Riemannian SPD learning to represent and characterize fixational oculomotor Parkinsonian abnormalities

Juan Olmos, Antoine Manzanera, Fabio Martínez

Highlights

Riemannian SPD learning to represent and characterize fixational oculomotor Parkinsonian abnormalities

Juan Olmos, Antoine Manzanera, Fabio Martínez *

- A deep learning strategy that pools convolutional representations into compact descriptors that then feed Riemannian layers.

- A novel digital biomarker to quantify Parkinson's Disease (PD) from ocular fixation video recordings.

- An explainability strategy that highlights relevant regions over fixational eye video sequences.

- The proposed approach reveals remarkable performance on an extra dataset with more controlled conditions.

Riemannian SPD learning to represent and characterize fixational oculomotor Parkinsonian abnormalities

Juan Olmos[a], Antoine Manzanera[b], Fabio Martínez [*a]

[a] *Biomedical Imaging, Vision and Learning Laboratory (BIVL^{2}ab), Universidad Industrial de Santander (UIS), 680002, Colombia*
[b] *Unité d'Informatique et d'Ingénierie des Systèmes (U2IS), ENSTA Paris, Institut Polytechnique de Paris, Palaiseau, 91762, France*

Abstract

Parkinson's disease (PD) is the second most common neurodegenerative disorder, mainly characterized by motor alterations. Despite multiple efforts, there is no definitive biomarker to diagnose, quantify, and characterize the disease early. Recently, abnormal fixational oculomotor patterns have emerged as a promising disease biomarker with high sensitivity, even at early stages. Nonetheless, the complex patterns and potential correlations with the disease remain largely unexplored, among others, because of the limitations of standard setups that only analyze coarse measures and poorly exploit the associated PD alterations. This work introduces a new strategy to represent, analyze and characterize fixational patterns from non-invasive video analysis, adjusting a geometric learning strategy. A deep Riemannian framework is proposed to discover potential oculomotor patterns aimed at withstanding data scarcity and geometrically interpreting the latent space. A convolutional representation is first built, then aggregated onto a symmetric positive definite matrix (SPD). The latter encodes second-order statistics of deep convolutional features and feeds a non-linear hierarchical architecture that processes SPD data by maintaining them into their Riemannian manifold. The complete representation discriminates between Parkinson and Healthy (Control) fixational observations, even at PD stages 2.5 and 3. Besides, the proposed geometrical representation exhibit capabilities to statistically differentiate observations among Parkinson's stages. The developed tool demonstrates coherent results from explainability maps back-propagated from output probabilities.

Keywords: Keywords:
Oculomotor patterns, Parkinson's Disease classification, Symmetric Positive Definite pooling, Deep non-Euclidean learning, Riemannian manifold

1. Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disorder, affecting more than 10 million people globally nowadays [1]. This incurable disease is today explained by progressive degeneration of dopamine neurotransmitters, affecting the nervous system and producing in consequence alterations in the patient's movement [2]. Currently, there is no definitive disease biomarker, and the clinical diagnosis and prognosis are typically limited to observational analysis and coarse scales to stratify motor disabilities [2, 3]. PD symptoms affect the quality of life and can span decades, therefore timely treatment and a correct characterization are essential to slow down motor impairments and disabilities [2]. Additionally, there is a wide range of motor manifestations and disease phenotypes to characterize the disease, such as: tremor in hands, disabilities during gait, and trunk rigidity [4]. Nonetheless, such motor impairments have proven to be unsuitable for precise quantification of the disease progression, and they are mostly detected at an advanced stage of the disease [4].

Recently, some works have highlighted ocular tremor as a distinctive manifestation in PD patients, with sufficient sensitivity to capture abnormalities, even at the early stages of the disease [4, 5, 6]. However, most existing methods capture ocular patterns from sophisticated and intrusive devices that simplify the eye dynamics to displacement trajectories, limiting the understanding of the disease progression [7]. Consequently, today there is no clear evidence of such descriptors as a

Email addresses: jaolmosr@correo.uis.edu.co (Juan Olmos), antoine.manzanera@ensta-paris.fr (Antoine Manzanera), famarcar@saber.uis.edu.co (Fabio Martínez [*])

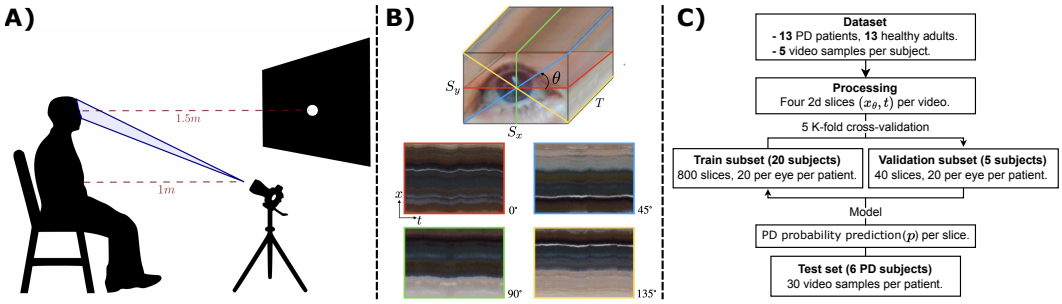

Figure 1: A) Recording set up of the oculomotor fixation task. B) Video slices representation of eye movements during the fixation task. C) Flowchart of the study.

support to disease evaluation, diagnosis, and follow-up in clinical routines.

