Chained Regularization for Identifying Brain Patterns Specific to HIV Infection

Ehsan Adeli*, Dongjin Kwon*, Qingyu Zhao*, Adolf Pfefferbaum*, Natalie M. Zahr*, Edith V. Sullivan*, Kilian M. Pohl*

*a Department of Psychiatry & Behavioral Sciences, Stanford University, Stanford, CA 94305, USA
b Center for Health Sciences, SRI International, Menlo Park, CA 94025, USA

Abstract

Human Immunodeficiency Virus (HIV) infection continues to have major adverse public health and clinical consequences despite the effectiveness of combination Antiretroviral Therapy (cART) in reducing HIV viral load and improving immune function. As successfully treated individuals with HIV infection age, their cognition declines faster than reported for normal aging. This phenomenon underlines the importance of improving long-term care, which requires better understanding of the impact of HIV on the brain. In this paper, automated identification of patients and brain regions affected by HIV infection are modeled as a classification problem, whose solution is determined in two steps within our proposed Chained-Regularization framework. The first step focuses on selecting the HIV pattern (i.e., the most informative constellation of brain region measurements for distinguishing HIV infected subjects from healthy controls) by constraining the search for the optimal parameter setting of the classifier via group sparsity ($\ell_{2,1}$-norm). The second step improves classification accuracy by constraining the parameterization with respect to the selected measurements and the Euclidean regularization ($\ell_2$-norm). When applied to the cortical and subcortical structural Magnetic Resonance Images (MRI) measurements of 65 controls and 65 HIV infected individuals, this approach is more accurate in distinguishing the two cohorts than more common models. Finally, the brain regions of the identified HIV pattern concur with the HIV literature that uses traditional group analysis models.

*Corresponding author: kilian.pohl@sri.com

Preprint submitted to NeuroImage August 13, 2018
1. Introduction

Despite the success of highly active antiretroviral therapy (HAART) and combination antiretroviral therapy (cART) in extending longevity of individuals infected with the Human Immunodeficiency Virus (HIV), neurocognitive impairments still commonly occur [1, 2, 3]. Structural Magnetic Resonance Imaging (MRI) has often been used to determine the neural correlates of cognitive and motor deficits in HIV infection, indicating, for example, specific relationships between regional brain volume deficits [4, 5], memory compromise [6], and accelerated brain aging in HIV infected adults [7]. Neurocognitive and motor impairments in HIV infection, however, are similar to those reported in other age-related diagnoses [8]. To improve diagnostic specificity of MRI in HIV, this manuscript proposes a novel machine learning method and applies it to the morphometric measurements extracted from structural MRI scans collected from HIV infected and healthy control (CTRL) participants.

Conventional HIV MRI studies typically test for group differences (with respect to the CTRL cohort) by separately analyzing each image measurement for the impact of HIV [4, 9, 6, 10, 2, 1]. Separate analysis of measurements may lead to contradicting or inconclusive findings [11]. By contrast, our proposed analysis is a type of machine learning framework that considers all image measurements together to identify the subset of measurements (called patterns) specific to HIV and then relates the significance of the pattern to its accuracy in distinguishing individuals with HIV from CTRLs. A popular approach for identifying patterns uses sparse classifiers [12, 13, 14, 15, 16, 8], which assume that only a few measurements are informative for distinguishing cohorts. After identifying a pattern, the corresponding measures are often applied to a second (non-sparse) approach, which focuses only on improving classification accuracy [17, 18, 19, 20, 21, 22]. This two-step regularization procedure assumes that measurements selected by the sparse classifier define the unique, optimal pattern for distinguishing the two cohorts [23, 24, 17, 25]. This assumption, however, is gener-
ally not true because the redundancy in information across image measurements allows
for multiple solutions [19]. As the two steps are based on different classification ap-
30 proaches, the pattern identified by the sparse classifier of the first step are generally not
optimal for the non-sparse approach of the second step.

Herein, we propose an approach (denoted as Chained-Regularization) that uses
the same classifier first to identify a pattern and then, using the pattern, to distinguish in-
35 dividuals; however, different constraints guide the parameterization of the classifier in
each step. Our proposed algorithm models the selection of the most informative image
measurements in the first step by confining parameterization of the classifier through
group sparsity ($\ell_{2,1}$-norm) regularization [26, 8]. Group sparsity extends the concept
of the $\ell_1$-norm [27, 28, 16] of identifying a few informative measurements for comb-
ining measurements into groups and then identifying a small number of groups [27].

The grouping can be used for explicit modeling of relationships between measurements
[29]. In this work, each measurement from the regions of interest (ROIs) is grouped
with its counterpart in the other brain hemisphere given our assumption that HIV in-
fection affects the brain bilaterally. In the second step of Chained-Regularization, the
classifier is trained on just the selected individual measurements with the search for
the optimal parameter setting being constrained via Euclidean ($\ell_2$-norm) regulariza-
45 tion. The logic of this approach is that the $\ell_{2,1}$ regularization generally improves the
accuracy of classifiers in the presence of a large number of uninformative or redund-
ant image measurements (as it is often the case of neuroimaging studies), while the
$\ell_2$ regularization improves the accuracy of classifiers in the event that all provided im-
40 age measurements are informative [17, 18]. Our chained-regularization scheme, which
uses a sequential dependency approach to identify a pattern to be applied for deter-
mining group membership of individuals, is different from chain-regularization [30],
a concept used in physics to describe group of objects interacting with each other in a
chain.

We implement Chained-Regularization within a multiple kernel learning (MKL)
framework [31, 18]. MKL is based on the assumption that samples (i.e., individual par-
50 ticipants) that are similar to each other should be assigned to the same cohort (e.g.,
HIV). Similarity between two samples is measured through a pairwise comparison of
Figure 1: Training of the Chained-Regularization approach: The first step (top, denoted as Selection Step) selects the image measurements informative for distinguishing HIV from controls, while the second step (bottom, denoted as Reweighing Step) focuses on improving the accuracy by reweighing the selected measures for classifying the samples. Note, both steps are based on the same classifier but differ in regularizing (or constraining) its parameterization.

the corresponding image measurements. This comparison is defined by a set of metrics \( i.e., \) linear and nonlinear kernel functions, each capturing a unique characteristic across image measurements. The MKL algorithm now determines the combination of metrics and image measurements \([18]\) that lead to the highest classification accuracy (see Figure 1). It thus omits the simplifying assumption of most other classifiers that the discriminating characteristics of all image measurements are best captured by a single metric (as in \([18, 32, 33, 31, 34]\)).

In summary, our analysis makes two novel contributions: (1) We propose Chained-Regularization within the MKL framework, which, in our experiments, is significantly more accurate than single-step and other two-step approaches. (2) To the best of our knowledge, this is the first study to examine both linear and non-linear supervised learning approaches to identify patterns that discriminate HIV infected from healthy control brains.

The rest of the paper is organized as follows: Section 2 introduces the materials (the data set), preprocessing, the proposed chained regularization and the experimental setup. Appendix A provides additional technical details of the proposed method.
Table 1: Demographic information (‘svol’ = supratentorial volume).

