IOP Publishing Phys. Med. Biol. 69 (2024) 215036 https://doi.org/10.1088/1361-6560/ad8857

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
1 September 2024

REVISED
3 October 2024

ACCEPTED FOR PUBLICATION
17 October 2024

PUBLISHED
29 October 2024

PAPER

Explainable AI for automated respiratory misalignment detection in PET/CT imaging

Yazdan Salimi[1], Zahra Mansouri[1], Mehdi Amini[1], Ismini Mainta[1] and Habib Zaidi[1,2,3,4,*]

[1] Division of Nuclear medicine and Molecular Imaging, Geneva University Hospital, Geneva, Switzerland
[2] Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
[3] Department of Nuclear Medicine, University of Southern Denmark, Odense, Denmark
[4] University Research and Innovation Center, Óbuda University, Budapest, Hungary
* Author to whom any correspondence should be addressed.

E-mail: habib.zaidi@hug.ch

Keywords: PET/CT, image quality, misalignment artifact, deep learning, segmentation

Supplementary material for this article is available online

Abstract

Purpose. Positron emission tomography (PET) image quality can be affected by artifacts emanating from PET, computed tomography (CT), or artifacts due to misalignment between PET and CT images. Automated detection of misalignment artifacts can be helpful both in data curation and in facilitating clinical workflow. This study aimed to develop an explainable machine learning approach to detect misalignment artifacts in PET/CT imaging. *Approach.* This study included 1216 PET/CT images. All images were visualized and images with respiratory misalignment artifact (RMA) detected. Using previously trained models, four organs including the lungs, liver, spleen, and heart were delineated on PET and CT images separately. Data were randomly split into cross-validation (80%) and test set (20%), then two segmentations performed on PET and CT images were compared and the comparison metrics used as predictors for a random forest framework in a 10-fold scheme on cross-validation data. The trained models were tested on 20% test set data. The model's performance was calculated in terms of specificity, sensitivity, F1-Score and area under the curve (AUC). *Main results.* Sensitivity, specificity, and AUC of 0.82, 0.85, and 0.91 were achieved in ten-fold data split. F1_score, sensitivity, specificity, and AUC of 84.5 vs 82.3, 83.9 vs 83.8, 87.7 vs 83.5, and 93.2 vs 90.1 were achieved for cross-validation vs test set, respectively. The liver and lung were the most important organs selected after feature selection. *Significance.* We developed an automated pipeline to segment four organs from PET and CT images separately and used the match between these segmentations to decide about the presence of misalignment artifact. This methodology may follow the same logic as a reader detecting misalignment through comparing the contours of organs on PET and CT images. The proposed method can be used to clean large datasets or integrated into a clinical scanner to indicate artifactual cases.

1. Introduction

The advent of positron emission tomography (PET)/computed tomography (CT) was a great leap forward in clinical oncology, enabling non-invasive, *in-vivo*, and quantitative evaluation of different pathologies (Maldonado *et al* 2007). The increased uptake of ^{18}F-fluorodeoxyglucose (^{18}F-FDG) in cancer cells led to rapid adoption of ^{18}F-FDG PET/CT in clinical oncology for diagnosis, staging/restaging, and monitoring of treatment response for various cancer types (Czernin *et al* 2007). The combination of PET with CT enforces the synergistic effect of integrating anatomical, biological, and functional information and provides the necessary anatomical data to correct the PET signal for photon attenuation within the patient's body (Kinahan *et al* 1998). However, CT-based attenuation correction (CTAC) can introduce artifacts into the

reconstructed PET images, reducing overall image quality and quantitative accuracy, and increasing the risk of misinterpretation and erroneous clinical decision-making (Shiri *et al* 2023a).

The artifacts introduced to PET images might originate from artifacts present in CT images translating to PET through the CTAC procedure (e.g. artifacts due to the presence of contrast agents, truncation, and metallic objects), or artifacts arising from the misalignment and mismatch between PET and CT images (Cook *et al* 2004, Blodgett *et al* 2011). Focusing on the latter, sequential scanning, and different duration of the scans for both imaging modalities leads to capturing a snapshot of the motion by fast CT scan, and an average of the motion reflected in the slow PET scan (Kyme and Fulton 2021). Voluntary or involuntary motion, including bulk movement of the head and limbs, or muscular movement of the spine or jaw can be simple, such as translations and rotations or include more complex movements, such as twisting of the spine (Montgomery *et al* 2006, Gu *et al* 2010). Involuntary movements are common among pediatric, elderly, or demented patients and can be managed using fixation devices to position patients (Beyer *et al* 2005) or corrected post-acquisition by registering PET and CT images (Yang *et al* 2020, Shiri *et al* 2023a). The involuntary motions on the other hand involve cardiac motion (Lamare *et al* 2014), thoracic and abdominal movements due to respiration (Visvikis *et al* 2006), and peristalsis or bowel movements (Nakamoto *et al* 2004). It can also include involuntary movements of the whole body or specific muscles resulting from different neurological conditions (Dinelle *et al* 2006). Among these, respiratory motion is of importance due to the prevalence and implications of its induced artifacts.

Modern multi-ring CT scanners with helical acquisition mode are fast enough to perform scans during a full-inspiration breath-hold (Kyme and Fulton 2021). In contrast, PET is acquired during tidal respiration with ∼3 min per bed position. Unlike CT, it does not reflect a snapshot position of thoracic and abdominal organs (Kinahan *et al* 1998, Blodgett *et al* 2011). This misalignment between the two imaging modalities increases blurring and reduces the contrast of organs and lesions around the lung-diaphragm interface in the CTAT PET image (Xu *et al* 2011). Noticeable anomalies appear particularly in two regions: First, in the interface of the liver and lung, the upper part of the liver in the PET image may appear in the bases of the lungs in the CT image, resulting in a curvilinear cold artifact in the reconstructed PET image. Second, in the region between the lung and left ventricle, the uptake of the left ventricle in the PET image may overlap the lung tissue in the CT image (Sun and Mok 2012). Previous studies show more than 40% of the studies in the thoracic region suffer from this misalignment between PET and CT images (Gould *et al* 2007, Shiri *et al* 2023b).

Respiratory motion artifacts would be clinically significant when suspicious lesions appear in the diaphragmatic regions, within or adjacent to an artifact. Lesions, partially or wholly can be overlooked, assigned to an incorrect organ (Blodgett *et al* 2011), or undergo inaccurate quantification (McCall *et al* 2010, Geramifar *et al* 2013). McCall *et al* reported up to 35% and 10% quantification errors due to patient movement for tumors of 5 and 10 mm size, respectively (2010). In another study, Erdi *et al* observed up to 30% change in standardized uptake value (SUV), 9 mm error in location and 21% size shrinkage of lung lesions in PET images resulting from respiratory motion (Erdi *et al* 2004). A phantom study performed by Pevsner *et al* has shown up to 75% underestimation of the maximum activity concentration in surrogate lung lesions (Pevsner *et al* 2005). These errors can also propagate to further stages of patient's treatment, such as focal radiation therapy with dose escalation when PET is used to derive tumor volume (Lamare *et al* 2022). The aforementioned errors result in misdiagnosis and incorrect decision-making by clinicians, adversely impacting patients' prognosis. A number of solutions have been suggested over the past three decades to correct for respiratory motion, mainly relying on either breathing instructions to patients or on external devices, such as motion tracking devices to enable gated PET acquisitions. However, factors such as the trade-off between patient's radiation dose and comfort, scanning duration, complexity, workload, cost, and computational burden have prevented these techniques from widespread adoption in the clinic (Lamare *et al* 2022). The recent emergence of data-driven motion compensation approaches is also driving the field and resulted in a number of innovative developments (Kyme and Fulton 2021). On the other hand, fully automated identification of respiratory misalignment artifact (RMA) artifacts is of great value, as it can alert the physicians to conduct their interpretation by analyzing the non-AC image. Furthermore, in severe cases it can be used to call for instant reperforming of image acquisition of a single bed covering the artifacted region.

