MRI-BASED CLASSIFICATION OF NEUROPSYCHIATRIC SYSTEMIC LUPUS ERYTHEMATOSUS PATIENTS WITH SELF-SUPERVISED CONTRASTIVE LEARNING

Francesca Inglese^{*1}, Minseon Kim^{*2+},

Gerda M. Steup-Beekman³, Tom W.J. Huizinga³,

Mark van Buchem¹, Jeroen de Bresser¹, Daeshik Kim², and Itamar Ronen^{1†}

¹Department of Radiology, Leiden University Medical Center,

²School of Electrical Engineering, Korea Advanced Institute of Science and Technology,

³Department of Rheumatology, Leiden University Medical Center

ABSTRACT

Introduction/Purpose: Systemic lupus erythematosus (SLE) is a chronic auto-immune disease with a broad spectrum of clinical presentations, including heterogeneous and uncommon neuropsychiatric (NP) syndromes. Accurate diagnosis of neuropsychiatric SLE (NPSLE) is challenging due to lack of clinically useful biomarkers. Despite structural brain abnormalities on MRI in NPSLE being a common finding, a robust link between structural abnormalities and NPSLE has not been established, thus their contribution to the distinction between NPSLE patients and patients in which the NP symptoms are not primarily attributed to SLE is limited. Self-supervised contrastive learning algorithms do not require labels, and have been shown to be useful in classification tasks in rare diseases with limited number of datasets. The aim of our study was to apply self-supervised contrastive learning on T₁-weighted images acquired from a well-defined cohort of SLE patients to distinguish between SLE patients with NP symptoms due to the disease (NPSLE) or and SLE patients with similar symptoms due to other causes (non-NPSLE).

Subjects and Methods: 163 patients were included. We used 3T MRI T₁-weighted images registered to the MNI152 template. The training set comprised 68 non-NPSLE and 34 NPSLE patients. During the training procedure, we applied random geometric transformations (cropping, left-right flipping and rotations) between iterations to enrich our data sets. Our ML pipeline consisted of convolutional base encoder and linear projector. To test the classification task, the projector was removed and one linear layer was measured. We trained the encoder and projector with the Normalized Temperature-scaled Cross Entropy Loss (NT-xent) loss function. We performed a Monte Carlo validation that consisted of 6 repeated random sub-samplings each using a random selection of a small group of samples from each group.

Results: In the 6 trials described above, between 79% and 83% of the patients were correctly classified as NPSLE or non-NPSLE. For a qualitative evaluation of spatial distribution of the common features found in the NPSLE population, Gradient-weighted Class Activation Maps (Grad-CAM) were examined voxel-wise. Thresholded Grad-CAM maps show areas of common features identified for the NPSLE cohort, with no such communality found for the non-NPSLE group.

Discussion/conclusion: The self-supervised contrastive learning model was effective in capturing diagnostic brain MRI features from a limited but well-defined cohort of SLE patients with NP symptoms. The interpretation of the Grad-CAM results is not straightforward, but points to involvement of the lateral and third ventricles, periventricular white matter and basal cisterns. We believe that the common features found in the NPSLE population in this study indicate a combination of tissue loss, local atrophy and to some extent that of periventricular white matter lesions, which are commonly found in NPSLE patients and appear hypointense on T₁-weighted images.

Keywords Lupus Erythematosus, Systemic; Magnetic Resonance Imaging; Cohort Studies; Unsupervised Machine Learning; Neuroimaging.

^{*}Equal contribution + Current affiliation: Graduate School of AI, Korea Advanced Institute of Science and Technology

[†]Correspondence to: Itamar Ronen, LUMC/Radiology C3Q, Albinusdreef 2, Leiden 2333ZH, The Netherlands, i.ronen@lumc.nl

1 Introduction

Systemic lupus erythematosus (SLE) is a female-predominant auto-immune disease with a broad spectrum of clinical presentations and multi-organ involvement. SLE is characterized by the production and deposition of several autoantibodies [1]. The involvement of the central nervous system (CNS) in SLE leads to a series of non-specific neuropsychiatric (NP) manifestations in between 12-95% of SLE patients [2]. These NP symptoms widely range in terms of severity and prognostic implications [3]. NP events in SLE can be directly associated with the disease (NPSLE) or can be explained by another aetiology (non-NPSLE). NP symptoms are associated with an increased mortality and reduced quality of life within the SLE population [4]. The diagnosis of NPSLE is also difficult due to the heterogeneous nature of NP syndromes, which include headache, seizures, anxiety and psychosis [5], and the large variation in the attribution of the NP symptoms across studies and institutions highlights the difficulty in unequivocally diagnose NPSLE. In clinical practice, it is important to correctly classify NP events, since the therapeutic approach is defined based on this classification. A study performed in our center reported that about 15% of NP events attributed to SLE (NPSLE) during the first patient evaluation were reclassified after reassessment as non-NPSLE [6]. This discrepancy highlights the pressing need for biomarkers that will contribute to more reliably distinguish between NPSLE and non-NPSLE early in the diagnostic process.

Two different underlying mechanisms are thought to play a role in the pathophysiology of NPSLE. One is the inflammatory mechanism, where the blood-brain barrier (BBB) or the blood-cerebrospinal fluid (BCSF) barrier [7] is compromised due to presence of pro-inflammatory factors. Subsequently, auto-antibodies can enter the brain and trigger an inflammatory process that results in focal or diffuse tissue damage. The second proposed mechanism is the thrombotic or ischemic mechanism, where vascular injury and occlusion are present [8]. Due to the lack of a gold standard for the diagnosis, the best strategy so far for diagnosing NPSLE remains a multidisciplinary expert consensus after standardized evaluation of complaints and a complete battery of tests, including brain magnetic resonance imaging (MRI) [6]. Despite conventional brain MRI being the method of choice for clinical evaluation of SLE patients experiencing NP events, morphological changes and brain lesions observed in these patients do not clearly correlate with the clinical symptoms and disease outcome, underscoring the clinical-radiological paradox encountered with many NPSLE patients, defined by the presence of lesions in the absence of symptoms of NPSLE or viceversa [9].

