Predicting distant metastases in soft-tissue sarcomas from PET-CT scans using constrained hierarchical multi-modality feature learning

Yige Peng1,2, Lei Bi1,2,∗, Ashnil Kumar1,3, Michael Fulham1,2,4, Dagan Feng1,2,5 and Jinman Kim1,2,∗

1 The School of Computer Science, The University of Sydney, Australia
2 The ARC Training Centre for Innovative BioEngineering, Australia
3 The School of Biomedical Engineering, The University of Sydney, Australia
4 The Department of Molecular Imaging, Royal Prince Alfred Hospital, Australia
5 The Med-X Research Institute, Shanghai Jiao Tong University, People’s Republic of China
∗ Authors to whom any correspondence should be addressed.
E-mail: lei.bi@sydney.edu.au and jinman.kim@sydney.edu.au

Keywords: distant metastases, PET-CT, radiomics, soft-tissue sarcomas, convolutional neural networks

Supplementary material for this article is available online

Abstract

Objective. Positron emission tomography-computed tomography (PET-CT) is regarded as the imaging modality of choice for the management of soft-tissue sarcomas (STSs). Distant metastases (DM) are the leading cause of death in STS patients and early detection is important to effectively manage tumors with surgery, radiotherapy and chemotherapy. In this study, we aim to early detect DM in patients with STS using their PET-CT data.

Approach. We derive a new convolutional neural network method for early DM detection. The novelty of our method is the introduction of a constrained hierarchical multi-modality feature learning approach to integrate functional imaging (PET) features with anatomical imaging (CT) features. In addition, we removed the reliance on manual input, e.g. tumor delineation, for extracting imaging features.

Main results. Our experimental results on a well-established benchmark PET-CT dataset show that our method achieved the highest accuracy (0.896) and AUC (0.903) scores when compared to the state-of-the-art methods (unpaired student’s t-test p-value < 0.05).

Significance. Our method could be an effective and supportive tool to aid physicians in tumor quantification and in identifying image biomarkers for cancer treatment.

1. Introduction

Soft-tissue sarcomas (STSs) are a heterogeneous group of malignant tumors that arise in connective tissue in the body (Toro et al 2006). About 25% of STSs patients develop distant metastases (DM) that are the main cause of death (Billingsley 1999, Stojadinovic et al 2002). DM refers to the spread of tumors from the site of origin, usually, to other structures and viscera in the body. Tumors can spread to other locations in the same organ but these tend to produce local problems. For high-grade STSs, 50% of STS patients will develop DM and the median survival is 1.6 months after they developed DM (Billingsley 1999). It is hoped that the early identification of patients at a high risk of developing DM will enable the introduction of more effective therapies (Komdeur et al 2002, Vallières et al 2015).

Positron emission tomography-computed tomography, using the PET radiopharmaceutical 18F-Fluorodeoxyglucose (FDG PET-CT), is regarded as the imaging modality of choice for the staging and assessment of STSs (Fuglø et al 2012). FDG PET-CT provides formation about the metabolic activity of the tumor coupled to the anatomical data that identifies where the abnormal metabolism is located. In figure 1, we provide two examples of STS patients with tumors in the muscles of the thighs, where although the images appear similar, only 1 of the 2 patients developed DM.

Radiomics has been widely used to correlate data from medical images to disease outcomes such as the development of DM, overall survival, disease-free survival. The fundamental hypothesis in radiomics is that
medical images contain information that can reflect underlying pathophysiology, and such information can be used to inform clinical decision-making (Lambin et al. 2012). Traditional radiomics methods have 3 main steps: (i) manual segmentation and annotation of regions of interest (ROIs), (ii) extracting hand-crafted (HC) radiomic features (e.g. intensity, texture, shape) and, (iii) building predictive models such as support vector machines to correlate extracted features with the clinical outcomes. A large number of studies have exploited single-modality images, such as magnetic resonance images (MRI) (Corino et al. 2018), CT (Peeken et al. 2019), and PET (Hao et al. 2018). There are few radiomics studies where multi-modality data (e.g. PET-CT, PET-MRI) were used for patient outcome prediction in STS (Vallières et al. 2015) and head-and-neck cancer (Vallieres et al. 2017). The performance of these methods, however, relies on a priori skillset in tumor delineation, hand-crafting image features and tuning a large number of parameters for the predictive models.

Recent advances in convolutional neural networks (CNNs) have inspired CNN-based radiomics. CNNs extract high-level semantic information in an end-to-end manner, which reduces the need for prior knowledge in HC radiomic features definition and other manual input (Hosny et al. 2018b, Afshar et al. 2019). A number of studies show that the deep features have promising performance in outcome prediction for head-and-neck cancer patients (Diamant et al. 2019), neoadjuvant chemoradiation response prediction for locally advanced rectal cancer patients (Fu et al. 2020), preoperative meningiomas grading (Zhu et al. 2019) and glioblastoma multiforme survival prediction (Lao et al. 2017). However, these approaches were designed for single-modality images. There are limited studies reported on CNN-based radiomics approaches for multiple combined imaging modalities (Chen et al. 2019, Peng et al. 2019). The underlying assumption is that CNN’s abstract level features and HC texture features are complementary (Wang et al. 2018), and so combining them could provide more accurate results. In both studies, using CNNs alone is sub-optimal and performance is enhanced by incorporating traditional radiomics components such as HC feature extraction. The sub-optimal performance of CNN alone is due to the extraction of the imaging features only from the last convolutional layer that contained high-level semantic information and texture information, from shallow layers, is neglected. Therefore, these existing CNN-based radiomics studies suffer from the limitations of traditional radiomics methods. Meanwhile, they fail to fully exploit the potential of CNNs that have the ability to capture complementary information from multi-modality medical images.

