Non-Redundant Combination of Hand-Crafted and Deep Learning Radiomics: Application to the Early Detection of Pancreatic Cancer

Rebeca Vétil[1,2], Clément Abi-Nader[2], Alexandre Bône[2], Marie-Pierre Vullierme[3], Marc-Michel Rohé[2], Pietro Gori[1], and Isabelle Bloch[1,4]

[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2] Guerbet Research, Villepinte, France
[3] Department of Radiology, Hospital of Annecy-Genevois, Université de Paris, France
[4] Sorbonne Université, CNRS, LIP6, Paris, France
`rebeca.vetil@guerbet.com`

Abstract. We address the problem of learning Deep Learning Radiomics (DLR) that are not redundant with Hand-Crafted Radiomics (HCR). To do so, we extract DLR features using a VAE while enforcing their independence with HCR features by minimizing their mutual information. The resulting DLR features can be combined with hand-crafted ones and leveraged by a classifier to predict early markers of cancer. We illustrate our method on four early markers of pancreatic cancer and validate it on a large independent test set. Our results highlight the value of combining non-redundant DLR and HCR features, as evidenced by an improvement in the Area Under the Curve compared to baseline methods that do not address redundancy or solely rely on HCR features.

Keywords: Early Diagnosis · Pancreatic Cancer · Radiomics · Variational Autoencoders · Mutual Information

1 Introduction

Computational methods in medical imaging hold the potential to support radiologists in the early diagnosis of cancer, either by detecting small-size abnormal neoplasms [14], or even earlier in the disease course by recognizing indirect signs of malignancy. Such signs are usually subtle and organ-dependent, thus requiring a time-consuming and demanding clinical assessment. For example, in the case of pancreatic cancer, radiologists analyze the overall shape of the organ, check for fat replacement and note whether the pancreas shows atrophy and/or senile characteristics [7, 18, 19]. The identification of cancerous signs using automated tools can be based on radiomics, which are descriptors of texture and shape of a medical image, computed based on spatial relationships between voxels and their intensity distribution [11, 12]. Radiomics can be divided into two categories: (i) Hand-Crafted Radiomics (HCR), which are based on predefined mathematical formulas [11, 12]; (ii) Deep Learning Radiomics (DLR), estimated

using deep neural networks [10, 23], which may unveil additional complex relationships between voxels. HCR are generally extracted by open-source frameworks such as pyradiomics [24]. While such tools facilitate the standardization of the HCR, they only provide a limited number of predefined features. On the other hand, DLR features are typically extracted using either discriminative or generative models. Discriminative models frequently rely on one or multiple simple CNNs [3–5, 13, 20]. To prevent overfitting, some methods extract DLR by utilizing pretrained models trained on large datasets like ImageNet [3, 5, 20]. The deep neural networks commonly employed for computing these DLR features consist of multiple layers, with each layer producing potential features as its output. As a result, the choice of the layers to retain varies, with each method employing different heuristics to identify them [5, 20]. In the realm of generative models, auto-encoder (AE) networks are widely used [2]. AEs encode an image in a latent vector that is subsequently used to reconstruct the original image. This latent vector is considered to encapsulate the most descriptive features of the input image, making it a natural choice for representing the DLR [10, 21].

The two types of radiomics are complementary: the computation of DLR is data-driven, which ensures that the extracted features are adapted to a specific problem or type of data. On the other hand, the predefined and generic definitions of HCR may make them less adapted for a given specific task, but favors generalization and interpretability. Therefore, it has been recently proposed to combine HCR with DLR, arguing that this approach could result in an improved feature set for predictive or prognostic models [2]. The literature reports two main approaches to perform this combination: decision-level methods that train separate classifiers on DLR and HCR before aggregating their predictions [3, 5, 16], and feature-level methods that concatenate the two types of radiomics in a single feature vector which is then leveraged by a classifier [4, 13, 20]. These approaches extract HCR and DLR features independently, without guaranteeing complementarity between the two sets of features. As a result, the extracted DLR may be highly redundant with the HCR, limiting the value of their combination.

Given this context, we propose to extract DLR features that will complement the information already contained in the HCR. Our contributions are two-fold:

- A deep learning method, based on the VAE framework [9], that extracts non-redundant DLR features with respect to a predetermined set of HCR. This is achieved by minimizing the mutual information between the two types of radiomics during the training of the VAE. The resulting HCR and DLR features are leveraged to predict early markers of cancer.