Nowadays, computer-aided diagnosis systems have integrated machine learning mechanisms, allowing, among others, the analysis of motor information to guide PD classification [8]. These methods typically use information from sensors to approximate clinical variables like postural instabilities during gait, voice irregularities, cerebrospinal data, or eye movement tracking [9, 8]. Also, some alternatives have used markerless setups with deep learning strategies to recover uncontrolled Parkinsonian patterns during locomotion, showing promising results regarding the discrimination between Control and Parkinson patients [10]. Despite the recovery of kinematic descriptors correlated with the disease, such strategies require a vast amount of labeled data to properly learn discriminative patterns, which is a limitation in the medical domain [8]. Hence, the design of new tools to support PD characterization should consider motor descriptors that learn high pattern variability over a set of limited observations and allow exploring unknown motor correlations associated with the disease. Second-order compact descriptors have shown a higher discriminative capability [11] by pooling feature observations into symmetric positive definite (SPD) matrices. Nevertheless, such descriptors imply particular non-Euclidean processing constraints since they lie within a Riemannian manifold [12].

This work presents a novel digital biomarker that encodes PD fixational eye patterns by learning a deep Riemannian representation. The proposed method pools convolutional features into an SPD matrix of second-order statistics. The proposed markerless strategy uses spatio-temporal slices computed from video sequences that record micro-tremor patterns during an ocular fixation experiment. A projection of this compact descriptor into SPD layers allow the model to perform Riemannian learning by preserving the geometry of input SPD data, and achieving higher discrimination between Parkinsonian and Control classes. A preliminary version of this work appeared in [13]. This extended version performs a comprehensive analysis of the results, evaluating the model's ability to detect Parkinsonian patterns. Furthermore, a study is conducted to assess the model's ability to aid disease stratification by evaluating its sensitivity to distinguish among different stages of PD. The proposed architecture supported the stratification of relative early stages (2.5 and 3), which do not present advanced motor symptoms (in stage 2.5), such as impairment of balance, and only present moderate symptoms (in stage 3) related to postural instability and gait motor abnormalities. This work also introduces an interpretability strategy that recovers explainability maps, highlighting important spatiotemporal regions that support PD prediction. Furthermore, a new dataset is included to assess the generalization capability of the proposed approach.

2. Dataset

A retrospective study was conducted with 13 patients diagnosed with PD (average age of 72.3 ± 7.4) and 13 Control subjects (average age of 72.2 ± 6.1). To include inter-subject variability, the study incorporates PD subjects with different disease degree progression. The modified Hoehn-Yahr rating scale was used to categorize PD patients with the aid of a physical therapist. A total of five patients were categorized in stage 2.5, six patients in stage 3, and two in stage 4.

To record oculomotor patterns, the participants were invited to observe a fixed stimulus in front of them (at 1.5m). The screen was positioned at the same height as the eyes to avoid extra efforts of subjects. From 1m ahead and at a lower altitude, we recorded the upper face of the participants, see Fig. 1.A). Then, over the screen was projected a white spot disk. Participants received the instruction to fix their gaze on the white spot during

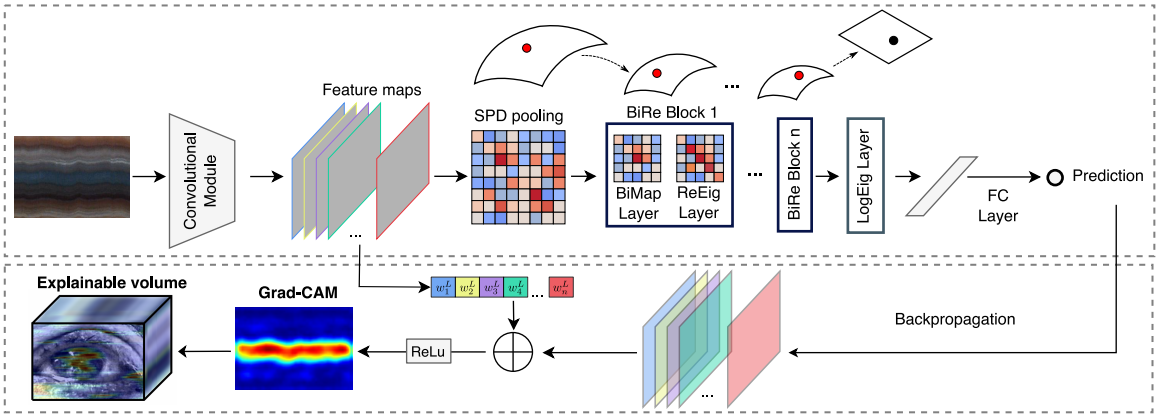

Figure 2: **Pipeline and network architecture.** Top: architecture of the mixed Convolutional / SPD (ConvSPD) model. Bottom: interpretation of the model using GradCAM maps from the convolutional module and reconstructed explainability video.

the recording. We used a standard optical camera with a temporal resolution of 60 fps. For each patient, we cropped recorded sequences to create five video samples of individual eyes for each patient, where each one constitutes a fixation period of five seconds, and the spatial resolution was cropped to 210×140 pixels, centering the first frame to the center of the pupil.

To recover tremor observations from each video clip, spatio-temporal information was captured in 2D (x_θ, t) video slices (see Fig. 1.B)). For this purpose, each video is considered as a volume $\{I(x,y,t)\}_{x=1,y=1,t=1}^{S_x,S_y,T}$, with spatial and temporal dimensions $S_x \times S_y$ and T (number of frames) respectively. To obtain the slices, we choose four radial directions $\theta \in \{0°, 45°, 90°, 135°\}$ around the center and cut the video volume along each direction. This way, each 2D image slice (x_θ, t) records temporal variations along the x_θ axis. The resulting image slices record subtle eye displacements, capturing potential differential oculomotor patterns related to the disease. For each patient, a total of 40 2D slice samples were collected, where each slice was extracted from video samples from each eye in the chosen four directions. The data used in this study is publicly available and can be accessed through the repository [1], which also provides details regarding the source codes used for data preparation.

