Automation of breast cancer dosimetry with Vision Transformers

Author: Ana Matas López, amataslo29@alumnes.ub.edu
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Bruno Juliá Diaz, brunojulia@ub.edu & Pedro Gallego Franco, pgallego25@gmail.com

Abstract: A hybrid model combining a Vision Transformers encoder with convolutional layers is proposed for breast cancer dosimetry. Its predictions were compared with a baseline U-Net trained under the same conditions. The results suggest improved performance in OAR metrics, while maintaining acceptable accuracy in PTV dose prediction. However, no statistically significant differences were found, so further research is still needed to explore the full potential of Transformers in radiotherapy.

Keywords: Ionizing Radiation, Linear Accelerator, Vision Transformers, Convolutional Neural Networks, Self-Attention Mechanism

SDGs: This study is related to the sustainable-development goals (ODS) 3, 4 and 9 (see page 6)

I. INTRODUCTION

Breast cancer accounts for 28.9% of all cancers in women and remains the leading cause of cancer-related death among women in Spain, according to the Spanish Society of Medical Oncology (SEOM) [1]. Beyond its clinical and economic burden on the healthcare system, breast cancer also has a significant social impact, affecting the quality of life of patients and their families.

Once cancer is detected, patients enter a clinical workflow that may involve various treatments, including chemotherapy, surgery, and radiotherapy, with the overall goal of ensuring a good quality of life after therapy. Specifically, in radiotherapy, the aim is to eradicate cancerous cells while minimizing the dose delivered to surrounding healthy tissues, thereby reducing radiation-induced complications and toxic effects.

The two main techniques used for breast cancer treatment planning are Three-Dimensional Conformal Radiation Therapy (3D-CRT) and Intensity-Modulated Radiation Therapy (IMRT). For both, treatment planning is a time-consuming and manual process that depends heavily on the expertise of clinical users, including radiation oncologists, radiation therapists, and medical physicists. This subjectivity can lead to considerable variability in the resulting plans, potentially compromising the overall quality and consistency of the treatment. In this context, artificial intelligence (AI) has enabled significant progress in automating and accelerating the planning process by learning from large datasets and providing fast, uniform, and high-quality dose predictions [2].

Nowadays, the convolutional neural network (CNN) U-Net architecture has been widely used in medical applications, particularly in image segmentation and dose distribution prediction [3]. Its structure, which reduces and then restores the image size while keeping connections between layers, enables the preservation of local spatial features. However, this design limits the ability to capture global relationships within the image. On the other hand, the Transformer architecture, known for its self-attention mechanism, allows the model to learn both long-range and global dependencies [4]. These capabilities can be especially useful for dose prediction tasks, where non-local correlations exist between different regions of the body due to the sequential nature of computed tomography (CT) image slices [5]. Recently, Transformers have been adapted for image analysis through Vision Transformers (ViTs), which process images by dividing them into non-overlapping patches and adding positional encodings to retain spatial information [6, 7].

Although the ViT architecture was originally introduced in 2020, its application in radiotherapy is still emerging [8]. Recent studies have explored its use in predicting toxicity outcomes, early treatment response, and tumor segmentation in medical imaging. Despite these advances, many areas in radiotherapy remain unexplored with ViTs, partly due to the large dataset requirements and the high computational cost in order to provide an efficient training. To enable broader adoption of this architecture in radiotherapy, more research is essential, along with its integration into well-established neural network designs supported by the scientific community. Hybrid models can combine the strengths of both architectures and help unlock the full potential of ViTs in clinical applications.

In this study, a hybrid model based on ViTs and CNNs is proposed for dose prediction in breast cancer radiotherapy. To evaluate its performance, the results will be compared against a baseline U-Net model trained under the same conditions. Therefore, the main objective is to evaluate whether Transformer-based architectures can provide results that are clinically comparable or superior to those obtained with conventional models like U-Net.

II. METHODS

A. Radiation

A fundamental aspect of radiotherapy is achieving an accurate dose distribution that eradicates malignant cells while minimizing treatment-related side effects.