- Validation of the proposed approach in the case of pancreatic cancer, using 2319 training and 1094 test subjects collected from 9 medical institutions with a split performed at the institution level. This is all the more important as most combination approaches have been solely evaluated in a cross-validation setting on mono-centric data [3, 5, 16].
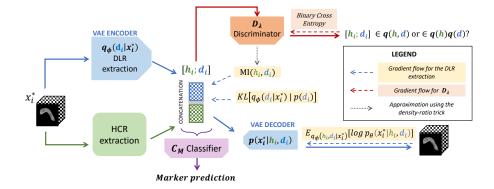
Fig. 1. Overview of our method. Starting from a masked image, Hand-Crafted Radiomics (HCR) are calculated analytically, while Deep Learning Radiomics (DLR) are extracted by the encoder of a VAE. These two types of radiomics are subsequently combined and given to the decoder for image reconstruction. The independence of HCR and DLR is enforced by the minimization of the Mutual Information (MI). The latter is approximated by the density-ratio trick [8], involving a discriminator $\mathcal{D}_\lambda$. Following the training of the VAE, a classifier $\mathcal{C}_M$ can be trained using both the HCR and DLR features to predict a specific marker of interest.

2   Method

Our method, illustrated in Figure 1, relies on a generative model that recreates a 3D input image from the concatenation of HCR and DLR features. Feature extraction is done analytically for the HCR and through a VAE encoder for the DLR. Independence between the features is encouraged through the minimization of their mutual information, which is estimated by a discriminator relying on the density-ratio trick [8]. Finally, the resulting features are given to a classifier for cancer marker prediction.

Generative framework. Let $x \in \mathbb{R}^V$ be a 3D image acquired via a standard imaging technique, and $y \in \{0,1\}^V$ the corresponding binary segmentation mask of a given organ, with V the number of voxels. In order to focus on a specific organ and facilitate the extraction of specific features, we work on the masked image $x^* = x \times y$. We postulate the existence of a generative model enabling us to create an image x^* from a low-dimensional representation space $[h, d]$ where $h \in \mathrm{R}^{N_h}$ and $d \in \mathrm{R}^{N_d}$ represent the HCR and DLR features with N_h and N_d being the number of hand-crafted and deep features, respectively. Assuming that x^* follows an independent and identically distributed Gaussian distribution, and that f_θ is a non-linear function mapping the concatenation of vectors $[h, d]$ to the masked image x^*, we hypothesize the following generative process:

$$p_\theta(x^* \mid y, h, d) = \prod_{v=1/y_v=1}^{V} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{(x_v^* - f_\theta([h,d])_v)^2}{2\sigma^2} \tag{1}$$

HCR and DLR features computation. We place ourselves within the VAE framework [9] and assume that $p(d)$ follows a Gaussian distribution with zero mean and identity covariance. HCR features are calculated analytically, while DLR features are computed by introducing the approximate posterior distribution $q_\phi(d \mid x^*)$. We hypothesize $q_\phi(d \mid x^*) \sim \mathcal{N}(\mu_\phi(x^*), \sigma_\phi^2(x^*)\mathbf{I})$, and maximize a lower bound of the marginal log-likelihood $\log p_\theta(x^* \mid y)$. We obtain the following loss function:

$$\mathcal{L}_{\text{VAE}} = - \mathbb{E}_{q_\phi(d|x^*)}[\log(p_\theta(x^* \mid y, h, d))] + KL[q_\phi(d \mid x^*) \mid p(d)] \qquad (2)$$

where KL refers to the Kullback-Leibler divergence.

Mutual Information Minimization. To promote the independence between HCR and DLR features, we propose to minimize their Mutual Information (MI), expressed here as $KL[q(h, d) \mid q(h)q(d)]$, where $q(h, d)$ represents the joint distribution of the DLR and HCR features, and $q(h)q(d)$ the product of their marginal distributions. These terms involve mixtures with a large number of components, making them intractable. Moreover, obtaining the direct Monte Carlo estimate necessitates processing the entire dataset in a single pass. Thus, we sample from these distributions to compute the MI: to sample from $q(h, d)$, we randomly choose an image x_i^*, extract its HCR features h_i as well as its DLR features d_i using the VAE encoder, and concatenate them. Samples from $q(h)q(d)$ are obtained by concatenating vectors h_k and d_j with $k \neq j$. Finally, to compute the MI, we need to compute the density-ratio between $q(h, d)$ and $q(h)q(d)$. To do so, we resort to the density-ratio trick [8], which consists in introducing a discriminator $\mathcal{D}_\lambda([h, d])$ able to discriminate between samples from $q(h, d)$ and samples from $q(h)q(d)$. Thus, we obtain:

$$KL[q(h, d) \mid q(h)q(d)] = \mathbb{E}_{q(h,d)}\left[\log \frac{q(h, d)}{q(h)q(d)}\right] \approx \sum_i \text{ReLU}\left(\left[\log \frac{\mathcal{D}_\lambda(h_i, d_i)}{1 - \mathcal{D}_\lambda(h_i, d_i)}\right]\right).$$
$$(3)$$

where the ReLU function forces the estimate of the MI to be positive, which prevents from back-propagating wrong estimates of the density-ratio. $\mathcal{D}_\lambda$ implementation is detailed in Section 6.1 of the appendix.

Optimization. The final loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{VAE}} + \kappa KL[q(h, d) \mid q(h)q(d)] \qquad (4)$$

This loss function is composed of two terms: the left-hand term, which is the common VAE loss function and promotes the reconstruction of the masked image while regularizing the approximate posterior distribution; and the right-hand term which minimizes the MI between $q(h, d)$ and $q(h)q(d)$, and enforces the extraction of DLR features which are not redundant with HCR features. The importance of the MI in the loss function is weighted by κ, which we empirically set to 1 (see Section 6.2 of the appendix for more details). To ensure that the density-ratio is well-estimated, as explained in [8], we opt for an alternate optimization scheme between the VAE model and the discriminator $\mathcal{D}_\lambda$: every 5
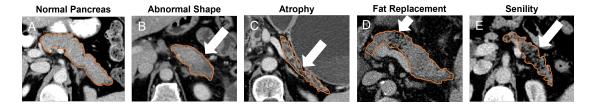
Fig. 2. Portal CT scans showing early markers of pancreatic cancer. Pancreas are delineated in orange. (A) shows a normal pancreas. White arrows indicate an abnormal enlarged tail (B), a parenchymal atrophy (C), fat replacement in the neck of the pancreas (D) and senile characteristics (E).

epochs, we freeze the optimization of the VAE, train the discriminator for 150 epochs, and continue the optimization of the VAE model.

Early cancer markers prediction. Once the VAE model is trained, DLR can be extracted and leveraged to predict cancer markers. We propose to train, for each marker of interest, a classifier $\mathcal{C}_M$ based on the concatenation of HCR and DLR extracted by our model. Unlike VAE training, which is unsupervised and task-agnostic, $\mathcal{C}_M$ training is supervised and specific to a cancer marker.

3   Experiments

We illustrate our method on the pancreas, for which we aim to predict four early markers of abnormality that manifest prior to the onset of visible lesions:

(i) *Abnormal shape:* Changes in the shape of the pancreas can be associated with pancreatic cancer as the tumor growth can lead to various structural changes in the pancreas [15, 25];

(ii) *Atrophy:* Pancreatic atrophy may signal pancreatic cancer [19] and can indicate small isodense lesions [26];

(iii) *Fat replacement:* Fat replacement is characterized by the accumulation of fat within the pancreas and is associated with various metabolic diseases, pancreatitis, pancreatic cancer, and precancer [7, 17, 19]. While this mainly modifies the texture, severe fat replacement can also affect the shape by inducing lobulated margins;

(iv) *Senility:* Anatomical changes in the pancreas, such as pancreatic atrophy, fatty replacement and fibrosis have been documented in elderly individuals and increase the susceptibility of individuals to pancreatic cancer [7, 18].

These early signs are illustrated in Figure 2.

Dataset. Data were obtained from our private cohort and split into two independent datasets $\mathcal{D}^{Train}$ and $\mathcal{D}^{Test}$, containing 2319 and 1094 abdominal portal CT scans from six and three independent medical centers, respectively. The reference labels regarding the early markers previously described were obtained

based on the assessment of the CT scan by a pool of 7 radiologists. Reference labels were collected for 676 cases of $\mathcal{D}^{Train}$ and all the subjects from $\mathcal{D}^{Test}$.

Preprocessing. For all the subjects, pancreas segmentation masks were obtained using a segmentation model derived from the nnU-Net [6] and manually reviewed by radiologists. The CT images and corresponding masks were resampled to $1 \times 1 \times 2$ mm^3 in the (x, y, z) directions, and centered in a volume of size $192 \times 128 \times 64$ voxels. Images intensities were clipped to the $[0.5, 99.5]$ percentiles and standardized based on the percentiles, mean and standard deviation of the pancreas intensities in $\mathcal{D}^{Train}$.