Nowadays, we are experiencing a paradigm shift in medicine, driven by the rapid and widespread adoption of artificial intelligence (AI) for various tasks (Zaidi and El Naqa 2021). AI applications have spread to every corner of medicine, specifically medical imaging, encompassing image acquisition, reconstruction, data corrections, segmentation, as well as the use of medical images in diagnosis, prognosis and outcome prediction (Langlotz *et al* 2019, Sanaat *et al* 2021, Salimi *et al* 2023a, 2024a). The ultimate bridge between research and real-life application of AI in medicine relies on two key factors: generalizability (Willemink *et al* 2020) and explainability (Reddy 2022). Currently, access to data is commonly limited to patient populations

with a particular ethnicity from a single institution and geographical region. This often results in models with excellent performance on own dataset, but lower performance across other datasets (Soffer *et al* 2019). Heterogenous, vast, and inclusive curated datasets with high-quality images and labels are essential for developing generalizable and robust models suitable for commercialization and clinical exploitation (Park and Han 2018). However, available large datasets often prioritize quantity over quality and are sourced from multiple origins, each collecting data for specific applications. This frequently results in low-quality images, which can diminish the overall performance of the models trained on them or, in severe cases, likely leading to erroneous outcomes (Mayer-Schönberger and Ingelsson 2018, Redman 2018). Furthermore, curating large datasets is time-consuming and labor-intensive, with researchers spending most of their time on this task (van Ooijen 2019). This highlights the demand for automated AI-driven algorithms to conduct task-specific quality assurance and curate large datasets prior to enrollment in further developments (Amini *et al* 2023, Shiri *et al* 2023a, Salimi *et al* 2024b).

Another factor preventing acceleration of commercialization and widespread adoption of AI models in clinical setting is non-explainability and lack of transparency of most of the developed models (Amann *et al* 2020). As evidence-based decision making is one of the main principles of precision medicine, physicians show reluctancy and hesitation in using black-box AI-based models in daily practice (Kundu 2021, Yoon *et al* 2022). Explainable AI have become a hot topic in recent years, paving the way for deployment of AI in daily clinical practice (Reddy 2022). In this work, we introduce a fully automated AI-driven model to identify respiratory motion artifact in ^{18}F-FDG PET images using an explainable and transparent methodology. The proposed model can be used in daily practice to instantly identify images suffering from RMA artifact to prevent erroneous interpretations, and/or call for second image acquisition/reconstruction. Moreover, this fully automated model can be used to curate large datasets and deliver clean thoracic/abdominal datasets to be used further in the development of AI-driven models for various clinical tasks.

2. Material and methods

2.1. Dataset

A total number of 1216 PET/CT images were included in this study acquired on two clinical scanners (Siemens Healthineers, Knoxville, USA). Table 1 summarizes the patient population demographic parameters as well as acquisition and reconstruction parameters.

First, PET/CT images were visually labeled for the presence of respiratory misalignment. Then, four anchor moving organs were delineated using either PET or CT images as input. The agreement between segmentation masks was estimated using common segmentation evaluation metrics, a feature selection algorithm was used to select the most important features, then the selected metrics were fed to a random forest machine learning model to predict the presence or absence of respiratory motion artifact in PET images. Figure 1 summarizes the steps followed in this study.

2.2. Visual assessment and labeling of PET/CT images

All PET/CT image pairs were visualized using in house developed MatLab 2022-based software and were classified into two categories corresponding to images presenting with misalignment artifact (RMA) and without artifact (No RMA) depending on whether respiratory misalignment between PET and CT in the chest/abdomen interval region is present or not. *The reader checked three images, including non-corrected PET, attenuation and scatter corrected PET, and CT side by side and using fusion of each of these two images with toggling between modalities was possible. In clinical setting, all three images are visualized together to check the presence of RMA*. The labels were recorded for the next steps.

2.3. Organ segmentation

We used previously trained nnU-Net (Isensee *et al* 2021) deep learning models to segment four organs, including the liver, heart, spleen, and lungs on both PET and CT images. Two different models were available, a model that uses non-corrected PET (PET-NC) images as input, called PET-nnU-Net (Salimi *et al* 2024c) generating liver-PET, spleen-PET, heart-PET, and lungs-PET segmentations, and a second model using CT images as input called CT-nnU-Net (Salimi *et al* 2023b) generating liver-CT, spleen-CT, heart-CT, and lungs-CT segmentations. The two models were used to generate organ segmentation masks. The models were trained in a five-fold data split strategy and all five folds were inferenced on each corresponding image, i.e. five CT segmentation model models were ensembled on each CT image and five PET segmentation models were ensembled on each PET image. It should be emphasized that PET-NC images were converted to SUV units before using them as input to PET-nnU-Net models. Details about the performance of PET-nnU-Net and CT-nnU-Net models can be found in the above-mentioned references. Supplementary

Table 1. Patient demographics and PET/CT scanning parameters of clinical studies used in this study.

Group	No RMA	RMA
Numbers #	553	663
Pitch Factor	0.8	0.8
kVp	109.1 ± 11.3 (100.0–140.0)	109.1 ± 11.0 (100.0–140.0)
Manufacturer	Siemens Biograph Vision, Siemens Biograph mCT	Siemens Biograph Vision, Siemens Biograph mCT
$CTDI_{vol}$ (mGy)	4.28 ± 2.43 (0.30–20.85)	4.22 ± 2.05 (0.79–20.21)
Age (Years)	59.9 ± 17.1 (18.0–93.0)	62.1 ± 15.642 (18–96.0)
Gender	Female: 279, Male: 249, Unknown: 25	Female: 373, M: 273, Unknown: 17
Patient Size (m)	1.67 ± 0.13 (1.44–2.0)	1.67 ± 0.12 (1.51–1.97)
Patient Weight (Kg)	69.4 ± 16.2 (48.7–164.0)	69.6 ± 15.9 (51.3–147.4)
Tube Current (mA)	130.6 ± 53.2 (26.4–514.1)	129.3 ± 45.4 (32.0–361.3)
PET acquisition time/bed (sec)	150 ± 49 (97–648)	152 ± 48 (75–658)
PET Reconstruction Method	OSEM3D + PSF + TOF 5i5s, OSEM3D + TOF 2i21s	OSEM3D + PSF + TOF 5i5s, OSEM3D + TOF 2i21s

figure 1 shows an example of all organs that can be segmented by both CT-nnU-Net and PET-nnU-Net models. However, the above-mentioned four anchor organs were selected for the next steps.

2.4. Metrics extraction

Two PET-based and CT-based segmentation masks were compared using common image segmentation evaluation metrics, including the Dice coefficient, Jaccard index, mean surface distance (MSD), Hausdorff distance (HD), and segment volume difference (mL) between the two segmentation masks. In other words, segmentation metrics were calculated for liver-PET vs liver-CT, spleen-PET vs spleen-CT, heart-PET vs heart-CT, and lungs-PET vs lungs-CT. In addition, the overlap between the lung-CT segmentation and liver-PET masks (Lung-CT/Liver-PET overlap volume) and overlap between lung-CT and spleen-PET masks (lung-CT/spleen-PET overlap volume) were extracted by implementing image processing algorithms representing the misalignment between PET and CT images. In total, 22 metrics were extracted and used in the following steps.

2.5. Feature selection and Random Forest machine learning in 10-fold cross-validation

A total of 1216 images were split randomly into cross-validation (80%, 972 cases) and test (20%, 224 cases) sets. The cross-validation set was used to train models in 10-fold data split, and the trained models were tested on 244 separate unseen test set. In the next step, an Analysis of Variance (ANOVA) feature selection method was used to sort the 22 features by importance and removing the less correlated features to prevent overfitting. It should be noted that feature selection was performed using only the cross-validation set to prevent data leakage. The ten most important features were selected and fed into the model in the next steps of training random forest machine learning models.

A random forest machine learning model was implemented using scikit-learn python library (Buitinck *et al* 2013) with the following training parameters: 10-fold random stratified data split, number of estimators equal to 100, maximum depth equal to 10. The minimum number of samples required to split an internal node equal to 1 and minimum number of samples required to be at a leaf node equal to 5. At every fold, 90% of the data were used as training whereas 10% were used as unseen test set. The whole data were used as test once during 10-fold cross-validation data split. Ten different models were trained and the training hyperparameters saved and used for the next step, i.e. testing on test set. All 10 models were inferenced on the unseen test-set and the decisions were ensembled by averaging the probabilities indicated by every 10 models. The decisions taken by the model trained on each fold were recorded and compared with the ensembled model.