The chosen treatment modality in this study is External Beam Radiotherapy (EBRT) with 6 MV photon beams (X-ray). They have been selected due to their favorable balance between penetration depth and dose distribution. This makes them suitable for treating breast cancer, where both superficial and moderately deep malignant cell regions must be effectively irradiated.

These high-energy photons are generated by a linear accelerator (Clinac 2100), which accelerates electrons to relativistic speeds and directs them onto a tungsten target, a material with a high atomic number. Upon impact, the sudden deceleration of the electrons produces X-ray photons through the Bremsstrahlung process, which are not monoenergetic.

The selected energy level of the X-rays is widely adopted in clinical practice because it remains below the threshold for significant photoneutron production, which typically occurs at energies of 10 MV or higher. As a result, the risk of neutron-induced secondary effects is minimized. Neutron contamination arises from photonuclear reactions, specifically through the absorption of high-energy photons by atomic nuclei. Therefore, patient safety is enhanced by avoiding these interactions.

Consistent with what has been stated, a hypofractionated schedule has been adopted, delivering a total dose of 40.05 Gy in 15 fractions of 2.6 Gy each, as commonly used in breast cancer radiotherapy.

B. Neural Network

The main objective of this work is to construct a dose prediction model based on a Transformer architecture. Vision Transformers have been introduced as an alternative to traditional convolutional neural networks for image analysis tasks. ViTs take as input a sequence of image patches along with positional encoding, allowing all patches to be analyzed simultaneously. This enables the model to establish relationships between patches regardless of their spatial distance, in contrast to CNNs, which rely on local operations to extract spatial features.

The model was trained and evaluated using the high-level Keras API embedded in TensorFlow. The dataset includes 200 breast cancer patients, each with a planning computed tomography (CT) volume of 128 × 128 pixels and 2 channels, Planning Target Volume (PTV) and Organs at Risk (OARs). Training used the ADAM optimizer with an initial learning rate of 0.001 and random parameter initialization, running for 150 epochs. The loss function used was the Mean Squared Error (MSE), computed as the average of the squared differences between predicted and reference dose values at each voxel.

1. *Model Architecture*

1.1. Patch embedding block: The input CT image is divided into non-overlapping patches of 8×8 pixels. After being flattened, each patch becomes a 128-dimensional vector ($8 \times 8 \times 2 = 128$). Taking into account the size of the input image, this results in a total of 256 patches per image ($(128/8)^2 = 256$) , forming a sequence of 256 vectors that are linearly projected into a 64-dimensional embedding space. Positional encodings are added to each patch to indicate its original location within the image, allowing the model to retain spatial information before the sequence is processed by the ViT encoder.

1.2. ViT encoder: After the patch embedding, the sequence is passed through 8 ViT encoder layers. Each layer consists of a Multi-Head Self-Attention (MHSA) module with 4 attention heads, and a Multi-Layer Perceptron (MLP) module composed of two hidden layers (2048 and 1024 units), each containing a Dense layer with GELU activation. In addition, Layer Normalization (LN) and residual connections were added before and after each module.

1.3. Convolutional-based decoder: The output sequence of the encoder is reshaped into a 2D representation of shape $16 \times 16 \times 64$. This representation is then passed through a convolutional decoder composed of three transposed convolutional layers with stride 2 and ReLU activation, which progressively upsample the spatial dimensions to 32×32, then 64×64, and finally 128×128. A final convolutional layer with linear activation outputs the predicted dose distribution as a single-channel image of shape $128 \times 128 \times 1$, matching the original image resolution.

2. *Model Evaluation*

To evaluate the performance of the model, the results were compared with those obtained from a baseline model provided by the *Hospital de Sant Pau.* This model is based on a U-Net architecture, and its training setup was carried out under the same conditions as the one developed in this study, employing the ADAM optimizer with the same learning rate and loss function. The architectural diagrams can be found in Appendix.

