Representation learning in PET scans enhanced by semantic and 3D position specific characteristics

Theodoros P. Vagenas, Maria Vakalopoulou, Christos Sachpekidis, Antonia Dimitrakopoulou-Strauss, and George K. Matsopoulos, Senior Member, IEEE

***Abstract*— Representation learning methods that discover task and/or data-specific characteristics are very popular for a variety of applications. However, their application to 3D medical images is restricted by the computational cost and their inherent subtle differences in intensities and appearance. In this paper, a novel representation learning scheme for extracting representations capable of distinguishing high-uptake regions from 3D ^{18}F-Fluorodeoxyglucose positron emission tomography (FDG-PET) images is proposed. In particular, we propose a novel position-enhanced learning scheme effectively incorporating semantic and position-based features through our proposed Position Encoding Block (PEB) to produce highly informative representations. Such representations incorporate both semantic and position-aware features from high-dimensional medical data, leading to general representations with better performance on clinical tasks. To evaluate our method, we conducted experiments on the challenging task of classifying high-uptake regions as either non-tumor or tumor lesions in Metastatic Melanoma (MM). MM is a type of cancer characterized by its rapid spread to various body sites, which leads to low survival rates. Extensive experiments on an in-house and a public dataset of whole-body FDG-PET images indicated an increase of 10.50% in sensitivity and 4.89% in F1-score against the baseline representation learning scheme while also outperforming state-of-the-art methods for classifying MM regions of interest. The source code will be available at `https://github.com/theoVag/Representation-Learning-Sem-Pos`.**

***Index Terms*— FDG-PET images classification, Metastatic Melanoma, Position Encoding Block, Representation Learning, Semantic sampling, VICReg**

I. Introduction

Metastatic Melanoma (MM) is a form of cancer that can spread throughout the body by giving metastases at many different sites and organs and is associated with very low survival rates [1]. Its metastatic nature and the highly variable tumor structure make MM very difficult to treat. Immunotherapy with novel immune checkpoint inhibitors and targeted immunotherapies have recently shown significant improvement in managing this disease and extending the patient's survival [2]. In the era of personalized medicine, tight therapy monitoring is required in order to early predict a patient's response to treatment and tailor it accordingly. Two image modalities that are often used in cancer monitoring are the Positron Emission Tomography (PET) in conjunction with Computed Tomography (CT). More specifically, ^{18}F-Fluorodeoxyglucose (FDG) PET images indicate high uptake of FDG in tumor regions as compared to surrounding tissue [3]. However, there is also a physiologically enhanced FDG accumulation in organs like the brain and liver as well as in the urinary tract due to the exception of the tracer. Clinicians can extract quantitative biomarkers from the FDG images, such as Standardized Uptake Values (SUV), Metabolic Active Tumor Volume (MATV) or Total Lesion Glycolysis (TLG) [4].

However, the extraction of such biomarkers requires the localization of tumor lesions from the FDG-PET images. Manual segmentation of these images constitutes a labor-intensive and time-consuming procedure that can significantly stall the clinical workflow due to the large number and the heterogeneity of tumor lesions of MM. In order to accelerate the procedure many semi-automatic segmentation techniques, such as fixed or adaptive thresholds, have been initially utilized in the clinical environment [4]. With the recent advances in Artificial Intelligence (AI) many Deep Learning (DL) techniques have been proposed for the segmentation and classification of Regions of Interest (ROIs) in medical images and lesions' localization [5]. Such DL-based segmentation techniques for the delineation of tumor lesions from whole-body FDG-PET images using variants of UNet [6], GANs [7] and Transformers [8] have been proposed. Classification tasks concerning the treatment evaluation [9] or the characterization of ROIs [10] or patients suffering from cancer [11] have been evaluated with promising results. However, training deep neural networks necessitates large annotated databases of images that cannot be obtained efficiently. In addition, many of the previous segmentation techniques resulted in a large number of false positive regions, which limited their clinical utilization. In

T. P. Vagenas and G. K. Matsopoulos are with the School of Electrical and Computer Engineering, National Technical University of Athens, 15773 Athens, Greece (e-mail: tpvagenas@biomed.ntua.gr, gmatsopoulos@biomed.ntua.gr)

Maria Vakalopoulou is with CentraleSupélec, University of Paris-Saclay, France and Archimedes/Athena RC, Greece. (e-mail: maria.vakalopoulou@centralesupelec.fr).

Christos Sachpekidis (CS) and Antonia Dimitrakopoulou-Strauss (ADS) are with Clinical Cooperation Unit Nuclear Medicine, German Cancer Research Center, 69120 Heidelberg, Germany (e-mail: c.sachpekidis@dkfz-heidelberg.de, a.dimitrakopoulou-strauss@dkfz.de)

this paper, we propose a data-efficient representation learning based framework capable of accurately separating true tumor lesions from non-tumor related uptake.

Recently, representation learning through Self-Supervised Learning (SSL) has become very popular due to its ability to learn useful representations without the need for large, manually annotated datasets. Our method employs VICReg, an SSL framework, but incorporates supervision to guide the learning process. Therefore, we present SSL methods because they are closely related to our approach and provide essential context, as understanding their strengths and challenges, especially in the medical imaging domain, is crucial to our motivation. Commonly in SSL techniques the model is initially pretrained on unlabeled data through the comparison of augmented views of the images to learn invariant and discriminative semantical representations by minimizing their distance [12]. One of the most common methodologies in SSL is contrastive learning, which aims to learn representations of the data where similar samples are close while dissimilar samples are far apart [13]. Other approaches, such as the VICReg [14], aim to apply regularizations into the loss function to avoid collapse and improve learning by promoting representations' variance and decorrelating the feature dimensions.

Although SSL has been applied to medical imaging, its application to medical images has to alleviate limitations opposed by their nature. High inter-class similarity and low-intensity variation in medical images hinder the extraction of discriminative features. Strong augmentations risk distorting diagnostic information, while weak augmentations may lead to trivial solutions. The problem is compounded by imbalanced datasets with fewer disease-related images and potential bias introduced during training, which limits model generalizability. To this end, defining good pretext tasks for the training is crucial for the quality of representations. Furthermore, enriching the representations with spatial information beyond the intensity details can significantly enhance the extracted features. The requirement of representation learning for large batch sizes can oppose significant computational burden in using the whole 3D medical images without downsampling [12], [15], [16]. Our method enhances semantic features through trainable layers, yielding robust representations with minimal computational overhead. This efficient representation learning scheme employs class-related sampling and integrates both semantic and positional features to generate useful representations from FDG-PET ROIs.

In this paper, we propose a novel position-enhanced representation learning scheme to produce representations combining both semantic and 3D position information capable of classifying ROIs of high uptake from FDG-PET images from MM patients as tumor lesions or not. We introduce the Position Encoding Block (PEB), which calculates and integrates a position vector that includes the centroid location, the bounding box size and the rotation vector, into a VICReg [14] based framework that supports the learning of robust representations that preserve their variety. These positional details, integrated via the PEB, are critical for diagnosis as they capture anatomical variations and typical tumor morphology, thereby enhancing the quality of the extracted representations. This methodology paves the way for the further utilization of meaningful representations in enhancing DL-based classification accuracy in 3D medical images, such as whole-body FDG-PET images of cancer patients. Our main contributions are summarized as follows:

- The development of a novel representation learning method to enhance the detection of MM tumor regions by learning embeddings, which can encode tumors' characteristics from the 3D FDG-PET in the clinical workflow.
- A new 3D position-aware Position Encoding Block (PEB) inside the representation learning method to introduce spatial information to the feature vector. The combination of semantic and position information in the representation vector provides enhanced representations where the semantic features are influenced, amplified or suppressed, according to the studied region's anatomical position, size and rotation.

Our method was extensively evaluated on both an in-house and a public dataset of high-uptake regions from FDG-PET images from MM patients. The proposed model achieved superior results against state-of-the-art methods previously used for MM classification. Comparison with multiple and diverse schemes supports the additive value of the PEB in the majority of them with an increase of AUC ranging from approximately 1% to 5%. The proposed representation learning method, when combined with a two-layer Multilayer Perceptron (MLP) classifier trained on the same dataset as the supervised backbone network without position information, achieved an improvement of 4.04% in the F1-score, 2.83% in Balanced Accuracy (B.Acc.) and 3.2% in Area Under The Curve (AUC) compared to it. Additionally, considering settings with limited annotated regions, when using only the 15% of the training set, the proposed method outperformed the supervised ResNet18 version without position for 5.70% in F1-score, 6.47% in B.Acc. and 5.18% in AUC.

II. Related Works

Tumor detection and quantification is crucial for monitoring and assessing treatments of MM. Earlier studies for segmenting tumor lesions from whole-body FDG-PET images included threshold-based techniques and their combination with manual corrections [4]. FDG-PET imaging quality for MM patients was evaluated using SUV values and comparisons with liver and background values in [17]. In [6], the authors presented a public dataset including FDG-PET images of MM patients among other types and evaluated nnUNet for the segmentation task also aiming to reduce the number of false positive regions. In our previous work [18], a clustering-based segmentation method was combined with radiomics features and neural networks to identify tumor lesions from the FDG-PET/CT images. In this work, we propose a representation learning method to produce representations that can accurately distinguish tumors from non-tumor regions by suppressing many false positive regions and automatically without opposing restrictions related to the region's size.

