Advanced Densely Connected System with Embedded Spatio-Temporal Features Augmentation for Immunotherapy Assessment

Francesco Rundo
ADG Central R&D
STMicroelectronics
Catania, Italy
francesco.rundo@st.com

Giuseppe Luigi Banna
Medical Oncology Department
Queen Alexandra Hospital
Portsmouth, UK
giuseppe.banna@nhs.net

Francesca Trenta
IPLAB
University of Catania
Catania, Italy
francesca.trenta@unict.it

Sebastiano Battiato
IPLAB
University of Catania
Catania, Italy
battiato@dmi.unict.it

2021 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-3900-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IJCNN52387.2021.9534078

***Abstract*—In medical field, the term "immunotherapy" refers to a form of cancer treatment that uses the ability of body's immune system to prevent and destroy cancer cells. In the last few years, immunotherapy has demonstrated to be a very effective treatment in fighting cancer diseases. However, immunotherapy does not work for every patients and moreover, certain types of immunotherapy drugs could have side effects. With this regard, scientific researchers are investigating for effective ways to select the patients who are more likely to respond to the treatment. Hence, pre-clinical data confirmed that, sometimes, the composition of immune system cells infiltrating the tumor micro-environment may interfere with the efficacy of immunotherapy treatments. In this work, we developed a 3D Deep Network with a downstream classifier for selecting and properly augmenting features from chest-abdomen CT images toward improving cancer outcome prediction. In our work, we proposed an effective solution to a specific type of aggressive bladder cancer, called Metastatic Urothelial Carcinoma (mUC). Our experiment results achieved high accuracy confirming the effectiveness of the proposed pipeline.**

***Index Terms*—Spatio-Temporal Data, Deep Convolutional Network, Radiomics.**

I. Introduction

Immunotherapy refers to a treatment that uses the patient's own immune system to fight diseases such as cancer [1], [2]. In medical oncology, immunotherapy has been widely used in several experimental medical use-cases, in association with traditional protocols including chemotherapy and radiation therapy. This usage increase due to the very promising outcomes that the immunotherapy treatment has achieved in recent years. Malignant cells have the characteristic of reproducing very quickly at the expense of healthy cells. The currently adopted immunotherapic approaches can act in two directions: (1) By stimulating the body's immune system to make it more effective in recognizing and destroying cancer cells; (2) By introducing molecules - such as proteins - into the immune system to enhance the defense system, making it more "intelligent" in detecting and eliminating cancer cells. In the last few years, the scientific community has focused its research efforts on delivering innovative strategies by analyzing visual features from medical images. Large collections of medical images applied to train deep neural networks have gradually become a very promising topic both in the research community and in the medical industry. However, the image analysis process is often time-consuming and require expert evaluations. Even simple tasks, such as image classification or segmentation requires a lot of efforts from expert clinicians. Hence, there is a strong need to develop automatic solutions for medical image analysis, especially in the medical oncology field. These considerations underlying the effort to study and develop new efficient algorithms in order to predict cancer disease by analyzing medical images. More recently, the growing interest in deep learning technologies has led to the development of several approaches to accomplish these tasks. Hence, we propose the usage of a Deep Learning pipeline consisting of a Densely Connected Network with Non-Local Self Attention module suitable to perform advanced image classification. Recently, Deep Learning approaches have achieved outstanding performances in 2D medical image analysis task. However, researchers have recently investigated how to apply Deep Learning methods effectively in order to process 2D or 3D information from sliced images. With this regard, Densely Connected Networks have been proposed to process medical images, considering their apparent ability, compared with respecto to the other neural backbones, with volumetric data in medical imaging. In fact, the main benefit of 3D CNNs is their improved flow of information and gradients throughout the network, which makes them easy to train. Using these technologies, we developed an effective pipeline to improve the process of predicting immunotherapy treatment outcome from CT-scan, making, in this way, the healthcare system more efficient. The remainder of the paper is structured as follows. In Section II we present the related works. In Section III we describe the proposed pipeline while in Section IV we report the experiments we made for validating the proposed approach. Finally, in section V we report the conclusions.