Currently, MRI features can only contribute in a limited way in the diagnostic process, mostly in the way of exclusion. Several studies have shown that patients with SLE have more white matter hyperintensities (WMH) and more atrophy and infarcts compared to controls [10, 11, 12, 13, 14]. These findings per se, albeit indicative of robust presence of structural abnormalities in NPSLE, are not useful yet for the diagnostic process. Mostly – basic metrics such as global atrophy and lesion count/load do not lead to a specific diagnosis. It is therefore imperative to further explore neuroimaging biomarkers that can help clinicians to differentiate between NPSLE and non-NPSLE patients, and further down the line, also help in the stratification of NPSLE patients based on their clinical phenotype.

Deep learning has been shown to be useful in diagnostic tasks related to clinical neuroimaging data both in diseases with overt brain damage such as stroke, as well as in diseases in which brain alterations are not directly detectable in standard radiological observation [15]. Deep learning models can extract significant features that are relevant in clinical diagnosis and can distinguish between patient populations even when the brain alterations are not visibly overt [16]. Classification tasks, however, require a large number of data sets, as well as trained clinicians to generate labels to aid the categorization process. Dementia [17, 18, 19, 20] and psychiatric disorders [21, 22] have been natural targets for using deep neural networks as imaging data for these diseases are widely available. NPSLE, on the other hand, is a sub-category of SLE, which in itself is categorized as an orphan/rare disease (prevalence of 1-5 in 10,000, source: www.orpha.net) and thus the amount of data available is limited. This makes a supervised ML approach impractical for studying brain abnormalities in NPSLE in a single-center study.

Recently, self-supervised learning approaches, which train the model on unlabeled data by providing self-generated labels from the data themselves, have become popular in image classification. In non-medical applications, self-supervised learning approaches were applied in the prediction of the rotation angles of objects [23], colorization of gray-scale images [24] and solving randomly generated Jigsaw puzzles [25]. Instance-level identity preservation with contrastive learning has shown to be effective in learning rich representations for classification [26, 27, 28]. In this context, self-supervised learning approaches are more suitable for dealing with limited data sets, such as the one presented in this work. Self-supervised learning in biomedical imaging has been implemented in several instances, among which screening of 2-dimensional chest x-ray images [29], in the evaluation of cardiac time-series data [30], in tissue segmentation of brain lesions [31], in segmentation of the renal dynamic contrast-enhanced MRI [32], in robust and accelerated reconstruction of quantitative and B_0 -inhomogeneity-corrected R^*_2 maps from multi-gradient recalled echo MRI data [33] and in quality enhancement of compressed sensing MRI of vessel wall [34].

In this study, we hypothesized that a self-supervised learning approach would be effective for the classification tasks in our limited patient population, in particular in the distinction between two important diagnostically different SLE patient groups: NPSLE and non-NPSLE patients. To test this hypothesis, we applied a self-supervised method to 3D structural MRI data with the aim of distinguishing and classifying such data for NPSLE and non-NPSLE patients.

2 Methods

2.1 Patient population

The Leiden University Medical Center (LUMC) is the national referral center in the Netherlands for SLE patients with NP complaints. SLE Patients are referred to the outpatient clinic if they present with NP manifestations. In this retrospective study we included 216 patients with SLE recorded between May 2007 and April 2015. Of these, 28 patients were excluded because of undefined diagnosis, 3 patients were excluded because of motion artefacts in the MRI scan, 20 patients were excluded because of brain infarcts over 1.5 cm and 2 patients were excluded due to the presence of other diseases (one for a brain tumor and one for a large arachnoid cyst). This resulted in a total of 163 patients included in this study. The medical ethics committee of Leiden-The Hague-Delft approved of the study and all included patients signed for informed consent.

All patients came to the clinic for a full one-day visit and underwent an identical standardized assessment that included a brain MRI scan [35] and a combination of multidisciplinary medical assessments and extensive complementary tests, necessary for deciding whether the NP-events are attributed to SLE [8]. Attribution of NP symptoms to SLE was established during a multidisciplinary consensus meeting. This diagnostic process is described in detail previously [35]. NP events were classified according to the 1999 ACR nomenclature for NPSLE [36].

During an intake interview, information about gender, age, and SLE disease duration was provided by the patients and verified by their medical records. During the evaluation, SLE activity and damage indexes were scored for each patient: the SLE disease activity was defined using the Systemic Lupus Erythematosus Disease Activity Index 2000 (SLEDAI-2K) [37]; SLE irreversible damage was determined through the Systemic Lupus International Collaborating Clinics/American College of Rheumatology damage index (SDI) [38].

2.2 MRI protocol

All patients were scanned, according to a standardized scanning protocol, on a Philips Achieva 3T MRI scanner (Philips Healthcare, Best, The Netherlands) with a body transmit RF coils and an 8-Channel head coil array. The sequence used for this project was a 3D T₁-weighted gradient echo scan (voxel size = $1.17 \times 1.17 \times 1.2\text{mm}^3$; TR/TE = 9.8/4.6ms).

2.3 MRI preprocessing

All the T₁-weighted images were registered to a standard brain template, the Montreal Neurological Institute standard template (MNI152), using FNIRT (FMRIB's non-linear image registration tool) [39, 40], using affine registration with 12 degrees of freedom.