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>sex</th>
<th>Age (years)</th>
<th>svol ($\times 10^6$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIV</td>
<td>65</td>
<td>20</td>
<td>45</td>
<td>51.81 ± 8.44</td>
</tr>
<tr>
<td>Matched CTRL</td>
<td>65</td>
<td>20</td>
<td>45</td>
<td>51.76 ± 8.44</td>
</tr>
<tr>
<td>CF CTRL</td>
<td>180</td>
<td>102</td>
<td>78</td>
<td>43.36 ± 18.92</td>
</tr>
</tbody>
</table>

Section 3 compares our approach to other implementations on the HIV data set and reports on its identified pattern specific to HIV. Section 4 provides an in depth discussion about the findings of the previous section and their relevance with respect to the HIV literature. The paper concludes with Section 5.

2. Materials and Methods

2.1. Participant Information

Data used in this study are from 65 HIV infected individuals and 245 CTRL subjects. For classification, we match 65 CTRLs to the 65 HIV cohort. Specifically, for each HIV subject, one subject is selected from the CTRL cohort, such that they have the same sex and a minimal difference in their ages. We refer to the matched samples as ‘matched CTRL group’. The remaining 180 CTRL subjects, referred to as Confounding Factors CTRL group (CF CTRL group), are used for analysis of the confounding factors and minimizing their effects. Table 1 shows the demographic information of participants in all groups, and Figure 2 plots their age distributions. All 310 participants are tested for HIV, viral load, and CD4 T-cell count. HIV infected individuals had a CD4 count $> 100$ cells/µL and a Karnofsky score $> 70$ [35]. Data from these subjects were used in previous studies [10, 4, 9].

2.2. Structural MRI Data Acquisition

Imaging data are acquired from each participant on a 3T General Electric (GE) SIGNA HDx system using an 8-channel Array Spatial Sensitivity Encoding Technique.
Figure 2: Age distribution of the participants: HIV (left), Matched CTRL (middle), and CF CTRL (right).

(ASSET) coil for parallel and accelerated imaging. Furthermore, Inversion Recovery-Spoiled Gradient Recalled (IR-SPGR) echo sequence (TR=7.068ms, TI=300ms, TE = 2.208ms, flip angle=15°, matrix=256 × 256, slice dimensions=1.2 × 0.9375 × 0.9375mm, 124 slices) are collected in the sagittal plane.

2.3. MRI Data Preprocessing and Feature Extraction

Preprocessing of the T1-weighted (T1w) MR images involves noise removal [36], computing signal-to-noise ratio (SNR) [37] and correcting field inhomogeneity via N4ITK (Version 2.1.0) [38]. Next, the brain mask is segmented by majority voting [39] across maps extracted by FSL BET (Version 5.0.6) [40], AFNI 3dSkullStrip (Version AFNI_2011_12_21_1014) [41], FreeSurfer mri-gcut (Version 5.3.0) [42], and the Robust Brain Extraction (ROBEX) method (Version 1.2) [43], applied to bias and non-bias corrected T1w images. The refined mask is then used to repeat image inhomogeneity correction.

We further apply the cross-sectional approach of FreeSurfer (Version 5.3.0) [44, 45] to the skull-stripped T1w MRI of each subject in order to measure the mean curvature (MeanCurv), surface area (SurfArea), gray matter volume (GrayVol), and average thickness (ThickAvg) of 34 bilateral cortical Regions Of Interest (ROIs) [2 hemispheres × 4 measurement types × 34 ROIs = 272], the volumes of 8 bilateral sub-cortical ROIs (i.e., thalamus, caudate, putamen, pallidum, hippocampus, amygdala, accum-bens, cerebellar cortex) [2 × 8 = 16], the volumes of 5 subregions of the corpus cal-losum (posterior, mid-posterior, central, mid-central and anterior), and the combined volume of all white matter hypointensities [5 + 1 = 6]. White matter hypointensities
are defined according to Freesurfer as voxels inside the white matter with signal intensities lower than a threshold level [46]. Finally, volumes of the left and right lateral and third ventricles \(2 \times 2 = 4\) are measured by non-rigidly aligning the SRI24 atlas [47] to the T1w MRI of the subject via ANTS (Version: 2.1.0) [48]. This procedure thus extracts 298 measures from each brain MRI.

For the entire matched data set, each of these 298 brain measures are normalized using their z-scores [49]. To avoid using any data for testing the model, the z-scores are parameterized by computing the mean and standard deviations of measurements across the CF CTRL cohort. Based on this distribution, the z-scores are then computed for each subject of the matched CTRL and HIV groups. Furthermore, the segmentations are used to compute the supratentorial volume (svol) for each subject. As in [50], svol is used to approximate brain size.

2.4. Confounding Factors

For each of the 298 measures, we compute the Pearson correlation between the corresponding z-scores of the 180 subjects of the CF CTRL group and the factors, i.e., age, sex, svol, race, and SNR. Some of the measures are significantly correlated with age, sex, and svol (p-value < 0.05). For each measurement, a general linear model (GLM) [51] is parameterized with respect to corresponding z-scores to omit the effect of the confounding factors. Specifically, for each image measure \(m \in \{1, \ldots, 298\}\), the following GLM is fit across the subjects \(i \in \{1, \ldots, 180\}\) of the CF CTRL group with the corresponding z-score \(v_i^m\) as the observation and age \((f_i^{age})\), sex \((f_i^{sex})\), and svol \((f_i^{svol})\) as the confounding factors:

\[
v_i^m \sim \beta_{m,0} + \beta_{m,1} f_i^{age} + \beta_{m,2} f_i^{sex} + \beta_{m,3} f_i^{svol}.
\]  (1)

After obtaining the optimal regression coefficients \((\hat{\beta}_{m,0}, \hat{\beta}_{m,1}, \hat{\beta}_{m,2}, \hat{\beta}_{m,3})\) across all subjects, the model is applied to the HIV and matched CTRL dataset. Specifically, the residual explained by each subject’s individual confounding factors multiplied by the regression coefficients is removed from the initial observation, i.e., the residual
scores $x_i^m$ is defined as
\[ x_i^m := v_i^m - (\hat{\beta}_{m,0} + \hat{\beta}_{m,1} f_{i}^{\text{age}} + \hat{\beta}_{m,2} f_{i}^{\text{sex}} + \hat{\beta}_{m,3} f_{i}^{\text{svol}}). \] (2)

2.5. Pattern Extraction and Classification

In this section, the proposed Chained-Regularization technique is outlined. For the interested reader, Appendix A derives the Chained-Regularization approach in detail. Based on the residual scores of the matched data set, the accuracy of the proposed Chained-Regularization framework (denoted as $\ell_{2,1}\ell_{2}$-reg; see also Figure 1) in correctly labeling HIV infected and health control subjects is measured via 10-fold (nested) cross-validation (see Figure 4). With respect to each (testing) fold, the training of $\ell_{2,1}\ell_{2}$-reg on the remaining data starts with the Selection Step, i.e., extract the informative pattern for classifying samples. The training then proceeds with the Reweighing Step, i.e., finding the optimal parameterization of the classifier based on that pattern. On the testing fold, we record the labeling of subjects according to the trained $\ell_{2,1}\ell_{2}$-reg. This procedure between training and testing is repeated until the labeling across all 10 testing folds are generated. Based on those labelings, we compute the Accuracy of prediction (i.e., the percentage of the testing subjects that are classified correctly into their respective classes), specificity (SPE), sensitivity (SEN) and area under the receiver operating characteristic (ROC) curve (AUC). Note, our MKL-based mapping function outputs a continuous value (more details in Appendix A) from which a binary class label is derived via thresholding. By changing the threshold, we can create the ROC curve and hence calculate the AUCs. In addition, we apply the Fisher’s exact test [52] to ensure that implementation is significantly better than chance ($p$-value < 0.01). The remainder of this section provides further details about training of $\ell_{2,1}\ell_{2}$-reg.

