A. Measurements of Reference TMTV and Dmax

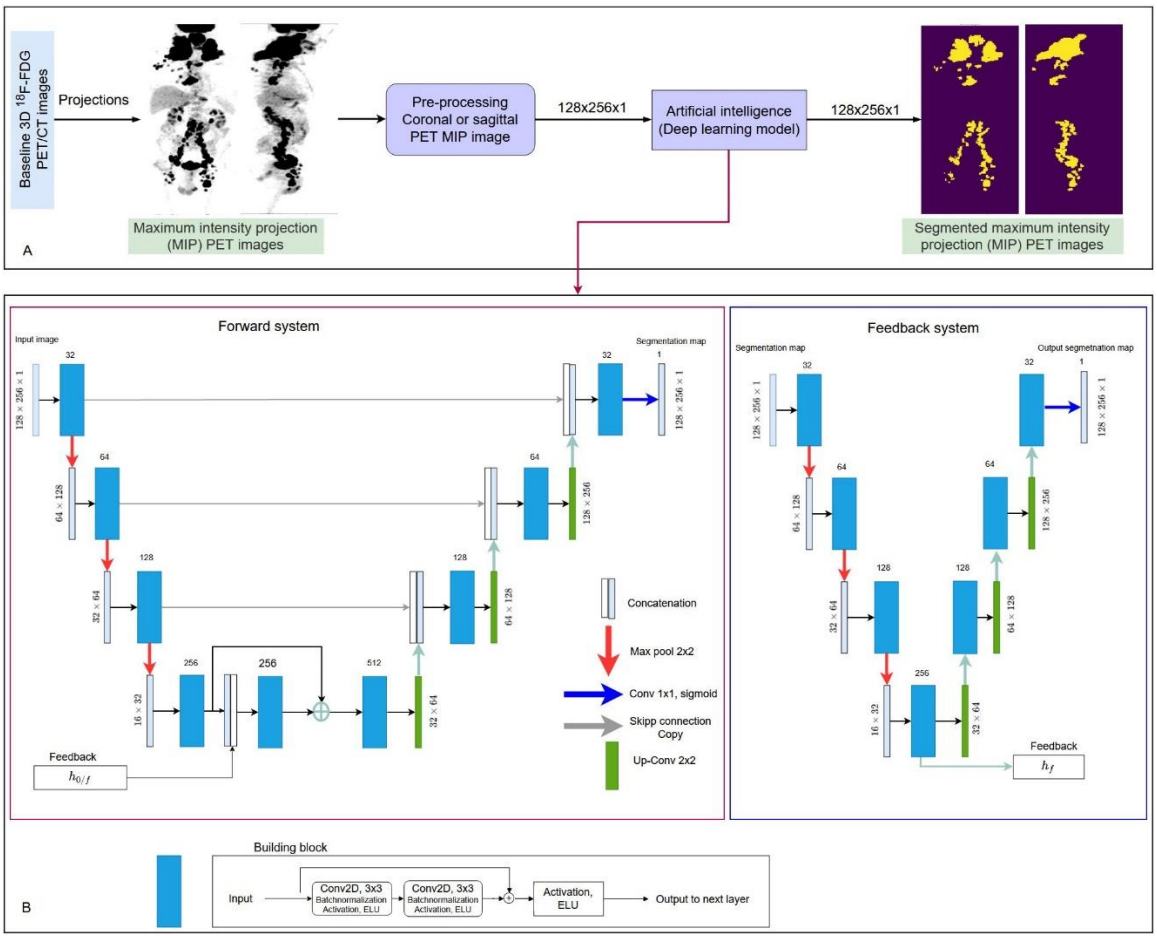
For the REMARC cohort, the lymphoma regions were automatically identified in the 3D PET images as described in (*1*,*2*). A SUVmax 41% threshold segmentation was then applied on these regions, and the results were visually checked by an expert nuclear medicine physician to exclude physiological lesions and to manually add missed lesions whenever needed as described in (*3*).

The LNH073B data were processed by a nuclear medicine physician using the LIFEx software (*4*): hypermetabolic regions were first automatically detected by selecting all voxels with an SUV greater than 2 included in a region greater than 2 mL, and a 41% SUVmax thresholding of the resulting regions was used. Like in the REMARC cohort, the expert removed the regions corresponding to physiological uptakes and added pathological regions missed by the algorithm.

