

# BRAIN SUBTLE ANOMALY DETECTION BASED ON AUTO-ENCODERS LATENT SPACE ANALYSIS: APPLICATION TO *DE NOVO* PARKINSON PATIENTS

Nicolas Pinon<sup>1</sup>   Geoffroy Oudoumanessah<sup>\*1,2,3</sup>   Robin Trombetta<sup>1</sup>  
 Michel Dojat<sup>2</sup>   Florence Forbes<sup>3</sup>   Carole Lartizien<sup>1</sup>

<sup>1</sup> Univ. Lyon, CNRS UMR 5220, Inserm U1294, INSA Lyon, UCBL, CREATIS, France

<sup>2</sup> Univ. Grenoble Alpes, Inserm U1216, CHU Grenoble Alpes, Institut des Neurosciences, France

<sup>3</sup> Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, France

## ABSTRACT

Neural network-based anomaly detection remains challenging in clinical applications with little or no supervised information and subtle anomalies such as hardly visible brain lesions. Among unsupervised methods, patch-based auto-encoders with their efficient representation power provided by their latent space, have shown good results for visible lesion detection. However, the commonly used reconstruction error criterion may limit their performance when facing less obvious lesions. In this work, we design two alternative detection criteria. They are derived from multivariate analysis and can more directly capture information from latent space representations. Their performance compares favorably with two additional supervised learning methods, on a difficult *de novo* Parkinson Disease (PD) classification task.

**Index Terms**— Anomaly detection, Neuroimaging, Deep learning, Parkinson Disease, Mixture models, One-Class SVM.

## 1. INTRODUCTION

Most recent success of deep supervised learning, in the context of medical image analysis, critically depends on the availability of large sets of annotated images. The performance of supervised learning methods, on tasks such as anomaly detection, is then limited when the studied pathology is rare or when a fine expert annotation is required. A typical example is that of *de novo* (just diagnosed) Parkinson Disease (PD) patients, for which brain structural abnormalities are subtle and hardly visible in standard T1w or diffusion MR images. A natural alternative to supervised methods is *outlier detection* or *Unsupervised Anomaly Detection* (UAD). This formalism requires only the manual identification of "normal" data to construct a tractable model of normality, while *outliers* are then automatically detected as samples deviating from this normal model. Different categories of UAD methods have been applied to medical image segmentation or detection tasks. They mainly differ in the features

used to learn the normal model and the score computed to assess the distance to this model, which in brain anomaly detection is typically assessed at the voxel level. Illustrations include several auto-encoders (AE) architectures that have been compared in [1]. These AE models are trained to perform a "pretext" task on normal images consisting in the reconstruction of these images. For an arbitrary image, voxel-wise anomaly scores are then computed as the reconstruction errors, *i.e.* the differences between the image voxels and the reconstructed ones. Such errors are expected to be much larger for unseen voxels from patient images, provided the chosen architecture has initially well captured the normal subjects main features. To further investigate the importance of the normal model construction, building on this standard deep UAD formalism, we recently compared different auto-encoders architectures for the detection of subtle anomalies in the diffusion parametric maps of *de novo* PD patients [2]. This comparison included an auto-encoder (AE), taking 2D transverse slices as input, and the adaptation of a patch-based siamese auto-encoder (SAE) proposed in [3]. Our results demonstrated encouraging performance with the SAE model slightly outperforming the AE, thus indicating that patches may indeed be advantageous, in particular for their ability to capture local spatial neighborhood information around each voxel. However, as regards the detection score, the study also confirmed recent observations outlining the limitations of the reconstruction error scores for the detection of very subtle abnormalities [4]. In this work, we propose to investigate other detection procedures combining 1) enhanced normal models and 2) scoring rules derived from multivariate statistics. Following the approach reported in [3], we consider a patch-based approach but propose to perform the detection step in the latent space of the auto-encoder. More specifically, latent space representations of the normal images are extracted from the patch-based SAE of [2], and then used as features to build a normal model. Two types of models are considered, a non parametric discriminative one class support vector machine (OC-SVM) [5] and a parametric generative mixture model [6] (see Figure 1 and next section for details). So doing, the

\*Has equally contributed as the first author.

hope is to combine the representation power of patch-based AE networks to extract relevant and subtle features, with the efficiency of multivariate statistical models. These two combinations are then compared to a baseline UAD model based on the reconstruction error and to two standard *supervised* CNN, namely 3D ResNet and DenseNet.

