Detection of Transient Neurotransmitter Response Using Personalized Neural Networks

Ivan S. Klyuzhin, Connor W. J. Bevington, Ju-Chieh (Kevin) Cheng, and Vesna Sossi

***Abstract*—Measurement of transient neurotransmitter response (TNR), such as stimulus-related dopamine release, from dynamic positron emission tomography (PET) images typically suffers from limited detection sensitivity and/or high false positive rates. In this work, we perform voxel-level TNR detection using artificial neural networks (ANN), and compare their performance to previously used standard statistical tests. We use acquired and simulated [^{11}C]raclopride (a D2 receptor antagonist) human data to train and test different ANN architectures for the task of dopamine release detection. A distinguishing feature of our approach is the use of "personalized" ANNs that are designed to operate on the dynamic image from a specific subject and scan. Training of personalized ANNs was performed using simulated data that have been matched with an acquired image in terms of the signal and noise. In our tests of TNR detection performance, the F-test of the linear parametric neurotransmitter PET (lp-ntPET) model fit residuals was used as the reference method. For a moderate TNR magnitude, the areas under the receiver operating characteristic curves were 0.77–0.79 for the best ANNs and 0.64 for the F-test. At a fixed false positive rate of 0.01, the true positive rates were 0.8–0.9 for the ANNs and 0.6 for the F-test. When applied to a real image, the ANNs identified an additional TNR cluster compared to the F-test. These results demonstrate that personalized ANNs may offer a greater detection sensitivity of dopamine release and other types of TNR compared to previously used methods.**

***Index Terms*—Neurotransmitter response, dopamine release, lp-ntPET, neural network, brain imaging, PET.**

I. Introduction

NEUROLOGICAL and psychiatric disorders are often associated with alterations in multiple neurotransmitter systems. For example, in Parkinson's disease, degeneration of the dopaminergic system is known to be largely responsible for the motor symptoms, while changes in the serotonergic [1] and cholinergic [2] systems are mostly related to mood and cognition. In subjects with obsessive-compulsive disorder, the binding of the serotonin transporter in the insular cortex was found to be decreased compared to healthy controls [3]. In populations prone to risk taking and other compulsive behaviors, impaired reward processing has been found to be associated with abnormal stimulus-induced dopamine release [4]. Probing the function of the neurotransmitter systems in these and other disorders can provide important insights into disease etiology and mechanisms.

Changes in the synaptic neurotransmitter levels in response to a stimulus can be estimated using dynamic positron emission tomography (PET) imaging with receptor radioligands. The most widely used approach is to compute the difference in the non-displaceable binding potential (ΔBP_{ND}) [5] between a baseline (BL) and a post-stimulus scan. A limitation of this method is that it assumes that the neurotransmitter system is in a steady state; thus, when there is a transient change in neurotransmitter concentration, also termed transient neurotransmitter response (TNR), factors such as the magnitude and duration of the TNR become compounded in the measured ΔBP_{ND} value [6]. To address this shortcoming, a linear parametric neurotransmitter PET model (lp-ntPET) has been developed [7], [8], which can be applied to data from a single scan. By fitting the lp-ntPET model to the regional time-activity curves (TACs), one can determine the time course of the neurotransmitter response more precisely, distinguishing the TNR magnitude from its duration.

Applying lp-ntPET on a voxel level can reveal a detailed spatial distribution of TNR [8]. However, such application becomes problematic in high-resolution PET due to a typically high image noise originating from a small number of acquired counts per voxel. The problem is exacerbated by the requirement to acquire dynamic PET data at a relatively high temporal sampling rate (1–3 minutes per frame), further reducing the number of counts per voxel. Fitting lp-ntPET on such data will almost always produce non-zero TNR, regardless of the truth [9]. To identify voxels that are more likely to contain true TNR, the F-test is performed, followed by thresholding at a chosen significance level, between two model fits: a BL model that does not account for TNR, typically the simplified reference tissue model (SRTM) [10], and lp-ntPET. However, the use of the F-test in this application may suffer from several limitations; first, the F-test assumes normally-distributed samples, while the noise in iteratively-reconstructed PET images is known to be closer to a gamma distribution [11]. Secondly, it does not fully account for the different noise levels present in different frames. Finally, it does not exploit any possibly useful prior assumption about the TAC shape.

In this work, we investigate the use of artificial neural networks (ANN) for the task of voxel-level TNR detection as a substitute for the F-test. ANNs are highly efficient at encapsulating both the signal and noise properties of the training data, which yields state-of-the-art performance in

This work was supported by the Natural Sciences and Engineering Research Council of Canada grant number 240670-13, and Canadian Institutes of Health Research grant number 125989.

I. S. Klyuzhin is with the Division of Neurology, Department of Medicine, University of British Columbia, Vancouver, BC, V6T 1Z3 Canada (e-mail: ivansk@physics.ubc.ca).

J-C. Cheng, C. W. J. Bevington, V. Sossi are with the Department of Physics and Astronomy, University of British Columbia, Vancouver, BC, V6T 1Z4 Canada. J-C. Cheng is also with the Pacific Parkinson's Research Centre, University of British Columbia, Vancouver, BC, V6T 2B5 Canada.

many domains. Importantly, unlike the F-test, noise learned by the ANN is not required to have any particular distribution. We have previously demonstrated that denoising autoencoders outperform conventional methods of dynamic PET denoising [12]. On the other hand, the ANN output was found to be rather sensitive to the degree of similarity between the data to be denoised and the training set, i.e. TAC shapes not present in the training set were not processed correctly. Indeed, the over-specificity of ANNs to the signal and noise distributions of the training set is most commonly regarded as a limitation, since ANNs are typically intended to operate on images of different subjects, obtained under different scanning conditions.

A distinguishing feature of our approach is that we attempt to use the high specificity of ANNs to our advantage, by intending/training ANNs to only operate on a single dynamic data set obtained for a particular subject and scan. We term ANNs trained and applied in such fashion "personalized" ANNs (pANN), although it is understood that such networks are scan-specific, rather than subject-specific. Training a pANN to detect TNR in a particular scan involves a procedure where a reference-region TAC and governing kinetic parameters are first determined from the scan, and then used to generate well-matched simulated data for pANN training. Using the scan-specific reference TAC for training data generation is intended to maximize the sensitivity of ANN to the presence of TNR signal in noisy voxel TACs. As long as the employed ANN architecture is relatively small, the network can be readily trained using the simulated data and applied to the acquired dynamic image within minutes, without imposing significant limitations on the image processing and analysis throughput.

To demonstrate the method performance, we used acquired data from a scan with [^{11}C]raclopride (RAC, a dopamine D2 receptor antagonist) where transient dopamine release was induced mid-scan by means of a gambling task. We evaluated several feed-forward ANN architectures for the task of TNR detection: a fully-connected neural net, a convolutional neural net, as well as two architectures based on denoising autoencoders where the TNR signals were treated as anomalies in BL TACs. All networks were trained to output the non-calibrated confidence of TNR presence in a given voxel's TAC. In order to objectively compare the pANNs to the F-test, dynamic test images with known ground truth were generated using the reference TAC and kinetic parameter distributions obtained from the acquired RAC scan. Several performance metrics were evaluated and compared between the methods: 1) areas under the receiver operating characteristic (ROC) curves, for voxels in uniform TNR regions and in small TNR clusters; 2) areas under the ROC curves (AUCs) for different BP_{ND} values, TNR magnitudes and cluster sizes; 3) the sensitivity of TNR detection, fractions of detected TNR clusters, and Dice coefficients at fixed false positive (FP) rates; 4) the size distributions of FP clusters. Finally, we applied the pANNs and the F-test to detect TNR clusters in the acquired image, and compared the results.