In this work, we propose a new CNN-based method to predict the development of DM and we refer to the proposed method as constrained hierarchical multi-modality feature learning (CHMFL). Our contributions are as follows:

(i) We used a constrained feature learning (CFL) module to spatially guide the learning process to focus on the important semantic regions (e.g. tumors). Our formulation of the module means that it can target the...
functional hot spots in PET within the anatomical context of CT. The CFL allows the CNN to automatically
detect and focus on the tumor, while conventional radiomics and other CNN-based radiomics methods
with single or multi-modality imaging data require manual annotation as the input to constrain the feature
extraction process within the tumor region.

(ii) We designed a hierarchical multi-modality feature learning (HMFL) module that derives optimal radiomic
features by integrating complementary features across modalities at different scales. Our formulation of the
module combined multi-modality features from different scales in an iterative manner. In comparison,
existing multi-modality radiomics methods extract imaging features separately from the individual imaging
modalities and fuse the features later or integrate the multi-modal images at an earlier stage. The hierarchical
combination of features enabled a more complex and flexible fusion of PET and CT features, e.g. low-level
PET texture features from a shallow layer with semantic CT features from a deeper layer.

We previously reported preliminary work in a conference paper, using an unconstrained 3D CNN with
traditional radiomic features to predict DM from FDG PET-CT (Peng et al 2019). This current work differs from
the previous work in that we: (i) introduce a CFL module to remove the reliance on annotated ROIs during the
testing stage; (ii) use a hierarchical multi-level learning module to derive discriminative radiomics features from
multiple imaging modalities and, (iii) have carried out a more exhaustive assessment of the new method in
comparison with previous and more recent studies.

2. Methods

2.1. Materials and pre-processing
We used a public PET-CT STSs dataset from the Cancer Imaging Archive (Clark et al 2013, Vallières et al 2015).
This dataset has 51 patients with histologically proven, extremity primary STS. Each patient had 4 imaging
modalities FDG PET, CT and T1-weighted and T2-weighted with fat-suppression (T2FS) MR scans. The gross
tumor volume was manually annotated slice-by-slice on T2FS MR scans by an expert radiation oncologist and
registered to PET and CT images. DM was confirmed later by biopsy or on additional imaging studies; 3 patients
who developed local tumor recurrence were excluded. Hence there were 24 patients with and 24 patients
without DM. The reconstructed PET image slices had a size of 96 × 96 with a pixel size of 3.91–5.47 mm² while
CT image slice had a size of 512 × 512 with a pixel size of 0.98 mm². Both PET and CT had a slice thickness of
3.27 mm. All the PET images contained standardized uptake value (SUV) data where the SUV reflects FDG
uptake and is based on the dose of FDG and the body mass of the patients. Isotropic voxel resampling was
applied to PET and CT images to ensure that the spatial dimensions were the same in all directions. After that, a
fixed-sized bounding box of 112 × 112 × 144 mm³ was determined based on the size of the largest tumor in
this dataset, which was used to extract the input imaging volumes. Finally, we used standardization and contrast
enhancement to minimize the influence of both high and low-frequency noise (Foracchia et al 2005).
Specifically, for PET images, the input volumes were standardized by the mean and standard deviation values of
the whole dataset, so as to adjust all the ROIs to a notionally common scale based on the metabolism intensity of
tumor regions. For the CT images, the input volumes were standardized by the mean and standard deviation values of
each patient individually due to the characteristics of the Hounsfield scale in CT images.

2.2. Overview of the proposed method
In figure 2, we outlined our CHMFL architecture. The volumetric PET and CT images were pre-processed and
then fed separately into two identical branches. Each branch has multiple downsampling convolutional layers
for feature extraction (as shown within the yellow PET and blue CT feature maps in figure 2). The feature maps
derived after each convolutional layer were adaptively pooled and then concatenated into a single feature vector
to facilitate HMFL. The CFL module used several upsampling convolutional layers to guide the network to focus
on the important regions (e.g. the tumor). This process also incorporated the fine-grained features forwarded
from the HMFL module at each level. Finally, the derived multi-modality PET-CT features (as shown in the left
lower part of figure 2) were fed into three fully connected layers for DM prediction.

