other hand, the highest performances were obtained when using dose maps as input in dimensions of 64 × 64 and shallower CNN architectures, i.e., ResNet10, for which the best-achieved performance was 80.2% (80.4%, without data augmentation). Addressing the imbalance using weights in the cost function delivered the best AUC (80.4%) among these two results. Still considering dose maps-based results, ResNet50 obtained the lowest performance among all the experiment's configurations, i.e., around 22.5% for the lower result and 70.9% for the highest, which corresponded to the dimension 128 × 128 input dimension and without data augmentation. In the remaining experimental configurations, ResNet50 achieved results between 40% and 60% AUC.

Considering the results where dose maps and clinical variables were jointly used as inputs, the ResNet10 architecture outperformed the remaining evaluated architectures with 80.9% AUC. Such a result was reached without data augmentation and weighted cross-entropy to alleviate the imbalance. Although still low, It is worth highlighting the significant increase obtained by ResNet50's performance when clinical variables and dose maps were added.

The highest results, e.g., 80.2%, 80.4%, and 80.9%, were obtained with dimension 64 × 64, no data augmentation, and with the shallowest architecture, namely ResNet10. On the other hand, regardless of the use of data augmentation and *weighted cross-entropy*, consistently lowest performances were achieved with the larger dose maps size (256 × 256) and the deepest CNN architecture, ResNet50, with AUCs from 23.8% to 52.9% with only dose maps, and from 37.8% to 44% when both types of data were forwarded through the proposed prediction model. The results obtained by the DenseNet121 architecture were between those achieved by ResNet10 and ResNet50, i.e., in the range 64.7% to 76.9%.

4 Discussion

Our results suggest it is feasible to achieve satisfactory prediction of APT events in lung cancer patients following RT by training a CNN with co-registered dose maps as input. Interestingly, the overall accuracy reached by our best models (around 80% AUC) is similar to that reached through the previous approach [3], which relied on a combination of clinical and dosimetry features extracted from the lung area identified as most correlated with APT events. We should, however, emphasize that this level of accuracy could be reached without any clinical variables as inputs, as the results were not significantly improved using both dose maps and clinical variables, compared to the use of dose maps alone, whereas the use of clinical variables alone led to poor predictive power (60% AUC or less). Another important observation is that the choice of the CNN architecture, the use of balancing and data augmentation, as well as the choice of the input dose maps resize had a strong impact on the final performance. The best results were obtained with shallower architectures (ResNet 10), without data augmentation or balancing, and using smaller dose maps (64 x 64). We also tested using the original size of dose maps as input (512 x 512), but the results were even poorer, so we did not include them.

Considering the results obtained by larger CNN architectures, e.g., ResNet50, it is well known that the number of training labeled samples required for efficient learning increases with the depth of the network. Since ResNet50 contains significantly more trainable parameters than the other variants of ResNet, it is not surprising that it performs worse given the amount of available data. By contrast, the few configurations that led to the best performance in the testing (around 80% AUC) corresponded to the shallowest architectures. Additionally, with respect to this last point, the performance (around 70% AUC) of DenseNet deserves to be highlighted, as even though DenseNet encompasses more layers than ResNet50, its performance was consistently superior, which may be due to the dense blocks to alleviate the vanishing gradient problem.

Although we focused here on our first objective, we include in Figure 2 a preliminary result of using saliency maps generated with Grad-CAM [9] to illustrate how these could be further exploited in our future work to provide some interpretability of the CNN prediction output, as well as to confirm the location of the area most responsible for the APT event occurrence. The figure shows an overlaid representation of the phantom used for the co-registration of dose maps and the average of saliency maps computed by Grad-CAM. Note that the average saliency maps were computed only over the samples correctly classified by the CNN model, separately