One limitation of the suggested approach could be the lack of all four organs PET and CT segmentations, although the PET-nnU-Net and CT-nnU-Net models generate all four organs in a single inference, the end-user might have access to a limited number of segmented organs. To evaluate the performance of the proposed method using limited organ segmentation masks, we trained multiple RF models with the same hyperparameters using a single organ information in 6 various combinations i.e. input #1: lungs-CT vs lungs-PET metrics, input #2: liver-CT vs liver-PET metrics, input #3: spleen -CT vs spleen -PET metrics, input #4: heart -CT vs heart -PET metrics, input #5: liver-CT vs liver-PET metrics + lungs-CT vs lungs-PET metrics + lung-CT/liver-PET overlap volume, and input #6: spleen-CT vs spleen-PET metrics + lungs-CT vs lungs-PET metrics + lung-CT/spleen -PET overlap volume. Table 2 summarizes the adopted metrics.

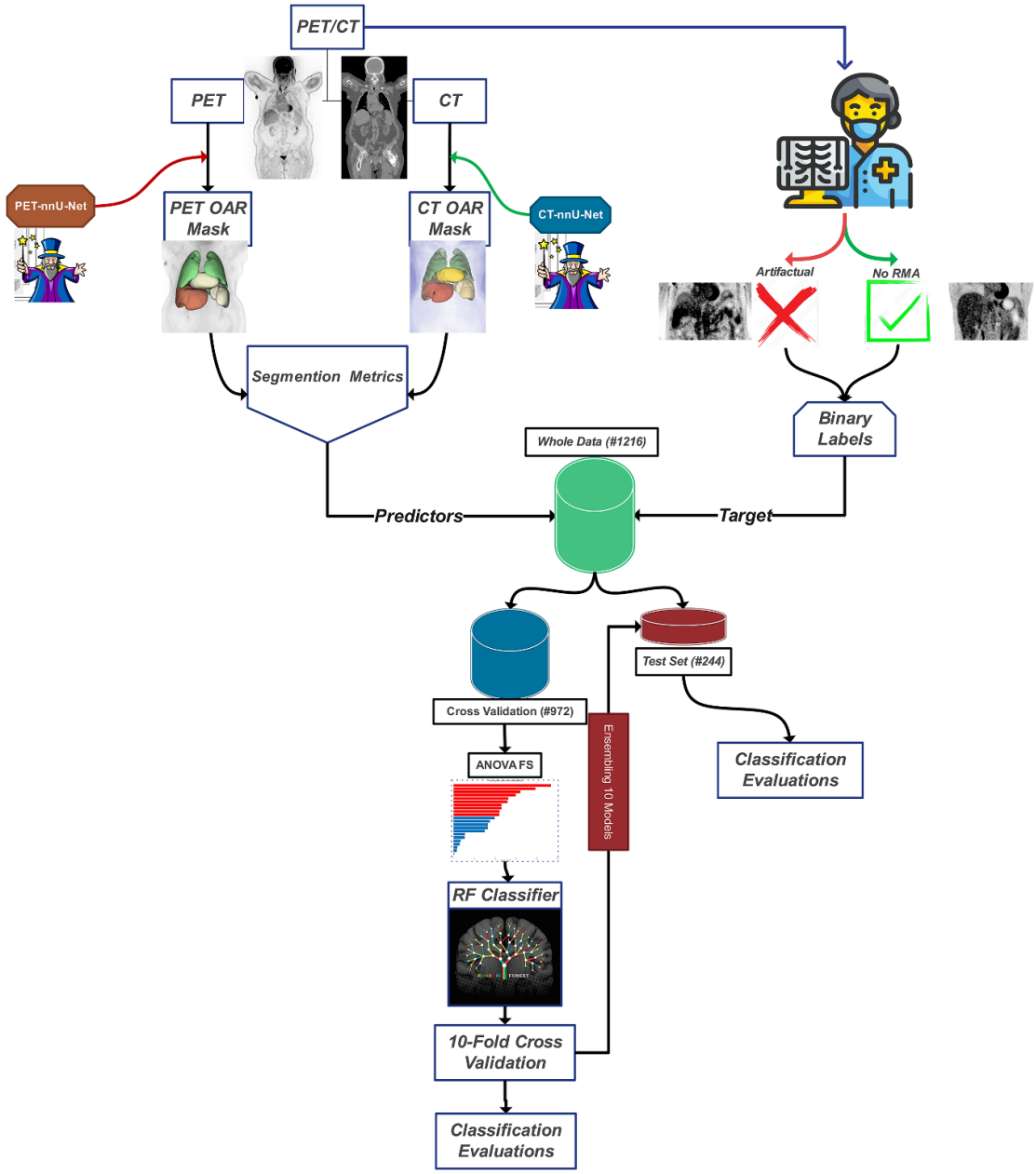

Figure 1. Workflow and processing steps followed in this study protocol.

Table 2. Metrics used for each included combination.

Number	Included metrics
Input #1	lungs-CT vs lungs-PET metrics
Input #2	liver-CT vs liver-PET metrics
Input #3	spleen-CT vs spleen-PET metrics
Input #4	heart-CT vs heart-PET metrics
Input #5	liver-CT vs liver-PET metrics + lungs-CT vs lungs-PET metrics + lung-CT/liver-PET overlap volume
Input #6	spleen-CT vs spleen-PET metrics + lungs-CT vs lungs-PET metrics + lung-CT/spleen-PET overlap volume

2.6. Evaluation strategy

The ML model decisions and probabilities were recorded and used to evaluate the performance of the model in terms of sensitivity, specificity, F1-score, accuracy, balanced accuracy (BAC, average of sensitivity and specificity) and area under the curve (AUC). Finally, confusion matrix and receiver operating curve (ROC curve) were drawn for model evaluation. The evaluation was performed for both cross-validation and test sets.

Table 3. Comparison of the 22 image segmentation metrics between two classes (No RMA vs RMA). MSD: Mean Surface Distance. HD: Hausdorff distance, VD: Volume difference.

Group	No RMA	RMA
Spleen Dice	0.852 ± 0.123	0.727 ± 0.208
Spleen Jaccard	0.756 ± 0.136	0.605 ± 0.212
Spleen MSD (mm)	3.6 ± 13.5	7.1 ± 20.2
Spleen HD (mm)	11.3 ± 25.5	20.7 ± 32.6
Spleen VD (mL)	4.8 ± 46.3	20.8 ± 60.2
Heart Dice	0.924 ± 0.042	0.887 ± 0.053
Heart Jaccard	0.861 ± 0.053	0.8 ± 0.077
Heart MSD (mm)	2.7 ± 4.7	3.8 ± 5.3
Heart HD (mm)	19.9 ± 79.6	23.9 ± 80.2
Heart VD (mL)	1.1 ± 47.9	−22.8 ± 67.2
Lungs Dice	0.934 ± 0.027	0.904 ± 0.038
Lungs Jaccard	0.877 ± 0.04	0.828 ± 0.058
Lungs MSD (mm)	1.7 ± 1.2	2.7 ± 1.5
Lungs HD (mm)	7.2 ± 21.1	13.6 ± 9.1
Lungs VD (mL)	−28.6 ± 165.8	−313.7 ± 336.4
Liver Dice	0.927 ± 0.034	0.862 ± 0.074
Liver Jaccard	0.865 ± 0.049	0.763 ± 0.102
Liver MSD (mm)	2.1 ± 1.4	4.2 ± 2.9
Liver HD (mm)	7.956 ± 7.4	17.15 ± 15.2
Liver VD (mL)	15.4 ± 118.9	142.9 ± 162.7
Lung-CT/Spleen-PET Overlap Volume (mL)	9.9 ± 16.6	39.3 ± 37.9
Lung-CT/Liver-PET Overlap Volume (mL)	35.5 ± 49.0	163.5 ± 118.8

3. Results

Of the 1216 images, 553 images (∼45%) belonged to No RMA class whereas the rest, i.e. 663 images (∼ 55%) belonged to the RMA class. Table 3 compares the measured 22 metrics between the two groups (RMA and No RMA).

Figure 2 shows an example of PET and CT segmentations for a case labeled as No RMA depicting a good alignment between the two segmentation masks showing excellent performance of both CT-nnU-Net and PET-nnU-Net models on the corresponding images.