AI-based classification methods have been widely proposed for identifying and categorizing tumor-related regions

or patients from FDG-PET/CT images. Convolutional Neural Networks (CNNs) and, in particular, ResNet architectures [19] were used to classify FDG-PET/CT patient images into malignant, benign, and equivocal examinations, or regions concerning lung cancer and lymphoma in [20] and [21] respectively. In this direction, FDG-PET images solely were utilized in [22], to classify whole-body scans as normal or pathological using a lightweight fully convolutional solution. However, identifying all ROIs inside the scans instead of simply finding normal and abnormal scans can be crucial, as it aids diagnosis by enabling the calculation of metrics, e.g. Total Metabolic Tumor Volume (TMTV), which depends on precise segmentation. A two-step procedure for the segmentation of high uptake regions and their classification as prostate cancer-related was combined with a 3D DenseNet model for PSMA PET/CT images in [23]. In [10], a CNN combined with two dense layers was applied to classify suspicious regions concerning lung cancer and lymphoma. Previous works for FDG-PET/CT images limited their effectiveness by employing either 2D UNet or 2D transformer [24], which loses the 3D spatial information, or by focusing on segmenting ROIs in only specific anatomical regions such as lungs. Furthermore, using only supervised learning requires large annotated databases that are difficult to acquire in medical tasks.

Nevertheless, there has been limited research on AI-driven classification to support the MM clinical workflow. Although during the autoPET challenge, segmentation of FDG-PET images for different cancers has been studied, limited works focused on the MM case while also the presented results, in terms of DSC overlap, false positive regions and their generalizability, need further improvement before methods' application to the clinical workflow [6]. A pilot study for predicting treatment response and mortality of MM patients used PET/CT and PET/MRI imaging and combined manually calculated features with CNNs [25]. In [26] hyperprogression of MM lesions was predicted by extracting radiomics features from the manually segmented ROIs in FDG-PET/CT. However, manual segmentation in whole-body 3D images is a labor-intensive and time-consuming procedure that poses challenges for its integration into clinical workflows. The paper [27] proposes a threshold-based segmentation to extract high-uptake regions and a CNN-based method to identify tumorous patches from FDG-PET/CT images for MM. Based on this work, they also presented a method using thresholding and finetuning via nnUNet to segment MM patients' images and calculate features for treatment outcome prediction [9]. In general threshold-based techniques are not robust against the tumor heterogeneity of the whole body FDG-PET images as regions often exhibit large intensity variations and lack well-defined boundaries. Furthermore, they extracted small patches resulting in a loss of the spatial context including the global positioning in the body and the sizing details. Our method can efficiently learn to produce useful representations that combine the regions' semantic and spatial information to effectively reduce false positives and classify them as tumor lesions or not.

Supervised DL techniques require large databases of annotated images, the construction of which is not cost-effective due to the labor-intensive annotation procedure. Self-supervision has recently been studied to alleviate this requirement by learning useful representations without manual annotations. In [28], a self-supervised method including translation-aware features and adversarial and construction loss for supporting the detection of pathologies in Optical Coherence Tomography and X-ray images was proposed to classify 2D patient images. For lung tumor classification from CT, SSL-based schemes utilized a UNet to restore samples' images from their augmentations [29] or a contrastive scheme for samples from different anatomical regions [30]. However, focusing only on a specific region, such as the lung, and validation of only one dataset from one source restricts the application to a group of regions with limited variability. On the contrary, our work exploits each region's location, sizing, and rotation to produce better representations of the regions inside the whole-body images of two datasets. The authors in [31] utilized a mask autoencoder for pretraining and self-supervision training aiming to extract anatomy-dependent noise for low-dose PET/CT images. The study [32] presents a UNet-based architecture with Atrous Spatial Pyramid Pooling (ASPP) for multi-scale feature extraction in lymphoma tumor segmentation from PET/CT images. These weak supervision schemes might lead to suboptimal performance for infrequent cases in the dataset while training with high computational demands limits their scalability to whole-body 3D images. To address this, our approach uses small boxes to reduce memory usage, allowing larger batch sizes. Building on this procedure, we sampled pairs of tumors and non-tumors through a tailored learning scheme to alleviate the imbalance in the dataset and learn semantic features.

Recently, research works aiming at combining anatomical or positional features in the context of medical images have been presented. For classification, [33] utilizes positional embeddings of horizontal and vertical strips within the transformer architecture, but in a fully supervised scheme. Hierarchical contrastive learning to compare the representations from coarse anatomical structures to finer ones in [34], focused on splitting the 2D X-ray images into non-overlapping patches which can limit its utilization in 3D images and classification of small regions inside them. The [35] introduces positional details by assigning a label to 2D slices related to the slice number in a volume in a contrastive learning scheme restricting the backbone model to 2D feature extraction and considering only the relative position of the 2D slice. Similar to this, [36] proposed 2D contrastive learning for CT classification, utilizing weak supervision through kernel loss and positional data by tracking the slice's location. Moreover, the [37] aimed at learning a position vector as a pretext with a boundary-based reconstruction to capture spatial information for segmentation tasks. Anatomical information has also been studied for SSL through sampling patches from the same or different anatomical regions, based on an atlas and a registration method, to contrast them [30]. Although extracting neighboring patches with similar anatomical information as a pretext task was applied to low-dose CT [38], the method focused on denoising and the position was used in terms of neighboring patches. Learning the spatial relationship between imaging planes was

applied in [39] for segmenting and classifying larger or specific structures such as the heart and the knee. One recent work incorporated positional priors for SSL pretraining for 3D CT organ segmentation through establishing a pretext task, that predicts class assignments of random cropped regions based on their position according to the building blocks [40]. The [41] attempts to use as a pretext task a regression task to recognize the part of the body for organ segmentation. Most of the previous works suggest introducing position as a pretext task but predicting spatial details alone cannot be linked directly to producing features concurrently aware of anatomy/position and class-specific information. In our work, an enhanced position vector is embedded into a semantically aware representation with learnable modules to regulate its participation in the final representation and adapt the features' importance according to the anatomical position inside the whole body image.

III. Materials and Methods

A. Overview

The main architecture, as depicted in Fig. 1, consists of the variant of the VICReg [14] scheme enhanced by the PEB, which combines the semantic features extracted from the CNN backbone and the spatial features. The augmentation and sampling module initially creates pairs of regions with the same class, (1) for a tumor pair or (2) for a non-tumor pair, to be inserted in the representation learning framework. The pair of regions passes through the backbone and the region's position is calculated and integrated with the output of the backbone by the PEB. The enhanced vectors are inserted into the projection head, and finally, its output is used to compute the VICReg-based loss function.

B. Dataset

Private dataset: The in-house dataset consists of FDG-PET/CT whole-body images from 44 patients diagnosed with Metastatic Melanoma Stage IV [42]. The study was approved by the Ethical Committee of the University of Heidelberg (S-107/2012) and the Federal Agency for Radiation Protection (Bundesamt für Strahlenschutz, Z 5–22463/2–2012-016). Whole-body PET/CT images were acquired on a dedicated PET/CT system (Biograph mCT, S128, Siemens Co., Erlangen, Germany). Each PET image underwent attenuation correction, and the reconstruction of images was performed iteratively using a matrix of (400 × 400) pixels and a voxel size of 2.04 × 2.04 × 4 mm^3. Experts produced the ground truth segmentation masks, delineating the tumoral regions semi-automatically by applying manual corrections to the masks extracted by the clustering-based segmentation [18]. In total, 3161 regions (389 tumors and 2772 non-tumors) were indicated in 44 scans. For the experiments on the private dataset, due to its small size, 50% of the dataset was used as a test, and the remaining 50% for training supplemented by the AutoPET training patients.

AutoPET grand challenge: This dataset originates from the public AutoPET grand challenge dataset [6] which includes 1025 whole-body FDG-PET/CT images from patients with NSCLC, melanoma, and malignant lymphoma, as well as negative controls. From this dataset, we selected only the first study from the patients diagnosed with malignant melanoma to meet the study's particular goals. This resulted in a dataset of 177 FDG-PET/CT images. PET was obtained by a Siemens Biograph mCT clinical scanner using a 3D-ordered subset expectation maximization algorithm (matrix size 400 × 400 and voxel size of 2.04 × 2.04 × 3 mm^3). In total, 15890 regions (2023 tumors and 13867 non-tumors) were indicated in 177 scans. High uptake regions' extraction from the datasets was done based on the unsupervised clustering of our previous research work [18]. Large organs, the Brain, Heart, Kidneys, and Bladder, were manually excluded from the dataset to focus on ROIs, which were more challenging to classify. The autoPET dataset was divided into train/validation and test splits with 80% and 20% of the patients, respectively, by considering the ratio of tumors/non-tumor regions to be approximately constant across the splits. All FDG-PET images were initially transformed to Standardized Uptake Values (SUV) normalized by body weight (SUVbw) with the following equation [43]:

$$\mathrm{SUV(g/mL)} = \frac{\text{Tissue Radioactivity (Bq/mL)}}{[\text{ Injected Dose (Bq)}/\text{ Weight (g)}]} \quad (1)$$

where tissue activity was decay-corrected to account for the time elapsed from injection to acquisition and weight refers to the body weight.

C. Proposed representation learning scheme

1) Augmentation and Sampling module: Most commonly, representation learning and SSL methods require pairs of images or augmented views of the same image to learn useful representations. However, in the context of medical images, transformations alone are not adequate to learn representations useful for classification [15]. Furthermore, in medical images, such as FDG-PET images, intensity variations are more subtle and object boundaries are less prominent. This fact limits the capabilities of the SSL to extract useful information in contrast to the natural images that include discriminative regions and characteristics. We created pairs of regions of the same class to drive the training scheme and the network to learn to produce class-related representations through semantically aware comparisons as the pretext task. The use of these pairs enhances the initially unsupervised representation learning by guiding the model to focus on data similarities instead of explicit labels. In addition, this sampling can be easily adapted to many SSL techniques as demonstrated in TABLE I and it works well in combination with the proposed PEB.