The predicted dose distributions maps were compared against the reference dose distributions for both models.

Dice Similarity Coefficients (DSCs), which measure the overlap between predicted and clinical isodose regions, were computed to assess the spatial concordance. Additionally, Dose-Volume Histograms (DVHs), graphical representations that show the proportion of a volume receiving at least a given dose, were generated to evaluate dose conformity within the PTV and OARs. For the comparison of dose values, the following metric was used:

$$\frac{M_{\text{Clinical}} - M_{\text{Predicted}}}{M_{\text{Prescription}}} \times 100\ (\%)$$

A total of 10 patients from the test set, previously treated in the hospital, were selected to evaluate both models. These patients were not included in the training or validation sets, ensuring that the models had no prior exposure to them. This separation is crucial to objectively evaluate the generalization performance of the models. It is important to emphasize that the main objective is to quantify each model's deviation from the clinical reference. To assess whether the prediction errors differ significantly, the Wilcoxon signed-rank test was applied. A statistically significant result indicates that one of the models consistently provides more accurate predictions. Since multiple comparisons were performed (one per metric), the Bonferroni correction was applied to control the risk of false positives.

III. RESULTS

Table I presents the differences between the predicted and clinical dose values for the PTV and OARs, including standard deviations. For the OARs, the hybrid ViT & Conv model shows a consistently lower discrepancy across all evaluated metrics compared to the U-Net, especially for the lung V20Gy, where the difference is reduced by approximately 3%. On the other hand, no improvement is observed for the PTV, the U-Net yields a slightly better result than the hybrid model. Among all evaluated metrics, no difference was found to be statistically significant.

TABLE I: Average differences in PTV and OARs for some metrics between the clinical and predicted plans, including the standard deviation, for each model.

		Differences (%)	
ROI	**Metric**	**ViT & Conv**	**U-Net**
Breast PTV	D95%	3.22 ± 0.03	1.44 ± 0.03
	D_{mean}	1.43 ± 0.01	0.68 ± 0.01
Heart	V25Gy	2.13 ± 0.19	2.50 ± 0.20
	D_{mean}	0.62 ± 0.06	1.62 ± 0.06
Ipsi Lung	V20Gy	10.90 ± 0.19	14.05 ± 0.20
	D_{mean}	2.35 ± 0.04	2.43 ± 0.04

The DSC values across isodose volumes between clinical and predicted dose distributions are shown in Figure 1 for both models. As observed, values above 0.80 are achieved across most of the studied dose range, with particularly high agreement in the intermediate dose regions, which are of greatest clinical significance. A sharp decline in DSC is observed beyond approximately 95% of the prescribed dose. However, this drop is not relevant to the study as the clinical criteria ensure that 95% of the volume must be covered by at least 95% of the dose. Beyond this threshold, the dose distribution becomes much more variable, and no strict correspondence with the clinical plan is expected. When comparing both models, the U-Net exhibits a slight peak in the high-dose region, whereas the hybrid model shows a smoother overall profile and reduced discrepancies in the low-dose range.

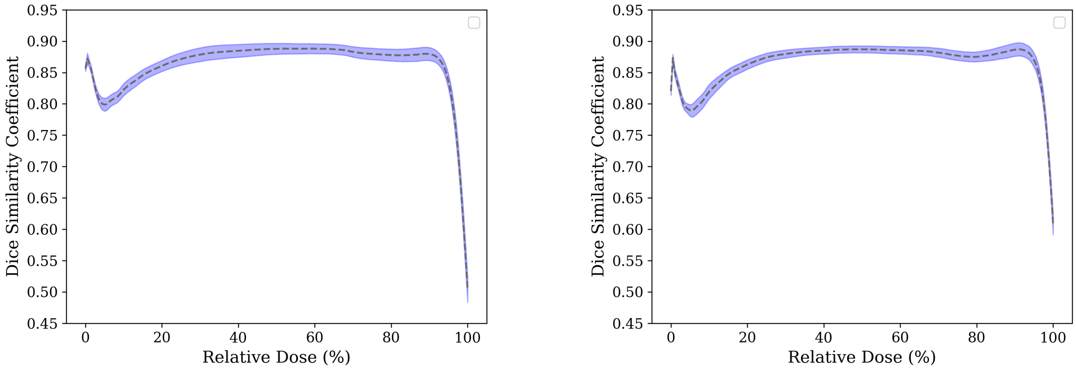
Another comparison between the clinical and predicted metrics is shown in a box plot, Figure 2. Regarding the PTV, both models yield predictions very close to 0%, with a slight tendency toward underestimation. In contrast, when analyzing the OARs, differences become more pronounced; the U-Net model exhibits greater overall variability, with the highest dispersion observed in the ipsilateral lung V20 metric. By contrast, the hybrid model demonstrates more consistent performance across most structures, although a few outliers are present, most notably in the heart.