For both cohorts, the physicians were blinded to the patient outcomes. The 3D lymphoma regions validated by experts were used to compute the baseline TMTV and Dmax (based on the centroid of the lymphoma regions) (*5*).

B. Final Network Architecture and Training

The deep learning model was trained from the REMARC data using a five-fold cross-validation technique. It was then tested on another independent cohort, LNH073B. The architecture of the network was inspired by (*6*). The model consists of an encoder and a decoder network with a skipped connection between the two paths and external fully connected network-based feedback. Lymphoma regions are often scattered over the whole body, and information could easily be lost in the successive convolution and pooling operations. To alleviate this scenario, we have used residual CNN as a building block (*7*) in all encoder and decoder components of the deep learning model (Figure 1). It can ease training and facilitate information propagation from input to the output of the network architecture. The input and output dimensions of the network were 128×256×1.



Supplemental Figure 1. Components of the proposed convolutional neural network (CNN). A) Overview of the deep learning model, inputs, and outputs. The coronal or sagittal PET MIP images are provided as independent inputs to the deep learning model. The corresponding segmented regions having the same size as the input image are the output. B) Deep learning model architecture. The building block is the convolutional building block of the deep learning model. Each 2D CNN (Conv2D) with a kernel size of 3x3 was followed by batch normalization and activation function. We have used the exponential linear unit (ELU) activation function, except it was a sigmoid activation function at the output layers. After the convolutional building block in the encoder, we applied a 2x2 max pooling operation with stride 2 for downsampling. Before the convolutional building block, we used a 2x2 up-convolutional layer in the decoder. The deep learning model will be publicly available upon publication [GitHub].

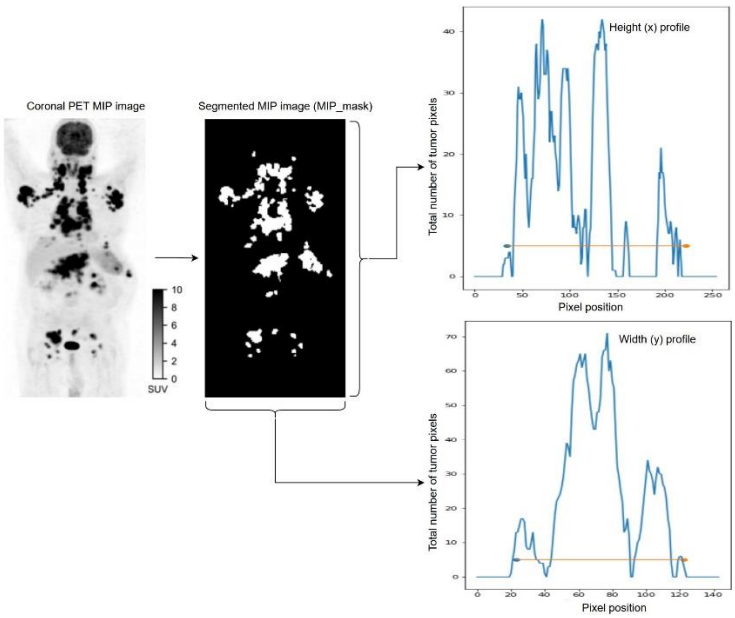
Pre-processing. All available 3D PET images and the corresponding expert-validated 3D lymphoma segmented regions were resized in to 4 x 4 x 4 mm^3 voxel size. The resized 3D images were then padded

or cropped to fit into a 128x128x256. The resized and cropped image were projected into sagittal and coronal views. The input and output image dimensions to the network were 128x256x1.

Training. The model was trained with a batch size of 32 for 1000 epochs and 300 early stop criteria. Different augmentation techniques, including flipping and rotation, were considered and tested but did not improve the results, so we did not use any data augmentation to produce the final model. The deep learning model neural network weights were updated using a stochastic gradient descent algorithm, ADAM optimizer (*8*), with a learning rate of 1e-4. All other parameters were Keras default values. A sigmoid output activation function was used to binarize the image into the lymphoma region and non-lymphoma region. We used the average of the Dice similarity coefficient ($Loss_{Dice}$) and binary cross-entropy ($Loss_{binary\ cross\text{-}entropy}$) as a loss function defined by:

$$loss = 1/2\ (\ Loss_{binary\ cross-entropy} + Loss_{Dice})$$

The model was implemented with Python, Keras API, and Tensorflow backend. The data was processed using the Python 3.8.5 package, including Numpy, Scipy, Pandas, and Matplotlib. We did not apply any post-processing method for the segmentation metrics. To compute the surrogate biomarkers from the AI-based segmented images, regions with less than 4.8 cm^2 were removed. The deep learning model will be publicly available upon publication at [GitHub].