II. Related Works

Several Deep Learning based solutions have been proposed in the scientific community, dealing with the development

of automatic pipelines for the medical treatment outcome assessment [3]–[5]. In [3], the authors evaluated the performance of the most effective Machine Learning (ML) algorithms to predict the mortality after a radical cystectomy in a large dataset of bladder cancer patients. The results have shown that the Regularized Extreme Learning Machine outperformed other methods in terms of accuracy. Different solutions for cancer image-lesion segmentation enabling 2D or 3D convolution networks have been proposed in scientific literature [5]–[9]. The performance indicators related to the segmentation stage of metastatic lesions from CT imaging are significant and promising, although it refers to lesions of the same type (visceral or lymphatic) [5]–[9]. Further interesting deep pipelines for estimating the response to such cancer treatments based on quantitative data analysis can be found in [10]–[12]. The Deep Learning architecture proposed in [10] is a modified version of the AlexNet backbone [11]. The authors applied the proposed deep architecture to learn visual features from segmented CT slices in order to assess the chemotherapic treatment. The experimental results pointed out the effectiveness of the proposed deep network on estimating treatment outcome [10]. In [13], the authors introduced a novel deep pipeline for detecting immunotherapy outcomes learning ad-hoc visual features generated by a stack of encoders. With an accuracy of 86.05%; Specificity of 89.29%, and Sensitivity of 80.00% , the proposed pipeline showed promising results. Further investigations by the same authors have been reported in further works [14] and [15]. Through the use of such innovative deep pipelines including self-attention [14] and visual data augmentation [15] the authors retrieved promising results regarding to the prediction of the immunotherapy treatment outcome as confirmed by the performance indexes that will be later compared in this paper. The pipeline herein described outperforms the previous proposed solutions confirming the progress in classification performance.

III. The proposed pipeline: Description

The proposed Deep Learning pipeline is devised to predict the outcome of the immunotherapy treatment by extracting discriminating features from CT-scan images. As shown in Fig. 1, the proposed method comprises an innovative 3D Densely Connected Convolutional Network with a self-attention mechanism and a genetic-driven spatio-temporal data augmentation layer. This study was triggered by the results achieved by our previous work [14], [15], where a 3D Non-Local Neural Network was implemented to improve bladder cancer outcome prediction. In this paper, we demonstrated the correlation between the patient's response to immunotherapy treatment and the augmented hierarchical spatio-temporal features generated by our designed deep architecture. We conducted our analysis by only considering the RECIST (ver. 1.1) CT-scan compliant lesions from CT-scan device imaging software. The Response Evaluation Criteria in Solid Tumors (RECIST) is a medical evaluating criteria defined by scientific community to assess tumor response [16]. Specifically, the RECIST 1.1 guidelines provide more stringent criteria than previous version regarding lesion measurement, the augmented definition of disease progression, the selection of bone lesions and cysts as target lesions, etc. Specialists in medical oncology have followed the aforementioned criteria in order to define an objective and robust criterion to classify the patient who presents a disease progression from the one who instead responds positively to cancer treatment. During clinical trial, the involved physicians selected the lesion, i.e. the Region of Interest (ROI) $S_R(x, y)$ with a dimension of $M \times N$ pixels. Then, the segmented CT-scan RECIST 1.1 cancer lesion $S_R(x, y)$ will be processed by the Genetic-driven Reinforcement Learning block in order to generate the spatio-temporal Volume of Interest (VOI) of $T_D \times M_D \times N_D$ to be fed to the 3D Densely Connected Network. In this work, we process a set of $32 \times 64 \times 64$ (VOI) discriminat features. In addition, an implicit "attention" module is developed in order to identify the visual feature which represent the most significant parts of the RECIST 1.1 compliant lesion. In particular, we have extended the proposed 3D convolutional architecture by designing a series of Non-Local Blocks [17]. These blocks are based on the idea to capture long-range dependencies at different scales. The proposed architecture aims to arrange the analyzed problem of immunotherapy treatment outcome estimation as a classical binary classification task. With this regard, the proposed architecture classifies the patients into two different classes: Class 1, includes patients which shows complete response (CR) or partial response (PR) or stable disease (SD)) and Class 2, including the patients who showed disease progression (PD). As introduced, the described pipeline is schematized in Fig. 1. The network architecture used in this study consists of a sequence of 3D dense blocks. The first convolution layer processes the input volume (VOI) with a size of $32 \times 64 \times 64$ pixels using a kernel size of $3 \times 3 \times 3$ pixels. The output of this layer is processed by 6 dense blocks composed by [6, 8, 8, 8, 8, 6] 3D layers respectively, that also has a kernel of $3 \times 3 \times 3$ in size. The output is followed by a ReLU non-linear activation function. Moreover, each dense block is preceded by [0, 1, 2, 3, 4, 5] Embedded Gaussian Non-Local blocks [17], respectively. Finally, a transition-down layer with a $2 \times 2 \times 2$ max pooling completes the block. In short, the input VOI is processed by the described blocks (both dense and non- local) generating the feature maps which will gradually decrease (in dimension) until they becomes a one-dimensional vector having length of 736×1. The resulting feature map traverses three fully connected (FC) layers followed by RELU. The final layer consists of a fully connected layer which outputs 2 values to a softmax layer that converts the values to a range from 0 to 1. In fact, the output of the proposed architecture can be interpreted as the likelihood of an input VOI being classified into 1 of the 2 categories. In particular, classical negative log-likelihood have been used as loss function.

