2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)
April 13-16, 2021, Nice, France

INTERPRETABLE DEEP MODEL FOR PREDICTING GENE-ADDICTED NON-SMALL-CELL LUNG CANCER IN CT SCANS

C. Pino[⋆] *S. Palazzo*[⋆] *F. Trenta*[†] *F. Cordero*[⋆†] *U. Bagci*[⋆⋆] *F. Rundo*[††]
S. Battiato[†] *D. Giordano*[⋆] *M. Aldinucci*[⋆†] *C. Spampinato*[⋆]

[⋆] PeRCeiVe Lab, Department of Electrical, Electronic and Computer Engineering, University of Catania, Italy
[†] iCTLab, University of Catania, Italy
[⋆†] Department of Computer Sciences, University of Torino, Italy
[⋆⋆] Center for Research in Computer Vision, University of Central Florida, Orlando, FL, USA
[††] STMicrolectronics, ADG Central R&D,Catania

2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-1246-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISBI48211.2021.9433832

ABSTRACT

Genetic profiling and characterization of lung cancers have recently emerged as a new technique for targeted therapeutic treatment based on immunotherapy or molecular drugs. However, the most effective way to discover specific gene mutations through tissue biopsy has several limitations, from invasiveness to being a risky procedure. Recently, quantitative assessment of visual features from CT data has been demonstrated to be a valid alternative to biopsy for the diagnosis of gene-addicted tumors. In this paper, we present a deep model for automated lesion segmentation and classification as gene-addicted or not. The segmentation approach extends the 2D Tiramisu architecture for 3D segmentation through dense blocks and squeeze-and-excitation layers, while a multi-scale 3D CNN is used for lesion classification. We also train our model with adversarial samples, and show that this approach acts as a gradient regularizer and enhances model interpretability. We also built a dataset, the first of its nature, consisting of 73 CT scans annotated with the presence of a specific genomics profile. We test our approach on this dataset achieving a segmentation accuracy of 93.11% (Dice score) and a classification accuracy in identifying oncogene-addicted lung tumors of 82.00%.

Index Terms— XAI, 3D Tiramisu, Lesion Classification

1. INTRODUCTION

In the last decade, evidence of the role of some genes (such as Epidermal Growth Factor Receptor, EGFR) as a therapeutic target in advanced stages of lung cancers has emerged. In these cases, immunotherapy [1] and molecular target drugs [2] appear to be more effective than standard chemotherapy alone, showing that tumor genetic characterization and tailored therapeutic treatment may increase survival chances. The current gold standard technique for identifying such mutation profiles – deep sequencing – is performed on tissue biopsy, with the related major drawbacks, from the complexity of detecting tumors making the outcome operator-dependent, to invasiveness and tumor dissemination [3]. Computed tomography (CT) is routinely used for treating lung cancers and it is also gaining an important role for mutation identification: correlation between low-level textures and mutational status in non-small-cell lung cancer (NSCLC) has been observed [4]. These findings opened a non-invasive way in tumor genetic profiling, with an expected positive impact on diagnosis and treatment outcomes.

Deep learning has significantly surpassed traditional visual feature extraction in a variety of tasks, including medical image analysis, by learning task-specific features directly from data. Under these premises, we propose a deep learning model for automated lesion segmentation and mutation identification in CT scans of patients affected by NSCLC. We advocate a segmentation model, designed as a 3D extension of the Tiramisu network [5], to extract cancer lesions, which are further processed by a deep 3D classifier that predicts whether an extracted lesion is gene-addicted or not. Performance analysis carried out on a dataset of 73 CT scans, which so far is the only dataset of this nature, demonstrates the accuracy of our approach, outperforming several baselines.

We also target the problem of *interpretability*, necessary to explain model decisions to physicians, by training the classification network on adversarial attacks. Indeed, adversarial robustness entails feature interpretability, according to *explainable AI* theory [6] and as confirmed by our results. Our approach is thus able to genomically characterize lung cancers, while providing interpretable decisions.

2. RELATED WORK

This paper tackles the problem of predicting gene mutation from the visual analysis of CT scans. The topic is the mainstream of recent research in radiology — *radiogenomics* — that correlates quantitative analysis of imaging data (mainly

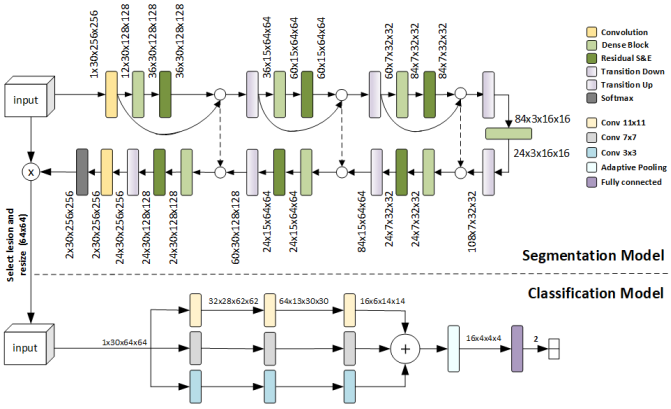

Fig. 1. **Lesion segmentation and classification networks**. Output layers for deep supervision are not shown for readability. Zoom-in for details about feature maps.