Extracting HCR and DLR. 32 HCR features were extracted utilizing the pyradiomics library [24], focusing exclusively on shape and first-order intensity features (see Section 6.3 in the appendix for the comprehensive list). Complementary DLR features were extracted using the VAE model of Section 2. The architecture followed the U-Net [22] encoder-decoder scheme without skip connections. The number of convolutional layers and the convolutional blocks were automatically inferred thanks to the nnU-Net self-configuring procedure [6] (see Section 6.4 in the appendix for details). The model was trained on $\mathcal{D}^{Train}$ for 1000 epochs. The dimension of DLR features d was set to 32, resulting in a final latent space dimension for the VAE of 64. Data augmentation consisting of rotation and cropping was applied during training.

Predicting early cancer markers. For each marker, a logistic regression was trained based on the concatenation of HCR and DLR features extracted from the subjects in $\mathcal{D}^{Train}$ for whom reference labels were available. The logistic regression was regularized using $\mathcal{L}_2$ penalty, with a default regularization coefficient of 1. Final predictions for $\mathcal{D}^{Test}$ were derived by ensembling models obtained through a four-fold cross-validation setup.

4   Results

Quantitative results. To demonstrate the usefulness of extracting DLR with MI minimization, two VAEs were trained. Both followed the same procedure (detailed in Figure 1) but differed only in the presence or absence of the MI minimization term in their loss function. Then, several logistic regression models with different inputs were trained in order to assess the effect of combining HCR and DLR features. In total, the following experiments were run:

- **HCR only:** H_{32} **and** H_{64}. These two experiments use the 32 basic HCR features described in Section 6.3 of the appendix, and H_{64} uses a further 32 HCR gray-level features calculated by the pyradiomics library [24] and selected by recursive feature elimination.
- **DLR only:** D_{32}^{MI} **and** D_{32}. 32 DLR features extracted by a VAE with and without MI minimization, respectively;
- **HCR + DLR:** HD_{64}^{MI} **and** HD_{64}. 32 basic HCR features + 32 DLR features extracted by a VAE with and without MI minimization, respectively.

Thus, the logistic regressions of H_{32}, D_{32} and D_{32}^{MI} used vectors of size 32, while those of H_{64}, HD_{64} and HD_{64}^{MI} used vectors of size 64. Prediction results for each of the four cancer markers are presented in Table 1.

Table 1. Pancreatic cancer marker prediction. For each experiment, we report the means and standard deviations of the AUC (in %) obtained by bootstrapping with 10000 repetitions. For each line, first and second best results are in bold and underlined, respectively. The last row shows the difference in AUC compared with H_{32}, averaged over the different markers. DLR and HCR refer to Deep Learning Radiomics and Hand-Crafted Radiomics, respectively.

	HCR only		**DLR only**		**HCR + DLR**	
	H_{32}	H_{64}	D_{32}	D_{32}^{MI}	HD_{64}	HD_{64}^{MI}
Abnormal Shape	68.38±0.07	68.11±0.07	67.66±0.07	**72.41±0.07**	71.2±0.07	70.07±0.07
Atrophy	81.05±0.06	<u>81.57±0.05</u>	74.08±0.07	79.08±0.06	80.82±0.06	**82.57±0.06**
Fat Replacement	<u>70.55±0.07</u>	69.78±0.08	65.96±0.08	65.74±0.07	69.28±0.08	**71.05±0.07**
Senility	71.63±0.08	70.21±0.08	70.18±0.07	69.1±0.08	72.28±0.08	**72.44±0.07**
δ w.r.t H_{32}	-	-0.48±0.07	-3.43±0.07	-1.32±0.07	<u>0.49±0.07</u>	**1.13±0.07**

The comparison between H_{32} and H_{64} showed that adding 32 gray-level HCR features was not beneficial as results were similar, or even decreased: for instance, for senility, the AUC went from 71.63 % (H_{32}) to 70.21 % (H_{64}). On average, the AUC of H_{64} lost -0.48 points compared with H_{32}. These experiments demonstrated the power of the 32 basic HCR features, and the need to find complementary features that would add value.

Then, for almost all markers, H_{32} outperformed D_{32} and D_{32}^{MI}, meaning that no VAE, whether trained with or without MI minimization, managed to automatically extract 32 DLR features as informative as the 32 basic HCR features used by H_{32}. For texture-related markers, such as fat replacement and senility, MI minimization did not produce clear differences. On the other hand, on shape-related markers, the DLR features learned by D_{32}^{MI} were shown to be more relevant than those learned by D_{32} with a basic VAE. Thus, on average, DLR features were better when extracted by a VAE trained with MI minimization, but still proved less informative than HCR features.