2.4 Machine learning pipeline architecture

We followed the self-supervised framework introduced by Chen et al. [26], where an encoder network f_θ was used to project the image into a feature space, followed by two-layer multi-perceptron (MLP) p_π projector that projected the features into latent vector z . In our work we modified this approach to address the fact that 3D MRI data requires more classification parameters than 2D natural images. Therefore, we designed the encoder network f_θ using three convolutional layers with batch normalization and a max pooling layer. To test the representation feature, we changed the projector layer into a linear layer which had the same output size as the number of classes, which in our case equals 2 – NPSLE and non-NPSLE. Subsequently, we fine-tuned the linear layer with a training set. Finally, we tested the accuracy of the trained encoder and linear layer. Figure 1 shows the architecture described above.

2.5 Preprocessing

In 2D natural images, Chen et al. [26] used stochastic data augmentation t , randomly selected from the family of augmentation T , including random cropping, random color distortion, and random flip. In our 3D MRI data, we used stochastic data augmentation, by performing random selection from the set of augmentations we applied to our data. These included: random cropping, random flipping along the z axis and random in-plane rotation. The random

MRI-based Classification of Neuropsychiatric Systemic Lupus Erythematosus Patients with Self-supervised Contrastive Learning

Figure 1: **Architecture of ML pipeline: encoder, projector and linear module.** In the training phase, an encoder and projector were used to project the images into representation space with a latent vector z . In the fine tuning and test phase, the projector was changed with a linear module and predict the class.

Figure 2: **3D data transformation.** Three types of geometrical transformations were applied to the 3D MRI data augment the data set: cropping of the image, left-right flips, and in-plane rotations.

crop was applied up to 15 voxels along the three axes. For the random flip, only left-right flips were applied based on the foot-head axis (coronal plane). For rotation, random rotation of angles up to 45 degrees was applied in the left-right, anterior-posterior and foot-head directions. To test the robustness of our method with respect to different data augmentation strategies, we tried three different augmentation settings: one only with crop, one with crop, flip and rotation. The geometric transformations are depicted in Figure 2.

2.6 Contrastive loss

The contrastive loss function $\mathcal{L}_{\text{con}}$ is defined as follows,

$$\mathcal{L}_{\text{con}, \theta, \pi}(x, \{x_{\text{pos}}\}, \{x_{\text{neg}}\}) := -\log \frac{\sum_{\{z_{\text{pos}}\}} \exp(\text{sim}(z, \{z_{\text{pos}}\})/\tau)}{\sum_{\{z_{\text{pos}}\}} \exp(\text{sim}(z, \{z_{\text{pos}}\})/\tau) + \sum_{\{z_{\text{neg}}\}} \exp(\text{sim}(z, \{z_{\text{neg}}\})/\tau)}, \quad (1)$$

where $z, \{z_{\text{pos}}\}$, and $\{z_{\text{neg}}\}$ are corresponding 128 dimensional representation vectors (z) of x obtained by the encoder and projector $z = p_{\pi}(f_{\theta}(x))$. The $\text{sim}(u, v) = u^T v / \|u\| \|v\|$ denotes cosine similarity between two vectors and τ is a temperature parameter [26]. We trained the encoder and projector with the contrastive loss function, *NT-xent* which maximizes the similarity between each transformed sample.

2.7 Patients selection for validation of the ML pipeline

To determine the accuracy of our study, six trials were performed. In each trial, the training set consisted of 68 non-NPSLE and 34 NPSLE patients randomly chosen from within the total patient population (163 subjects). In order to have an equal number of NPSLE and non-NPSLE patients for the training procedure, we used the images of NPSLE twice in every epoch. Therefore, a total of 136 images were used to train the model. In each trial, the test set consisted

MRI-based Classification of Neuropsychiatric Systemic Lupus Erythematosus Patients with Self-supervised Contrastive Learning

Table 1: **Patient population characteristics.** Sex, age and SLE clinical variables are described for each trial. SLEDAI= systemic Lupus Erythematosus Disease Activity Index 2000. SDI= Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index.

(a) Patient population characteristics for the training sets.

	Trial 1 (n=102)	Trial 2 (n=102)	Trial 3 (n=102)	Trial 4 (n=102)	Trial 5 (n=102)	Trial 6 (n=102)
Sex (female,%)	90%	91%	91%	93%	90%	89%
Age (mean±SD)	44±13	41±12	41±12	41±12	41±14	43±13
SLE duration (mean±SD)	7±8	7±8	7±7	8±8	7±8	7±8
SLEDAI-2K (mean±SD)	2±2	2±2	2±2	2±2	2±2	2±2
SDI (mean±SD)	1±1.2	0.8±1	0.8±1	0.8±1	0.9±1.1	0.9±1.1

(b) Patient population characteristics for the testing sets.

	Trial 1 (n=18)	Trial 2 (n=18)	Trial 3 (n=18)	Trial 4 (n=18)	Trial 5 (n=18)	Trial 6 (n=18)
Sex (female,%)	100%	89%	94%	89%	100%	100%
Age (mean±SD)	34±9	46±16	41±10	47±15	38±11	37±13
SLE duration (mean±SD)	6±7	5±6	9±10	3±5	8±8	7±8
SLEDAI-2K (mean±SD)	2±1	1±1	2±2	2±2	3±2	3±1
SDI (mean±SD)	0.83±0.9	1.6±1.5	1±1.2	0.8±1.3	0.7±1.1	0.6±0.8

of 9 non-NPSLE patients and 9 NPSLE patients randomly chosen within the total patient’s population (163 subjects) excluding the training set. The demographic and clinical characteristics of the patients selected in each trial are shown in Tables 1.