The manually annotated class labels from each dataset were used to select the pairs of regions of the same class. The class-specific sampling strategy was able to alleviate the highly imbalanced dataset by inserting pairs of tumors and pairs of non-tumor regions in each batch. With this method, the initially relatively small number of tumor lesions led to a substantially larger pool of pairs of regions from which we can sample to train the network. In this regard, the framework aims to maximize the agreement of different regions with the same semantic meaning and learn a space where tumors

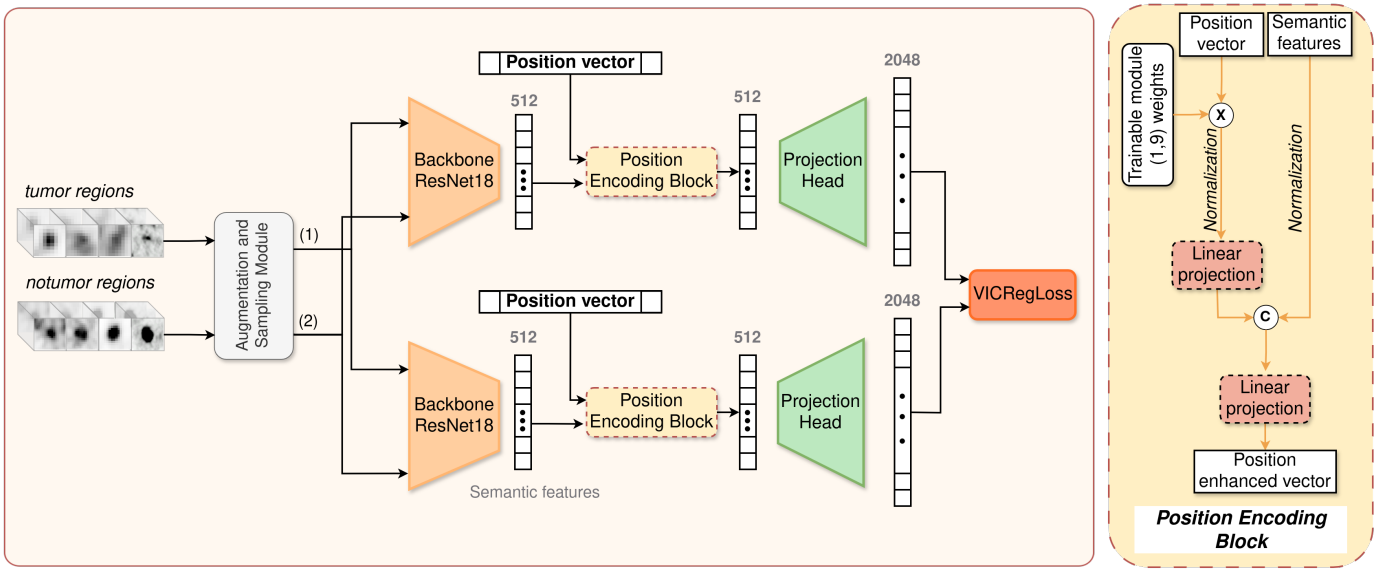


Fig. 1: The left part presents the proposed representation learning scheme with the sampling module and the PEB, while the right one shows the PEB architecture, which fuses the position information with the original embedding.

are separated from non-tumor regions. In the final setting, we apply random sampling of pairs of regions with a higher probability for tumors and include pairs with augmented views with a 30% rate. This approach introduces easier examples alongside the most challenging pairs of regions from the same class. Five main augmentations were applied to the ROIs to create an augmented version of the initial image: random affine transformations, flip and rotate, gaussian noise and histogram shifts with a probability of 80% each.

2) Main representation learning architecture: The main representation learning architecture utilized to produce representations from the ROIs is based on VICReg, first introduced in [14]. This architecture is suitable for extracting semantic information from ROIs in medical imaging and for our specific task of distinguishing high-uptake regions due to its main properties. Firstly, the invariance term drives the representations of images of the same meaning closer by minimizing their distance. Secondly, its variance regularization property can mitigate the collapsing to trivial representations problem of SSL techniques, which is a crucial challenge in medical images where the differences in intensity and appearance are subtle. Thirdly, the covariance term of the VICReg framework decorrelates the features inside the embedding leading to representations that preserve the useful information and avoid redundant information in their features. Furthermore, the VICReg architecture is favorable against other representation learning techniques because it does not require large batch sizes or a memory bank that could oppose excessive computational burden.

As it is depicted in Fig. 1, the main architecture is composed of two branches, including an encoder, which is the backbone that extracts the semantic features, the PEB that combines the semantic and spatial information in one enhanced vector, the projection head and the VICReg loss function. Initially, the two ROIs, which include a pair of an image and its augmentation or a pair of ROIs, I and I', belonging to the same class as described earlier are inserted into the same backbone, a ResNet18 in our case, to extract the semantic information. Concurrently, each ROI's position vector with respect to the initial 3D scan is calculated, and the two vectors are inserted in the PEB. The two representations, semantic and location, are combined into one, resulting in one representation per image, $y = f_\theta(I)$ and $y' = f_\theta(I')$. Next, the enhanced vectors are inserted into the projection head in order to be projected and mapped to the z embeddings, $z = h_\phi(y)$ and $z' = h_\phi(y')$ that will be used for the calculations in the loss function. The projection head, which consists of two MLPs, expands the dimension to four times the dimension of the features (features' dimension is 512).

3) Loss function: The VICReg-based loss function, as defined in [14], aims to minimize the distance between embeddings of the same images while simultaneously ensuring that the variance of each dimension in the vector across the batch remains above a threshold. Concurrently, it decorrelates the features to preserve the information in the produced representations. The main loss function is calculated by the following equations.

Invariance loss: The invariance loss is calculated by the mean of the squared distance between the pairs of the outputs of the two branches of the architecture. This term aims to bring the representations of the views of each pair of regions closer to each other. Given an image x_i in a batch and y_i its pair (P) drawn from the same class, their representations, produced by one of the two branches. are denoted Z_x and Z_y, respectively.

$$s\left(Z_x, Z_y\right) = \frac{1}{n}\sum_i dist^2(z_{x_i}, z_{y_i}), \qquad dist = \|z_{x_i} - z_{y_i}\|_2 \tag{2}$$

where each branch includes n vectors of size d, z_{x_i} and z_{y_i} denotes representations of two regions of the same class.

Variance loss: For the variance term v, the standard deviation for each one of the d dimensions of the batch of vectors is used in a hinge function as calculated below:

$$v(Z) = \frac{1}{d}\sum_{j=1}^{d} \max\left(0, \gamma - \sigma\left(z[,j], \epsilon\right)\right) \tag{3}$$

where σ denotes the standard deviation regularized with ϵ (a small value to avoid unstable calculations), $z[,j]$ denotes the values from the j dimension of the Z which equals to a vector with the j_{th} variable of all representations in the batch Z. Finally, the target standard deviation value along each j dimension of the batch is $\gamma = 1$.

Covariance loss: The covariance matrix of Z is defined as:

$$C(Z) = \frac{1}{n-1}\sum_{i=1}^{n}\left(z_i - \bar{z}\right)\left(z_i - \bar{z}\right)^T \tag{4}$$

where $\bar{z}$ is the mean of z_i. The covariance loss term $c(z)$ equals the sum of the squared off-diagonal elements of $C(z)$ divided by the number of dimensions d. By driving the off-diagonal elements of the covariance matrix toward zero, it decorrelates the features and reduces redundant information.

$$c(Z) = \frac{1}{d}\sum_{i\neq j}[C(Z)]_{i,j}^2 \tag{5}$$

Weighted average: The loss of each batch is calculated as the weighted average of the three terms (predefined $\lambda = 25, \mu = 25, \nu = 1$).

$$\begin{aligned}\ell\left(Z_x, Z_y\right) = \lambda s\left(Z_x, Z_y\right) + \mu\left[v(Z_x) + v\left(Z_y\right)\right] + \\ + \nu\left[c(Z_x) + c\left(Z_y\right)\right]\end{aligned} \tag{6}$$

The Loss function is calculated for the entire dataset D.

$$Loss = \sum_{i\in D}\sum_{x,y\sim P}\ell\left(z_{x_i}, z_{y_i}\right) \tag{7}$$

D. Position vector

This section presents the position vector that enhances the semantic features from the backbone. In medical images, the characterization of ROIs is not only affected by the intensity fluctuations but also by the region's location inside the body, the region's size, and the rotation according to the axes. Positional details can adapt the features according to the spatial context as ROIs have different shape or texture characteristics in the different anatomical regions. We calculated a position vector based on these three observations: *(i)* The position of a region inside the body in FDG-PET plays an important role in its characterization. Some regions contain tumor lesions with less frequency than others. Different positions will enable different features inside the embedding with more participation. *(ii)* The size of the bounding box of a region can also provide useful information for the classification. For example, regions that extend to a very small distance in one axis may refer to artifacts. *(iii)* The three angles that the region creates with respect to the x, y, and z axes in the three-dimensional space complement the position of the centroid and provide a more detailed position in the space.

Considering the above, we calculate and insert into our network a position vector for each region with the following structure:

$$[c_x, c_y, c_z, s_x, s_y, s_z, \theta_x, \theta_y, \theta_z] \tag{8}$$

For each ROI, the centroid (c) is computed from its coordinates inside the image, and the size (s) is determined from the bounding box dimensions using straightforward measurements. The calculation of the rotation vector with the axes angles is presented below:

$$R = \begin{bmatrix} R_{00} & R_{01} & R_{02} \\ R_{10} & R_{11} & R_{12} \\ R_{20} & R_{21} & R_{22} \end{bmatrix} \tag{9}$$

$$\theta_x = \arctan 2\left(R_{21}, R_{22}\right) \tag{10}$$

$$\theta_y = \arctan 2\left(-R_{20}, \sqrt{R_{21}^2 + R_{22}^2}\right) \tag{11}$$

$$\theta_z = \arctan 2\left(R_{10}, R_{00}\right) \tag{12}$$

where R is the rotation matrix, and θ_x, θ_y, θ_z are the angles with the axes.