II. Methods

A. Method overview

The pipeline for training a pANN for TNR detection is illustrated in Fig. 1a. The acquired RAC image was denoised using a generic (non-personalized) method, and a tissue-derived input function (reference TAC) was measured from the image. The distributions of kinetic parameters in the non-specific and specific tracer binding regions were estimated using SRTM2 [13]. Using the reference TAC, kinetic parameter distributions, and randomly generated volumetric map of non-specific and specific binding regions (with and without TNR) (Fig. 1b), we applied the lp-ntPET model to generate a noise-free dynamic (4D) image; the lp-ntPET parameters corresponding to TNR were set according to the scan protocol and stimulus timing. The dimensions, temporal sampling, and voxel size of the generated image matched those of the acquired image.

The noise-free image was forward-projected using the system model of the PET scanner, Poisson noise was added to the projection data, and the data were reconstructed and post-processed with the same methods as the acquired data to obtain a noisy dynamic image. Voxel TACs from the noisy image were used in the pANN training as inputs, and voxel classes (BL/TNR) were used as targets. In the case of autoencoder training, both noisy and corresponding noise-free TACs were used. We then applied the pANN to the acquired dynamic image and simulated test images to obtain a parametric class score image. The class score image was thresholded to obtain the TNR cluster map. Below we provide a detailed description of these steps.

B. Image acquisition and post-processing

The RAC scan of a healthy control subject was performed on the high resolution research tomograph (HRRT, Siemens) after a bolus injection of 20 mCi of RAC. The scan duration was 75 min, and the acquired data were histogrammed into 30 temporal frames (frame durations 5×1 min, 5×2 min, 20×3 min). At 36 minutes into the scan (18th frame), the subject was presented with a stimulus (Vancouver gambling task [14]) to induce transient dopamine release in the striatum.

The acquired data were reconstructed using the ordered-subset expectation maximization (OSEM) algorithm with incorporated highly-constrained backprojection (HYPR) operator [15]; the HYPR kernel size was $(5.0\ \text{mm})^3$ and 16 OSEM subsets were used. The image dimensions were 256×256×207 voxels, with an isotropic voxel size of $(1.22\ \text{mm})^3$. After reconstruction, HYPR post-processing with a $(5.0\ \text{mm})^3$ kernel size was applied to further reduce noise. The subject was also scanned on a Philips Achieva 3T scanner to provide anatomical reference. The acquired T1-weighted image was segmented using Freesurfer 6.0 [16] to obtain masks of the cerebellum and striatum.

C. Kinetic models for dynamic PET Data

As mentioned above, we used two models to first fit the acquired TACs and then to simulate RAC kinetics. SRTM was used to analyze the BL state:

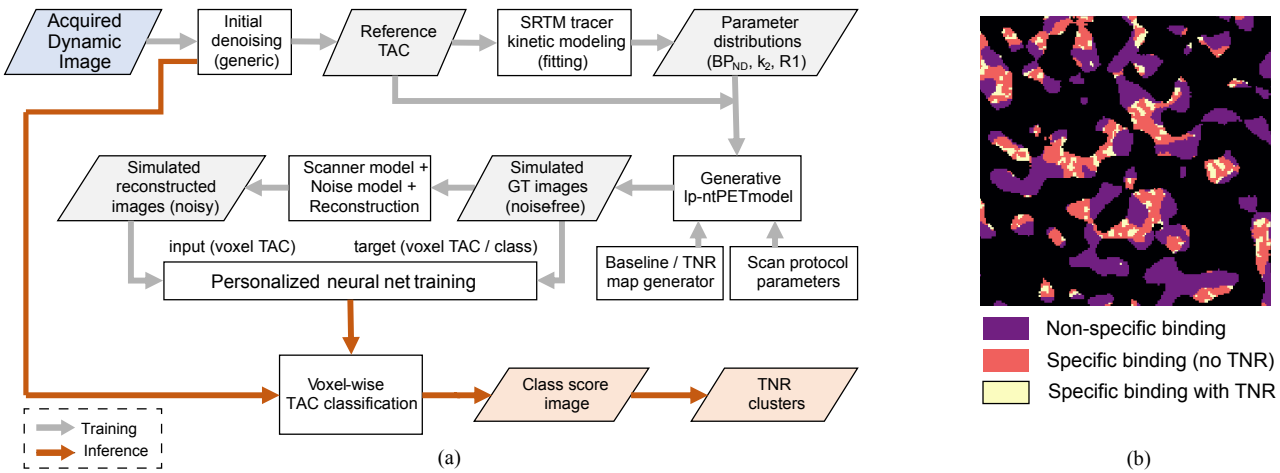

Fig. 1. (a) The pipeline for training a pANN for voxel-level TNR detection. (b) A ranomly-generated volumetric map (single plane is shown) with regions denoiting non-specific and specific tracer binding, as well as regions of specific binding with TNR.

$$C_T(t) = R_1 C_R(t) + \left[k_2 - \frac{R_1 k_2}{1 + BP_{ND}}\right] C_R(t) \otimes \exp\left[-\frac{k_2 t}{1 + BP_{ND}}\right] \quad (1)$$

where $C_T(t)$ is the measured (fitted) TAC, t is the frame time, k_2 is the tissue-to-plasma rate constant, R_1 is the relative delivery in the target region compared to the reference region, $C_R(t)$ is the tissue-derived input function (reference region TAC); $\otimes$ denotes the temporal convolution. From this equation, a noise-free TAC with no TNR can be generated given the input function and the triplet of values BP_{ND}, k_2, R_1.

To model TACs with TNR, we used the lp-ntPET model:

$$C_T(t) = R_1 C_R(t) + k_2 \int_0^t C_R(u)du - \frac{k_2}{1 + BP_{ND}} \int_0^t C_T(u)du - \gamma \int_0^t C_T(u)h(u)du \quad (2)$$

where the last term accounts for time-varying TNR, and γ is related to the magnitude TNR. $h(t)$ is the time-course of TNR modeled as the function

$$h(t) = H(t - t_D)\tau^{\alpha} e^{-\alpha(\tau - 1)}, \quad \tau = (t - t_D)/(t_P - t_D) \quad (3)$$

where H is the Heavside step function, t_D and t_P are the start and peak times of TNR, respectively; the parameter α determines the sharpness of the response.

D. Training image generation

The cerebellum, defined from a mask produced by Freesurfer, was used as reference region to determine the tissue input TAC $C_R(t)$. SRTM2 was applied on the voxel level, and the joint distributions of BP_{ND}, k_2 and R_1 in the non-specific (cerebellum) and specific (striatum) binding regions were evaluated. To avoid biased parameter values due to the presence of TNR, the SRTM2 was fit only to the TACs limited to the first 36 minutes, i.e. prior to the stimulus. The 1st and 99th percentiles of the measured BP_{ND} distribution were (-0.45, 0.89) in the cerebellum, and (0.78, 5.72) in the striatum. The corresponding percentiles for k_2 were (0.0015, 0.009) in the cerebellum, and (0.0022, 0.0065) in the striatum. The ratio of R_1/k_2 for the subject was 260.0 for all voxels; thus we set $R_1 = 260 \times k_2$.

To generate simulated dynamic images for pANN training, we started with a random volumetric map, which labeled 3D regions as corresponding to non-specific, specific and zero tracer binding (i.e. cold regions) (Fig. 1b). Voxels in these regions were assigned triplets of (BP_{ND}, k_2 and R_1) values drawn from the respective distributions obtained from the acquired image. Following the value assignment, an anisotropic diffusion filter [17] was applied, which produced smoothly varying local gradients of the kinetic parameters (magnitudes and directions) within the specific and non-specific binding regions, while preserving high-contrast edges between the regions. Further detail can be found in [12].