## 2. UAD PIPELINE

The proposed framework for unsupervised brain anomaly detection is depicted on Figure 1. The central AE is first trained to learn the representation space of normal samples and reconstruct pseudo-normal images. The standard setting consists of computing *reconstruction error* maps (as the difference between the input and output images) on which anomalous unseen regions are expected to exhibit poor reconstructions or equivalently high anomaly scores. In this work, we also investigate two other outlier detection rules based respectively on a *generative* and a *discriminative* model designed to capture information from the AE latent space.

### 2.1. Latent space feature extraction

To construct an efficient normal model, we consider a patch-based network to enrich the latent space with local information at the voxel level. Leveraging the architecture proposed in [3], we use a SAE [7] composed of two replica of an auto-encoder sharing the same weights, associated to the following loss term :

$$L_{SAE}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{t=1}^2 \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|_2^2 - \alpha \cdot \cos(\mathbf{z}_1, \mathbf{z}_2)$$

which balances two objectives 1) decoding the representation  $\mathbf{z}$  learned from the encoder fed with patches  $\mathbf{x}$  into a reconstruction  $\hat{\mathbf{x}}$  that is close from the original patch  $\mathbf{x}$  and 2) having close (in the sens of the cosine similarity)  $\mathbf{z}$  for similar patches<sup>1</sup>.

### 2.2. Outlier detection in the latent space

As an alternative to the reconstruction error  $\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$  between a patch  $\mathbf{x}$  and its reconstruction  $\hat{\mathbf{x}}$ , we present below two outlier detection procedures built from a collection of normal patch representations  $(\mathbf{z}_i)_{1 \leq i \leq n}$  to account for normality in the latent space.

#### A discriminative approach: One-Class SVM.

The goal of the OC-SVM [8] is to construct a decision function  $f$ , positive on the estimated support of the distribution of normal samples  $\mathbf{z}_i$ , negative elsewhere and null on the frontier. The training samples from the normal class are first mapped to a higher dimensional space via a feature map  $\phi(\cdot)$  associated with a kernel  $k$  such that  $k(\mathbf{z}_i, \mathbf{z}_j) = \phi(\mathbf{z}_i) \cdot \phi(\mathbf{z}_j)$ . As the problem is linear in this redescription space, the parameters  $\mathbf{w}$  and  $\rho$  of the hyperplane  $\mathbf{w} \cdot \phi(\mathbf{z}) - \rho = 0$  are

obtained by solving a convex optimization problem aiming at maximizing the distance of the hyperplane from the origin.

The decision function can then be expressed as  $f(\mathbf{z}) = \mathbf{w} \cdot \phi(\mathbf{z}) - \rho$ . In a typical scenario, samples with negatives scores of  $f$  would be considered outliers. During inference,  $\mathbf{z}$  extracted from patches can be evaluated by the decision function to get an anomaly score corresponding to their distance to the hyperplane. This anomaly score, attributed to the central voxel of each patch, then provides an anomaly score map for the whole image. An ensemble of OC-SVM scores, trained on different  $\mathbf{z}_i$  is used to provide a more robust anomaly map.

#### A generative approach: multivariate mixtures.

While OC-SVM estimates only the support of the normal model, the goal here is to estimate the full normal distribution. To this end, we use a mixture model distribution  $p$ , denoted by  $\mathcal{MMST}$ , whose individual components are multiple scale t-distributions ( $\mathcal{MST}$ ):