E. Enhanced representations with the Position Encoding Block (PEB)

On the right side of Fig 1, the PEB is presented. This module takes as inputs the semantic features vector from the backbone and the position vector calculated for each ROI and outputs the final position enhanced vector, which is a representation effectively combining the semantic and position information. Initially, the position vector is multiplied by trainable weights of the same size, which, during training, regulate the weight/importance of each value of the position vector to the formulation of the combined vector. Layer normalization is applied to both the semantic and location feature vectors to normalize activations across channels in each layer and enhance training stability. A linear layer to linearly project the position vector to a space with a dimension equal to the dimension of the semantic feature vector is applied after the multiplication. The two feature vectors are concatenated (symbol C in Fig. 1) across the channel dimension, resulting in a vector of $2 \cdot 512 = 1024$ size. In the first part of the PEB, the semantic and position vectors are concatenated to create a space where the position and semantic information lie in separate dimensions. In this setup, the semantic and the position information are encoded in different batches of dimensions without mixing them up. Then the linear layer mixes these dimensions and projects them to a space with half dimensions. This is done for two reasons, first to facilitate implementation while maintaining a consistent vector size of 512 and secondly the weights of this layer effectively combine the information by assigning each feature from the position and semantic vector different importance in creating the final vector.

We evaluated the strength of our representations on the task of tumor classification. In Fig. 2, the overall downstream

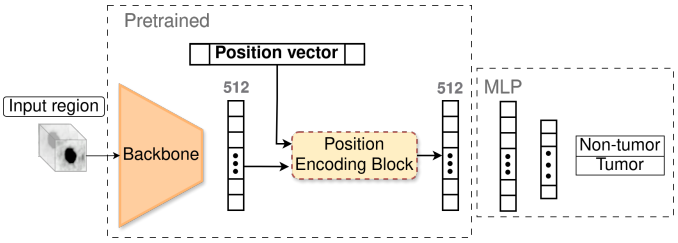


Fig. 2: Overview of the downstream task for ROI classification. Every input region passes through our pre-trained model, and a simple multi-layer perception in order to be classified into a tumor or nontumor class.

task is presented. The ROI is inserted into the backbone to extract the initial semantic features. Concurrently, the position vector, including the ROI's centroid, size, and rotation details, is computed. Both vectors pass through PEB to output the final representation. An MLP classifier is then trained for ROI classification. Please note that any other classifier or task could also be used to evaluate our pretrained model. During inference, labels are not required, and the position vector for any region or patch can be computed at minimal cost.

F. Implementation details

The proposed framework is implemented using the PyTorch library [44] on an NVIDIA GeForce RTX 3060 GPU with 12GB memory. For the representation learning, the encoders in the two branches of our model are shared using the same ResNet18 [19] architecture and the same weights parameters. The projection head that maps the representations to the embeddings where the loss function is calculated, includes two fully connected layers with $4 \times 512 = 2048$ neurons while batch normalization and ReLU activation function are included only in the 1st hidden layer [14]. For training, a batch size of 128 was used for a maximum of 2000 epochs, with early stopping applied after 600 epochs of no improvement. The network was optimized using LARS [45] optimizer with an initial learning rate of 0.5, weight decay 1e-6 and momentum 0.9 to enable the model's training with this batch size. A cosine warmup scheduler with 60 warmup epochs was also used to enhance training. For the implementation of the representation learning schemes, Lightly library [46] was adapted.

For the classification scheme, the classification head consists of 2 linear layers (with size 128 and 2) followed by ReLU and Softmax activation functions, respectively. We employed a Focal loss function with $\alpha = 0.55$ and $\gamma = 2.0$ to enhance the training of difficult samples. The model was trained with an AdamW optimizer [47], learning rate 1e-4 and early stopping with patience of 200 epochs. The former data augmentations were applied to alleviate the highly imbalanced dataset. The images were standardized to have a mean value equal to zero and a standard deviation equal to one to facilitate training. During the downstream classification task for each region, one bounding box centered on the region's centroid and with a size of 16 in each dimension was created, leading to a final box size of 16x16x16. Then, these boxes are fed into the network to learn the representations and classify them as tumor lesions or not.

For the external validation in the private dataset, the slice thickness is resampled to $3mm$, and the images were cropped at the approximate height of thighs to match the imaging of the autoPET dataset. We split the autoPET dataset in train/validation/test to keep the unseen test set for evaluation. The pretraining of the representation learning method is done on the training/validation set of the autoPET which includes a larger number of patients. Afterwards, the classification MLP was trained on the public training set or the private training set respectively, and evaluated on the test sets.

IV. Experiments and Results

A. Evaluation scheme and metrics

Two validation schemes were employed to assess the classification performance of the proposed method, tailored to each dataset's inherent properties. The public dataset included more patients and regions, while the private dataset with fewer patients was used to evaluate the model's generalizability. For the classification evaluation, F1-score, sensitivity, precision, specificity, B.Acc. (simple accuracy was not discriminative due to the imbalanced dataset) and AUC will be provided. Statistical significance was validated by applying the Wilcoxon signed-rank test (with $\alpha = 0.05$) for the pair of models that achieved the best metrics in each experiment (the best and second best models are denoted with bold font and underline respectively, while * symbol in tables denotes $p - value < 0.05$).

B. Results

1) Comparison of state-of-the-art image-based representation techniques with and without position vector: Some of the most representative image-based representation techniques will be compared in Table I. The following methods were used to present a comprehensive comparison of representation methods as they include a contrastive learning scheme (SimCLR [48]), clustering-based methods (SWAV [49], SMOG [50]), methods based on memory banks (MOCO [51], BYOL [52]), the well established DINO method [53], a supervised variation of SSL [54] and the baseline VICReg [14] scheme.

Multiple metrics are presented to provide insights for the validation in the highly imbalanced dataset of regions while the semantic sampling strategy, training parameters and the backbone, ResNet18, were kept constant. Each pair of the table shows the original scheme with the sampling module, and the same scheme with the addition of the proposed positional module. From these comparisons, we can conclude that position enhancement improved the classification accuracy in terms of F1-score and most other metrics. The largest improvements were observed in SWAV, MOCO, and VICReg, where F1-score increased by 2–5%, while smaller gains of 1–1.5% were noted in SIMCLR, Supervised, and DINO with the addition of position. Finally, SMOG and BYOL metrics indicated a very small increase between baseline and position-enhanced variation. In Table I green arrow indicates an increase in the metric with the addition of PEB while the red a decrease. All models showed

statistically significant differences between their original and position-enhanced versions, except for MOCO. The three conclusive metrics, F1-score, B.Acc. and AUC increased with PEB in all models except for a slight reduction in BYOL's B.Acc. We have to note that in this setting the VICReg+pos outperformed the supervised contrastive learning in all metrics except specificity and precision which will be discussed in the following paragraph. Taking into consideration the best classification results among all the representation learning schemes, VICReg with the PEB achieved the best results with an F1-score of 75.34% and also a high sensitivity (sens) of 80.58% together with a precision (prec) of 70.74%, B.Acc. of 87.75% and AUC 96.29%.

TABLE I: Comparison of representation learning methods with and without the PEB.

Method	Sens (%)	Spec (%)	Prec (%)	F1 (%)	B.Acc (%)	AUC (%)
Supervised [54]	72.70	**95.63**	**71.76**	72.23	84.17	93.89
Supervised+pos	80.58↑	93.95↓	67.03↓	73.18↑	87.27↑	96.06↑*
SIMCLR [48]	76.90	95.11	70.60	73.62	86.01	93.74
SIMCLR+pos	**82.41**↑	94.15↓	68.26↓	74.67↑	**88.28**↑	95.37↑*
SWAV [49]	76.12	94.15	66.51	70.99	85.13	93.79
SWAV+pos	76.38↑	95.27↑	71.15↑	73.67↑	85.83↑	95.75↑*
SMOG [50]	72.97	95.43	70.92	71.93	84.20	94.27
SMOG+pos	78.48↑	94.19↓	67.34↓	72.48↑	86.34↑	95.40↑*
DINO [53]	74.28	95.11	69.88	72.01	84.70	93.96
DINO+pos	80.05↑	94.23↓	67.93↓	73.49↑	87.14↑	95.47↑*
BYOL [52]	71.13	94.55	66.58	68.78	82.84	93.08
BYOL+pos	69.29↓	95.23↑	68.93↑	69.11↑	82.26↓	94.58↑*
MOCO [51]	69.82	95.51	70.37	70.09	82.67	93.83
MOCO+pos	76.12↑	95.03↓	70.05↓	72.96↑	85.57↑	94.57↑
VICReg [14]	70.08	95.59	70.82	70.45	82.84	91.64
VICReg+pos	80.58↑	94.91↓	70.74↓	**75.34**↑	87.75↑	**96.29**↑*

pos: proposed scheme with Position Encoding Block.
*: Statistical significant differences

In Table II, supervised contrastive learning and its PEB-enhanced variant are compared with the proposed scheme using a decision threshold selected from a grid search over values from 0.3 to 0.9 in increments of 0.05, ensuring comparable precision and specificity with the supervised method. This approach enables a more accurate assessment of the remaining metrics. The proposed VICReg scheme outperformed supervised contrastive learning with our PEB by 0.83% and standard supervised contrastive learning by a larger margin of 3.60% in the F1-score. The proposed scheme achieved the highest values across all metrics, demonstrating its superiority.

TABLE II: Comparison between supervised contrastive and the proposed scheme.