CT), through the extraction of morphological, shape and texture features [7], to tumor genomics information for a less invasive and safer diagnosis than through tissue biopsy. In particular, genetic characterization of NSCLC and therapy progression estimated through feature extraction from imaging data have been explored with promising results [4, 8]. At the same time, deep learning has been applied recently to tumor analysis from CT imaging data, for evaluating tumor progression in response to treatment [9, 10], but, to our knowledge, no method exists for predicting gene mutation. In [9], a 3D segmentation model is trained to estimate the volume of lung cancer lesions to quantify immunotherapy effectiveness by evaluating tumor progression; in [10], a 5-layer CNN with locally-connected blocks is used to distinguish between bladder lesions in different stages with and without complete chemotherapy responses. Thus, recent radiology research has demonstrated that CT analysis is an effective strategy for assessing mutational status of cancers. Following this trend, we here face the problem of visual feature learning for the identification of the mutational status in lung cancer.

3. METHOD

Our deep model (shown in Fig. 1) consists of: a *segmentation network* to extract lesions from a CT scan, followed by a *lesion classifier* to discriminate gene-addicted lesions from non-gene-addicted ones. The segmentation model is based on the *Tiramisu* network [5], i.e., a 2D fully-convolutional DenseNet [11] based on the U-Net architecture [12], consisting of a downsampling path for feature extraction and an upsampling path for output generation, with skip connections that help to preserve high-resolution details by reusing feature maps between the two paths. The main differences between our segmentation model (shown on top in Fig. 1) and the standard Tiramisu network are: 1) We employ 3D (rather than 2D) convolutional layers to process the whole CT sequence; 2) Residual squeeze-and-excitation layers [13] are used to emphasize relevant features and improve the representational power of the model; 3) We apply deep supervision [14] to the upsampling layers to deal with the vanishing gradient problem.

The input to the model consists of a subset of resized slices (for memory issues) from the raw CT scan, i.e., $256\times256\times30$, with the last dimension equal to the maximum depth among annotated lesions. This 3D input sample is initially fed to a standard 3D convolutional layer to expand the feature dimensions. The resulting feature maps then traverse the downsampling path of the model, going through a sequence of dense blocks, residual squeeze-and-excitation layers, and transition-down layers (which internally employ 3D max-pooling). In the downsampling path, the feature maps at the output of each residual squeeze-and-excitation layer are concatenated with the input features of the preceding dense block to encourage feature reuse. At the end of the downsampling path, the *bottleneck* of the model consists of a dense block composed by three dense layers. The upsampling path is symmetric to the downsampling one, with two main differences: 1) skip connections from the downsampling path concatenate feature maps at the corresponding layers of the upsampling path; 2) transition-up layers are implemented by transposed convolutions; the output of these layers are used for deep supervision. Finally, one convolutional layer followed by a softmax layer is applied to the output of each decoder layer in order to obtain a 2-channel segmentation map (lesion/non-lesion voxels), on which loss is computed for deep supervision. The model output is the binary segmentation produced by comparing lesion/non-lesion voxel likelihoods of the last decoder layer (as shown in Fig. 1). Our lesion classifier, whose architecture is shown in Fig. 1 (bottom), is a multi-scale 3D CNN that receives a crop of the CT scan corresponding to a lesion identified by the upstream segmentation model, and predicts whether the tumor is gene-addicted or not. Each branch of the model consists of a cascade of 3D depthwise separable convolutional layers with ReLU non-linearities, sharing the number of feature maps but differing by kernel size (3, 7 and 11). Padding is added to ensure that feature map sizes are consistent across the three branches. After multi-scale analysis, the output feature maps are summed and rescaled to a volume of size $4\times4\times4$ through adaptive max pooling. A final fully-connected layer maps the pooled features to a vector of size 2, with gene-addicted and non-gene-addicted class scores.