2.8 Training details

We used three convolutional layers as the base encoder network and 2-layer multi-layer perceptron with 128 embedding dimensions as the projection head. All models were trained by minimizing the final contrastive loss with a temperature of $\tau = 0.5$. For the rest, we followed similar optimization steps as in SimCLR [26]. We trained with 1,000 epochs under the stochastic gradient descent (SGD) base Layer-wise Adaptive Rate Scaling (LARS) optimizer [41], a cosine annealing learning rate and a gradual warmup scheduler [26]. We used a weight decay of $1e - 6$ and momentum of 0.9. We used linear warm-up for the first 10 epochs until a learning rate of 1.0 was achieved and decay with cosine decay schedule. We used a batch size of 16. Furthermore, we used global batch normalization, which shared the parameters over the multiple GPUs. We trained the encoder with contrastive loss. Then, we fine-tuned the encoder along a linear layer to optimize the cross-entropy loss with learning rate 0.001 during 35 epochs.

2.9 Gradient class activation mapping

There are several methods for interpreting predictions in deep learning [42, 43, 44, 45, 46, 47]. We chose to use gradient class activation mapping (grad-CAM) [42]. Grad-CAM uses gradient information flowing into the last convolutional layer of the model to assign importance values to each parameter for the prediction of the model. We extended the grad-CAM to deal with 3D convolutional neural networks. Since we are interested in a classification for a specific disease, and images are all aligned to the same space (MNI 152), it was possible to average the grad-CAM maps of test samples to obtain a qualitative measure of the brain regions where significant features were found after averaging across subjects in each trial. For this purpose, for each trial, the grad-cam maps of the samples which were correctly classified in the same class were averaged. Thresholded averaged grad-CAM maps were then overlaid on the MNI template for display.

2.10 Statistics

For evaluation of the performance of the algorithm at each trial, we used classification *accuracy*, *precision* and *recall* as quantitative measures for the performance of the algorithm. An image is considered correctly classified when the model classifies it into the correct class: NPSLE or non-NPSLE. Accuracy was defined as the ratio between the total number of correctly classified samples and the total number of test sets. The coefficient of variation for the three augmentation

Table 2: Results of accuracy, precision and recall.

(a) **Classification accuracy per trial, for all three data augmentation strategies.** Data are shown as percentage. Numbers in parentheses stand for the number of correctly classified non-NPSLE and NPSLE subjects, respectively.

Random data augmentation Set	T	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Percentage (SD)
Crop	$\{T_1\} \in T$	61.11 (6,5)	77.78 (8,6)	77.78 (5,9)	100.0 (9,9)	88.89 (7,9)	94.44 (9,8)	83.35 (14.05)
Crop, Flip	$\{T_1, T_2\} \in T$	61.11 (7,4)	66.67 (6,6)	94.44 (8,9)	83.33 (7,8)	94.44 (8,9)	100.00 (9,9)	83.33 (16.10)
Crop, Flip, Rotation	$\{T_1, T_2, T_3\} \in T$	66.67 (6,6)	72.22 (8,5)	77.78 (6,8)	100.0 (9,9)	77.78 (8,6)	79.63 (7,8)	79.63 (11.47)

(b) **Classification precision and recall per trial, for all three data augmentation strategies.** Data are shown as percentage, and each pair of numbers represents the precision and recall values, respectively.

Random data augmentation Set	T	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Percentage (SD)
Crop	$\{T_1\} \in T$	45.5/ 55.6	42.9/ 66.7	64.3/ 100.0	50.0/ 100.0	56.3/ 100.0	44.4/ 88.9	50.6 (8.0)/ 85.2 (19.0)
Crop, Flip	$\{T_1, T_2\} \in T$	36.4/ 44.4	50.0/ 66.7	52.9/ 100.0	53.3/ 88.9	52.9/ 100.0	50.0/ 100.0	49.3 (6.0)/ 83.3 (23.0)
Crop, Flip, Rotation	$\{T_1, T_2, T_3\} \in T$	50.0/ 66.7	38.5/ 55.6	57.1/ 88.9	50.0/ 100.0	42.9/ 66.7	52.9/ 100.0	48.6 (7)/ 79.7 (19)

strategies was also calculated by taking the ratio of the mean accuracy and the standard deviation. Precision was defined as the fraction of correctly predicted NPSLE samples out of the total test samples (NPSLE + non-NPSLE). The recall, defined as the fraction of relevant items selected from the interest class, as calculated as the fraction of correctly predicted NPSLE samples out of the total number of NPSLE samples.

To assess differences in mean accuracy, precision and recall among the three different geometric transformation approaches we used here, repeated-measures analysis of variance (ANOVA) was performed using the Statistical Package for the Social Sciences (SPSS) version 25 (IBM corporation, Armonk, NY, USA).

3 Results

3.1 Classification performance

Table 2a shows the individual and mean accuracy of the classification results for the six trials, defined as the percent of correctly classified NPSLE/non-NPSLE patients out of the total tests within a trial. Results are given for three different strategies for data augmentation that included one (random crop), two (random crop + flip) and three (crop + flip + rotate) transformations. The accuracy(SD) of the six trials in each augmentation strategy was 83.35% (14.05%), 83.33% (16.10%) and 79.63% (11.47%). No significant differences were found when comparing the accuracy of the classification across the three data augmentation options. The coefficient of variation for the three augmentation strategies ranged between 0.14 and 0.19.

Table 2b shows the individual and mean precision and recall of the 6 classification trials. Precision is defined as percent of correctly predicted NPSLE cases out of the total number of tests within a trial, and recall is defined as the fraction of correctly predicted NPSLE samples out of the total number of NPSLE test samples within a trial. The same three sets of geometric transformation for data augmentation were used also here. The precision (SD) across trials was 50.6% (8%), 49.3% (6%), and 48.6% (7%) for trials using only crop, crop + flip, and crop + flip + rotation transformation, respectively. Recall (SD) were 85.2% (19%), 83.3% (23%), and 79.7% (19%) for trials using only crop, crop + flip, and crop + flip + rotation transformation, respectively.