2.3. CFL module
Our CFL module was designed to guide the learning process to focus on semantically important regions at
both the training and inference stages. This was achieved by gathering and assembling the complementary
information from multi-modality PET-CT images to obtain a 2-channel volumetric segmentation output. We
used 4 transposed convolutional blocks to expand the spatial support from the feature maps at a lower scale for
upsampling. These upsampling blocks at different levels shared similar structures (see CFL module in table 1 for
Meanwhile, the multi-modality PET-CT features extracted from the HMFL module were forwarded to the upsampling blocks by horizontal connections (see Figure 2). In this way, we gathered fine-grained detail for tumor contour prediction that would be otherwise lost in the downsampling path. In turn, tumor regions were emphasized in the HMFL module by the backpropagation process. Moreover, in order to avoid the vanishing gradient problem with network deepening, a residual learning was formulated after the concatenation of forwarded PET-CT features and the corresponding upsampled feature maps at each level: the concatenated feature map was processed through several convolutional layers and non-linearities, then added to the output of the last nonlinearity within the residual learning.

During the training stage, two loss functions were employed for different tasks. A pixel-wise cross-entropy loss was used to compare the predicted segmentation output with the ground-truth tumor annotation. Another cross-entropy loss was used for DM prediction. Given a weight $w$ for our CFL module $0 \leq w \leq 1$, the total loss $L$ was defined as follows:

$$L = \alpha \cdot L_{seg} + (1 - \alpha) \cdot L_{DM}$$

where $\alpha$ is a weighting factor to balance the two loss terms. The pixel-wise cross-entropy loss $L_{seg}$ is calculated as:

$$L_{seg} = -\frac{1}{N} \sum_{i=1}^{N} \left( y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i) \right)$$

where $y_i$ is the ground-truth label, $\hat{y}_i$ is the predicted probability, and $N$ is the number of pixels in the input image. The DM prediction loss $L_{DM}$ can be defined as:

$$L_{DM} = -\frac{1}{N'} \sum_{i=1}^{N'} \left( y_{DM, i} \log \hat{y}_{DM, i} + (1 - y_{DM, i}) \log (1 - \hat{y}_{DM, i}) \right)$$

where $y_{DM, i}$ is the ground-truth DM label, $\hat{y}_{DM, i}$ is the predicted probability for DM, and $N'$ is the number of DM pixels in the input image.

Table 1. Network architecture used in the HMFL and CFL Module.

| Layers          | Details (kernel size, stride, padding, …) | Output size (batch size, channel number, …) |
|-----------------|------------------------------------------|---------------------------------------------|
| **HMFL module** | Input transition Conv3d (5 x 5, 1, 2); BatchNorm; ELU | 1 x 16 x 112 x 112 x 144 |
| Down_Conv_1     | Conv3d (2 x 2, 2, 0); BatchNorm; ELU;     | 1 x 32 x 56 x 56 x 72 |
| Down_Conv_2     | Conv3d (2 x 2, 2, 0); BatchNorm; ELU;     | 1 x 64 x 28 x 28 x 36 |
| Down_Conv_3     | Conv3d (2 x 2, 2, 0); BatchNorm; ELU;     | 1 x 128 x 14 x 14 x 18 |
| Down_Conv_4     | Conv3d (2 x 2, 2, 0); BatchNorm; ELU;     | 1 x 256 x 7 x 7 x 9 |
| **CFL module**  | Up_Conv_1 ConvTranspose3d (2 x 2, 2, 0); BatchNorm; ELU | 1 x 128 x 14 x 14 x 18 |
| Up_Conv_2       | ConvTranspose3d (2 x 2, 2, 0); BatchNorm; ELU | 1 x 64 x 28 x 28 x 36 |
| Up_Conv_3       | ConvTranspose3d (2 x 2, 2, 0); BatchNorm; ELU | 1 x 32 x 56 x 56 x 72 |
| Up_Conv_4       | ConvTranspose3d (2 x 2, 2, 0); BatchNorm; ELU | 1 x 16 x 112 x 112 x 144 |
| Output transition Conv3d (5 x 5, 1, 2); BatchNorm; ELU | 1 x 2 x 112 x 112 x 144 |
| Conv3d (1 x 1, 1, 0); SoftMax |                               |
\[ L = -(1 - w)^2 \sum_{m=1}^{M=2} p_{1,m} \log q_{1,m} - w \sum_{n=1}^{N} p_{2,n} \log q_{2,n}, \]  

where \( p_{1,m} \) represents the target probability of developing DM, \( q_{1,m} \) (the output of this network) represents the predicted probability of developing DM, and \( M \) denotes the number of output neurons generated by the last fully connected layer in this network. \( q_{2,n} \in Q \) is the predicted binary segmentation volume, \( p_{2,n} \in P \) is the ground-truth binary annotation image and \( N \) denotes the total number of image voxels. \( w \) is a weight to balance the two losses.