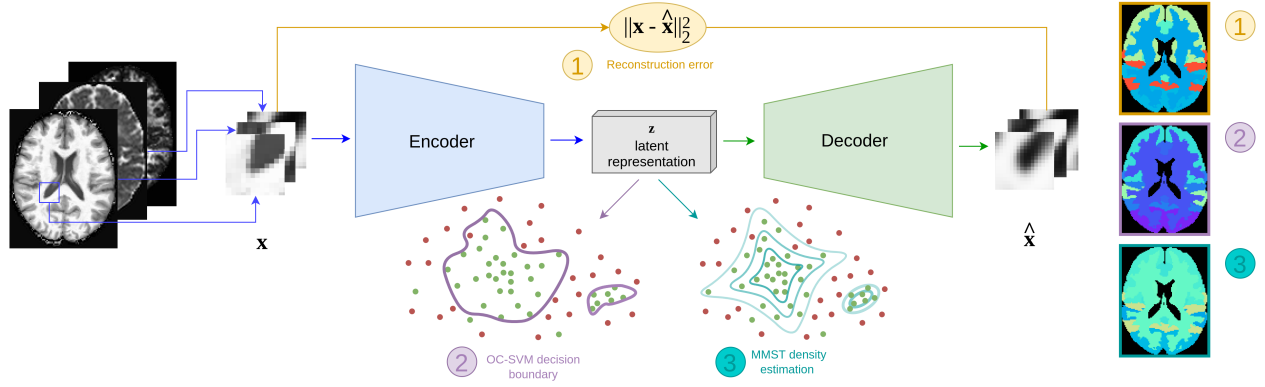
$$p(\mathbf{z}; \Theta) = \sum_{k=1}^K \pi_k \mathcal{MST}(\mathbf{z}; \theta_k) \\ \text{with } \Theta = (\pi_k, \theta_k)_{1 \leq k \leq K}, \pi_k \in [0, 1] \text{ and } \sum_{k=1:K} \pi_k = 1$$

$\mathcal{MST}$  distributions are generalizations of the multivariate t-distribution that extend its Gaussian scale mixture representation. The standard univariate scale variable is replaced by a  $M$ -dimensional scale variable  $(W_m)_{1 \leq m \leq M} \in \mathbb{R}^M$  where  $M$  denotes the latent space dimension. This allows a richer variety of shapes beyond elliptical distributions. The scale variable  $W_m$  for dimension  $m$  can be interpreted as accounting for the reliability of this dimension and is typically small when  $\mathbf{z}$  is far from the mean parameter. The specific definition can be found in [9]. Given a learning set of  $(\mathbf{z}_i)_{1 \leq i \leq n}$ , the estimation of the model parameter denoted by  $\hat{\Theta}_n$  is theoretically feasible using a standard expectation-maximization (EM) algorithm but is too time and memory costly in practice when the amount of data is large. In this work, we therefore resort to an *online* version of EM [10] that we derived for our  $\mathcal{MMST}$  model as detailed in [11]. Finally given a latent representation of a patch  $\mathbf{z}$ , we can use the scale variables to derive a measure of proximity  $f$  to the learned normal model:  $f(\mathbf{z}) = \max_{1 \leq m \leq M} \bar{w}_m^{\mathbf{z}}$ , with  $\bar{w}_m^{\mathbf{z}} = \mathbb{E}[W_m | \mathbf{z}; \hat{\Theta}_n]$ , where the expectation is computed for the learned  $\mathcal{MMST}$  model and is typically larger when at least one dimension of  $\mathbf{z}$  is well explained by the model. This measure of proximity, available for each voxel, provides in turn an anomaly score map for the whole image.

### 2.3. Post-processing of the anomaly maps

A threshold value (the *abnormality threshold*) set to an extreme quantile (eg. in the range of [90%, 100%]) of the anomaly scores distribution in the normal train samples was derived for each method (reconstruction error, encoder + OC-SVM and encoder +  $\mathcal{MMST}$ ) and applied to the test patient and test control dataset. The resulting binary anomaly maps can serve to identify suspect regions. To help evaluate the localization of these anomalies, two atlases were consid-

<sup>1</sup>In the case of learning on brain MR images, "similar" patches means that the patches are located in the same place in the brain, which is possible because all MRI's are registered to a common atlas beforehand.



**Fig. 1:** The trained encoder extracts latent representation  $\mathbf{z}$  of patches, used by 1) a decoder to compute reconstruction error in the image space 2) OC-SVM and 3)  $\mathcal{MMST}$  to perform outlier detection in the latent space. Anomaly maps representing the percentage of abnormal voxels per brain structures are shown on the right, warm colors corresponding to the highest percentages.

ered and fused: the Neuromorphometrics atlas [12] which segments the brain into 8 macro-regions and the MNI PD25 atlas [13] which is specifically designed for PD patients exploration and delineates 8 relevant subcortical structures (see Fig 2). The percentage of anomalous voxels was computed for each of these regions of interest leading to region-wise anomaly maps as depicted on the right of Figure 1.