No significant differences were found in accuracy ($p=0.531$), precision ($p=0.845$) and recall ($p=0.686$) among trials using the 3 different geometric transformations.

3.2 Common features in NPSLE - Grad-CAM results

Figures 3a through 3c show the thresholded averaged grad-CAM maps for the 6 trials. Threshold was set at three different levels: 0.75, 0.85 and 0.95. Each of the three figures shows the results following data augmentation with crop only (3a), crop + flip (3b) and crop + flip + rotate (3c). Common features that show up on the grad-CAM maps were found only in the NPSLE cohort. The areas generated by the grad-CAM maps with threshold set at 0.75 are too generic to report on specific brain regions, while the higher thresholds of 0.85 and 0.95 reveal more parcellated maps indicating local involvement. Brain regions included in the grad-CAM maps with threshold above 0.85 are lateral ventricles and periventricular white matter, as well as third ventricle and basal cisterns (Figure 4). There were no specific regions that contributed to the model's prediction of non-NPSLE.

(a) Grad-CAM results of NPSLE patients for the trials augmented with "crop". Each row displays results from one trial. Columns display three different Grad-CAM thresholds: 0.75 (left), 0.85 (middle) and 0.95 (right).

MRI-based Classification of Neuropsychiatric Systemic Lupus Erythematosus Patients with Self-supervised Contrastive Learning

(b) Grad-CAM results of NPSLE patients for the trials augmented with “crop” and “LR flip”. Each row displays results from one trial. Columns display three different Grad-CAM thresholds: 0.75 (left), 0.85 (middle) and 0.95 (right).

(c) Grad-CAM results of NPSLE patients for the trials augmented with “crop”, “LR flip” and “rotation”. Each row displays results from one trial. Columns display three different Grad-CAM thresholds: 0.75 (left), 0.85 (middle) and 0.95 (right).

Figure 3: Grad-CAM results with different data augmentation set.

Figure 4: Brain areas that correspond with regions highlighted in the grad-CAM maps with threshold above 0.85.

4 Discussions

In this study, we designed a self-supervised machine-learning pipeline for classification of T1-weighted MRI images, aimed at distinguishing between images of NPSLE patients and those of non-NPSLE patients. The accuracy of the classification algorithm, based on 6 repeated trials, was significantly above random choice, and practically independent of the augmentation strategy. The mean classification accuracy into the two classes, NPSLE and non-NPSLE, across the three augmentation strategies ranged between 79% and 83%. Within-augmentation variability across trials was well contained: the coefficient of variation for the three augmentation strategies ranged between 0.14 and 0.19, indicating a good repeatability of the accuracy of the classification, despite the significant heterogeneity of clinical measures and the symptoms in the NPSLE population.

The overall precision of our model, representing the fraction of correctly predicted NPSLE out of the total number of NPSLE predictions, was lower, at about 51.8%. This indicates a relatively high rate of false negatives (non-NPSLE subjects identified as NPSLE). This may indicate that the structural brain changes characteristic of NPSLE and were picked up by the classification algorithm can also be found in non-NPSLE patients. MRI abnormalities, such as lesions, local atrophy and other diffuse abnormalities have been found in non-NPSLE and as well as in SLE patients without NP when compared with healthy controls [10, 11, 12, 13, 14]. It is thus not surprising that some of the features picked up in the training of the algorithm were erroneously attributed to non-NPSLE patients. It remains to be seen whether with increased population size (via, for example, a multicenter effort), or additional MRI modalities that reflect better other aspects of structural changes in the disease (for example, FLAIR T2 images that report on white matter hyperintensities) will contribute to the precision of the classification and limit the false-positives. Conversely, the recall, defined as the fraction of total relevant results (correctly predicted NPSLE patients) out of the NPSLE group, averaged at about 83% indicating a relatively low rate of false negatives, i.e. NPSLE patients that were not classified as such. This corroborates the meaningfulness of the features found by the ML algorithm and provides support to their link to brain changes in NPSLE population.

We studied the relationship between brain alterations in NPSLE and the common features identified by the classification algorithm with grad-CAM, a commonly used visualization tool that provides a coarse localization map highlighting important regions in the image for the classification task. As of recent, Grad-CAM maps have been applied to medical imaging modalities, including a successful application aimed at grading gliomas based on MR images [48, 49]. While grad-CAM maps do not provide quantitative statistical information in the way that statistical parametric maps do, they do indicate communality in the features that led to the classification. In particular, basal cistern, lateral ventricles, third ventricle and periventricular white matter seemed to be features that discriminate NPSLE patients from non-NPSLE. Studies on structural MRI pre [50] and post-mortem [51] of brains of NPSLE patients showed small focal lesions and white matter hyperintensities concentrating on periventricular white matter, as well as ventricular dilation. Higher occurrence of periventricular and deep WMH lesions was reported in SLE patients compared to controls but without stratification for NPSLE versus non-NPSLE patients [52]. Overall, the etiology of periventricular WMH in NPSLE is not well understood, not is it fully investigated in other diseases with prevalence of periventricular WMH. In older adults, periventricular WMH appear to be associated with impaired cognitive function, in particular with working memory, and

MRI-based Classification of Neuropsychiatric Systemic Lupus Erythematosus Patients with Self-supervised Contrastive Learning

are linked to disruption of long distance white matter connections. Some characterization of periventricular WMH in older adults was provided by diffusion tensor imaging and pathological observations and revealed that periventricular WMH are mostly characterized by gliosis and myelin loss [53]. Further investigation on the role of periventricular WMH in NPSLE patients is necessary to confirm their role and their importance in the disease process. There is no consistent reporting on direct involvement of the basal cisterns in SLE or NPSLE, barred few case reports [54, 55]. A plausible explanation for the involvement of the basal cisterns is the effect of atrophy, leading to increase in CSF volume in the basal cisterns as well as in the lateral ventricles.