### 2.4. HMFL module

We used 5 convolutional blocks for multi-modality image feature extraction (more details of the HMFL module are provided in table 1). PET and CT images were processed separately by the identical PET and CT branches. Within each convolutional block, the output feature map of the 3D convolutional layer was defined as:

\[ F = \mathcal{W} * X + b, \]

where \( F \) is the input to the convolution layer, \(*\) is the convolution operation, \( \mathcal{W} \) denotes the learned weights, and \( b \) is the learned bias. A batch normalization layer and a nonlinear activation function ELU were also added. By performing a 3D convolution with a kernel size of \((I, J, K)\), the value at the location \((x, y, z)\) of the feature map \( F \) was determined from its neighborhood:

\[ F(x, y, z) = \sum_{i} \sum_{j} \sum_{k} W(i, j, k) * X(x + i, y + j, z + k) \]

with \(-\left\lceil \frac{I}{2} \right\rceil \leq i \leq \left\lceil \frac{I}{2} \right\rceil, -\left\floor{\frac{J}{2}} \leq j \leq \left\floor{\frac{J}{2}}, -\left\lceil \frac{K}{2} \right\rceil \leq k \leq \left\lceil \frac{K}{2} \right\rceil \].

For HMFL, we firstly concatenated PET and CT feature maps to include multi-modality context information at each scale of the convolutional layers. After concatenation, an adaptive pooling layer was used to project the fused feature map into a single vector. This combination of feature maps from different scales could obtain both diverse texture details from shallow layers and high-level semantic layers, which can be defined as:

\[ F_{fusion} = F \left( \bigcup_{l=1}^{L=4} APL(F_{ct}^{l} \otimes F_{ct}^{l}) \right) \otimes F(F_{pet}^{5} \otimes F_{ct}^{5}), \]

where \( APL \) denotes the adaptive max-pooling layer, and \( L \) is the number of convolutional layers for multi-modal PET-CT feature extraction, and \( \otimes \) represents the concatenation operation.

Multi-modality feature maps at each scale were concatenated into a single fully connected layer and processed with additional two fully connected layers. ReLU layers and dropout layers with a probability of 0.5 were added after each fully connected layer to reduce overfitting.

### 2.5. Implementation details

Our method was implemented with PyTorch (Paszke et al 2017) and ran on an 11GB NVIDIA GeForce GTX 1080Ti GPU. The learning rate was set to 0.0001 and the batch size was set to 1. Our model was initialized using the approach presented in He et al (2015), and adaptive-moment-estimation (Adam) (Kingma and Ba 2014) was used for network optimization. We have further conducted an experiment where we adopted data augmentation techniques at the training stage, i.e. randomly rotation (90°, 180°, or 270°) in the axial axis and randomly flip in one of all three axes (axial, sagittal and coronal), and we named this comparison experiment as CHMFL_Agumented. The training was terminated when no further changes in the total loss. In our method, the total loss was generally stable after 200 epochs and our CNN model took approximately five hours to fine-tune with. In addition, our model took around 10 s to inference 8 patients; this time is similar to the existing 3D based CNN models.

### 2.6. Experimental setup

The experiments that we carried out compared our proposed method to:

(a) state-of-the-art radiomics methods that were separated into 3 categories:

(i) Traditional radiomics: We used the method proposed by Vallieres et al (2017), Multiresolution auto-correlation handcrafted (HC) clinical texture features were extracted and included: the grey-level co-occurrence matrix, grey-level run-length matrix, grey-level size zone matrix, and neighborhood grey-tone difference matrix features. We performed a stepwise forward feature selection scheme with multivariable analysis. The optimal conventional feature set contained 25 different radiomics features. A random forest (RF) classifier was trained with these texture features. We refer to this method as HC + RF;
(ii) CNN-based radiomics method. We reimplemented the method of Diamant et al (2019) based on the technical details from their paper that used 2D image slices as the input data. This method has four main operational layers: 2D convolutional, nonlinearity (PReLU), max-pooling, and fully connected layers for classification. We refer to this method as CNLPC.

(iii) Hybrid methods that combine CNNs with traditional radiomic components:

1. The 3DMCL method proposed by Peng et al (2019), which had two branches to separately extract features from PET and CT volumes, and then deep features were combined with hand-crafted radiomics features and fed into fully connected layers to make a final prediction.

2. A hybrid predictive model, comprised of a many-objective radiomics model and a 3D CNN, which used spatial contextual information from PET-CT images that was designed by Chen et al (2019). The output of the 2 components was fused through an evidential reasoning approach to predict lymph node metastases in head-and-neck cancers. We refer to this method as MOR + 3D CNNs.

(b) Different imaging modalities and different CNN dimensions: We used all the PET-CT image slices containing tumor regions as the input for the 2D CNNs-based comparison methods. We used the PET-CT volumes, based on the bounding box, as the input for the 3D CNNs-based comparison methods. The PET-CT input slices were directly obtained from the 3D volumes that went through the same pre-processing steps in section 2.1. The 2D and 3D CNNs used a similar architecture to the 3DMCL method that was suggested by Peng et al (2019) and the state-of-the-art method in this DM prediction problem for STS patients.

(c) Individual components of the proposed method:

(i) CFL—our proposed method without HMFL module, used PET and CT images as input;

(ii) Mask + HMFL—our proposed method without CFL module, used PET, CT, and tumor label images as input (2-channel PET-label image and 2-channel CT-label image).