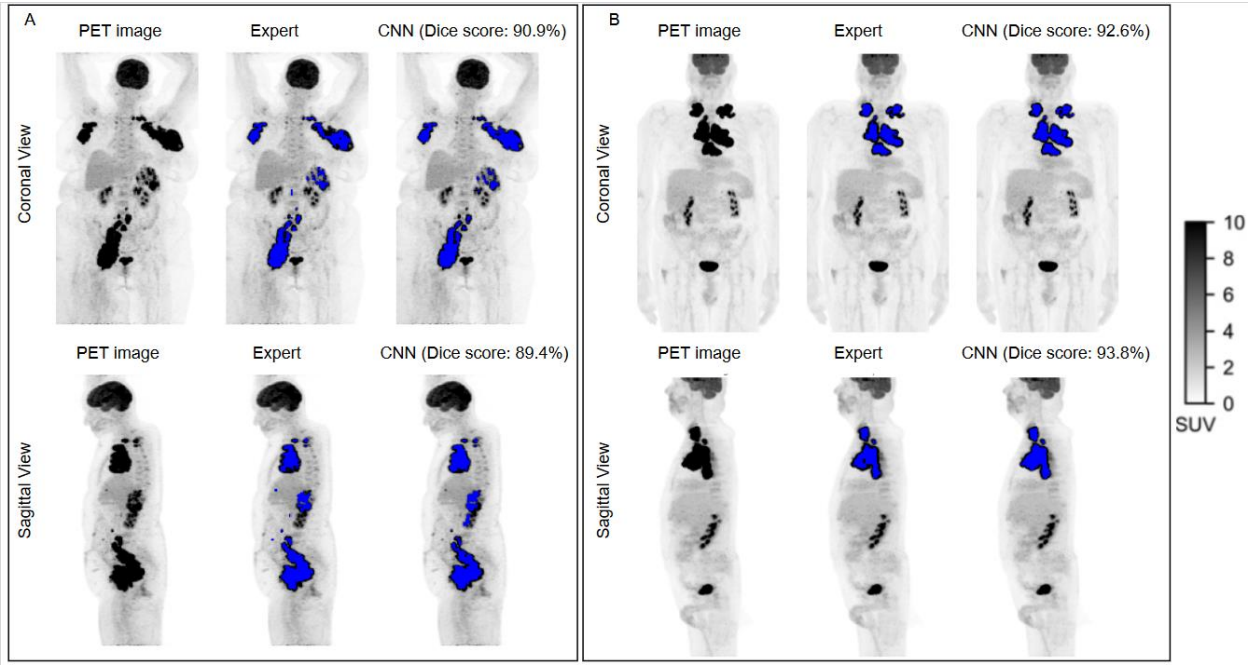
Supplemental Figure 2. Illustration of the calculation of the tumor dissemination feature. For the given PET MIP image, we created two profiles corresponding to the sum of the signal in the x and y directions, respectively. The horizontal line shows the distances between the 2% percentiles and the 98% percentiles. It was the same for the sagittal PET MIP image. Pixel positions with zero total number of tumor pixels (often at the beginning and end of the pixel positions) are not considered for the percentile calculation.

