A MULTIMODAL MODEL TO QUANTIFY AND CHARACTERIZE NEUROMOTION PATTERNS RELATED WITH PARKINSON DISEASE

JOHN EDINSON ARCHILA VALDERRAMA

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
PROGRAMA DE DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN
BUCARAMANGA
2025

A MULTIMODAL MODEL TO QUANTIFY AND CHARACTERIZE NEUROMOTION PATTERNS RELATED WITH PARKINSON DISEASE

JOHN EDINSON ARCHILA VALDERRAMA

Trabajo de Grado para optar al título de:
Doctor en Ciencias de la Computación

Director:
Fabio Martínez Carrillo
Doctor en Ingeniería de Sistemas y Computación

Codirector:
Antoine Manzanera
Ph.D spécialité Signal et Image

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2025

ACKNOWLEDGEMENTS

I want to express my deepest gratitude to Professor Fabio Martínez. Thanks to his guidance, I was able to strengthen skills that are valuable both in research and in life, such as perseverance, resilience, learning from failure, goal planning, and mental toughness. So many hours of meetings, so many academic and experiential discussions, so much advice from Professor Fabio are giving their first results in my academic path.

I want to express my deepest gratitude to Professor Antoine Manzanera. His constant support, scientific expertise, and kindness were fundamental to the initiation, development, and completion of this work.

I would also like to express my gratitude to Dr. Ivan Peña, a neurologist who has always been willing to evaluate and characterize patients.

I would like to thank the Fundación del Parkinson y del Adulto Mayor (FAMPAS) for their unwavering support, always being available to assist us by providing a space in their facilities for patient recordings.

I would like to thank Professor Paula for providing us with a recording space at the Physiotherapy School. I would also like to express our gratitude to Ana, a master's student in Physiotherapy, for her unwavering support during the recordings.

I would like to thank the research group Biomedical Imaging, Vision and Learning Laboratory ($BivL^2ab$) for their invaluable support during the recording sessions and seminars. These activities beautifully combined elements of academia and family. A special thanks to Juan Olmos, who was always willing to offer his help and support.

I would like to thank my colleagues at the high school who occasionally covered my classes for a few minutes while I completed virtual meetings with my research advisors.

I would like to express my heartfelt gratitude to my family, who were always there to offer words of encouragement and a helping hand. Special thanks to my mom, dad, and my brother for their unwavering support.

I would like to thank God for granting me the strength to begin and complete this project. During challenging moments, I found solace and inspiration in His words.
I dedicate this work to my daughter Eimy Elizabeth. I love her to the stars, and she loves me to the kittens (she adores kittens!).

ABBREVIATIONS

Bi	Bilinear
BiLSTM	Bidirectional Long Short Term Memory Network
BiMap	Bilinear Mapping
BiRe	Bilinear Rectification
CART	Classification and Regression Trees
CNN	Convolutional Neural Networks
DF	Deep Features
GBT	Gradient Boosted Tree
H&Y	Hoehn and Yahr
KF	Kinematic Features
KS	Kolmogorov Smirnov
LogEig	Logarithm maps
MCC	Mattews Correlation Coefficient
MDS-UPDRS	Movement Disorder Society Unified Parkinson's Disease Rating Scale
PCA	Principal Component Analysis
PD	Parkinson's Disease
PIGD	Postural Instability Gait Difficulty
PK	Parkinson Subjects
R	Rigidity
ReEig	Rectification Eigenvalues
RGB	Red Green Blue
RM	Riemannian Mean
SAS	Simpson Angus Scale
SNJs	Square Wave Jerks

SPD	Symmetric Positive Definitive
SVM	Support Vector Machine
TD	Tremor Dominant
UPDRS-ME	Unified Parkinson's Disease Rating Scale Motor Examination
VGG	Visual Geometry Group
VOG	Video Oculography Technique

CONTENTS

page

LIST OF FIGURES

page

LIST OF TABLES

page

RESUMEN

TÍTULO: Un modelo multimodal para cuantificar y caracterizar patrones neuromotores relacionado con la enfermedad de Parkinson [*]

AUTOR: John Edinson Archila Valderrama [**]

PALABRAS CLAVE: Parkinson, red multimodal riemanniana, clasificación multimodal, representación de extremo a extremo.

DESCRIPCIÓN: La enfermedad de Parkinson es el segundo trastorno neurológico más común, asociado con una deficiencia de dopamina que afecta principalmente la función motora. La evaluación clínica actual depende en gran medida de la experiencia de los especialistas, lo que conduce a diagnósticos subjetivos que pueden limitar tratamientos tempranos y efectivos. A pesar de las propuestas computacionales para apoyar el diagnóstico, el problema de investigación sigue abierto debido a la alta variabilidad de los síntomas, el conjunto limitado de observaciones y el uso de modalidades individuales. Además, la caracterización y estratificación de las observaciones relacionadas con el Parkinson aún presentan desafíos significativos. En esta tesis, se propuso un enfoque geométrico multimodal para apoyar el diagnóstico a partir de observaciones oculomotoras y de marcha, proporcionando una clasificación multimodal de pacientes con Parkinson. Siguiendo una lógica progresiva, en una primera aproximación, las observaciones en video se codificaron como matrices de covarianza y se resumieron utilizando la media geométrica Riemanniana como descriptor de video. Este enfoque fue evaluado con 26 pacientes (78 videos), reportando un 96% de precisión para fusiones temprana y tardía. A continuación, se diseñó una segunda aproximación en forma de una red multimodal Riemanniana capaz de aprender patrones de segundo orden, alcanzando un 96% de precisión para fusiones temprana e intermedia, y un 92% para fusión tardía en una cohorte de 32 pacientes (512 videos), superando a los enfoques unimodales. En tercer lugar, se propuso una representación geométrica multimodal Riemanniana de extremo a extremo para clasificar afectaciones motoras clave, proporcionando apoyo a las escalas de estratificación de la enfermedad. Este método se evaluó en una cohorte de 32 pacientes (512 videos), demostrando capacidades para clasificar bradicinesia ocular (93% de precisión), bradicinesia durante la marcha (90% de

---

[*] Trabajo de investigación

[**] Facultad de Ingenierías Fisicomecánicas Escuela de Ingeniería de Sistemas e Informática. Director: Fabio Martínez Carrillo, Ph.D. Codirector: Antoine Manzanera, Ph.D.

precisión), freezing de la marcha (83% de precisión) y otros ítems de las escalas.

El trabajo desarrollado demostró que construir y aprender descriptores geométricos multimodales puede ser efectivo en escenarios con datos limitados, contribuyendo al apoyo diagnóstico de la enfermedad de Parkinson. Además, los hallazgos destacan la necesidad de enfoques multimodales para abordar la naturaleza multifactorial y la alta variabilidad sintomática de la enfermedad.

ABSTRACT

TITLE: A multimodal model to quantify and characterize neuromotion patterns related with Parkinson disease. *

AUTHOR: John Edinson Archila Valderrama

**

KEYWORDS: Parkinson, Riemannian multimodal network, multimodal classification, end-to-end representation.

DESCRIPTION: Parkinson's disease is the second most common neurological disorder, which is associated with dopamine deficiency, primarily affecting motor function. Current clinical evaluation highly depends on the expertise of specialists leading to subjective diagnosis that may limit early and effective treatments. Despite computational proposals to support diagnosis, the research problem remains open due to the high variability of symptoms, the limited set of observations, and the use of single modalities. Furthermore, the characterization and stratification of Parkinson's disease observations still present significant challenges. In this thesis, a geometric multimodal approach for diagnosis support was proposed from oculomotor and gait observations, providing a multimodal classification of Parkinson's patients. Following a progressive logic, in a first approximation, video observations were encoded as covariance matrices, and summarized using a Riemannian geometric mean, as a video descriptor. This approach was evaluated with 26 patients (78 videos), reporting 96% accuracy for early and late fusion. Next, a second approximation was designed as a Riemannian multimodal network capable of learning second-order patterns, achieving 96% of accuracy for early and intermediate fusion, and 92% for late fusion in a cohort of 32 patients (512 videos), outperforming unimodal approaches. Thirdly, a geometric multimodal Riemannian end-to-end representation was proposed to classify key motor impairments, providing support for disease stratification scales. This method was evaluated in a cohort of 32 patients (512 videos), demonstrating capabilities to classify ocular bradykinesia (93% accuracy), bradykinesia during gait (90% accuracy), freezing of gait (83% accuracy), and other scale items. The developed work demonstrated that building and learning multimodal geometric descriptors can be effective in scenarios with limited data, contributing to

---

* Research work

** Faculty of Physics-Mechanics Engineering. School of Systems Engineering and Informatics. Advisor: Fabio Martínez, Ph.D. Co-advisor: Antoine Manzanera, Ph.D.

diagnostic support for Parkinson's disease. Additionally, the findings underscore the necessity of multimodal approaches to address the multifactorial nature and high symptomatic variability of the disease.

1. INTRODUCTION

1.1. PARKINSON'S DISEASE

Parkinson's disease (PD) is the second most common neurodegenerative disorder, affecting more than 6.1 million people, with an estimated incidence rate of 3%, particularly among individuals aged 80 and older [1, 2]. Today, PD has no cure and there is no definitive biomarker for early and effective diagnosis, which in consequence commonly leads to a late and ineffective treatment, representing a substantial public health challenge [2].

Regarding the etiology of PD, the disease is associated with a progressive loss of dopamine, a neurotransmitter essential for motor control and mood regulation [3]. Consequently, PD is expressed as progressive motor impairments, evolving according to dopamine deficit pathways that impact specific brain regions. Figure 1 illustrates a rough roadmap that relates the affected brain regions with the associated motor manifestations, as well as, the current protocols followed by specialists to characterize PD [4].

Firstly, dopamine deficit is accentuated in the substantia nigra producing very subtle impairments, difficult to associate with the disease, including sleep problems, loss of smell, and affected eye micromovements [2] (yellow region in Figure 1). These impairments are challenging to detect and quantify, so the current clinical scales and diagnostic methods generally

---

[1] Valery L FEIGIN; Emma NICHOLS, and et al ALAM Tahiya. "Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016". In: *The Lancet Neurology* 18.5 (2019), pp. 459–480. DOI: 10.1016/S1474-4422(18)30499-X.

[2] Tianbai LI and Weidong LE. "Biomarkers for Parkinson's disease: how good are they?" In: *Neuroscience bulletin* 36.2 (2020), pp. 183–194.

[3] Anthony E LANG. "The progression of Parkinson disease: a hypothesis". In: *Neurology* 68.12 (2007), pp. 948–952.

[4] Christopher H HAWKES; Kelly DEL TREDICI, and Heiko BRAAK. "A timeline for Parkinson's disease". In: *Parkinsonism & related disorders* 16.2 (2010), pp. 79–84.

fail to capture these primary disease symptoms. In a second phase of dopamine deficit, there exist evidence that principal affected regions are the temporal lobe, which generates tremor, rigidity, and/or akinesia, mainly unilaterally [5, 4] (green region in Figure 1). During this phase, there exist unilateral manifestation, which are key for specialist to observe early disease manifestations. From this stage, emerge coarse clinical scales to associate observed motor manifestations with the dopamine deficit, which originates from the substantia nigra, and affects the nucleus basalis of Meynert [4]. For instance, the Hoehn and Yahr (H&Y) scale categorizes from visual observations, motor impairments related to laterality during gait, allowing the coarse classification of disturbance levels between zero and five (strongest motor manifestations) [6]. Alternatively, the MDS-UPDRS part III scale includes the H&Y indexes, and completes it with intermediate levels coming from hand tremor or limb rigidity observations, facial expression, quantifying items observed by a specialist (from the minimum (0) to the severe (4) stage) [7, 8].

During PD evolution, in the third phase, dopamine deficiency affects the prefrontal cortex, manifesting as increased bilateral impairment in limb movement and posture [4, 9] (blue region in Figure 1). During such phase, could be observed loss in walking balance, could be

---

[5] P RIEDERER et al. "Lateralisation in Parkinson disease". In: *Cell and tissue research* 373 (2018), pp. 297–312.

[6] Christopher G GOETZ 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". In: *Movement disorders* 19.9 (2004), pp. 1020–1028.

[7] Christopher G GOETZ et al. "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results". In: *Movement disorders: official journal of the Movement Disorder Society* 23.15 (2008), pp. 2129–2170.

[8] CG GOETZ et al. "MDS-Unified Parkinson's Disease Rating Scale (MDS-UPDRS)". in: *Available from the International Parkinson and Movement Disorder Society website: https://www. movementdisorders. org/MDS/MDS-Rating-Scales/MDS-Unified-Parkinsons-Disease-Rating-Scale-MDS-UPDRS. htm* (2008).

[9] Moria DAGAN et al. "The role of the prefrontal cortex in freezing of gait in Parkinson's disease: insights from a deep repetitive transcranial magnetic stimulation exploratory study". In: *Experimental brain research* 235 (2017), pp. 2463–2472.

observed loss in walking balance, following H&Y scale, and the MDS-UPDRS part III scale, which in turn complements the analysis with tremors in hands, facial expressions, and voice impairment. At more advanced level, during the fourth phase, the primary motor and sensory areas are affected (red area), showing postural instability, falls, dependence, and cognitive impairment (for example, speech impairment) [4, 10]. In this stage, H&Y scale corresponds to 4 and 5 level affectation. MDS-UPDRS part III is associated with specific observations of postural instability, bradykinesia, and rigidity.

It should be noted that the manifestation of several of these symptoms appears after 20 years, making diagnosis less precise [4]. Despite a coarse explanation of the disease pathways, the PD report has variability in terms of intensity, onset phenotypes, presence / absence, and progression of these symptoms, being a multifactorial disorder without sufficient understanding [11, 6, 7]. Besides, motor scales have allowed standardize observations but the analysis and the evaluation and patient stratification highly depends on the expertise of the professional which could be prone to errors [12, 13]. In consequence with multifactorial disease nature, and evident limitations on current analysis, there is demanding more sensitive, multi-modal diagnostic tests with enhanced resolution to detect motor changes across disease stages, particularly for facilitating the development of neuroprotective therapies, which

---

[10] Jacob J CROUSE et al. "Postural instability and falls in Parkinson's disease". In: *Reviews in the Neurosciences* 27.5 (2016), pp. 549–555.

[11] Eduardo TOLOSA et al. "Challenges in the diagnosis of Parkinson's disease". In: *The Lancet Neurology* 20.5 (2021), pp. 385–397.

[12] Bart POST et al. "Unified Parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable?" In: *Movement disorders : official journal of the Movement Disorder Society* 20.12 (2005), 1577—1584. DOI: 10.1002/mds.20640.

[13] "Metric properties of nurses' ratings of parkinsonian signs with a modified Unified Parkinson's Disease Rating Scale". In: *Neurology* 49.6 (1997), pp. 1580–1587. DOI: 10.1212/WNL.49.6.1580. eprint: https://n.neurology.org/content/49/6/1580.full.pdf.

remains an unmet necessity in Parkinson's disease management [14]. Furthermore, the scale lacks an analysis of how each modality contributes to the final score, which hinders personalized diagnosis. This section highlighted the challenges associated with the diagnosis and multifactorial nature of the disease. The remainder of the chapter is structured as follows: Section 1.2 describes the most dominant motor impairments related to gait and eye movement. It also presents the computational methods proposed to classify unimodal and multimodal approaches, emphasizing the features and challenges involved in Parkinson's classification. Section 1.3 outlines the research problem, proposes the general objective that guides the study, and formulates the research question. Section 1.4 discusses the challenges associated with multimodal Parkinson's classification, as well as the contributions made in this research. Additionally, it describes the academic outputs produced throughout this journey. Finally, Section 1.5 summarizes the contents of each of the following chapters and the appendices of the book, providing a global overview of the thesis.

1.2. RELATED WORKS

Parkinson's disease causes several motor and non-motor manifestations, at different stages, that could be recorded from different sensors, and affect different human functionalities [15]. These manifestations allow to represent and quantitatively evaluate the disease progression or the effect of a particular treatment. The most common and standarized PD motor disabilities are principally observed from walking. However, such observations lack of sufficient sensitivity to early disease detection, and also to precisely characterize and score the

---

[14] Shen-Yang LIM and Ai Huey TAN. "Historical perspective: the pros and cons of conventional outcome measures in Parkinson's disease". In: *Parkinsonism & related disorders* 46 (2018), S47–S52.

[15] JONES RACHEL. "Biomarkers: casting the net wide". In: *Nature* 466.7310 (2010), S11–S12. DOI: https://doi.org/10.1038/466S11a.

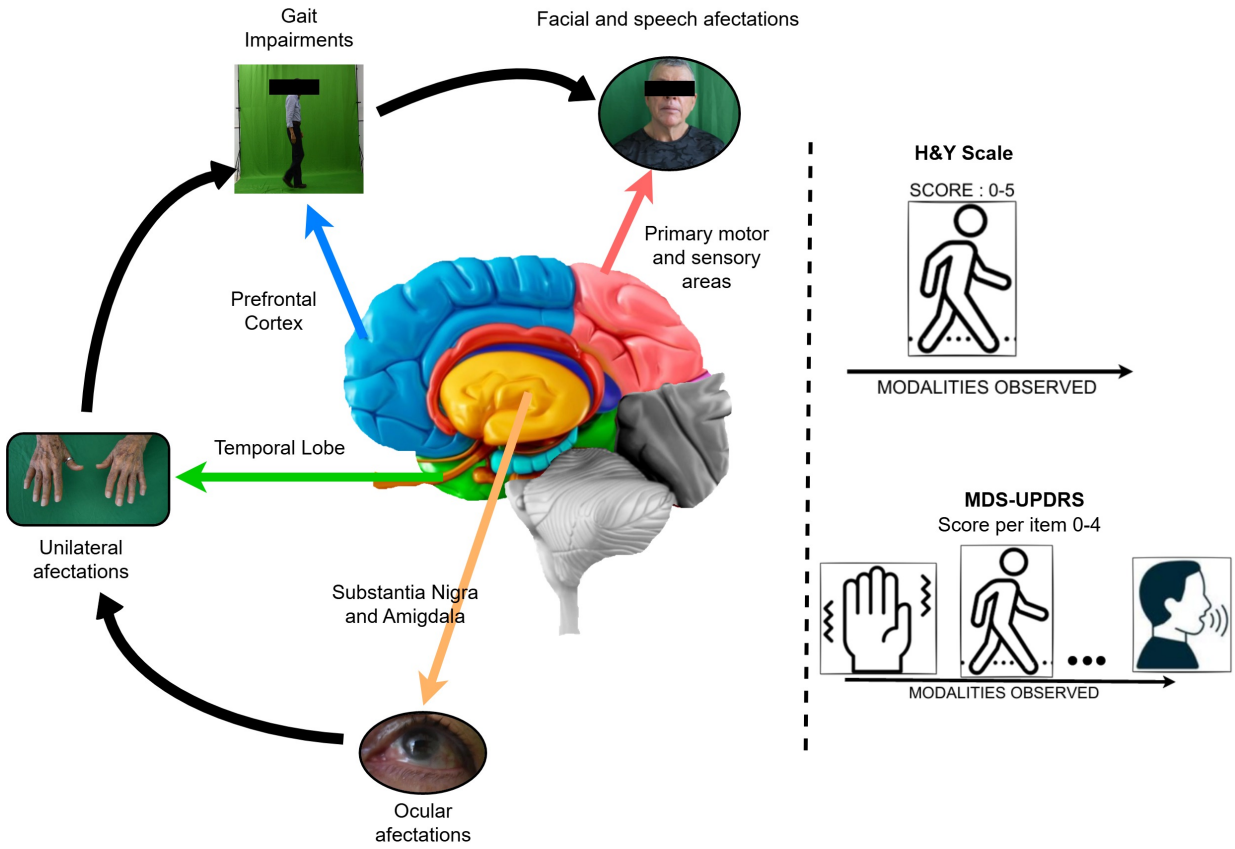

Figure 1. Tentative explanatory pathway to link brain regions affected by dopamine deficiency with the primary motor manifestations observed in patients, along with insights gathered from various sources of clinical information (modalities) based on measurement scales.

disease progression [16]. Complementary, ocular motions have been suggested as potential biomarkers of PD disease, allowing for differentiation between healthy people and patients in the early or intermediate stages of the disease [17]. Different studies have shown that ocular motions can be significantly altered by a deficit of dopamine, which makes them promising

[16] Michela TINELLI; Panos KANAVOS, and Federico GRIMACCIA. "The value of early diagnosis and treatment in Parkinson's disease: a literature review of the potential clinical and socioeconomic impact of targeting unmet needs in Parkinson's disease". In: (2016).

[17] George T GITCHEL et al. "Experimental support that ocular tremor in Parkinson's disease does not originate from head movement". In: *Parkinsonism & related disorders* 20.7 (2014), pp. 743–747.

descriptors of earlier PD stages [18, 19].

Due to the multifactorial nature of the disease, the integration of key modalities during movement enables better performance in diagnosis prediction. In this work, the computation of both motor modalities (walking and ocular motion) is carried out from a non-invasive perspective, allowing to capture natural patient movements. In the next subsections, we briefly describe each modality and current state-of-the-art strategies used to capture and process the related motion patterns.

1.2.1. Gait analysis. The most common motor disabilities have been quantified from walking observations [15, 20]. Particularly, in PD patients, gait is affected by patterns related to postural instability, tremors, rigidity, and bradykinesia [21]. These symptoms appear in the intermediate and late stages. For gait characterization, in the literature, alternatives using inertial (IMU) [22] or force sensors located on the feet soles [23] have been proposed. These classical alternatives produce temporal signals dedicated principally to lower limb analysis

---

[18] Merel S. EKKER et al. "Ocular and visual disorders in Parkinson's disease: Common but frequently overlooked". In: *Parkinsonism & Related Disorders* 40 (2017), pp. 1–10. DOI: 10.1016/j.parkreldis.2017.02.014.

[19] Panagiota TSITSI et al. "Fixation duration and pupil size as diagnostic tools in Parkinson's disease". In: *Journal of Parkinson's Disease* 11.2 (2021), pp. 865–875.

[20] Anat MIRELMAN et al. "Gait impairments in Parkinson's disease". In: *The Lancet Neurology* 18.7 (2019), pp. 697–708. DOI: 10.1016/S1474-4422(19)30044-4.

[21] Ana Beatriz Ramalho Leite SILVA et al. "Premotor, Nonmotor And Motor Symptoms Of Parkinson's Disease: A New Clinical State Of The Art". In: *Ageing Research Reviews* (2022), p. 101834.

[22] Elham RASTEGARI; Sasan AZIZIAN, and Hesham ALI. "Machine Learning and Similarity Network Approaches to Support Automatic Classification of Parkinson's Diseases Using Accelerometer-based Gait Analysis". In: *Proceedings of the 52nd Hawaii International Conference on System Sciences*. 2019.

[23] Lazzaro di BIASE et al. "Parkinson's disease wearable gait analysis: kinematic and dynamic markers for diagnosis". In: *Sensors* 22.22 (2022), p. 8773.

[24]. Nonetheless, these approaches may be intrusive, alter the character of the gait, and lack sensitivity. Other approaches based on IMU sensors can involve multiple kinematic variables to consider more complex models that include multiple parts of the body, but such an approach remains intrusive and requires sophisticated calibration systems [25].

Alternatively, the use of common cameras greatly simplifies the clinical protocol, considerably reduces the discomfort of patients, and potentially improves the evaluation of abnormal patterns related to PD [26, 27]. For instance, video analysis strategies are used to extract the skeleton of a person to approximate kinematic joint analysis from locomotion[28]. The naturality of movement is not affected, but this simplification can limit the characterization of motion patterns. As a more global alternative, a 3D convolutional neural network was introduced to classify and highlight salient regions of the limbs and other parts of the body[29]. However, these saliency maps only provide qualitative information and still need to be correlated with the quantitative values of stages for a more comprehensive interpretation of the disease.