There are several limitations and challenges associated with our study, and we hope to address some of them in our future efforts. Characterization of the overlapping features between the false negatives and “true” NPSLE subjects will certainly require special attention in future studies, with significantly larger populations of SLE and non-NPSLE patients, and with an inclusion of control groups of SLE patients without NP, as well as of healthy controls.

Despite the fact that there were common regions highlighted by Grad-CAM in the NPSLE group, it is premature to claim at this point that these regions are indeed clinically significant, and a broader investigation is required. In the current investigation we used only one MRI modality, namely T₁-weighted images, a modality that is highly sensitive to volumetric structural changes, but less sensitive and less specific to lesions, infarcts, microbleeds and hemorrhages, which all result in local hypointensities. Thus, it is imperative to continue the investigation with a more multimodal approach, with additional modalities that will add more sensitivity and specificity to a variety of structural changes commonly found in SLE and NPSLE. From the algorithm perspective, self-supervised machine learning is identity preserving, and it is therefore possible to add a variety of MRI modalities to the process in the hope of significantly improving the classification performance.

In deep learning classification tasks, a large amount of data sets is essential. The number of samples (overall number of patient data sets) that were included in our study (163) was limited compared to typical numbers of samples used in classification tasks, typically in the tens of thousands of cases [29]. NPSLE is a rare and highly heterogeneous disease and it is therefore not a natural target for unsupervised machine learning based classification. The goal of this study was to establish a retrospective link between (known) diagnosis and structural brain differences between two classes of samples: NPSLE and non-NPSLE patients. Within the limits of a single-center study we benefitted from the maximum number of patients available in the Netherlands, as well as from the most comprehensive diagnostic process for NPSLE (being a national referral center for the disease). When attempting to use a 2 class-wise supervised learning approach, the model diverged, possibly due to the limited number of data sets and the inconspicuousness of the visual features. Based on the fact that self-supervised learning is known to perform well even with a small data sets [56, 57], we opted for self-supervised learning, in the hope that following the classification, common features for two independent classes will emerge, coinciding with the NPSLE and non-NPSLE patient groups. Eventually, the algorithm was only able to find common features within one class of patients (NPSLE), and no common features were found in the non-NPSLE group in our trials. We cannot claim as a certainty that there are no common features to the non-NPSLE group, but the positive result we obtained for the classification of the NPSLE group supports the notion that in our classification task it was more effective to allow the algorithm to learn the visual representation in a self-supervised manner using similarity across images, rather than providing a-priori class labels.

To conclude: we set the stage for classification of brain imaging data of NPSLE and non-NPSLE patients using deep neural networks, achieving relatively high average accuracy across repeated trials. We showed that self-supervised learning is capable of capturing common image features in one class of subjects (NPSLE), a task that is impossible to accomplish with supervised learning, that can capture relatively inconspicuous visual information by cross-entropy loss in the MRI images, and is advantageous when only a limited number of data sets is available. The method is modular and can accommodate additional imaging modalities to be included in the classification, and can be easily applied to other studies of rare diseases that suffer from similar limitations.

Funding

Author MK was supported by funding provided by Brain Korea 21+ Project, BK Electronics and Communications Technology Division, KAIST in 2019.

Acknowledgement

We thank Dr. Sahar Yousefi for her critical reading of the manuscript and her useful comments on the ML methodological aspects. We also thank Jihoon Tack for his help in implementing the deep learning model and useful comments on the ML development aspects.