An extra dataset was captured only for test purposes, using semi-controlled conditions but from other institutions. We used a chin rest to improve head stabilization in an updated protocol. A total of 6 PD patients (average age of 70.16 ± 6.8) were recorded. This dataset has undergone the same preprocessing as the previous sequences, and then slice samples were fed to the trained architecture. This allows evaluation of the proposed method's generalization under different conditions, such as illumination changes and recording angles. The entire dataset, including the test data, was approved by an Ethic Committee, and each participant filled out a written informed consent. To our knowledge, no public dataset contains more eye movement videos for Parkinson's prediction.

3. Proposed Method

In this work, we analyze fixational oculomotor abnormalities from spatio-temporal video slices by first, learning a deep convolutional representation, whose patterns are encoded into an SPD embedding, which is then exploited through a non Euclidean learning framework that preserves Riemannian manifold properties. The proposed representation is learned from an end-to-end scheme, taking advantage of convolutional hierarchical representation encoded in top layers as second-order statistics, allowing to discriminate Parkinsonian fixational patterns from Control subjects. Figure 2 summarizes the proposed method. The repository including the source code for training the proposed model is available at [1].

3.1. Deep features encoding

The first part consists in decomposing spatio-temporal slices into a hierarchical convolutional bank of

[1] https://gitlab.com/bivl2ab/research/SPDLearning-PD-FixationalPatterns

filter responses (see Fig. 2). This Convolutional module is composed of several layers, hierarchically organized to progressively expand the time × space receptive field, as well as the semantic level w.r.t. PD classification. At the end of this module, we obtain a bank of N feature maps $\{\mathbf{F}^{(i)}\}_{i=1}^{N}$ with dimension $W \times H$.

3.2. Non-linear SPD pooling

In a typical approach, a low-dimensional descriptor is embedded at the top of the convolutional layers to carry out class discrimination. Nonetheless, this abrupt dimension reduction may lose significant feature relationships that may highlight associated class patterns. Recently, SPD embeddings have gained attention because of the capability to recognize patterns and encode compact descriptors of high dimensional observations [14]. We decided to endow the proposed network with an intermediate SPD pooling layer dedicated to capturing second-order statistics in a compact matrix embedding that summarizes relevant information from the features bank $\mathbf{X}_{k-1} = \{\mathbf{F}^{(i)}\}_{i=1}^{N}$. We re-organized them in a matrix $\mathbf{X}_{k-1} \leftarrow [\mathrm{vec}(\mathbf{F}^{(1)}) \ \cdots \ \mathrm{vec}(\mathbf{F}^{(N)})]$. Then, the proposed SPDpool layer computes

$$\mathbf{X}_k = f_{SPDpool}(\mathbf{X}_{k-1}) = \frac{1}{W \times H}\mathbf{X}_{k-1}^{\top}\mathbf{X}_{k-1}$$

The output is a $N \times N$ Gram matrix where $\mathbf{X}_k(i, j)$ records the inner product (correlation) between the i-th and j-th feature map. This matrix has been used to estimate statistical discrepancy [15]. The resultant SPD embedding allows encoding relevant relationships in the previous CNN module w.r.t the significant Parkinsonian patterns.

3.3. Riemannian Module Structure

The resultant SPD embeddings form a Riemannian manifold that must be addressed using non-Euclidean learning to preserve the geometric structure. To learn patterns from such SPD representation and discriminate samples according to supervision labels, we follow the transformations on the SPD manifold as described by [12].

Firstly, a bilinear mapping (*BiMap*) layer is used to transform SPD matrices into a new bank of more compact (lower dimension) SPD matrices through bilinear mapping:

$$\mathbf{X}_k = f_{BiMap}(\mathbf{X}_{k-1}) = \mathbf{W}_k\mathbf{X}_{k-1}\mathbf{W}_k^{\top},$$

where, $\mathbf{X}_{k-1} \in S_{++}^{d_{k-1}}$ is the input SPD matrix of the previous layer $(k-1)$, while $\mathbf{W}_k \in \mathbb{R}_*^{d_k \times d_{k-1}}$ is the transformation matrix (connection weight) that generates the new SPD matrix $\mathbf{X}_k \in S_{++}^{d_k}$. This layer is used after applying the *SPDpool*. Similar to a CNN, the sizes of the SPD matrices decrease after the BiMap layer, i.e. $d_k < d_{k-1}$. The connection weights $\mathbf{W}_k$ involved in this layer should be an orthogonal full-rank matrix in order to generate a consistent output $\mathbf{X}_k$ SPD matrix. This implies that the weights matrices lie in the compact *Stiefel manifold* $St(d_k, d_{k-1}) = \{\mathbf{W} \in \mathbb{R}^{d_k \times d_{k-1}} | \mathbf{W}\mathbf{W}^{\top} = \mathbb{I}_{d_k}\}$ [12]. Additionally, this constraint aids the optimization to achieve optimal solutions on $St(d_k, d_{k-1})$ during the learning process, avoiding data degeneration problems [12].