Figure 3 shows an example of a PET/CT image labeled as RMA in visual assessment and the corresponding segmentations generated using PET and CT images as input to the models. Both models' performance was excellent and could detect the misalignment between two segmentations. The lung-CT/liver-PET overlap and lungs-CT/spleen-PET overlap segmentation visualization is also depicted in figure 3.

3.1. Feature selection results

ANOVA feature selection method sorted the features by importance and the ten most correlated features of lung-CT/liver-PET overlap volume, liver Jaccard, liver Dice, lungs' volume difference, lungs-CT/spleen-PET overlap, lungs Jaccard, liver MSD, heart Jaccard, liver volume difference, and lungs' Dice were selected for training the ML model.

Figure 4 shows the Gaussian fit to two selected metrics of heart HD and lungs-CT/liver-PET overlap volume showing the difference between the two groups (RMA and No RMA). As presented in figure 4, there is a good match between heart HD between the two groups, while lungs-CT/liver-PET overlap volume can separate the two groups. The remaining Gaussian fits for all 22 metrics are presented in supplementary figure 2.

3.2. Classification results

Cross-validation. Random forest F1_score, sensitivity, specificity, precision, accuracy, BAC, and AUC of 84.5, 83.9, 87.7, 85.1, 86.0, 85.8, and 93.2, respectively, on average were achieved over all 10-fold cross-validation data split. Figure 5 shows the confusion matrix and ROC curve for the RF model.

All classification results by details for 10-folds cross-validation are summarized in table 4 showing a consistent and robust performance of RF model over 10-folds. The cross-validation confusion matrix and ROC curve for every single fold can be found in supplementary figures 3 and 4, respectively.

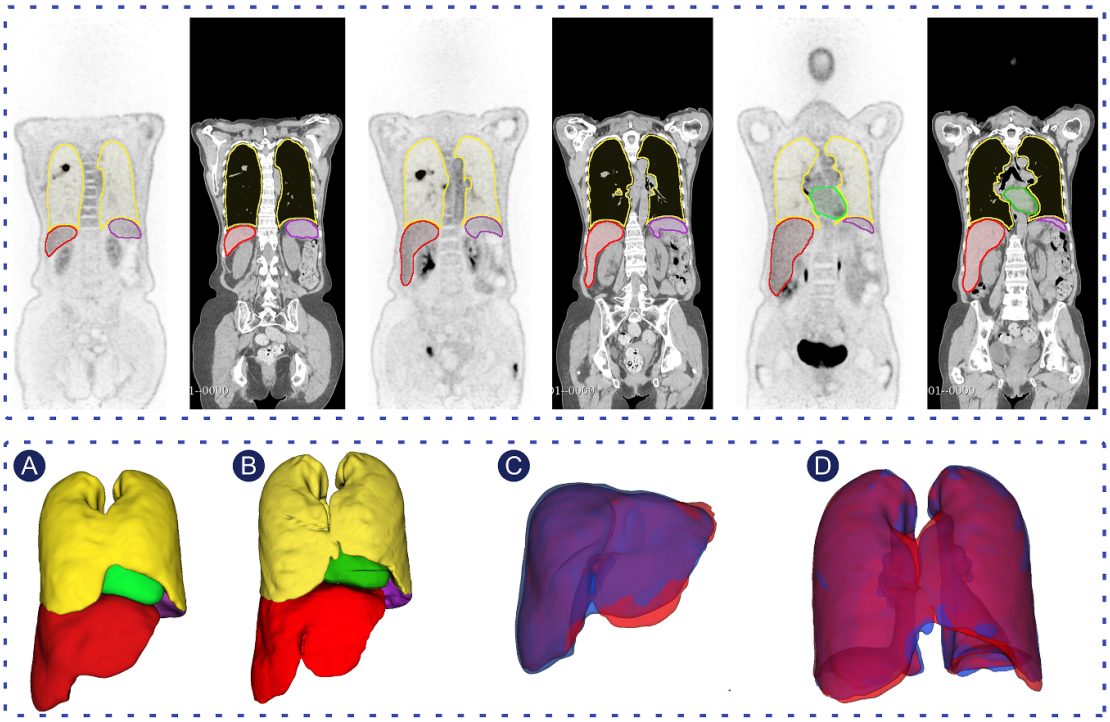

Figure 2. Representative PET/CT image labeled as No RMA. Top row: coronal slices of PET images and PET-nnU-NET segmentation output along with corresponding CT coronal slices and CT-nnU-Net segmentation output. Bottom row: (A) PET-nnU-NET 3D rendered image, (B) CT-nnU-Net 3D rendered image, (C) Liver-PET and Liver-CT 3D overlay rendered image, and (D) Lungs-PET and Lungs-CT 3D overlay rendered image.

3.3. Separate unseen test set evaluations

F1_score, sensitivity, specificity, precision, accuracy, BAC, and AUC on separate unseen dataset were 82.3, 83.8, 83.5, 80.9, 83.6, 83.6, and 90.1, respectively. Figure 6 presents the ROC curve and confusion matrix on the test set for the ensembled classification results. The detailed results of inference for each of the 10 models are summarized in table 5 showing that the performance of all folds were comparable while the ensembled decision specificity and sensitivity were more balanced with a higher BAC and closer sensitivity and specificity. The separate unseen test confusion matrix and ROC curves are presented in supplementary figures 5 and 6.

As reported by Nakamoto *et al* (2004), bowel motion could cause under/overestimation of SUV due to misalignment between PET and CTAC images. Figure 7 shows a PET/CT image labeled with severe respiratory mismatch between PET and CT images resulting from respiratory motion impacting the abdominal region. The abdominal organs including the colon and kidney have moved, and this movement can be captured by the approach we implemented to detect chest/abdomen interval motion.

Supplementary table 1 summarizes performance results of models using a limited number of segmentation masks available from input #1 to input #6 for both cross-validation and test set data. The highest accuracy was achieved using input #2 (liver segmentation on PET and CT) and input #5 (liver and lung segmentations on PET and CT) with accuracies of 83.5 and 85.1 on cross-validation set and 78.6 and 80.8 on the separate unseen test set.

Figure 8 shows another possible application the proposed approach to detect head motion and misalignment between PET and CT images.

4. Discussion

Misalignment between PET and CT images could cause attenuation and scatter correction errors, resulting in PET quantification bias through SUV under/overestimation, even missing lesions in the affected areas (Shiri *et al* 2023a). This misalignment could be due to either voluntary or involuntary respiratory, bowel, and cardiac movements. Detecting this kind of misalignment could be beneficial in clinical practice as well as in data curation approaches for large data management in retrospective machine learning and deep learning tasks. A high prevalence of respiratory misalignment was observed in our study, which is in agreement with previous studies (Shiri *et al* 2023b). Salimi *et al* (2024c) reported higher performance in ^{68}Ga-PSMA PET