In Table 1, we reported the details of the overall architecture.

A. The CT-scanner Image Processing Software Block

In this section, we go through the details of pre-processing steps for CT images. In the first stage, the radiologists per-

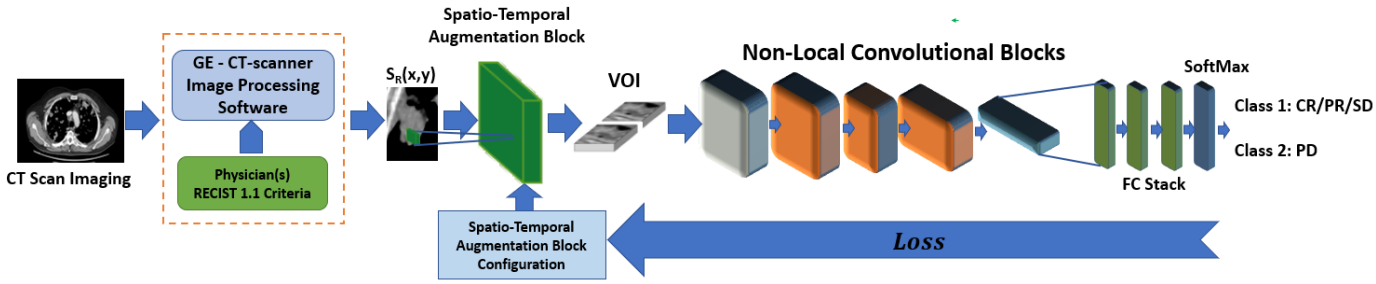

Fig. 1. The proposed Densely Connected Network with Non-Local Self Attention and Data Augmentation System

TABLE I
THE LAYERS SPECIFICATION OF THE PROPOSED DEEP ARCHITECTURE

Block	Output Size	*Layer(s) Description*	*Layers Numbers*
Convolution	$32 \times 16 \times 64 \times 64$	$3 \times 3 \times 3$ conv.	1
Dense Block	$128 \times 16 \times 64 \times 64$	Batch Normalization ReLU $3 \times 3 \times 3$ depth-wise conv. $1 \times 1 \times 1$ point-wise conv.	6
Transition Layer	$128 \times 8 \times 32 \times 32$	$1 \times 1 \times 1$ conv. $2 \times 2 \times 2$ maxpool	1
Dense Block	$256 \times 8 \times 32 \times 32$	[...]	8
Transition Layer	$256 \times 4 \times 16 \times 16$	$1 \times 1 \times 1$ conv. $2 \times 2 \times 2$ maxpool	1
Dense Block	$384 \times 4 \times 16 \times 16$	[...]	8
Transition Layer	$384 \times 2 \times 8 \times 8$	$1 \times 1 \times 1$ conv. $2 \times 2 \times 2$ maxpool	1
Dense Block	$512 \times 2 \times 8 \times 8$	[...]	8
Transition Layer	$512 \times 1 \times 4 \times 4$	$1 \times 1 \times 1$ conv. $2 \times 2 \times 2$ maxpool	1
Dense Block	$640 \times 1 \times 4 \times 4$	[...]	8
Transition Layer	$640 \times 1 \times 2 \times 2$	$1 \times 1 \times 1$ conv. $2 \times 2 \times 2$ maxpool	1
Dense Block	$736 \times 1 \times 2 \times 2$	[...]	6
Transition Layer	$736 \times 1 \times 1 \times 1$	$1 \times 1 \times 1$ conv. $2 \times 2 \times 2$ maxpool	1
Fully Connected	350	FC, ReLU	1
Fully Connected	250	FC, ReLU	1
Fully Connected	250	FC, ReLU	1
Classification	2	FC, Softmax	1

formed a pre-processing step of the CT-scan images turning off regions of non-interest in order to capture image lesions according to the required guidelines [14]–[16]. Specifically, the experts used a proper CT scanner device to create ground truths referring to the correct immunotherapy treatment outcomes. To minimize the subjective impact of the physician-driven manual choice of the visual RECIST 1.1. lesion, the pipeline validation phase is performed by using the k-fold cross-validation. Hence, we selected all RECIST 1.1 compliant lesions from CT-scans of each patient, from time to time. This procedure ensure that the evaluation of performance is carried out taking into account each possible setup of the the physician's choice of the input CT-scan lesion. The output of this block is the ROI image $S_R(x, y)$.