Other approaches have specialized in characterizing specific gait impairments such as: gait

---

[24] Qinghui WANG; Wei ZENG, and Xiangkun DAI. "Gait classification for early detection and severity rating of Parkinson's disease based on hybrid signal processing and machine learning methods". In: *Cognitive Neurodynamics* (2022), pp. 1–24.

[25] Dante TRABASSI et al. "Machine learning approach to support the detection of Parkinson's disease in IMU-based gait analysis". In: *Sensors* 22.10 (2022), p. 3700.

[26] Rachneet KAUR et al. "A Vision-Based Framework for Predicting Multiple Sclerosis and Parkinson's Disease Gait Dysfunctions—A Deep Learning Approach". In: *IEEE Journal of Biomedical and Health Informatics* 27.1 (2022), pp. 190–201.

[27] Peipei LIU et al. "Quantitative assessment of gait characteristics in patients with Parkinson's disease using 2D video". In: *Parkinsonism & Related Disorders* 101 (2022), pp. 49–56.

[28] Mohamed CHERIET et al. "Multi-speed transformer network for neurodegenerative disease assessment and activity recognition". In: *Computer Methods and Programs in Biomedicine* 230 (2023), p. 107344.

[29] Luis C GUAYACÁN and Fabio MARTÍNEZ. "Visualising and quantifying relevant parkinsonian gait patterns using 3D convolutional network". In: *Journal of biomedical informatics* 123 (2021), p. 103935.

independence following a specific item (3.10) of the MDS-UPDRS Part III [30, 31]. These methods compute landmarks that, from 2D [30] and 3D poses [31], together with kinematic variables, allow for the stratification of the disease stage (normal, slight, mild, and moderate levels). Some studies have also considered the stratification of gait across multiple scales to estimate distance and angle calculations between limbs and use regressions to generate a report based on the gait items of the UPDRS and the SAS scale [32]. However, stratification on the basis of solely the gait modality may hinder the prediction of patients in early stages, where gait-related motor impairments are difficult to perceive.

1.2.2. Oculomotor patterns. In recent studies, oculomotor patterns have been established as potential Parkinson's disease biomarkers, reporting a strong correlation with dopamine deficiency, which makes them candidates to support early detection and diagnosis of the disease, even in the absence of additional motor symptoms [33, 18]. Besides, ocular motions have allowed the differentiation between healthy people and patients in the early or intermediate stages of the disease [17].

To characterize such patterns, different experiments were proposed in the literature, allowing to measure the eye capability response and the control of eye movement [34]. For instance,

---

[30] Samuel RUPPRECHTER et al. "A clinically interpretable computer-vision based method for quantifying gait in parkinson's disease". In: *Sensors* 21.16 (2021), p. 5437.

[31] Mandy LU et al. "Vision-based estimation of MDS-UPDRS gait scores for assessing Parkinson's disease motor severity". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer. 2020, pp. 637–647.

[32] Andrea SABO et al. "Estimating parkinsonism severity in natural gait videos of older adults with dementia". In: *IEEE journal of biomedical and health informatics* 26.5 (2022), pp. 2288–2298.

[33] Han LI et al. "Abnormal eye movements in Parkinson's disease: From experimental study to clinical application". In: *Parkinsonism & Related Disorders* (2023), p. 105791.

[34] Andrzej W PRZYBYSZEWSKI et al. "Multimodal learning and intelligent prediction of symptom development in individual Parkinson's patients". In: *Sensors* 16.9 (2016), p. 1498.

an experiment has been carried out to evaluate the ocular fixation, *i.e.,* the ability to stabilize the gaze at a given point. Actually, it was found that in control patients, eyes register small involuntary movements called microsaccades at intervals of 1 to 2 Hz, while for Parkinson patients, the fundamental frequency of movements is around 5.7 Hz [35]. Hence, these kinds of eye patterns could be determinant to characterize PD patterns even in very early stages (including asymptomatic patients) [36].

Also, in smooth pursuit tasks, alterations in PD are linked to foveal motion and the capacity to follow linear trajectories and changes in direction [37]. For PD patients, a decrease in tracking speed and an increase in the frequency of involuntary movements have been observed [37, 38]. Despite the importance of eye movement characterization, scales used in clinical routines such as the modified H&Y and MDS-UPDRS part III do not consider this type of movement. This may be due to the current sophisticated setup used to capture such movements and the difficulty establishing a consensus between specialists to define a scale of affectation from these observations[39, 40].

To reduce the complexity of calibration and facilitate use in routine clinical practice, computer

---

[35] George T GITCHEL; Paul A WETZEL, and Mark S BARON. "Pervasive ocular tremor in patients with Parkinson disease". In: *Archives of neurology* 69.8 (2012), pp. 1011–1017.

[36] A.J. LARRAZABAL; C.E. GARCÍA CENA, and C.E. MARTÍNEZ. "Video-oculography eye tracking towards clinical applications: A review". In: *Computers in Biology and Medicine* 108 (2019), pp. 57–66. DOI: 10.1016/j.compbiomed.2019.03.025.

[37] Karen FREI. "Abnormalities of smooth pursuit in Parkinson's disease: A systematic review". In: *Clinical parkinsonism & related disorders* 4 (2021), p. 100085.

[38] RA ARMSTRONG. "Oculo-visual dysfunction in Parkinson's disease". In: *Journal of Parkinson's disease* 5.4 (2015), pp. 715–726.

[39] Zheng ZENG et al. "A robust gaze estimation approach via exploring relevant electrooculogram features and optimal electrodes placements". In: *IEEE Journal of Translational Engineering in Health and Medicine* (2023).

[40] Oliver BREDEMEYER et al. "Oculomotor deficits in Parkinson's disease: Increasing sensitivity using multivariate approaches". In: *Frontiers in Digital Health* 4 (2022), p. 939677.

vision-based methods have been proposed to analize and code oculomotor patterns [41, 42, 43]. For example, kinematic features such as amplitudes and reaction times have been quantified. Then, using an ensemble of classifiers, the overall diagnosis is predicted through binary classification [41]. Alternatively, kinematic features are quantified to calculate the statistical significance between stages [42] or for multistage classification on the MDS-UPDRS scale [43]. These studies have evidenced the importance of oculomotor patterns, complenting scales and other observations, but today, there is not a protocol that include these measures and descriptors.

1.2.3. Multimodal approaches on PD. Considering the variability on PD symptoms, the multifactorial nature of the disease, and the challenges on the characterizations, there exist multiple efforts that have proposed strategies that integrates different PD observations. Multimodal computational approaches consider several symptoms in patients, enabling the ability

---

[41] Donald C BRIEN et al. “Classification and staging of Parkinson’s disease using video-based eye tracking”. In: *Parkinsonism & Related Disorders* 110 (2023), p. 105316.

[42] Johnathan REINER et al. “Oculometric measures as a tool for assessment of clinical symptoms and severity of Parkinson’s disease”. In: *Journal of Neural Transmission* 130.10 (2023), pp. 1241–1248.

[43] Nils A KOCH et al. “Eye movement function captured via an electronic tablet informs on cognition and disease severity in Parkinson’s disease”. In: *Scientific Reports* 14.1 (2024), p. 9082.

to support the diagnosis between control subjects and Parkinson's patients [44, 45, 46, 47]. For example, voice frequency and handwriting features have been classified independently using ensemble classifiers. The output diagnosis is considered Parkinson if at least one modality predicts the disease [44]. However, voice and handwriting modalities may primarily appear in the intermediate stages of the disease, being limited for patients who predominantly experience gait disturbances rather than tremors. Other approaches have considered other motion modes to leverage the complementarity of different motor impairments, such as gait, speech, and tapping modalities through the use of sensors such as gyroscopes and accelerometers. These signals have been used to train deep convolutional models and achieve discrimination between patients and controls subjects [45].

Recently, gait and eye movement modalities have been coded from an infrared eye tracker and an inertial gait device to estimate statistically significant kinematic variables for each modality. To differentiate motor patterns between PD patients and control subjects, a multivariate logistic regression model was set using the resulting kinematics associated with saccade velocity, maximum trunk sway, and average turn angular velocity [46]. In general, previous approaches have carried out binary classification, which can be insufficient to support the clinical and personalized characterization of PD.

Alternatively, some multimodal studies group patients into a stage of the disease according

---

[44] Hung N PHAM et al. "Multimodal detection of Parkinson disease based on vocal and improved spiral test". In: *2019 International Conference on System Science and Engineering (ICSSE)*. IEEE. 2019, pp. 279–284.

[45] John PRINCE; Fernando ANDREOTTI, and Maarten DE VOS. "Multi-source ensemble learning for the remote prediction of Parkinson's disease in the presence of source-wise missing data". In: *IEEE Transactions on Biomedical Engineering* 66.5 (2018), pp. 1402–1411.

[46] Han LI et al. "Combined diagnosis for Parkinson's disease via gait and eye movement disorders". In: *Parkinsonism & Related Disorders* 123 (2024), p. 106979.

[47] John ARCHILA; Antoine MANZANERA, and Fabio MARTINEZ. "A multimodal Parkinson quantification by fusing eye and gait motion patterns, using covariance descriptors, from non-invasive computer vision". In: *Computer Methods and Programs in Biomedicine* (2021), p. 106607.

to an observation scale, providing an estimate of its progression level. Recent works have proposed stratified predictions via common rating scales [48, 49]. For example, voice and gait have been analyzed using microphones, gyroscopes, and accelerometers. Graphs were constructed from these signals for further classification via a random forest. This classification identifies levels as low, intermediate, and severe [48]. Alternatively, handwriting, gait, and speech have been measured by accelerometers, gyroscopes, and tablets. These signals are subsequently represented by frequencial features and have been used to train deep convolutional models to classify control subjects from early, intermediate and advanced stages [49]. Although stratification provides more specialized information than binary prediction does, it still lacks more relevant diagnostic details. Therefore, identifying and characterizing PD symptoms, such as bradykinesia in patients is vital, as it is a core symptom of the disease. Additionally, other motor impairments, such as posture and gait disturbances, are crucial in characterizing the diagnosis [50]. Nonetheless, current approaches only consider global classification score without additional output related to motion items.

1.2.4. Covariance in Multiple Contexts. The use of covariance matrices as second-order representations with symmetric positive-definite (SPD) properties has become an efficient alternative for capturing complex internal relationships within data. This approach has proven particularly effective in domains where data exhibit intrinsic non-Euclidean structures,

---

[48] K Deepa RAJ et al. "A Visibility Graph Approach for Multi-stage Classification of Parkinson's Disease Using Multimodal Data". In: *IEEE Access (2024)* (2024).

[49] Juan Camilo VÁSQUEZ-CORREA et al. "Multimodal assessment of Parkinson's disease: a deep learning approach". In: *IEEE journal of biomedical and health informatics* 23.4 (2018), pp. 1618–1630.

[50] Stefano CAPRONI and Carlo COLOSIMO. "Diagnosis and differential diagnosis of Parkinson disease". In: *clinics in geriatric medicine* 36.1 (2020), pp. 13–24.

such as satellite image classification [51], machine learning applied to acoustic signals [52], brain signal analysis [53, 54], emotion [55], facial [56], and motor intent recognition [57].

Information encoded through SPD matrices can be represented in a compact and robust manner, offering reduced computational cost [58] and high discriminative capacity [52, 53]. For classification tasks using machine learning or deep neural networks, various methods project these matrices from the Riemannian manifold to tangent spaces, where conventional classification algorithms can be applied [51, 52, 53]. However, in recent years, new architectures have emerged that operate directly on Riemannian manifolds by learning geometric representations and applying exponential and logarithmic operations [59]. Architectures such as SPDNet represent a promising research direction in the classification of structured data. Their main

---

[51] Sara AKODAD et al. “Ensemble learning approaches based on covariance pooling of CNN features for high resolution remote sensing scene classification”. In: *Remote Sensing* 12.20 (2020), p. 3292.

[52] Jie-Ru HUANG; Mei-Chen LIU, and Jin-Min ZHOU. “Piano Soundboard Classification Based on Intelligent Neural Network and Multi-feature Fusion Algorithm”. In: (2025).

[53] Abdulhamit SUBASI and Saeed MIAN QAISAR. “EEG-based emotion recognition using modified covariance and ensemble classifiers”. In: *Journal of Ambient Intelligence and Humanized Computing* 15.1 (2024), pp. 575–591.

[54] Alexandre BARACHANT et al. “Riemannian geometry applied to BCI classification”. In: *International conference on latent variable analysis and signal separation*. Springer. 2010, pp. 629–636.

[55] Xianhua ZENG et al. “Deep hybrid manifold for image set classification”. In: *Image and Vision Computing* 143 (2024), p. 104935.

[56] Ruiping WANG et al. “Covariance discriminative learning: A natural and efficient approach to image set classification”. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 2496–2503.

[57] Rongrong FU et al. “Dynamical differential covariance based brain network for motor intent recognition”. In: *IEEE Sensors Journal* 24.5 (2024), pp. 6515–6522.

[58] Oncel TUZEL; Fatih PORIKLI, and Peter MEER. “Region covariance: A fast descriptor for detection and classification”. In: *European conference on computer vision*. Springer. 2006, pp. 589–600.

[59] Zhiwu HUANG et al. “A Riemannian network for SPD matrix learning”. In: *Thirty-first AAAI conference on artificial intelligence*. 2017.

advantages lie in their ability to preserve complex statistical relationships and their compatibility with deep learning models adapted to non-Euclidean manifolds. This perspective has proven especially valuable in domains characterized by limited data environments, such as biomedical applications for example prostate cancer pattern detection [60] and other uses such as emotion recognition [55].

1.3. THESIS STATEMENT

1.3.1. The Research problem. Until today, there is no cure for Parkinson disease, and epidemiology studies report a dramatic geographic expansion [1]. Even worse, there is no definitive PD biomarker that allows an early diagnosis and a proper monitoring of the disease. In fact, the current diagnosis and characterization of PD is principally based on the observation and quantification of motor disabilities.

Currently, the standard system to stratify the disease is the MDS-UPDRS part III (Movement Disorder Society Unified PD Rating Scale), that includes more than 15 motion observation items and ranges the PD in five grades, from normal (no impairment) to severe stage (disability). Despite this disease stratification scale, the motor task analysis is dependent of physicians observations, which are biased by their experience, and this affects the inter-rater reliability of the scoring scales [61].

Besides, there are no definitive models that determine the integration of such observation to output a specific scale of the disease. This challenge is inherent to the disease because PD is multi-factorial, with several phenotypes (different disease manifestations) that result in a high variability in time occurrence of the symptoms, and also in the specific motor impairments.

---

[60] Edward SANDOVAL; Juan OLMOS, and Fabio MARTÍNEZ. “RIEMAE: Riemannian Masked Autoencoder for Classifying Malignant Prostate Cancer Patterns”. In: *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2025, pp. 1–4.

[61] Renee M HENDRICKS; Mohammad T KHASAWNEH, et al. “An investigation into the use and meaning of Parkinson’s disease clinical scale scores”. In: *Parkinson’s Disease* 2021 (2021).

Also, the scale is restricted to observed alterations, which may miss prodromal patterns or slight motor abnormalities that could support early diagnosis. Current analysis is also limited in terms of sensitivity to measure progression of the motor impairments, which may impact the assessment of patients and the systematic quantification of disease progression. In fact, several studies have reported patients with different motor affections that are labeled at the same PD level, limiting the personalized treatment and follow-up.

In modern medical diagnostics, computer-aided systems now incorporate advanced machine-learning techniques. These approaches often analyze sensor data from various modalities, such as gait analysis, vocal assessments, and eye movement tracking, to derive clinical insights. Moreover, innovative methods using markerless technology, and deep learning have proven promising in distinguishing between control subjects and PD patients. However, these strategies, typically supervised, require extensive labeled datasets to accurately identify discriminative patterns, a significant challenge in the medical field [62].

Nonetheless, much of these approaches operate only for binary discrimination, using single modalities. This fact may limit the global characterization of Parkinson population since the patient may develop motor impairments at different corporal segments and with very different manifestation such as rigidity during walking (slow movement) or hand tremor (fast and uncontrolled movement). This limitation may hinder the study of different phenotypes of the disease. The proposal of multimodal strategies is then key to personalize PD characterization, better represent the motor impairments and robustly follow the progression and effects of particular treatments. Such integration is nonetheless challenging because of the multidimensional nature of the different symptom observation sources. Also, the available data to

---

[62] Minja BELIĆ et al. “Artificial intelligence for assisting diagnostics and assessment of Parkinson’s disease—A review”. In: *Clinical neurology and neurosurgery* 184 (2019), p. 105442.

learn representations are very rare [29, 63] and take a long time to obtain, particularly considering the multifactorial nature of the disease, and that much of the motor impairments can be shared by different disease levels. The modelling of motor patterns recorded in videos may be carried out from current deep learning strategies. However such strategies generally demand a huge amount of data information to deal with action variability. For instance, for natural image purposes, the classification task is carried out over 1.2 million of images in training and 9 million in training [64], while the action recognition architectures are trained with thousands of videos (DeepMind Kinetics have more than $650\,000$ videos clips) [65]. For such reason, new learning representations are needed, that may take advantage of data geometry to deal with variability while being trained on reasonably limited datasets.

Following these problem statements, we propose to address the following research question: How to learn a multimodal Parkinsonian representation to characterize and quantify motor impairments associated to the disease, with the capability to discriminate among different stages, and which can be trained with a limited amount of data?

1.3.2. General objective. To propose a multimodal computational representation to quantify, characterize, and classify Parkinsonian motion patterns.

1.4. PARKINSON CHALLENGES AND THESIS CONTRIBUTIONS

Today, Parkinson analysis and characterization involve multiple challenges related to disease variability and heterogeneity, expressed in different onset symptoms, region affectations, and

---

[63] Ali SAAD et al. "A preliminary study of the causality of freezing of gait for Parkinson's disease patients: Bayesian belief network approach". In: *International Journal of Computer Science Issues* 10.3 (2013), pp. 88–95.

[64] Jia DENG et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[65] Will KAY et al. "The kinetics human action video dataset". In: *arXiv preprint arXiv:1705.06950* (2017).

of course with different speed-up severity conditions. Additionally, acquisition setups and proposed modelling of the disease are limited to cover the complex spectrum of the disease. Hence, technically, there exist also limitations on data acquisition, discovering on PD relationships, coding of descriptors, and the design and evaluation of strategies to effectively support current diagnosis protocols. In this thesis, we cover some of these challenges and partially contribute in topics related to the quantification and coding of motor alterations to support Parkinson evaluation, diagnosis, and follow-up. We now summarize some of the tackled challenges and achieved academic products, with the following contributions:

- **Covariances representation to encode motion patterns.**

  Encoding multimodal motor patterns presents a significant challenge due to the specific characteristics of each type of movement, such as temporal resolution, spatial resolution, and movement granularity. Furthermore, integrating these sources into a unified geometric representation poses an additional challenge, as it requires identifying key features that preserve both spatial and temporal structures while reducing the high dimensionality of video sequences. In this thesis, we partially contributed to this challenge by introducing a novel geometric data representation framework designed to encode motion patterns through the use of covariance matrices. These matrices lie on the Riemannian manifold of symmetric positive definite (SPD) matrices, which provides a mathematically principled space capable of capturing the nonlinear relationships and dynamic dependencies inherent to complex motor behaviors. Such a covariance modelling was established according to three proposals:

  – Covariances coded from Optical flow kinematic features. In such a case, the multimodal covariances are built from such kinematics, capturing interpretable aspects of motor dynamics. The methodology and the validation setup is reported in Chapter 2, Appendix A, and Appendix B. In Chapter 2 is introduced a general framework of covariance classification using gait and ocular fixation patterns.

In Appendix A, the covariance descriptors are included into a recurrent architecture for time-interval classification. In Appendix B is reported how these kinematic covariances are tested in two complementary modalities: speech and face motion.

  - Pre-trained deep features, obtained from convolutional neural networks trained on large-scale image datasets, which encode intermediate-level semantic representations of motion. In such a case the covariance descriptor is built from a bank of learned deep features to increase representation dimensionality. In Chapter 3 is reported the methodological framework of this pre-trained deep covariances and its impact in the classification of PD.
  - Manifold-learned features, where deep models are specifically designed or adapted to operate within the Riemannian domain, enabling end-to-end training that preserves the manifold structure throughout the learning process. In Chapter 4 is reported the end-to-end architecture that allows to learn the best discriminative features, which are thereafter encoded in SPD descriptors to conduct the classification. It should be noted that prediction in such a case is carried out for different items of standard PD motion scales.

By encoding these features into SPD matrices, we achieve compact and discriminative representations that are robust to noise and inter-subject variability. Furthermore, our method was evaluated across multiple motor modalities, primarily eye movement and gait. However, as shown in Appendix B, distinctive results were also obtained when analyzing facial expression and voice. In this regard, the proposed descriptors are versatile in classifying different sources of movement through the use of symmetric positive definite matrices.

- **Multimodal architectures to model Parkinsonian findings.**

  The wide spectrum of symptoms to characterize and support diagnosis of Parkinson

poses a challenge and demands multimodal representation capable of processing data from different modalities. Such a challenge requires the implementation of fusion techniques, ensuring discrimination between the various stages of the disease.

Specific contributions in this subject are:

  - The development of novel multimodal geometric strategies to encode motor impairments to support classification between Parkinsonian and control subjects (see Chapter 3 and Appendix B).
  - These representations evidenced discriminative capability for binary classification and also were validated regarding the capacity to differentiate between various stages of disease progression (see Appendix A).
  - The geometric multimodal representation was also extended to support simultaneous motor impairments, which follow current clinical assessment guidelines (see Chapter 4). In this context, we fused gait impairments and ocular patterns to support a standardized multi-item protocols to enhance interpretability and clinical applicability.

- **Validation and Clinical Parkinson support.**

  Parkinson is a multifactorial disease that, depending on phenotypic nature, may have significant variations regarding motor impairments, scales of affectation, and stages in which they appear. Today, there is not enough available data about motor Parkinson observations. Hence, a persistent challenge in the analysis and evaluation of proposed computational strategies is the clinical validation under real-world conditions. Current proposals are scarce, with protocols dedicated to one modality, and most of the time, their use are restricted for other researchers. In such a line, this document contributes in the following points:

  - During this thesis was acquired a video dataset with motor impairments related to gait and eye movement, accompanied by labeled information such as stratification

according to the H&Y scale, four gait-related observations based on the MDS-UPDRS scale, and an eye movement assessment. Also, the dataset includes demographic variables and the medical treatment received by each patient. The dataset was developed progressively through the inclusion of incremental patient cohorts, as follows:

* Model introduced in Chapter 2 was trained and validated from 26 participants with three videos recorded for each: one for gait and two for eye movements.
* Model in Chapter 3 was trained and validated from 32 participants with a total of 16 videos recorded per subject, eight for gait and eight for eye movements.
* Additionally, for Chapter 4, the information of 32 patients was updated with the characterization of motor impairments, integrating both standardized clinical scales and non-standardized neurological observations. This information allowed a comprehensive validation of the proposed geometric multimodal representations, focusing on their capacity not only to predict the overall presence of Parkinson's disease but also to identify specific motor impairments such as gait disturbances, ocular bradykinesia, freezing episodes, and postural instability.

A complete description of the constructed dataset is reported in Appendix C. This dataset, that was also built as part of the international ECOS-Nord Project C23MS01 "Parkinson", was approved by an ethical committee.

1.4.1. Academic Products. This subsection reports academic products obtained around the thesis work. It is included journal and conference papers, as well as collaborations with other researchers, demonstrating a significant impact in the fields of biomedical engineering and artificial intelligence. Additionally, we received recognition through various awards for our academic contributions, underscoring the significance of our research in both the academic community and clinical practice.