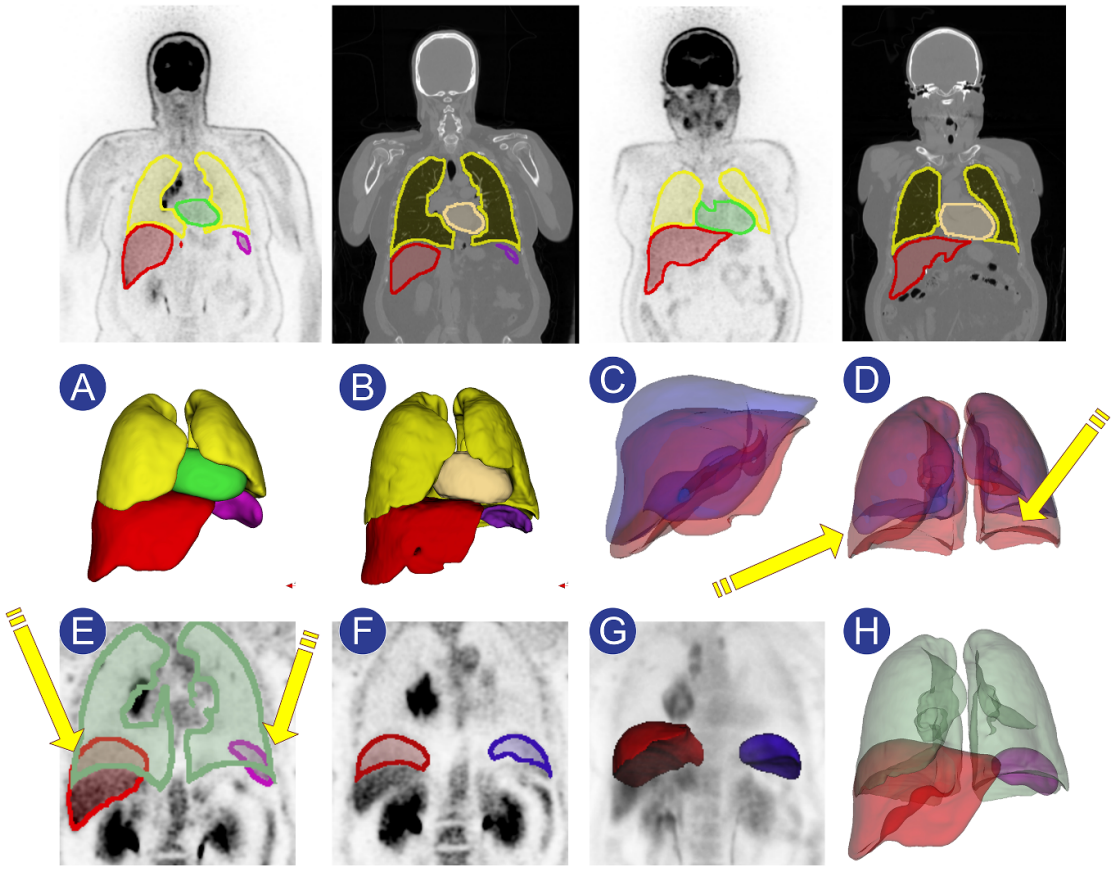

Figure 3. PET/CT and segmentation masks of an image with RMA. Top row: coronal slices of PET image and PET-nnU-Net output next to the CT image and CT-nnU-Net output. Bottom row: (A) PET-nnU-NET 3D rendered image, (B) CT-nnU-Net 3D rendered image, (C) Liver-PET and Liver-CT 3D overlay rendered image, and (D) lungs-PET and lungs-CT 3D overlay rendered image. The yellow arrows indicate area of misalignment between lungs-CT and lungs-PET segmentation masks. (E) Lung-CT (green) overlap mask with spleen-PET (dark blue) and liver-PET (res) segmentations. (F) Overlap visualization on PET image. (G) 3D rendered image of PET image with the overlap segmentations as mask. (H) Lung-CT, liver-PET, and spleen-PET 3D visualization.

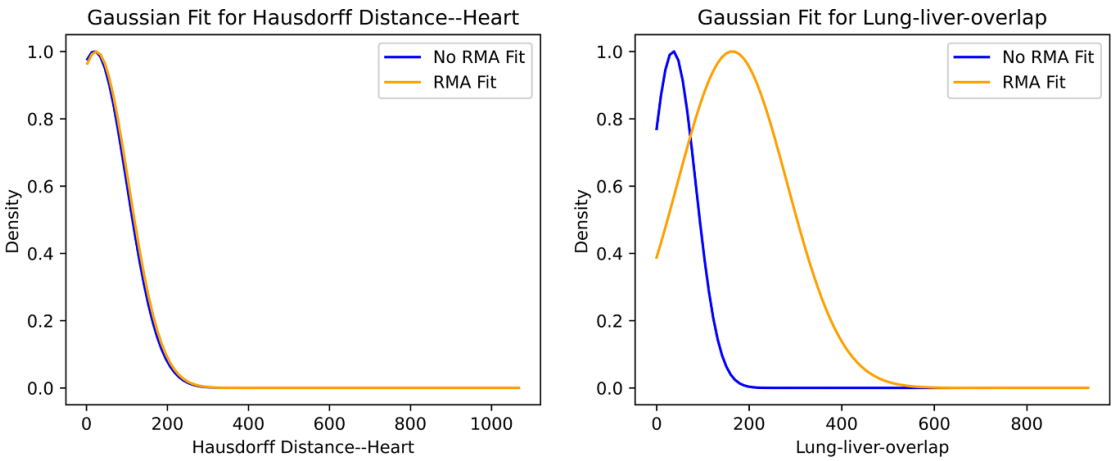

Figure 4. Example of Gaussian fit for two selected metrics: heart HD and lungs-CT/liver-PET overlap volume, demonstrating the ability of the latter as a good separator between the two groups.

organ segmentation compared to Yazdani *et al* (2024) study, likely due to excluding PET/CT studies presenting with misalignment from training data in the former study.

Respiratory motion artifact could be clinically significant when suspicious lesions are present in the diaphragmatic regions, within or adjacent to an artifact. Lesions, partially or wholly can be overlooked,

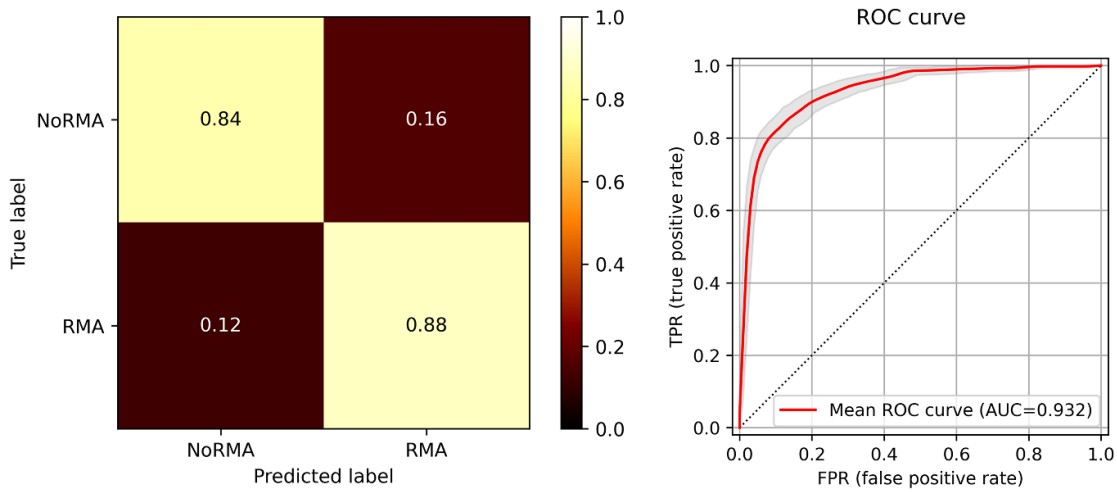

Figure 5. ROC curve and normalized confusion matrix for cross-validation data set.

Table 4. Summary of 10-fold cross-validation results for each fold.

Fold #	F1_score	Sensitivity	Specificity	Precision	Accuracy	BAC	AUC
0	84.8	86.7	84.9	83.0	85.7	85.8	92.7
1	78.2	75.6	84.9	81.0	80.6	80.2	90.1
2	82.8	81.8	86.8	83.7	84.5	84.3	91.8
3	82.2	84.1	83.0	80.4	83.5	83.6	93.7
4	80.9	81.8	83.0	80.0	82.5	82.4	90.3
5	88.6	88.6	90.6	88.6	89.7	89.6	94.8
6	85.4	79.5	94.3	92.1	87.6	86.9	93.1
7	84.1	84.1	86.8	84.1	85.6	85.4	94.9
8	88.4	86.4	92.5	90.5	89.7	89.4	96.2
9	89.9	90.9	90.6	88.9	90.7	90.7	95.0
Overall	84.5	83.9	87.7	85.1	86.0	85.8	93.2