B. The Spatio-Temporal Data Augmentation Block

In this section, we present the proposed method for generating high-level spatio-temporal discriminant features. The generative model consists of an enhanced transient-response 2D Cellular Non-linear Networks (2D-CNN) [18]. The paradigm of this enhanced 2D-CNN is capable of performing the spatio-temporal behavior. In 2D-CNN, the cell denotes the basic unit. Any cell in a CNN is connected only to its neighbor cells. The CNN cells can interact directly with each other within a finite radius [19]. Cells not directly connected together may affect each other indirectly because of the propagation effects of the dynamics of CNNs. Specifically, in our proposed transient-response 2D-CNN every single cell of the system dynamically evolves from the initial state along the trajectory that converges -in a time-transient session- to a specific steady-state [15]–[18], creating a state-controlled template with a transient-response stage. Formally, we defined the following equations:

$$\begin{aligned} C\frac{dx_{ij}(t)}{dt} &= -\frac{1}{R_x}x_{ij} + \\ &+ \sum_{C(k,l)\in N_r(i,j)} A_1(i,j;k,l)y_{kl}(t) + \\ &+ \sum_{C(k,l)\in N_r(i,j)} A_2(i,j;k,l)u_{kl}(t) + \\ &+ \sum_{C(k,l)\in N_r(i,j)} A_3(i,j;k,l)x_{kl}(t) + \\ &+ \sum_{C(k,l)\in N_r(i,j)} D_1(i,j;k,l)(y_{ij}(t), y_{kl}(t)) + \\ &+ K_b \\ &1 \le i \le M, 1 \le j \le N \end{aligned} \tag{1}$$

$$y_{ij}(t) = \frac{1}{2}(|x_{ij}(t)+1| - |x_{ij}(t)-1|) \tag{2}$$

$$\begin{aligned} N_r(i,j) = \{C_r(k,l); (max(|k-i|, |i-j|) \le r)\} \\ (1 \le k \le M, 1 \le l \le N)) \end{aligned} \tag{3}$$

As reported, we developed a 2D-CNN which takes the segmented RECIST 1.1 lesion $S_R(x, y)$ as input u_{kl} and state x_{kl}. In Eq. (1)-(3), the $Nr(i, j)$ denotes the neighborhood of each 2D-CNN cell $C(i, j)$, taking into account a radius r. The variable $y_{ij}(t)$ represents the generated hierarchical feature. $A_1(i, j; k, l)$, $A_2(i, j; k, l)$, $A_3(i, j; k, l)$, $D_1(i, j; k, l)$ indicates the cloning templates while K_b denotes the bias. They are randomly initialized for each of the defined 32 setup

in combination with a 3×3 binary mask $A_1^{B^v}(i,j;k,l)$ and $A_2^{B^v}(i,j;k,l)$ for the $v-th$ template matrices $A_1^v(i,j;k,l)$ and $A_2^v(i,j;k,l)$ with $v = 1, 2...32$. During the training step, we retro-propagated the temporal dynamics of the overall loss $L(t)$ to this block, defining the elements of the matrices $A_1^v(i,j;k,l)$ and $A_2^v(i,j;k,l)$. The configuration of the matrices is performed by using the proposed Reinforcement Learning algorithm. More in detail, we determined the optimal policy P_o that optimize the cumulative discount reward R:

$$P_0 = argmax_{P_0} \; E[\sum_{t \geq 0} \gamma^t R(\cdot|s_t, a_t)|P_0] \tag{4}$$

Where γ denotes a proper discounted coefficient in (0,1). In order to evaluate the goodness of a state s_t (specific setup of the $v-th$ cloning templates $A_1^v(i,j;k,l)$ and $A_2^v(i,j;k,l)$) and the goodness of a coupled state-action (s_t, a_t), we defined the Value function $V^{P_0}(s_t)$ and the Q-value function $Q^{P_0}(s_t, a_t)$ respectively:

$$V^{P_0}(s_t) = E[\sum_{t \geq 0} \gamma^t R(.|s_t)|P_0] \tag{5}$$

$$Q^{P_0}(s_t, a_t) = E[\sum_{t \geq 0} \gamma^t R(.|s_t, a_t)|P_0 \;] \tag{6}$$

while the reward function is so defined:

$$\begin{aligned} R = -(\frac{\partial L(A_m^v(\cdot), D_1^v(\cdot), K_b^v(\cdot), A_p^{B^v}(\cdot), B_v, v, t)}{\partial t})^2 \\ m = 1,2,3; p = 1,2; v = 1,2,...32 \end{aligned} \tag{7}$$