We improve the base classification model by enforcing *interpretability* through robustness to adversarial attacks [6]. To accomplish this, we employ a white-box attack mechanism – Projected Gradient Descent (PGD) [15] – based on adversarial training [16] that aims at generating adversarial examples, through input perturbations, to make the model misclassify lesions during training. Formally, let x_t be the perturbed input data after t iterations of the algorithm, starting from the

original input $x = x_0$. Each PGD iteration adds a perturbation to the data at the previous iteration and projects the new data to a point within a L_2 hypersphere with radius ε around x_0, to ensure similarity to the original input. The iterative perturbation process can be described as:

$$x_{t+1} = \Pi_{x_0,\varepsilon}(x_t + \varepsilon \, \text{sign}(\mathcal{L}_{\text{XE}}(x_t, y)) \quad (1)$$

where t is the iteration number, y the label of the input sample, and $\mathcal{L}_{\text{XE}}(x_t, y)$ the binary cross-entropy loss of the lesion classifier. $\Pi_{x_0,\varepsilon}(\hat{x})$ is defined as:

$$\Pi_{x_0,\varepsilon}(\hat{x}) = \begin{cases} x_0 + \frac{\hat{x}-x_0}{||\hat{x}-x_0||_2}\varepsilon & \text{if } ||\hat{x} - x_0||_2 > \varepsilon \\ \hat{x} & \text{otherwise} \end{cases} \quad (2)$$

During training, we feed the PGD-generated samples to the model with the original y label. This procedure forces the model to learn features that are less susceptible to input perturbations, resulting in increased robustness and interpretability.

4. EXPERIMENTAL RESULTS

Dataset and annotations. We built a CT dataset with 73 CT scans of lung cancer patients, diagnosed by biopsy. For each biopsy, the presence of a specific genomics profile is noted, with 21 out of 73 lesions defined as oncogene-addicted, i.e., characterized by a single dominant driver gene (either EGFR or ALK or ROS-1 or BRAF). The remaining 52 cases are not oncogene-addicted. An expert oncologist annotated each CT scan by drawing the contours of the biopsied lesion and selecting a subset of slices containing it. All CT scans were acquired with the following scanning parameters: 120 kV tube voltage, reference current of 50 mA, a spatial resolution of 512×512 pixels, and 100 slices 3 mm thick each.

Training procedure. We train the segmentation and classification models separately with 5-fold cross-validation: each test fold contains 5 randomly-selected non-gene-addicted CTs and 5 gene-addicted scans. During training, the segmentation model receives a tensor of size 256×256×30, obtained by resizing the height (sagittal axis) and width (frontal axis) of the original CT scan and using the 30 slices (along the longitudinal axis) around the annotated lesion as provided by the radiologists. We compute the Dice loss on the output of the segmentation network as well as on the intermediate decoder layers for deep supervision. Mini-batch gradient descent (batch 2) is performed for minimizing segmentation loss, using Adam (μ = 0.001, β_1 = 0.5, β_2 = 0.999).

The classification model is trained on the individual lesions identified by the segmentation model, padded by 5 voxels (to account for under-segmentation), and resized to 64×64×30. Although multiple lesions may be segmented, the biopsied lesion is generally the most extended in the considered slices: for this reason, we select the largest mask from the segmentation output and provide it as input to the classification model. As loss function, we employ the standard binary cross-entropy loss, weighed by 1.5 for gene-addicted lesions and 0.5 for non-gene-addicted ones to account for class imbalance in the training set. Perturbed samples for adversarial attacks are generated by setting ε = 0.1 in the PGD algorithm.

Model	DICE Score (%)
3D-UNet [17]	86.57 ± 7.36
Base model	90.47 ± 6.21
Base model + SE	92.46 ± 5.83
Base model + SE + DS	93.11 ± 4.23

Table 1. **Lesion segmentation results**

Lesion segmentation results. Segmentation accuracy is reported in terms of the Dice coefficient. Since each CT scan may have multiple lesions, but only the largest lesion (the one on which biopsy is carried out) is annotated by radiologists, we assess the Dice coefficient for that lesion only. We compare the performance of our approach to a state-of-the-art 3D-UNet [17]. We also perform ablation studies by evaluating the performance of our "base model" – without squeeze-and-excitation layers (SE) and deep supervision (DS) – and by adding them one at a time. Results, given in Tab. 1, show that our approach outperforms the baselines as well as a standard state-of-the-art method on our dataset.

Lesion classification results. We then evaluate the accuracy of the proposed lesion classifier. Due to the lack of established literature approaches tackling this problem, we compare our method to the following baselines:

— **Classification without segmentation**, i.e., our multi-scale lesion classification model, operating on the entire CT scans.

— **Bottleneck features from the segmentation model**, i.e., forwarding the bottleneck features (size 24×3×16×16 = 18,432) of the segmentation model to three fully-connected layers of size 1024, 1024 and 2 for lesion classification.