References

- [1] John G Hanly, Elizabeth Kozora, Steven D Beyea, and Julius Birnbaum. Nervous system disease in systemic lupus erythematosus: current status and future directions. *Arthritis & rheumatology*, 71(1):33–42, 2019.
- [2] Hanna Ainiala, J Loukkola, J Peltola, Migraine Korpela, and A Hietaharju. The prevalence of neuropsychiatric syndromes in systemic lupus erythematosus. *Neurology*, 57(3):496–500, 2001.
- [3] Noa Schwartz, Ariel D Stock, and Chaim Putterman. Neuropsychiatric lupus: new mechanistic insights and future treatment directions. *Nature Reviews Rheumatology*, 15(3):137–152, 2019.
- [4] Amy E Ramage, Peter T Fox, Robin L Brey, Shalini Narayana, Matthew D Cykowski, Mohammad Naqibuddin, Margaret Sampedro, Stephen L Holliday, Crystal Franklin, Daniel J Wallace, et al. Neuroimaging evidence of white matter inflammation in newly diagnosed systemic lupus erythematosus. *Arthritis & Rheumatism*, 63(10):3048–3057, 2011.
- [5] John G Hanly. Diagnosis and management of neuropsychiatric sle. *Nature Reviews Rheumatology*, 10(6):338–347, 2014.
- [6] César Magro-Checa, Els J Zirkzee, Liesbeth JJ Beart-van de Voorde, Huub A Middelkoop, Nic J van der Wee, Menno V Huisman, Jeroen Eikenboom, Nyika D Kruij, Mark A van Buchem, Tom WJ Huizinga, et al. Value of multidisciplinary reassessment in attribution of neuropsychiatric events to systemic lupus erythematosus: prospective data from the leiden npsle cohort. *Rheumatology*, 56(10):1676–1683, 2017.
- [7] Sivan Gelb, Ariel D Stock, Shira Anzi, Chaim Putterman, and Ayal Ben-Zvi. Mechanisms of neuropsychiatric lupus: The relative roles of the blood-cerebrospinal fluid barrier versus blood-brain barrier. *Journal of autoimmunity*, 91:34–44, 2018.
- [8] César Magro-Checa, Elisabeth J Zirkzee, Tom W Huizinga, and Gerda M Steup-Beekman. Management of neuropsychiatric systemic lupus erythematosus: current approaches and future perspectives. *Drugs*, 76(4):459–483, 2016.
- [9] César Magro-Checa, Gerda M Steup-Beekman, Tom W Huizinga, Mark A Van Buchem, and Itamar Ronen. Laboratory and neuroimaging biomarkers in neuropsychiatric systemic lupus erythematosus: where do we stand, where to go? *Frontiers in medicine*, 5:340, 2018.
- [10] J Luyendijk, SCA Steens, WJN Ouwendijk, GM Steup-Beekman, ELEM Bollen, J Van Der Grond, TWJ Huizinga, BJ Emmer, and MA van Buchem. Neuropsychiatric systemic lupus erythematosus: lessons learned from magnetic resonance imaging. *Arthritis & Rheumatism*, 63(3):722–732, 2011.
- [11] H Ainiala, P Dastidar, J Loukkola, T Lehtimäki, M Korpela, J Peltola, and A Hietaharju. Cerebral mri abnormalities and their association with neuropsychiatric manifestations in sle: a population-based study. *Scandinavian journal of rheumatology*, 34(5):376–382, 2005.
- [12] Simone Appenzeller, Aline Daiane Carnevale, Li Min Li, Lilian TL Costallat, and Fernando Cendes. Hippocampal atrophy in systemic lupus erythematosus. *Annals of the rheumatic diseases*, 65(12):1585–1589, 2006.
- [13] Simone Appenzeller, Leonardo Bonilha, Pablo A Rio, Li Min Li, Lilan Tereza Lavras Costallat, and Fernando Cendes. Longitudinal analysis of gray and white matter loss in patients with systemic lupus erythematosus. *Neuroimage*, 34(2):694–701, 2007.
- [14] Elizabeth Kozora and Christopher M Filley. Cognitive dysfunction and white matter abnormalities in systemic lupus erythematosus. *Journal of the International Neuropsychological Society: JINS*, 17(3):385, 2011.
- [15] Li Zhang, Mingliang Wang, Mingxia Liu, and Daoqiang Zhang. A survey on deep learning for neuroimaging-based brain disorder analysis. *Frontiers in neuroscience*, 14, 2020.
- [16] Jyoti Islam and Yanqing Zhang. Early diagnosis of alzheimer’s disease: A neuroimaging study with deep learning architectures. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1881–1883, 2018.
- [17] Kangan Oh, Young-Chul Chung, Ko Woon Kim, Woo-Sung Kim, and Il-Seok Oh. Classification and visualization of alzheimer’s disease using volumetric convolutional neural network and transfer learning. *Scientific Reports*, 9(1):1–16, 2019.
- [18] Taeho Jo, Kwangsik Nho, and Andrew J Saykin. Deep learning in alzheimer’s disease: diagnostic classification and prognostic prediction using neuroimaging data. *Frontiers in aging neuroscience*, 11:220, 2019.
- [19] Daniel Stamate, Richard Smith, Ruslan Tsygancov, Rostislav Vorobev, John Langham, Daniel Stahl, and David Reeves. Applying deep learning to predicting dementia and mild cognitive impairment. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 308–319. Springer, 2020.

MRI-based Classification of Neuropsychiatric Systemic Lupus Erythematosus Patients with Self-supervised Contrastive Learning