Method	Th	Sens (%)	Spec (%)	Prec (%)	F1 (%)	B.Acc (%)
Supervised [54]	0.50	72.70	95.63	71.76	72.23	84.17
Supervised+pos	0.65	74.02	96.44	76.01	75.00	85.23
VICReg+pos	0.60	**75.33***	**96.44**	**76.33***	**75.83***	**85.88***

pos: proposed scheme with Position Encoding Block. Th: Threshold
*: Statistical significant differences

In Table III the proposed method was compared with Volume Contrast (VoCo) which is a recent contrastive learning method that utilizes position information in pretraining for 3D medical images. In these results, we can observe that even though VoCo aims to learn representations by predicting the sub-volumes' position, in our dataset its performance in 4 out of 6 metrics, e.g. F1-score 66.29%, was lower than our proposed. The supervised version of VoCo achieved a 2% increase in the F1-score but it could not reach the proposed 75.34%. In VoCo, supervision was a segmentation-based approach, a fact that could limit its contribution to the classification task. For comparison with the weakly supervised positional 2D (WSP-contr) method [36], we employed the ResNet18 backbone from the study, adjusting only the temperature parameter after experimentation to better suit our dataset. Additionally, the 2D classification results were transformed in the same 3D region set using an overlap threshold of 0.5 to declare an identified region. The observed decrease in performance can be attributed to the limitations of the 2D slice position that cannot model all 3D spatial information. Furthermore, 2D slice-based modeling risks overlooking small tumor regions or erroneously classifying large non-tumorous areas as tumors.

TABLE III: Comparison on related representation learning techniques with positional encoding.

Method	Sens (%)	Spec (%)	Prec (%)	F1 (%)	B.Acc (%)	AUC (%)
VoCo [40]	61.52	96.21	71.43	66.10	78.86	88.09
VoCo supervised	59.95	**97.54***	**78.97***	68.15	78.74	86.31
WSP-Contr [36]	71.35	92.26	57.48	63.67	81.81	87.99
Proposed	**80.58***	94.91	70.74	**75.34***	**87.75***	**96.29**

*: Statistical significant differences

2) Quantitative and Qualitative comparison with state-of-the-art on the downstream task: In this section, we present experiments comparing classification performance on both the autoPET dataset and the private dataset. For the classification of regions from PET/CT images for malignant melanoma a CNN variation (CNN Var) with two convolutional blocks followed by two linear layers was used in [27]. Furthermore, a related variant (CNN Var2) designed for general FDG-PET scans [22] was utilized with the number of filters (32,64,128,265). A CNN based on the AlexNet architecture has also previously been utilized for PET/CT classification in [10]. For our experiments, we utilized fewer layers because our regions were of size 16, which means that we can downsample fewer times, leading to an architecture with 5 convolutional layers where the first two and the last one include Maxpool layers. A DenseNet variant [23] previously used for classifying uptake in PSMA PET/CT images was also compared in this study. The preprocessing steps were the same for all the compared models to provide a fair comparison.

In Table IV, the previously mentioned techniques for PET images, a supervised ResNet18 model and the proposed method were compared in classifying regions on the autoPET dataset. The CNN Var [27], CNN Var2 [22] and the AlexNet [55] variations achieved very similar results with an F1-score of 72.43%, 73.28% and 73.06% in the classification task. The DenseNet Var while achieving relatively high Precision, it underperformed in the other metrics due to failing to identify many true tumors. In terms of F1-score, B.Acc., and AUC, the proposed model outperformed all the others by achieving 75.34%, 87.75%, and 96.29% while also indicating the best

sensitivity by identifying 80.58% of the true tumor lesions. The proposed framework was superior against the CNN Vars and the AlexNet Var, showing a 2.06%-2.91% increase in F1-score, and the supervised ResNet18 classification with approximately a 4% increase. Statistical significance was confirmed for the best performing models, where the proposed model achieved higher value in 4 out of 6 metrics.

TABLE IV: Comparison with classification methods in the autoPET dataset

Method	Sens (%)	Spec (%)	Prec (%)	F1 (%)	B.Acc (%)	AUC (%)
CNN Var [27]	75.85	94.87	69.30	72.43	85.36	91.38
CNN Var2 [22]	66.93	**97.60***	**80.95***	73.28	82.26	93.01
AlexNet Var [55]	77.95	94.59	68.75	73.06	86.27	94.83
DenseNet Var [23]	65.09	96.52	74.03	69.27	80.80	89.94
ResNet18	75.33	94.51	67.69	71.30	84.92	93.09
Proposed	**80.58***	94.91	70.74	**75.34***	**87.75***	**96.29***

*: Statistical significant differences

In Table V, the same experimental comparisons were implemented for the private dataset in order to provide further insights into the proposed model's generalizability and the effectiveness of the generated representations on similar datasets with different acquisition parameters. For the proposed methods, Proposed+Ft and Proposed w/o Ft, the weights from the training on the autoPET dataset were loaded to test their applicability to the second dataset and fine-tuning was applied to the private training set in the 1st case. AlexNet variation, DenseNet and ResNet18 models achieved similar lower metrics with an F1-score of 67%, B.Acc. of 79%-80%, and AUC of 89%-91% approximately. The CNN variation achieved an approximate F1-score of 69.5%, B.Acc. of 82%, and AUC of 90.96%, probably due to the less complex model, which was trained efficiently in the smaller dataset. Better performance was indicated by the CNN Var2 with an F1-score of 71.23% and AUC of 93.55%. As an external validation, the proposed scheme only pre-trained on the autoPET set (referred to as "Proposed w/o Ft") was evaluated on the private test set. In this regard, both the pre-trained and finetuned proposed method enhanced the classification performance, outperforming the other methods in most metrics. The proposed method with finetune indicated similar to the autoPET experiments metrics with F1-score, B.Acc., and AUC of 75.46%, 86.58%, and 93.80%, respectively. The proposed pre-trained only on the autoPET dataset without finetuning also achieved high evaluation metrics, suggesting that our model generalizes well with out-of-distribution samples.

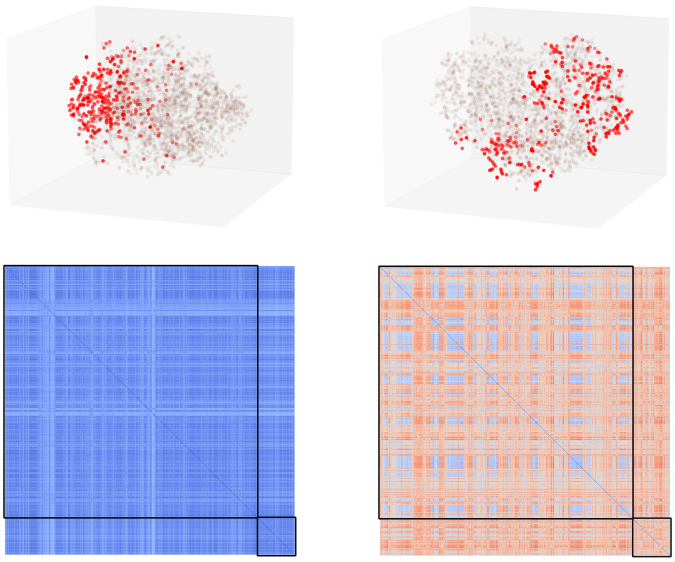
Fig. 3 presents an example where it can be observed that the proposed method identified most tumor lesions while also suppressing some of the challenging false positive regions that the other identified as tumors.

3) Interpretation of the learned representations: In this section, visualization of the learned representations is conducted through t-SNE [56] which can visualize the high dimensional vectors to the 3D/2D space by preserving the structure of the high dimensions. The representations of the same class should be close to each other and separated from the representations of the opposite class. In Fig. 4, the representations of ROIs from the test set are presented. Regions of class tumors and non-tumor regions are presented with red and grey color, respectively. In the 1st row, the t-SNE representations of both baseline VICReg and the proposed indicate separation between the classes, with one or two regions of concentrated tumor representations, while there is also a region where tumors and non-tumors are located very close to each other. The last are the regions that are the most difficult to classify.

TABLE V: Comparison with classification methods in the private dataset

Method	Sens (%)	Spec (%)	Prec (%)	F1 (%)	B.Acc (%)	AUC (%)
CNN Var [27]	69.95	94.63	69.19	69.57	82.29	90.96
CNN Var2 [22]	71.04	95.10	71.43	71.23	83.07	93.55
AlexNet Var [55]	62.84	**96.05***	**73.25***	67.65	79.44	89.02
DenseNet Var [23]	66.12	95.01	69.54	67.79	80.56	90.34
ResNet18	64.48	95.30	70.24	67.24	79.89	91.34
Proposed + Ft	78.14	95.01	72.96	**75.46**	86.58	93.80
Proposed w/o Ft	**83.61**	91.81	63.75	72.34	**87.71**	**94.37**

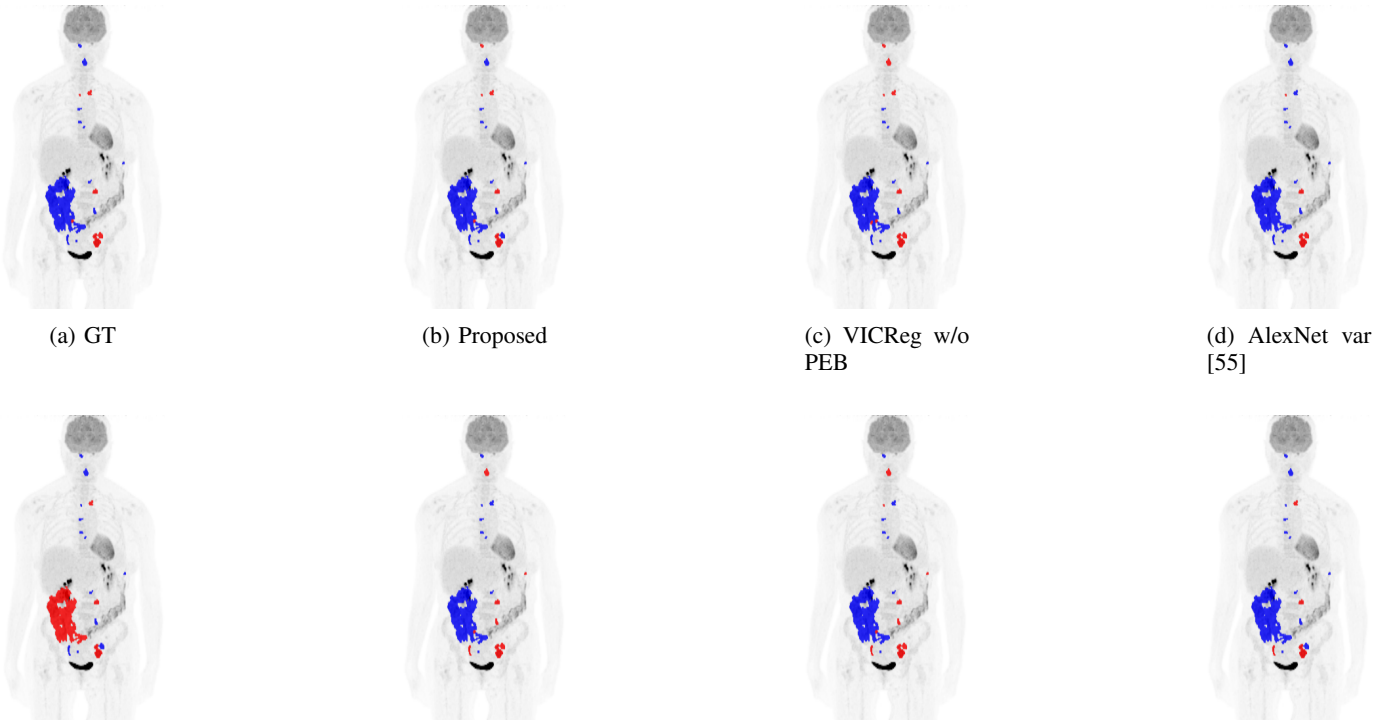
Proposed w/o Ft: only pre-trained on the autoPET. Ft:Finetune
*: Statistical significant differences