Then, after the BiMap, a eigenvalue rectification (*ReEig*) layer is implemented. This regularization is inspired by the rectified linear units (ReLU) and their effectiveness in improving non-linear learning [12]. Particularly, the ReEig layer is composed of a non-linear function to improve the training process by rectifying the eigenvalues of SPD matrices via:

$$\mathbf{X}_k = f_{ReEig}(\mathbf{X}_{k-1}) = \mathbf{U}_{k-1} \max(\varepsilon I, \mathbf{\Sigma}_{k-1})\mathbf{U}_{k-1}^{\top},$$

where $\mathbf{U}_{k-1}$ and $\mathbf{\Sigma}_{k-1}$ come from the eigenvalue decomposition $\mathbf{X}_{k-1} = \mathbf{U}_{k-1}\mathbf{\Sigma}_{k-1}\mathbf{U}_{k-1}^{\top}$. Here, ε is a non-negative rectification threshold. This operation tunes up the eigenvalues avoiding non-positiveness, preserving the SPD data structure, and consequently improving the discriminative performance. The BiMap and ReEig layers constitute the main layers of the network, with a BiMap layer always followed by a ReEig layer. We refer as *BiRe block* to the concatenation of these layers, see Figure 2.

At the end of the SPD BiRe blocks sequence, the Riemannian information is projected back into Euclidean space to perform the classification part. For that purpose, the LogEig layer projects the SPD data into a Euclidean space using the Riemannian logarithm map:

$$\mathbf{X}_k = f_{Log}(\mathbf{X}_{k-1}) = \mathbf{U}_{k-1} \log(\mathbf{\Sigma}_{k-1})\, \mathbf{U}_{k-1}^{\top}.$$

The resulting matrix $\mathbf{X}_k$ is a Euclidean squared matrix that facilitates the computation of the following output layers [12]. These layers consist of a Flatten layer, a fully connected layer, and the final output layer is a sigmoid operation.

3.4. Learning Scheme

From a binary cross-entropy loss function, a stochastic optimization algorithm using adaptive moment estimation (*Adam* algorithm) is implemented. However, two issues arise using Euclidean gradients and traditional backpropagation (BP) within the Riemannian layers. The first is that the eigenvalue decomposition of Eig

layers (*ReEig* and *LogEig*) is not well posed in the traditional BP. The second issue concerns the orthogonality constraint of the connection weights in the *BiMap* layers. The updated weight should ensure generating valid SPD matrices [12]. For the first issue, structured derivatives have been proposed as a solution [16]. Regarding the learning in BiMap layers, the steepest descent direction within the Stiefel manifold has been calculated [12, 13]. For that, it is necessary to calculate the tangent component of $St(d_k, d_{k-1})$, which is defined as the subtraction of the Euclidean gradient $\nabla L^{(k)}_{\mathbf{W}^t_k}$ by the normal component: $\widetilde{\nabla} L^{(k)}_{\mathbf{W}^t_k} = \nabla L^{(k)}_{\mathbf{W}^t_k} - \nabla L^{(k)}_{\mathbf{W}^t_k} \left(\mathbf{W}^t_k\right)^\top \mathbf{W}^t_k$. Now, the tangent component $\widetilde{\nabla} L^{(k)}_{\mathbf{W}^t_k}$ can be seen as the direction to update the connection weight $\mathbf{W}^t_k$. Finally, the retraction operation Γ over $St(d_k, d_{k-1})$ is used to map this weight back from the tangent space into $St(d_k, d_{k-1})$ via:

$$\mathbf{W}^{t+1}_k = \Gamma\left(\mathbf{W}^t_k - \alpha \widetilde{\nabla} L^{(k)}_{\mathbf{W}^t_k}\right). \tag{1}$$

The result is an updated weight $\mathbf{W}^{t+1}_k$ using a learning rate α, and the retraction operation is defined from the *QR* decomposition [17]. This way, the Riemannian weights are updated, and the loss can be back-propagated through the successive convolutional layers.

4. Spatio-temporal explainability maps

In this work, we also seek to discover the underlying characteristics of spatiotemporal regions that may explain particular disease prediction in order to complement the analysis. To do so, we implement an interpretability module into the proposed architecture based on a Gradient-weighted class activation mapping (GradCAM) (see in Fig. 2) [18]. This provides a visual explanation output as a new spatiotemporal slice that informs about the contribution of each space $\times$ time location w.r.t a particular prediction. More precisely, for an input 2D slice (x_θ, t), the model produces a bank of feature maps in their last convolutional layer $\{\mathbf{F}^{(k)}(i_x, i_t); 1 \leq k \leq N, 1 \leq i_x \leq W, 1 \leq i_t \leq H\}$, where N, W, H refer to the number of channels, space samples and time samples of the considered layer. Then, the output PD prediction probability p is related w.r.t changes at this convolutional level, by taking $\tilde{p} = \max\{p, 1-p\}$ and computing gradients of $\tilde{p}$ with respect to $\{\mathbf{F}^{(k)}(i_x, i_t)\}$. After applying a back-propagation from the chain rule including Riemannian derivatives, we calculate the global average of gradients of features from this layer, as follows:

$$w_k = \frac{1}{W \times H} \sum_{i_x=1}^{W} \sum_{i_t=1}^{H} \frac{\partial \tilde{p}}{\partial \mathbf{F}^{(k)}(i_x, i_t)}.$$

This way, w_k is interpreted as a weight that quantifies the global importance of the k-th feature map w.r.t the model prediction $\tilde{p}$. Then, explainability maps are weighted and rectified by the ReLu function

$$F_{GradCAM} = max\left(0, \sum_{k=1}^{K} w_k \mathbf{F}^{(k)}(i_x, i_t)\right).$$

The resultant maps are valuable to complement prediction assistance, helping to visually identify discriminative patterns from the video slices. In clinical scenarios, such tool can be used to support the diagnosis, highlighting important regions that suggest Parkinsonian abnormalities.