The state-of-the-art methods used in our comparisons were those mentioned in the related works, and they can be divided into three categories: We used a hold out 6-fold cross-validation approach for our method and the methods that we compared it to. The 48 PET-CT data were randomly divided into 6 equal-sized subsets and each subset had 8 PET-CT images. For each fold, 5 subsets were used to train the network and the remaining subset was used for testing. We repeated this process 6 times to assess the 48 PET-CT images. The results that we present in section 3 are the mean value across all 6 folds. Six established evaluation metrics were adopted, including accuracy, sensitivity, specificity, precision, F1 score, and area under the receiver—operating characteristic curve (AUC). For all experimental comparisons with our proposed CHMFL method, we computed the p-value with an unpaired student’s t-test.

3. Results

Our CHMFL achieved the overall best DM prediction performance with the highest accuracy (0.896), sensitivity (0.958), F1 score (0.902) and AUC (0.903)—see table 2 and figure 3; Our CHMFL’s specificity (0.833) and precision (0.852) ranked at the second place. In addition, our CFL module alone obtained the highest sensitivity (0.958), and our Mask + HMFL obtained the highest specificity (0.875) and precision (0.857).

When compared with methods using 2D or 3D CNNs with different modality imaging data (e.g. PET, CT, PET-CT)—see table 3 and figure 4, our CHMFL method outperformed all the comparison CNNs-based methods regardless of imaging modality and the kernel dimension of CNN. 3D CNNs performed better than those using 2D CNNs. Methods based on PET images outperformed methods based on CT images. We evaluated how the CFL module’s weight w affected the performance of CHMFL over three key evaluation metrics (e.g. accuracy, sensitivity and specificity). The result in figure 5 suggests that the best performance was achieved when the weight was 0.5.

Two example PET-CT studies are shown in figure 6 with corresponding visualization results of the extracted feature map with respect to three existing radiomics methods that outperformed other comparison methods except for our CFL and CHMFL methods. In figure 6, We generated class activation maps (Zhou et al 2016) to visualize the predicted class scores on any given image. Global average pooling layers were implemented to capture the spatial average of the feature map of each unit at the last convolutional layer. A channel-wise sum of...
value and

When compared to 3D-CNN-PET-CT

Our CFL automatically focused on ROIs that were semantically more important to the STSs DM predictions.

4.1. CFL module analysis

our CHMFL outperformed state-of-the-art radiomics methods.

modality PET-CT image features;

were able to correctly focus on tumor regions in the derived feature maps from tumor segmentation could be useful for patient outcome prediction;

Our main

4. Discussion

these feature maps was used to generate the final output (i.e. class activation maps). The discriminative object parts detected by the CNNs were highlighted.

The discriminative ability of both the 3DMCL and our CFL and CHMFL methods is depicted in figure 7, via t-distributed stochastic neighborhood embedding (t-SNE) (Maaten and Hinton 2008) visualization. t-SNE is an unsupervised, nonlinear technique primarily used for visualizing high-dimensional image features in a two or three-dimensional space, which allows for exploring the relationship of the extracted features.

4. Discussion

Our main findings are that: (i) our CFL module automatically identified the tumor and deep features learned from tumor segmentation could be useful for patient outcome prediction; (ii) our HMFL module derived multi-modality PET-CT image features; (iii) our method improved upon current single-modality methods and, (iv) our CHMFL outperformed state-of-the-art radiomics methods.

4.1. CFL module analysis

Our CFL automatically focused on ROIs that were semantically more important to the STSs DM predictions. When compared to 3D-CNN-PET-CT (figure 6(f)), both our CFL and CHMFL methods with the CFL module were able to correctly focus on tumor regions in the derived feature maps (figures 6(c) and (d)). In contrast, 3D-CNN-PET-CT falsely concentrated on many normal uptake regions, which resulted in a >10% decrease in AUC value and >4% decrease in both accuracy and F1 score (shown in table 3). Moreover, when compared to existing methods that segmentated tumor region before feature extraction and outcome prediction, such as 3DMCL.

---

**Figure 3.** Classification performance (measured in receiver operating characteristic (ROC) curve) of our CHMFL in comparison to other existing radiomics methods.