The specific binding regions were set to contain clusters of TNR, ranging in size from 100 to approximately 500 voxels. Each cluster had unique randomly assigned values of γ, t_D, t_p and α, set in accordance with the stimulus timing [8], [9], [18]:

- γ ranged from 0.1×10^{-3} to 1.0×10^{-3} s^{-1}.
- The range of t_D (TNR start time) was 36 $\pm$ 5 min.
- The range of t_P (TNR peak time) was from t_D+5 to t_D+15 min.
- α was randomly chosen between 0.25, 1.0 and 4.0.

Based on the assumption that the profiles of TNR in nearby voxels should be similar, the values of these parameters were set to be spatially constant within each cluster.

From the generated parametric images, we simulated noise-free TACs for each voxel using (2) and the reference TAC from

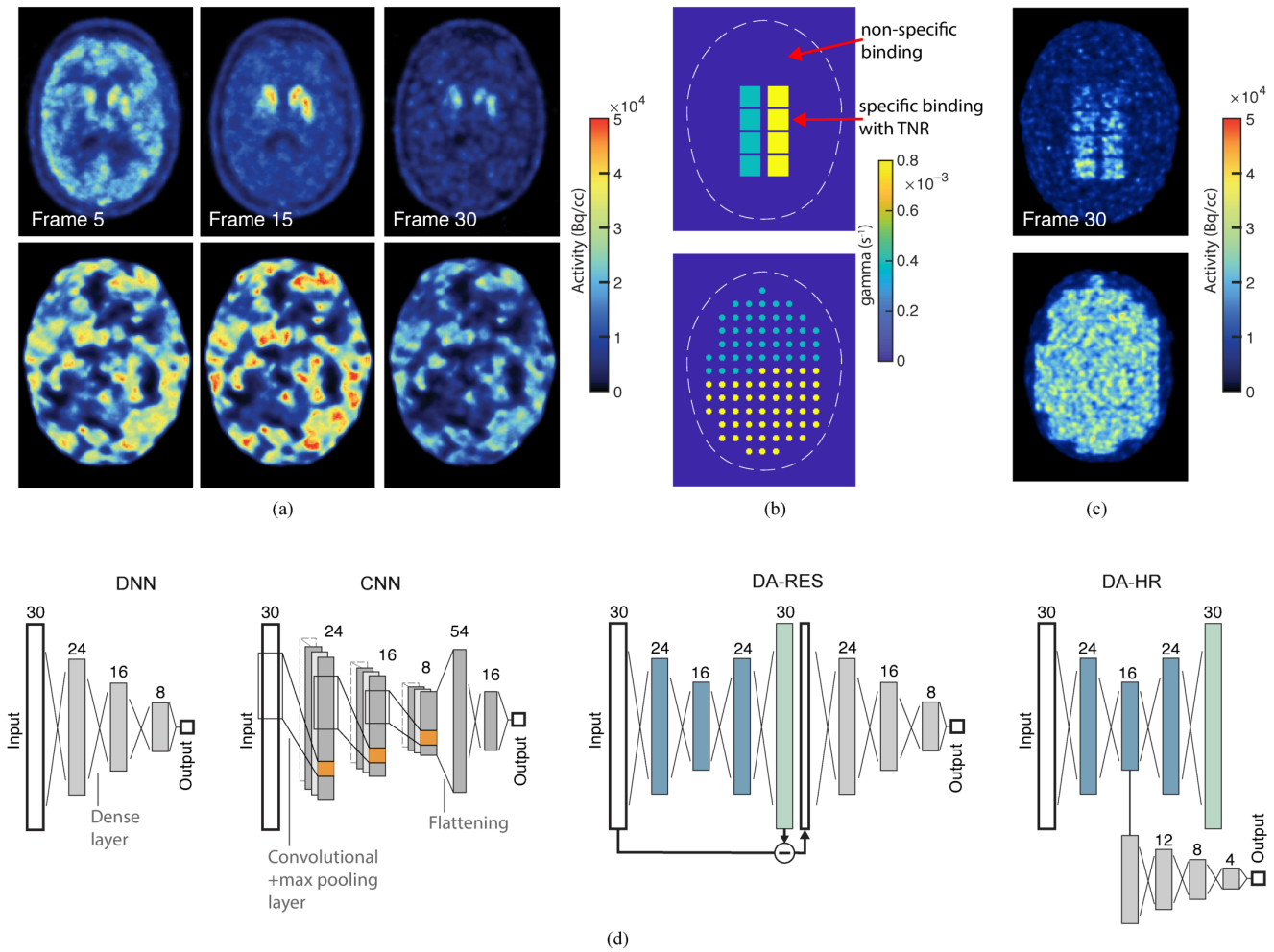

Fig. 2. (a) Example frames of the acquired RAC image (top) and simulated training images (bottom). (b) The parametric test images showing the regional distribution of γ values. (c) Example frames of the simulated dynamic test images. (d) Architectures of the tested ANNs for TNR detection. The autoencoder layers are shown in blue. The number of nodes per layer is shown above the layer blocks.

the acquired scan. For the voxels with no TNR, the parameter γ was set to zero. The simulated TACs were resampled into 30 time points, to yield dynamic images with frame number and durations consistent with those in the acquired RAC image. Likewise, the voxel size was set to (1.22 mm)^3, equal to the HRRT image voxel size.

To simulate intrinsic resolution blurring of the HRRT, each frame of the simulated noise-free image was smoothed with an isotropic 2.5 mm full-width at half-maximum (FWHM) Gaussian filter. Each frame was then forward-projected using the HRRT system matrix, and the effects of subject attenuation and detector normalization were included in the sinogram space. Poisson noise was added to the resulting 3D sinogram data, which were reconstructed and post-processed similarly to the acquired image (Fig. 2a).

E. Test image generation

Two parametric test images were used (Fig. 2b). The first image contained large regions with uniform TNR. The values of γ were set to be different in different regions, equal to 0.2×10^{-3}, 0.4×10^{-3}, 0.6×10^{-3} and 0.8×10^{-3}. The values of t_D and t_P were set in all regions to 36 and 46 minutes, respectively, and α was set to 1.0.

The second image contained spherical TNR clusters of various sizes, placed in a BL "background" with BP_{ND} set to 4.0. The cluster diameters were 3 voxels (3.6 mm), 5 voxels (6.0 mm) and 7 voxels (8.4 mm). The values of γ in the clusters were similarly set to 0.2×10^{-3}, 0.4×10^{-3}, 0.6×10^{-3} and 0.8×10^{-3}; the other TNR parameters were similar to those in the uniform regions. The procedure to generate noisy dynamic images for testing was similar to that for the training image; examples of individual frames are shown in Fig. 2c. The voxel TACs were generated using the measured reference TAC.

Initial experiments have shown that using a (5.0 mm)^3 HYPR reconstruction/post-processing kernel produced strong ceiling and floor effects when measuring the classification performance in the uniform and cluster TNR regions, respectively. Therefore, with simulated test images, the HYPR reconstruction and post-processing kernel size was reduced to (2.5 mm)^3; the ANN training data were adjusted accordingly.

F. Neural net architectures

Diagrams of the tested neural net architectures are plotted in Fig. 2d. The first architecture was a traditional dense neural net (DNN) with three hidden layers. The number of nodes in each layer was 30 (input), 24, 16, 8, 1 (output). All layers had rectified linear unit (ReLU) activations, except the output, which had sigmoid activation (ranging from 0 to 1).