C. Statistical Analysis Details

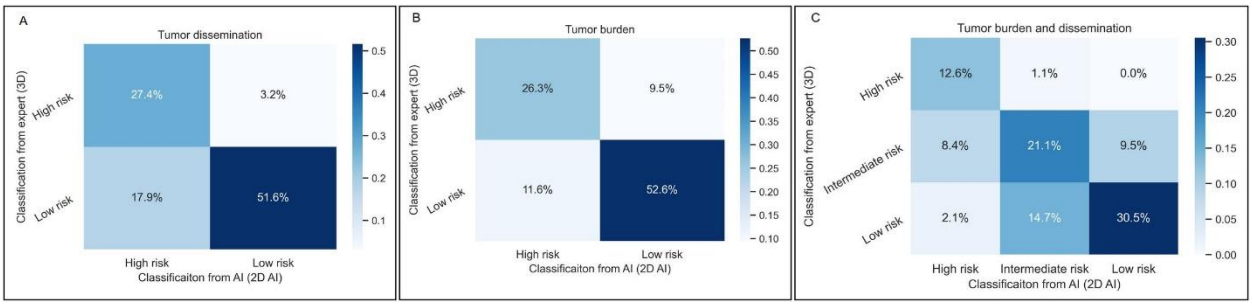
Univariate Analysis. For all biomarkers, we calculated a time-dependent area under the receiver operating characteristics curve (AUC) (*9*). Bootstrap resampling analysis was performed to associate confidence intervals to the Cox model hazard ratio and the time-dependent AUC. The bootstrapping involved 10,000 random samplings of the data with replacement. All statistical comparisons, except the Kaplan-Meier analysis, were made without discretizing the continuous values.

Multivariate Analysis. We estimated the survival functions using Kaplar-Meier estimates. For each PET-derived feature, we selected the optimal cut-off values for PFS and OS at the values that yielded the smallest P-value in the log-rank test between categories of a given study population. The cut-off values were constrained to be between the interquartile ranges of the TMTV or Dmax values. This procedure was the same for all measurements, namely for the 3D ^{18}F-FDG PET-based biomarkers (TMTV and Dmax) and

PET MIP-based biomarkers from the deep learning method (sTMTV and sDmax). A receiver-operating-characteristics (ROC) analysis was also used to define the optimal cut-off values that predict the occurrence of an event (progression-free survival or overall survival) by maximizing the sensitivity plus specificity minus one (i.e., sensitivity + specificity -1). It yielded nearly the same results as calculating the cut-off values using the log-rank test approach. For the TMTV, we obtained a cut-off value of 222 cm^3, which is close to the published values of 220 cm^3 (*1*). For uniformity of the comparison of the 2D and 3D PET features, we followed the same procedures for all features to compute the cut-off values.



Supplemental Figure 3. ^{18}F-FDG PET MIP images and segmentation results (blue color overlapped over the PET MIP images) by experts (MIP_masks) and by the CNN for four patients: (A) from the REMARC cohort, and (B) from the LNH073B cohort.



Supplemental Figure 4. Confusion matrices for classification of patients using PET features derived from using the expert-delineated 3D ^{18}F-FDG PET images (3D-expert) and from using the 2D PET MIP images using CNN (2D-AI) on LNH073B cohort. A) Two-risk-group classification using Dmax and sDmax, B) two-risk-group classification using TMTV and sTMTV, and C) three-risk-group classification using TMTV and Dmax (3D-expert), and sTMTV and sDmax (CNN).

References

1. Vercellino L, Cottereau AS, Casasnovas O, et al. High total metabolic tumor volume at baseline predicts survival independent of response to therapy. *Blood*. 2020;135:1396-1405.
2. Capobianco N, Meignan M, Cottereau A-S, et al. Deep-learning 18 F-FDG uptake classification enables total metabolic tumor volume estimation in diffuse large B-Cell lymphoma. *J Nucl Med*. 2021;62:30-36.
3. Cottereau A-S, Nioche C, Dirand A-S, et al. 18F-FDG PET dissemination features in diffuse large B-cell lymphoma are predictive of outcome. *J Nucl Med*. 2020;61:40-45.
4. Nioche C, Orlhac F, Boughdad S, et al. Lifex: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer Res*. 2018;78:4786-4789.
5. Cottereau A-S, Meignan M, Nioche C, et al. Risk stratification in diffuse large B-cell lymphoma using lesion dissemination and metabolic tumor burden calculated from baseline PET/CT†. *Ann Oncol*. 2021;32:404-411.
6. Girum KB, Crehange G, Lalande A. Learning with context feedback loop for robust medical image segmentation. *IEEE Trans Med Imaging*. 2021;40:1542-1554.
7. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016:770-778.
8. Kingma DP, Ba J. Adam: A method for stochastic optimization. *3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc*. December 2014:1-15.
9. Heagerty PJ, Lumley T, Pepe MS. Time-Dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. 2000;56:337-344.