Figure 3 presents clinical and predicted DVHs for a selected patient from the dataset. Overall, the predicted curves closely match the clinical references across all regions of interest. The hybrid model displays more similarities in both OARs, while the U-Net shows slightly better agreement with the clinical PTV curve. The corresponding dose distributions are shown in Figure 4. Visually, the predicted dose maps from both models are consistent with the clinical distribution. OAR masks are also included to illustrate the radiation distribution they receive, complementing the information provided by the DVHs.

IV. DISCUSSION

The ViT & Conv model demonstrate promising performance in predicting dose distributions. As shown in Table I and Figure 1, the hybrid model achieves lower relative errors in most metrics, particularly those related to OARs, while exhibiting slightly reduced conformity in the PTV. Nevertheless, since the errors remain below 5% for both D95% and D_{mean}, these deviations are considered acceptable in terms of prediction accuracy. The DVH curves and dose maps further confirm the consistency.

In this regard, the proposed model achieves results



(a) ViT & Convolutionals

(b) U-Net

FIG. 1: Analysis of DSC comparing isodose volumes between clinical and predicted dose distributions, including one standard deviation error.

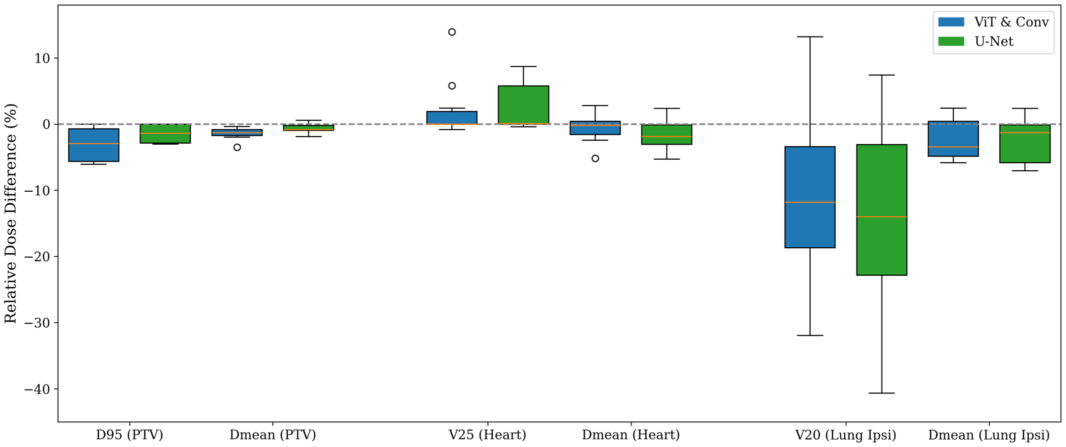


FIG. 2: Box plot showing the difference between predicted and clinical dosimetric metrics. Clinical indicators include PTV (D95, D105, Dmean), heart (Dmean, V25), and ipsilateral lung (Dmean, V20). The boxes indicate median and interquartile range (IQR). Whispers extend to 1.5 times the IQR and outliers are represented as points.