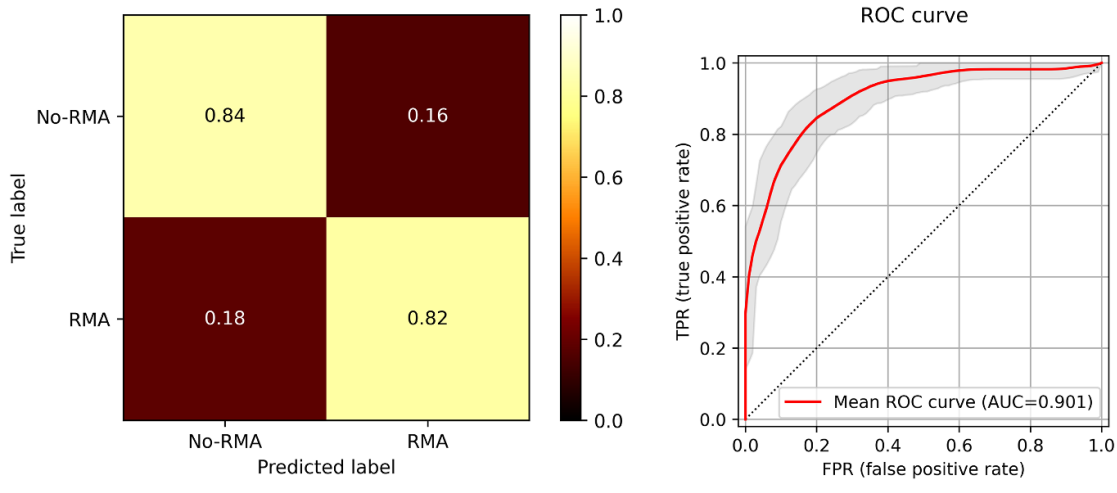

Figure 6. Confusion matrix and ROC curve for ensembled models on the 20% test set.

assigned to an incorrect organ (Blodgett *et al* 2011) or suffer from inaccurate quantification (McCall *et al* 2010, Geramifar *et al* 2013). Previous studies developed methodologies that can score overall image quality, reflecting a combination of the presence of image artifacts, noise and contrast patterns, as well as diagnostic confidence employing deep learning algorithms (Hopson *et al* 2020, Amini *et al* 2023, Qi *et al* 2023, Schwyzer *et al* 2023, Zhang *et al* 2023). They used PET images or processed PET volumes, such as maximum intensity projections (MIPs) as input to a deep learning classifier to assess image quality by nuclear medicine

Table 5. Detailed results of 10 models trained on 10-folds inferenced on the test set.

Fold #	F1_score	Sensitivity	Specificity	Precision	Accuracy	BAC	AUC
0	81.9	83.8	82.7	80.2	83.2	83.2	89.8
1	79.6	81.1	81.2	78.3	81.1	81.1	90.4
2	80.3	82.9	80.5	78.0	81.6	81.7	89.9
3	81.8	82.9	83.5	80.7	83.2	83.2	89.9
4	79.8	80.2	82.7	79.5	81.6	81.4	90.0
5	81.6	83.8	82.0	79.5	82.8	82.9	89.7
6	80.4	81.1	82.7	79.6	82.0	81.9	89.8
7	80.3	82.9	80.5	78.0	81.6	81.7	89.6
8	82.1	84.7	82.0	79.7	83.2	83.3	89.7
9	79.8	82.0	80.5	77.8	81.1	81.2	90.1
Ensembled	82.3	83.8	83.5	80.9	83.6	83.6	90.1

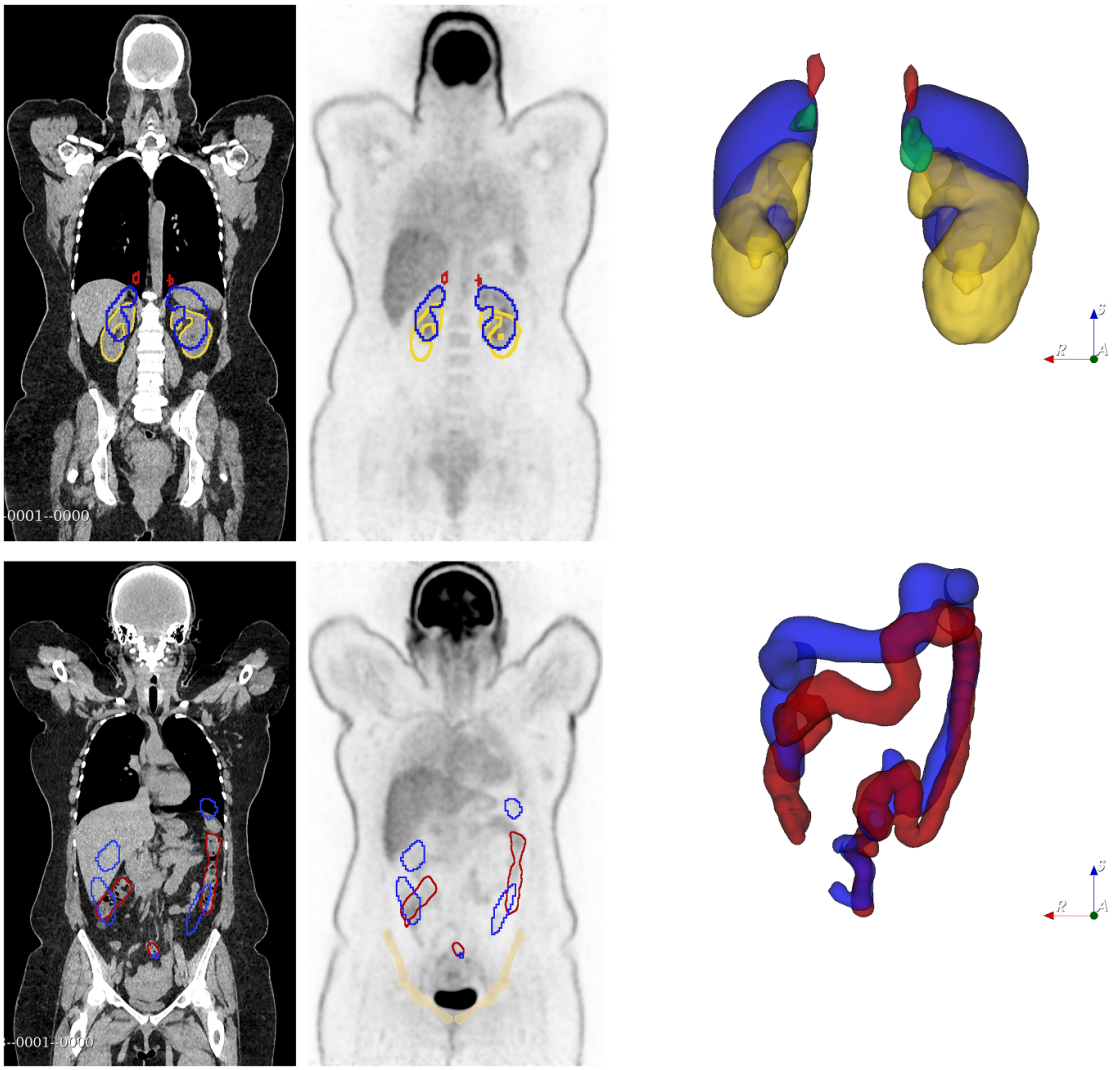

Figure 7. Coronal slices of CT and PET images as well as 3D rendered images. Top row: kidneys-CT (yellow) and kidneys-PET (blue) segmentation and the rendered 3D image showed the displacement of the kidneys between the two imaging modalities. Bottom row: Colon-CT (blue) and Colon-PET (dark red) segmentation overlay on coronal slices as well as 3D rendered display of organ segmentations depicting the changes in colon position.

physicians. However, the black box nature of deep learning makes the models less reliable and explainable. Our study focused on a single misalignment artifact using an explainable approach to detect the mismatch artifact in the chest/abdomen interval. We followed the same logical steps humans follow to detect misalignment between PET and CT images through delineation of four anchor moving organs in the thorax/abdomen region. Thanks to the very robust nnU-Net pipeline, we had access to CT-nnU-Net and

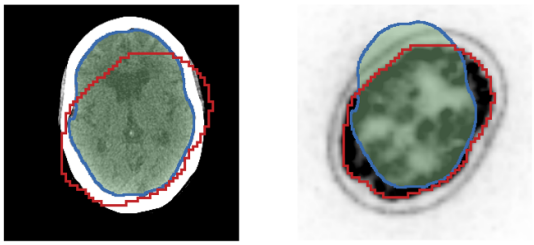

Figure 8. Brain PET/CT images. The patient moved his head between the two acquisitions. Blue: Brain-CT, Red: Brain-PET.