Inspired by [18, 13], the HIV specific pattern, identified in the Selection Step during training, is defined by the optimal ‘weight’ vector specifying a linear multivariate model defined by image measurements that correctly label subjects according to the MKL model. MKL classifies samples by learning the optimal pairings between kernels and image measurements. Finding the optimal pairing is described as a minimization
problem with respect to a weight vector, sparsity of which specifies the importance of pairings for class separation. We use 7 different kernels to build our multiple kernel learning model, including 3 kernel types [linear, histogram intersection kernel (HIK), and radial basis function kernel (RBF)] with different settings of their hyperparameters. These 7 kernels are defined in detail in Appendix A. Specific to our implementation, the optimal weight vector minimizes a cost function measuring classification accuracy and ‘group-sparsity’ associated with those weights. As also shown in Figure 3, group-sparsity is measured by first transforming the weight vector into a matrix so that each column represents a group and each group combines the weights associated with measurements from the same type and region (regardless of hemisphere). $\ell_2, 1$-norm is then applied to the matrix, i.e., the $\ell_2$-norm is applied to each column resulting in the column being reduced to a scalar value and then the $\ell_1$-norm is applied to the vector of those scalar values resulting in the entire matrix being reduced to a scalar value. Note, this computation generally penalizes weight vectors that select a larger number of groups, i.e., are not sparse on a group level.

The optimal ‘weight’ vector now depends on the weight $C$ of the term measuring classification accuracy and the weight $\lambda$ of the term measuring group sparsity within the MKL cost function (refer to Appendix A for more details). As in [13, 53], the search space for those two hyperparameters is $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}\}$. To identify the best hyperparameter setting, we perform 5-fold inner cross-validation 10 times. Each time, we randomly divide the training data into 5 validation folds. For each validation fold, we first train our implementation of MKL with respect to a specific hyperparameter setting on the remaining training data. For that setting, we then record the accuracy of the implementation on the validation fold and the identified pattern, i.e., regional scores associated with non-zero weights. We repeat this process for each hyperparameter setting and then only keep the pattern that is associated with the highest validation accuracy across all parameter settings. Repeating this process for the remaining 4 validation folds and 9 more inner-cross validations then results in a total of 50 ‘trials’. The Selection Step then defines the HIV specific pattern as the set of residual scores that were part of all 50 trial patterns. This multi-trial selection process is considered more robust than only relying on single run of a sparse classifier.
Measurements from the left hemisphere
Measurements from the right hemisphere

Figure 3: Illustration of feature grouping for group sparsity. (a) Regular sparsity ($\ell_1$-norm) operates on a vector that concatenates the measurements from the left and right hemispheres. (b) Group sparsity operates on the matrix formed by putting the features from the same ROIs of the left and right hemispheres in its columns.

[18, 23].

The **Reweighing Step** focuses on improving MKL’s classification accuracy when only relying on the residual scores of the HIV specific pattern. As training of the classifier is now confined to only informative image measurements, classification accuracy is generally improved by replacing the $\ell_{2,1}$-norm with the $\ell_2$-norm in the cost function of the MKL implementation. The **Reweighing Step** then performs parameter exploration of this MKL implementation via 5 fold inner cross-validation, *i.e.*, it records the hyperparameter setting that leads to the highest average validation accuracy across the 5 inner folds. The training of $\ell_{2,1}$-$\ell_2$-reg is completed by training MKL with the selected hyperparameter setting on the complete training data. Note, choosing the optimal hyperparameters without including any data from the testing fold yields more reliable and reproducible results [54] than tuning the hyperparameters without any inner validation folds.

The group sparsity (in the Selection Step) guarantees ‘bilateral selection’ of each type of ROI-specific measurement (*i.e.*, measurements on both left and right hemispheres are selected or neither one of them). The Reweighing Step then builds the final classifier relying on all selected individual measurements and the $\ell_2$-norm, which generates non-sparse classifiers that generalize well to unseen testing data [17, 55]. For the interested reader, Appendix A derives the Chained-Regularization approach in detail. Specifically, we first generalize the MKL approach of [13], which was specific to $\ell_1$-norm regularization, to regularizers that are convex and differentiable in $\mathbb{R}_{\geq 0}$. We then embed that approach into the proposed Chained-Regularization framework.
On the \( i \)th training iteration, the Selection Step selects the most informative measurements (i.e., the pattern) using \( \ell_{2,1} \)-regularization, and then the Reweighing Step uses that pattern to build the classifier with \( \ell_{2} \)-regularization. In the second step, inner cross-validation is used to choose the model hyperparameters. Next, the built classifier is used to calculate the accuracy scores on the corresponding testing fold (say Acc\(_i\)). The average accuracy for all folds is then reported (i.e., Acc = \( \frac{1}{10} \sum_{i=1}^{10} \text{Acc}_i \)).
2.6. Alternative Implementations

To motivate the specific implementation of the Chained-Regularization approach, the nested cross-validation of Chained-Regularization is repeated with different combinations of regularizers, i.e., using $\ell_1$-norm in the Selection Step and $\ell_2$-norm in the Reweighing Step (denoted by $\ell_1$-$\ell_2$-reg), using $\ell_{2,1}$-norm in the Selection Step and $\ell_1$-norm in the Reweighing Step (denoted by $\ell_{2,1}$-$\ell_1$-reg), and using $\ell_2$-norm in both steps (denoted by $\ell_2$-$\ell_2$-reg). In addition, the comparison includes an implicit model for the grouping of the ROI measurements by computing the average value of each group and then using the $\ell_1$-norm in the Selection Step and $\ell_2$-norm in the Reweighing Step (denoted as Avg $\ell_1$-$\ell_2$-reg). To demonstrate the advantages of Chained-Regularization, only the MKL approach is cross-validated, i.e., omitting the Reweighing Step as well as the repeated selection procedure. The corresponding Single-Step Regularization approaches are denoted as $\ell_1$-reg, $\ell_2$-reg and $\ell_{2,1}$-reg. Note, we omitted certain alternative implementations from the experimental setup to keep the comparison concise and informative. For example, one could implement Chained-Regularization using the $\ell_1$-norm in both steps. While this implementation produces similar accuracy score as $\ell_{2,1}$-$\ell_1$-reg, the approach most likely underestimates the impact of the disease on a small number of brain regions; a risk generally associated with sparse classifiers based on the $\ell_1$-norm [56]. Furthermore, note that training a MKL without regularization, constraint or a penalty term (in the reweighing step) is not feasible as the underlying minimization problem is then underdetermined [18], i.e., results in an unstable classifier.