Journal Papers

- Archila, J., Manzanera, A., & Martínez, F. (2022). A multimodal Parkinson quantification by fusing eye and gait motion patterns, using covariance descriptors, from non-invasive computer vision. Computer Methods and Programs in Biomedicine, 215, 106607. https://doi.org/10.1016/j.cmpb.2021.106607. **Status: published**.

- Archila, J., Manzanera, A. & Martínez, F (2025). A Riemannian multimodal representation to classify parkinsonism-related patterns from noninvasive observations of gait and eye movements. Biomed. Eng. Lett. 15, 81–93 https://doi.org/10.1007/s13534-024-00420-0 **Status: published**.

- Archila, J., Peña, I., Celis, L., Olmos, J., Manzanera, A., & Martínez, F. (2025). A multimodal gait and ocular geometric representation to generate a Parkinson progression report. Engineering Applications of Artificial Intelligence, 160, 111834. https://doi.org/10.1016/j.engappai.2025.111834 **Status: published**.

Conference Papers

- Archila, J., Manzanera, A., & Martínez, F. (2021). A recurrent approach for predicting Parkinson stage from multimodal videos. In 17th International Symposium on Medical Information Processing and Analysis (Vol. 12088, pp. 37-45). SPIE. https://doi.org/10.1117/12.2606293 **Status: published**.

- Archila, J., Manzanera, A., & Martínez, F. (2024). A Mixed audio-video SPD network for online classification of Parkinsonian speech patterns. 18th Ibero-American Conference on Artificial Intelligence. https://doi.org/10.1007/978-3-031-80366-6_10 **Status: published**.

Collaborations in Papers

- Valenzuela, B., Archila, J., Arevalo, J., Contreras, W., & Martinez, F. (2025). Integrating synchronized dysarthria and hypomimia deep patterns to quantify Parkinson disease. International Journal of Information Technology, 1-9. https://doi.org/10.1007/s41870-025-02647-1 **Status: published**.

- Cadena, J. P., Valderrama, J. E. A., Sierra-Jerez, F., Tarazona, A. M., & Carrillo, F. M. (2024). Hand Tremor Characterization from a Spatiotemporal Convolutional Representation. Ingeniería, 29(3), e21091-e21091. https://doi.org/10.14483/23448393.21091 **Status: published**.

- Ruano, J., Arcila, J., Romo-Bucheli, D., Vargas, C., Rodríguez, J., Mendoza, Ó., ... & Martínez, F. (2022). Deep learning representations to support COVID-19 diagnosis on CT slices. Biomédica, 42(1), 170-183.

  https://doi.org/10.1007/s13534-024-00420-0 **Status: published**.

Awards

- **Best Student Paper Award** (2021, December). A recurrent approach for predicting Parkinson stage from multimodal videos. In 17th International Symposium on Medical Information Processing and Analysis (Vol. 12088, pp. 37-45). SPIE.

- **Eloy Valenzuela Award** (2024, April) for best research work in the area of physical-mechanical engineering. For his participation in the work "Telerehabilitation: an alternative for the diagnosis of locomotor disorders in hard-to-reach areas". Awarded by the Industrial University of Santander.

1.5. THESIS OUTLINE

In this section, we present a graphical abstract that summarizes thesis contribution and facilitates the analysis of the evolution of research using different approaches based on the construction of symmetric positive definite (SPD) descriptors and the learning of geometric

representations. These approaches propose classification tasks between control subjects and patients with Parkinson's disease, such as general diagnosis, stage prediction according to the H&Y scale, prediction of motor impairments based on standardized scales, and non-standardized specialist observations. Below, we provide details of the contributions and findings from each chapter and appendix.

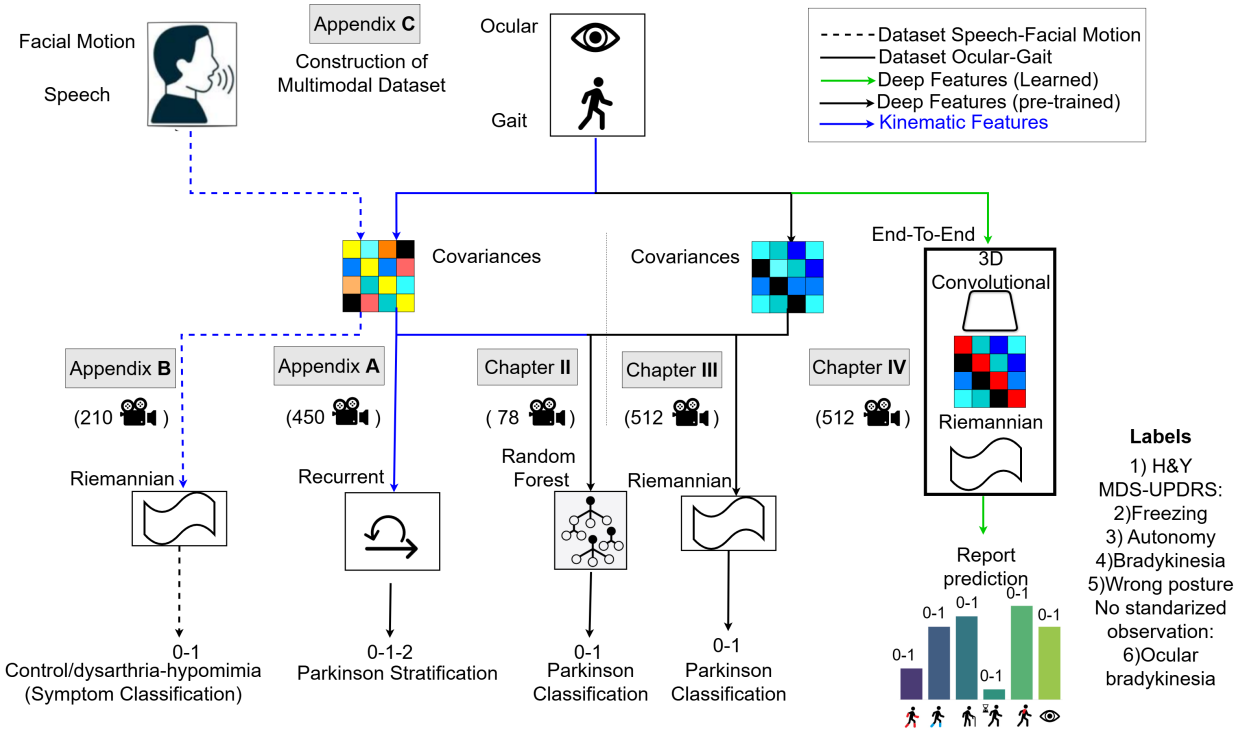

Figure 2. Graphical Abstract: The contributions and findings of each chapter and appendix of this research are presented.

Chapter 2 Covariance as a compact descriptor for classification. This chapter introduced a novel approach that captures movement abnormalities associated with PD from different sources, and assists in the disease quantification. The gait and eye fixation movement patterns are herein recorded and analyzed from video descriptors to characterize the disease. Each video sequence, at each modality, is represented by frame-covariance matrices that summarize responses of deep and kinematic features. These frame-covariances form

a video manifold that codes motion pattern modalities in a compact representation. Then, a geometrical Riemmanian mean is computed as a video descriptor. The same method is applied to both modalities to facilitate their interpretation, merging for any kind of (deep or kinematic) features. Such multimodal video descriptors are projected to a tangent plane to allow them to undergo linear operations. Then, the descriptor is mapped to a supervised machine learning strategy to obtain a PD classification. The proposed approach was validated from early (at the level of covariance descriptors) and late (at the level of output probabilities) integration of both modalities to understand better the capability of discrimination of the proposed video descriptors.

Chapter 3 Learning new representations based on covariance matrices. In this study, we presented a multimodal motion representation to capture, in a non-invasive manner, multifactorial motion impairments, geometrically characterize patterns associated with PD, and subsequently discriminate them with respect to a control population. In particular, sequences of large-scale movements such as gait, and small-scale movements such as smooth ocular motion are encoded using second-order relationships of their deep features through frame-level covariance. Then, the mean of the covariance matrices that compose the video, is calculated on the Riemannian manifold. Thus, this geometric representation is computed as a global descriptor for a video chunk. Next, geometrical deep learning is employed to refine a Riemannian descriptor, effectively capturing second-order patterns for Parkinson-based classification. This multimodal learning representation is investigated through various fusion methods that integrate localized impairments (ocular movement) and global movement impairments (gait). The assessed fusion methods are as follows: early fusion, which involves integrating the Riemannian means of each modality within the Riemannian manifold; intermediate fusion, which combines the Riemannian means mapped to Euclidean space and integrates them through a dense layer; and late fusion, which reduces to a linear weighting of the output probabilities through the learning of each modality.

Chapter 4 Quantifying motor affectations from end-to-end geometrical approach. This chapter introduces a multimodal geometric representation of gait and eye movement video sequences, that learns new representations from the Riemannian manifold to generate a motor report displaying predictions such as: bradykinesia affecting gait and ocular motions, lateral affectation of gait, autonomy during gait, freezing of gait and wrong posture. The videos are processed to compute 3D convolutional deep features, which are aggregated into covariance matrices that encode the relationships among these video features. Each covariance matrix represents a global descriptor of the video. Geometric deep learning techniques are then applied to refine and produce new symmetric positive definitive (SPD) representations, effectively capturing second-order patterns for classifying six Parkinson's disease motor impairments. This study explores multimodal learning representations through early and intermediate fusion methods, demonstrating how symmetric positive definitive (SPD) matrices can be effectively fused to encode Parkinsonian patterns across different impairments related with balance, mobility and key symptoms such as bradykinesia.

Appendix A A recurrent approach for predicting Parkinson. In section is presented a multimodal motion approach that codifies temporal patterns from gait and eyes motion, to characterize the stage of PD in the patients. The proposed approach achieves a markerless per-frame video analysis to extract temporal neuromotor patterns associated with PD, that identifies the most critical motion sequences in the patient. The spatial information from the video is compactly encoded as covariance matrices of deep features calculated from optical flow maps, which thereafter feed a recurrent net to learn temporal motion dependency of this covariance sequences. Finally, a prediction performed over time on video slices stratifies Parkinson's patients into different stages. This work opens perspectives toward new tools in motor therapies of patients, following the objective of identifying which movements and gestures are more significant of each PD stage.

Appendix B Exploring geometrical representations with other modalities. In this section was introduced a geometrical online learning method to support Parkinson classification considering multimodal sources (audio and video). Thus, characterizing dysarthria and hypomimia, the proposed approach use a set of video landmarks that, together with fundamental frequencies, form a compact covariance descriptor. From this second-order representation, geometrical learning is herein implemented to learn covariation patterns associated to the disease at different temporal intervals.

Appendix C Data and Parkinson patient characterization. The appendix describes the main characteristics considered during the construction of the dataset. Initially, a demographic and statistical analysis of the population and recorded patients were conducted. Subsequently, the protocol was developed with the assistance of a neurologist. Then, the recordings were carried out, and the neurologist's evaluations were included.

2. A MULTIMODAL PARKINSON QUANTIFICATION BY FUSING EYE AND GAIT MOTION PATTERNS, USING COVARIANCE DESCRIPTORS, FROM NON-INVASIVE COMPUTER VISION

2.1. ABSTRACT

Parkinson's disease (PD) is a motor neurodegenerative disease principally manifested by motor disabilities, such as postural instability, bradykinesia, tremor, and stiffness. In clinical practice, there exist several diagnostic rating scales that coarsely allow the measurement, characterization and classification of disease progression. These scales, however, are only based on strong changes in kinematic patterns, and the classification remains subjective, depending on the expertise of physicians. In addition, even for experts, disease analysis based on independent classical motor patterns lacks sufficient sensitivity to establish disease progression. Consequently, the disease diagnosis, stage, and progression could be affected by misinterpretations that lead to incorrect or inefficient treatment plans. This work introduces a multimodal non-invasive strategy based on video descriptors that integrate patterns from gait and eye fixation modalities to assist PD quantification and to support the diagnosis and follow-up of the patient. The multimodal representation is achieved from a compact covariance descriptor that characterizes postural and time changes of both information sources to improve disease classification. A multimodal approach is introduced as a computational method to capture movement abnormalities associated with PD. Two modalities (gait and eye fixation) are recorded in markerless video sequences. Then, each modality sequence is represented, at each frame, by primitive features composed of (1) kinematic measures extracted from a dense optical flow, and (2) deep features extracted from a convolutional network. The spatial distributions of these characteristics are compactly coded in covariance matrices, making it possible to map each particular dynamic in a Riemannian manifold. The temporal mean covariance is then computed and submitted to a supervised Random Forest

algorithm to obtain a disease prediction for a particular patient. The fusion of the covariance descriptors and eye movements integrating deep and kinematic features is evaluated to assess their contribution to disease quantification and prediction. In particular, in this study, the gait quantification is associated with typical patterns observed by the specialist, while ocular fixation, associated with early disease characterization, complements the analysis. In a study conducted with 13 control subjects and 13 PD patients, the fusion of gait and ocular fixation, integrating deep and kinematic features, achieved an average accuracy of 100% for early and late fusion. The classification probabilities show high confidence in the prediction diagnosis, the control subjects probabilities being lower than 0.27 with early fusion and 0.3 with late fusion, and those of the PD patients, being higher than 0.62 with early fusion and 0.51 with late fusion. Furthermore, it is observed that higher probability outputs are correlated with more advanced stages of the disease, according to the H&Y scale. A novel approach for fusing motion modalities captured in markerless video sequences was introduced. This multimodal integration had a remarkable discrimination performance in a study conducted with PD and control patients. The representation of compact covariance descriptors from kinematic and deep features suggests that the proposed strategy is a potential tool to support diagnosis and subsequent monitoring of the disease. During fusion it was observed that devoting major attention to eye fixational patterns may contribute to a better quantification of the disease, especially at stage 2.

The partial content of this work has been accepted and published in [47].

2.2. PROPOSED APPROACH

This work presents a novel multimodal methodology to capture and integrate movement abnormalities associated with Parkinson's disease by using as video representation a special Riemmanian manifold of temporal frame-covariance matrices.

The pipeline of the proposed approach is illustrated in figure 3. Firstly, videos that record particular modalities of interest are represented as a set of frame-covariance matrices. This

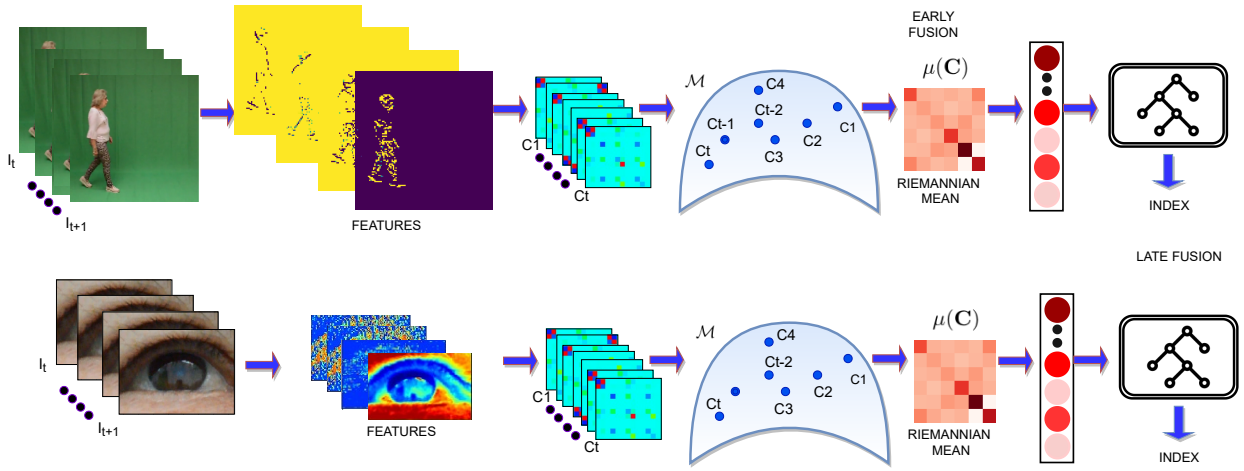

Figure 3. The pipeline of the proposed approach. Top: Gait. Bottom: Ocular fixation. In both modalities, we calculate the features for each frame along with the video, and then compute frame-level spatial covariance of the features. Finally, we summarize the information into a unique covariance matrix for the complete video (Riemann's mean). We propose two fusion approaches: early (concatenate the descriptors of each modality) and late (weight the probabilities of the two modalities to obtain a final probability).

set forms a special manifold that can be summarized by a Riemannian mean. The fusion of the two modalities can be performed at the video descriptor level (early fusion) or at the prediction level, after mapping in a supervised learning strategy as Random Forest.

2.2.1. Frame-level Representation. A markerless strategy is herein introduced by computing a spatially dense representation for each frame along the video sequence. The local representation refers to the set of features extracted for each frame I_t at time t, denoted F_t. The features in F_t aim at enhancing relevant motion and characteristics that could be discriminative to PD, allowing thereafter a proper coding of abnormal patterns. Each frame I_t is then represented by a set of N features. $F_t = \left\{f_{(1,t)}, f_{(2,t)}, \ldots, f_{(n,t)}\right\}$. In this work, two different schemes were evaluated to characterize each frame: kinematic features and deep features calculated using pre-trained networks. On the one hand, kinematic features are computed from a dense optical flow field. On the other hand, taking advantage of the expressivity of deep representations, each frame is processed by a filter bank composed of the

convolution kernels extracted from the first layers of a pre-trained convolutional network (see figure 4). The two next subsections detail these two sets of features.

2.2.2. Kinematic features from a Dense flow field. The optical flow is a 2d vector field that corresponds to the estimation of the apparent velocities of all pixels of the video between two consecutive frames. Such quantity is naturally relevant to characterize patients movement by computing kinematic local primitives on gait or eye fixation. To compute the optical flow, the video sequence is first pre-processed by computing a local entropy map to lower redundancy and enhance edges. The movement is then calculated using Farnebäck's method, [66] that uses a quadratic polynomial approximation of each pixel's neighborhood to estimate local velocity: $I_t(\mathbf{z}) \simeq \mathbf{z}^T \mathbf{A}_t \mathbf{z} + \mathbf{b}_t^T \mathbf{z} + c_t$, where $\mathbf{z} = (x, y)^T$ is the pixel position, and matrix $\mathbf{A}_t$, vector $\mathbf{b}_t$ and scalar c_t are estimated from the image I_t. The displacement vector $\mathbf{d}_t$ between I_t and I_{t+1} is then obtained as: $\mathbf{d}_t = -\frac{1}{2}\mathbf{A}_t^{-1}(\mathbf{b}_{t+1} - \mathbf{b}_t)$.

The obtained dense field $\mathbf{d}_t$ provides a rich and dense kinematic description of recorded video sequences. Hence, to characterize gait and eye motion patterns, a set of kinematic measures are extracted from the flow field. Specifically, we decompose the normalized velocity vector in unit tangential $\mathbf{T}_t$ and unit normal $\mathbf{N}_t$ components to obtain greater specificity of the velocity [67]. These two components allow determining a possible decrease in gait velocity in patients [68]. They also allow to determine the variation in eye movement focused on a fixed point, which could have a linear trend [69]. Similarly, scalar components of the acceleration are

---

[66] Gunnar FARNEBÄCK. "Two-frame motion estimation based on polynomial expansion". In: *Scandinavian conference on Image analysis*. Springer. 2003, pp. 363–370.

[67] Adel SALEH et al. "Exploiting the Kinematic of the Trajectories of the Local Descriptors to Improve Human Action Recognition." In: *VISIGRAPP (3: VISAPP)*. 2016, pp. 182–187.

[68] Shiho OKUDA et al. "Gait analysis of patients with Parkinson's disease using a portable triaxial accelerometer". In: *Neurology and Clinical Neuroscience* 4.3 (2016), pp. 93–97.

[69] Michele RUCCI and Martina POLETTI. "Control and functions of fixational eye movements". In: *Annual Review of Vision Science* 1 (2015), pp. 499–518.

calculated as the magnitudes of the tangential and normal components of the acceleration, respectively denoted a_t^T and a_t^N. These components determine the possible abrupt velocity changes due to the imbalances in the patient's gait [22]. Also, it is expected that the correlation between velocity and acceleration is lower in gait for patients with PD than in control subjects [68]. In the ocular fixation modality, the square wave jerks (SWJs) are inappropriate movements that occur through kinetic changes when the eye is dispersed from the fixed point. SWJs saccades have more frequency and magnitude in PD patients than control subjects [70]. In summary, kinematic features are encoded as the first order kinematics, corresponding to the horizontal and vertical components of the unit tangential vector $f_{(1,t)}$ and unit normal vector $f_{(2,t)}$ velocity, and as the second order kinematics, corresponding to the magnitude of tangential $f_{(3,t)}$ and normal acceleration $f_{(4,t)}$ [67]. The features based on the optical flow are illustrated on the left side of figure 2. Tangential and normal velocity is reflected in different parts of the body according to their components. The accelerations in $f_{(3,t)}$ and $f_{(4,t)}$ are less perceptible. However, they are visible on the feet and wrists.

2.2.3. Deep Features. The deep convolutional networks have recently demonstrated great capability to represent very complex visual patterns in classification and detection tasks, showing remarkable robustness to camera lens distortions, illumination changes or occlusions, among many others [71]. Nevertheless, proper end-to-end learning with these architectures requires a huge amount of data to carry out the training. Therefore, to achieve major flexibility modelling, we chose to represent each frame using learned features from pre-trained nets. These features are then computed from convolutional deep pre-trained nets as an alternative and complementary description of each frame. So the first layers learn a set

---

[70] Jorge OTERO-MILLAN et al. "Saccades during attempted fixation in parkinsonian disorders and recessive ataxia: from microsaccades to square-wave jerks". In: *PLoS One* 8.3 (2013), e58535.

[71] Samer HIJAZI; Rishi KUMAR, and Chris ROWEN. "Using convolutional neural networks for image recognition". In: *Cadence Design Systems Inc.: San Jose, CA, USA* (2015), pp. 1–12.

of kernel filters that provide a rich representation of images, including non linear relations achieved by activation functions. Specifically, these filters process an input image I_t from the video, that can be either an RGB frame (3 channels), or an optical flow map (2 channels), through a set of S learned convolution filters $\mathbf{\Phi} = \{\Phi_k\}_{1 \leq k \leq S}$ to form a set of S features $\mathbf{F_t^k} = a_k\left(I_t * \Phi_k\right)$, where a_k is a non linear activation function. For instance, for optical flow frames, the learned filters can compute acceleration related maps that may contribute to describe Parkinson disease patterns from video sequences.

Classically, the deep Convolutional Neural Networks (CNN) calculate features from layer to layer in such a way that, if the layer l has n_l neurons (i.e. calculates n_l features), and each neuron has $d_l \times d_l$ weights (i.e. calculates a $d_l \times d_l$ convolution), then the layer l actually computes n_l convolutions of size $d_l \times d_l \times n_{l-1}$, where n_{l-1} is the number of features (or channels) of the previous layer.

More recently, other CNN approaches have been proposed that perform separable convolution in depth, *i.e.* for each layer, n_{l-1} (depth-wise) convolutions of size $d_l \times d_l$ are first applied, and then, n_l (point-wise) convolutions of size n_{l-1} are applied. This produces the same number of features and the same data size, while reducing the computational cost by an order of magnitude [72, 73]. This decomposition allows less redundancy at the output compared to standard convolution [74]. Besides, the experiments on different data sets show that a network with separable layers requires fewer data to achieve similar or better performance compared to the dense-layer architecture [74].

In this work we evaluated both the conventional and the separable architectures (such as

---

[72] Andrew G HOWARD et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications". In: *arXiv preprint arXiv:1704.04861* (2017).

[73] Mark SANDLER et al. "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

[74] Laurent SIFRE and Stéphane MALLAT. "Rigid-motion scattering for image classification". In: *Ph. D. thesis* (2014).