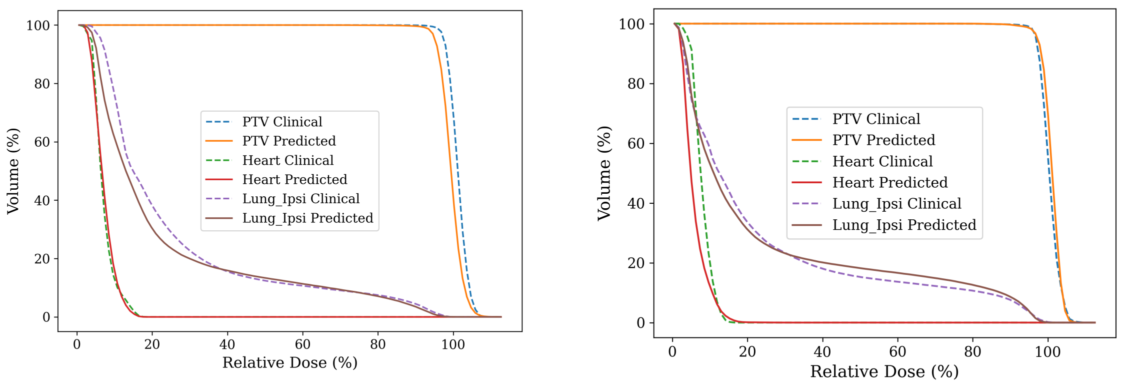
comparable to the widely used U-Net architecture. Although some metrics appear to improve with the hybrid model, the differences are not statistically significant, and therefore it cannot be conclusively stated that one model outperforms the other. Since transformer-based models are more computationally expensive to train, a training with a larger dataset could potentially enhance the results obtained.

One limitation worth noting is that the evaluation was restricted to a single baseline model. Future work should investigate more complex ViT variants, increase the size of the training dataset, and assess the effects of hyperparameter tuning to further optimize performance.

From a clinical perspective, the proposed model could be used as a starting point for automatic dose map generation, reducing the time required from dosimetrists during the planning process. This would allow professionals to focus on making small adjustments to the plan rather than designing it from scratch. This strategy could help increase the number of patients treated, while still ensuring high treatment quality.

V. CONCLUSIONS

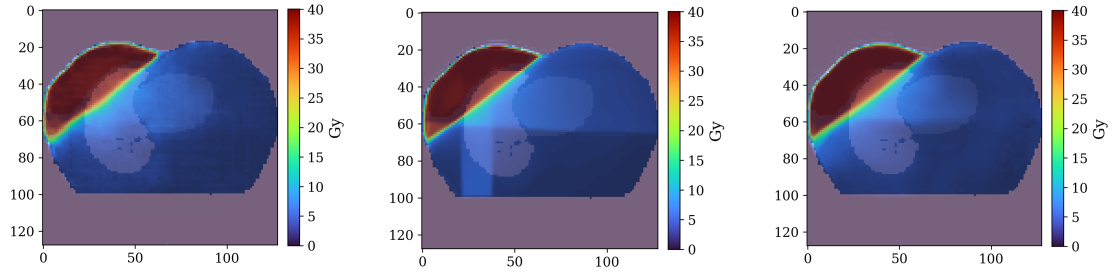
In conclusion, this study proposed a hybrid model combining a ViT encoder with convolutional layers as a decoder for dose prediction tasks in breast cancer radiotherapy. By comparing the model with well-established U-Net, the effectiveness of this new architecture has been demonstrated, achieving comparable performance in dose prediction. Nevertheless, the full potential of Transformers in medical physics remains underexplored due to the limited amount of available data. Further research could



(a) ViT & Convolutionals

(b) U-Net

FIG. 3: Comparison of the clinical (solid lines) and dose predicted (dash lines) DVH curves for a selected patient.



(a) ViT & Convolutionals

(b) Clinical

(c) U-Net

FIG. 4: Comparison of the clinical and predicted dose maps for a selected patient.

result in substantial improvements in both prediction accuracy and the reduction of clinical worktime associated with treatment planning.