In addition to variations of Chained-Regularization, the comparison includes conventional support vector machine (SVM) classifiers widely used in neuroimaging applications to highlight the benefits of Chained-Regularization in the context of MKL. The class of alternative SVM classifiers include linear SVM, SparseSVM [57], and sparse feature selection [20] followed by a linear SVM (SFS+SVM). In addition, t-test [20], elastic-net [24], and the mutual information based feature selector minimum-redundancy maximum-relevancy (mRMR) [58] are coupled with a linear SVM classifier to further evaluate the performance of the proposed feature selection technique.

For each implementation, the accuracy scores of the previous section are computed.
We also apply the DeLong test [59] to mark implementations that are significantly worse ($p$-value $< 0.01$) than the proposed $\ell_{2,1}\ell_{2}$-reg.

3. Results

3.1. Comparison

Classification results of the proposed and alternative methods are summarized in Table 2. The proposed Chained-Regularization technique ($\ell_{2,1}\ell_{2}$-reg) achieves the highest Accuracy (82.3%), SEN (0.84), and AUC (0.87). The SPE (0.82) is equivalent to ‘$\ell_1\ell_2$-reg’ and ‘Avg $\ell_1\ell_2$-reg’. All other implementations of the comparison (including $\ell_2\ell_2$-reg and $\ell_{2,1}\ell_1$-reg) not only received lower scores, but were also significantly worse than the proposed chained $\ell_{2,1}\ell_2$ regularization. The single step regularizers received higher scores in all four performance measures than the conventional approaches with the exception of SFS+SVM. The performance scores of SFS+SVM (Accuracy: 0.69%, SPE: 0.69, SEN: 0.70 and AUC: 0.73) were higher than those of $\ell_2$-reg and $\ell_{2,1}$-reg but lower than $\ell_1$-reg (Accuracy: 70.3%, SPE: 0.70, SEN: 0.70, AUC: 0.73), the single step regularization with the highest Accuracy and AUC. Finally, only conventional methods (i.e., t-test+SVM, mRMR+SVM, SparseSVM and SVM) produced classification results that were not significantly better than chance.

3.2. The HIV Pattern

For $\ell_{2,1}\ell_2$-reg (the most accurate approach in the comparison), Figure 5 shows the frequencies (normalized in the range $[0, 1]$) of the 298 image measurements selected by the Selection Step across the 10 runs of cross-validation on the whole matched data set considered for identifying the pattern. This figure shows the measurements with a selection frequency of 1 (selected all times), i.e., those that are actually used in the Reweighing Step of our method, with colors based on their measurement types. Note, the ordering of measurement types is arbitrary. We refer to this set of measurements as the HIV pattern. The remaining measurements are displayed in gray regardless of the type of measurement. Approximately 39% of all image measurements are selected in all runs. These measures define the HIV pattern.
Table 2: Classification results of different approaches summarized by Accuracy, specificity (SPE), sensitivity (SEN) and area under the ROC curve (AUC). The best score in each category is in bold. Methods are marked with †, if they were significantly worse than the proposed approach ($p < 0.01$ according to Delong’s Test [59]). Methods marked with ‡ are significantly better than chance ($p < 0.01$ according to the Fisher exact test [52]).

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
<th>SPE</th>
<th>SEN</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed $\ell_2,1$-$\ell_2$-reg†</td>
<td>82.3</td>
<td>0.82</td>
<td>0.84</td>
<td>0.87</td>
</tr>
<tr>
<td>Chained (Baseline)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\ell_1$-$\ell_2$-reg‡</td>
<td>81.9</td>
<td>0.82</td>
<td>0.79</td>
<td>0.86</td>
</tr>
<tr>
<td>Avg $\ell_1$-$\ell_2$-reg‡</td>
<td>79.7</td>
<td>0.82</td>
<td>0.77</td>
<td>0.85</td>
</tr>
<tr>
<td>$\ell_2$-$\ell_2$-reg†‡</td>
<td>73.1</td>
<td>0.74</td>
<td>0.73</td>
<td>0.76</td>
</tr>
<tr>
<td>$\ell_2,1$-$\ell_1$-reg†‡</td>
<td>72.5</td>
<td>0.72</td>
<td>0.73</td>
<td>0.76</td>
</tr>
<tr>
<td>Single Step Regularization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\ell_1$–reg†</td>
<td>70.3</td>
<td>0.70</td>
<td>0.70</td>
<td>0.75</td>
</tr>
<tr>
<td>$\ell_2,1$–reg†‡</td>
<td>69.7</td>
<td>0.70</td>
<td>0.68</td>
<td>0.73</td>
</tr>
<tr>
<td>$\ell_2$–reg†‡</td>
<td>68.7</td>
<td>0.64</td>
<td>0.70</td>
<td>0.71</td>
</tr>
<tr>
<td>Conventional Methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SFS [20]+SVM†‡</td>
<td>69.9</td>
<td>0.69</td>
<td>0.70</td>
<td>0.73</td>
</tr>
<tr>
<td>elastic-net [24]+SVM†‡</td>
<td>65.1</td>
<td>0.64</td>
<td>0.64</td>
<td>0.69</td>
</tr>
<tr>
<td>t-test [20]+SVM†</td>
<td>59.1</td>
<td>0.61</td>
<td>0.56</td>
<td>0.65</td>
</tr>
<tr>
<td>mRMR [58]+SVM†</td>
<td>59.6</td>
<td>0.56</td>
<td>0.61</td>
<td>0.64</td>
</tr>
<tr>
<td>SparseSVM [57]†</td>
<td>57.9</td>
<td>0.55</td>
<td>0.60</td>
<td>0.64</td>
</tr>
<tr>
<td>SVM†</td>
<td>56.7</td>
<td>0.57</td>
<td>0.56</td>
<td>0.60</td>
</tr>
</tbody>
</table>

To analyze the significance of each type of measurement (Mean Curvature, Surface Area, Gray Matter Volume, Average Thickness, and Subcortical ROI Volumes), we first create a baseline for comparison by performing 10-fold cross validation just on the Reweighing Step with the scores being confined to the HIV pattern, recording the testing accuracy for each fold, and then computing the mean and standard deviation in the accuracy score across all 10 folds. The results in an accuracy of 87.69% ± 1.69 (mean ± standard deviation), which we refer to as ‘All Measurements’ in Table 3. For each measurement type, we then omit the corresponding measures from the data,
perform 10-fold cross-validation of the Selection Step on this subset of data, record the pattern, and repeat the previous cross-validation of the Reweighing Step with respect to that pattern.