### 3. EXPERIMENTS

#### 3.1. Data description and splitting

T1-weighted and DTI MR scans from 54 healthy controls and 124 *de novo* PD patients were extracted from the PPMI database [14]. All retrieved images were acquired with the same MR scanner model (3T Siemens Trio Tim). Mean diffusivity (MD) and fractional anisotropy (FA) maps were computed from DTI using MRtrix3.0. All maps  $X$  (T1w, FA, MD) were normalized in intensity with

$$X_{\text{norm}} = \frac{X - 1\% \text{quantile}(X)}{99\% \text{quantile}(X) - 1\% \text{quantile}(X)}$$
 with  $\chi$  being the intensity distribution of train controls images of one modality. All maps were non-linearly registered onto the MNI atlas resulting in images of dimension  $121 \times 145 \times 121$  with a voxel size of  $1.5 \times 1.5 \times 1.5 \text{ mm}^3$ . As for the cross-validation, healthy controls dataset was divided into 10 folds following a bootstrap procedure [15], leading to each fold containing [39, 41] train controls and [13, 15] test controls. The same procedure was performed with PD patients, leading to each fold containing [36, 40] train patients and [82, 86] test patients. Special care was put into balancing the age and sex distribution of each fold.

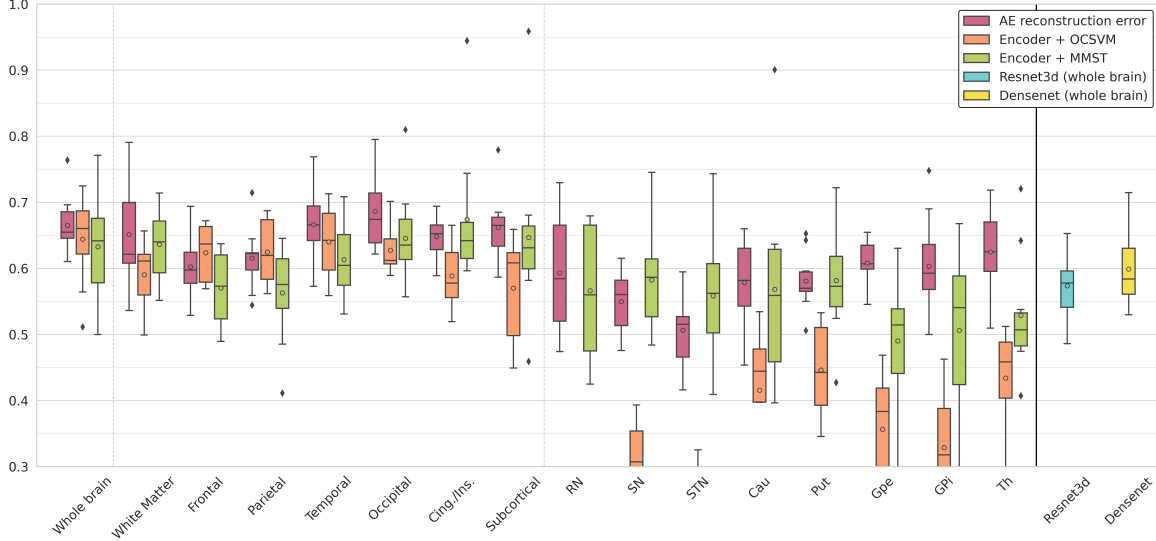
#### 3.2. Hyperparameters of the UAD pipeline

The encoder was composed of 4 convolutional blocks with kernel size (5, 5), (3, 3), (3, 3) and (3, 3), with strides respectively (1, 1), (1, 1), (3, 3) and (1, 1), number of filters respectively 3, 4, 12 and 16, no padding and GeLU activation. Each block was followed by a batch normalization block. The

decoder was the symmetric counterpart of the encoder. The input of the encoder consisted of the patches of each of the 3 modalities combined as channels. The SAE model was trained with [975000, 1025000] patches of size  $15 \times 15 \times 3$  (25 000 patches per subject). We used Adam optimizer [16] for 20 epochs, with default hyperparameters, best model selection based on validation loss and training batch size of 1000. An ensemble of five OC-SVM were trained, each with 500  $\mathbf{z}_i$  samples extracted from 500 random brain localizations from the train set and the mean of the 5 decision functions was used as the final anomaly score (note that this differs from [3] where one OC-SVM is trained per voxel). We used  $\nu = 0.03$  and a Gaussian kernel whose hyperparameter  $\frac{1}{\gamma}$  was set to the product of variance and dimension of the  $\mathbf{z}_i$ . For  $\mathcal{MMST}$ , we used  $K = 9$ . We set the *abnormality threshold* defined in section 2.3 to 98% (experiments have shown that the choice of this threshold has little influence on the final performance).