Acknowledgments

This work would not have been possible without the collaboration of the Medical Physics Department of *Hospital de la Santa Creu i Sant Pau*, which provided me with the opportunity to take part in scientific research in a professional and specialized environment. I am especially grateful to Pedro Gallego for his valuable guidance and continuous dedication throughout these past months. I would also like to extend my appreciation to Bruno Juliá, for his support and for facilitating the development of this work. Finally, I want to express my thanks to my parents, my partner, and my close friends for their constant encouragement and concern during this entire process.

[1] SEOM y REDECAN. Cáncer de mama 2022
[2] Liu J et al. An overview of artificial intelligence in medical physics and radiation oncology (September 2023)
[3] Mashayekhi M et al. Artificial intelligence guided physician directive improves head and neck planning quality and practice uniformity: A prospective study (2023 May)
[4] Vaswani A et al. Attention Is All You Need (June 2017)
[5] Jiao Z et al. TransDose: Transformer-based radiotherapy dose prediction from CT images guided by super-pixel-level GCN classification (October 2023)
[6] Dosovitskiy A et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (October 2020)
[7] He K, Gan C et al. Transformers in Medical Image Analysis: A Review (February 2022)
[8] Rane & Nitin. Transformers for Medical Image Analysis: Applications, Challenges, and Future Scope (November 2023)

Automatització de dosimetries de càncer de mama amb Vision Transformers

Author: Ana Matas López, amataslo29@alumnes.ub.edu
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Bruno Juliá Diaz, brunojulia@ub.edu & Pedro Gallego Franco, pgallego25@gmail.com

Resum: Es proposa un model híbrid que combina un codificador *Vision Transformer* amb capes convolucionals per a la dosimetria en el càncer de mama. Les prediccions obtingudes s'han comparat amb les d'un model de referència *U-Net* entrenat en les mateixes condicions. Els resultats suggereixen una millora en les mètriques dels OAR, mantenint alhora una conformitat acceptable en la predicció de la dosi al PTV. Tanmateix, no s'han observat diferències estadísticament significatives, així doncs, calen més investigacions per explorar tot el potencial dels *Transformers* en l'àmbit de la radioteràpia.

Paraules clau: Radiació ionitzant, Accelerador Lineal, *Vision Transformers*, Reds Neuronals Convolucionals, Mecanisme d'atenció-pròpia