**Table 2.** Classification performance comparisons with existing radiomics methods.

| Method                  | Accuracy | Sensitivity | Specificity | Precision | F1 score | AUC  |
|-------------------------|----------|-------------|-------------|-----------|----------|------|
| HC + RF                 | 0.750 (0.102)* | 0.792 (0.125)* | 0.708 (0.105)* | 0.731 (0.094)* | 0.760 (0.102)* | 0.726 (0.126)* |
| CNLPC                   | 0.729 (0.094)* | 0.792 (0.188)* | 0.667 (0.130)* | 0.703 (0.083)* | 0.745 (0.109)* | 0.783 (0.187)* |
| MOR + 3D CNNs           | 0.729 (0.111)* | 0.750 (0.224)* | 0.780 (0.204)* | 0.720 (0.112)* | 0.750 (0.129)* | 0.793 (0.174)* |
| 3DMCL                   | 0.854 (0.085)* | 0.917 (0.188)* | 0.792 (0.187)* | 0.815 (0.168)* | 0.863 (0.094)* | 0.854 (0.148)* |
| Mask + HMFL (Ours)      | 0.813 (0.094)  | 0.750 (0.129)  | 0.875 (0.224)  | 0.857 (0.174)  | 0.800 (0.094)  | 0.769 (0.139)  |
| CFL (Ours)              | 0.813 (0.102)  | 0.958 (0.102)  | 0.667 (0.209)  | 0.742 (0.145)  | 0.836 (0.066)  | 0.852 (0.155)  |
| CHMFL (Ours)            | 0.854 (0.051)  | 0.875 (0.102)  | 0.833 (0.129)  | 0.840 (0.074)  | 0.857 (0.034)  | 0.873 (0.184)  |
| CHMFL_Augmented (Ours)  | 0.896 (0.094)  | 0.958 (0.102)  | 0.833 (0.204)  | 0.852 (0.142)  | 0.902 (0.080)  | 0.903 (0.112)  |

---

*Note: Values in parentheses denote standard deviation, and the asterisk (*) indicates statistically significant differences (P < 0.05) compared to the HC + RF baseline method.
Figure 4. Classification performance (measured in ROC) of our CHMFL in comparison to methods using different modal image and convolutional layers.

Table 3. Classification performance comparisons with methods using different modal image and convolutional layers.

| Method                  | Accuracy | Sensitivity | Specificity | Precision | F1 score | AUC  |
|-------------------------|----------|-------------|-------------|-----------|----------|------|
| 2D-CNN-CT               | 0.583(0.102) | 0.708(0.292) | 0.458(0.292) | 0.567(0.094) | 0.630(0.153) | 0.503(0.254) |
| 2D-CNN-PET              | 0.729(0.051) | 0.542(0.129) | 0.917(0.209) | 0.867(0.174) | 0.667(0.051) | 0.656(0.139) |
| 2D-CNN-PET-CT           | 0.729(0.094) | 0.792(0.188) | 0.667(0.130) | 0.703(0.083) | 0.745(0.109) | 0.698(0.187) |
| 3D-CNN-CT               | 0.667(0.085) | 0.667(0.213) | 0.667(0.258) | 0.667(0.112) | 0.667(0.131) | 0.684(0.203) |
| 3D-CNN-PET              | 0.771(0.094) | 0.750(0.188) | 0.792(0.224) | 0.783(0.168) | 0.766(0.094) | 0.734(0.156) |
| 3D-CNN-PET-CT           | 0.792(0.105) | 0.792(0.209) | 0.792(0.204) | 0.792(0.143) | 0.792(0.111) | 0.773(0.148) |
| Mask + HMFL (Ours)      | 0.813(0.094) | 0.750(0.129) | **0.875 (0.224)** | 0.857 (0.174) | 0.800 (0.094) | 0.769 (0.139) |
| CFL (Ours)              | 0.813 (0.102) | **0.958 (0.102)** | 0.667 (0.209) | 0.742 (0.145) | 0.836 (0.066) | 0.852 (0.155) |
| CHMFL (Ours)            | 0.854 (0.051) | 0.875 (0.102) | 0.833 (0.129) | 0.840 (0.074) | 0.857 (0.034) | 0.873 (0.184) |
| CHMFL_Augmented (Ours)  | 0.896 (0.094) | **0.958 (0.102)** | 0.833 (0.204) | 0.852 (0.142) | **0.902 (0.080)** | **0.903 (0.112)** |

*p < 0.05, in comparison to our proposed CHMFL method derived from an unpaired student’s t-test. The results are presented in the form of ‘mean value (standard deviation)’.

…the state-of-the-art method for STS DM prediction, figure 6(e), both our CHMFL and CFL methods could accurately identify the entire sarcoma with more details. Although Mask + HMFL was forced to focus on the tumor region by incorporating an extra channel of tumor label image as input, there were still some false positive regions, e.g. as in the bottom case of figure 6(e) when there were similar tissues around the tumor. Without our CFL module, the Mask + HMFL method had a tendency of not predicting DM due to concentrating on more regions other than tumors. Although this contributed to a 4% increase in specificity and 1% increase in precision, there are >10% decrease in both AUC score and sensitivity when compared with our proposed CHMFL (shown in table 2). We also noted that automatically constraining the learning process to extract feature maps only from the tumor region can obtain more information that can potentially reflect underlying pathophysiology, such as the heterogeneity of STS, which is an important prognostic factor of DM development (Eary et al 2008). In addition, such an automated process removes the reliance on accurate manual tumor delineation during the inference stage while obtaining better overall performance.

4.2. HMFL module analysis

The inclusion of HMFL in our method further improved the performance of the CFL. Most existing CNNs-based radiomics methods, including 3DMCL, CNLPC and MOR + 3DCNNs, only leverage high-level features extracted from the last convolutional layer in their model, and therefore inherently disregarded the complementary PET and CT image features at the lower level of the network. In contrast, our method iteratively and hierarchically fused the multi-modality PET and CT image features across the different image scales, which
enabled more flexible and complex multi-modality information fusion. As an example, the feature map derived from our CHMFL method (figure 6(c)) captured more details inside the tumor and better predicted the tumor contour when compared with our CFL method (figure 6(d)).