PET-nnU-NET models. However, none of the models is 100% accurate and we expect errors depending on the model and the selected organs. Dice coefficients of 92 vs 97, 82 vs 96, 93 vs 97, and 91 vs 94 were reported for the liver, spleen, lungs, and heart PET-nnU-Net vs CT-nnU-Net, respectively. CT-nnU-Net (Salimi *et al* 2023b) model performance was superior to PET-nnU-Net model (Salimi *et al* 2024c) as reported earlier, and as such, the errors in both model segmentations were expected to overlap between the two groups (RMA and No RMA) as shown in figure 4 and supplementary figure 2. The Dice coefficient between organ-CT and organ-PET segmentation masks is due to errors in CT-nnU-Net and PET-nnU-Net models and the misalignment between PET and CT images. Although there is significant difference between the values summarized in table 3 according to Wilcoxon test (p-value < 0.05), a model trained using only single organ segmentation metrics underperformed the final model using the metrics selected after features selection on all 22 metrics. The best test set accuracy using only a single organ information was the use of liver information resulting in 78.3 accuracy. Adding lungs segmentation improved the accuracy to 80.7 which is still lower than using the 10 selected metrics by ANOVA test. To tackle this limitation, we included multiple metrics extracted from four anchor moving organs as well as the lungs-CT/liver-PET and lungs-CT/spleen-PET overlap volume to improve the decision-making performance. We used a random forest machine learning model to classify images into two groups using the selected metrics in a 10-fold data split strategy to eliminate the randomness in data split and tested the trained models on 20% of data as test set. We ensembled all 10 trained models on the test set to achieve more robust results on a separate unseen case and prevent wrong decisions by an overfitted model in a single fold. This method can be improved by providing a more robust segmentation model on both PET and CT imaging modalities and larger multi-centric studies. In terms of performance, our model performed reasonably well in both cross-validation and separate unseen test sets with accuracies higher than 83% and an AUC more than 90%.

Misalignment could cause attenuation and scatter correction errors. Besides, organ segmentations generated by CT images as a common automated organ segmentation approach in hybrid imaging are not well aligned with the real molecular information on PET images in case of mismatch, which can cause errors in time activity curve calculations and estimation of time integrated activities and organ absorbed doses for personalized radiation dosimetry, not only in the chest/abdomen interval and in moving abdominal regions as shown in figure 7. The proposed methodology could be used to detect misalignment in other organs, such as the colon and brain. Involuntary movement was reported for lung, spleen, heart, and liver up to 22 mm (Giraud *et al* 2001, Clifford *et al* 2002, Harada *et al* 2002, McLeish *et al* 2002, Allen *et al* 2004, Brandner *et al* 2006), the pancreas (Feng *et al* 2009), and kidneys (Brandner *et al* 2006, Yamashita *et al* 2014). We have tested our proposed methodology only on ^{18}F-FDG images emanating from a single clinical center. The applicability of this pipeline should be evaluated on other compounds and multicentric datasets.

One of the limitations of our study is that we considered visual labeling, which is the only available option in clinical setting, as the reference label for presence of artifacts. However, our approach provides access to volume changes in liver, spleen, and lung as a quantitative metric to evaluate the extent of misalignment as well as the segmentation mask, showing differences between lungs-PET and lungs-CT masks as presented in figures 3(C)–(E). The user might check the segmentation mask showing the changes in size, position, and shape of organs in PET and CT images.

5. Conclusion

We developed a fully automated explainable methodology to detect misalignment between PET and CT images in the chest/abdomen interval region using a multi-step deep learning-based segmentation and random forest machine learning pipeline that can be used in clinical setting to identify cases suffering from misalignment artifacts and in retrospective data curation on large datasets for deep learning applications.

The advantage of our approach is the use of explainable, easy to compute, and clinically relevant information such as volume change and Dice coefficient in the decision-making process. More accurate and robust PET an CT organ segmentation models would enhance the reliability of the proposed approach.

Data availability statement

The data used in this work is not available. All trained models are available on GitHub at: https://github.com/YazdanSalimi/PETCT-RMA-Detection.

The data cannot be made publicly available upon publication because no suitable repository exists for hosting data in this field of study. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgment

This work was supported by the Euratom research and training Programme 2019–2020 Sinfonia project under Grant Agreement No. 945196 and the Swiss National Science Foundation under Grant SNSF 320030-231742.

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This retrospective study was approved by the institutional ethics committee (CCER ID: 2023-00322) and the requirement to obtain informed consent was waived.

Consent to participate

Informed consent was waived in this retrospective study.

Conflict of interest

None of the authors have affiliations that present financial or non-financial competing interests for this work.

ORCID iDs

Yazdan Salimi https://orcid.org/0000-0002-1233-9576
Zahra Mansouri https://orcid.org/0000-0003-2087-9721
Mehdi Amini https://orcid.org/0000-0002-4370-680X
Habib Zaidi https://orcid.org/0000-0001-7559-5297

References

Allen A M, Siracuse K M, Hayman J A and Balter J M 2004 Evaluation of the influence of breathing on the movement and modeling of lung tumors *Int. J. Radiat. Oncol. Biol. Phys.* **58** 1251–7

Amann J, Blasimme A, Vayena E, Frey D and Madai V I 2020 Explainability for artificial intelligence in healthcare: a multidisciplinary perspective *BMC Med. Inform. Decis. Mak.* **20** 310

Amini M, Salimi Y, Sabouri M, Hajianfar G, Sanaat A, Hervier E, Mainta I, Rahmim A, Shiri I and Zaidi H 2023 *2023 IEEE Nuclear Science Symp., Medical Imaging Conf. and Int. Symp. on Room-Temperature Semiconductor Detectors (NSS MIC RTSD) (4–11 November 2023)* p 1

Beyer T, Tellmann L, Nickel I and Pietrzyk U 2005 On the use of positioning aids to reduce misregistration in the head and neck in whole-body PET/CT studies *J. Nucl. Med.* **46** 596–602

Blodgett T M, Mehta A S, Mehta A S, Laymon C M, Carney J and Townsend D W 2011 Pet/ct artifacts *Clin. Imaging* **35** 49–63

Brandner E D, Wu A, Chen H, Heron D, Kalnicki S, Komanduri K, Gerszten K, Burton S, Ahmed I and Shou Z 2006 Abdominal organ motion measured using 4D CT *Int. J. Radiat. Oncol. Biol. Phys.* **65** 554–60

Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A and Grobler J 2013 API design for machine learning software: experiences from the scikit-learn project (arXiv:1309.0238)

Clifford M A, Banovac F, Levy E and Cleary K 2002 Assessment of hepatic motion secondary to respiration for computer assisted interventions *Comput. Aided Surg.* **7** 291–9

Cook G J, Wegner E A and Fogelman I 2004 *Seminars in Nuclear Medicine* vol 34 (Elsevier) pp 122–33

Czernin J, Allen-Auerbach M and Schelbert H R 2007 Improvements in cancer staging with PET/CT: literature-based evidence as of september 2006 *J. Nucl. Med.* **48** 78S–88

Dinelle K, Blinder S, Cheng J-C, Lidstone S, Buckley K, Ruth T J and Sossi V 2006 *2006 IEEE Nuclear Science Symp. Conf. Record* vol 6 (IEEE) pp 3283–7