Finally, experiments HD_{64} and HD_{64}^{MI} showed that combining the two types of radiomics is beneficial since the average AUC gained 0.49 (HD_{64}) and 1.13 % (HD_{64}^{MI}) compared to H_{32}. Yet, results demonstrated that minimizing the redundancy produced the best results compared with all other approaches. Indeed, in HD_{64}, adding 32 DLR features produced variable results depending on the markers: compared to H_{32}, the AUC increased by a maximum of 2.82% for abnormal shape prediction, and dropped by a maximum of 1.27% for predicting fat replacement. On the other hand, HD_{64}^{MI} outperformed H_{32} on all prediction problems, meaning that the non-redundant DLR features systematically provided useful information.

Influence of the latent space. To explore the influence of the latent space dimension on the prediction performances, we replicated the $\text{HD}_{64}^{\text{MI}}$ experiment with increasing size L of the latent space, and reported prediction results in Table 2. Table 2 shows that increasing the latent space size resulted in lower classification performances. Specifically, a latent space size of 32 provided the most relevant DLR features.

Table 2. Pancreatic cancer marker prediction with varying latent space size. For each experiment, a VAE with Mutual Information (MI) minimization and latent space size L was trained. Predictions were obtained after training logistic regressions on 32 basic HCR features $+ L$ DLR features extracted by a VAE with MI minimization. We report the means and standard deviations of the AUC (in %) obtained on the test set by bootstrapping with 10000 repetitions. For each line, first best results are in bold. DLR and HCR refer to Deep Learning Radiomics and Hand-Crafted Radiomics, respectively.

	$L = 32$	$L = 64$	$L = 256$	$L = 512$	$L = 1024$	$L = 2048$
Abnormal Shape	**70.07**±**0.07**	69.02±0.07	68.87±0.07	69.91±0.07	69.33±0.07	68.68±0.07
Atrophy	82.57±0.06	82.28±0.05	81.77±0.06	**82.68**±**0.05**	80.9±0.06	80.21±0.06
Fat Replacement	**71.05**±**0.07**	70.91±0.07	70.23±0.08	70.45±0.08	69.55±0.07	68.96±0.08
Senility	**72.44**±**0.07**	72.02±0.07	70.38±0.08	71.65±0.08	72.03±0.07	69.6±0.08

Qualitative results. To visualize the effect of the extracted DLR features, we looked at the absolute value of the logistic regression weights for D_{32} and $\text{D}_{32}^{\text{MI}}$ in two ways. In Figure 3-A, the absolute value of these coefficients are displayed. The higher the absolute value of the coefficient, the higher its importance in the logistic regression prediction. When the MI was not minimized, HCR features had stronger importance than DLR ones. On the other hand, when we encouraged the independence between the two types of features through MI minimization, the contribution of DLR features to the prediction increased. Figure 3-B shows the number of DLR features among the k features with highest importance, for increasing values of k. $\text{HD}_{64}^{\text{MI}}$ and HD_{64} are shown in blue and orange, respectively. In addition, two extreme scenarios are shown: one where the logistic regression is predominantly influenced by the DLR features (in green), and another one where the logistic regression is primarily driven by the HCR features (in red). We can see that the blue curve approached the green curve, meaning that DLR features from $\text{HD}_{64}^{\text{MI}}$ contributed more to the outcome prediction. When the MI was not minimized, DLR features had less influence on the predictions as the orange curve approached the scenario in which DLR would be ignored.

Reconstruction performances. To explore the reconstruction performances of the VAE, we computed the average l_2 error per voxel between the original test images and their corresponding reconstructions. Upon applying nnU-Net's [6] automatic intensity normalization procedure, voxel intensities were observed to range from -3 to 2.3. Specifically, we employed a VAE with a latent space
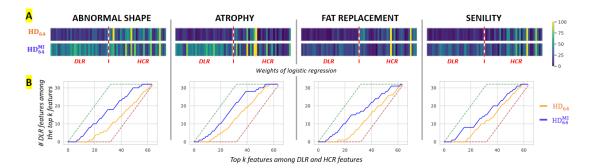
Fig. 3. Qualitative assessment of the Deep Learning Radiomics (DLR) and Hand-Crafted Radiomics (HCR) features through the coefficients of the logistic regressions. *A: Absolute value of the coefficients of the logistic regressions.* We plot, for each logistic regression corresponding to one marker, the absolute value of the coefficient for each of the 64 features. The first 32 features corresponded to DLR, while the 32 remaining features corresponded to HCR. *B: Number of DLR features among the top k features.* Dashed lines represent the extreme scenarios in which all 32 DLR are more informative than all 32 HCR (green) or all 32 HCR are more informative than all 32 DLR (red).