where $L(\cdot)$ denotes the loss of the overall pipeline which depends on the state s_t (2D-CNN setup: $A_1(i,j;k,l)$, $A_2(i,j;k,l)$, $A_3(i,j;k,l)$, $D_1(i,j;k,l)$ and bias K_b) and the actions a_t while the policy P_0 is defined by the update of the $A_1^{B^v}(i,j;k,l)$, $A_2^{B^v}(i,j;k,l)$ and B_v masks. For each training iteration t_γ, a classical Genetic algorithm through common $crossover$ and $mutation$ operations [20] applied to the binary mask B_v , selects the $v-th$ feature setup to modify (among to the 32 defined setup) and always with a set of $crossover$ and $mutation$ operations, it will change the binary masks $A_1^{B^v}(i,j;k,l)$ and $A_2^{B^v}(i,j;k,l)$ of the selected $v-th$ setup thus identifying the coefficient of the cloning templates $A_1(i,j;k,l)$, $A_2(i,j;k,l)$ which will be updated by means of a random update (action a_t) generating a new setup of spatio-temporal (time t_γ) cloning templates $A_1^v(i,j;k,l,t_\gamma)$, $A_2^v(i,j;k,l,t_\gamma)$. The others templates and bias remain unchanged with respect to initial configuration. Only setup that produce a decrease in the overall loss dynamic will be accepted while the others will be discarded. At the end of the training phase, for each RECIST 1.1 lesions, we have obtained a $32 \times 64 \times 64$ VOI which minimize the Loss of the whole pipeline. The flowchart of the proposed Data Augmentation block is reported in the following Fig. 2.

C. Dense Blocks

The developed architecture is a 3D Densely Connected Convolutional Neural Network (3D-DCNN) embedding separable convolution layers (both depth-wise and point-wise) [14]. In our pipeline, we adopted separable convolutions in order to yield effective results with fewer computational cost. The dense block consists of a sequence of dense layers, followed by a batch normalization layer and a 3D convolutional layer with a kernel of $3 \times 3 \times 3$ in size. Finally, a ReLU activation function concludes the block. Each dense block is followed by a transition down layer, aiming to half feature map dimension, and composed by a convolutional layer with kernel size of $1 \times 1 \times 1$ followed by a max pooling layer of kernel $2 \times 2 \times 2$. Finally, the output of dense blocks is then passed to Non-Local Blocks.

D. Self-Attention through Non-Local Blocks

Non-local blocks have been recently introduced [17], as a very promising approach for capturing space-time long-range dependencies and correlation on feature maps, resulting in a sort of "self-attention" mechanism. Self-attention through non-local blocks aims to enforce the model to extract correlation among feature maps by weighting the averaged sum of the features at all possible positions in the generated feature maps [17]. In our pipeline, non-local blocks operate on almost each convolution layer to extract feature in dependencies at multiple hierarchical levels through an holistic morphological modeling of the input RECIST 1.1 CT-scan image-lesion. Formally, given a generic deep network as well as a general Non-Local Block input data x (feature map), the employed non-local operation computes the corresponding response y_i (of the given Deep block) at a i location in the input data as a weighted sum of the input data at all positions $j \neq i$:

$$y_i = \frac{1}{\psi(x)} \sum_{\forall j} \zeta(x_i, \;\; x_j) \beta(x_j) \tag{8}$$

where $\zeta(\cdot)$ denotes a pairwise potential which describes the affinity or relationship between data positions at index i and j, respectively. $\beta(\cdot)$ is a unary potential modulating ζ according to input data. The sum is then normalized by a factor $\psi(x)$. The parameters of potentials $\zeta(\cdot)$ are learned during model's training and defined as follows:

$$\zeta(x_i, \; x_j) = e^{\Theta'(x_i)^T \;\; \Phi(x_j)} \tag{9}$$

Where Θ' and Φ are two linear transformations of the input data x with learnable weights $W_{\Theta'}$ and W_Φ.

$$\begin{aligned} \Theta'(x_i) &= W_{\Theta'} x_i \\ \Phi(x_j) &= W_\Phi x_j \\ \beta(x_j) &= \; W_\beta x_j \end{aligned} \tag{10}$$

For the $\beta(\cdot)$ function, we defined a common linear embedding (classical 1x1x1 convolution) with learnable weights W_β. The normalization function ψ is:

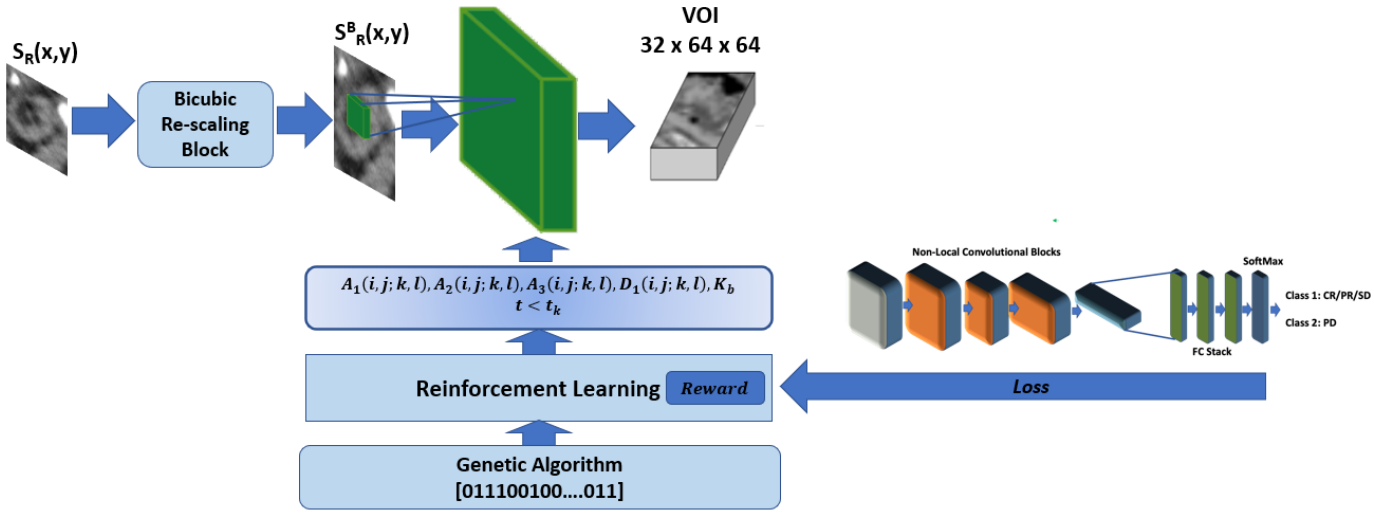

Fig. 2. The proposed Spatio-Temporal Data Augmentation Pipeline

$$\psi(x) = \sum_{\forall j} \zeta(x_i,\ x_j) \tag{11}$$

In Eqs (9)-(11) an Embedded Gaussian setup is reported [17]. The selection of the Embedded Gaussian is specifically recommended for 3D applications. The so processed features will be fed into the final block of the 3D-DCNN composed by a stack of fully connected with SoftMax for final binary classification.

E. The Classification layer

After processing the $32 \times 64 \times 64$ VOI, the network generates a 736×1 one-dimension visual embedding. The vector is fed into a stack of Fully Connected (FC) layers, comprising 3 FC layers with 350, 250, and 250, respectively. Finally, a Softmax layer is applied to compute the binary classification of the processed RECIST 1.1 compliant CT-scan lesion. The flowchart of the proposed pipeline is shown in Fig. 1.

F. Dataset: Recruitment and data pre-processing

We evaluate all the models on a dataset consisting of 106 chest-abdomen CT scans from patients affected by bladder cancer. More in detail, 43 target lesions of the 106 recruited is associated to a complete/partial response or a disease stabilization following immunotherapy treatment (CR: Complete response / PR: Partial Response / SD: Stable Disease as Class 1), while 63 lesions is associated to experienced disease progression despite anti-PD-L1 drug treatment (PD: progressive disease as Class 2). In addition, the dataset comprises subjects who are under the age of 60 (30%), male patients (91%), female patients (9%), subjects who have lymph node metastases (33%), subject who have various visceral metastatic lesions (67%). Patients enrolled in this study underwent CT examination after an histologically confirmation of bladder cancer disease. All procedures were carried out under the supervision of the clinicians and after receiving the patient's informed consent. This clinical study was performed under IRB "Catania 1 Ethical Committee" Nr. D4191C00068 and MO29983 approval. In addition, each enrolled participants were treated with a PD-L1 immunotherapy agent in the second-line setting. After the clinical trial, the experts have analyzed the target lesions from CT examinations images by using a General Electric multi-slice (64 slices) device with slice thickness of 2.5 mm; working current in the range 10 – 700 mA; working voltage: 120 kV ; pitch: 0.98. For training and validation, we used 76 target lesions subdivided into 28 belonging to Class 1 and 48 of Class 2. For the test set, we used a subset of 30 CT images (15 of Class 1 and 15 of Class 2). For each lesion, an augmented $32 \times 64 \times 64$ VOI was generated trough the Genetic-like Reinforment Learning block. To perform training, we defined a mini-batch size of 10, an initial learning rate of $3e^{-4}$, a number of epochs equal to 900 and the stochastic gradient descent with momentum (SGDM) algorithm as learning optimizer. As introduced, we performed k-fold cross-validation in order to provide a robust testing session. The evaluation was conducted on a workstation equipped with an Intel 16-Cores and NVIDIA GeForce RTX 2080 GPU.