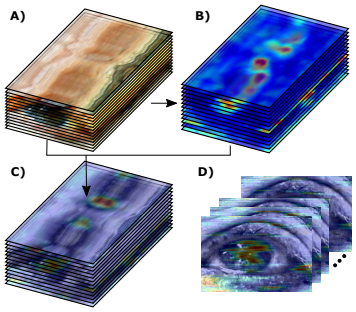

Figure 3: **A)** Video slices along the vertical axis ($\theta = 0°$). **B)** Corresponding 2D GradCAM explainability maps. **C)** Explainability maps over original video slices. **D)** Video reconstruction by stacking such maps.

Finally, from the previous explainability maps computed on the space $\times$ time slices, we can reconstruct an explainability video that highlights regions during the fixation task: Firstly, we calculate S_y slices varying the height of the cutting plane from 0 to $S_y - 1$. For each slice we calculate a 2D GradCAM map (see Fig. 3.B). To highlight regions over original slices we overlap each slice with their correspondent GradCAM map, as shown in Fig. 3.C). The stacking of the previous maps provides the explainability video, as illustrated in Fig. 3.D).

5. Experimental Setup

Network Configuration. The architecture of each module and training were configured as follows:

- CNN module: weights were initialized from a pre-trained CNN architecture [19] with 8 convolutional blocks. From the 2^{nd}, 4^{th}, 6^{th} and 8^{th} block, this module outputs 64, 128, 256 and 512 feature maps of size 53×75, 27×38, 14×19 and 7×10 respectively.

- SPD pooling layer ($f_{SPDpool}$): Several experiments were run to provide the SPD embedding, computed either from the 2^{nd}, 4^{th}, 6^{th} or 8^{th} block.

- The Riemannian module: We compared shallow and deep modules with 3 BiRe blocks (m^{th}-3BiRe models) and 1 BiRe block (m^{th}-1BiRe models). A rectification threshold $\varepsilon = 10^{-4}$ was set. In order to output a PD probability estimation, the final layers correspond to one *LogEig* layer, a flatten layer, and a fully connected layer with a sigmoid function.

Each ConvSPD m^{th}-nBiRe model was trained during 50 epochs. The CNN module was trained with a learning rate of 10^{-4}, while the Riemannian module was trained with a learning rate of 10^{-3}, using the equation 1.

Validation. A 5-fold cross-validation scheme was set. Each fold was configured with 21 subjects for training ($840 = 21 \times 40$ 2D samples) and five ($200 = 5 \times 40$ 2D samples) for testing. The test sets contain at least two patients with PD in these fold splits. The 2D samples coming from one same patient were exclusively used for either training or testing during the splitting process. Each model's performance was measured using the average sensitivity, accuracy, and F1-score of the folds. Additionally, the size of the networks (number of parameters) and inference time were considered.

6. Evaluation and results

The proposed approach was comprehensively evaluated on the discrimination between Parkinsonian and Control subjects' fixation recordings. First, an ablation study was carried out to measure the respective impacts of depth convolutional and Riemannian BiRe blocks. The ConvSPD m^{th}-n models based on the m^{th} convolutional layer and using n BiRe blocks were compared, for $m \in \{2, 4, 6, 8\}$ and $n \in \{1, 3\}$. Table 1 summarizes the results from such configurations. The intermediate convolutional blocks provide the best results with an accuracy of 98.2% (±2.2), 98.6% (±1.4), and 98.6% (±1.6) for 4^{th}-3BiRe, 6^{th}-1BiRe, and 6^{th}-3BiRe model respectively. In general, all configurations reached scores above 95%, which shows the capability of deep non Euclidean representations to discriminate Parkinson from Control subjects by extracting tiny microtremor patterns from fixational exercises.

At the end of Table 1 are presented the results with purely convolutional architectures, where a global average pooling layer replaced the SPDpooling and Riemannian layers. Different convolutional models with m layers were tested, for $m \in \{2, 4, 6, 8\}$. The performance increases for models that comprise more layers, achieving an accuracy of 93.2% (±6.8) with the Conv-2^{nd} model and above 96% for Conv-6^{th} and Conv-8^{th} models. The number of parameters is presented in the table for a fair comparison w.r.t ConvSPDmodels of the same size. This evidences the contribution by the SPD modules, which provide an accuracy gain of about 0.9-3.3% compared to convolutional models. Moreover, the highest gain is obtained for smaller architectures, as seen comparing the ConvSPD 2^{nd}-1BiRe and Conv-2^{nd} models. Additionally, the standard deviations of the Conv-m^{th} models are higher than those of the ConvSPD models.

In addition, we studied the capability of each model to differentiate among different disease stages. For this purpose, we stratified the data into Control, Stage 2.5, Stage 3, and Stage 4 subsets. Then, we recover the output probabilities related to the disease class for each sample from the validation test. This experiment was run with all ConvSPD models, and the unique model that produced statistically different distributions among subsets was the ConvSPD 4^{th}-3BiRe. The proposed method was designed and trained for binary PD / Control classification, but retrieving an output PD probability that allows analyzing the behavior for samples labeled with different PD stages. As it happens, we are interested in assessing the stratification performance using the stage labels of the patients. Figure 4 shows the violin plots of PD output probabilities for each disease stage (including Control) group produced by the ConvSPD 4^{th}-3BiRe net. As expected, clear discrimination is observed between the control population and the PD groups. Besides, within PD stages, we observe that stage 4 sample probabilities were narrowly distributed around p = 1. Interestingly, there exist statistical differences among the three-stage considered groups. For this experiment, all PD patients were correctly predicted, and only a few sample slices were misclassified (11 slices for one patient at stage 3 and three slices for another PD patient at stage 2.5). To measure the statistical difference, we implemented the Kolmogorov-Smirnov (KS) test that validates the capability of the proposed approach to separate classes. The KS test is

Table 1: Classification results (%) for the different m^{th}-nBiRe ConvSPD models. For comparison, the bottom part shows the performance for purely convolutional networks, Conv-m^{th} meaning keeping only the m^{th} first convolutional layers.