(a) VICReg (b) Proposed

Fig. 4: t-SNE representations and distance heatmap for the VICReg and the proposed scheme. Red and grey color denotes the tumor and non-tumor regions, respectively.

In the second row, a heatmap with the distances between each vector pair, sorted by class, is presented. The vectors of class non-tumor are shown first, and the tumors are shown in the last. In this figure, regions of the same class inside the boxes should have a small distance (closer to the blue color) and regions of the opposite class a large distance (red). The baseline VICReg presents a relatively smaller distance inside the tumor class on the right bottom square, but there is not a clear separation from the regions of the class non-tumor. However, in the position-enhanced proposed scheme, we can observe spatially closer regions being formed and on the right bottom, there is a visually observable block of regions with a lower distance between them and a higher distance from the other representations.



(a) GT (b) Proposed (c) VICReg w/o PEB (d) AlexNet var [55]

(e) DenseNet var [55] (f) CNN var [27] (g) CNN var2 [22] (h) ResNet18

Fig. 3: Classification comparison example. Blue and red regions indicate non-tumor and tumor regions, respectively. The proposed scheme indicates overall better sensitivity while also suppressing false positive identification.

C. Ablation study

1) Ablation study on backbone selection and positional components: The choice of the backbone can significantly affect the results because it is responsible for the extraction of the semantically discriminative features from the initial ROIs. We experimented with some of the most common backbones for image classification, ResNet variants (ResNet10, ResNet18, ResNet50) [19], Squeeze-and-Excitation Networks (SENet) [57], EfficientNet B0H2 [58] and Vision Transformer (ViT) [59]. In Table VI, the classification results in the autoPET dataset are presented for each one of the tested backbones. We can observe that ResNet18 achieved the best F1-score (75.34%), which can summarize the performance due to our imbalanced dataset and, concurrently, large sensitivity, which is important in clinical applications. The ResNet18, which is a simpler structure than ResNet50 and ViT, achieved the best results in our case. This is likely because we feed the network 3D ROIs with dimensions of 16 voxels, which exhibit less complex appearances, allowing ResNet18 to efficiently extract information and avoid overfitting.

Table VII summarizes the metrics from the ablation study on the position components. The experiments included the VICReg baseline without position, the inclusion of solely the region's centroid, the combination of centroid and size and the proposed position vector. We can observe that the insertion of the centroid improved the results except for the precision, while the addition of size produced slight improvements. Finally, the complete position vector indicated the best values in four out of six metrics with statistical significance confirmed, while it showed similar values in the rest.

TABLE VI: Comparison of the backbone networks.

Method	Sens (%)	Spec (%)	Prec (%)	F1 (%)	B.Acc (%)	AUC (%)
ResNet10	74.28	95.59	72.01	73.13	84.94	96.06
ResNet18	80.58	94.91	70.74	**75.34***	**87.75***	**96.29***
ResNet50	75.07	**95.96***	**73.90***	74.48	85.51	94.49
SENet	73.49	95.31	70.53	71.98	84.40	94.33
EfficientNet	77.95	95.03	70.55	74.06	86.49	95.56
ViT	**80.84***	92.51	62.22	70.32	86.68	93.91

*: Statistical significant differences

TABLE VII: Ablation study on the position vector's components.

Method	Sens (%)	Spec (%)	Prec (%)	F1 (%)	B.Acc (%)	AUC (%)
VICReg	70.08	**95.59***	**70.82***	70.45	82.84	91.64
Cent	79.79	94.01	67.56	73.16	86.90	94.71
Cent+Size	80.05	94.26	68.54	73.85	87.16	95.42
Cent+Size+Rot	**80.58***	94.91	70.74	**75.34***	**87.75***	**96.29***

*: Statistical significant differences

2) Impact of the loss weights on the training: In Table VIII different values have been tested for the weights in the loss function. These values regulate the weight of the invariance, variance and covariance term of the loss function in the training procedure. The experimental values were selected in relevance with the initial VICReg [14]. In the 1st and 2nd rows, the requirement for both variance and covariance terms is supported. We can also observe that equal weights in the

former two terms achieved better results. The best results were achieved for the values of $\lambda = \mu = 25$, $\nu = 1$ where the model's performance was superior in all metrics.

TABLE VIII: Ablation study on the loss weights λ, μ, ν.

λ	μ	ν	Sens (%)	Spec (%)	Prec (%)	F1 (%)	B.Acc (%)	AUC (%)
25	0	1	68.24	92.79	59.09	63.34	80.52	84.84
25	25	0	76.64	91.83	58.87	66.59	84.24	93.29
10	10	1	76.12	93.35	63.60	69.30	84.73	94.65
50	50	1	80.58	93.63	65.88	72.49	87.10	95.74
25	10	1	78.48	93.91	66.30	71.88	86.20	94.94
10	25	1	78.22	92.63	61.83	69.06	85.42	94.48
25	25	1	**80.58***	**94.91***	**70.74***	**75.34***	**87.75***	**96.29***

*: Statistical significant differences

3) Evaluating the impact of the PEB's location and the sampling module: In Table IX an ablation study concerning the usage of the position vector in the framework is presented. In the first row, the position vector is introduced only in the classification stage while the pretraining stage is performed without it. Although adding the PEB enhances the classification accuracy, the results indicate that the VICReg representation learning with the PEB leads to better training and better predictive features. The second row shows that employing the PEB solely during representation learning but removing it during classification training, results in lower performance. In the third row, a fully supervised training of the architecture with the position vector is presented. The training performance was inferior compared to the implementation of semantic sampling with the PEB, due to the poor training of the combined ResNet18 with the PEB in a fully supervised manner. Training the model in two steps, where the representation extraction is learned from comparing class-related pairs including their positional details and then adding the small MLP module on top of it for the classification task was superior.

TABLE IX: Ablation study on the PEB block location.

Method	Sens (%)	Spec (%)	Prec (%)	F1 (%)	B.Acc (%)	AUC (%)
PEB only-on-clf[1]	77.43	94.27	67.35	72.04	85.85	94.72
PEB only-on-rep[2]	73.23	94.75	68.05	70.54	83.99	93.33
PEB supervised	76.90	93.03	62.74	69.10	84.97	94.30
Proposed	**80.58***	**94.91***	**70.74***	**75.34***	**87.75***	**96.29***

*: Statistical significant differences
[1]: only-on-clf: PEB only in the classification stage
[2]: only-on-rep: PEB only in the representation learning

In Table X an ablation study for the sampling module and the PEB is provided. When utilizing either the sampling module (row 2) or the PEB (row 3) individually, the model's performance improves compared to the baseline VICReg (row 1), achieving F1-scores of 70.45% and 71.73%, respectively, versus the baseline F1 of 66.51%. A comparison of AUC values using DeLong's test between VICReg with and without sampling did not show statistically significant differences (p = 0.1509). Finally, the combination of both modules (row 4) demonstrates its superiority by achieving the highest values in most metrics including the most indicative ones: F1, B.Acc. and AUC. Our statistical tests indicate that our proposed PEB module has a higher impact on the performance than the sampling module however, their combination reports the best performance over four different metrics and the second-best in the other two.

TABLE X: Ablation on the semantic and positional modules.

Method	S	Pos	Sens (%)	Spec (%)	Prec (%)	F1 (%)	B.Acc (%)	AUC (%)
VICReg	–	–	74.02	92.59	60.39	66.51	83.30	92.89
VICReg	✓	–	70.08	**95.59***	**70.82***	70.45	82.84	91.64
VICReg	–	✓	76.90	94.27	67.20	71.73	85.59	95.48
Proposed	✓	✓	**80.58**	94.91	70.74	**75.34**	**87.75**	**96.29**

*: Statistical significant differences
S: Sampling module

4) Evaluating multimodal PET/CT approaches for Representation Learning: Table XI presents an investigation into whether PET, CT, or their combination in multimodal approaches yields stronger representations. The early fusion, where PET and CT were inserted as two channels in the model, slightly increased sensitivity and B.Acc. but resulted in many false positives, lowering precision, probably due to CT where tumors in some areas are not easily distinguishable. The late fusion approach, where PET and CT passed through two backbones and their output was concatenated with the PEB to produce the final vector, performed even worse. The single PET modality was superior in the three conclusive metrics while having less computational burden.