IV. Experimental Results

We compared our proposed model with different architectures. Table II reports the testing results in terms of accuracy, sensitivity and specificity, of our proposed method in comparison with other similar approaches. As expected, referring only to classical deep architecture, Table II highlights the out-performance of 3D architectures compared to 2D ones, confirming 3D ResNet-101 as the most performing among those tested (Accuracy: $0,857 \pm 0,0487$ - Sensitivity: $0,847 \pm 0,044$ - Specificity: $0,867 \pm 0,065$). An interesting result that deserves to be highlighted is related to the use of augmentation techniques that would seem to significantly increase the performance of the classification architectures. Specifically,

TABLE II
EXPERIMENTAL PERFORMANCE BENCHMARKING (MEAN ± STANDARD DEVIATION (STD))

Model	**Metrics**					
	Accuracy		***Sensitivity***		***Specificity***	
	Mean	***STD***	***Mean***	***STD***	***Mean***	***STD***
2D DenseNet-201	$0,830$	$0,036$	$0,847$	$0,069$	$0,813$	$0,067$
2D ResNet-101	$0,827$	$0,040$	$0,817$	$0,048$	$0,837$	$0,070$
2D VGG-19	$0,775$	$0,071$	$0,827$	$0,082$	$0,723$	$0,115$
2D ResNet-18 + Aug. [15]	$0,918$	$0,0439$	$0,917$	$0,064$	$0,921$	$0,045$
3D Resnet-101	$0,857$	$0,0487$	$0,847$	$0,044$	$0,867$	$0,065$
3D DenseNet-201	$0,840$	$0,047$	$0,833$	$0,051$	$0,847$	$0,065$
3D DenseNet-NLB [14]	$0,913$	$0,035$	$0,923$	$0,062$	$0,904$	$0,033$
Proposed	**0,936**	**0,040**	**0,936**	**0,059**	**0,937**	**0,049**
Proposed w/o NLB	$0,913$	$0,033$	$0,916$	$0,047$	$0,911$	$0,048$
Proposed w/o Gen-driv RL	$0,901$	$0,038$	$0,903$	$0,050$	$0,900$	$0,550$

the data augmentation pipeline developed by the authors in [15] allows to bring the performance of the ResNet-18-based deep backbone to higher values (Accuracy: $0,918 \pm 0,043$ - Sensitivity: $0,917 \pm 0,064$ - Specificity: $0,921 \pm 0,045$) even than 3D architectures, confirming that the use of these methodologies allows to generate features more discriminating than those obtained from the simple space-temporal analysis. Similar analysis are valid for self-attention techniques which, as is evident from the results of the 3D-DenseNet with Non-Local Block reported in Table II and described in [14], allow to increase the overall performance of the deep classifiers thanks to a greater correlation between the space-time dependencies of the features. Therefore, the high performance of our proposed pipeline (Accuracy: $0,936 \pm 0,040$ - Sensitivity: $0,936 \pm 0,059$ - Specificity: $0,937 \pm 0,049$) appears in line with the scientific evidence documented so far, combining in the same model both self-attention and augmentation techniques. The significantly degraded performances obtained by the same proposed architecture without Non-Local Block (w/o NLB) or Genetic-driven Reinforcement Learning block (w/o Gen-driv RL) further confirm the claims described so far.

V. CONCLUSION

In this study, we designed a radiomics pipeline based on innovative Deep Learning backbone which shown very promsing results in predicting the outcomes of bladder cancer affected patients treated with immunotherapy. In this context, we built a DenseNet embedding a Self Attention module to process the segmented and augmented cancer image-lesions in order to correlated the cancer imaging with the effectiveness treatment. More in details, we observed that our proposed pipeline enables to non-invasively predict immunotherapy treatment response trough chest-abdomen CT-scan imaging of bladder cancer diagnosed patients, providing remarkable results in terms of accuracy, sensitivity and specificity. Compared to other models, the delivered DenseNet with Self-Attention mechanism has better performance in classification task. Moreover, the developed Spatio-Temporal Data Augmentation pipeline prevent over-fitting issues and improve the overall generalization ability of the implemented architecture. The developed pipeline was designed for the embedded STA1295 platform with OpenCV and YOCTO Linux O.S. [15]. We are working on a validation of the proposed methodology in a large, randomized and multicenter clinical trial that allows us to test our approach on a more significant medical dataset.

REFERENCES

[1] H. O. Alsaab, S. Sau, R. Alzhrani, K. Tatiparti, K. Bhise, S. K. Kashaw, and A. K. Iyer, "Pd-1 and pd-l1 checkpoint signaling inhibition for cancer immunotherapy: mechanism, combinations, and clinical outcome," *Frontiers in pharmacology*, vol. 8, p. 561, 2017.

[2] K. R. Spencer, J. Wang, A. W. Silk, S. Ganesan, H. L. Kaufman, and J. M. Mehnert, "Biomarkers for immunotherapy: current developments and challenges," *American Society of Clinical Oncology educational book*, vol. 36, pp. e493–e503, 2016.