Model	Sensitivity	Accuracy	F1-score	Parameters	Inference Time
2^{nd}-1BiRe	95.7 ± 7.0	96.5 ± 4.2	96.6 ± 3.7	0.161M	17.13ms
2^{nd}-3BiRe	92.4 ± 9.6	95.5 ± 5.4	95.3 ± 5.1	0.160M	16.64ms
4^{th}-1BiRe	95.8 ± 5.5	97.2 ± 3.1	97.2 ± 2.7	0.695M	36.02ms
4^{th}-3BiRe	**97.4** ± 3.8	**98.2** ± 2.2	98.2 ± 1.8	0.694M	32.14ms
6^{th}-1BiRe	**98.5** ± 2.2	**98.6** ± 1.4	98.6 ± 1.5	2.832M	89.63ms
6^{th}-3BiRe	**98.0** ± 2.8	**98.6** ± 1.6	98.7 ± 1.3	2.827M	75.31ms
8^{th}-1BiRe	97.1 ± 5.8	97.2 ± 3.5	97.4 ± 3.2	11.373M	281.99ms
8^{th}-3BiRe	96.6 ± 6.0	97.3 ± 3.9	97.6 ± 3.5	11.353M	243.13ms
Conv-2^{nd}	92.8 ± 10.8	93.2 ± 6.8	93.7 ± 5.8	0.158M	4.1ms
Conv-4^{th}	95.3 ± 7.0	95.3 ± 4.5	95.7 ± 3.6	0.683M	4.18ms
Conv-6^{th}	94.9 ± 10.7	96.6 ± 6.2	96.7 ± 5.9	2.783M	4.21ms
Conv-8^{th}	95.2 ± 9.0	96.4 ± 5.1	96.6 ± 4.7	11.177M	4.24ms

a non-parametric method that measures the agreement between two distributions. In such a way, it allows testing the hypothesis that two sample probability distributions are not significantly different. In such case, the proposed approach (configuration ConvSPD 4^{th}-3BiRe) finds significant differences between early stage 2.5 and 3 ($p < 0.07$), stage 2.5 and 4 ($p < 10^{-3}$), and between stage 3 and 4 ($p < 10^{-4}$). However, this p-value only indicates significant differences of probability distribution among the stages. It brings insights into the model's potential to discriminate among classes, but further analysis of the network and its components is necessary.

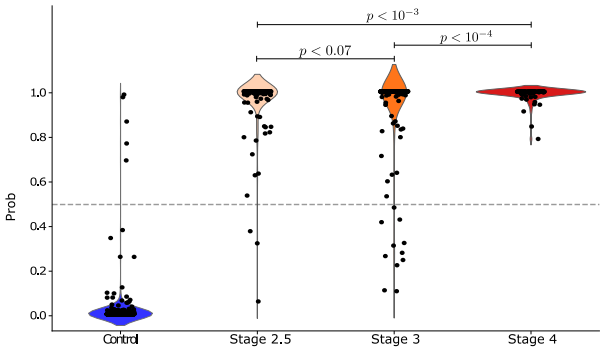

Figure 4: Distributions of Parkinson probability outputs by the best binary classifier, for the different stage groups.

We also evaluated the interpretability of the proposed approach by generating synthetic videos from explainability maps. In order to recover spatiotemporal regions with major association with output predictions, we computed explainability maps calculating GradCAM maps of slices samples. Then, a stacked block of explainability maps, recovered from a particular patient video, was used to reconstruct a synthetic video that highlights regions with significant disease correlation. The explainability slices and videos were herein designed as a possibility to analyze computational decisions, which can aid the patient analysis during the clinical examination. Figure 5 summarizes the achieved results on slices and synthetic videos from output predictions of Parkinsonian and control fixational patterns. The explainability maps and synthetic videos were also recovered from the purely convolutional model as a baseline. In Figure 5.A) is presented the average GradCAM heatmap of slice samples for different ConvSPD and purely convolutional models. This in order to synthesize the most relevant region in general regarding the output predictions for each model. Here, the maps obtained using Riemannian representations (ConvSPD models) show evident importance in the center region, where the response of fixational abnormalities occurs. In addition, for both Control and PD, the ConvSPD 4^{th}-3BiRe achieve a better ability to focus on such region. Despite this, analysis from average maps hampers the clarity of abnormalities. In this respect, we analyzed single slices samples from all individuals. For instance, the last column of Figure 5.A shows a sample slice of a PD patient and a healthy subject. In general, we observe a continuous and regular pattern along the temporal axis for control samples. In contrast, for PD patients, the maps typically report attention peaks along the temporal dimension. In fact, we analyze such particular regions into retrieved explainability videos, finding that the associated temporal intervals with a major probability of PD have a notable tremor and abrupt motion of the iris. On the other hand, Figure 5.B) illustrates the synthetic videos reconstructed from the recovered explainability maps