#### 3.3. Performance evaluation of the UAD models

Performance of the three methods was evaluated as in [6, 2]. The percentage of abnormal voxels in the whole brain or per region of interest derived from the post-processing of the anomaly score maps (see section 2.3) was employed to classify the test controls and test patients as healthy or pathological (PD). By varying a threshold on this metric, we can draw a ROC curve from the test population, and derive the best-achievable *g-mean score* defined as  $\sqrt{\text{Sensitivity} \times \text{Specificity}}$ . *g-mean score* is used as a performance metric to compare the different classification models. In the absence of reference annotations of the brain structures affected by the pathology, this pretext classification task allows to indirectly evaluate if the anomalies detected by the UAD models are characteristic of the pathology. It was computed either considering the percentage of anomalies in the whole brain, or in each of the regions of interest of the Neuromorphometric and MNI PD25 atlases.



**Fig. 2:**  $g$ -mean score of the 3 UAD and 2 CNN models. For UAD models, we consider anomaly % on the whole brain and per region, including the 8 subcortical structures from the MNI PD25 atlas: substantia nigra (SN), red nucleus (RN), subthalamic nucleus (STN), globus pallidus interna and externa (GPi, GPe), thalamus, putamen and caudate nucleus.

### 3.4. Comparison with supervised approaches

We compared classification performance of the three UAD models (reconstruction error, encoder + OC-SVM and encoder +  $MMST$ ) to that of two standard supervised 3D convolutional networks: 3D ResNet with 18 layers [17] and DenseNet-264 [18]. Each of these 2 CNN took as input the whole 3D T1w, MD and FA brain images combined as channels. A dense layer was added at the end of each network in order to have a one-dimensional output for classification. For each fold, the models were trained on 75% of train controls and train patients, the remaining 25% being kept for validation. Training was performed with Adam optimizer [16] for 300 epochs with default hyperparameters and a batch size of 8. Note that the train patients described in section 3.1 were only used for training of these two supervised networks. These two models were evaluated on the same test patient dataset as used for the UAD models thus enabling a fair comparison.

## 4. RESULTS

The  $g$ -mean score of each method is reported in Figure 2. We notice that the 3 UAD models achieve a median  $g$ -mean score around 0.65 on the whole brain, and in the range [0.6, 0.7] when only considering certain macro-regions (e.g. temporal or occipital lobe). For subcortical structures (e.g. RN or SN), performance drop to the range [0.5, 0.6] and even lower for some methods (especially Encoder + OC-SVM). At this stage of the PD progression, these subcortical structures seem slightly impacted. Note that the supervised methods, Resnet3D and Densenet, provide on the whole brain a median  $g$ -mean score in the range [0.55, 0.6], lower than the

UAD models considered in this study.

## 5. DISCUSSION AND CONCLUSION

Auto-encoders have shown to be a reference method regarding unsupervised anomaly detection [1] but have also shown limits when used for very subtle anomalies [4]. We have investigated whether an analysis of the latent space could improve these performances compared to a classical reconstruction error approach. We used two methods based on different paradigms: One-Class SVM (*discriminative*) and Mixture of Multiple scaled t-distributions (*generative*). It is clear from the supervised networks results that the proposed task, discriminating *de novo* PD from controls, is very hard: the supervised methods performances fall below the unsupervised methods ones, validating our approach. As seen with the performance of the reconstruction error, we found that the latent space UAD methods are strong competitors but do not surpass the former. In comparison with [2] where only diffusion was used, we report that the addition of T1w images does not improve significantly the performances.

We also demonstrated that using a patch-based encoder, as a feature extractor to feed a  $MMST$  model, gave promising results as it allows capturing some spatial context, which was lacking in [6]. Finally, the discrimination of PD based only on subcortical structures seems not feasible, as reported in [19] for substantia nigra, at an early stage of the pathology.

Future work includes investigating whether the combination of reconstruction error and latent space anomaly maps can increase the classification performance. We aim to extract 3D features with the auto-encoders and complete the multi-modal approach by adding T2w and T2\*w images as in [20].

## 6. ACKNOWLEDGMENTS

G. Oudoumanessah was financially supported by the AURA region. This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011012813R1 made by GENCI. It was partially funded by French program “Investissement d’Avenir” run by the Agence Nationale pour la Recherche (ANR-11-INBS-0006).