TABLE XI: Ablation study on the modalities PET and CT.

Method	Sens (%)	Spec (%)	Prec (%)	F1 (%)	B.Acc (%)	AUC (%)
PET	80.58	94.91	**70.74***	**75.34***	87.75	**96.29***
PET+CT early fusion	**82.94***	93.59	66.39	73.75	**88.27***	95.71
PET+CT late fusion	74.54	**95.11***	69.95	72.17	84.83	95.59

*: Statistical significant differences

5) Influence of the training set size on model performance: In Table XII, the performance of the proposed method was compared against the supervised ResNet18 for different training set sizes. Each row shows the metrics achieved for 15%, 30%, 50% and 100% of the training set. It can be observed that the proposed method trained with 50% of the initial data achieved very high metrics, F1-score of 72.84%, which is less than a 3% difference from the full dataset and AUC values of approximately 95% while the supervised ResNet18 dropped at F1-score of 68% and AUC of 91.59%. In the 30% of the dataset, the proposed model outperformed the fully supervised in 2 out of 3 metrics while it also preserved these values for the 15% of the dataset. Statistical significance was confirmed for the paired comparisons of each model for each data percentage. These results support the model's efficiency in settings with limited training samples where it can learn better representations.

V. Discussion

In this work, we developed a representation learning method incorporating semantic and position information to produce representations capable of distinguishing true tumor lesions from non-tumor high-uptake regions from FDG-PET images focused on MM patients. A class-specific sampling strategy is

TABLE XII: Comparison results for different training set sizes.

	Supervised ResNet18			Proposed		
Method	**F1(%)**	**B.Acc(%)**	**AUC(%)**	**F1(%)**	**B.Acc(%)**	**AUC(%)**
15%	63.50	81.79	89.92	69.20	**88.26**	95.10*
30%	68.42	81.74	92.49	68.25	88.18	95.17*
50%	68.03	80.73	91.59	72.84	86.03	95.51*
100%	71.30	84.92	93.09	**75.34**	87.75	**96.29***

*: Statistical significant differences

developed to compare ROIs with the same semantic meaning and extract discriminative features. The semantic representation vector extracted by the backbone was then effectively fused through the PEB with the position information of each region inside the patient's whole-body image. A VICReg-based scheme is used to pretrain the ResNet18 backbone and the PEB by utilizing both augmented views and pairs of ROIs from the same class. This approach generates rich representations by preventing collapse to trivial solutions during training and by decorrelating feature dimensions.

MM can give multiple metastases across the whole body, making the manual segmentation of the tumor lesions a time-consuming task, leading to crucial delays in the clinical workflow. In this regard, DL-based segmentation methods for the automatic delineation of tumors have recently been proposed. A major concern with these methods is the high number of false positives, where regions with high uptake are not tumor-related and may lead to assessment errors. The proposed method was evaluated on a dataset containing high-uptake regions from both tumor lesions and non-tumor-related uptake. The reported metrics indicate high accuracy in distinguishing these regions, outperforming existing classification systems used for FDG-PET images. Most previous methods relied on large annotated datasets for supervision, but obtaining such data is time-consuming and impractical for clinical workflows. In this regard, the proposed method can effectively be trained by employing the semantic aware sampling strategy which can generate pairs of ROIs with the same diagnostic meaning and the enhanced VICReg based training scheme. Experimental results (Table XII) have shown that the model can distinguish the ROIs with an F1-score of 69.20% and Balanced Accuracy of 88.26% even with only 15% of the training dataset while outperforming the fully supervised ResNet18 and the compared classification methods when using 100% of the dataset. The experimental results on two datasets (Tables IV and V) support the model's effectiveness and its generalization capabilities with significant implications in accelerating and enhancing the clinical workflow.

Although representation learning schemes, such as SSL, have been extensively studied to alleviate the requirement for large annotated databases for natural images, they are less examined in the medical domain due to many challenges. One limitation is the difficulty of selecting pairs for contrastive learning in medical images where the datasets are imbalanced or include a small number of patients. To this end, heavy augmentations have been applied to support the pretext task of comparing ROIs, but in medical images, they can alter the anatomical and semantic characteristics. Selecting a suitable pretext task for pretraining is challenging, and the effectiveness of SSL depends on its alignment with the target application. Furthermore, ROIs with different clinical information can present very subtle differences in the grayscale intensities. These challenges have been mitigated in this work firstly by introducing semantic aware sampling where pairs of regions with the same diagnostic class are used to enable the network to learn class-aware discriminative representations. This sampling strategy also alleviated the negative impact of the imbalanced dataset by generating more tumor lesion pairs than would be possible using only the limited number of tumorous ROIs. Finally, considering the small variations in the intensity and appearance of the ROIs from PET images led us to utilize the VICReg scheme, which enables the representation learning scheme to avoid collapse to trivial solutions and to decorrelate features leading to strong representations. The impact of the proposed class aware sampling module was supported by the experiment in Table X where it substantially increased the F1-score to 70.45% and precision to 70.82% against the usage of the augmentations. Moreover, its combination with the PEB further enhanced overall performance.

Representation learning schemes usually rely on utilizing large batch sizes, which cannot be applied to large 3D whole-body medical images due to restrictions in computational resources and GPU memory. The proposed method utilizes smaller high-uptake ROIs, enabling efficient network training with large batch sizes. Using only ROIs without additional information leads to loss of spatial information which is important for the classification task. A comparative analysis of the baseline model (without PEB) and the proposed model with PEB demonstrates its minimal computational overhead. The GPU memory usage increased by approximately 0.9%, training time per epoch rose by about 0.54%, inference time per patient remained unchanged at 0.09 seconds, and the number of parameters grew from 38,409,024 to 38,938,953 (an increase of 1.38%).

In medical images, apart from the intensity variations, the location inside the patient's body, their size, and their general position are crucial details for diagnosis. Some regions present in general high-uptake but it doesn't correspond to tumors. Positional details are critical for diagnosis, as they account for variability in the shape and texture of ROIs across anatomical regions. For example, physiological uptake in tissues of the liver, and lymphoid areas complicates the identification of malignant lesions, especially in high-metabolism regions [60]. In this direction and considering our previous work [18], which indicated that for regions with different sizes, different features could lead to better diagnostic accuracy, we inserted the position vector. The position vector consists of three components: the region's centroid, size, and rotation, which together provide detailed spatial context. This design is validated by the ablation study in Table VII, where the PEB increased the F1-score by 4.89% and AUC by 4.65%, as well as by the PEB location ablation results in Table IX. In this regard, the proposed PEB's design and implementation can combine the semantic and position vectors through learnable parameters and projections to produce a representation vector where different semantic features are affected by the position vector. The combination of semantic information with the position information achieved

superior classification accuracy in most of the representation learning methods, compared in Table I, supporting the additive knowledge of the PEB.

Although the proposed framework presented results indicating its superiority, there are some limitations to be addressed. First, the input to the neural network was selected to be of size (16,16,16) to incorporate adequate information and reduce noise for both large and smaller ROIs. The region's size is held constant because this study focused on the ROI sampling strategy and the PEB. In addition to that, even though the proposed framework works well with limited annotated data, it requires the availability of at least a small portion of them for the semantic aware sampling. This was an assumption to enhance the representation with discriminative features for the task of classification and alleviate the restrictions opposed by the medical images. Thirdly, in order to study the position vector's adaptability to a second dataset, we conducted experiments where the backbone was pretrained in the public dataset and used for the classification of the second dataset. The origin point of the images should be approximately in the same region, and the spacing should be similar for both datasets. We mitigated this constraint by automatically cropping and resampling the ROIs of the private dataset, which led to very promising results. A future direction could be to further adapt the position vector conditioned on the input images by automatically locating reference points, such as liver center or learnable points. Finally, the representations of suspicious regions from our proposed method could be combined to produce representations of the patients that will describe their current medical condition or stage. These patient representations may then assist in automatic treatment assessment and monitoring of disease progression between follow-up scans, ultimately promoting personalized treatment.

VI. Conclusion

Representation learning is commonly utilized for its ability to extract task-specific features; yet, its application in 3D medical imaging is constrained by high computational costs and subtle differences in image appearance. Detecting tumor lesions in 3D FDG-PET images of MM patients requires either manual delineation by experts or leveraging large annotated databases to train deep neural networks with many limitations such as the resulting false positive regions. The proposed framework can mitigate these requirements by efficiently handling fewer annotated images through a representation learning scheme based on VICReg. This framework combines the effective region sampling based on class-related semantic details with the PEB to generate rich representations that integrate semantic and positional information. This integration is essential for accurately classifying high-uptake regions in FDG-PET images as tumor lesions or non-tumor regions. Extensive evaluation on both a public and a private dataset supports the proposed methods' applicability and high accuracy.

References

[1] M. Ralli *et al.*, "Immunotherapy in the treatment of metastatic melanoma: Current knowledge and future directions," *Journal of immunology research*, vol. 2020, 2020.

[2] C. Sachpekidis, V. Weru, A. Kopp-Schneider, J. C. Hassel, and A. Dimitrakopoulou-Strauss, "The prognostic value of [18f]fdg pet/ct based response monitoring in metastatic melanoma patients undergoing immunotherapy: comparison of different metabolic criteria," *European journal of nuclear medicine and molecular imaging*, 2023.

[3] A. Guerrisi *et al.*, "Novel cancer therapies for advanced cutaneous melanoma: The added value of radiomics in the decision making process–a systematic review," *Cancer Medicine*, vol. 9, p. 1603, 3 2020.

[4] J. van Sluis, E. C. de Heer, M. Boellaard, M. Jalving, A. H. Brouwers, and R. Boellaard, "Clinically feasible semi-automatic workflows for measuring metabolically active tumour volume in metastatic melanoma," *European journal of nuclear medicine and molecular imaging*, vol. 48, pp. 1498–1510, 5 2021.