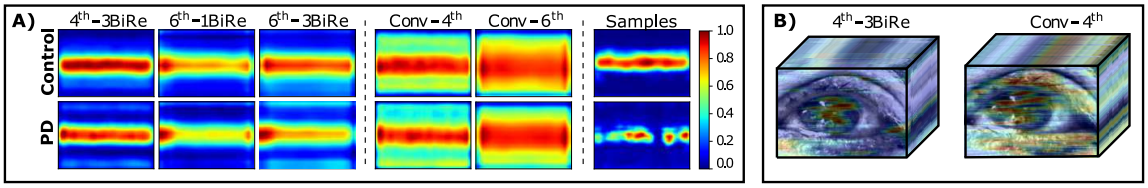

Figure 5: **A)** Comparison of GradCAM heatmaps over horizontal slices ($\theta = 0°$) averaged for each class, computed from the last layer of the convolutional module of the ConvSPD models, and from the last convolutional layer of the Conv models. The last column shows GradCAM map extracted from 4^{th}-3BiRe model, but for one single PD patient and a Control slice sample. **B)** Synthetic video reconstruction from explainability maps.

for both: Riemannian and purely convolutional strategies. The videos recovered from the proposed approach focus mainly on the region of the iris, which in time, considers in some cases, the blinking patterns. In contrast, the convolutional strategies produce sparse patterns without coherence in many regions responsible for producing predictions. In that sense, recovered videos using the proposed method may support observational analysis of the disease, during a clinical examination.

Finally, concerning the capability of the proposed approach in terms of generalization and robustness to different conditions, we evaluate the best model on the extra dataset with a total of 5 patients. The proposed approach correctly predicted all samples on this extra dataset, achieving perfect accuracy, F1-score, and sensitivity scores with an average PD probability prediction (p) of 99.94%. This result reveals the robust capabilities of the method to discriminate PD patients, even in different conditions and capture setups. It should be noted that this extra dataset only considers six extra patients, and the capture setups include head stabilization.

7. Discussion

This paper presented a deep learning-based strategy to characterize oculomotor fixational patterns, based on a convolutional representation encoded in a compact SPD descriptor that operates in a Riemannian manifold. The proposed approach has revealed stable convergence and robustness to learn patterns from limited training data, a common case in medical contexts [13]. The method achieves 98.2% of accuracy using four convolutional blocks and three Riemannian BiRe structures (4^{th}-3BiRe model) in a real scenario with 13 PD patients and 13 Control subjects. From an ablation study, we observed that intermediate layers aid the network to capture spatio-temporal patterns associated to PD. We also note that more stacked BiRe blocks only improve models that use intermediate representations. This might suggest that the learning of the Riemannian module depends on the quality of the SPD layer encoded statistics. Besides, the proposed method outperforms purely convolutional models, with a higher level of confidence. The high standard deviation on purely convolutional models evidences a lack of robustness of these methods on different validation sets. Regarding the higher inference time, it must be said that SPD layers involve numerical operations that have not been optimized yet in existing frameworks. Whatever, such times (< 1 sec) are perfectly acceptable in the routine evaluation of PD patients.

In a clinical context, an important factor to include tools in standard analysis protocols is the capability to explain outputs to support decisions coherently. Beyond predictions, the strategies should be interpretable regarding the link between the model's inputs and its decision. In such a sense, an additional contribution of the proposed strategy is the adaptation of an explainability mechanism to highlight the main spatiotemporal regions of ocular movements that contribute to the prediction. For this, during the test phase, a stack of slices was computed from each particular video and projected to the proposed approach to retrieve explainability maps. From these maps, a synthetic explainability video was reconstructed from slices, returning a video that highlights the main eye regions associated with the prediction of the proposed approach. As expected, the explainability maps from the Riemannian representation retrieve a more coherent representation that may support observational analysis to discover abnormal oculomotor patterns associated with a particular PD-diagnosed patient. For instance, for the Control subjects, the highlighted regions are more regular, which in synthetic videos translates into a constant focus in the subject's iris region. On the contrary, for patients with the disease, the maps have peaks of highlighting in some temporal fragments, which coincides with ocular tremor movements on original raw videos. Particularly, we observed this Parkinsonian pattern on

slices samples misclassified as control. The explainability video volumes highlight skin regions such as the eye rims and eyebrows. In contrast, although convolutional schemes bring competitive prediction scores, their retrieved Grad-CAM maps are sparse, without significant attention region around the eye. To our knowledge, this is the first time that explainability maps have been retrieved from SPD Riemannian representation produced from a convolutional backbone. As a perspective, measuring singularities from explainability maps could produce a better quantification of oculomotor abnormalities.

Additionally, we evaluated the sensitivity of discrimination across different stages of PD based on output PD predictions from 2D samples. Here the 4^{th}-3BiRe model produced statistically different probability distributions among the PD stages. As expected, the distribution of the control subject samples exhibits statistically different distinctions with all stages of the disease. Likewise, in advanced stages (stage 4), the model easily predicts all 2D slice samples closely spread around higher scores. However, for the early stages (2.5 and 3), the distributions were more sparse and challenging to distinguish, which can be explained by the unclear frontiers between the H&Y rating scales. Indeed, some clinicians suggest that the progression is not linear and that an increment in the scale does not necessarily imply a higher degree of general motor dysfunction [20]. On the other hand, analyzing the video sequences of the stage 3 patient whose 11 2D samples were misclassified, we observed a slight loss of sharpness in some video fragments, which affects the correct encoding of fixational alterations. Despite this, it was possible for the binary classification task to correctly classify all the subjects by averaging the predictions from the samples. Remarkably, the proposed approach has the ability to detect early stages samples at levels 2.5 and 3 of the H&Y rating scale. Particularly, this scale characterizes patients with non-advanced motor symptoms like level 2, which can be challenging for physicians to detect and differentiate in clinical evaluations. Similarly, level 3 is only associated with postural instability but not necessarily with tremor impairments. Finally, the model achieves remarkable generalization over head-stabilized extra data, supporting the hypothesis that PD quantification is performed based on ocular abnormalities, not head motion [7]. Further evaluations of the proposed approach should consider larger datasets that include stratified populations with motor PD impairments. Additionally, patients at earlier and prodromal stages should also be considered. Also, as perspectives, longitudinal studies could be used to validate the capability of the proposed approach to support progression quantification at each particular patient. In the literature, a method was proposed to quantify tremor patterns from 2D eye slices using standard machine learning techniques and representing features from a pre-trained bank of filters [21]. In reported results, such an approach achieved an accuracy of 87.7% using the same dataset, but following a leave one out cross validation, representing 10% lower than the proposed method. This fact may be associated with the general deep feature representation that is not optimized to capture the associated ocular abnormalities. In fact, the proposed approach during the Riemannian learning reaches an optimal representation to discriminate between control and Parkinsonian patterns.