Erdi Y E *et al* 2004 The CT motion quantitation of lung lesions and its impact on PET-measured SUVs *J. Nucl. Med.* **45** 1287–92
Feng Y, Liu X, Ma J, Lu Z and Chen W 2009 *2009 Annual Int. Conf. IEEE Engineering in Medicine and Biology Society* vol (IEEE) pp 2680–3
Geramifar P, Zafarghandi M S, Ghafarian P, Rahmim A and Ay M R 2013 Respiratory-induced errors in tumor quantification and delineation in CT attenuation-corrected PET images: effects of tumor size, tumor location, and respiratory trace: a simulation study using the 4D XCAT phantom *Mol. Imaging Biol.* **15** 655–65
Giraud P *et al* 2001 Conformal radiotherapy (CRT) planning for lung cancer: analysis of intrathoracic organ motion during extreme phases of breathing *Int. J. Radiat. Oncol. Biol. Phys.* **51** 1081–92
Gould K L, Pan T, Loghin C, Johnson N P, Guha A and Sdringola S 2007 Frequent diagnostic errors in cardiac PET/CT due to misregistration of CT attenuation and emission PET images: a definitive analysis of causes, consequences, and corrections *J. Nucl. Med.* **48** 1112–21
Gu S, McNamara J E, Mitra J, Gifford H C, Johnson K, Gennert M A and King M A 2010 Body deformation correction for SPECT imaging *IEEE Trans. Nucl. Sci.* **57** 214–24
Harada T, Shirato H, Ogura S, Oizumi S, Yamazaki K, Shimizu S, Onimaru R, Miyasaka K, Nishimura M and Dosaka-Akita H 2002 Real-time tumor-tracking radiation therapy for lung carcinoma by the aid of insertion of a gold marker using bronchofiberscopy *Cancer* **95** 1720–7
Hopson J B, Neji R, Reader A J and Hammers A 2020 *2020 IEEE Nuclear Science Symp. and Medical Imaging Conf. (NSS/MIC)* (*31 October–7 November 2020*) pp 1–3
Isensee F, Jaeger P F, Kohl S A A, Petersen J and Maier-Hein K H 2021 nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation *Nat. Methods* **18** 203–11
Kinahan P E, Townsend D, Beyer T and Sashin D 1998 Attenuation correction for a combined 3D PET/CT scanner *Med. Phys.* **25** 2046–53
Kundu S 2021 AI in medicine must be explainable *Nat. Med.* **27** 1328
Kyme A Z and Fulton R R 2021 Motion estimation and correction in SPECT, PET and CT *Phys. Med. Biol.* **66** 18TR02
Lamare F, Bousse A, Thielemans K, Liu C, Merlin T, Fayad H and Visvikis D 2022 PET respiratory motion correction: quo vadis? *Phys. Med. Biol.* **67** 03TR2
Lamare F, Le Maitre A, Dawood M, Schäfers K, Fernandez P, Rimoldi O and Visvikis D 2014 Evaluation of respiratory and cardiac motion correction schemes in dual gated PET/CT cardiac imaging *Med. Phys.* **41** 072504
Langlotz C P *et al* 2019 A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/the academy workshop *Radiology* **291** 781–91
Maldonado A, González-Alenda F, Alonso M and Sierra J 2007 PET-CT in clinical oncology *Clin. Transl. Oncol.* **9** 494–505
Mayer-Schönberger V and Ingelsson E 2018 Big data and medicine: a big deal? *J. Intern. Med.* **283** 418–29
McCall K C, Barbee D L, Kissick M W and Jeraj R 2010 PET imaging for the quantification of biologically heterogeneous tumours: measuring the effect of relative position on image-based quantification of dose-painting targets *Phys. Med. Biol.* **55** 2789
McLeish K, Hill D L, Atkinson D, Blackall J M and Razavi R 2002 A study of the motion and deformation of the heart due to respiration *IEEE Trans. Med. Imaging* **21** 1142–50
Montgomery A J, Thielemans K, Mehta M A, Turkheimer F, Mustafovic S and Grasby P M 2006 Correction of head movement on PET studies: comparison of methods *J. Nucl. Med.* **47** 1936–44
Nakamoto Y, Chin B B, Cohade C, Osman M, Tatsumi M and Wahl R L 2004 PET/CT: artifacts caused by bowel motion *Nucl. Med. Commun.* **25** 221–5
Park S H and Han K 2018 Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction *Radiology* **286** 800–9
Pevsner A, Nehmeh S A, Humm J L, Mageras G S and Erdi Y E 2005 Effect of motion on tracer activity determination in CT attenuation corrected PET images: a lung phantom study *Med. Phys.* **32** 2358–62
Qi C, Wang S, Yu H, Zhang Y, Hu P, Tan H, Shi Y and Shi H 2023 An artificial intelligence-driven image quality assessment system for whole-body [(18)F]FDG PET/CT *Eur. J. Nucl. Med. Mol. Imaging* **50** 1318–28
Reddy S 2022 Explainability and artificial intelligence in medicine *Lancet Digit. Health* **4** e214–e5
Redman T C 2018 If your data is bad, your machine learning tools are useless *Harv. Bus. Rev.* **2** (available at: https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless)
Salimi Y, Hajianfar G, Mansouri Z, Sanaat Y, Amini M, Shiri I and Zaidi H 2024a Organomics: a novel concept reflecting the importance of PET/CT healthy organ radiomics in non-small cell lung cancer prognosis prediction using machine learning *Clin. Nucl. Med.* **49** 899–908
Salimi Y, Mansouri Z, Hajianfar G, Sanaat A, Shiri I and Zaidi H 2024b Fully automated explainable abdominal CT contrast media phase classification using organ segmentation and machine learning *Med. Phys.* **51** 4095–104
Salimi Y, Mansouri Z, Shiri I, Mainta I and Zaidi H 2024c Deep learning-powered CT-less multi-tracer organ segmentation from PET images: a solution for unreliable CT segmentation in PET/CT imaging *medRxiv* (https://doi.org/10.1101/2024.08.27.24312482)
Salimi Y, Shiri I, Akavanallaf A, Mansouri Z, Arabi H and Zaidi H 2023a Fully automated accurate patient positioning in computed tomography using anterior-posterior localizer images and a deep neural network: a dual-center study *Eur. Radiol.* **33** 3243–52
Salimi Y, Shiri I, Mansouri Z and Zaidi H 2023b Deep learning-assisted multiple organ segmentation from whole-body CT images *medRxiv* (https://doi.org/10.1101/2023.10.20.23297331)
Sanaat A, Shiri I, Arabi H, Mainta I, Nkoulou R and Zaidi H 2021 Deep learning-assisted ultra-fast/low-dose whole-body PET/CT imaging *Eur. J. Nucl. Med. Mol. Imaging* **48** 2405–15
Schwyzer M, Skawran S, Gennari A G, Waelti S L, Walter J E, Curioni-Fontecedro A, Hofbauer M, Maurer A, Huellner M W and Messerli M 2023 Automated F18-FDG PET/CT image quality assessment using deep neural networks on a latest 6-ring digital detector system *Sci. Rep.* **13** 11332
Shiri I *et al* 2023b Differential privacy preserved federated transfer learning for multi-institutional (68)Ga-PET image artefact detection and disentanglement *Eur. J. Nucl. Med. Mol. Imaging* **51** 40–53
Shiri I, Salimi Y, Hervier E, Pezzoni A, Sanaat A, Mostafaei S, Rahmim A, Mainta I and Zaidi H 2023a Artificial intelligence-driven single-shot PET image artifact detection and disentanglement: toward routine clinical image quality assurance *Clin. Nucl. Med.* **48** 1035–46
Soffer S, Ben-Cohen A, Shimon O, Amitai M M, Greenspan H and Klang E 2019 Convolutional neural networks for radiologic images: a radiologist's guide *Radiology* **290** 590–606
Sun T and Mok G S 2012 Techniques for respiration-induced artifacts reductions in thoracic PET/CT *Quant. Imaging Med. Surg.* **2** 46

van Ooijen P M 2019 Quality and curation of medical images and data *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks* pp 247–55

Visvikis D, Lamare F, Bruyant P, Boussion N and Le Rest C C 2006 Respiratory motion in positron emission tomography for oncology applications: problems and solutions *Nucl. Instrum. Methods Phys. Res.* A **569** 453–7

Willemink M J, Koszek W A, Hardell C, Wu J, Fleischmann D, Harvey H, Folio L R, Summers R M, Rubin D L and Lungren M P 2020 Preparing medical imaging data for machine learning *Radiology* **295** 4–15

Xu Q, Yuan K and Ye D 2011 Respiratory motion blur identification and reduction in ungated thoracic PET imaging *Phys. Med. Biol.* **56** 4481

Yamashita H *et al* 2014 Individually wide range of renal motion evaluated by four-dimensional computed tomography *Springerplus* **3** 1–7

Yang J, Sohn J H, Behr S C, Gullberg G T and Seo Y 2020 CT-less direct correction of attenuation and scatter in the image space using deep learning for whole-body FDG PET: potential benefits and pitfalls *Radiology* **3** e200137

Yazdani E, Karamzadeh-Ziarati N, Cheshmi S S, Sadeghi M, Geramifar P, Vosoughi H, Jahromi M K and Kheradpisheh S R 2024 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* **24** 30

Yoon C H, Torrance R and Scheinerman N 2022 Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? *J. Med. Ethics* **48** 581–5

Zaidi H and El Naqa I 2021 Quantitative molecular positron emission tomography imaging using advanced deep learning techniques *Annu. Rev. Biomed. Eng.* **23** 249–76

Zhang H, Liu Y, Wang Y, Ma Y, Niu N, Jing H and Huo L 2023 Deep learning model for automatic image quality assessment in PET *BMC Med. Imaging* **23** 75