The second architecture was a one-dimensional convolutional neural net (CNN), with three convolutional layers followed by a dense layer and a regression layer with sigmoid activation. The number of neurons (or filters) in each layer was the same as the number of nodes in the DNN; a max pooling layer with a downscale factor of 2 followed each convolutional layer. The span of convolution windows was set to 1/2 of that layer's input size (i.e. for the first layer the span was equal to 15).

The third architecture was based on a denoising autoencoder, with an auxiliary neural net operating on the input-output residuals (denoted DA-RES). This method was inspired by the use of autoencoders for anomaly detection, where the difference between the inputs and outputs (e.g. mean squared error) is often used as the anomaly score [19]. Within the context of our work, TNR signals can be treated as "anomalies" in normal BL TACs. However, since we had simulated examples of such anomalies, instead of using an unsupervised anomaly metric we performed supervised training a small auxiliary neural net. The DA consisted of 5 fully-connected feed-forward layers, as illustrated. The number of nodes in each layer was 30 (input), 24, 16, 24, 30 (output), and all layers had ReLU activations.

In the fourth architecture, the auxiliary network was set to operate on the DA's hidden representation (bottleneck) layer (denoted DA-HR). This design was inspired by a previous work on autoencoder-based anomaly detection, where a one-class support vector machine [20] or one-class neural net [21] were trained using the DA's hidden data representation. With DA trained on BL TACs, TNR TACs were expected to produce anomalous activations in the bottleneck nodes. Similarly to DA-RES, we perform supervised two-class training of the auxiliary network, which was intended to detect when the activation pattern was different from the normal distribution.

Using cross-validation, we evaluated the performance of DNNs/CNNs with different number of hidden/convolutional layers. The validation dataset was constructed by randomly picking 20% of the training data, and ten cross-validation folds were used. We found that there was very little benefit in using more than three hidden/convolutional layers in terms of the validation error. We also tested different convolution window spans in the CNNs equal to 1/3, 1/2 and 2/3 of the previous layer size, and found that the validation error was 1% lower on average in the latter two cases. Once the number of hidden layers was determined, we trained the networks for testing using the entire training data. For DNN/CNN training, noisy BL/TNR TACs were used as inputs, and the TAC classes (0 for BL, 1 for TNR) were used as targets. The number of BL and TNR training samples was 100,000 each, and training was performed for 300 epochs. The DA-based architectures were trained in two phases. In phase one, the DA network was trained using only BL TACs: noisy TACs were used as inputs, and noise-free TACs were used as targets. Training was performed for 300 epochs using 100,000 samples. In the second phase, the autoencoder weights and biases were fixed, and training of the auxiliary network was performed. In this phase a mixture and BL and TNR TACs was used, and TAC classes were used as targets. The number of training samples was 100,000 for each class, and training was performed for 200 epochs. We used the ADAM optimizer with the initial learning rate of 0.001; mean squared error was the optimized loss function.

G. Reference method

The following previously proposed [8] method to detect voxels with TNR was used here as the "reference". The noisy TAC from a given voxel was fitted with the SRTM2 and lp-ntPET models, described by (1) and (2). The weighted residual sum of squares (wRSS) was computed for the two fits, and the F-score was computed using the equation

$$F = \frac{n - p_2}{p_2 - p_1} \frac{wRSS_1 - wRSS_2}{wRSS_2} \tag{4}$$

where n is the number of time samples in the TAC, the subscript designates the model (1 for SRTM2, 2 for lp-ntPET), and p is the number of parameters in the model. An empirically-determined threshold is applied to the F values to determine voxels with TNR. The F score can then be treated as the decision function of a classifier, and compared to other anomaly detection methods (e.g. by comparing the ROC AUC).

H. Method testing and comparison

Using the simulated test images, we computed several objective metrics that quantified the TNR detection performance, and compared them between DNN, CNN, DA-RES, DA-HR and the F-test. The ROC AUC was measured for the task of BL/TNR classification of individual voxels. For voxels taken from the uniform BL/TNR regions, the AUC was measured for different BP_{ND} and γ values; for voxels from the cluster regions, the AUC was measured for different γ values and cluster sizes.

For voxels in the uniform regions, we also measured the TNR detection sensitivity at fixed FP rates of 0.05 and 0.01. Likewise, for TNR clusters, the detected cluster fraction, cluster specificity, and Dice coefficient (using the detected and ground truth cluster shapes) were measured at FP rates of 0.05 and 0.01. These metrics were obtained and compared between the methods for different values of BP_{ND}, γ and cluster sizes.

Using the uniform BL regions, we measured the distribution of sizes of FP clusters, and estimated the cluster size threshold to reject 95% and 99% of FP clusters.

Finally, we applied the neural nets and the F-test to detect dopamine release in the acquired RAC image. Since in this case we don't know the ground truth TNR distribution, we simply compare the TNR clusters detected by the neural nets to those detected by the F-test.

III. Results

A. ROC analysis

The ROC AUC values measured for voxels in the uniform TNR regions are plotted in Fig. 3a, for different values of BP_{ND} and γ. The graphs demonstrate that the methods DNN, CNN and DA-RES had consistently higher AUCs compared to the F-test; on the other hand, DA-HR performed similarly or worse than the F-test. For γ values of 0.4×10^{-3} and greater, the ANN methods offered only a small improvement over the F-test, likely due to the ceiling effect. For γ=0.2×10^{-3}, the biggest improvement compared to the F-test was found for BP_{ND}=4.

The ROC AUC values measured for voxels in the TNR clusters are plotted in Fig. 3b. Here the AUC values were generally much lower compared to those for the uniform regions, likely due to the presence of the partial volume effect (i.e. TNR TACs from small clusters become mixed with BL TACs from the surrounding background). The graphs demonstrate that DNN, CNN and DA-RES consistently and substantially outperform the F-test, with different cluster sizes and γ values; the AUC values of DNN and DA-RES are practically identical, while the values for CNN are marginally lower. DA-HR performed on average similarly to the F-test. None of the methods could effectively detect TNR in 3-voxel clusters with γ less than 0.8×10^{-3}.

Examples of the ROC curves for voxels in the uniform regions are plotted in Fig. 4a for γ=0.2×10^{-3} and 0.4×10^{-3}. The curves demonstrate that there is indeed a significant improvement with DNN or CNN over the F-test (DA-RES curves were similar to those of DNN). The shapes of the ROC curves imply that all ANNs except DA-HR would offer a substantially better true positive rate, i.e. sensitivity, at practically any FP rate. Of particular interest are low FP rates of 0.05 and 0.01; in these regimes the F-test had better true positive to FP trade-off compared to DA-HR.

Examples of the ROC curves for voxels in the TNR clusters are plotted in Fig. 4b for 3- and 5-voxel clusters, and γ = 0.6×10^{-3}. The curve shapes reveal that the ANN AUC gains over the F-test are mostly coming from the intermediate FP rates (0.2–0.8), which are unlikely to be acceptable in practical applications.

B. TNR detection sensitivity

Images showing the results of TNR detection in the uniform and cluster test regions are plotted in Fig. 4c and 4d, respectively. The images demonstrate the improved sensitivity of DNN and CNN compared to the F-test. DA-RES images (not shown) were similar to the DNN images. For the γ values of 0.4×10^{-3} and greater, most voxels in the uniform regions were classified correctly. Most of misclassified TNR voxels were near the edges, likely due to the partial volume effect (from the combined intrinsic resolution blurring and denoising). Interestingly, the locations of the false negative voxels were very similar in the F-test and ANN-produced TNR images, suggesting that there were locations in the input image that were consistently difficult to correctly classify for all methods. The detected TNR clusters (Fig. 4d) were noticeably more prominent in the DNN and CNN images compared to the F-test and DA-HR.