8. Conclusions

This paper introduced a novel PD digital biomarker from fixational oculomotor patterns using a mixed deep-learning strategy. The approach integrates convolutional and Riemannian modules to discriminate between Parkinson and Control populations, and can retrieve useful information for medical experts by calculating explainability maps that emphasize relevant spatiotemporal regions from input video sequences. The tool could effectively aid PD diagnosis, even at H&Y stages two and three. Future work involves studying the generalization of the approach in different clinical settings, assessing training robustness with unbalanced scenarios, and adapting the method to raw volumetric inputs from complete video sequences.

Acknowledgments

The recorded dataset was possible thanks to the support of the Parkinson foundation FAMPAS and the institution *Asilo San Rafael*.

Funding

Ministry of Science Technology and Innovation (MINCIENCIAS), project: *Caracterización de movimientos anormales del Parkinson desde patrones oculomotores, de marcha y enfoques multimodales basados en visión computacional*. Code 92694.

References

[1] Parkinson's foundation, Understanding Parkinson's: statistics, from `www.parkinson.org/understanding-parkinsons/` (2023).

[2] W. Poewe, K. Seppi, C. M. Tanner, G. M. Halliday, P. Brundin, J. Volkmann, et al., Parkinson disease, Nature Reviews Disease Primers (2017) 1–21.
[3] E. Tolosa, A. Garrido, S. W. Scholz, W. Poewe, Challenges in the diagnosis of Parkinson's disease, The Lancet Neurology 20 (5) (2021) 385–397.
[4] R. S. Weil, A. E. Schrag, J. D. Warren, S. J. Crutch, A. J. Lees, H. R. Morris, Visual dysfunction in Parkinson's disease, Brain: A Journal of Neurology 139 (11) (2016) 2827–2843.
[5] F. Chan, I. T. Armstrong, G. Pari, R. J. Riopelle, D. P. Munoz, Deficits in saccadic eye-movement control in Parkinson's disease, Neuropsychologia 43 (2005) 784–796.
[6] G. T. Gitchel, P. A. Wetzel, M. S. Baron, Pervasive ocular tremor in patients with Parkinson disease, Archives of Neurology (2012).
[7] G. T. Gitchel, P. A. Wetzel, A. Qutubuddin, M. S. Baron, Experimental support that ocular tremor in Parkinson's disease does not originate from head movement, Parkinsonism & related disorders 20 (7) (2014) 743–747.
[8] M. Belić, et al., Artificial intelligence for assisting diagnostics and assessment of Parkinson's disease—a review, Clinical neurology and neurosurgery 184 (2019).
[9] W. Wang, J. Lee, F. Harrou, Y. Sun, Early detection of Parkinson's disease using deep learning and machine learning, IEEE Access (2020).
[10] L. C. Guayacán, E. Rangel, F. Martínez, Towards understanding spatio-temporal Parkinsonian patterns from salient regions of a 3d convolutional network, in: Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2020, pp. 3688–3691.
[11] M. Harandi, et al., Dimensionality reduction on SPD manifolds: The emergence of geometry-aware methods, IEEE transactions on pattern analysis and machine intelligence 40 (2017) 48–62.
[12] Z. Huang, L. Van Gool, A Riemannian network for SPD matrix learning, AAAI Conference on Artificial Intelligence (2017).
[13] J. Olmos, A. Manzanera, F. Martínez, An oculomotor digital parkinson biomarker from a deep Riemannian representation, in: International Conference on Pattern Recognition and Artificial Intelligence, Springer, 2022, pp. 677–687.
[14] Q. Wang, J. Xie, W. Zuo, L. Zhang, P. Li, Deep CNNs meet global covariance pooling: Better representation and generalization, IEEE transactions on pattern analysis and machine intelligence (2020).
[15] Y. Li, N. Wang, J. Liu, X. Hou, Demystifying neural style transfer, arXiv preprint arXiv:1701.01036 (2017).
[16] C. Ionescu, O. Vantzos, C. Sminchisescu, Matrix backpropagation for deep networks with structured layers, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2965–2973.
[17] P.-A. Absil, R. Mahony, R. Sepulchre, Optimization algorithms on matrix manifolds, Princeton University Press, 2009.
[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.
[19] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[20] C. G. Goetz, W. Poewe, O. Rascol, et al, Movement disorder society task force report on the Hoehn and Yahr staging scale: status and recommendations the movement disorder society task force on rating scales for Parkinson's disease, Movement disorders 19 (9) (2004) 1020–1028.
[21] I. Salazar, S. Pertuz, W. Contreras, F. Martínez, A convolutional oculomotor representation to model Parkinsonian fixational patterns from magnified videos, Pattern Analysis and Applications 24 (2) (2021) 445–457.