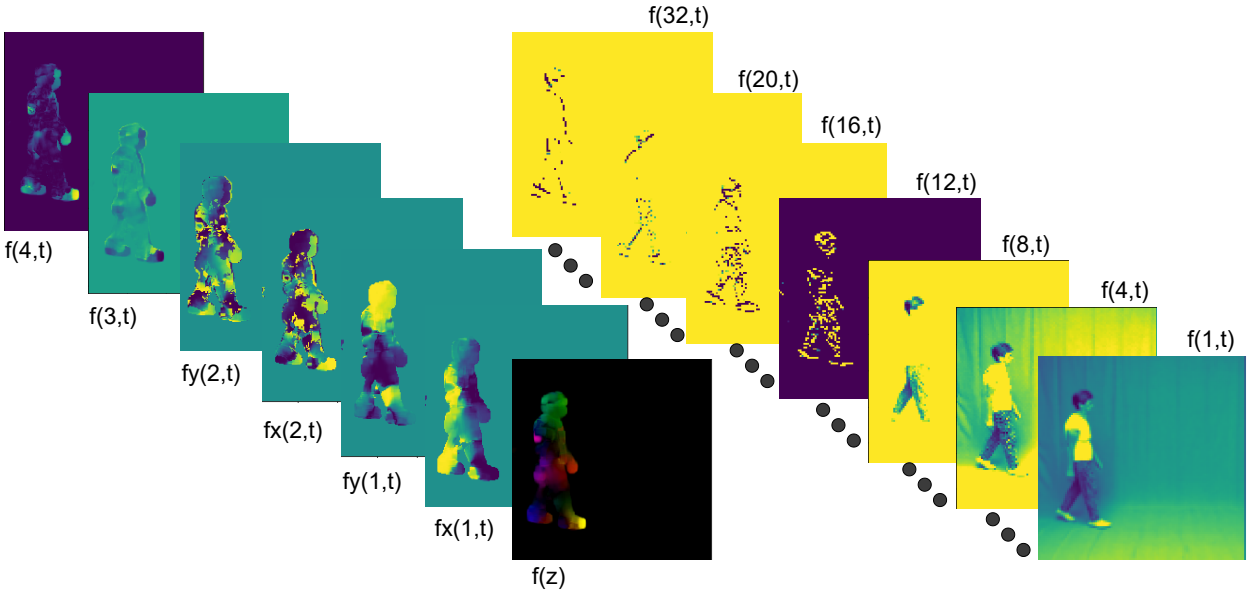

Figure 4. Left: Kinematic Features from the optical flow. The first order kinematics features are the horizontal and vertical components of the unit tangential $f_{(1,t)}$ and unit normal $f_{(2,t)}$ velocity. The second-order kinematics features are the magnitude of normal $f_{(3,t)}$ and tangential $f_{(4,t)}$ acceleration. Right: some of the 32 deep features coming from the fourth layer of MobileNet V2.

MobileNet). Figure 4 (right) shows feature maps extracted from the gait modality. This figure highlights the expression of the different features within different parts of the body, and the co-variation of the different features, which justifies the interest of a covariance based representation.

2.2.4. Riemaniann space of covariance descriptors. The integration of gait and eye motion patterns is expected to provide a more sensitive description of Parkinson's patterns and a better quantification of the disease. Because computed video descriptors are represented as covariance matrices, a natural fusion can be done for the different modalities. A compact integration can be achieved by aggregating gait and eye descriptors. In this work two levels, early and late fusion were evaluated. Considering the fact that Parkinson's disease has a typical unilateral involvement, we considered one covariance descriptor for each eye, and one covariance for the gait, to recover the whole motion spectrum to characterize

each patient. The description starts then by computing, for each frame t, a spatial covariance matrix C_t relative to the set of feature maps $F_t = \{f_{(1,t)}, \ldots, f_{(n,t)}\}$, where the n features can be the kinematic features, the deep features, or the union of all features. The covariance matrix is computed as:

$$C_t(i,j) = \mathbb{E}\left((f_{(i,t)} - \mathbb{E}(f_{(i,t)}))(f_{(j,t)} - \mathbb{E}(f_{(j,t)}))\right)$$

where the expectation $\mathbb{E}$ is calculated over the $W \times H$ points of each feature map $f_{(i,t)} \in \mathbb{R}^{W \times H}$, where W and H represent the width and height of the feature maps, respectively.
Then, a very compact representation is obtained for each frame, allowing to model complex patterns from a low temporal dimensional manifold: indeed, the covariance matrices lie on a half cone space, the actual dimension of the covariance matrix being given by $\dim(C) = \frac{n(n+1)}{2}$ where n is the number of features.
Such measures allow to summarize motion characteristics that may be typical of Parkinsonian patterns from a global (gait) and from a local (eye fixation) evaluation. Then, for a given video sequence, the frame feature maps are summarized as a sequence of spatial covariance matrices, represented as $\mathbf{C} = (C_1, C_2, C_3, \ldots, C_N)$. These symmetric positive matrices C_i are part of a non Euclidean space which is a Riemmanian manifold $\mathcal{M}$[75], and then Euclidean metrics is not suitable to compute temporal statistics on $\mathbf{C}$.
To make such measures, each covariance point should be projected to a tangent plane to the manifold (logarithmic operation). Accordingly, a projected covariance could be mapped to the original Riemannian manifold, which corresponds to the exponential operation. Particularly, the mean in $\mathcal{M}$ of a set of covariance matrices $\mathbf{C}$ can be iteratively found by optimization, where the mean μ is the point (covariance matrix) with minimum distance ρ among the sample covariance matrices [75]. Thus, the geometrical mean can be expressed as:

---

[75] P Thomas FLETCHER and Sarang JOSHI. "Riemannian geometry for the statistical analysis of diffusion tensor data". In: *Signal Processing* 87.2 (2007), pp. 250–262.

$$\mu_{t+1} = \exp_{\mu_t}\left(\frac{1}{k}\sum_{i=1}^{k}\log_{\mu_t}(C_i)\right)$$

where μ_0 is the initial guess and μ_{t+1} it the $(t+1)$ approximation of the geometric mean. This expression requires in each iteration the computation of the matrix function $\log_{\mu_t}$ and $\exp_{\mu_t}$, expressed as:

$$\begin{aligned}\log_{\mu_t}(C_i) &= {\mu_t}^{\frac{1}{2}}\log\left({\mu_t}^{-\frac{1}{2}}C_i{\mu_t}^{-\frac{1}{2}}\right){\mu_t}^{\frac{1}{2}}\\ \exp_{\mu_t}(C_i) &= {\mu_t}^{\frac{1}{2}}\exp\left({\mu_t}^{-\frac{1}{2}}C_i{\mu_t}^{-\frac{1}{2}}\right){\mu_t}^{\frac{1}{2}}\end{aligned} \tag{1}$$

where ${\mu_t}^{\frac{1}{2}} = \exp(\frac{1}{2}\log(\mu_t))$ and $\log(\mu_t) = \sum_t \log(\lambda_t)\Lambda_t^T$ (in the same way): $\exp(\mu_t) = \sum_t \exp(\lambda_t)\Lambda_t^T$), where Λ and λ are the eigenvectors and eigenvalues of the matrix μ, respectively. Here $\exp$ and $\log$ are the corresponding functions of matrices that extend the real exponential and logarithmic functions. As frame-covariance samples, the geometrical mean covariance has the dimension of $\mu_t \in \mathbb{R}^{d\times d}$, being symmetric ($\mu_t = {\mu_t}^\top$) and positive ($\det(\mu_t) > 0$). Therefore the final descriptor has the dimension of $\dim(\mu_t) = \frac{n(n+1)}{2}$ where n is the number of features. Finally, each video descriptor is defined as the Riemannian mean of frame-level covariance matrices. In this work, the implementation of the covariance mean has been computed as described in Algorithm 1.

[1] $\mathbf{C} = (C_1, C_2, C_3, \ldots, C_N)$ start with: $\mu_0 = C_1$ $X_k = \frac{1}{N}\sum_{i=1}^{N}\log_{\mu_k}(C_i)$ $\mu_{k+1} = \exp_{\mu_k}(X_k)$ $||X_k|| < \varepsilon$ $[\mu_{k+1}]$

This global statistic provides a compact representation that summarizes the main tendencies observed along the different phases of the action and naturally reduces noise or error artifacts that could suddenly appear in one frame. Furthermore, because the global descriptor ignores the temporal relations between the different frames, it is also invariant to the phase of the action.

2.2.5. Parkinson prediction using Covariance descriptors. The covariance mean, that represents a global measure for any video, can be used as a patient signature to quantify the level of Parkinson's disorder or to automatically classify between Parkinson and Control motion patterns. For this classification, a supervised strategy can be implemented to learn patterns from classes and to build a disease space, where new samples can be projected to be automatically labeled with a particular class. Nonetheless, these supervised algorithms generally operate under a Euclidean metric. To project each covariance into a Euclidean space, a logarithmic projection was carried out as $\log(C_i) = \Sigma \log(\lambda_i)\Sigma^T$, that defines a space with reference to the identity [75].

In this work was implemented the Random Forest as a supervised strategy due to the demonstrated effectiveness to represent very complex problems over discrete space, to address overfitting problems, and to be less sensitive to atypical data [76, 77]. Specifically in this work, a set of Riemannian mean descriptors of study subjects $C_1, C_2, ...C_i$ with the disease stage notations $y \in \{0, 1\}$ are used to learn boundaries between Parkinson and control. Then, the Random forest defines a set of decision trees, into a Bootstrap aggregating strategy through an ensemble learning, that allows to obtain multiple classifications and maximize the expected mean prediction. For so doing, the strategy randomly selects a set of covariance features to build different tree versions. Each tree is constructed using the (CART) technique based on a recursive procedure that seeks to obtain the division of the data that minimizes the label variance from each node [78]. The final prediction is made by averaging the predictions of the individual trees, as follows: $\hat{y} = \sum_{i=1}^{B} \frac{\theta_i}{B}$, where B is the number of trees.

---

[76] Jehad ALI et al. "Random forests and decision trees". In: *International Journal of Computer Science Issues (IJCSI)* 9.5 (2012), p. 272.

[77] Md Zahangir ALAM; M Saifur RAHMAN, and M Sohel RAHMAN. "A Random Forest based predictor for medical data classification using feature ranking". In: *Informatics in Medicine Unlocked* 15 (2019), p. 100180.

[78] Leo BREIMAN et al. *Classification and regression trees*. CRC press, 1984.

2.2.6. Fusion modalities. The different Parkinsonian observations can be fused once the videos are described by the set of frame covariances, forming the special sequence manifold. In this work, two different levels of fusion are proposed, evaluated over the two motions of interest, *i.e.,* gait and eye fixation, described as follows:

Early Fusion. In the early fusion approach, we propose a joint representation (J^e) of Riemann's descriptors: in ocular fixation (C^e) (each eye separately) and gait (C^g) per $i\text{-}patient$. $J_i^e = [C_i^{e_{left}}, C_i^{e_{right}}, C_i^g]$
This descriptor represents the covariation among selected features to represent videos (kinematic and/or deep features) in the different actions of the same person. The formed descriptor J_i is then used to build a classification space with the Random forest strategy. Then, during training, different tree versions can be formed by grouping different features from different modalities. Finally, those hybrid random trees are used for classifying an unknown subject from his eyes and gait sequences.

Late Fusion. A second fusion alternative proposed in this work was to learn independent classification spaces using each modality separately. In such case, each mean covariance, for gait $C_i^g \rightarrow RF^g$ and for eyes $(C_i^{e_{left}}, C_i^{e_{right}}) \rightarrow RF^e$ is used to learn independent modality trees, resulting in two different random forest models: (RF^g, RF^e).
In such case, each of the specialized random forest provides its own probability of disease. Then, we model the resulting probability P_f as a linear combination of the probabilities of each classifier by: $P_f = wP_g + (1-w)P_e$, where P_e and P_g are the ocular fixation and gait probabilities respectively, and w is a modality importance weight in the final disease prediction.

2.3. EXPERIMENTAL SETUP

2.3.1. Data. A total of 26 participants was included in this study: 13 control subjects (average age of 72.2 ± 6.1) and 13 PD patients (average age of 72.3 ± 7.4). The PD patients were diagnosed in the second or third stage of the disease by a physician using standard protocols of the Hoehn-Yahr scale. This study was approved by the Ethics Committee of Universidad Industrial de Santander and written informed consent was obtained. Regarding the motion modalities, the following protocols were applied:

- For eye fixational recording, the patients observed a fixed spotlight projected on a screen with a dark background, for an average duration of 6 seconds. The eye region was manually cropped (210×140 pixels) to obtain the sequences of interest.
- For gait, markerless sagittal-plane videos were recorded with a spatial resolution of 520×520 pixels and a temporal resolution of $60\ fps$. The locomotion was recorded along a 6 meter displacement, for an average duration of 6 seconds. For each participant, one video for gait and one video for each eye were recorded, resulting in a total dataset of 78 videos.

To evaluate the performance of the proposed approach a cross-validation leave-one-patient-out was implemented with the multimodal dataset. In such a scheme at each iteration, one patient is left out to test and the remaining ones (25 subjects in our particular experiment) are used for training. For these experiments, the Parkinsonian patients correctly classified were counted as true positive (TP) and the correct control patients were identified as true negative (TN). Then a set of metrics was used to fully understand the performance of the approach in its different configurations. The metrics herein implemented are the sensitiviy ($sen = \frac{TP}{TP+FN}$), specificity ($spec = \frac{TN}{FP+TN}$), accuracy ($acc = \frac{TP+TN}{TP+FP+FN+TN}$), precision ($prec = \frac{TP}{TP+FP}$), the F1-score ($F_1 = \frac{2 \times prec \times sen}{prec+sen}$) and Matthews correlation coefficient (MCC), defined as: $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}$. Also, the confusion matrices (TP, TN, FP, FN)

were calculated to assess the effectiveness of the classifier by giving the same weighting to each of the four groups [79, 80, 81].

2.3.2. Parameters tuning. The proposed approach was adjusted at different stages to optimize the representation w.r.t. description and quantification of Parkinsonian disease patterns. According to each stage, the following parameters were set:

- **Training Characteristics.** The evaluation was carried out within a leave-one-patient-out cross-validation scheme to mitigate the risk of overfitting. The hyperparameter were fixed for all experiments before the cross-validation scheme. In each iteration, the model was trained on the data of 25 out of the 26 participants and evaluated on the excluded subject. We consider this strategy ensured that the performance of each configuration was assessed on unseen data, preserving the independence between training and testing sets. The final configuration was selected based on the average performance across all folds. Consequently, the reported results reflect a robust evaluation framework that prevents information leakage and minimizes overfitting, despite the limited size of the dataset.

- **Kinematic features.** In both modalities the video sequences were processed with the Farnebäck optical flow with 5 scales and 3×3 window size, to obtain velocity fields at each frame. From each computed frame was then computed a total of 6 kinematics, namely the horizontal and vertical components of the tangential and normal unit velocity

---

[79] Davide CHICCO and Giuseppe JURMAN. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC genomics* 21.1 (2020), p. 6.

[80] Jinxin LIU; Burak KANTARCI, and Carlisle ADAMS. "Machine learning-driven intrusion detection for Contiki-NG-based IoT networks exposed to NSL-KDD dataset". In: *Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning*. 2020, pp. 25–30.

[81] Vytautas ABROMAVIČIUS et al. "Two-stage monitoring of patients in intensive care unit for sepsis prediction using non-overfitted machine learning models". In: *Electronics* 9.7 (2020), p. 1133.

vectors, along with the tangential and normal acceleration magnitudes.

- **Deep features.** The deep features were taken from the first layers output using two different pre-trained nets: the VGG16 (standard convolutions) and the MobileNet V2 (depth wise separate convolutions). For VGG net was selected the first to fourth deep layers, that count from 64 (kernel size of 3×3) to 128 (kernel size of 3×3) filters. Then, each activation output from this net has spatial size from (112×112) to (224×224). For MobileNetV2, the second to fifth layer were selected. Each layer counts a total of 32 filters and activation output with spatial size of (112×112).

- **Riemann descriptor.** A covariance mean is calculated for each video. Three descriptors are obtained for each patient, for each eye and for the gait independently. The resulting covariance descriptor, for each video, is formed by only 6 scalar motion features and/or 32 deep features. The descriptor fusing all modalities corresponds to a vector whose dimension can vary from 108 to $4\,332$.

- **Random Forest.** The classifier was trained using the bootstrap aggregating strategy with optimization metrics based on entropy criterion. A comprehensive evaluation of different sets of trees and numbers of samples per leaf in the Random Forest was carried out to get the best classification results in each modality and using different types of fusion.

2.4. RESULTS

The proposed approach was firstly evaluated with respect to the features capabilities to describe each motion mode, to find the best configuration to proceed with the further multimodal analysis. Also, two different versions of multimodal information fusion were evaluated. The next subsections summarize the reported results in each of the considered evaluation phases.

Features	Acc		Descriptor
	Gait	Eye	Size
DF-VGG/1st	0.923	0.807	4096
DF-VGG/2nd	**0.961**	0.807	4096
DF-VGG/3rd	0.923	**0.846**	4096
DF-VGG/4th	0.923	**0.846**	16384
DF-MobileNetV2/2nd	0.923	0.769	1024
DF-MobileNetV2/3rd	0.923	0.807	1024
DF-MobileNetV2/4th	**0.961**	**0.846**	1024
DF-MobileNetV2/5th	0.884	0.769	1024
KF-vel	0.576	0.730	16
KF-vel-acel	**0.923**	**0.923**	36

Table 1. Comparison of the accuracy obtained by each feature alone, in each modality

	Eye-DF	Eye-KF	Gait-DF	Gait-KF
sen	0.769	**1**	0.923	**1**
spec	**0.926**	0.846	**1**	0.846
prec	**0.909**	0.866	**1**	0.866
acc	0.846	**0.923**	**0.961**	0.923
F1-s	0.824	**0.928**	**0.959**	0.928
MCC	0.700	**0.856**	**0.925**	0.856

Table 2. Scores in ocular fixation (Eye) and gait modality using only deep features (DF) or only kinematic features (KF)

2.4.1. Feature evaluation In this work a frame-level representation from kinematic (velocity and/or acceleration, computed from a dense optical flow) and/or deep (using the first layer output from two different nets: VGG16 and/or MobileNetV2) features was considered. These features were evaluated independently over each motion mode to select the best configuration based on their capability in this task. Table 1 summarizes the individual performances of the different features.

	Eye-DF		Eye-KF		Gait-DF		Gait-KF	
	PK	C	PK	C	PK	C	PK	C
PK	10(76.9%)	3(23.1%)	13(100%)	0	12(92.3%)	1(7.7%)	13(100%)	0
C	1(7.7%)	12(92.3%)	2(15.4%)	11(84.6%)	0	13(100%)	2(15.4%)	11(84.6%)

Table 3. Confusion matrices per modality using only kinematic (KF) or deep features (DF), for Parkinson (PK) and Control (C) subjects.

The best PD prediction capability of deep features (DFs) has been reported for gait sequences, which may be associated with recovering particular postures during locomotion. The features extracted from the fourth layer of MobileNet resulted in the best features for the two motion modes, achieving an average accuracy of 0.96 ± 0.19 and 0.84 ± 0.36, for gait and eye fixation, respectively.

Regarding kinematic patterns, very compact covariance descriptors of 36 scalar values were obtained by integrating the velocity and acceleration patterns. Table 1 also illustrates the performance of such kinematic patterns in two different versions, using only velocities, and accelerations. For kinematics, the complete descriptor results in the best representation option in both modes, while remaining extremely compact in size.

Hence, from this study, the DFs computed from the fourth layer of MobileNet (DF-Mobilenet/4th) and the complete kinematic descriptor (KF-vel-acel) were selected. A more exhaustive evaluation was then carried out for these two configurations. The results are reported in Table 2. DFs achieved a remarkable performance for the different metrics considered. Regarding fixational eye patterns, the kinematic features KFs had higher MCC and accuracy scores, but DFs had a major specificity.

Table 3 displays the confusion matrices for each modality. In ocular fixation, the covariation of DFs presents a classification error of 23.1% in Parkinsonian patients and 7.7% in control subjects. However, the covariation of the kinematic characteristics of eye movements

	Early Fusion KF-DF	**Early Fusion DF**	**Early Fusion KF**
sen	1	0.923	0.846
spec	1	1	1
prec	1	1	1
acc	1	0.961	0.923
F1-s	1	0.959	0.916
MCC	1	0.925	0.856

Table 4. Early fusion scores for the two modalities, using kinematic (KF), deep (DF) or joint features (KF-DF)

reduces the classification error of patients to zero but doubles the classification error of control subjects. This fact may be associated to the capability of kinematic patterns to recover tiny micro-tremor in Parkinsonian patients but introducing artifacts in control subjects. The complementarity of the two types of feature is also observed in the gait modality, so that the DFs classified with zero error in the control subjects, but with 7.7% error in the PD patients. Similarly, the KFs reduce the error in the patient classification to zero but increase the classification error in 15.4% of control subjects. In fact, the kinematic locomotion of control subject results highly variable and therefore such representation may be unable to cover the whole spectrum of possible movements. In contrast, the Parkinsonian patients have locomotion signatures that can be properly recovered from kinematic descriptors but with increasing variability of postural configurations, representing a limitation for a deep feature representation.

	Late Fusion KF-DF	**Late Fusion DF**	**Late Fusion KF**
sen	1	0.923	0.846
spec	1	1	0.923
prec	1	1	0.916
acc	1	0.961	0.884
F1-s	1	0.959	0.879
MCC	1	0.925	0.771

Table 5. Late fusion scores for the two modalities, using kinematic (KF), deep (DF) or joint features (KF-DF).

	Early Fusion (KF-DF)		**Early Fusion (DF)**		**Early Fusion (KF)**		**Late Fusion (KF-DF)**		**Late Fusion (DF)**		**Late Fusion (KF)**	
	PK	C	PK	C	PK	C	PK	C	PK	C	PK	C
PK	13(100%)	0	12(92.3%)	1(7.7%)	11(84.6%)	2(15.4%)	13(100%)	0	12(92.3%)	1(7.7%)	11(84.6%)	2(15.4%)
C	0	13(100%)	0	13(100%)	0	13(100%)	0	13(100%)	0	13(100%)	1(7.7%)	12(92.3%)

Table 6. Confusion matrices for the different fusion modes, using kinematic (KF), deep (DF), or joint (KF-DF) features, for Parkinson (PK) and Control (C) subjects.

To better illustrate the discriminatory behavior of the motion descriptors constructed, a low-dimensional space was built from a projection of resultant covariance matrices using KFs ans DFs features. The first three components were plotted using principal component Analysis

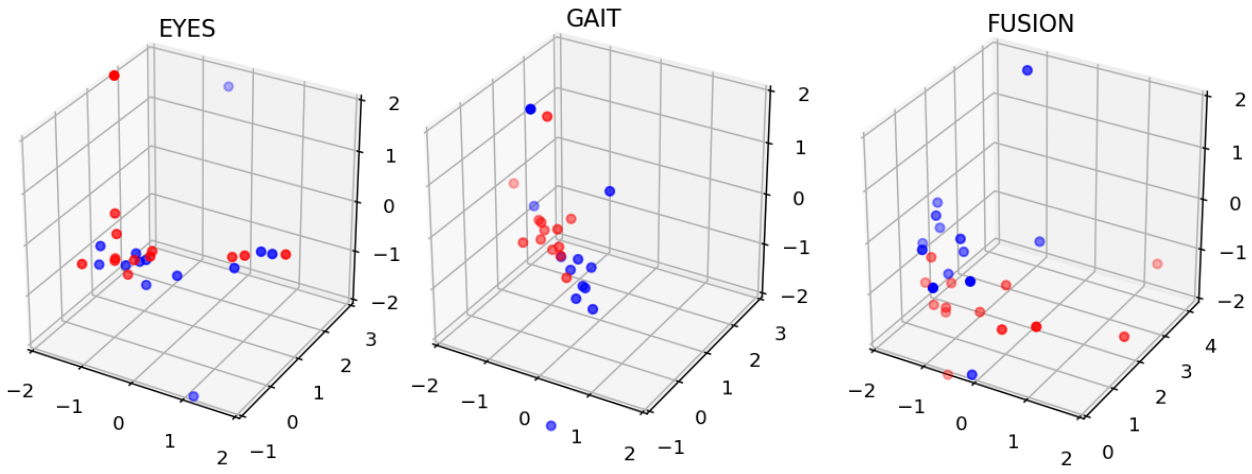

Figure 5. Projection over the three principal components of sample descriptors, for eyes, gait and fusion modalities. Red and blue points represent Parkinson and Control patients, respectively.