- [20] Silvia Basaia, Federica Agosta, Luca Wagner, Elisa Canu, Giuseppe Magnani, Roberto Santangelo, Massimo Filippi, Alzheimer's Disease Neuroimaging Initiative, et al. Automated classification of alzheimer's disease and mild cognitive impairment using a single mri and deep neural networks. *NeuroImage: Clinical*, 21:101645, 2019.
- [21] Daniel Durstewitz, Georgia Koppe, and Andreas Meyer-Lindenberg. Deep neural networks in psychiatry. *Molecular psychiatry*, 24(11):1583–1598, 2019.
- [22] Weiming Lin, Tong Tong, Qinquan Gao, Di Guo, Xiaofeng Du, Yonggui Yang, Gang Guo, Min Xiao, Min Du, Xiaobo Qu, et al. Convolutional neural networks-based mri image analysis for the alzheimer's disease prediction from mild cognitive impairment. *Frontiers in neuroscience*, 12:777, 2018.
- [23] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *International Conference on Learning Representations (ICLR)*, 2018.
- [24] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.
- [25] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.
- [26] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.
- [28] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.
- [29] J Zhang, Y Xie, Z Liao, G Pang, J Verjans, W Li, Z Sun, J He, Y Li, C Shen, et al. Viral pneumonia screening on chest x-ray images using confidence-aware anomaly detection. *arxiv 2020. IEEE Trans Med Imaging.*, 2021.
- [30] Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals. *arXiv preprint arXiv:2005.13249*, 2020.
- [31] Nicolau Goncalves, Janne Nikkilä, and Ricardo Vigario. Self-supervised mri tissue segmentation by discriminative clustering. *International journal of neural systems*, 24(01):1450004, 2014.
- [32] Wenjian Huang, Hao Li, Rui Wang, Xiaodong Zhang, Xiaoying Wang, and Jue Zhang. A self-supervised strategy for fully automatic segmentation of renal dynamic contrast-enhanced magnetic resonance images. *Medical physics*, 46(10):4417–4430, 2019.
- [33] Max Torop, Satya VVN Kothapalli, Yu Sun, Jiaming Liu, Sayan Kahali, Dmitriy A Yablonskiy, and Ulugbek S Kamilov. Deep learning using a biophysical model for robust and accelerated reconstruction of quantitative, artifact-free and denoised images. *Magnetic resonance in medicine*, 84(6):2932–2942, 2020.
- [34] Da-in Eun, Ryoungwoo Jang, Woo Seok Ha, Hyunna Lee, Seung Chai Jung, and Namkug Kim. Deep-learning-based image quality enhancement of compressed sensing magnetic resonance imaging of vessel wall: Comparison of self-supervised and unsupervised approaches. *Scientific Reports*, 10(1):1–17, 2020.
- [35] Elisabeth JM Zirkzee, Gerda M Steup-Beekman, Eduard Lem Bollen, Esther Baptist, Thomas M Slee, Menno V Huisman, Huub Am Middelkoop, Jasper Luyendijk, Mark A van Buchem, Tom Wj Huizinga, et al. Prospective study of clinical phenotypes in neuropsychiatric systemic lupus erythematosus; multidisciplinary approach to diagnosis and therapy. *The Journal of Rheumatology*, 39(11):2118–2126, 2012.
- [36] Matthew H Liang, Michael Corzillius, Sang Cheol Bae, Robert A Lew, Paul R Fortin, Caroline Gordon, David Isenberg, Graciela S Alarcón, Karin V Straaton, Judah Denburg, et al. The american college of rheumatology nomenclature and case definitions for neuropsychiatric lupus syndromes. *Arthritis and rheumatism*, 42(4):599–608, 1999.
- [37] Dafna D Gladman, Dominique Ibanez, and Murray B Urowitz. Systemic lupus erythematosus disease activity index 2000. *The Journal of rheumatology*, 29(2):288–291, 2002.
- [38] Dafna D Gladman, Charles H Goldsmith, Murray B Urowitz, Paul Bacon, Paul Fortin, Ellen Ginzler, Caroline Gordon, John G Hanly, David A Isenberg, Michelle Petri, et al. The systemic lupus international collaborating clinics/american college of rheumatology (slicc/acr) damage index for systemic lupus erythematosus international comparison. *The Journal of Rheumatology*, 27(2):373–376, 2000.
- [39] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

MRI-based Classification of Neuropsychiatric Systemic Lupus Erythematosus Patients with Self-supervised Contrastive Learning

- [40] Mark W Woolrich, Saad Jbabdi, Brian Patenaude, Michael Chappell, Salima Makni, Timothy Behrens, Christian Beckmann, Mark Jenkinson, and Stephen M Smith. Bayesian analysis of neuroimaging data in fsl. *Neuroimage*, 45(1):S173–S186, 2009.
- [41] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.
- [42] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [43] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.
- [44] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017.
- [45] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
- [46] Ian Covert, Scott Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33, 2020.
- [47] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.
- [48] Sérgio Pereira, Raphael Meier, Victor Alves, Mauricio Reyes, and Carlos A Silva. Automatic brain tumor grading from mri data using convolutional neural networks and quality assessment. In *Understanding and interpreting machine learning in medical image computing applications*, pages 106–114. Springer, 2018.
- [49] Danail Stoyanov, Zeike Taylor, Seyed Mostafa Kia, Ipek Oguz, Mauricio Reyes, Anne Martel, Lena Maier-Hein, Andre F Marquand, Edouard Duchesnay, Tommy Löfstedt, et al. *Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF 2018, and IMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings*, volume 11038. Springer, 2018.
- [50] Eyal Muscal and Robin L Brey. Neurological manifestations of systemic lupus erythematosus in children and adults. *Neurologic clinics*, 28(1):61, 2010.
- [51] William M Brooks, Wilmer L Sibbitt Jr, Mario Kornfeld, Rex E Jung, Arthur D Bankhurst, and Carlos A Roldan. The histopathologic associates of neurometabolite abnormalities in fatal neuropsychiatric systemic lupus erythematosus. *Arthritis & Rheumatism*, 62(7):2055–2063, 2010.
- [52] E Hachulla, U Michon-Pasturel, D Leys, JP Pruvo, V Queyrel, E Masy, J Arvieux, C Caron, F Brevet-Coupeé, PY Hatron, et al. Cerebral magnetic resonance imaging in patients with or without antiphospholipid antibodies. *Lupus*, 7(2):124–131, 1998.
- [53] Ludovica Griffanti, Mark Jenkinson, Sana Suri, Enikő Zsoldos, Abda Mahmood, Nicola Filippini, Claire E Sexton, Anya Topiwala, Charlotte Allan, Mika Kivimäki, et al. Classification and characterization of periventricular and deep white matter hyperintensities on mri: a study in older adults. *Neuroimage*, 170:174–181, 2018.
- [54] Kenji TSUSHIMA and Keishi KUBO. Tuberculous meningitis developed during treatment for systemic lupus erythematosus (sle). *Kekkaku (Tuberculosis)*, 74(1):27–32, 1999.
- [55] T Kawamata, M Kagawa, O Kubo, M Takeshita, H Ujiie, K Sato, and M Izawa. Clinicopathological studies of three cases of cerebral aneurysms associated with systemic lupus erythematosus. *No shinkei geka. Neurological surgery*, 19(7):633–639, 1991.
- [56] Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E Petersen, Yike Guo, Paul M Matthews, and Daniel Rueckert. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 541–549. Springer, 2019.
- [57] Ammara Masood, Adel Al-Jumaily, and Khairul Anam. Self-supervised learning model for skin cancer diagnosis. In *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 1012–1015. IEEE, 2015.