The TNR detection sensitivities for voxels in the uniform regions are given in Table I, for the fixed FP rate of 0.01. The DNN and DA-RES methods offered the most substantial gains in sensitivity compared to the F-test, especially for intermediate γ values of 0.4×10^{-3} and 0.6×10^{-3}. The greatest improvement was found with BP_{ND}=5 and γ=0.4×10^{-3}: the sensitivity was 0.62 for the F-test and 0.9 for DNN/DA-RES. On the other hand, the sensitivity of DA-HR was very low, particularly for BP_{ND}=3.

TABLE I
TNR Detection Sensitivity in Uniform Regions (FP rate = 0.01)

	γ (s^{-1})			
	0.2×10^{-3}	0.4×10^{-3}	0.6×10^{-3}	0.8×10^{-3}
	BP_{ND} = 3			
F-test	0.12	0.38	0.62	0.85
DNN	0.19	0.6	0.82	0.95
CNN	0.15	0.51	0.77	0.91
DA-RES	0.19	0.57	0.8	0.95
DA-HR	0.023	0.087	0.15	0.28
	BP_{ND} = 4			
F-test	0.15	0.6	0.82	0.97
DNN	0.23	0.81	0.96	0.99
CNN	0.22	0.78	0.95	1.0
DA-RES	0.22	0.79	0.95	0.99
DA-HR	0.055	0.35	0.35	0.53
	BP_{ND} = 5			
F-test	0.2	0.62	0.88	0.99
DNN	0.37	0.90	0.98	1.0
CNN	0.35	0.89	0.98	1.0
DA-RES	0.41	0.90	0.99	1.0
DA-HR	0.076	0.21	0.39	0.48

C. Detection of true TNR clusters

Using the test image with TNR clusters, we measured the average detected fraction of a TNR cluster as a function of the cluster size and γ value. The results are provided in Table II. Only those clusters where at least 7 central voxels were detected correctly entered the calculation; this explains the biased values for the 3-voxel clusters with 0.4×10^{-3}. For greater γ values and larger clusters, we found that DNN, CNN and DA-RES detected a larger fraction than the F-test by a factor of approximately 1.7. In the best-case scenario (γ=0.8×10^{-3}, cluster diameter 7 voxels), we could detect only about 50% of a TNR cluster. The Dice coefficient was measured for the detected clusters to take both the sensitivity and specificity into account, i.e. what fraction of the BL voxels around the TNR cluster was classified correctly (Fig. 5a). For all γ values, methods DNN, CNN and DA-RES had greater

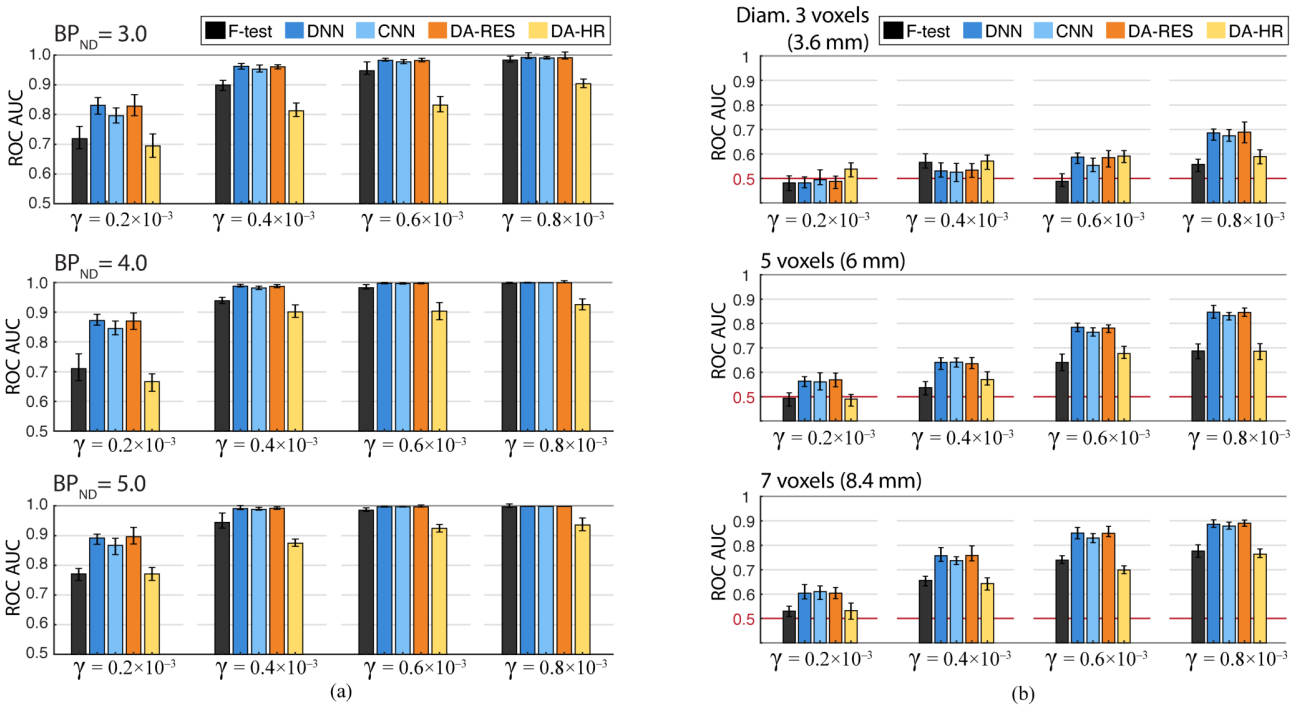

Fig. 3. Comparison of BL/TNR classification ROC AUC values between the tested methods. (a) AUC values measured in the uniform TNR regions. (b) AUC values measured in TNR clusters (BP_{ND}=4).

Dice coefficients than the F-test for 5- and 7-voxel TNR clusters.

TABLE II
FRACTIONS OF DETECTED TNR CLUSTERS (FP RATE = 0.01)

	Cluster diameter		
	3 vox. (3.6 mm)	5 vox. (6 mm)	7 vox. (8.4 mm)
	γ = 0.4×10⁻³		
F-test	**0.16** ± 0.11	**0.07** ± 0.094	**0.11** ± 0.089
DNN	**0.19** ± 0.23	**0.075** ± 0.078	**0.17** ± 0.13
CNN	**0.16** ± 0.17	**0.093** ± 0.085	**0.18** ± 0.14
DA-RES	**0.25** ± 0.27	**0.082** ± 0.084	**0.18** ± 0.13
DA-HR	**0.082** ± 0.038	**0.086** ± 0.081	**0.089** ± 0.077
	γ = 0.6×10⁻³		
F-test	**0.14** ± 0.083	**0.13** ± 0.12	**0.21** ± 0.11
DNN	**0.14** ± 0.12	**0.18** ± 0.16	**0.34** ± 0.17
CNN	**0.16** ± 0.16	**0.18** ± 0.16	**0.35** ± 0.16
DA-RES	**0.12** ± 0.11	**0.2** ± 0.17	**0.34** ± 0.17
DA-HR	**0.14** ± 0.11	**0.097** ± 0.084	**0.13** ± 0.099
	γ = 0.8×10⁻³		
F-test	**0.17** ± 0.12	**0.15** ± 0.12	**0.28** ± 0.16
DNN	**0.19** ± 0.13	**0.26** ± 0.15	**0.45** ± 0.17
CNN	**0.16** ± 0.08	**0.26** ± 0.15	**0.47** ± 0.17
DA-RES	**0.15** ± 0.11	**0.26** ± 0.15	**0.46** ± 0.17
DA-HR	**0.16** ± 0.12	**0.13** ± 0.12	**0.17** ± 0.12

With $\gamma = 0.2\times10^{-3}$, the detected cluster fraction was negligibly small with all methods.