dimension of $L = 32$ and MI minimization during training. The resulting reconstruction error was found to be $(4.4 \pm 1.4) \times 10^{-3}$, which was comparable to the l_2 error obtained from a VAE trained without MI minimization, amounting to $(4.1 \pm 1.4) \times 10^{-3}$. These observations suggest that the introduction of MI minimization did not significantly impact the quality of the reconstructed images, neither resulting in deterioration nor improvement. Additionally, Table 3 further explores the relationship between reconstruction performance and latent space sizes, demonstrating that increasing the latent space size did not have a discernible effect on the quality of the reconstructions.

Table 3. Reconstruction performances with varying latent space sizes. For each experiment, a VAE with Mutual Information minimization and latent space size L was trained. We report the l_2 error per voxel between the original image and its reconstruction, with voxel intensities varying in $[-3, 2.3]$.

	$L = 32$	$L = 64$	$L = 256$	$L = 512$	$L = 1024$	$L = 2048$
l_2 error $\times 10^3$	4.4±1.4	4.4±1.4	4.4±1.4	4.4±1.4	4.3±1.4	4.3±1.5

5   Discussion and conclusion

We presented a method to learn DLR features that are not redundant with HCR ones. The method was based on the well-known VAE framework [9] that

extracted DLR features from masked images in an unsupervised manner. The complementarity between the two types of radiomics features was enforced by minimizing their MI, and the resulting features were used to train classifiers predicting different cancer markers. Experiments in the case of four early markers of pancreatic cancer indicated that our method increased prediction performances with respect to two state-of-the-art approaches. These findings suggest that our approach holds potential to improve patient survival outcomes. Qualitative results confirmed the advantages of minimizing the MI during training, as it resulted in the generation of DLR features that were complementary to HCR features and more prominently utilized for marker prediction. These results were obtained on a large and independent test set, which is particularly important as radiomics models require robust validation strategies to ensure their generalization and reproducibility when applied to new datasets [1]. With this in mind, it might be interesting to further encourage this feature efficiency by imposing independence between the DLR features themselves. Another research avenue could be to simplify the proposed pipeline by developing an end-to-end network capable of performing both feature extraction and classification tasks within a unified framework. Achieving this objective would necessitate the simultaneous training of the feature extractor and multiple sub-networks for each classification task. However, this approach might pose challenges in terms of training complexity, particularly due to the presence of substantial class imbalances across the various classification tasks. Alternatively, another possibility is to train an end-to-end convolutional neural network (CNN). Although more direct in nature, this approach would entail the training of a separate CNN for each question, which could be computationally heavier compared to the calibration of a logistic regression based on a single feature extractor, as suggested in our current work. Future studies should also address the interpretability of the extracted DLR features, as this aspect was not covered in the present work.

6   Appendix

6.1   Estimating the Mutual Information

The Mutual Information (MI) is estimated following the density-ratio trick [8] which requires to train a discriminator $\mathcal{D}_\lambda$ predicting whether concatenated radiomics vectors $[h, d]$ come from $q(h, d)$ or $q(h)q(d)$. Samples for training $\mathcal{D}_\lambda$ are obtained following the procedure shown in Figure 4. In practice, $\mathcal{D}_\lambda$ is modeled as a 2-layer Multi Layer Perceptron with ReLu activation, which is trained by minimizing a binary cross-entropy (BCE) loss term. Once the discriminator is trained, the MI between HCR and DLR features can be approximated as follows:

$$\text{MI}(h, d) = \mathbb{E}_{q(h,d)}\left[\log \frac{q(h, d)}{q(h)q(d)}\right] \approx \sum_i \text{ReLU}\left(\left[\log \frac{\mathcal{D}_\lambda(h_i, d_i)}{1 - \mathcal{D}_\lambda(h_i, d_i)}\right]\right). \quad (5)$$
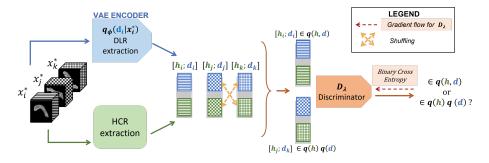
Fig. 4. Training $\mathcal{D}_\lambda$. Given three different input images x_i^*, x_j^* and x_k^*, the corresponding HCR and DLR features are computed: h_j, h_j, h_k and d_i, d_j, d_k. Samples from $q(h,d)$ are obtained by concatenating features of a same image (h_i and d_i for instance), while samples from $q(h)q(d)$ are obtained by concatenating h_k and d_j with $k \neq j$.