With respect to using subsets of the measurements, omitting Average Thickness from the data resulted in the pattern with the highest mean accuracy score (79.6% ± 1.96). Omitting Mean Curvature, Surface Area, or Volume from the HIV pattern resulted in accuracy scores that were significantly lower than those produced by All Measurements (or the HIV pattern). The same was true when confining classification to cortical gray matter volumes.

Beyond the type of measurements, Table 4 lists and Figure 6 visualizes the selected cortical regions. 35% of all cortical measurements are selected by our method. Furthermore, a total of 52% of the subcortical measurements are selected. Figure 7 shows the selected subcortical regions (i.e., hippocampus, amygdala, accumbens and cerebellar cortex) along with the white matter structures (i.e., corpus callosum posterior and mid-posterior) selected by our approach. In addition to these ROIs, hypointensity lesion volumes are also selected. Note, as also argued in [8], the coefficients computed by sparse classifiers simply parameterize a linear multivariate model (explained
Table 3: The mean ± standard of the classification Accuracy and area under the ROC curve (AUC) of the proposed method with different subsets of the 298 measurements. Entries marked with † are significantly worse (p < 0.01; Delong’s Test) compared to ‘All Measurements’.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Measurements</td>
<td>87.7 ± 1.69</td>
<td>0.87 ± 0.03</td>
</tr>
<tr>
<td>No Average Thickness</td>
<td>79.6 ± 1.96</td>
<td>0.78 ± 0.07</td>
</tr>
<tr>
<td>No Mean Curvature†</td>
<td>77.2 ± 1.92</td>
<td>0.84 ± 0.05</td>
</tr>
<tr>
<td>No Surface Area†</td>
<td>73.9 ± 1.77</td>
<td>0.79 ± 0.03</td>
</tr>
<tr>
<td>No Gray Matter Volume†</td>
<td>66.8 ± 1.49</td>
<td>0.74 ± 0.06</td>
</tr>
<tr>
<td>Only Cortical Measurements†</td>
<td>69.2 ± 2.10</td>
<td>0.75 ± 0.02</td>
</tr>
</tbody>
</table>

Figure 6: cortical ROIs selected by our proposed approach. (b-e) show the selected ROIs for each measurement type separately, while (a) visualizes the union of the four types.

in detail in Appendix A), which predicts the class labels. Thus, coefficients are informative with respect to feature selection but are not good indicators for differentiating the importance among the selected features with respect to identifying cohorts.

The comparison of different approaches (see Table 2) revealed that our Chained-Regularization approach was significantly better than confining analysis to any one single step (i.e., \( \ell_{2,1}\)-reg or \( \ell_2\)-reg). Our approach was also significantly better than alternative implementations of the Chained-Regularization that used the same type of regularizer for both steps (i.e., sparse regularizer (\( \ell_{2,1}\)-\( \ell_1\)-reg) or Euclidean regularizer (\( \ell_2\)-\( \ell_2\)-reg)). This finding underlines the importance of selecting two regularizers that complement each other for our approach, i.e., the first regularizer models the selec-
Table 4: Cortical surface ROIs and their measurement types selected by our method.

<table>
<thead>
<tr>
<th>ROI</th>
<th>GrayVol</th>
<th>MeanCurv</th>
<th>ThickAvg</th>
<th>SurfArea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bankssts</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Caudalanteriorcingulate</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Caudalmiddlefrontal</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Cuneus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entorhinal</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fusiform</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inferiorparietal</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inferiortemporal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isthmuscingulate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateraloccipital</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateralorbitofrontal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lingual</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medialorbitofrontal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middletemporal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parahippocampal</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paracentral</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parsopercularis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parsorbitalis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parstriangularis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pericalcarine</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Postcentral</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posteriorcingulate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precentral</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precuneus</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rostralanteriorcingulate</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rostralmiddlefrontal</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Superiorfrontal</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Superiorparietal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Superiortemporal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supramarginal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frontalpole</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temporalpole</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transversetemporal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insula</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Choosing alternative complementary regularizers by replacing the $\ell_{2,1}$-norm in the Selection Step with other sparse regularizers (while leaving the Reweighing Step un-
changed) results in non-significantly lower accuracy scores compared with the proposed approach. Unlike other implementations of the Chained-Regularization, our proposed $\ell_2,1$ regularizer was the only one that explicitly modeled the bilateral effect of HIV on the brain by grouping measurements across hemispheres. This additional modelling constraint simplifies the classification task and results in higher performance scores.

With the exception of SFS+SVM (which is also significantly less accurate than our proposed method), the worst performing methods are common (two-step) approaches that used different methods for feature selection and classification. Such approaches view pattern identification and classification as two disconnected machine learning tasks [20, 60, 61]. Thus, the optimal pattern identified in the first step is generally not optimal for the classifier in the second step, which would explain the low accuracy scores.

All implementations led to SPE, SEN, and AUC values similar to their Accuracy score (which is being maximized). This concurrence emerged because our data set is well balanced between the two cohorts. The residual scores of the imaging measures further minimizes the risk of biasing analysis towards one cohort.

Note that $\ell_{2,1}+\ell_2$-reg iteratively runs several trials of nested cross-validation in the Selection Step for reliable selection of the relevant features (i.e., the HIV pattern). However, training based on this procedure is computational expensive, the training
time is insignificant in comparison with the years it took to acquire the data. With
the implementations done only on a single computing core of a machine with an
Intel® Core™ i7-4712HQ CPU @ 2.30 GHz with 16 Gigabytes of memory, using
Matlab R2017a, it took approximately 6 hours to train the model and search all possible
settings for the hyperparameters and tune them. Note, our implementation was
not optimized and therefore computation times may be improved. Furthermore, the
training is done only once, after which the model parameters are saved and run on
test data. The testing time of \(\ell_2,1-\ell_2\)-reg is less than 0.01 second, which is similar to
the other implementations of this comparison. Note that the increase in the running
time of our method (compared to single-step methods) is mainly due to the constant
number of trials that we repeat the method to get a more robust pattern selected, \(i.e.,\)
the increase in the running time is not exponential to the number of subjects or mea-
surements. Therefore, the method is scalable for larger number of inputs. However, if
the number of measurements dramatically increases, the approach faces the so-called
‘Small-Sample-Size’ problem, a common issue in machine learning [62]. This problem
arises when the number available samples (\(N\)) is far fewer than the number of features
\((d)\) extracted from them \(i.e.,\ N << d\). Under these settings, all machine learning and
pattern recognition methods fail to identify the intrinsic space of the samples.

The HIV pattern identified by the proposed Chained-Regularization technique com-
poses of approximately 39% of the 298 measurements that were selected in all 50 train-
ing runs (see Figure 5). This frequent selection of such a large number of measurements
does not contradict the sparsity constraint of MKL but rather is due to grouping of mea-
asurement and the accounting done by the Chained-Regularization approach. Classifiers
relying on group sparsity \((\ell_{2,1})\) tend to select more measurements than those relying
solely on sparsity \((\ell_1)\). Furthermore, our method marks a measurement as infor-
mative if it is selected by MKL at least once in connection with one of the 7 kernels
\(i.e.,\) Linear, HIK, and RBF with 5 instances of its hyperparamter setting). That those
kernel-measurement pairs are actually sparsely selected by MKL is shown in Figure 8,
which lists each pair separately. By doing so, our Chained-Regularization approach
avoids underestimating the impact of the disease to a small region of the brain as com-
monly done by sparse classifiers [56].