(PCA). Figure 5 illustrates the resultant projections in a 3D geometrical space for descriptors that correspond to eye, gait, and the fusion of both modalities. As expected, the low-dimensional representation of such geometrical means is useful for analyzing the grouping of points labeled with the same class. It appears that a discrimination rule can be easily implemented for the three descriptors.

2.4.2. Early fusion classification A first multimodal motion integration was achieved by concatenating (early fusion) the gait covariance descriptor and the eye fixation covariance descriptors (one for each eye). This fusion from mean covariance matrices allows a straightforward integration of the main features that characterize Parkinsonian patterns during the sequence. In this study, gait and eye integration was validated using KFs and DFs. Also, it was considered a descriptor that integrates both features (KF-DF). The KF covariances have a dimension of 6×6 (KF). The KF-DF covariances have a dimension of 38×38 (KF-DF), for each video descriptor of gait, eyes, and fusion.

Table 4 summarizes the scores achieved from this early integration using different sets of features. For all the studied subjects, the proposed approach achieved a perfect score by using

a mixed representation of kinematic and deep features. The postural representation obtained from DFs together with the kinematic description was sufficient to distinguish PD and control patients. Furthermore, covariance coding is a highly interpretable descriptor that can be used to recover salient information for each motion modality. Finally, the independent use of deep or kinematic features already achieves a good performance. Additionally, confidence intervals (95%) for specificity and sensitivity were calculated using the Clopper–Pearson method in the early fusion approach. For specificity, all three types of features (kinematic, pre-trained deep features, and their combination) yielded intervals of [0.75, 1]. For sensitivity, the confidence intervals were [0.54, 0.98] for kinematic features, [0.64, 0.99] for pre-trained deep features, and [0.75, 1] for the combined features. The narrowest intervals corresponded to the combined features. This result indicates that, based on the current sample, the value of the specificity and sensitivity would fall within the interval [0.75, 1].

A more detailed analysis was carried out by computing the confusion matrices for the different kinds of fusion (see in Table 6). Regarding independent sets of features, the deep features achieve a better integration with only one false negative. For KFs, two PD subjects were incorrectly classified as controls. This behavior could be associated with patients reporting small changes in gait because of the early stage of the disease.

A feature importance analysis was conducted to measure the contribution of each feature in the descriptor, ordered with respect to the impurity reduction at each split during the Random Forest training. In this experiment a full descriptor that included KFs and DFs was considered. The total descriptor dimension corresponds to the three covariance matrices (one for gait and two for eyes), i.e. $3 \times 38 \times 38 = 4\,332$. In Figure 6, the classification performance of the proposed descriptor was illustrated by selecting an incremental set of the most important features, according to the ranking performed by the Random Forest classifier. In such cases, using only 20% of the most important features, the proposed descriptor achieved an average score of 65%. From the observed results, we can hypothesize that each feature contributes approximately equally to the final prediction.

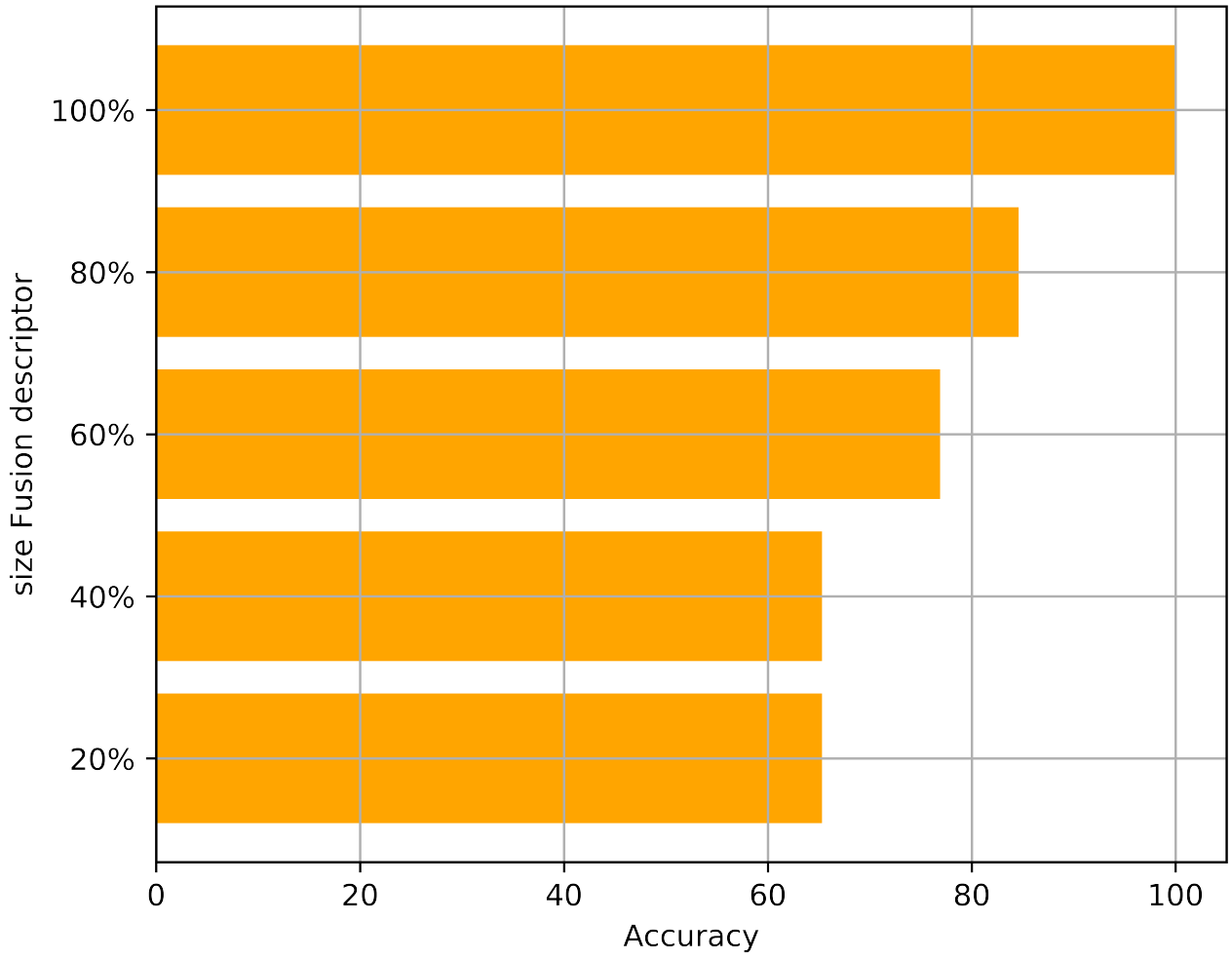

Figure 6. Feature importance analysis using Random forest with the general descriptor that includes kinematic and deep features. In the experiment was defined sets with the percentage of the most important features.

2.4.3. Late fusion classification The second multimodal integration proposed corresponds to building a discrete space classification for each modality independently, and then to fusing the probabilities obtained in each space. Each discrete space is obtained from a Random Forest, and each predicted sequence is projected onto the respective modality to obtain a corresponding probability of PD. The final probability P_f is the linear weighting of the ocular fixation probability P_e and gait probability P_g, as: $P_f = wP_g + (1-w)P_e$.

In the parameter study, the mode weight parameter w was varied from 0.1 to 0.9, and the best score was obtained by setting w to 0.4, which means 40% for the gait and 60% for the eye fixation modalities. This experiment is summarized in Figure 7 which plots the distributions of probability prediction for PD patients, for different values of w. For $w = 0.4$, the prediction for patients diagnosed with the disease has remarkable confidence and the outliers from the distribution reveal a probability prediction higher than 0.5. In contrast, $w < 0.4$ induces a significant variability in the probability prediction, with at least one example reported as false-negative (outlier point lower than 0.5). On the other hand, $w > 0.4$ increases the median of the predicted probability, which highlights the eye fixation contribution, but with false-labeled samples. Two outliers (that still obtain a positive prediction) with $w = 0.4$ correspond to patients in an early stage of the disease. This analysis highlighted the potential of eye fixational patterns as early PD biomarkers, while postural gait motion acts as a complementary cue, and can strengthen the disease analysis and quantification.

Table 5 summarizes the results of the proposed late fusion using different configurations of frame-level features to compute the Riemannian mean. This multimodal integration also results successful to discriminate between PD and control patients. In addition, the configuration of the classifier from a rich representation of both features is more effective than a single feature type to achieve perfect scores. However, we have calculated the confidence intervals (95%) for sensitivity using the Clopper–Pearson method for each group of features: for kinematic features [0.54, 0.98], for pre-trained deep features [0.64, 0.99], and for the combined features [0.75, 1]. Additionally, confidence intervals for specificity were also calculated for each type of feature: for kinematic features [0.64, 0.99], for pre-trained deep features [0.75, 1], and for the combined features [0.75, 1], suggesting that the narrowest intervals are associated with the combined features, and that the values of sensitivity and specificity metrics with this type of features are likely to fall within the range of 0.75 to 1.
A more detailed analysis can be obtained from the confusion matrices (see Table 6). The

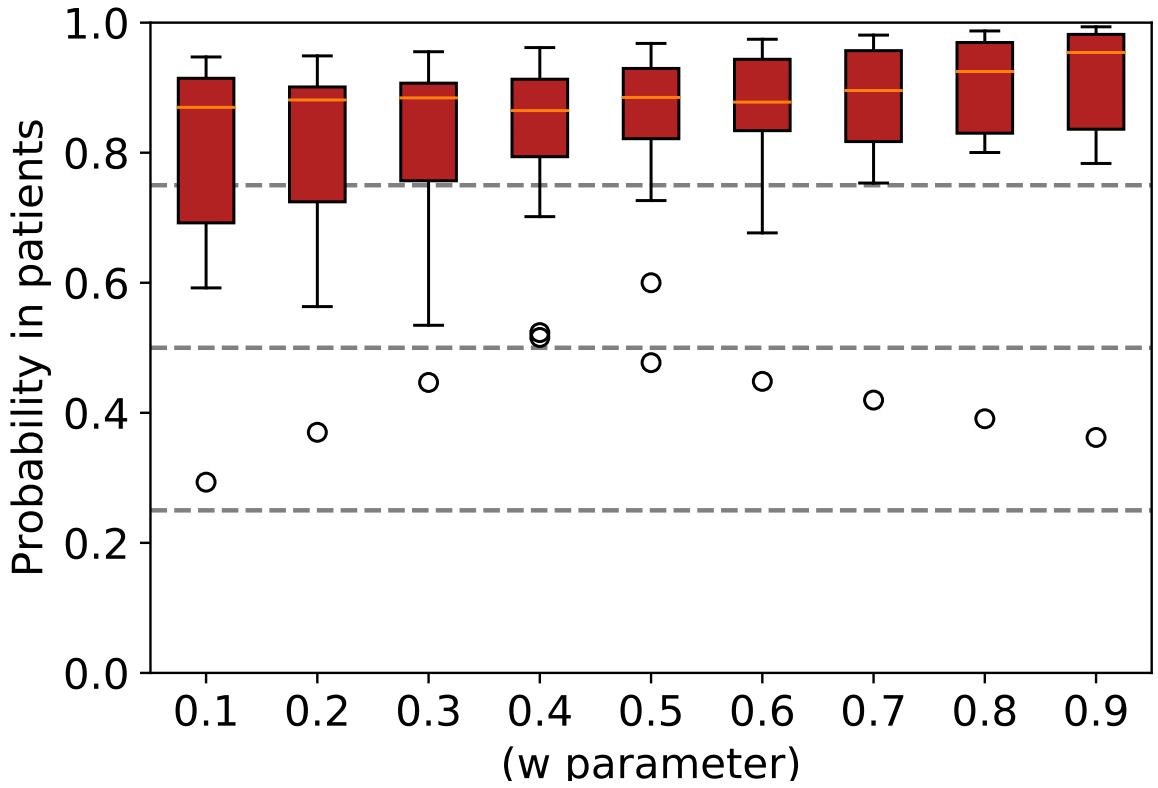

Figure 7. Distribution of probability predictions for Parkinson patients, for different w values. For $w = 0.4$, the outliers get a probability prediction higher than 0.5, achieving a proper classification. In contrast, all other values of w induce at least one false-negative.

use of only DFs shows comparable results with respect to early fusion. Nevertheless, late fusion obtains two false positives, *i.e.,* control patients that were classified as PD patients. Finally, a comprehensive experiment was carried out to analyze the probability outputs from the random forest classifier in early and late fusion for each patient. Figure 8 displays the PD probability, as outputted by the Random Forest classifier for each patient in the two fusion schemes. Control patients remain on probabilities lower than 0.3, which is a fairly confident index for binary classification. In addition, for PD patients, the probabilities are generally close to one. It is important to clarify that the default value of 0.5 was used for the calculation of the metrics, and that the dashed lines serve a purely illustrative purpose. Therefore,

no threshold selection criterion was applied in order to avoid bias associated with such selection. Additionally, the normality of the predicted probability distributions was evaluated following the Shapiro–Wilk normality test, separately to the control and patient groups, considering both the early fusion and late fusion models. The results for early fusion indicated that the predicted probabilities for the control group followed a Gaussian distribution (since $\rho > 0.05$, specifically $\rho = 0.9262$), whereas the Parkinson group did not exhibit normality (since $\rho < 0.05$, specifically $\rho = 0.0238$). Similarly, for the late fusion model, the predicted probabilities for the control group followed a normal distribution ($\rho = 0.55$), while the results for the Parkinson group rejected the normality hypothesis ($\rho = 0.0133$). Given that the distributions for the Parkinson group were not normally distributed, a non-parametric test was employed to compare the distributions. Consequently, the Kolmogorov–Smirnov test was applied to the Parkinson and control groups. The results showed a statistically significant difference in the probability distributions between the control and Parkinson predictions ($\rho = 1.9 \times 10^{-7}$) for both the early fusion and late fusion approaches. Interestingly, the three patients (p4, p5, and p6), in the early stage of the disease (second stage, according to the annotation of an expert following the H&Y scale), had gait locomotion patterns very similar to control subjects of the same age. Typically, patient p5 (second stage) had a lower probability than the patient p1 that had been categorized as third stage, according to the H&Y scale. For such patients, the oculomotor description appears to be the most discriminant with respect to the PD. For the other patients, the gait modality offers a greater contribution to the final probability because the gait patterns show more pronounced impairments than in the early stages. However, these patients had a higher PD probability in the early fusion scheme, which could indicate that early covariance integration leads to a better representation of the disease and could be more effective.

2.5. DISCUSSION AND CONCLUDING REMARKS

A novel approach was introduced for fusing motion modalities captured in markerless video sequences. In this study, a population of 26 patients, distributed as Parkinson (13 patients) and Control (13 patients), was considered. For each patient multiple sequences were recorded during gait locomotion from the sagittal view, without any invasive device. In addition, eye fixation patterns were recorded with a standard camera, and with weakly controlled conditions. Each frame of the sequence was first represented with DFs and/or KFs, computed from pre-trained convolutional networks and a dense optical flow, respectively. In this work, the DFs were computed from conventional (VGG) and separable architectures (MobileNetV2). Regarding KF, the horizontal and vertical components of the unit tangential vector and unit normal vector velocity were considered, as well as the magnitude of tangential and normal acceleration. Then, frame-level covariance coding was carried out to represent instantaneous posture and kinematics, which were thereafter summarized in a Riemannian temporal mean covariance. In multiple experiments, the proposed approach showed remarkable results, in configurations for early fusion (average accuracy of 100%) and late fusion (average accuracy of 100%). In the best configurations, the proposed approach combines KFs and DFs to achieve a more robust representation at the frame level. In addition, eye fixation had a high discrimination power and was therefore weighted with more importance in multimodal fusion.

These mean covariance matrices are very compact, ranging from 6 to 38 features. They are used as motion descriptors that can be fused (early or late) through a random forest classifier. A main limitation on covariance representation occurs when the prediction is based on only KFs or DFs, for both types of fusion. In such case, the kinematic information properly models Parkinsonian patterns in both modes but results insufficient to cover the whole variability for control subjects. In contrast, the proposed approach obtained excellent results by integrating both types of KFs and DFs in a total population of 26 subjects recorded three times, including 13 patients diagnosed with PD. Moreover, 95% confidence intervals for sensitivity

and specificity were calculated using the Clopper–Pearson method under the early and late fusion approaches. The narrowest intervals were observed when combining kinematic and deep features, suggesting a more stable and generalizable performance. Specifically, these intervals ranged from 0.75 to 1 for both sensitivity and specificity metrics, indicating that, based on the current sample, the model is likely to correctly classify patients (sensitivity) or controls (specificity) with a performance between 75% and 100%. This reinforces the improvement of combined features, which exhibit narrower confidence intervals, compared to independent features with broader intervals. It turns out that complementary features can deal with proper modelling of disease patterns, but also covering the control population. According to the random forest probabilistic results, early fusion was the best option, achieving an accuracy of 100%, using 38 features. Thus, the combination of both types of features significantly improves the differentiation between the motor patterns of control subjects and patients with PD.

Nowadays, the quantification of patients strongly depends on medical expertise based only on coarse motor scales and reported indices. These scales only consider strong motion changes, which limits the sensitivity to monitor the progression of the disease or to make an early diagnosis, and often produces high variance in final scores associated with a particular patient [12]. To overcome this issue, an approach is proposed to better monitor the disease, taking advantage of the markerless quantification of known patterns such as gait, but also integrating new biomarkers of the disease, such as tremor from the eyes. The integration of these motions results naturally from covariance frames, and markerless capture may offer potential applications in other types of non-controlled scenario. As demonstrated by the results, the combination of DFs and KFs make it possible to distinguish between control and PD patterns, with extremely small size descriptors (between 108 and $4\,332$), making real-time recognition possible, which is promising for clinical scenarios.

Multimodal approaches have been previously reported to analyze PD better: for instance, integration of walking patterns, speech signal analysis, and controlled writing experiments

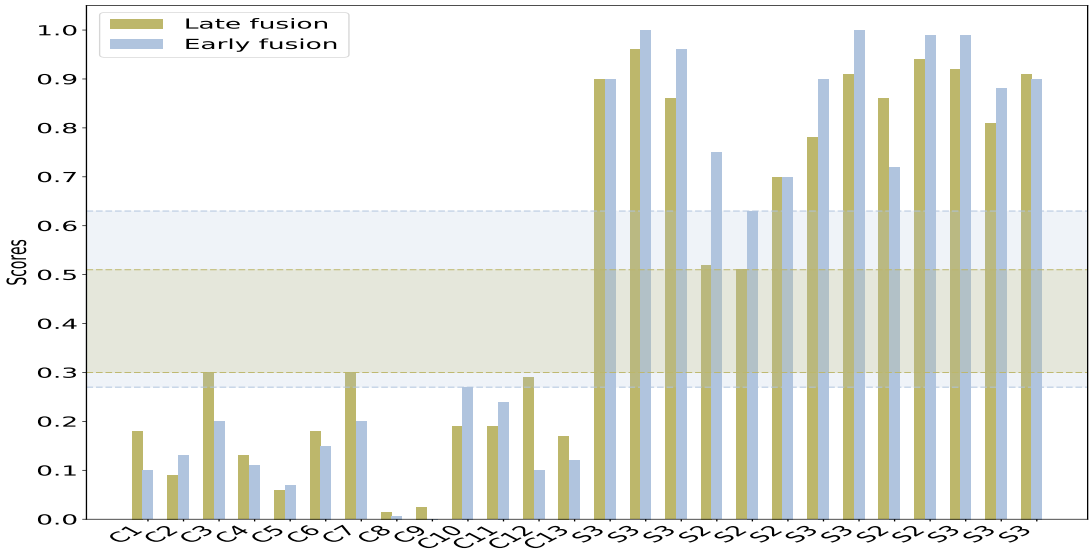

Figure 8. Probabilities of Parkinson with 13 control subjects (C) and 13 patients (P) considering kinematics and deep features in the two fusion's types.The wider horizontal blue stripe shows a higher confidence range for early fusion than for late fusion (horizontal green stripe).

[82, 83, 84]. Nonetheless, these methodologies depend on sophisticated capture devices and tedious experiments, making their use in routine clinical practice difficult. More recently, markerless strategies based on deep learning representations from video-sequences have

---

[82] H. N. PHAM et al. "Multimodal Detection of Parkinson Disease based on Vocal and Improved Spiral Test". In: *2019 International Conference on System Science and Engineering (ICSSE)*. 2019, pp. 279–284. DOI: 10.1109/ICSSE.2019.8823309.

[83] J. C. VASQUEZ-CORREA et al. "Comparison of User Models Based on GMM-UBM and I-Vectors for Speech, Handwriting, and Gait Assessment of Parkinson's Disease Patients". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 6544–6548. DOI: 10.1109/ICASSP40776.2020.9054348.

[84] J. C. VÁSQUEZ-CORREA et al. "Multimodal Assessment of Parkinson's Disease: A Deep Learning Approach". In: *IEEE Journal of Biomedical and Health Informatics* 23.4 (2019), pp. 1618–1630. DOI: 10.1109/JBHI.2018.2866873.

emerged [85]. In the latter work, an average accuracy of 90% was reported for the task of classifying PD with respect to the control population. These strategies have characterized gait patterns from end-to-end learning representations, but remain dependent on a huge and balanced training set to compute spatio-temporal patterns that discriminate patients with PD from the control population. Similarly, Lin *et. al* proposed a microwave motion detector to characterize tremor patterns using a non-invasive device, but it required complex calibration processes and protocols to capture motion signs [86]. Other studies have reported the sensitivity of the disease associated with eye motion patterns. In these studies for instance, in a population of 112 patients and only two controls (two from 60 control subjects), an ocular tremor with an average fundamental frequency of 5.7 Hz and an average magnitude of 0.27 in the horizontal plane and 0.33 in the vertical plane has been found. This shows the potential characteristics of PD biomarkers related to eye patterns [35]. These fundamentals have been explored to propose video strategies that recover and learn eye fixation patterns, making possible the representation of disease in weakly controlled scenarios [87]. In this case an approach proposed in previous work [87] achieved an average accuracy of 95%, in a population with 13 control subjects and 13 patients. Despite the remarkable results, this approach is limited by focusing on eye analysis and, losing other signs that may complement disease characterization.

Our work is based on a simple video system with a standard camera that provides objective information to the specialist in supporting the diagnosis and treatment of the disease.

---

[85] Luis Carlos GUAYACÁN; Edgar RANGEL, and Fabio MARTÍNEZ. "Towards understanding spatio-temporal parkinsonian patterns from salient regions of a 3D convolutional network". In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2020, pp. 3688–3691.

[86] Chia-Hung LIN et al. "Tremor Class Scaling for Parkinson Disease Patients using an Array X-Band Microwave Doppler based Upper Limb Movement Quantizer". In: *IEEE Sensors Journal* (2021).

[87] Isail SALAZAR et al. "A convolutional oculomotor representation to model parkinsonian fixational patterns from magnified videos". In: *Pattern Analysis and Applications* 24.2 (2021), pp. 445–457.