6.2   Influence of the hyperparameter κ

The final loss function for training our model is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{VAE}} + \kappa KL[q(h,d) \mid q(h)q(d)] \qquad (6)$$

where κ is a hyperparameter weighting the importance of the the mutual information in the total loss function. Table 4 reports prediction results obtained with different values of κ. According to these results, κ was set to 1 in all our experiments.

Table 4. Cancer marker prediction scores for different values of κ. For each experiment, we report the means and standard deviations of the AUC (in %) obtained by bootstrapping with 10000 repetitions. For each line, best result is in bold.

	$\kappa = 0.01$	$\kappa = 0.1$	$\kappa = 1$	$\kappa = 10$
General Shape	70.44±0.07	70.01±0.07	70.07±0.07	**71.03±0.07**
Atrophy	80.82±0.05	81.43±0.06	**82.57±0.06**	80.77±0.06
Fat Replacement	69.52±0.08	70.5±0.07	**71.05±0.07**	68.65±0.08
Senility	**73.14±0.08**	72.36±0.08	72.44±0.07	72.38±0.08

6.3   HCR features extraction

32 HCR features were extracted using the pyradiomics library [24]:

– **14 shape features** describing the size and shape of the pancreas

- Mesh Volume
- Voxel Volume
- Surface Area
- Surface Area to Volume ratio
- Sphericity
- Maximum 3D diameter
- Maximum 2D diameter in the axial plane
- Maximum 2D diameter in the coronal plane
- Maximum 2D diameter in the sagittal plane
- Major Axis Length
- Minor Axis Length
- Least Axis Length
- Elongation
- Flatness
- **18 first-order intensity features** describing the intensities distribution within the organ
  - Energy
  - Total Energy
  - Entropy
  - Minimum
  - 10th percentile
  - 90th percentile
  - Maximum
  - Mean
  - Median
  - Interquartile Range
  - Range
  - Mean Absolute Deviation
  - Robust Mean Absolute Deviation
  - Root Mean Squared
  - Skewness
  - Kurtosis
  - Variance
  - Uniformity

More details about each feature can be found on the online documentation.

6.4   Model Architecture

As detailed in Figure 5, the proposed variational autoencoder (VAE) followed a 3D encoder-decoder architecture. The network topology (number of convolutions per block, filter sizes) was chosen based on the nnU-Net self-configuring procedure [6], resulting in $1,110,240$ trainable parameters. The VAE was trained on 1000 epochs with a batch size of size 32. Every 5 epochs, the VAE was frozen and the discriminator $\mathcal{D}_\lambda$ was trained for 150 epochs with a batch size equal to the total training dataset. The VAE and $\mathcal{D}_\lambda$ were optimized using two independent Adam optimizers with a learning rate of 10^{-3}.
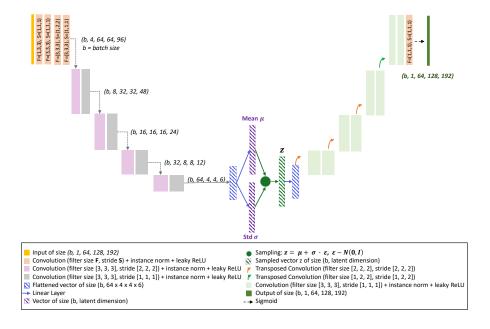
Fig. 5. Architecture of the proposed VAE

References

1. Aerts, H.J., Velazquez, E.R., Leijenaar, R.T., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications **5**(1), 4006 (2014)
2. Afshar, P., Mohammadi, A., Plataniotis, K.N., Oikonomou, A., Benali, H.: From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities. IEEE Signal Processing Magazine **36**(4), 132–160 (2019)
3. Antropova, N., Huynh, B.Q., Giger, M.L.: A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Medical Physics **44**(10), 5162–5171 (2017)
4. Chen, S., Qin, J., Ji, X., Lei, B., Wang, T., Ni, D., Cheng, J.Z.: Automatic scoring of multiple semantic attributes with multi-task feature leverage: a study on pulmonary nodules in CT images. IEEE Transactions on Medical Imaging **36**(3), 802–814 (2016)
5. Huynh, B.Q., Li, H., Giger, M.L.: Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. Journal of Medical Imaging **3**(3), 034501–034501 (2016)
6. Isensee, F., et al.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
7. Khoury, T., Asombang, A.W., Berzin, T.M., Cohen, J., Pleskow, D.K., Mizrahi, M.: The clinical implications of fatty pancreas: a concise review. Digestive Diseases and Sciences **62**, 2658–2667 (2017)
8. Kim, H., Mnih, A.: Disentangling by factorising. In: International Conference on Machine Learning. pp. 2649–2658. PMLR (2018)

9. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: 2nd International Conference on Learning Representations, ICLR (2014)
10. Kumar, D., Wong, A., Clausi, D.A.: Lung nodule classification using deep features in CT images. In: Conference on Computer and Robot Vision. pp. 133–138. IEEE (2015)
11. Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S.A., Schabath, M.B., Forster, K., Aerts, H.J., Dekker, A., Fenstermacher, D., et al.: Radiomics: the process and the challenges. Magnetic Resonance Imaging **30**(9), 1234–1248 (2012)
12. Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R.G., Granton, P., Zegers, C.M., Gillies, R., Boellard, R., Dekker, A., et al.: Radiomics: extracting more information from medical images using advanced feature analysis. European Journal of Cancer **48**(4), 441–446 (2012)
13. Lao, J., Chen, Y., Li, Z.C., Li, Q., Zhang, J., Liu, J., Zhai, G.: A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. Scientific Reports **7**(1), 10353 (2017)
14. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
15. Liu, F., Xie, L., Xia, Y., Fishman, E., Yuille, A.: Joint shape representation and classification for detecting PDAC. In: International Workshop on Machine Learning in Medical Imaging. pp. 212–220. Springer (2019)
16. Liu, S., Xie, Y., Jirapatnakul, A., Reeves, A.P.: Pulmonary nodule classification in lung cancer screening with three-dimensional convolutional neural networks. Journal of Medical Imaging **4**(4), 041308–041308 (2017)
17. Majumder, S., Philip, N.A., Takahashi, N., Levy, M.J., Singh, V.P., Chari, S.T.: Fatty pancreas: should we be concerned? Pancreas **46**(10), 1251 (2017)
18. Matsuda, Y.: Age-related morphological changes in the pancreas and their association with pancreatic carcinogenesis. Pathology International **69**(8), 450–462 (2019)
19. Miura, S., Kume, K., Kikuta, K., Hamada, S., Takikawa, T., Yoshida, N., Hongo, S., Tanaka, Y., Matsumoto, R., Sano, T., et al.: Focal parenchymal atrophy and fat replacement are clues for early diagnosis of pancreatic cancer with abnormalities of the main pancreatic duct. The Tohoku Journal of Experimental Medicine **252**(1), 63–71 (2020)
20. Paul, R., Hawkins, S.H., Balagurunathan, Y., Schabath, M., Gillies, R.J., Hall, L.O., Goldgof, D.B.: Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. Tomography **2**(4), 388–395 (2016)
21. Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., Yang, G.Z.: Deep learning for health informatics. Journal of Biomedical and Health Informatics **21**(1), 4–21 (2016)
22. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
23. Shafiee, M.J., Chung, A.G., Khalvati, F., Haider, M.A., Wong, A.: Discovery radiomics via evolutionary deep radiomic sequencer discovery for pathologically proven lung cancer detection. Journal of Medical Imaging **4**(4), 041305–041305 (2017)
24. Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.: Computational radiomics system to decode the radiographic phenotype. Cancer Research **77**(21), e104–e107 (2017)

25. Vétil, R., Abi-Nader, C., Bône, A., Vullierme, M.P., Rohé, M.M., Gori, P., Bloch, I.: Learning shape distributions from large databases of healthy organs: applications to zero-shot and few-shot abnormal pancreas detection. In: International Conference on Medical Image Computing and Computer Assisted Intervention, Part II. pp. 464–473. Springer (2022)
26. Yamao, K., Takenaka, M., Ishikawa, R., Okamoto, A., Yamazaki, T., Nakai, A., Omoto, S., Kamata, K., Minaga, K., Matsumoto, I., et al.: Partial pancreatic parenchymal atrophy is a new specific finding to diagnose small pancreatic cancer ($\leq$ 10 mm) including carcinoma in situ: comparison with localized benign main pancreatic duct stenosis patients. Diagnostics **10**(7), 445 (2020)