**4.3. Evaluation of CNN-based methods with different image modalities and different convolutional layers**

There was a marked difference in performance between PET-CT CNNs and CNNs with PET alone or CT alone. Further, PET-based methods outperformed CT-based methods. This was expected since PET images provided metabolism information of tumors, while CT can only provide the anatomical information, and tumor regions are not always visible in CT (as exemplified in figure 1). The relatively lower performance of 2D CNNs, when compared to 3D CNNs counterparts is attributed to the fact that volumetric image features derived from 3D CNNs are better to discriminate the spatial information within the tumor that is associated with the DM development, e.g. volumetric tumor shape and size (Hosny et al 2018a). In contrast, 2D CNNs-based methods (e.g. 2D-CNN-CT and 2D-CNN-PET) have limited representation capability of tumor characteristics in two dimensions with few axial slices. Therefore, it would be better to incorporate 3D CNNs with multi-modality imaging data when the computational power is available, which allows achieving better performance (as shown in table 3).

**4.4. Comparison of CHMFL with existing methods**

Our CHMFL method obtained the best overall performance when compared with the existing radiomics methods. HC + RF method achieved competitive performance overall the evaluation metrics except the AUC score when compared with CNLPC and MOR + 3DCNNs. Unfortunately, the performance of HC + RF was reliant on effective feature handcrafting and tuning a large number of parameters, which may limit its generalizability to different datasets. The performance improvement from 3DMCL to CNLPC and MOR + 3DCNNs was likely due to the use of multi-modality PET and CT images providing complementary
information. When compared to the second-best performing method 3DMCL, our method achieved much higher specificity (as shown in table 2). 3DMCL is reliant on using single-level image features for prediction, which results in 3DMCL overfits to the positive prediction of DM, which were unable to discriminate the tumors. As exemplified in figure 7 our CHMFL had greater separability between the patients with/without DM than both 3DMCL and 3D-CNN-PET-CT, where only a few cases were not properly separated.

5. Conclusions

We proposed a CHMFL method for predicting the development of DM. Our results with a public dataset of soft-tissue sarcomas showed that our method was capable of better identifying PET-CT radiomic features in primary tumors that were associated with the development of DM, when compared to the state-of-the-art radiomics methods.

5.1. Limitations and future work

Our focus in the current study was to investigate the prediction of distant tumor spread (metastatic disease) in patients with STSs from PET-CT images. Predicting the presence of DM as a binary classification is an abstraction of a time to event prediction problem (i.e. estimating the point at which an event occurs). The time to event problem is a more complicated modelling challenge than binary classification and may require different methodological approaches. In the public dataset we used, all the patients have a 7 year follow-up period for outcome observation and DM was generally confirmed within 4 years after diagnosis of primary STS; this was appropriate for binary classification. The public dataset is small (n = 51) and thus there was no separate held-out data used only for testing. We reported only the mean results across all validation experiments. The results may be different with a held-out cohort in a much larger dataset. We are actively working on characterizing and annotating a much larger soft-tissue sarcoma dataset. Moreover, we have not generalized our results to other tumor types or where other imaging modalities are employed. In future work we intend to evaluate our approach in non-small cell lung cancer and lymphomas, using PET-CT, and also include other parameters such as local tumor recurrences and long-term survival. In lymphomas there are generally multiple sites of disease and disease recurrence occurs unpredictably and so analyzing multiple lesions will be necessary to attempt to predict where the disease will occur. We would like to adapt our approach to such a situation and this will require multiple bounding boxes.

Acknowledgments

This work was supported in part by the Australian Research Council (ARC) grants (DP200103748, IC170100022).

ORCID iDs

Yige Peng https://orcid.org/0000-0001-5549-2688
Lei Bi https://orcid.org/0000-0001-9759-0200
References

Afshar P, Mohammad Al A, Platanitis K N, Oikonomou A and Benali H 2019 From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities IEEE Signal Process Mag. 36 132–60

Billingsley KG 1999 Multifactorial analysis of the survival of patients with distant metastasis arising from primary extremity sarcoma Cancer Cytopathol 85 389–95

Chen L, Zhou Z, Shei D, Zhang Q, Shah J, Pham N-L, Jiang S and Wang J 2019 Combining many-objective radiomics and 3D convolutional neural network through evidential reasoning to predict lymph node metastasis in head and neck cancer Phys. Med. 64 075011

Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppell P, Moore S, Phillips S, Maffitt D and Pringle M 2013 The cancer imaging archive (TCIA): maintaining and operating a public information repository J. Digit. Imaging 26 1045–57

Corino V D A, Montin E, Messina A, Casali P G, Gronchi A, Marchianò A and Mainardi L T 2018 Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions J. Magn. Reson. Imaging 47 829–40

Diamant A, Chatterjee A, Villiers M, Shenouda G and Seuntjens J 2019 Deep learning in head & neck cancer outcome prediction Sci. Rep. 9 2764