The proposed approach proved effective in classification tasks by achieving perfect classification scores in multimodal configurations. In addition, it is robust in assigning categorical probabilities to positive classified patients. Furthermore, many of the classic patterns are only detectable in advanced stages of the disease, restricting the analysis to the advanced Parkinson population. In contrast, the proposed approach considers gait patterns but also uses new eye fixational patterns, which have recently shown a major sensitivity to very early stages of the disease.

In summary, the proposed method integrates a local and global representation from KFs and DFs, which are effectively combined into covariance matrices to form a special Riemmanian manifold for each video sequence. The Riemannian mean from the manifold is easily integrated among different sequences and motion modes. The validation reported in this work should be extended in future studies to evaluate the ability of the proposed approach to distinguish different levels of the disease. In addition, patient diagnosis should be carried out on more sophisticated scales, such as the UPDRS-ME, to better correlate symptoms, to address the inter-experts variability, and to define a relevant inclusion of this tool within a clinical routine.

Despite the remarkable results of deep learning networks, these representations remain dependent on a large set of data to address the variability of samples, with a stratified condition among classes of the problem. In the biomedical context, meeting such requirements is difficult and leads to unnatural implementations to support routine treatments. For instance, the deep representation proposed by Guayacan et. al [85] achieved PD discrimination through an end-to-end learning scheme from 3D video analysis. However, this approach has reported average accuracy around 90%. These results show some limitations in operating with spatiotemporal maps, which may be associated with insufficient data for training. In addition, this type of approach has restrictions to include additional modalities. In contrast, recent advances in representation of deep learning strategies are exploited using DFs that can generalize the representation of input images without requiring any additional training.

The proposed scheme has the advantage of being robust in PD discrimination, but also very compact (video descriptors with a size between 108 and $4\,332$ scalar values). In addition, the proposed approach can integrate several motion modalities without any change in the pipeline of the method, which may benefit to a broader analysis from the clinical domain.
The proposed approach was validated through a limited study of 26 subjects, due to difficulties to acquire data from more patients. The main issue around data is the quantification of populations with comparable demography characteristics that allow to address methodologies related with the capability to discriminate disease patterns. Hence, as perspectives is proposed a further validation with larger datasets and stratified patients according to the progression of the disease. This validation may be useful to discover new multimodal patterns that enhance the sensitivity of clinical scales like the UPDRS and may better measure disease progression or the effectiveness of a particular treatment. Also, the inclusion of new modalities may enrich disease representation and impact as a tool to support early diagnosis.
The use of covariance descriptors was extended to classify spatiotemporal motor patterns in patients at early and intermediate stages. This approach utilizes a recurrent architecture. For more details, please refer to Appendix A.

3. A RIEMANNIAN MULTIMODAL REPRESENTATION TO CLASSIFY PARKINSONISM-RELATED PATTERNS FROM NONINVASIVE OBSERVATIONS OF GAIT AND EYE MOVEMENTS

3.1. ABSTRACT

Parkinson's disease is a neurodegenerative disorder principally manifested as motor disabilities. In clinical practice, diagnostic rating scales are available for broadly measuring, classifying, and characterizing the disease progression. Nonetheless, these scales depend on the specialist's expertise, introducing a high degree of subjectivity. Thus, diagnosis and motor stage identification may be affected by misinterpretation, leading to incorrect or misguided treatments. This work addresses how to learn multimodal representations based on compact gait and eye motion descriptors whose fusion improves disease diagnosis prediction. This work introduces a noninvasive multimodal strategy that combines gait and ocular pursuit motion modalities into a geometrical Riemannian Neural Network for PD quantification and diagnostic support. Markerless gait and ocular pursuit videos were first recorded as Parkinson's observations, which are represented at each frame by a set of frame convolutional deep features. Then, Riemannian means are computed per modality using frame-level covariances coded from convolutional Deep features. Thus, a geometrical learning representation is adjusted by Riemannian means, following early, intermediate, and late fusion alternatives. The adjusted Riemannian manifold combines input modalities to obtain PD prediction. The geometrical multimodal approach was validated in a study involving 13 control subjects and 19 PD patients, achieving a mean accuracy of 96% for early and intermediate fusion and 92% for late fusion, increasing the unimodal accuracy results obtained in the gait and eye movement modalities by 6 and 8%, respectively. The proposed method was able to discriminate Parkinson's patients from healthy subjects using multimodal geometrical configurations based on covariances descriptors. The covariance representation of video

descriptors is highly compact (with an input size of 625 and an output size of 256 (1 BiRe)), facilitating efficient learning with a small number of samples, a crucial aspect in medical applications.

The partial content of this work has been accepted and published in [88].

3.2. PROPOSED APPROACH

This work presents a multimodal strategy to capture and integrate PD motion impairments using a Riemannian manifold of spatiotemporal frame-level covariance matrices. Videos of each modality are represented as a set of frame-level covariance matrices. These parameters are further summarized through the Riemannian mean. Subsequently, we propose a multimodal Riemannian approach that explores the integration of motion modalities at early, intermediate, or late stages. The general pipeline is illustrated in Figure 9.

3.2.1. Learning Riemannian video mean descriptors. In this work, we consider a spatially dense representation, recovering for each frame a bank of deep features computed from a given layer of a convolutional network. Since adequate end-to-end training of deep architectures demands a huge amount of data, we decided to use pretrained convolutional networks with fixed weights for feature map extraction. Thus, a selected layer has a bank of learned convolution kernels, scalar biases, and nonlinear functions $\{\Phi_k, b_k, \mathbf{a}_k\}_{1\leq k\leq S}$, that decompose the image I_t into S feature maps $\mathbf{F_t} = \left\{\mathbf{F_t^k}\right\}_{1\leq k\leq S}$, where $\mathbf{F_t^k} = \mathbf{a}_k\left(I_t * \Phi_k + b_k\right)$. Then, for each frame t, a covariance matrix C_t with size $S \times S$ is computed as: $C_t(i,j) = \mathbb{E}\left(\left(\mathbf{F_t^i} - \mathbb{E}(\mathbf{F_t^i})\right)\left(\mathbf{F_t^j} - \mathbb{E}(\mathbf{F_t^j})\right)\right)$, where $\mathbb{E}$ denotes the spatial expectation, computed over the $W \times H$ values of each feature map. As a symmetric matrix, each frame covariance has dimensions of $\dim(C_t) = \frac{S(S+1)}{2}$.

---

[88] John ARCHILA; Antoine MANZANERA, and Fabio MARTÍNEZ. "A Riemannian multimodal representation to classify parkinsonism-related patterns from noninvasive observations of gait and eye movements". In: *Biomedical Engineering Letters* 15.1 (2025), pp. 81–93.

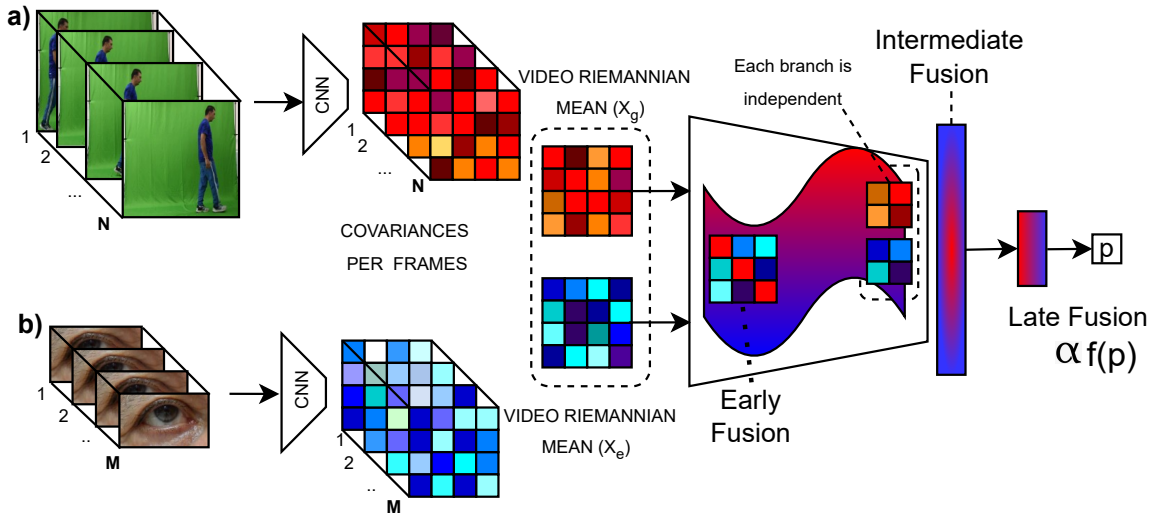

Figure 9. Riemannian video means: Gait and Ocular smooth motion modalities are processed and combined through Riemannian means, and then compact representations are learned by early fusion (on BiMap layers), intermediate fusion (on dense layers) or late fusion (at the probability level).

Hence, we compute a Riemannian temporal mean $\mathcal{X}(\mathbf{C})$, for each input video chunk of N frames, represented as the set of N covariance matrices $\mathbf{C} = \{C_t\}_{1 \leq t \leq N}$. Symmetric positive definite (SPD) matrices like Covariance in Riemannian manifolds, basic operations such as average cannot be computed by linear combination, but through an iterative gradient descent optimization, defined as: $\mathcal{X}_0 = C_1$, $\mathcal{X}_{n+1} = \exp_{\mathcal{X}_n}(\Theta_n)$ and $\Theta_n = \frac{1}{N}\sum_{t=1}^{N} \log_{\mathcal{X}_n}(C_t)$ [75]. Each video mean $\mathcal{X}(\mathbf{C})$ represents the point of the Riemannian manifold that minimizes the average distance to the covariance matrices of $\mathbf{C}$, thus minimizing geometric distortion. A Riemannian mean $\mathcal{X}^g(\mathbf{C}^g)$ (resp. $\mathcal{X}^e(\mathbf{C}^e)$) is then computed for each gait (resp. eye movement) sequence.

3.2.2. Riemannian Multimodal Representation. The main contribution of this work is the integration of multimodal PD impairments into a compact geometric manifold that computes second-order representations. Here we integrated the following complementary modalities:

- *Gait* is a key modality in the analysis of Parkinson's disease, since PD patients present alterations such as bradykinesia (slowness of movement), rigidity, balance problems, coordination, and posture issues. Additionally, gait is an activity closely related to the quality of life of patients, and changes in this complex locomotor dynamics can be used to assess disease severity and treatment response.

- On the other hand, changes in *Ocular movement* can be detected years before common motor symptoms appear, making it a potential early biomarker for the disease. Oculomotor impairments include reduced saccadic velocity and deteriorated smooth pursuit. These ocular biomarkers can improve diagnostic accuracy and speed.

Integrating these two modalities allows them to cover a wide range of early impairments, as well as cardinal symptoms like bradykinesia, encompassing afflictions present in the early, intermediate, and advanced stages of the disease. Then, we explore three fusion alternatives. First, an early integration was developed from second-order matrices computed from videos. Second, we explore the independent learning of modalities by distinct branches, preserving the data geometry, which is further integrated into a unified manifold. In the third version, fully independent Riemannian branches are integrated from the output probabilities. The details of each scheme are described in the next subsections.

Early Geometric Fusion. Eye and gait videos are represented as the mean SPD covariance matrices $\mathcal{X}^m$, where $m = \{g, e\}$ either gait (g) or eye (e) modality. These mean matrices summarize spatiotemporal patterns as correlations between deep computed features. Nonetheless, all the components of the matrix may not have the same significance concerning the Parkinson classification task. To perform early fusion, in this work, we first learn a new SPD matrix for each modality following a bilinear mapping as: $\mathcal{X}_k^m = W_k \mathcal{X}_{k-1}^m W_k^T$, with

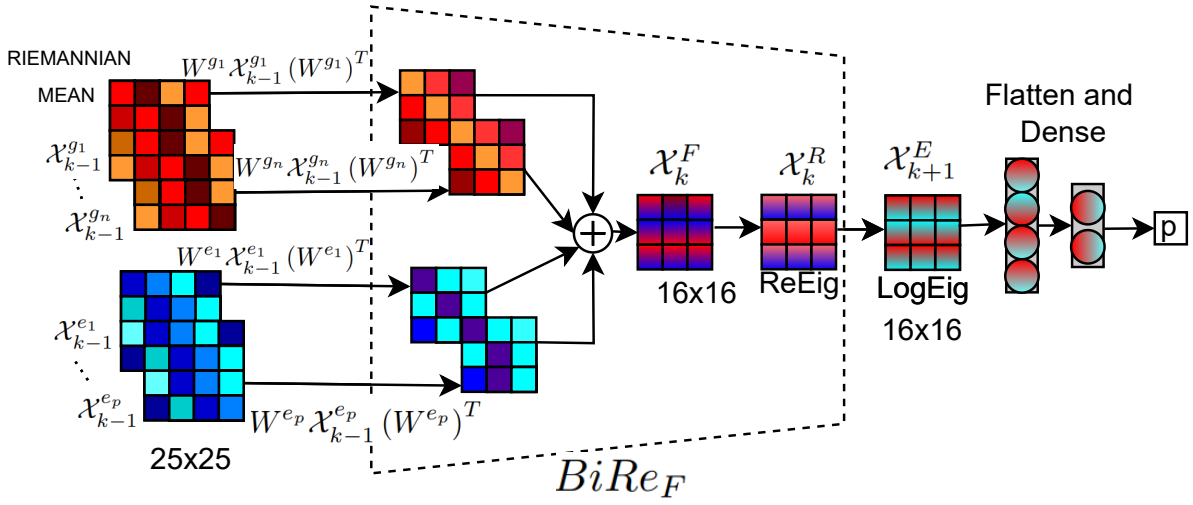

Figure 10. Early Fusion Modality: Using the Riemann means $\mathcal{X}^g$, where $\mathcal{X}^e$ represents the video chunk of each modality, the Bimap layer $W\mathcal{X}W^T$, is applied for each modality and the output matrices are integrated into a single symmetric matrix: $\mathcal{X}^F$. Subsequently, the ReEig layer is applied to maintain the geometry of the matrices: $\mathcal{X}^R$, and then the mapping to the Euclidean space is performed for classification: $\mathcal{X}^E$.

$\mathcal{X}_{k-1}^m \in \mathbb{R}_*^{d_{k-1}\times d_{k-1}}$, and $W_k \in \mathbb{R}_*^{d_k\times d_{k-1}}$ the weight matrix transformation [89]. Then the output matrix at step k, $\mathcal{X}_k^m \in \mathbb{R}_*^{d_k\times d_k}$. This projection outputs a new SPD matrix for each modality, whose components are optimised by the training process according to the classification of PD. Hence the early geometrical fusion (as illustrated in Figure 10 is formulated, for n gait video chunks and p eye pursuit video chunks, as follows:

$$\mathcal{X}_k^F = W^{g_1}\mathcal{X}_{k-1}^{g_1}(W^{g_1})^T + .. + W^{g_n}\mathcal{X}_{k-1}^{g_n}(W^{g_n})^T + W^{e_1}\mathcal{X}_{k-1}^{e_1}(W^{e_1})^T + .. + W^{e_p}\mathcal{X}_{k-1}^{e_p}(W^{e_p})^T$$

Then, $\mathcal{X}_k^F$ is a new SPD matrix that conjugates early fusion from $\{\mathcal{X}_{k-1}^{g_i}\}_i$ and $\{\mathcal{X}_{k-1}^{e_j}\}_j$ input matrices. To ensure SPD property, an eigenvalue rectification layer is carried out, as:

---

[89] Michael M BRONSTEIN et al. "Geometric deep learning: going beyond Euclidean data". In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 18–42.

$\mathcal{X}_k^R = U_k \max(\varepsilon I, \Sigma_k) U_k^T$ where U_k is the matrix formed by the eigen-vectors and Σ_k the diagonal matrix formed by the eigen-values of $\mathcal{X}_k^F$. Here, $\varepsilon > 0$ is a rectification threshold value, and I is the identity matrix. This operation adjusts the eigenvalues, avoiding negative values and improving discriminative performance. The resultant $\mathcal{X}_k^R$ can be projected again following a bilinear map (BI) with the consecutive rectification eigenvalue rectification (RE). This process, referred to as geometric learning layers $(BiRe_F)$, can be iterated several times. After each $BiRe_F$ layer, a Riemannian batch normalization algorithm is applied, to preserve the structure of the input descriptors [90]. At the end of all considered (n_B) $BiRe_F$ blocks, a LogEig layer is implemented to project the output features from the Riemannian manifold to a Euclidean space and carry out the classification task. This is done by the Riemannian logarithm map, defined as: $\mathcal{X}_{k+n_B}^E = \mathbf{log}(\mathcal{X}_{k-1+n_B}^R) = U_{k-1+n_B}^R \log(\Sigma_{k-1+n_B}^R)(U_{k-1+n_B}^R)^T$. This projection is followed by standard flattening, dense layers, and softmax to carry out the PD classification.

Intermediate Fusion from independent Geometrical branches. As an alternative geometric fusion, from each SPD input $\mathcal{X}_{k-1}^m$ can be learned a geometrical branch using BiRe blocks. Then, outputs are fused at an intermediate level. In particular, the SPD input associated with each chunk is corresponds to a single Riemannian mean and is processed independently through its sequence of BiRe blocks.

After their n_B BiRe blocks, Riemannian logarithm maps $\mathbf{log}$ are computed for each chunk and concatenated: $\left[\mathbf{log}(\mathcal{X}_{k-1+n_B}^{Rg_l}), \ldots, \mathbf{log}(\mathcal{X}_{k-1+n_B}^{Re_l})\right]$.

Then the produced matrices are flattened, concatenated and processed by a dense layer $\mathcal{D}$. The resultant descriptor is computed as:

$$\mathcal{F}_{int} = \mathcal{D}\left[\mathbf{log}(\mathcal{X}_{k-1+n_B}^{Rg_l}), \ldots, \mathbf{log}(\mathcal{X}_{k-1+n_B}^{Re_l})\right].$$

---

[90] Daniel BROOKS et al. "Riemannian batch normalization for SPD neural networks". In: *Advances in Neural Information Processing Systems* 32 (2019).

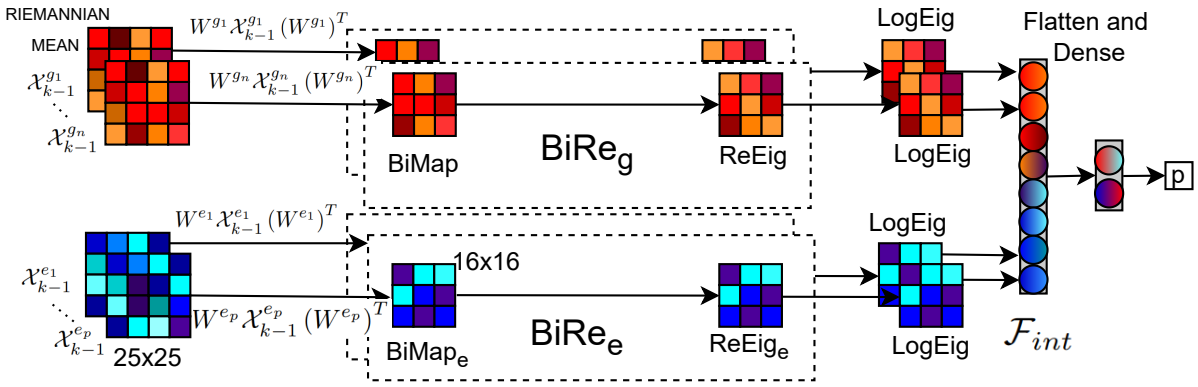

Figure 11. Intermediate Fusion Modality: Each Riemannian mean associated with a video chunk from gait or eye pursuit is independently processed through its Riemannian branch. Then, all the outputs, once mapped to the Euclidean space by their LogEig layer, are flattened and concatenated to be input to the same fully connected network.

As for the early fusion, at the end of the network, a softmax layer computes the probabilities of the Parkinson's disease and control classes.

Late fusion from modality-wise probability outputs. As a third multimodal integration, this work explored late fusion by combining the probability output of two independent geometrical branches. The approach considers different branches, that specialize in learning geometrical manifolds, independently, for each modality and for each video chunk. To fuse multiple Parkinson's probabilities, the importance α of each modality is computed based on the relative consistency index (Ci) between probabilities among the different video chunks. The rationale is that prediction probabilities should be similar for different Riemannian means of the same sequence. This index is computed on gait modality because locomotion is particularly valuable in the diagnosis and differentiation of Parkinson's disease for the reason that it encompasses the assessment of motor symptoms such as tremors, postural instability, and bradykinesia. These distinctive motor symptoms can provide crucial diagnostic clues

and help distinguish Parkinson's disease [91]. Then, a fusion is represented as a linear combination of probabilities. The consistency index compares the probabilities of the different Riemann means for the gait mode.

A value near zero indicates high consistency in gait prediction. Therefore it implies a high weighting of the gait modality $\alpha = 1 - Ci$, and then $1 - \alpha = Ci$ for the eye modality. In the particular case where only 1 Riemannian mean per modality is used, a value of $\alpha \in [0, 1]$ is set as a hyperparameter before training.

3.3. EXPERIMENTAL SETUP

3.3.1. Multimodal data. This work evaluates the ability of a multimodal approach that integrates gait and eye smooth pursuit videos to discriminate Parkinsonian from normal movement patterns. To do so, this study included 13 control subjects (average age of 72.2 ± 6.1) and 19 PD patients (average age of 72.3 ± 7.4). PD patients were categorized at different stages of the disease following the Hoehn-Yahr scale. According to observations carried out by an expert physician, one patient was categorized in the first stage (unilateral motor symptoms), and two patients were categorized in the second stage (motor impairment on both sides but no postural instability). The remaining sixteen patients were categorized in the third stage, exhibiting bilateral patterns related to (motor impairment on both sides and balance impairment). During this study, the participants were given the following motion instructions:

- *Gait*: In this exercise, participants were invited to walk in a straight line, and the camera was positioned to capture them from their sagittal view. The entire locomotion was recorded over a 5-meter walking distance, with an average video duration of 5 seconds per recording. Each video had a spatial resolution of 520×520 pixels and a temporal

[91] Grazia CICIRELLI et al. "Human Gait Analysis in Neurodegenerative Diseases: a Review". In: *IEEE Journal of Biomedical and Health Informatics* (2021).

resolution of 60 frames per second (fps). Figure 25 depicts the postural configuration during the gait exercise, with a uniform green background in the scene. The purpose of this configuration is to facilitate the observation of the impact of disease on posture, bradykinesia, and hypokinesia (reduction in the amplitude of movements). Additionally, from the sagittal view, it is possible to detect if the patient exhibits impairment on either side.

- *Ocular Smooth Motion*: In this scenario, patients were instructed to focus on a spotlight projected onto a screen with a dark background with an average video duration of 5 seconds per recording. The height of the monitor was carefully adjusted to align the center of the screen with the center of the pupillary plane, as illustrated in Figure 1. The motion of the spotlight was controlled both horizontally (from right to left and vice versa) and vertically (from top to bottom and vice versa). Subsequently, the recorded video was manually cropped (210×140 pixels) to a region of interest around the eye. In the eye movement analysis, 8 videos are considered, 4 for the left eye and 4 for the right eye. The proposed setup based on smooth eye movement facilitates the detection of oculomotor impairments, such as bradykinesia and other movement disorders, by the specialist.

A total of sixteen videos were acquired for each participant, encompassing both gait and ocular smooth motion exercises. The association of a gait video and an eye movement video during the training and test of our models is related to laterality since asymmetric symptoms are an important aspect of the disease. Then, a video that captures the right side of the patient from the sagittal view during gait will always be associated with a video of the right eye. In both modalities, we utilized the same camera, a conventional Nikon D3200, which offered a spatial resolution of 1280×720. Consequently, the entire dataset for this study comprises a total of 512 videos. This study was approved by the Ethics Committee, and written informed consent was obtained from each participant.