To measure the robustness and interpretability of each approach, we also evaluate the average Frobenius norm of the Jacobian matrix, $||\mathcal{J}(x)||_F$, which estimates how the model is affected by input perturbations. Results are reported in Tab. 2 and show that our approach outperforms both baselines in terms of accuracy and robustness. We can notice that using the segmented lesion leads (last two lines in Tab. 2) to better robustness and interpretability (as demonstrated by the achieved Jacobian norm values). This can be explained by the fact that using lesion-masked information during classification prevents the representations learned by the classifier from being inconsistent with lesions, which is one of the main reasons for scarce generalization and overfitting in these applications [18]. Additionally, adversarially training our classifier yields even lower Jacobian norm values, at the expense of classification accuracy (82% vs 78%). The accuracy decrease is expected, since adversarial training forces the model to avoid using non-generalizable features, but en-

Model	Accuracy (%) ↑	$\|\|\mathcal{J}(\mathbf{x})\|\|_{\mathbf{F}}$ ↓
No segm	58.00 ± 4.47	5.60 ± 4.27
Bottleneck	66.00 ± 5.47	11.06 ± 3.75
Ours	82.00 ± 8.36	0.58 ± 0.37
Ours + adv. training	78.00 ± 4.47	0.35 ± 0.36

Table 2. **Classification results**. Mean classification accuracy (in percentage) and Frobenius norm of the Jacobian matrix.

hances generalization capabilities, thus it has to be preferred.

5. CONCLUSION

Visual analysis of CT scans for characterizing NSCLC cancer is emerging as a valid alternative to biopsy calling for automated analysis methods. We here propose a deep learning approach to identify oncogene-addicted tumor lesions from lung CT scans. Our model consists of a segmentation module for lesion localization and a classification module for gene-addiction prediction, trained with adversarial examples in order to increase feature robustness and interpretability. Experimental results show that our approach achieves promising classification performance (about 80%), providing at the same time detailed interpretation maps of model decisions.

6. COMPLIANCE WITH ETHICAL STANDARDS

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of University Turin, Italy (04/17/2019, No. 32).

7. CONFLICTS OF INTEREST

The authors declare that they have no conflict of interest.

8. ACKNOWLEDGEMENTS

This work has been partially funded by: a) the H2020 *DeepHealth* project: Deep-Learning and HPC to Boost Biomedical Applications for Health (G.A. 825111) and b) the University of Catania under the "Piano per la Ricerca 2016/2018".

9. REFERENCES

[1] Choi Young Rak et al., "The diagnostic efficacy and safety of endobronchial ultrasound-guided transbronchial needle aspiration as an initial diagnostic tool," *Korean J Intern Med*, vol. 28, no. 6, pp. 660–667, 2013.

[2] W. Lian and Y. Ouyang, "CT-guided aspiration lung biopsy for EGFR and ALK gene mutation analysis of lung cancer," *Oncol Lett*, vol. 13, pp. 3415–3422, 2017.

[3] M. Yasukawa et al., "Clinical Implications of Transbronchial Biopsy for Surgically-resected Non-small Cell Lung Cancer," *In Vivo*, vol. 32, pp. 691–698, 2018.

[4] Subba Digumarthy et al., "Can CT radiomic analysis in NSCLC predict histology and EGFR mutation status?," *Medicine*, vol. 98, 2019.

[5] Simon Jégou et al., "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *CVPRW 2017*.

[6] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry, "Robustness may be at odds with accuracy," in *ICLR 2019*.

[7] Lei Yang et al., "Can CT-based radiomics signature predict KRAS/NRAS/BRAF mutations in colorectal cancer?," *European radiology*, vol. 28, pp. 2058–2067.

[8] Muhammad Shafiq-ul Hassan et al., "Voxel size and gray level normalization of ct radiomic features in lung cancer," *Scientific reports*, vol. 8, no. 1, pp. 1–9, 2018.

[9] Yi Yang et al., "Development and validation of a deep learning model to assess tumor progression to immunotherapy.," American Soc. of Oncology - 2019.

[10] Cha Kenny et al., "Bladder cancer treatment response assessment in CT using radiomics with deep-learning," *Scientific reports*, vol. 7, no. 1, pp. 1–12, 2017.

[11] Gao Huang et al., "Densely connected convolutional networks.," in *CVPR*, 2017, vol. 1, p. 3.

[12] Olaf Ronneberger et al., "U-NET: Convolutional networks for biomedical image segmentation," in *MICCAI 2015*.

[13] Jie Hu et al., "Squeeze-and-excitation networks," *arXiv:1709.01507*, vol. 7, 2017.

[14] Qi Dou et al., "3D Deeply Supervised Network for Automatic Liver Segmentation from CT Volumes," in *MICCAI 2016*.

[15] Aleksander Madry et al., "Towards deep learning models resistant to adversarial attacks," *arXiv:1706.06083*, 2017.

[16] Christian Szegedy et al., "Intriguing properties of neural networks," *arXiv:1312.6199*, 2013.

[17] Özgün Çiçek et al., "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *MICCAI 2016*.

[18] Becks Simpson et al., "Gradmask: Reduce overfitting by regularizing saliency," in *MIDL*, 08–10 Jul 2019.