D. FP cluster analysis

FP clusters were analyzed in the uniform BL regions with BP_{ND}=4, and single-voxel FP rates set to 0.05 an 0.01. Smaller FP clusters are preferable, since those can be rejected based on the cluster size thresholding [8], without sacrificing larger true positive clusters. The cumulative distribution functions of FP clusters with respect to their size are plotted in Fig. 5b. At the FP rate equal to 0.05, the 95%-tiles (99%-tiles) of the FP cluster sizes in voxels were 25 (52) for the F-test, 50 (121) for DNN, 42 (93) for CNN, 58 (106) for DA-RES, and 63 (103) for DA-HR. Thus, the ANN methods produced FP clusters that were more than twice the size compared to the F-test.

However, at the FP rate equal to 0.01, the 95%-tiles (99%-tiles) were 16 (30) for the F-test, 17 (31) for DNN, 16 (53) for CNN, 17 (24) for DA-RES, and 23 (55) for DA-HR. Thus, the FP clusters produced by the DNN and DA-RES were nearly equal in size to those produced the F-test. Thus, at the FP rate of 0.01, the methods DNN and DA-RES produced FP clusters of the same size as the F-test, while offering a substantially greater sensitivity.

E. TNR detection in the acquired image

Applying the TNR detection methods to the segment of the acquired RAC image defined by the striatal mask produced parametric class-score images shown in Fig. 6a. A higher class score corresponds to a greater confidence that the voxel has non-zero TNR. On the left side of the striatum, the ANN methods detected similar clusters to the F-test. Namely, four high-score regions were detected, two in the superior caudate and two in the superior putamen. However, the DNN, CNN and DA-RES methods had a greater contrast of regions with

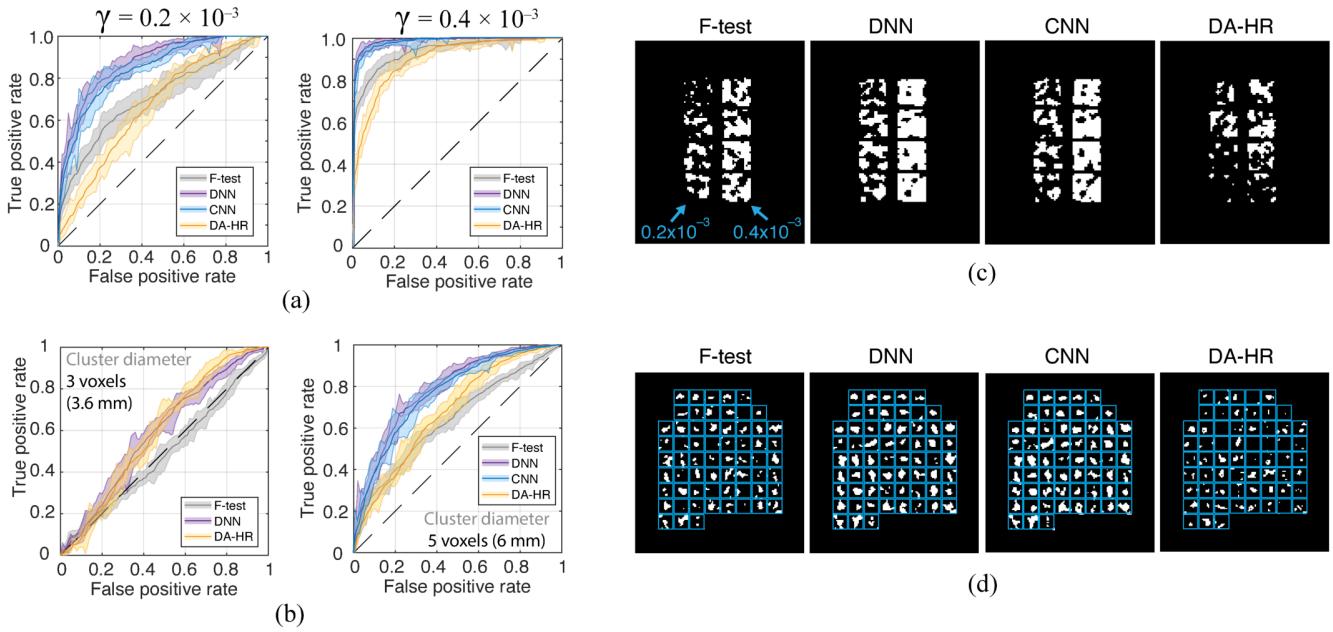

Fig. 4. (a) Examples of ROC curves for the tested methods measured in the uniform TNR regions (BP_{ND}=4). Shaded regions represent the 95% confidence intervals. (b) Examples of ROC curves measured in the TNR cluster region (BP_{ND}=4, γ=0.6×10^{-3}). Shaded regions represent the 95% confidence intervals. The CNN curve for 3-voxel clusters was omitted for clarity. (c) Detected TNR image in uniform TNR regions (FP rate 0.01). Numbers indicate γ values. (d) Detected TNR image in the TNR cluster region (FP rate 0.01). Rectangular cells are drawn around the locations of ground truth TNR clusters.

high class scores as, for example, in the inferior putamen. On the right side of the striatum, DNN, CNN and DA-RES also detected similar regions to those from the F-test, however they also detected a new high-score region in the posterior putamen not present in the F-test image.

The results of thresholding the class score images to obtain the final TNR classification at FP rate of 0.01 are plotted in Fig. 6b. The detected TNR clusters in the superior caudate/putamen were similar in size between the F-test and the DNN/CNN/DA-RES methods. The newly detected TNR cluster in the posterior putamen was connected with the caudate cluster in the DNN image, and disconnected in the CNN and DA-RES images.

To test whether the additional ANN-detected cluster in the posterior putamen was a true positive or a FP, we measured the mean regional TAC in that cluster (according to DA-RES), and fitted it with SRTM and lp-ntPET (6c). A significantly better fit was obtained with lp-ntPET, with the following fitted parameter values: γ=0.226×10^{-3} s^{-1}, t_D=31 min, t_P=43.5 min, α=0.25, BP_{ND}=4.46. These numbers suggest that there was likely transient dopamine release in the posterior putamen during the scan, which was detected by the pANNs and missed by the F-test.

IV. Discussion

Traditionally, the detection of TNR from dynamic PET data has been achieved by comparing the SRTM and lp-ntPET model fits using an F-test. Our results demonstrate that pANNs trained to classify single-voxel TACs outperform the F-test for TNR detection, in terms of several objective metrics. Specifically, for a given FP rate, the networks DNN, CNN and DA-RES had higher sensitivity than the F-test in uniform TNR regions and in TNR clusters. At the FP rate of 0.01, the FP cluster sizes produced by DNN and DA-RES were similar to that of the F-test. When tested on a real image, all ANNs detected same TNR clusters as the F-test; an additional TNR cluster was detected by the DNN, CNN and DA-RES networks, which was verified to likely contain TNR by direct lp-ntPET model fitting. This outcome provides an initial validation of the approach where a neural net is trained entirely on simulated data and applied to detect TNR in real data.

We attribute the better pANN performance to several factors. First, as universal approximators the ANNs can accurately capture the input data distribution. In the process of training, the ANNs have presumably encapsulated the differences between lp-ntPET and SRTM models on a frame-by-frame level. Second, better performance may be attributed to the better handling of frame-dependent noise properties. Finally, we believe that training the ANNs using a specific input function (reference TAC), what we call ANN "personalization", contributes to their greater sensitivity to detect TNR.