To assess the performance of the proposed method, a leave-one-patient-out cross-validation

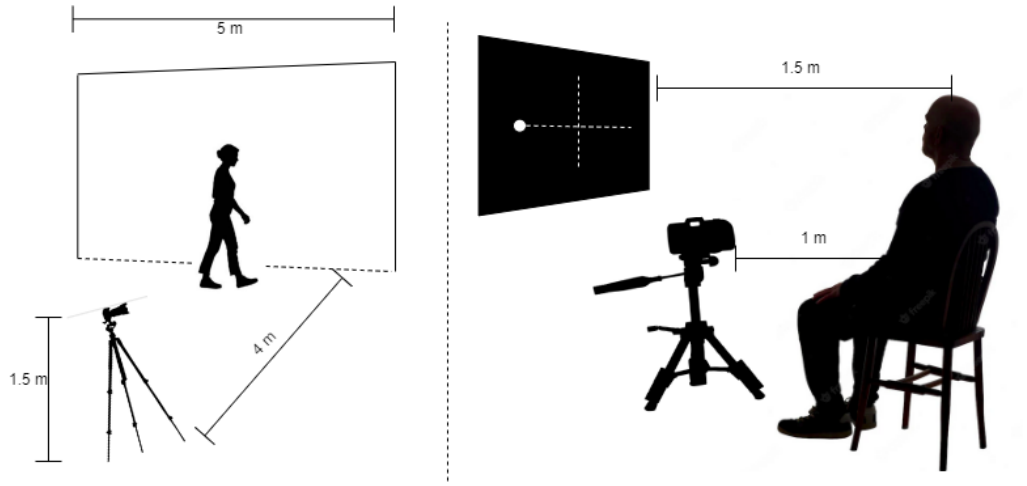

Figure 12. Acquisition setup of gait and ocular smooth motion modalities in our markerless approach.

was employed with the multimodal dataset. In this approach, at each iteration, one patient is excluded from testing, and the remaining ones (31 subjects) are used for training. For these experiments, Parkinsonian patients (resp. Control persons) correctly identified were counted as true positives (TP) (resp. true negatives, TN). Subsequently, a set of metrics was utilized to comprehensively evaluate the performance of the method in its different configurations. The metrics calculated here are accuracy ($acc = \frac{TP+TN}{TP+FP+FN+TN}$), sensitivity ($sen = \frac{TP}{TP+FN}$), precision ($prec = \frac{TP}{TP+FP}$), and the F1-score ($F_1 = \frac{2 \times prec \times sen}{prec+sen}$).

3.3.2. Parameter Tuning. The suggested method was fine-tuned at various phases to enhance the depiction of the characterization and measurement of Parkinsonian disease patterns. At each phase, the following parameters were explored:

- **Training Characteristics.** The evaluation was carried out using a subject-wise leave-one-patient-out cross-validation scheme. The hyperparameter were fixed for all experiments before the cross-validation scheme. In each iteration, the model was trained using data from 31 participants and evaluated on the excluded 1 participant. This

procedure ensured that performance was always assessed on unseen data by strictly separating the training and testing sets. The final configuration was selected based on the average performance across the different metrics throughout all iterations.

- **Deep features.** 8 different options were evaluated, considering the outputs of 4 different layers of the pre-trained MobileNet V2, from the 2^{nd} to the 5^{th} one, and considering either 25 or 32 filters per layer.

- **Number of Riemannian Means per video.** To analyze the ability to learn new Riemannian representations from videos of gait and eye movement modalities, computing 1 or 2 Riemannian means per video was compared. For two Riemannian means, each descriptor is associated with the duration of half the video.

- **Deep of Geometrical Network.** The contribution of the depth of the geometric architecture is related to the model's ability to learn more efficient geometric representations. For this purpose, experiments were conducted with 1 and 2 BiRe layers, reducing the descriptor dimension to 16×16 and 12×12, respectively.

- **Euclidean Last layers.** All experiments were performed with 2 dense layers with sizes of 64 and 2, respectively. At the output, a softmax layer outputs the prediction of the video chunk for intermediate and early fusion. Each modality branch possesses its softmax layer for late fusion, which are linearly combined with a learned alpha parameter.

3.4. RESULTS

The initial evaluation phase of the proposed approach involves an assessment of the capabilities of the deep features to characterize each modality. In this study, we considered a video-level representation using deep features, independently evaluated within each modality, with one single Riemannian mean computed for the whole video. Table 7 provides a

Deep Features	Acc Gait	Acc Eyes	Size Descriptor
MobileNetV2/2nd	0.78	0.78	1024
MobileNetV2/3rd	0.8	0.82	1024
MobileNetV2/4th	0.87	0.82	1024
MobileNetV2/5th	0.84	0.86	1024
MobileNetV2/2nd	0.82	0.79	625
MobileNetV2/3rd	**0.87**	**0.88**	**625**
MobileNetV2/4th	0.87	0.78	625
MobileNetV2/5th	0.85	0.83	625

Table 7. Comparison of the accuracy obtained by deep features taken from different layers in each modality, with one single Riemannian mean for the whole video

summary of the individual performance results, with Riemannian mean descriptors of dimensions $1\,024$ and 625, computed for 32 and 25 feature maps, respectively. Remarkably, the features retrieved from MobileNetV2/3rd (25 feature maps) were sufficient to achieve the best performance for both motion modes. They achieved an average accuracy of 0.87 ± 0.1 for gait and 0.88 ± 0.2 for smooth ocular motion.

The second experiment aimed to evaluate early fusion. In such cases, the proposed approach was evaluated using 2, 3 or 4 Riemannian means (RMs) that integrate both modalities (gait and eye movement). Additionally, the models were computed considering 1 or 2 BiRe layers to assess the contribution of the Riemannian networks. Table 8 summarizes the scores associated with the different metrics as a function of the number of BiRe layers and the number of RMs at the input. The first computed results were with 1 RM for each video modality and one BiRe layer. This proposed configuration has a high sensitivity (91%) and an F1 score of 0.91. Interestingly, one early integration layer was sufficient to achieve a remarkable score. Notably, adding a gait RM achieved a perfect score for precision and specificity. Additionally, a more redundant representation with 4 RMs (2 for gait and 2 for eye movement) decreases the classification performance.

Third, we evaluated the intermediate fusion, where each independent branch generates a positive definite symmetric matrix, without considering the weighting of the other input modal-

RM input	BiRe	sen	spec	acc	prec	F1-s
2RM (1G-1E)	1	0.91	0.86	0.89	0.91	0.91
	2	0.82	0.77	0.8	0.83	0.82
3RM (2G-1E)	**1**	**0.93**	**1**	**0.96**	**1**	**0.96**
	2	0.93	0.99	0.95	0.99	0.96
3RM (1G-2E)	1	0.93	0.91	0.92	0.94	0.93
	2	0.87	0.86	0.87	0.85	0.86
4RM (2G-2E)	1	0.75	0.82	0.78	0.86	0.8
	2	0.78	0.77	0.77	0.83	0.8

Table 8. Early fusion, using 2, 3, or 4 Riemannian means (RMs) of gait (G) and eyes (E), with 1 or 2 BiRe layers.

RM input	BiRe	sen	spec	acc	prec	F1-s
2RM (1G-1E)	1	0.93	0.81	0.88	0.88	0.9
	2	0.82	0.77	0.8	0.83	0.82
3RM (2G-1E)	**1**	**0.95**	**0.99**	**0.96**	**0.99**	**0.97**
	2	0.92	0.99	0.95	0.99	0.95
3RM (1G-2E)	1	0.92	0.86	0.9	0.9	0.91
	2	0.85	0.91	0.88	0.93	0.89
4RM (2G-2E)	1	0.72	0.82	0.76	0.85	0.75
	2	0.69	0.82	0.74	0.85	0.76

Table 9. Intermediate fusion, using 2, 3, or 4 Riemannian means (RM) of gait (G) and eyes (E), with 1 or 2 BiRe layers.

ity matrices. Table 9 summarizes the results of varying the number of RMs and the number of BiRe layers. The influence of the fusion was analyzed considering 1 RM for each modality; the values of sensitivity (93%) and F1-score 0.9 with 1 single BiRe layer were highlighted. Again, a decreasing performance behavior is observed when using 2 BiRe layers.

The specificity, accuracy, and precision increased between 2% and 5% when additional RMs were integrated for eye movements (1 for gait and 2 for smooth eye movement). On the other hand, the performance of 3 RMs (2 for gait and 1 for smooth eye movement) with 1 and 2 BiRe layers achieved the highest intermediate fusion results for sensitivity (95%), specificity (99%), accuracy (96%), precision (99%), and F1-score (0.97).

RM input	BiRe	sen	spec	acc	prec	F1-s
2RM (1G-1E)	**1**	**0.92**	**0.92**	**0.92**	**0.94**	**0.93**
	2	0.83	0.75	0.80	0.84	0.84
3RM (2G-1E)	1	0.9	0.88	0.86	0.92	0.93
	2	0.87	0.8	0.84	0.86	0.86
3RM (1G-2E)	1	0.87	0.89	0.89	0.92	0.89
	2	0.79	0.85	0.81	0.88	0.83
4RM (2G-2E)	1	0.66	0.86	0.74	0.88	0.75
	2	0.69	0.81	0.78	0.84	0.75

Table 10. Late fusion, using 2, 3, or 4 Riemannian means (RM) of gait (G) and eyes (E), with 1 and 2 BiRe layers.

Fourth, the late fusion results are summarized in Table 10. In contrast to early and intermediate fusion, the best results are associated with the minimum number of RMs (1 for each modality; see Tables 8 and 9) with specificity, sensitivity, and accuracy values of 92%, precision of 94% and F1 scores of 0.93 with 1 BiRe layer. These results were obtained with a predefined alpha value of 0.7. This means a weighting of 70% for the gait mode and 30% for the smooth eye movement mode. Analogous to the early and intermediate fusion, the worst performance is associated with 4 RMs (2 for each modality) with a reduction between 6% and 26% of the calculated metrics.

In a complementary way, Table 11 shows the best results of the three fusion approaches. Interestingly, the intermediate fusion approach has the highest predictive value for videos associated with Parkinson's disease patients (94.7%). Although the late fusion approach presents lower values, this configuration provides decent results with compact representations, *i.e.,* with the minimum number of Riemannian means (1 for each modality).

Additionally, multimodal strategies were compared with unimodal geometrical representations, for gait and eyes, independently (see Table 12). The results of early and intermediate fusion are obtained with the two RMs associated with gait and one RM associated with eye movement. The results Results show higher F1-score than those of unimodal approaches. In the late fusion approach, simple linear weighting of the probabilities slightly outperforms uni-

Configuration		2G-1E 1BiRe		2G-1E 2BiRe	
		PK	**C**	**PK**	**C**
Early	**PK**	**141 (92.7%)**	11(7.3%)	141 (92.7%)	11(7.3%)
	C	0 (0%)	**104 (100%)**	1 (1%)	103 (99%)
Intermediate	**PK**	**144 (94.7%)**	8 (5.3%)	140 (92.1%)	12 (7.9%)
	C	1 (1%)	**103 (99%)**	1 (1%)	103 (99%)
Late	**PK**	**140 (92.1%)**	12 (7.9%)	135 (88.8%)	17 (11.2%)
	C	8 (7.7%)	**96 (92.3%)**	26 (25%)	78 (75%)

Table 11. Confusion matrices (percentages expressed w.r.t. the number of sequences) for the different fusion modes, using 3 Riemannian means with one or two BiRe layers at different levels of fusion: early, intermediate, or late.

modal approaches; however, integrating modalities in Riemannian space or Euclidean space is more robust than simple linear weighting of probabilities. Additionally, confidence intervals were calculated for the unimodal geometric approaches (Gait with 2 Riemannian models, Ocular with 1 Riemannian model) and the multimodal geometric approaches (early and intermediate fusion with 3 Riemannian models, and late fusion with 2 Riemannian models). The narrowest sensitivity intervals were observed in the gait, early fusion, and intermediate fusion approaches, ranging from [0.74, 0.99]. In contrast, wider intervals were found for late fusion [0.69, 0.98] and ocular movement [0.64, 0.96]. Specificity confidence intervals were also calculated; the early and intermediate fusion approaches yielded the narrowest intervals [0.75, 1], whereas late fusion and ocular movement resulted in wider intervals [0.64, 0.99]. The poorest result was associated with the gait modality [0.54, 0.98]. Consequently, the best outcomes in terms of both sensitivity and specificity were associated with the early and intermediate fusion geometric approaches.

3.5. DISCUSSION AND CONCLUDING REMARKS

Today, the assessment of PD patients relies heavily on the expertise of healthcare professionals, primarily using coarse motor scales such as the H&Y, the MDS-UPDRS, and re-

Modality	RM Input	sen	spec	acc	prec	F1-s
Gait	1	0.92	0.81	0.87	0.87	0.89
	2	0.97	0.81	0.9	0.88	0.92
Eyes	1	0.86	0.9	0.88	0.92	0.89
	2	0.83	0.9	0.86	0.92	0.87
Early Fusion	2G-1E	0.93	1	0.96	1	0.96
Intermediate	2G-1E	0.95	0.99	0.96	0.99	0.97
Late Fusion	1G-1E	0.92	0.92	0.92	0.94	0.93

Table 12. Gait, ocular, and multimodal fusion scores, using 1, 2 or 3 Riemannian means and 1 BiRe

ported indices such as the Clinical Impression of Severity Index for PD (CISI-PD)[92][93][94]. To overcome such subjectivity, this work introduced a multimodal geometric approach that fuses gait and eye movement patterns, exploring early (from video covariances), intermediate (from dedicated mode paths), and late integration (from output probabilities) methods to classify Parkinsonian patterns.

The results showed a robust representation when covariance inputs were fused into early and intermediate stages ($100 - 99\%$ specificity and $100 - 99\%$ precision for early and intermediate fusion). Additionally, the late fusion alternative demonstrated competitive performance (92% specificity and 94% precision), combining the output probabilities from a linear rule with the advantage of interpretability provided for each branch. This interpretability can be valuable for clinicians and researchers to gain insights into disease characteristics. However, the narrowest 95% confidence intervals for sensitivity and specificity were observed in the early

---

92 Margaret M HOEHN and Melvin D YAHR. "Parkinsonism: onset, progression, and mortality". In: *Neurology* 17.5 (1967), pp. 427–427.

93 Christopher G GOETZ et al. "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results". In: *Movement disorders: official journal of the Movement Disorder Society* 23.15 (2008), pp. 2129–2170.

94 Pablo MARTÍNEZ-MARTÍN et al. "The clinical impression of severity index for Parkinson's disease: international validation study". In: *Movement Disorders: Official Journal of the Movement Disorder Society* 24.2 (2009), pp. 211–217.

and intermediate fusion approaches [0.75, 1]. In contrast, the wider intervals seen in the unimodal and late fusion approaches may reflect greater susceptibility to data fluctuations, potentially limiting their stability in clinical applications. It is worth noting that the lower performance of the unimodal gait approach in terms of specificity highlights the limitations of relying on a single modality, even when the input features effectively capture disease-related motor patterns. It is also important to acknowledge that a lower bound of 0.75 is not particularly high. This suggests that, while these models are promising, there remains a non-negligible degree of uncertainty regarding their ability to consistently maintain high performance across different samples or larger populations. During the validation of the proposed approach, it was also found that two Riemannian means were effective in describing gait, coding directed movements, periodic patterns, and accelerations at the beginning and end of locomotion. Regarding the smooth eye movement, a single Riemannian mean was sufficient to describe the observed patterns of eye movement. This fact may be associated with recorded uniform eye patterns, stability, and quasi-stationarity in gaze direction.

Current alternatives in the state-of-the-art for supporting PD classification primarily rely on isolated methods, for instance, analyzing gait that carries much of the PD motor impairments [9596, 97, 98, 29]. For gait, kinematic analysis has been classically conducted from marker reflectors and systems with multiple cameras [96, 97]. For example, in a study involving 20 patients and 20 control subjects, spatiotemporal measurements were taken, including step length,

---

[95] Tianpeng LI et al. “Automatic timed up-and-go sub-task segmentation for Parkinson’s disease patients using video-based activity classification”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26.11 (2018), pp. 2189–2199.

[96] Rebecca BAN et al. “Dynamic gait stability in people with mild to moderate Parkinson’s disease”. In: *Clinical Biomechanics* 118 (2024), p. 106316.

[97] Michela RUSSO et al. “Kinematic and Kinetic Gait Features Associated With Mild Cognitive Impairment in Parkinson’s Disease”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2024).

[98] Jannis van KERSBERGEN et al. “Camera-based objective measures of Parkinson’s disease gait features”. In: *BMC research notes* 14 (2021), pp. 1–6.

velocity, cadence, and phase duration of gait. Statistical analysis identified the most significant variables, highlighting step velocity and amplitude [96]. Another study using reflectors and motion capture systems measured limb flexion and extension, achieving results above 80% accuracy with classic methods such as SVM and GBT (Gradient Boosted Tree) [97]. These strategies nonetheless are supported by marker setups, placed on different body parts, that may affect the naturalness of the patient's movement and hinder their use in routine clinical practice.

Hence, markerless approaches have been introduced as an alternative to kinematic gait analysis through computer vision [95, 98]. For instance, a computer vision strategy has been processed to estimate key postural joints during locomotion and encode spatiotemporal information, which thereafter is projected to an SVM classifier for PD prediction, reporting a sensitivity of 93.1% (a study with 24 PD patients) [95]. Alternatively, the correlation between kinematic variables such as walking speed, step length, and mediolateral sway was calculated in a study with 19 patients and 8 control subjects, reporting statistical differences $(p \leq 0.001)$ for step length and average speed [98]. These works reduced gait complexity by using only postural analysis, which may lose Parkinsonian spatiotemporal details, proper to dynamic postural changes. Other computer vision approaches use dense representations, taking advantage of whole video recordings, for instance using 3D convolutional networks for classification and identification of key movement regions during gait. A convent for Parkinsonian gait characterization achieved an accuracy of 88% for 11 control subjects and 11 PD patients [29]. These recent approaches have evidenced key advantages in analyzing gait patterns and standing out patterns associated with PD. However, although gait is an important modality for observing motor impairments, this modality is principally dedicated to characterizing intermediate and advanced stages

Regarding early and even prodromal PD characterization, eye movement analysis has gained

attention in the literature for recent finding correlations with this disease [19, 99, 100, 101, 87]. Typically, oculomotor patterns are measured from electrooculography protocols, involving electrodes to measure signals [99, 100]. For example, latencies, velocities, and angular velocities have been measured from such protocols to establish significant differences between PD patients and control subjects (a study with 45 patients and 30 control subjects) [100]. These approaches however are restricted to standard measures without exploring potential relationships to find new PD patterns. Alternatively, a traditional machine learning technique employing vertical electro-oculography, supported by SVM, was employed to quantify time-frequency features, achieving a sensitivity rate of 69.7% and specificity of 87.1% for 27 PD patients [99]. Besides, video-oculography protocols have been less sensitive to noise compared to electrooculography protocols. By measuring kinematic variables such as pupil reaction and deviation [19] or jerks, latency movements, and gain of horizontal smooth pursuit [101], statistically significant differences were found between Parkinson's patients and control subjects. However, video-oculographic protocols generally require complex calibration methods and additional training for the specialist. Moreover, using only kinematic trajectories can simplify this complex movement, hindering predictive capability. Therefore, methods have been proposed that compute deep features using simple protocols based on a conventional video camera. For example, deep pre-trained features from ocular videos of 13 patients and 13 control subjects have been encoded using covariance matrices, subsequently performing classification with SVM, achieving an accuracy above 90% in diagnosis prediction [87]. Despite remarked advances, these isolated unimodal alternatives may lose the PD pheno-

---

[99] Sajjad FARASHI. "Analysis of vertical eye movements in Parkinson's disease and its potential for diagnosis". In: *Applied intelligence* 51.11 (2021), pp. 8260–8270.

[100] Olivier RASCOL et al. "Abnormal ocular movements in Parkinson's disease: evidence for involvement of dopaminergic systems". In: *Brain* 112.5 (1989), pp. 1193–1214.

[101] JianYuan ZHANG et al. "Eye movement especially vertical oculomotor impairment as an aid to assess Parkinson's disease". In: *Neurological Sciences* 42 (2021), pp. 2337–2345.

typing, and therefore it is required the monitoring of multiple modalities to establish a robust characterization of the disease.

Some multimodal alternatives to support the multifactorial character of the disease have been introduced integrating gait patterns, speech signals, and controlled writing experiments [102, 49]. Particularly, convolutional neural networks were employed to model the spectrograms of speech and kinematics patterns associated with writing and gait, achieving an accuracy of 95% considering 42 patients and 40 control subjects [49]. This methodology addresses the integration of modalities, without an exhaustive analysis about the individual contribution of each modality. Also, handcrafted features were computed from gait, speech, and kinematics from writing to adjust Gaussian Mixture Models, emphasizing the significance of features from each modality [102]. However, these methods are invasive, requiring inertial sensors and sophisticated capture protocols to obtain an appropriate representation, such as the need for a completely silent environment for voice recording or the manipulation of tablets for writing results can be challenging for some patients. Additionally, motor impairments associated with these modalities mainly occur in intermediate stages where motor impairments are highly noticeable. Other methods consider synchronized modalities, for instance, using voice and facial expression, and using BiLSTM recurrent networks, Rubiano-Cruz *et.al* achieved accuracies around 80% in a study with 13 PD patients and 13 control subjects. However, the selection of these two modalities presents difficulties because they can vary according to language, accent, culture, and the individual's emotions, complicating the process of data reproducibility and normalization. In the intermediate stages, patients may have severe difficulties in speech and facial expression, making monitoring more challenging [103].

---

[102] Juan Camilo VÁSQUEZ-CORREA et al. "Comparison of user models based on GMM-UBM and i-vectors for speech, handwriting, and gait assessment of Parkinson's disease patients". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 6544–6548.

[103] Ricardo Andres RUBIANO-CRUZ. "Detection of Parkinson's Disease with Multimodal Deep-Learning". In: (2024).

Recently, a method to quantify gait and eye movement modalities was proposed, using an infrared eye tracker and an inertial gait device to estimate kinematics associated with both modalities, respectively. Such an approach calculated the variables with the greatest statistical significance. A multivariate logistic regression model was built from saccade velocity, max trunk sway, and mean turn angular velocity, to discriminate motor patterns between patients and control subjects [46]. However, this approach relies on advanced capture equipment and laborious experiments, making their implementation in routine clinical practice challenging. Additionally, the measurement of only kinematic variables through sensors located on specific parts of the body may limit the movement characterization. Besides, the proposed model has difficulty identifying nonlinear relationships between the variables under study, resulting in the loss of potentially useful information. Recently, another strategy has integrated eye fixation and gait classification based on computing the covariance of deep features for random forest classification. This approach achieved a sensitivity of 92% with 13 PD patients [47]. Nonetheless, this approach is limited to exploiting and learning a compact geometry from coded descriptors. In contrast, the proposed approach achieved a high range of sensitivity between 92% and 95% at the fusion level (late 92%, intermediate 95% and early 92%). Additionally, a contribution analysis revealed that in late fusion, a weighted 70% for gait mode and 30% for smooth eye movement were sufficient for multimodal classification. These findings are linked to the fact that the majority of patients (16 out of 19) fall under the Hoehn and Yahr scale, where bilateral gait impairment and postural instability are notably prevalent. The introduced method can learn independent geometrical branches with F1-s (97%) that can interpolate and generate new covariance patterns from the learning process. From this covariance representation, the descriptors are very compact (with an input size of 625 and an output size of 256 (1 BiRe)), which allows effective learning with few samples, a critical point in medical applications. The proposed approach, however, should be analyzed in a larger cohort of patients to establish statistical significance among the affected population. Also, it is important as a perspective, to design mechanisms that output disease stages according to

observational scales, allowing them to be used in disease progression. An extension of the use of learning new representations to classify motor impairments in Parkinson's disease, combining other modalities such as voice and facial expressions, was proposed in Appendix B. For further details, please refer to the Appendix B section.