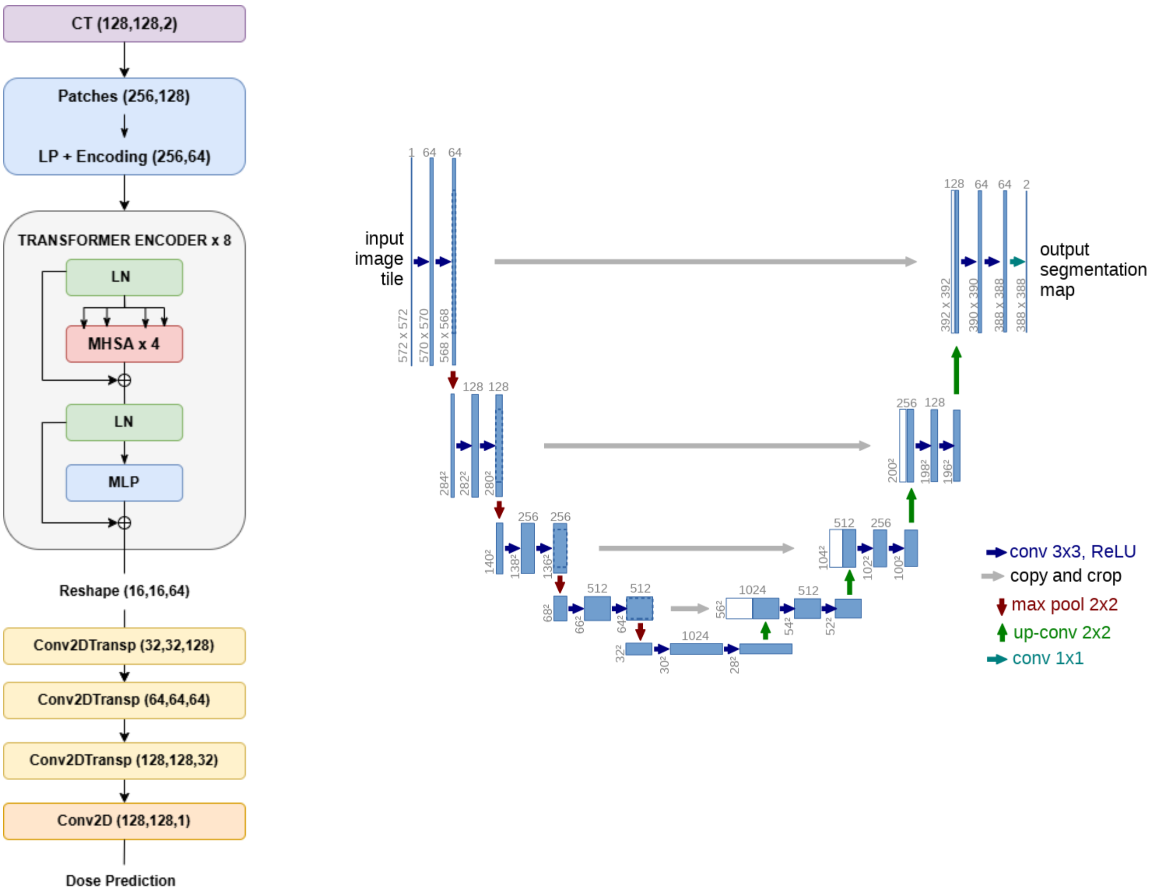
ODSs: Aquest TFG està relacionat amb els Objectius de Desenvolupament Sostenible (ODGs) 3, 4 i 9

Objectius de Desenvolupament Sostenible (ODSs o SDGs)

1. Fi de les desigualtats		10. Reducció de les desigualtats	
2. Fam zero		11. Ciutats i comunitats sostenibles	
3. Salut i benestar	X	12. Consum i producció responsables	
4. Educació de qualitat	X	13. Acció climàtica	
5. Igualtat de gènere		14. Vida submarina	
6. Aigua neta i sanejament		15. Vida terrestre	
7. Energia neta i sostenible		16. Pau, justícia i institucions sòlides	
8. Treball digne i creixement econòmic		17. Aliança pels objectius	
9. Indústria, innovació, infraestructures	X		

El contingut d'aquest TFG, part d'un grau universitari de Física, es relaciona amb l'ODS 3, i en particular amb la fita 3.d, que parla de reforçar la capacitat dels països en matèria d'alerta primerenca, reducció de riscos i gestió dels riscos per a la salut nacional i mundial, ja que contribueix a un tractament més eficaç del càncer de mama, que afecta gairebé un 30% de les dones espanyoles. També es pot relacionar amb l'ODS 4 i la seva fita 4.4, orientada a augmentar les competències tècniques i professionals entre els joves, ja que es promou una investigació en el sector del *Machine Learning* i la Intel·ligència Artificial dins d'un ambient universitari. Finalment, cal mencionar l'ODS 9 i la fita 9.5, basada en l'augment de la despesa en investigació i desenvolupament dels sectors públic i privat, perquè es suggereix l'estudi de models híbrids d'intel·ligència artificial per a la radioteràpia en tot tipus d'hospitals.

Appendix A: Models Architectures



(a) ViT & Convolutionals Architecture

(b) U-Net Architecture, *Ronneberger O et L. U-Net: Convolutional Networks for Biomedical Image Segmentation (May 2015)*

FIG. 5: Diagram illustrating the architectures used for dose map prediction. (a) A hybrid model combining Vision Transformers and Convolutional layers. (b) A conventional U-Net architecture as described in Ronneberger et al., 2015.