19
As indicated by Table 4, combining the four different types of measurements used in our analysis is essential for creating a highly accurate HIV pattern. Of all measurement types, the Mean Curvature is the most frequent measurement type present in the pattern. However, when performing classification without the Mean Curvature (see Table 3), the drop in accuracy is less than when omitting Surface Area and Gray Matter Volume scores. The opposite is true for the regional gray matter volumes, which are least often selected, but whose omission from the pattern results in the largest drop in accuracy. This observation acknowledges that the number of times any type of measurement is part of a selected pattern does not necessarily indicate how important that type is for characterizing the disease.

4. Discussion

The measurements composing the HIV pattern identified by the most accurate approach, the proposed Chained-Regularization, are in agreement with the literature, which suggests that HIV infection is associated with volume deficits in cortical, subcortical, and white matter regions [63, 64, 65, 66, 67, 68, 69, 70, 71, 72]. As identifi-
fied using our automated, machine learning method (Table 4), the literature indicates that cortical areas affected in HIV relative to healthy controls include frontal, cingulate, sensorimotor, and parietal regions [9, 73, 74, 75, 76, 77]. For other cortical regions identified herein, reports on the effects of HIV are relatively less common: [9] lists temporoparietal regions; [71, 78] include effects of HIV on thinning of the temporal cortices; [72] describes effects on insula; [79] lists parahippocampal cortex. That our methods identified regions not commonly reported in the HIV literature (e.g., Caudalanteriorcingulate, Isthmuscingulate, Lateralorbitalfrontal, Parsopercularis, and Frontalpole) may be due to our inclusion of cortical measures such as mean curvature, thickness, and surface area, which are not typical metrics used in the HIV literature. Instead, the imaging literature usually focuses on the effects of HIV on gray matter volume (see the following for exceptions [80, 81, 82]). Indeed, in studies that assess cortical thickness rather than cortical volume, HIV effects are evident in areas such as the insula, orbitofrontal, temporal, and cingulate cortices [78, 83], similar to the ones identified here.

As also confirmed by our results, white matter is notably affected by HIV infection. Damage to myelin sheathes may be reflected in lower than normal white matter volume and greater prevalence of white matter hyperintensities [84] (deemed “hypointensities” by FreeSurfer [44]). Indeed, examination of brain microstructural integrity using DTI has detected subtle HIV-related differences from controls (e.g., lower fractional anisotropy and higher mean diffusivity) in myelin and axonal integrity [85, 86, 87, 88, 89, 90], even in normal-appearing white matter [91, 92, 93].

Subcortical regions frequently reported in the literature to have significantly smaller volumes in HIV subjects relative to healthy controls include hippocampus and basal ganglia structures [64, 94, 95, 96, 4]. Regarding the basal ganglia and limbic structures, our results specifically identify the accumbens and amygdala, whereas the literature more frequently cites the caudate, putamen, and pallidum (e.g., [65, 94, 95, 96]).

Although our approach does not identify the thalamus, a structure as particularly susceptible to HIV despite other reports (e.g., [9, 83, 4, 6]), our scheme does note cerebellum as a significant contributor to diagnosis differences. This inclusion is consistent with several other studies report HIV-related gray matter volume deficits in the cerebel-
lum [68, 69, 97]. The functional consequences of HIV effects on the cerebellum have been reported [98]; yet the cerebellum is generally underappreciated in the imaging literature as common analysis methods are designed with the neocortex in mind and may be suboptimal for the analysis of the cerebellum.

One of the main limitations of the proposed study is the large imbalance between the number of HIV patients and CTRL subjects. We addressed this issue by matching (and hence balancing) a subset of the CTRLs to the HIV group. However, this greatly reduced the number of samples used for extracting the pattern and testing, and thus the power of the analysis. To preserve the power of the provided data, expanding Chained-Regularization for explicit modeling of the imbalance between the cohorts can be a direction for future work.

5. Conclusion

We presented Chained-Regularization, a two-step-approach to identifying disease-specific patterns and performing pattern-based classification that, unlike the state-of-the-art, uses the same classification model for first identifying informative measures and then improving the accuracy of the classification based on the selected measures. Our choice of classification approach was a generalized version of the MKL method proposed by [13]. In the Selection Step, parameterization of MKL was confined by groups sparsity ($\ell_{2,1}$-norm) and in the Reweighing Step, the parameterization was penalized by the Euclidean ($\ell_2$-norm) regularization. This implementation was more accurate than alternative implementations and significantly better than common (two-step) approaches using different methods for feature selection and classification.

The Chained-Regularization approach identified a number of brain regions comporting with the literature and designated a few novel regions that (to our knowledge) have not been previously described in the HIV literature. These regions would benefit from further investigation as an improved understanding of the diseases remains critical for advancing the long-term care for the large number of HIV infected patients, who (even with suppressed viral loads) can suffer from cognitive disorders associated with HIV. Our current contribution in improving this understanding is in providing
Table A.5: Notations: Note that throughout this paper, we refer to matrices with bold capital letters (e.g., $A$), vectors with small bold letters (e.g., $a$), and scalars or functions with all non-bold letters. $a_{ij}$ is the scalar in row $i$ and column $j$ of $A$, while $a_i$ the $i^{th}$ row and $a_j$ the $j^{th}$ column of $A$.

435 an automated, impartial approach for identifying key brain regions implicated in HIV infection.

Appendix A. Multiple Kernel Learning for Feature Selection and Classification

The MKL of [13, 18] classifies samples by learning the optimal pairings between kernels and image measurements. Finding the optimal paring is described as a minimization problem with respect to a weight vector (denoted $\alpha$), sparsity of which specifies the importance of pairings for class separation. To make the minimization problem tractable, the search for the optimal weight vector is constrained by a regularization term $R(\alpha)$. The minimization problem is furthermore characterized by a prediction function $f(x, \alpha)$ that maps the image measurements $x$ of a sample to a label or cohort (i.e., $y$). The max-margin term $\|f(\cdot, \cdot)\|_{2,1}^{2}$ (or $\|f\|_{2,1}^{2}$ for short) then mea-
sures the distance between the ‘support vectors’ of the classes \(i.e.,\) HIV and CTRL as defined by \(f(\cdot, \cdot)\) (see [53, 99] for detailed definition and see Table A.5 for the notations). Introducing the loss function \(L(y, f(x, \alpha))\) for measuring the difference between the predicted and actual label of a sample, the final term of the minimization problem computes that difference across all training samples, \(i.e.,\) \(L(y, X, f, \alpha) := \sum_{m=1}^{N} L(y_m, f(x_m, \alpha))\). Thus, the regularized MKL approach is completely defined by

\[
\min_{f \in \mathcal{H}, \alpha} \frac{1}{2} \|f\|_{\mathcal{H}}^2 + C \cdot L(y, X, f, \alpha) + \lambda R(\alpha),
\]

\(s.t. \ \alpha \geq 0,\)

where \(C\) and \(\lambda\) are trade-off hyperparameters, the constraint \(\alpha \geq 0\) is needed to efficiently solve the minimization problem (similar to [13, 18]), and \(\mathcal{H}\) is a Reproducing Kernel Hilbert Space (RKHS) [100]. Note that \(\alpha \geq 0\) guarantees that the search for the optimal parameters is done in the space of non-negative values, in which we can define flexible (convex and smooth) regularization functions. The regularizers \(\ell_2,1\) and \(\ell_1\) are only smooth in the domain of non-negative values [18, 27]. For more details, refer to Section Appendix A.1.