## 7. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available by the Parkinson Progression Markers Initiative (PPMI). Ethical approval was not required as confirmed by the license attached with the open access data.

## 8. REFERENCES

- [1] C. Baur, S. Denner, B. Wiestler, et al., “Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study,” *Medical Image Analysis*, vol. 69, 2021.
- [2] V. Muñoz-Ramírez, N. Pinon, F. Forbes, et al., “Patch vs. Global Image-Based Unsupervised Anomaly Detection in MR Brain Scans of Early Parkinsonian Patients,” in *Machine Learning in Clinical Neuroimaging*, Cham, 2021, pp. 34–43.
- [3] Z. Alaverdyan, J. Jung, R. Bouet, and C. Lartizien, “Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: Application to epilepsy lesion screening,” *Medical Image Analysis*, vol. 60, 2020.
- [4] F. Meissen, B. Wiestler, G. Kaissis, and D. Rueckert, “On the pitfalls of using the residual as anomaly score,” in *Medical Imaging with Deep Learning*, 2022.
- [5] M. El Azami, A. Hammers, J. Jung, et al., “Detection of lesions underlying intractable epilepsy on t1-weighted mri as an outlier detection problem,” *PLOS ONE*, vol. 11, no. 9, pp. 1–21, 09 2016.
- [6] A. Arnaud, F. Forbes, N. Coquery, et al., “Fully automatic lesion localization and characterization: Application to brain tumors using multiparametric quantitative mri data,” *IEEE TMI*, vol. 37, no. 7, 2018.
- [7] J. Bromley, I. Guyon, Y. LeCun, et al., “Signature verification using a ”siamese” time delay neural network,” in *NIPS*, 1993, p. 737–744.
- [8] B. Schölkopf, R. C. Williamson, A. Smola, et al., “Support Vector Method for Novelty Detection,” in *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K. Müller, Eds., 1999, vol. 12.
- [9] F. Forbes and D. Wraith, “A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering,” *Stat. and computing*, vol. 24, no. 6, pp. 971–984, 2014.
- [10] O. Cappé and E. Moulines, “On-line expectation–maximization algorithm for latent data models,” *JRSS B*, vol. 71, no. 3, pp. 593–613, 2009.
- [11] G. Oudoumanessah, M. Dojat, and F. Forbes, “Unsupervised scalable anomaly detection: application to medical imaging,” Research report, Oct. 2022, <https://hal.archives-ouvertes.fr/hal-03824951>.
- [12] R. Bakker, P. Tiesinga, and R. Kotter, “The Scalable Brain Atlas: Instant Web-Based Access to Public Brain Atlases and Related Content,” *Neuroinformatics*, vol. 13, pp. 353–366, 2015.
- [13] Y. Xiao, V. Fonov, S. Beriault, et al., “Multi-contrast unbiased MRI atlas of a Parkinson’s disease population,” *Int J Comput Assist Radiol Surg*, vol. 10, pp. 329–341, 2015.
- [14] Kenneth Marek, Sohini Chowdhury, Andrew Siderowf, et al., “The parkinson’s progression markers initiative (ppmi) - establishing a pd biomarker cohort,” *Annals of Clinical and Translational Neurology*, p. 1460–1477, 2018.
- [15] Russell A. Poldrack, Grace Huckins, and Gael Varoquaux, “Establishment of Best Practices for Evidence for Prediction: A Review,” *JAMA Psychiatry*, pp. 534–540, 2019.
- [16] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” in *ICLR (Poster)*, 2015.
- [17] D. Tran, H. Wang, L. Torresani, et al., “A closer look at spatiotemporal convolutions for action recognition,” in *2018 IEEE CVPR*, 2018, pp. 6450–6459.
- [18] G. Huang, Z. Liu, G. Pleiss, et al., “Convolutional networks with dense connectivity,” *IEEE PAMI*, 2019.
- [19] J. Prasuhn, M. Heldmann, T. F. Münte, and N. Brüggemann, “A machine learning-based classification approach on Parkinson’s disease diffusion tensor imaging datasets,” *Neurological Research and Practice*, vol. 2, no. 1, pp. 46, Dec. 2020.
- [20] S. Sivaranjini and C. M. Sujatha, “Deep learning based diagnosis of Parkinson’s disease using convolutional neural network,” *Multimedia Tools and Applications*, vol. 79, no. 21–22, pp. 15467–15479, June 2020.