Eary J F, O’Sullivan F, O’Sullivan J and Conrad E U 2008 Spatial heterogeneity in sarcoma 18F-FDG uptake as a predictor of patient outcome J. Nucl. Med. 49 1973–9

Forbach M, Grisin E and Ruggeri A 2005 Luminosity and contrast normalization in retinal images Med. Image Anal. 9 179–90

Fu J, Zhong X, Li N, Van Dams R, Lewis J, Sung K, Raldow A C, Jin J and Qi X S 2020 Deep learning-based radiomic features for improving neoadjuvant chemoradiation response prediction in locally advanced rectal cancer Phys. Med. Biol. 65 75001

Fuglø H M, Jørgensen S M, Loft A, Hovgaard D and Petersen M M 2012 The diagnostic and prognostic value of 18F-FDG PET-CT in the initial assessment of high-grade bone and soft tissue sarcoma. A retrospective study of 89 patients Eur. J. Nucl. Med. Mol. Imaging 39 1416–24

Hao He et al 2018 Shell feature: a new radiomics descriptor for predicting distant failure after radiotherapy in non-small cell lung cancer and cervix cancer Phys. Med. Biol. 63 95007

He K, Zhang X, Ren S and Sun J 2015 Delving deep into rectifiers: surpassing human-level performance on imagenet classification Proc. IEEE Int. Conf. on Computer Vision (Piscataway, NJ: IEEE) pp 1026–34

Hosny A, Parmar C, Coroller T P, Grossmann P, Zeleznik R, Kumar A, Busink J, Gillies R J, Mak R H and Aerts H J W L 2018a Deep learning for lung cancer prognosis: a retrospective multi-cohort radiomics study PLoS Med. 15 e1002711

Hosny A, Parmar C, Quackenbush J, Schwartz L H and Aerts H J W L 2018b Artificial intelligence in radiology Nat. Rev. Cancer 18 500–10

Kingma D P and Ba J 2014 Adam: a method for stochastic optimization arXiv:1412.6980

Komdeur R, Hoekstra H J, Van den Berg E, Molenaar W M, Praz E, de Vries E G E and van der Graaf W T A 2002 Metastasis in soft tissue sarcomas: prognostic criteria and treatment perspectives Cancer Metastasis Rev. 21 167–83

Lambin P et al 2012 Radiomics: extracting more information from medical images using advanced feature analysis Eur. J. Cancer 48 441–6

Luo J, Chen Y, Li C-G, Li Q, Zhang J, Liu J and Zhai G 2017 A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme Sci. Rep. 7 10353

van der Maaten L and Hinton G 2008 Visualizing data using t-SNE J. Mach. Learn. Res. 9 2579–605

Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L and Lerer A 2017 Automatic differentiation in pytorch

Peeken J C, Bernhofer M, Spraker M B, Pfeiffer D, Devecka M, Thamer A, Shouman M A, Ott A, Nüsslin F and Mayr N A 2019 CT-based PET-CT soft-tissue sarcoma studies Int. J. Cancer 145 500–08

Peng Y, Bi L, Guo Y, Feng D, Fulham M and Kim J 2019 Deep multi-modality collaborative learning for distant metastases prediction in PET-CT soft-tissue sarcoma studies 2019 41st Annual Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC) (Berlin, Germany, 23–27 July 2019) (Piscataway, NJ: IEEE) pp 3658–68

Pistikopoulo A, Leon D H Y, Hooi A, Jacques D P, Lewis J and Brennan M F 2002 Analysis of the prognostic significance of microscopic margins in 2084 localized primary adult soft tissue sarcomas Ann. Surg. 235 424

Toro JR, Travis LB, Wu H J, Zhu K, Fletcher C D M and Devesa S S 2006 Incidence patterns of soft tissue sarcomas, regardless of primary site, in the surveillance, epidemiology and end results program, 1973–2001: an analysis of 26 758 cases Int. J. Cancer 83 2922–30

Vallieres M, Freeman C R, Skamene S R and El-Nagib I 2015 A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities Phys. Med. Biol. 60 5471

Vallieres M, Kay-Rivest E, Perrin L J, Liem X, Furstoss C, Aerts H J W L, Khaouam N, Nguyen-Tan P F, Wang C-S and Sultanem K 2017 Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer Sci. Rep. 7 1–14

Wang S, Hou Y, Li Z, Dong J and Tang C 2018 Combining convnets with hand-crafted features for action recognition based on an HMM-SVM classifier Multimed. Tools Appl. 77 18983–98

Zhou B, Khoala A, Lapedriza A, Oliva A and Torralba A 2016 Learning deep features for discriminative localization Proc. IEEE Conf. on Computer Vision and Pattern Recognition (Las Vegas, NV, 27–30 June 2016) (Piscataway, NJ: IEEE) pp 2821–9

Zhu Y, Man C, Gong L, Dong D, Yu X, Wang S, Fang M, Wang S, Fang X and Chen X 2019 A deep learning radiomics model for preoperative grading in meningioma Eur. J. Radiol. 116 128–34