[3] G. Wang, K.-M. Lam, Z. Deng, and K.-S. Choi, "Prediction of mortality after radical cystectomy for bladder cancer by machine learning techniques," *Computers in biology and medicine*, vol. 63, pp. 124–132, 2015.

[4] S. S. Garapati, L. Hadjiiski, K. H. Cha, H.-P. Chan, E. M. Caoili, R. H. Cohan, A. Weizer, A. Alva, C. Paramagul, J. Wei *et al.*, "Urinary bladder cancer staging in ct urography using machine learning," *Medical physics*, vol. 44, no. 11, pp. 5814–5823, 2017.

[5] Z. Hasnain, J. Mason, K. Gill, G. Miranda, I. S. Gill, P. Kuhn, and P. K. Newton, "Machine learning models for predicting post-cystectomy recurrence and survival in bladder cancer patients," *PloS one*, vol. 14, no. 2, 2019.

[6] K. H. Cha, L. Hadjiiski, R. K. Samala, H.-P. Chan, E. M. Caoili, and R. H. Cohan, "Urinary bladder segmentation in ct urography using deep-learning convolutional neural network and level sets," *Medical physics*, vol. 43, no. 4, pp. 1882–1896, 2016.

[7] M. Gordon, L. Hadjiiski, K. Cha, H.-P. Chan, R. Samala, R. H. Cohan, and E. M. Caoili, "Segmentation of inner and outer bladder wall using deep-learning convolutional neural network in ct urography," in *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134. International Society for Optics and Photonics, 2017, p. 1013402.

[8] X. Ma, L. Hadjiiski, J. Wei, H.-P. Chan, K. Cha, R. H. Cohan, E. M. Caoili, R. Samala, C. Zhou, and Y. Lu, "2d and 3d bladder segmentation using u-net-based deep-learning," in *Medical Imaging 2019: Computer-Aided Diagnosis*, vol. 10950. International Society for Optics and Photonics, 2019, p. 109500Y.

[9] E. Shkolyar, X. Jia, T. C. Chang, D. Trivedi, K. E. Mach, M. Q.-H. Meng, L. Xing, and J. C. Liao, "Augmented bladder tumor detection using deep learning," *European urology*, vol. 76, no. 6, pp. 714–718, 2019.

[10] E. Wu, L. M. Hadjiiski, R. K. Samala, H.-P. Chan, K. H. Cha, C. Richter, R. H. Cohan, E. M. Caoili, C. Paramagul, A. Alva *et al.*, "Deep learning approach for assessment of bladder cancer treatment response," *Tomography*, vol. 5, no. 1, p. 201, 2019.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[12] K. H. Cha, L. Hadjiiski, H.-P. Chan, A. Z. Weizer, A. Alva, R. H. Cohan, E. M. Caoili, C. Paramagul, and R. K. Samala, "Bladder cancer treatment response assessment in ct using radiomics with deep-learning," *Scientific reports*, vol. 7, no. 1, pp. 1–12, 2017.

[13] F. Rundo, C. Spampinato, G. L. Banna, and S. Conoci, "Advanced deep learning embedded motion radiomics pipeline for predicting anti-pd-1/pd-l1 immunotherapy response in the treatment of bladder cancer: Preliminary results," *Electronics*, vol. 8, no. 10, p. 1134, 2019.

[14] F. Rundo, G. L. Banna, L. Prezzavento, F. Trenta, S. Conoci, and S. Battiato, "3d non-local neural network: A non-invasive biomarker for immunotherapy treatment outcome prediction. case-study: Metastatic urothelial carcinoma," *Journal of Imaging*, vol. 6, no. 12, p. 133, 2020.

[15] F. Rundo, G. L. Banna, F. Trenta, C. Spampinato, L. Bidaut, X. Ye, S. Kollias, and S. Battiato, "Advanced non-linear generative model with a deep classifier for immunotherapy outcome prediction: A bladder cancer case study," in *ICPR Workshop 2020 - Artificial Intelligence for Healthcare Applications (AHIA 2020)*, 2021, accepted.

[16] E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. 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.

[17] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[18] T. Roska and L. O. Chua, "Cellular neural networks with non-linear and delay-type template elements and non-uniform grids," *International Journal of Circuit Theory and Applications*, vol. 20, no. 5, pp. 469–481, 1992.

[19] L. O. Chua and L. Yang, "Cellular neural networks: Theory," *IEEE Transactions on circuits and systems*, vol. 35, no. 10, pp. 1257–1272, 1988.

[20] A. Sehgal, H. La, S. Louis, and H. Nguyen, "Deep reinforcement learning using genetic algorithm for parameter optimization," in *2019 Third IEEE International Conference on Robotic Computing (IRC)*. IEEE, 2019, pp. 596–601.