The most common approach to train ANNs for medical image analysis is to diversify the training set, in the attempt to train the network to handle images of varying properties (different subjects, noise levels, resolutions, etc.). In contrast, our approach entails training a separate neural net to process each individual image. This approach becomes feasible when simulated image data can be obtained that closely match the acquired data. What exactly constitutes a "close match" is yet to be determined; in this work, using the same reference TAC, reconstruction and post-processing methods as with the real

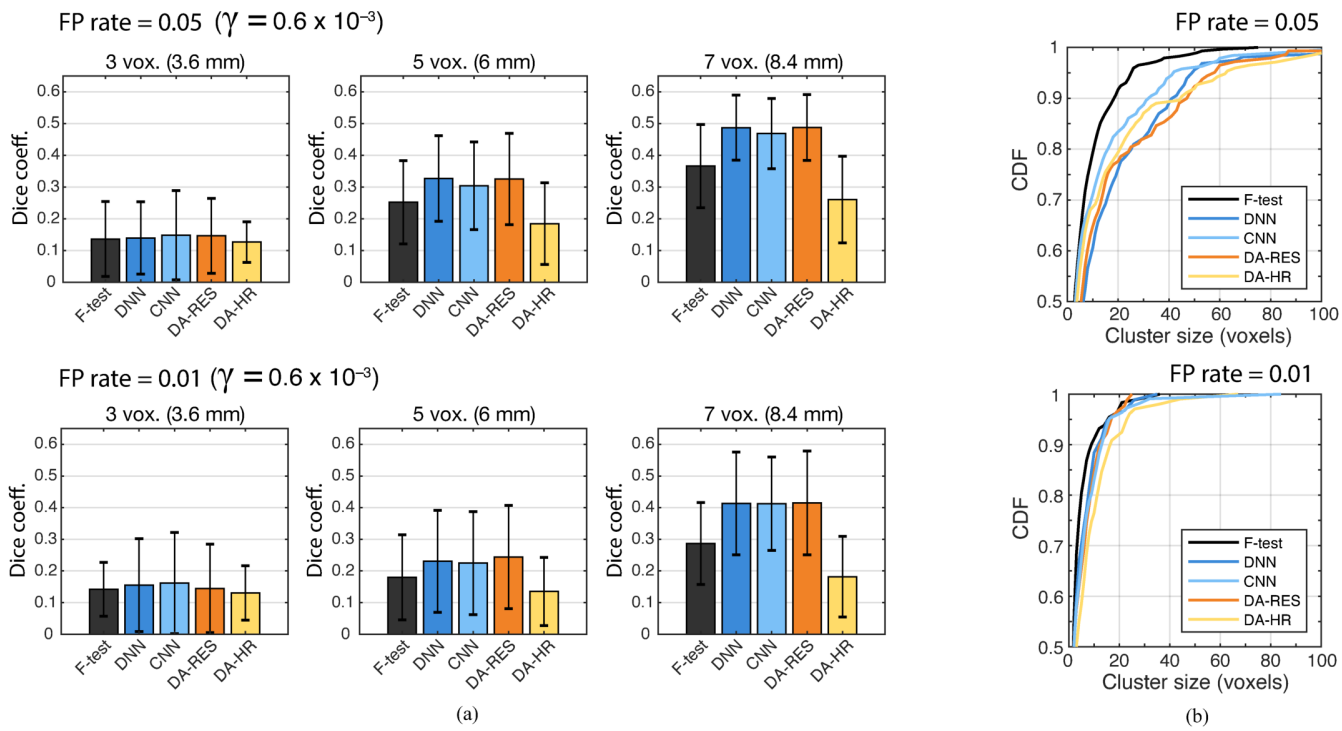

Fig. 5. (a) Dice coefficients for the detected TNR clusters at FP rates of 0.05 and 0.01. (b) Cumulative distribution functions (CDF) of FP clusters according to their volume, at FP rates of 0.05 and 0.01 (BP_{ND}=4).

data appears to have been sufficient. The network training time is not a limiting factor, as small networks such as the ones used here can be trained on modern graphics processing units within minutes. An interesting observation is the fact that by personalizing the ANNs there is no apparent need to worry about generalization and algorithm fairness, i.e. known and unknown biases related to subject age, weight, and other relevant attributes [22].

Interestingly, DNN and CNN performed rather similarly despite the different architectures. This may be a consequence of our ANNs being relatively small and shallow. Likewise, the practically identical performance of DNN and DA-RES demonstrates that there is no benefit in performing autoencoder-based TAC denoising prior to classification. The relatively worse performance of DA-HR compared to the other ANNs likely stems from data compression in the hidden representation. We speculate that within the DA bottleneck layer, a portion of TNR-related information is removed or marginalized due to the fact that the DA is trained using only the BL data. Therefore, hidden TAC representation would have less available information, i.e. fewer salient features related to TNR compared to a network that is trained on the full input. Thus, using a hidden DA representation specifically for TNR detection may be suboptimal. However, this architecture could be attractive in situations where the primary objective is different (e.g. denoising) to detect unexpected changes in the input distribution. In a recent study not related to medical imaging, an autoencoder was trained by optimizing a modified cost function that enabled the bottleneck layer to capture anomalies more efficiently [21]; it may be of interest to adopt this technique for medical image processing and analysis.

Despite the fact that pANNs outperformed the F-test in TNR detection, the overall sensitivity of TNR detection remained relatively low. Continued effort is required to improve the quality of image reconstruction, and to develop methods of post-processing. To reduce the number of detected FP clusters, a lower bound on the TNR cluster size can be imposed [8], at the cost of reducing the spatial "resolution" of detection.

A limitation of our study is that we only tested the pANNs on data from one real scan. At the time of this study, only one RAC scan with transient dopamine release has been acquired at our center. A more thorough applicability analysis needs to be performed when additional scans become available. Another limitation is that we only tested ANNs with one-dimensional inputs; this was done in order to fairly compare ANNs to the F-test. Using ANNs that operate on 4D data is likely to lead to further improvements in TNR detection accuracy.

To summarize, TNR identification using pANNs trained on simulated data is a viable approach, under the caveats discussed above. Although the ANN architectures tested here were relatively simple, the findings may extend to more advanced methods; current results can serve as a guide for further development. Future work will focus on testing deep convolutional spatio-temporal ANNs, and comparing them to conventional methods that incorporate more complex analysis of voxel clusters.

References

[1] J. F. Fu, I. Klyuzhin, S. Liu, E. Shahinfard, N. Vafai, J. McKenzie, N. Neilson, R. Mabrouk, M. A. Sacheli, D. Wile, M. J. McKeown, A. J. Stoessl, and V. Sossi, "Investigation of serotonergic Parkinson's