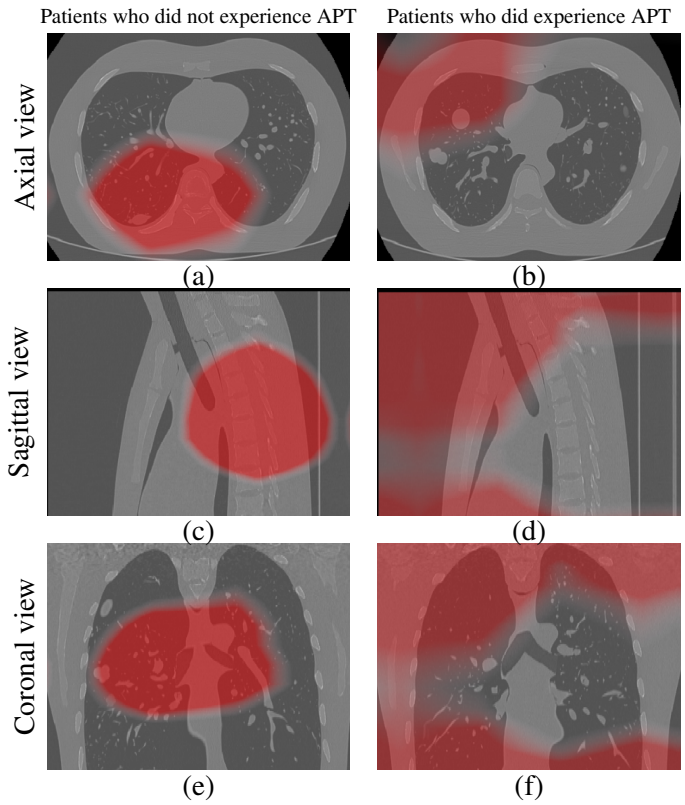

Figure 2: Saliency maps representations computed over ResNet10 architecture using Grad-CAM algorithm. The average of all saliency maps is overlaid on the phantom used for co-registration to visualize the localization of the parts of the dose maps relied upon by the CNN in axial (a/b), sagittal (c/d), and coronal (e/f) views for classifying patients without/with APT respectively.

for cases with and without APT. According to these saliency maps, prediction of the absence of APT events is achieved by the CNN by relying mostly on the dose information contained in lower lung areas, whereas prediction of the occurrence of APT events is achieved by relying on features located in higher lung areas but also strangely highlights top and bottom (coronal/sagittal view) of dose maps where the dose is near-zero, which is a bit puzzling. More work is necessary to understand better the explainability value of these saliency maps (including the comparison with other methods such as integrated gradient with or without smoothgrad) [10].

5 Conclusion

This study evaluated five 3D CNNs for predicting toxicity events based on 3D RT planning dose maps and associated clinical features. We also reported on the potential of Grad-CAM saliency maps to provide some interpretability of the toxicity predictions. Our results suggest that CNN can perform satisfactorily in this classification task. However, only a small part of the configurations of CNN architectures we tested led to good predictive performance. There is clearly room for improvement regarding the generalization capacity of these models. This could be achieved by better preprocessing the training data, using additional data augmentation techniques, or by simplifying the models regarding the number of parameters to avoid overfitting.

6 Acknowledgments

This study was funded by the CHIST-ERA grant [CHIST-ERA-19-XAI-007] with project acronym INFORM, by the General Secretariat for Research and Innovation (GSRI) of Greece, National Science Centre (NCN) of Poland [2020/02/Y/ST6/00071] and Agence Nationale de la Recherche (ANR) of France.

References

[1] F. Lucia, V. Bourbonne, D. Visvikis, et al. “Radiomics analysis of 3D dose distributions to predict toxicity of radiotherapy for cervical cancer”. *Journal of Personalized Medicine* 11.5 (2021), p. 398.

[2] V. Bourbonne, R. Da-Ano, V. Jaouen, et al. “Radiomics analysis of 3D dose distributions to predict toxicity of radiotherapy for lung cancer”. *Radiotherapy and Oncology* 155 (2021), pp. 144–150.

[3] V. Bourbonne, F. Lucia, V. Jaouen, et al. “Development and prospective validation of a spatial dose pattern based model predicting acute pulmonary toxicity in patients treated with volumetric arc-therapy for locally advanced lung cancer”. *Radiotherapy and Oncology* 164 (2021), pp. 43–49.

[4] V. Bourbonne, F. Lucia, V. Jaouen, et al. “Combination of Radiomics Features and Functional Radiosensitivity Enhances Prediction of Acute Pulmonary Toxicity in a Prospective Validation Cohort of Patients with a Locally Advanced Lung Cancer Treated with VMAT-Radiotherapy”. *Journal of Personalized Medicine* 12.11 (2022), p. 1926.

[5] K. He, X. Zhang, S. Ren, et al. “Deep residual learning for image recognition”. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-Decem (2016), pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[6] G. Huang, Z. Liu, L. Van Der Maaten, et al. “Densely connected convolutional networks”. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[7] D. P. Kingma and J. Ba. “Adam: A method for stochastic optimization”. *arXiv preprint arXiv:1412.6980* (2014).

[8] Y. Ganin, E. Ustinova, H. Ajakan, et al. “Domain-adversarial training of neural networks”. *Journal of machine learning research* 17.59 (2016), pp. 1–35.

[9] R. R. Selvaraju, M. Cogswell, A. Das, et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[10] L. Brocki, W. Marchadour, J. Maison, et al. “Evaluation of importance estimators in deep learning classifiers for Computed Tomography”. *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. Springer. 2022, pp. 3–18.