[5] H. Jiang *et al.*, "A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation," *Computers in Biology and Medicine*, vol. 157, p. 106726, 5 2023.

[6] S. Gatidis *et al.*, "A whole-body fdg-pet/ct dataset with manually annotated tumor lesions," *Scientific Data*, vol. 9, 12 2022.

[7] T. Shi, H. Jiang, M. Wang, Z. Diao, G. Zhang, and Y.-D. Yao, "Metabolic anomaly appearance aware u-net for automatic lymphoma segmentation in whole-body pet/ct scans," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 5, pp. 2465–2476, 2023.

[8] E. Yazdani *et al.*, "Automated segmentation of lesions and organs at risk on [68ga] ga-psma-11 pet/ct images using self-supervised learning with swin unetr," *Cancer Imaging*, vol. 24, no. 1, p. 30, 2024.

[9] I. Dirks, M. Keyaerts, I. Dirven, B. Neyns, and J. Vandemeulebroucke, "Development and validation of a predictive model for metastatic melanoma patients treated with pembrolizumab based on automated analysis of whole-body [18f]fdg pet/ct imaging and clinical features," *Cancers*, vol. 15, 8 2023.

[10] L. Sibille *et al.*, "18f-fdg pet/ct uptake classification in lymphoma and lung cancer by using deep convolutional neural networks," *Radiology*, vol. 294, pp. 445–452, 2020.

[11] J. Zhang, Y. Wang, J. Liu, Z. Tang, and Z. Wang, "Multiple organ-specific cancers classification from pet/ct images using deep learning," *Multimedia Tools and Applications*, vol. 81, pp. 16 133–16 154, 5 2022.

[12] S. C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung, and A. S. Chaudhari, "Self-supervised learning for medical image classification: a systematic review and implementation guidelines," *npj Digital Medicine 2023 6:1*, vol. 6, pp. 1–16, 4 2023.

[13] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies 2021, Vol. 9, Page 2*, vol. 9, p. 2, 12 2020.

[14] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," in *ICLR*, 2022.

[15] C. Zhang, H. Zheng, and Y. Gu, "Dive into the details of self-supervised learning for medical image analysis," *Medical Image Analysis*, vol. 89, p. 102879, 10 2023.

[16] S. C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung, and A. S. Chaudhari, "Self-supervised learning for medical image classification: a systematic review and implementation guidelines," *npj Digital Medicine 2023 6:1*, vol. 6, pp. 1–16, 4 2023.

[17] C. Sachpekidis, L. Pan, A. Kopp-Schneider, V. Weru, J. C. Hassel, and A. Dimitrakopoulou-Strauss, "Application of the long axial field-of-view pet/ct with low-dose [18f]fdg in melanoma," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 50, pp. 1158–1167, 3 2023.

[18] T. P. Vagenas *et al.*, "A decision support system for the identification of metastases of metastatic melanoma using whole-body fdg pet/ct images," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, pp. 1397–1408, 3 2023.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] K. Kawauchi *et al.*, "A convolutional neural network-based system to classify patients using fdg pet/ct examinations," *BMC Cancer*, vol. 20, pp. 1–10, 3 2020.

[21] D. Wallis, M. Soussan, M. Lacroix, P. Akl, C. Duboucher, and I. Buvat, "An [18f]fdg-pet/ct deep learning method for fully automated detection of pathological mediastinal lymph nodes in lung cancer patients," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 49, p. 881, 2 2022.

[22] A. Berenbaum *et al.*, "Performance of ai-based automated classifications of whole-body fdg pet in clinical practice: The clariti project," *Applied Sciences 2023, Vol. 13, Page 5281*, vol. 13, p. 5281, 4 2023.

[23] Y. Li *et al.*, "An Automated Deep Learning-Based Framework for Uptake Segmentation and Classification on PSMA PET/CT Imaging

of Patients with Prostate Cancer," *Journal of Imaging Informatics in Medicine*, Apr. 2024.

[24] D. Nishigaki *et al.*, "Vision transformer to differentiate between benign and malignant slices in 18f-fdg pet/ct," *Scientific Reports 2024 14:1*, vol. 14, pp. 1–11, 4 2024.

[25] T. Küstner *et al.*, "Development of a hybrid-imaging-based prognostic index for metastasized-melanoma patients in whole-body 18f-fdg pet/ct and pet/mri data," *Diagnostics (Basel, Switzerland)*, vol. 12, 9 2022.

[26] H. S. Gabryś *et al.*, "Pet/ct radiomics for prediction of hyperprogression in metastatic melanoma patients treated with immune checkpoint inhibitors," *Frontiers in Oncology*, vol. 12, p. 977822, 11 2022.

[27] I. Dirks, M. Keyaerts, B. Neyns, and J. Vandemeulebroucke, "Computer-aided detection and segmentation of malignant melanoma lesions on whole-body 18f-fdg pet/ct using an interpretable deep learning approach," *Computer Methods and Programs in Biomedicine*, vol. 221, p. 106902, 6 2022.

[28] H. Zhao *et al.*, "Anomaly detection for medical images using self-supervised and translation-consistent features," *IEEE Transactions on Medical Imaging*, vol. 40, pp. 3641–3651, 12 2021.

[29] H. Huang, R. Wu, Y. Li, and C. Peng, "Self-supervised transfer learning based on domain adaptation for benign-malignant lung nodule classification on thoracic ct," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, pp. 3860–3871, 8 2022.

[30] K. Yu, L. Sun, J. Chen, M. Reynolds, T. Chaudhary, and K. Batmanghelich, "Drasclr: A self-supervised framework of learning disease-related and anatomy-specific representation for 3d lung ct images," *Medical Image Analysis*, vol. 92, p. 103062, 2 2024.

[31] F. Zhao, D. Li, R. Luo, M. Liu, X. Jiang, and J. Hu, "Self-supervised deep learning for joint 3d low-dose pet/ct image denoising," *Computers in Biology and Medicine*, vol. 165, p. 107391, 10 2023.

[32] Z. Huang *et al.*, "Multi-scale feature similarity-based weakly supervised lymphoma segmentation in pet/ct images," *Computers in biology and medicine*, vol. 151, 12 2022.

[33] Y. Mo *et al.*, "Hover-trans: Anatomy-aware hover-transformer for roi-free breast cancer diagnosis in ultrasound images," *IEEE Transactions on Medical Imaging*, vol. 42, pp. 1696–1706, 6 2023.

[34] M. R. H. Taher, M. B. Gotway, and J. Liang, "Towards foundation models learned from anatomy in medical imaging via self-supervision," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14293 LNCS, pp. 94–104, 2024.

[35] D. Zeng *et al.*, "Positional contrastive learning for volumetric medical image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12902 LNCS, pp. 221–230, 2021.

[36] E. Sarfati, A. Bône, M. M. Rohé, P. Gori, and I. Bloch, "Weakly-supervised positional contrastive learning: Application to cirrhosis classification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14220 LNCS, pp. 227–237, 2023.

[37] Y. Zhang, P. Gu, N. Sapkota, H. Zheng, P. Liang, and D. Z. Chen, "A point in the right direction: Vector prediction for spatially-aware self-supervised volumetric representation learning," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 2023, pp. 1–5.

[38] Z. Chen, Q. Gao, Y. Zhang, and H. Shan, "Ascon: Anatomy-aware supervised contrastive learning framework for low-dose ct denoising," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14229 LNCS, pp. 355–365, 2023.

[39] T. Zhang, D. Wei, M. Zhu, S. Gu, and Y. Zheng, "Self-supervised learning for medical image data with anatomy-oriented imaging planes," *Medical Image Analysis*, vol. 94, p. 103151, 5 2024.

[40] L. Wu, J. Zhuang, and H. Chen, "Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22 873–22 882, 2024.

[41] Y. Tang *et al.*, "Body part regression with self-supervision," *IEEE Transactions on Medical Imaging*, vol. 40, no. 5, pp. 1499–1507, 2021.

[42] C. M. Breki *et al.*, "Fractal and multifractal analysis of PET/CT images of metastatic melanoma before and after treatment with ipilimumab," *EJNMMI Research*, vol. 6, no. 1, 2016.

[43] F. Orlhac, M. Soussan, J. A. Maisonobe, C. A. Garcia, B. Vanderlinden, and I. Buvat, "Tumor texture analysis in 18f-fdg pet: Relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis," *Journal of Nuclear Medicine*, vol. 55, pp. 414–422, 3 2014.

[44] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[45] Y. You, I. Gitman, and B. Ginsburg, "Large batch training of convolutional networks," *arXiv: Computer Vision and Pattern Recognition*, 2017.

[46] I. Susmelj, M. Heller, P. Wirth, J. Prescott, and M. E. et al., "Lightly," *GitHub. Note: https://github.com/lightly-ai/lightly*, 2020.

[47] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2017.

[48] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *37th International Conference on Machine Learning, ICML 2020*, vol. PartF168147-3, pp. 1575–1585, 2 2020.

[49] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.

[50] B. Pang, Y. Zhang, Y. Li, J. Cai, and C. Lu, "Unsupervised visual representation learning by synchronous momentum grouping," in *European Conference on Computer Vision*. Springer, 2022, pp. 265–282.

[51] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 9726–9735, 11 2019.

[52] J.-B. Grill *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[53] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[54] P. Khosla *et al.*, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 18 661–18 673.

[55] L. Sibille *et al.*, "18f-fdg pet/ct uptake classification in lymphoma and lung cancer by using deep convolutional neural networks," *Radiology*, vol. 294, pp. 445–452, 2020.

[56] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[57] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[58] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *ArXiv*, vol. abs/1905.11946, 2019.

[59] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[60] M. D. Vangu and J. I. Momodu, "F-18 fdg pet/ct imaging in normal variants, pitfalls and artifacts in the abdomen and pelvis," *Frontiers in Nuclear Medicine*, vol. 1, p. 826109, 1 2021.