In the above objective, function \(f\) is expressed in terms of the aggregated kernel function \(k(\cdot, \cdot, \alpha)\), the weight \(w_n\) of a training samples ‘\(n\)’ in the decision process, and bias parameter \(b\) [101, 53]:

\[
f(x, \alpha) := \sum_{n=1}^{N} w_n \cdot y_n \cdot k(x_n, x, \alpha) + b.
\]

As shown in Figure A.9 (and [13]), the aggregated kernel function \(k(\cdot, \cdot, \alpha)\) applies a set of subkernels \(\{k_1(\cdot, \cdot), \ldots, k_\kappa(\cdot, \cdot)\}\) to each single residual score and then computes a weighted average across all subkernels and residual scores with the weight

\[\text{Then, for this specific application, RKHS is defined as (note that } | \text{ means ‘such that’)}
\]

\[
\mathcal{H} := \left\{ f(\cdot, \cdot) | f(\cdot, \cdot) := \sum_{n=1}^{N} w_n y_n k(x_n, \cdot, \cdot) + b \text{ with } w \in \mathbb{R}^N \text{ and } b \in \mathbb{R} \right\}.
\]
defined by $\alpha$, i.e.,

$$k(x, x_n, \alpha) := \sum_{q=1}^{\kappa} \sum_{i=1}^{d} \alpha_{(i-1)\kappa + q} k_q(x^i, x^i_n).$$  \hspace{1cm} (A.3)$$

An efficient solution to Eq. (A.1) requires the subkernels to be positive semidefinite (PSD), which is a common constraint for kernel methods [101]. Note, that any linear combination (with non-negative coefficients) of PSD subkernels also results in PSD kernel (as in Eq. (A.3)). For our specific application, we choose three types of subkernels. The first one is a Linear (LIN) kernel, which is one of the simplest and most widely used kernels in machine learning:

$$k_{\text{LIN}}(x, x_n) := x^\top \cdot x_n^\top.$$  \hspace{1cm} (A.4)$$

As an alternative to the linear kernel, $k(\cdot, \cdot, \cdot)$ also includes the histogram intersection kernel (HIK) [102], a non-linear kernel popular for non-negative features. This kernel is applied to the absolute values of the residuals in $X$ (as in Eq. (2)), i.e.,

$$k_{\text{HIK}}(x, x_n) := \sum_{i=1}^{d} \min(|x^i|, |x^i_n|).$$  \hspace{1cm} (A.5)$$

Finally, the implementation includes several instances of the Radial Basis Function (RBF) or the Gaussian kernel [103], a popular, non-linear kernel that depends on the kernel hyperparameter $\sigma$:

$$k_{\text{RBF}}(x, x_n) := \exp \left( \frac{\|x - x_n\|^2}{2\sigma^2} \right).$$  \hspace{1cm} (A.6)$$

This kernel can be built by different values of its hyperparameter, $\sigma$. We build several instances of the RBF subkernel with respect to $\sigma \in \{10^{-2}, 10^{-1}, 1, 10, 10^2\}$. Doing so avoids hyperparameter tuning for this kernel, as MKL solves Eq. (A.1) with respect to $\alpha$ to select the pairs of subkernels and residual scores that best fit the data.

Assuming that $\mathcal{R}(\alpha)$ (which is explicitly defined later) is convex and differentiable for non-negative input values (i.e., $\alpha \geq 0$), the solution to Eq. (A.1) can be efficiently determined via Optimize-RMKL$(\cdot)$ (see Algorithm 1). Inspired by [13], “Optimize-RMKL$(\cdot)$”, iteratively solves the equation based on Block Coordinate Descent [104], i.e., by alternating between optimizing for $f$ and $\alpha$ until convergence. When optimizing for $f$ (with $\alpha$ being fixed), Eq. (A.1) reduces to a SVM that can be solved with
Algorithm 1 “Regularized multiple kernel learning” (RMKL), as in Eq. (A.1).

```
Optimize-RMKL (y, X, \( R(\cdot) \), C, \( \lambda \))

Input: Training features X, labels y, the regularization function \( R(\cdot) \), and hyperparameters C and \( \lambda \).

1: \( t \leftarrow 0, \alpha^0 = 1 \)
2: repeat
3: \( f^{t+1} \leftarrow \text{SVMSolver}(y, k(\cdot, \cdot, \alpha^t)). \)
4: \( \alpha^{t+1} \leftarrow \text{Solve (A.1) by using } f^* = f^{t+1} \text{ and regularization } R(\cdot), \text{ using PGD.} \)
5: \( \alpha^{t+1} \leftarrow \max(0, \alpha^t), t \leftarrow t + 1. \)
6: \( M \leftarrow \frac{1}{2} \| f^t \|_2^2 + C \mathcal{L}(y, X, f^t, \alpha^t) + \lambda R(\alpha^t). \)
7: until \( \| \alpha^{t+1} - \alpha^t \|_2 < 10^{-3}, \text{ or } M < 10^{-6}, \text{ or } t > 100 \)

Output: \( f^*, \alpha^t. \)
```

standard approaches (e.g., LIBSVM [105]). To determine the optimal feature-kernel weights \( \alpha \) (with \( f \) being fixed), projected gradient decent (PDG) [106] is applied to Eq. (A.1).