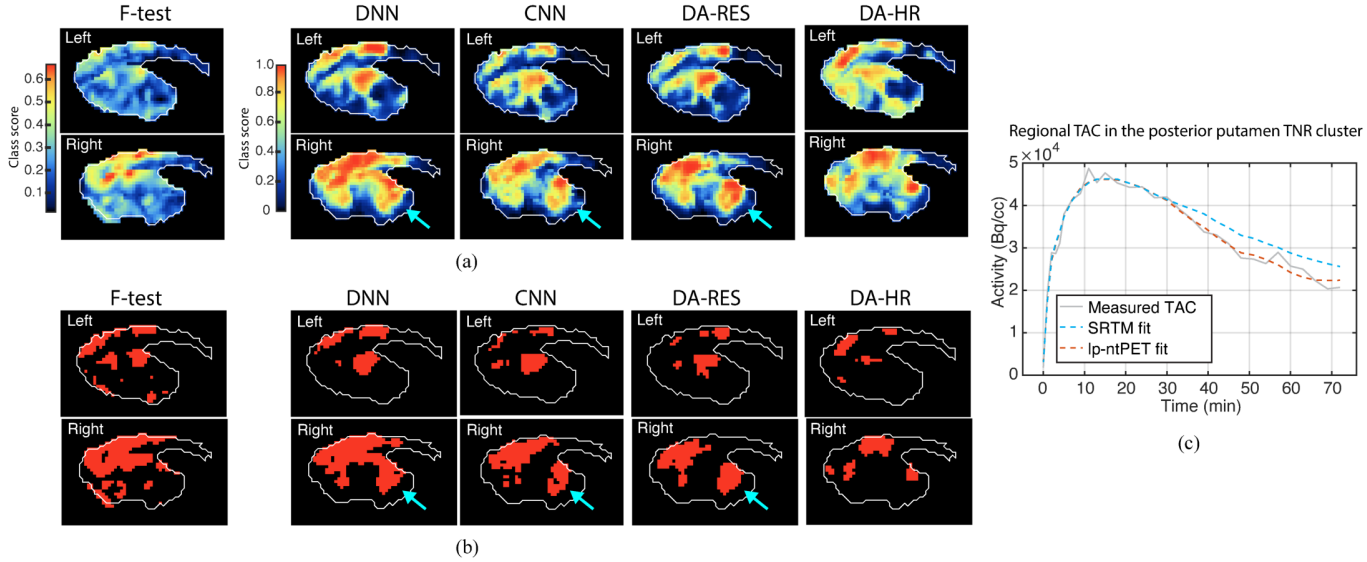

Fig. 6. (a) Parametric class score (baseline/TNR) images for the acquired RAC scan obtained with each method. For the F-test image, we plotted the logarithm of the F-score for achieve better scale resemblance to other methods. (b) Binary TNR maps obtained by thresholding the class score images. Arrows indicate the TNR cluster not detected by the F-test. (c) Measured regional TAC in the new TNR cluster detected by DA-RES, with SRTM and lp-ntPET fits.

disease-related covariance pattern using [11C]-DASB/PET," *NeuroImage: Clinical*, 2018.

[2] N. I. Bohnen, R. L. Albin, M. L. Müller, M. Petrou, V. Kotagal, R. A. Koeppe, P. J. Scott, and K. A. Frey, "Frequency of cholinergic and caudate nucleus dopaminergic deficits across the predemented cognitive spectrum of parkinson disease and evidence of interaction effects," *JAMA Neurology*, 2015.

[3] R. Matsumoto, M. Ichise, H. Ito, T. Ando, H. Takahashi, Y. Ikoma, J. Kosaka, R. Arakawa, Y. Fujimura, M. Ota, A. Takano, K. Fukui, K. Nakayama, and T. Suhara, "Reduced serotonin transporter binding in the insular cortex in patients with obsessive-compulsive disorder: A [11C]DASB PET study," *NeuroImage*, 2010.

[4] M. Dubol, C. Trichard, C. Leroy, A. L. Sandu, M. Rahim, B. Granger, E. T. Tzavara, L. Karila, J. L. Martinot, and E. Artiges, "Dopamine Transporter and Reward Anticipation in a Dimensional Perspective: A Multimodal Brain Imaging Study," *Neuropsychopharmacology*, 2018.

[5] R. B. Innis, V. J. Cunningham, J. Delforge, M. Fujita, A. Gjedde, R. N. Gunn, J. Holden, S. Houle, S. C. Huang, M. Ichise, H. Iida, H. Ito, Y. Kimura, R. A. Koeppe, G. M. Knudsen, J. Knuuti, A. A. Lammertsma, M. Laruelle, J. Logan, R. P. Maguire, M. A. Mintun, E. D. Morris, R. Parsey, J. C. Price, M. Slifstein, V. Sossi, T. Suhara, J. R. Votaw, D. F. Wong, and R. E. Carson, "Consensus nomenclature for in vivo imaging of reversibly binding radioligands," *J Cereb Blood Flow Metab*, vol. 27, no. 9, pp. 1533–1539, sep 2007.

[6] K. K. Yoder, C. Wang, and E. D. Morris, "Change in binding potential as a quantitative index of neurotransmitter release is highly sensitive to relative timing and kinetics of the tracer and the endogenous ligand," *Journal of Nuclear Medicine*, 2004.

[7] M. D. Normandin, W. K. Schiffer, and E. D. Morris, "A linear model for estimation of neurotransmitter response profiles from dynamic PET data," *NeuroImage*, vol. 59, no. 3, pp. 2689–2699, 2012.

[8] S. J. Kim, J. M. Sullivan, S. Wang, K. P. Cosgrove, and E. D. Morris, "Voxelwise lp-ntPET for detecting localized, transient dopamine release of unknown timing: Sensitivity analysis and application to cigarette smoking in the PET scanner," *Human Brain Mapping*, vol. 35, no. 9, pp. 4876–4891, 2014.

[9] S. Wang, S. Kim, K. P. Cosgrove, and E. D. Morris, "A framework for designing dynamic lp-ntPET studies to maximize the sensitivity to transient neurotransmitter responses to drugs: Application to dopamine and smoking," *NeuroImage*, vol. 146, no. May 2016, pp. 701–714, 2017.

[10] R. N. Gunn, A. A. Lammertsma, S. P. Hume, and V. J. Cunningham, "Parametric imaging of ligand-receptor binding in PET using a simplified reference region model." *NeuroImage*, vol. 6, no. 4, pp. 279–87, dec 1997.

[11] A. Teymurazyan, T. Riauka, H. S. Jans, and D. Robinson, "Properties of noise in positron emission tomography images reconstructed with filtered-backprojection and row-action maximum likelihood algorithm," *Journal of Digital Imaging*, 2013.

[12] I. S. Klyuzhin, J.-C. Cheng, C. Bevington, and V. Sossi, "Use of a Tracer-specific Deep Artificial Neural Net to Denoise Dynamic PET Images," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2019.

[13] Y. Wu and R. E. Carson, "Noise Reduction in the Simplified Reference Tissue Model for Neuroreceptor Functional Imaging," *Journal of Cerebral Blood Flow & Metabolism*, vol. 22, no. 12, pp. 1440–1452, dec 2002.

[14] M. V. Cherkasova, L. Clark, J. J. Barton, M. Schulzer, M. Shafiee, A. Kingstone, A. J. Stoessl, and C. A. Winstanley, "Win-concurrent sensory cues can promote riskier choice," *Journal of Neuroscience*, 2018.

[15] J. C. Cheng, J. Matthews, V. Sossi, J. Anton-Rodriguez, A. Salomon, and R. Boellaard, "Incorporating HYPR de-noising within iterative PET reconstruction (HYPR-OSEM)," *Physics in Medicine and Biology*, 2017.

[16] B. Fischl, "FreeSurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–7781, aug 2012.

[17] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, jul 1990.

[18] C. W. Bevington, I. S. Klyuzhin, L. Aceves, D. Doudet, J. C. K. Cheng, and V. Sossi, "Denoising and DA release: Effect of denoising on the ability to identify voxel-level neurophysiological response," in *2018 IEEE Nuclear Science Symposium and Medical Imaging Conf. Rec.*, Sydney, Australia, 2018.

[19] M. Sakurada and T. Yairi, "Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction," in *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis - MLSDA'14*. New York, New York, USA: ACM Press, 2014, pp. 4–11.

[20] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," *Pattern Recognition*, vol. 58, pp. 121–134, oct 2016.

[21] R. Chalapathy, A. K. Menon, and S. Chawla, "Anomaly Detection using One-Class Neural Networks," feb 2018.

[22] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, P. N. Ossorio, S. Thadaney-Israni, and A. Goldenberg, "Do no harm: a roadmap for responsible machine learning for health care," *Nature Medicine*, vol. 25, no. 9, pp. 1337–1340, sep 2019.