Appendix A.1. Chained \( \ell_{2,1}-\ell_2 \) Regularization

Given the training data as well as the search space \( \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3\} \) for both hyperparameters \( \lambda \) and \( C \) of Optimize-RMKL(\( \cdot \)), our Chained-Regularization approach (called Chained-Reg; see also Algorithm 2) now makes use of Optimize-RMKL(\( \cdot \)) in the Selection Step and the Reweighing Step. In the Selection Step, the regularizer \( R(\cdot) \) of Eq. (A.1) is defined by group sparsity (i.e., \( \ell_{2,1} \) norm) so that Optimize-RMKL(\( \cdot \)) identifies the kernel-measurement pairs that best distinguish the two cohorts. To explicitly define \( R(\cdot) \), we introduce the group matrix \( G(\alpha) \), in which each column represents a group according to bilateral dependencies of ROI measurements and the entries of the rows are the elements of \( \alpha \) corresponding to those measurements. Then, the first step of the Chained-Regularization defines \( R(\cdot) \) as

\[
R_{2,1}(\alpha_1) = \| G(\alpha_1) \|_{2,1}.
\]  

(A.7)

As mentioned, for the selection process to be reliable, we compute a distribution over the selected features by repeatedly solving Eq. (A.1) and then select residual scores based on that distribution. Specifically, the training data is split into 5 (inner)
folds based on random sampling. For each fold, the approach records the set of selected kernel-feature pairs associated with the most accurate (hyperparameter) setting of Optimize-RMKL(·). The accuracy of a setting is determined by parameterizing Optimize-RMKL(·) accordingly, training the approach on the remaining training data, and applying the resulting implementation to the inner fold. The entire process of splitting the training data and recording the set of selected kernel-feature pairs is repeated 9 more times to produce a total of 50 sets. The features selected in all trials then define the ‘selected measurement matrix’ $X'$. Note, this conservative threshold minimizes the chance to introduce another hyperparameter that requires tuning. However, one can implement other selection criteria if required by the application.

Given the selected measurements, the Reweighing Step of the Chained-Regularization again applies Optimize-RMKL(·) to solve Eq. (A.1), but now with respect to $X'$ and the regularizer defined by the Euclidean norm, i.e.,

$$R_2(\alpha_2) = ||\alpha_2||_2.$$  \hspace{1cm} (A.8)

To find the most accurate reweighing $\alpha_2$ of the selected measurements, 5-fold (inner)
cross-validation coupled with hyperparameter exploration is performed. The accuracy
of each hyperparameter setting is computed by first recording the classification re-
results on the fold not used for training Optimize-RMKL(·) and then averaging those re-
tsults across all folds. With respect to the most accurate setting of Optimize-RMKL(·),
Chained-Regularization returns the kernel function \( f \) and corresponding weight vector
\( \alpha_2 \) to define the classifier for the testing data.

We end the description of Chained-Regularization by noting that the proposed
approach is not specific to the two norms discussed here. As mentioned, Optimize-
RMKL(·) only requires that the chosen norms are convex and differentiable in \( \mathbb{R}^{d \times \kappa} \),
which, for example, \( \ell_1 \) and \( \ell_{2,1} \) norms are.

Acknowledgment

This research was supported in part by the NIH grants U01 AA017347, R37-
AA010723, K05-AA017168, K23-AG032872, K24-MH098759, and R01-MH113406.
It was also supported by the Creative and Novel Ideas in HIV Research (CNIHR) Pro-
gram through a supplement to the University of Alabama at Birmingham (UAB) Center
For AIDS Research funding (NIH P30 AI027767). This funding was made possible by
collaborative efforts of the Office of AIDS Research, the National Institute of Allergy
and Infectious Diseases, and the International AIDS Society.

References


associated neurocognitive disorders before and during the era of combination
antiretroviral therapy: differences in rates, nature, and predictors, Journal of
neurovirology 17 (1) (2011) 3–16.
Algorithm 2 Chained $\ell_{2,1}$-$\ell_2$ regularization for joint feature selection & classification.

\begin{algorithm}
\begin{algorithmic}[1]
\State \textbf{Chained-Reg} ($y_{trn}$, $X_{trn}$, $R_+()$, $R_{++}()$, $\zeta$, $C$s, $\lambda$s)
\EndState
\Statex \textbf{Input:} Training features $X_{trn}$, training labels $y_{trn}$, Regularization function for the first step $R_+()$, and for the second step $R_{++}(\cdot)$, number of the inner cross-validation folds, $\zeta$, number repetitions for feature selection, $T$, and the list hyperparameters values to search in, $C$s and $\lambda$s.

1: \textbf{for} $t \in \{1 \ldots T\}$ \textbf{do} \Comment{Selection Step}
2: \hspace{1em} Randomly split $X_{trn}$ to $\zeta$ folds.
3: \hspace{1em} \textbf{for} $\zeta' \in \{1 \ldots \zeta\}$ \textbf{do}
4: \hspace{2em} Define $X'_{val}$ & $y'_{val}$ by the data of fold $\zeta'$, and $X'_{trn}$ & $y'_{trn}$ by the remaining data.
5: \hspace{2em} \textbf{for} $C \in C$s \& $\lambda \in \lambda$s \textbf{do}
6: \hspace{3em} $f_1$, $\alpha_1$ \leftarrow \text{Optimize-RMKL($y'_{trn}$, $X'_{trn}$, $R_+()$, $C$, $\lambda$)}.
7: \hspace{3em} $acc_1[C,\lambda] \leftarrow $ Accuracy of the model defined by $f_1$, $\alpha_1$ on $X'_{val}$ and $y'_{val}$.
8: \hspace{2em} \textbf{end for}
9: \hspace{1em} $f^*_1, \alpha^*_1 \leftarrow $ Model that led to the best accuracy in $acc_1[\cdot,\cdot]$.
10: \hspace{1em} $P_1[T,\zeta] \leftarrow $ The selected features \textit{(i.e., the pattern)} defined by $\alpha^*_1 \neq 0$.
11: \hspace{1em} \textbf{end for}
12: \hspace{1em} \textbf{end for}
13: $P_2 \leftarrow $ The set of features that were always selected in all $P_1[\cdot,\cdot]$.
14: $X'_{trn} \leftarrow $ The reduced feature set according to $P_2$. \Comment{Reweighing Step}
15: \textbf{Split} $\hat{X}_{trn}$ to $\zeta$ folds.
16: \textbf{for} $C \in C$s \& $\lambda \in \lambda$s \textbf{do}
17: \hspace{1em} \textbf{for} $\zeta' \in \{1 \ldots \zeta\}$ \textbf{do}
18: \hspace{2em} Define $X''_{val}$ & $y''_{val}$ by the data of fold $\zeta'$, and $X''_{trn}$ & $y''_{trn}$ by the remaining data.
19: \hspace{2em} $f_2$, $\alpha_2$ \leftarrow \text{Optimize-RMKL($y''_{trn}$, $X''_{trn}$, $R_{++}()$, $C$, $\lambda$)}.
20: \hspace{2em} $acc_{in}[\zeta] \leftarrow $ Accuracy of the model defined by $f_2$, $\alpha_2$ on $X''_{val}$ and $y''_{val}$.
21: \hspace{2em} \textbf{end for}
22: \hspace{1em} $acc_2[C,\lambda] \leftarrow $ mean($acc_{in}[\cdot]$).
23: \hspace{1em} \textbf{end for}
24: $C^*, \lambda^* \leftarrow $ $C$ and $\lambda$ that led to the mean highest accuracy in $acc_2[\cdot,\cdot]$.
25: $f^*_2, \alpha^*_2 \leftarrow $ \text{Optimize-RMKL($y_{trn}$, $\hat{X}_{trn}$, $R_{++}()$, $C^*$, $\lambda^*$)}.
\textbf{Output:} $f^*_2, \alpha^*_2$.
\end{algorithmic}
\end{algorithm}