4. A MULTIMODAL GAIT AND OCULAR GEOMETRIC REPRESENTATION TO GENERATE A PARKINSON PROGRESSION REPORT.

4.1. ABSTRACT

Parkinson's disease (PD) is a progressive neurological condition, primarily associated with a deficiency in dopamine neurotransmitters, generating premotor, motor control, emotional, and executive dysfunctions. The characterization and PD diagnosis are principally based on the analysis of observed motion alterations, such as slowed movements (bradykinesia), wrong posture, and freezing of gait. Computational methods to support some of these observations remain limited to distinguishing between PD patients and control patients on the basis of protocols at advanced stages, according to current clinical PD guidelines. This work introduces a multi-item PD progression support compliant with both the modified H&Y scale and the MDS-UPDRS part III scale, which is based on a multimodal geometric representation that combines gait and oculomotor markerless video sequences. The proposed representation aims to evaluate gait autonomy, posture impairment, gait freezing, gait bradykinesia, bilateral gait impairment and non-included observations such as ocular bradykinesia. To do so, a 3D convolutional neural network (CNN) first captures spatiotemporal PD patterns. Then, a Riemannian network learns second order relationships among observed patterns, which are further fused at early or intermediate stages to output multi-item PD predictions. In a retrospective study with 13 control subjects and 19 diagnosed PD patients, the proposed approach (early fusion) achieved F1-scores of 95% for bilateral impairment, 94% for gait autonomy, 81% for freezing of gait, 92% for wrong posture, 91% for gait bradykinesia and 94% for ocular bradykinesia in PD patients. The proposed approach is a promising tool to support routine and standard clinical PD analysis.

The partial content of this work has been accepted and published in [104].

4.2. PROPOSED APPROACH

This work presents an end-to-end multimodal and geometric architecture, that supports a multiple item PD prediction, allowing support clinical PD characterization. The proposed approach starts from a convolutional 3D representation, and the resulting deep features are summarized in symmetric positive definite (SPD) matrices to deal with potential scarcity in training data. The SPD matrices are then processed by geometric layers whose function is to fuse PD modalities. The fusion of both modalities occurs at either the early or intermediate level (Figure 13). At the early fusion level, a unique SPD matrix is computed from the two motion modes, while at the intermediate level, an SPD matrix is computed for each modality, from which are learnt new geometric representations, that are subsequently fused in the dense layers to output multiple predictions to support motor scales. The details of the proposed approach are described in the next subsections.

4.2.1. 3D Convolutional Representation. During the initial phase, the 3D-CNN builds a latent representation of spatiotemporal patterns, enabling the emergence of relevant gait and ocular movement patterns across several hierarchical levels. Given a video sequence $\boldsymbol{S}$, each layer $\ell \in \{1, \ldots, L\}$ computes image transformations sequentially, using a bank of K_ℓ learnable 3D convolution filters Ψ_k^ℓ and scalar biases b_k^ℓ, commonly followed by non-linear activation functions $\mathbf{a}_k^\ell$ and pooling operation Π_ℓ, projecting in this way the input information onto a set of deep features by the following sequential process:

$$\Phi_k^\ell = \Pi_\ell \left(\mathbf{a}_k^\ell \left(\mathbf{\Phi}^{\ell-1} * \Psi_k^\ell + b_k^\ell \right) \right)$$

---

[104] John ARCHILA et al. “A multimodal gait and ocular geometric representation to generate a Parkinson progression report”. In: *Engineering Applications of Artificial Intelligence* 160 (2025), p. 111834. DOI: https://doi.org/10.1016/j.engappai.2025.111834.

with $\mathbf{\Phi}^{\ell} = \left\{\Phi_k^{\ell}\right\}_{k=1}^{K_\ell}$, $\mathbf{\Phi}^0 = \mathbf{S}$, and the resulting bank of 3D representations $\mathbf{\Phi}^L$ is the final result of the feature extraction process of the 3D-CNN, yielding a three-dimensional feature bank $\mathbf{\Phi^L} = \{\boldsymbol{F}^i\}_{i=1}^{N}$ (with $N = K_L$), which encompasses a set of attributes. This bank of features is calculated for each motion mode, allowing to capture Parkinsonian spatiotemporal patterns encountered in gait and ocular smooth motion.

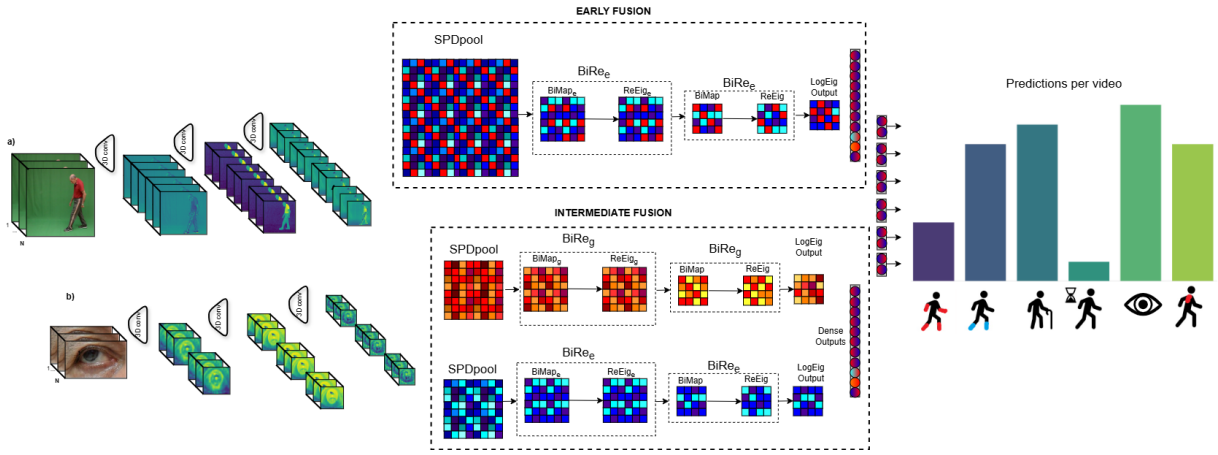

Figure 13. End-to-end multimodal and geometric architectures, that support multiple-item PD prediction. Two approaches were performed. Early Fusion: Gait and ocular smooth motion modalities are processed by the 3D-CNN and fused early through SPD pooling. Then further compact representations are learned by two BiRe blocks and mapping to the Euclidean space is performed to predict a symptom report of patients. Intermediate Fusion: Gait and Ocular smooth motion modalities are independently processed by their own 3D-CNN and Riemannian networks. Then their respective mappings to the Euclidean space are concatenated and input to the dense layers whose outputs form the predictions for the symptom report.

4.2.2. Geometrical fusion level

Early Riemannian fusion. The spatiotemporal patterns associated with Parkinson's disease, derived from ocular movement modalities (initial impairments) and gait modalities (intermediate and advanced impairments), can be summarized in a single matrix that calculates the covariance between the output volumes of both modalities. In this context, a Riemannian signature is obtained for each patient, taking into account impairments at different stages

of the disease. Let $\mathbf{\Phi_L^e}$ be the eye motion feature bank $\mathbf{\Phi_L^e} = \{\boldsymbol{F}^i\}_{i=1}^N$ and, be $\mathbf{\Phi_L^g}$ the gait feature bank $\mathbf{\Phi_L^g} = \{\boldsymbol{G}^i\}_{i=1}^N$, the first fusion alternative aims to retain and enhance the information captured by the 3D convolutional deep features of both modalities. To do this, a covariance-like symmetric positive definite (SPD) matrix is calculated. First, the $WH \times 2N$ rectangular matrix is built, whose column vectors are formed by the flattened feature maps of eye and gait: $\boldsymbol{R} = \big[vec(\boldsymbol{F}^1), \cdots, vec(\boldsymbol{F}^N), vec(\boldsymbol{G}^1), \cdots, vec(\boldsymbol{G}^N)\big]$, where $W \times H$ are the spatial dimensions of all feature maps. Then the $2N \times 2N$ SPD pooled matrix, input of the Riemannian network, is defined as: $\boldsymbol{C^0} = \boldsymbol{R}^T\boldsymbol{R}$. This layer highlights the relationships between features, thus strengthening the connections among patterns within each feature map of both modalities associated with PD.

To perform the geometrical learning, the SPD matrix $\boldsymbol{C^0}$ is transformed into a more compact SPD matrix using a sequence of Riemannian blocks, which are composed of two layers, the bilinear mapping (BiMap) layer and the Rectification Eigenvalues (ReEig) layer. Here, the BiMap computes: $\mathcal{C}^\ell = W_\ell \mathcal{C}^{\ell-1} W_\ell^T$, with $W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ is the weight matrix transformation ($d_\ell << d_{\ell-1}$) [105]. After a BiMap layer, the Riemannian block is completed with a ReEig layer, which, similarly to the ReLU function for Euclidean networks, applies a rectification threshold, by computing $\mathcal{C}^{\ell+1} = U^\ell \max(\varepsilon I, \Sigma^\ell)(U^\ell)^T$, where (Σ^ℓ, U^ℓ) respectively represent the diagonal matrix of eigenvalues and the matrix of eigenvectors of $\mathcal{C}^\ell$, and ε is a scalar threshold. At the end of the Riemannian module, a Riemannian logarithm map $\mathbf{log}$ is computed to map the resulting Riemannian descriptor back onto the Euclidean space [105]. Finally, this embedding is vectorized, concatenated, and thereafter processed by two dense layers $\mathcal{D}$. In the output layer, a softmax layers computes the probability of predictions of different PD symptoms and motor affectations to differentiate between PD patients and control subjects. The early fusion approach is illustrated in Figure 13.

---

[105] Zhiwu HUANG and Luc VAN GOOL. "A Riemannian Network for SPD Matrix Learning". In: *Association for the Advancement of Artificial Intelligence (AAAI)*. 2017.

Intermediate fusion from independent geometrical branches. Alternatively, gait and ocular movement can be identified with two types of motion: one with coarse granularity, involving limb coordination and balance of the trunk and head, and the other with fine granularity, which is based on smooth ocular movements in the vertical and horizontal directions. In this context, it is of particular interest that each movement independently learns new geometric representations within its respective branch. Finally, a subsequent fusion occurs in the Euclidean space through the dense layer. In the intermediate fusion, for each feature bank of gait $\mathbf{\Phi_L^g} = \{\boldsymbol{G}^i\}_{i=1}^N$, and of ocular movement $\mathbf{\Phi_L^e} = \{\boldsymbol{F}^i\}_{i=1}^N$, their respective $WH \times N$ rectangular matrices $\boldsymbol{R_g} = \left[vec(\boldsymbol{G}^1), \cdots, vec(\boldsymbol{G}^N)\right]$ and $\boldsymbol{R_e} = \left[vec(\boldsymbol{F}^1), \cdots, vec(\boldsymbol{F}^N)\right]$ are built. Each modality calculates its own $N \times N$ SPD pooling matrix $\boldsymbol{C_g^0} = \boldsymbol{R_g^T R_g}$ and $\boldsymbol{C_e^0} = \boldsymbol{R_e^T R_e}$. Next, for each branch, the geometric layers Bimap, ReEig, and the logarithmic mapping are computed as detailed in the previous subsection. Finally, these embeddings are vectorized and concatenated, to be processed by the same two dense layers $\mathcal{D}$. In the output layers, a Softmax layer computes the probability of discrimination between 6 motor affectations from control and PD patients. The intermediate fusion approach is illustrated in Figure 13.

4.2.3. Motor report prediction: A multitask learning. A main interest of this work is to develop a computational tool with the ability to support PD characterization, following standard scales under cardinal, postural, balance and mobility symptoms. Hence, the final geometrical layer C^L is mapped to a Euclidean space $\mathbf{log}(C^L)$ allowing operation over standard deep operations. In both early and intermediate representations, the respective $\mathbf{log}(C^L)$ is further processed by a dense layer (D). Then, multitask learning is herein defined by computing six independent predictions, taking as input the D layer. The global motor loss L is defined as:

$$L = \underbrace{\gamma_1 L_{gb} + \gamma_2 L_{ob}}_{\text{Cardinal}} + \underbrace{\gamma_3 L_{wp} + \gamma_4 L_{bg}}_{\text{Posture and Balance}} + \underbrace{\gamma_5 L_{fg} + \gamma_6 L_{ga}}_{\text{Mobility}}$$

where γ_i is an importance weight for each considered motor symptom. The cardinal symp-

toms the gait bradykinesia (L_{gb}) and ocular bradykinesia (L_{ob}), both of which are associated with slowness of movement, a reduction in amplitude (hypokinesia), the absence of movement (akinesia), or a progressive and repetitive reduction in motion. The motor loss also considers changes in posture and balance. Specifically, here, we consider labels associated with bilateral gait impairment (L_{bg}) and the wrong posture (L_{wp}) that involve observations associated with postural distortion (forward flexion of the trunk and symmetrical disruption of walking patterns decreasing velocity and increasing shuffling). Additionally, the freezing of gait L_{fg} and gait autonomy L_{ga} were considered mobility symptoms. These observations aim to characterize impaired autonomy, typically for starting or sustaining locomotion. All loss functions are based on cross-entropy.

The rationale for using six independent predictions is to drive the model to learn specific features for each impairment. This is particularly important in the context of Parkinson's disease, as a patient may exhibit ocular bradykinesia without necessarily having gait impairment or an incorrect posture. As an additional output, a heatmap is presented related to the six predictions, taking into account the affected laterality of each patient. For example, in Figure 14, the predictions of one PD patient and one control subject are shown. The heatmaps may help specialist identify the predominant impairments, creating a personalized profile for each subject. The rows represent the patient's impairments, listed from top to bottom as: bilateral impairment, freezing, gait autonomy, bradykinesia in gait, ocular bradykinesia, and poor posture, whereas the columns represent the 8 predictions obtained for each subject: odd numbers (right side for gait and right eye) and even numbers (left side for gait and left eye). In this example for the PD patient, greater impairment was observed on the left side.

4.3. EXPERIMENTAL SETUP

4.3.1. Multimodal data. The study involved 13 control subjects (average age 72.2 ± 6.1 years) and 19 PD patients (average age 72.3 ± 7.4 years). The PD progression severity was supported by a expert neurologist. In total, there are 5 PD patients in stage 1, 5 PD patients

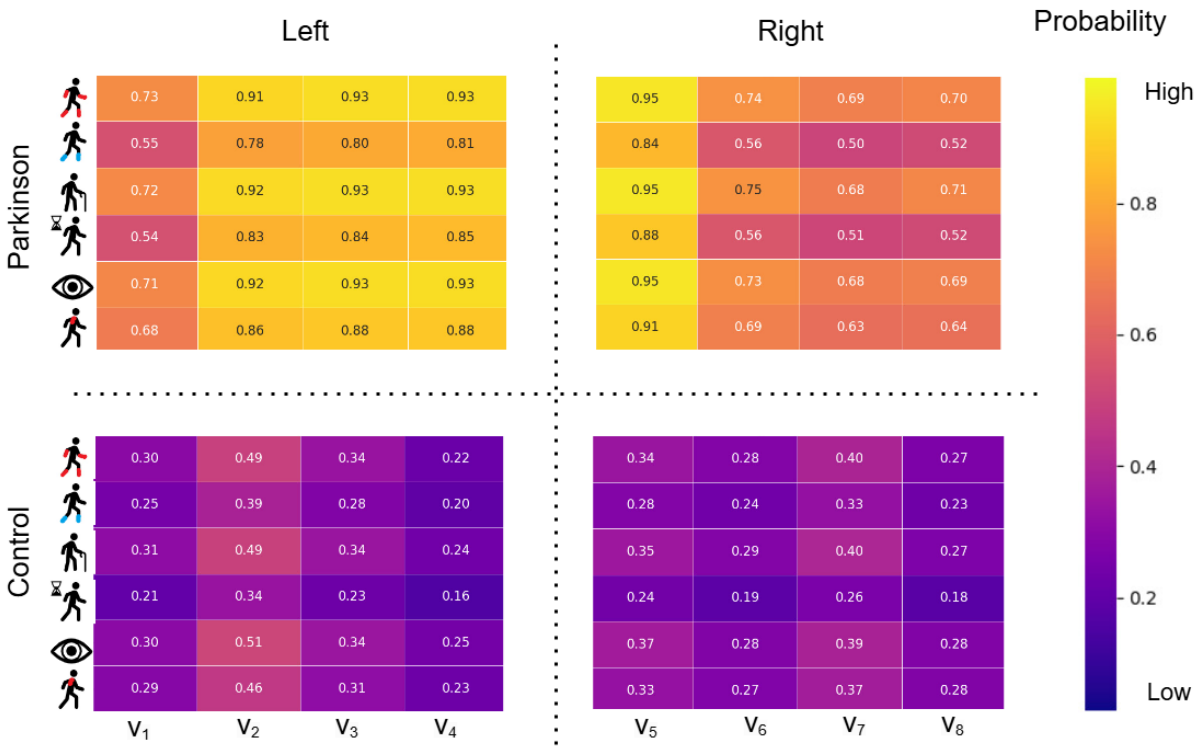

Figure 14. Graphical representation of the report. Above: motor report of each of the impairments of a patient on the right side and on the left side. Below: motor report of a control subject with right-side and left-side impairments. The rows represent the patient's impairments, listed from top to bottom as follows: bilateral impairment, freezing, gait autonomy, bradykinesia in gait, ocular bradykinesia, and poor posture. The columns represent the 8 multimodal predictions obtained for a subject (one per video).

in stage 1.5, 5 PD patients in stage 2, and 4 PD patients in stage 3 using the modified H&Y scale. The specialist conducted an evaluation of six motor impairments, including an item that predicts bilateral gait impairment based on observations from the modified H&Y scale. Furthermore, four items according to the MDS-UPDRS Part III scale were also evaluated (gait autonomy, freezing, posture, bradykinesia) and an additional and complementary detection of ocular bradykinesia was carried out. During the study, the participants were instructed to adhere to the following movement guidelines:

- **Gait.** The participants were invited to walk in a straight line while the camera captured them from a sagittal view. Locomotion was recorded over a 5-meter walk, with an average video duration of 5 seconds per recording. Each video had a spatial resolution of 520 × 520 pixels and a temporal resolution of 60 frames per second (fps). Figure 1 illustrates the postural configuration during the gait exercise, realized in front of a uniform green background.

- **Smooth ocular motion.** In this scenario, patients were instructed to maintain their gaze on a spotlight projected onto a screen with a dark background. The monitor's height was adjusted to align the center of the screen with the center of the pupillary plane, as depicted in Figure 1. The motion of the spotlight was controlled both horizontally (from right to left and vice versa) and vertically (from top to bottom and vice versa). The recorded video was subsequently manually cropped to a region of interest around the eye, with dimensions of 210 × 140 pixels.

Sixteen videos were obtained for each participant, encompassing both gait and ocular smooth motion exercises. For both modalities, the same camera, a conventional Nikon D3200, was utilized, providing a spatial resolution of 1280 × 720. Consequently, the entire dataset for this study comprises a total of 512 videos. This study received approval from the Ethics Committee, and written informed consent was obtained from each participant.

The proposed approach was validated under 5-fold cross-validation method taking at each iteration, 26 subjects for model training and 6 subjects for testing. For these experiments, the Parkinsonian patients (and the control participants) who were correctly identified were considered true positives (TP) and true negatives (TN), respectively. A set of metrics was subsequently utilized to comprehensively evaluate the performance of the method in its different configurations. The metrics calculated here are accuracy (acc), sensitivity (sen), precision ($prec$), specificity ($spec$) and the F1-score ($F1$-s). In this way, each metric is calculated for each binary label of classification report. The labels were based on predominant motor impairments: a laterality gait affectation (H&Y scale) and specific items of the MDS-UPDRS

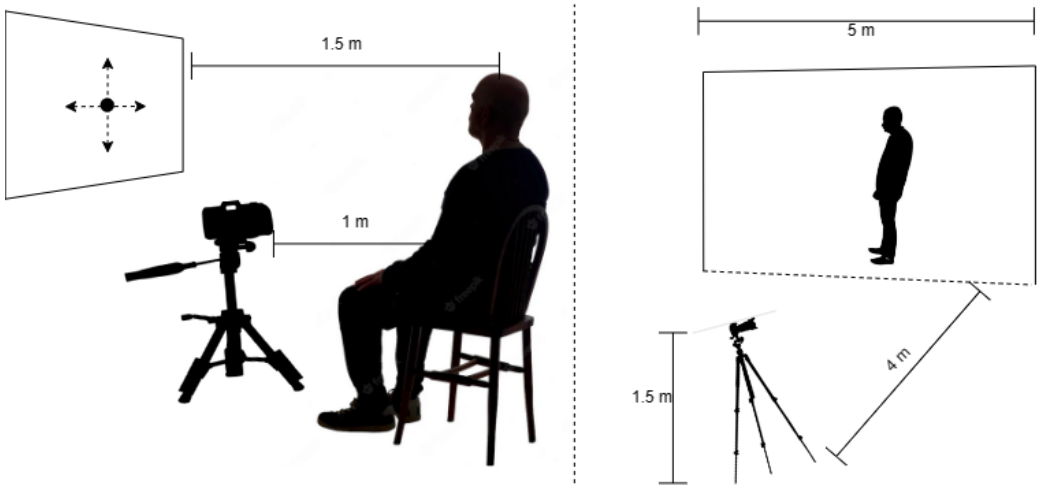

Figure 15. Gait and ocular smooth motion acquisition setups for our markerless video dataset.

part III scale (gait autonomy, freezing of gait, posture impairment, and gait bradykinesia). Additionally, non-standardized observations such as ocular bradykinesia were quantified.

4.3.2. Parameter Tuning. The following parameters were implemented:

- **Training Characteristics.** The evaluation was carried out using a subject-wise 5 fold cross-validation scheme. The hyperparameter were fixed for all experiments before the cross-validation scheme. In each iteration, the model was trained using data from 26 participants and evaluated on the excluded 6 participant. The model was trained with an early stopping criterion based on the minimum number of epochs required for loss reduction. To prevent overfitting, a dropout and a regularization technique were applied to the 3D convolutional layers. Additionally, data augmentation techniques, specifically horizontal flipping, were implemented during training. The hyperparameter were empirically selected before k-fold validation. The training and evaluation were conducted under a subject-wise leave-one-patient-out cross-validation scheme to ensure and preserve independence between training and testing sets

- **3D Convolutional Representation.** Experiments for each level of fusion were con-

ducted with 1, 2 or 3 3D convolutional layers, with 32, 64, and 128. output channels, respectively.

- **Geometrical Representation.** Experiments for each level of fusion were conducted with 1, 2 BiRe layers, with dimensions of 64×64 and 32×32. output matrices, respectively.

- **Variation of the SPD Pooling Matrix Dimension.** The output channels of the last 3D convolutional layer (N channels for gait and M channels for ocular motion) are related to the input matrix of the geometric phase. In early fusion, the input SPD matrix, which considers the combination of both modalities, has dimensions of $(N + M) \times (N + M)$. In intermediate fusion, the dimensions of the 2 matrices are $N \times N$, and $M \times M$ since SPD pooling is performed independently for each branch. Consequently, depending on the depth of the 3D-CNN, the resulting matrix from early fusion can have dimensions of (64, 64), (128, 128), or (256, 256). For intermediate fusion, the SPD pool matrix for each branch can have dimensions of (32, 32), (64, 64), or (128, 128).

4.4. RESULTS

This work introduced a geometric multimodal architecture to generate a motor scale report about Parkinsonism affectations observed from markerless videos of oculomotor tasks and gait. Hence, we consider two fusion alternatives of eye and gait observations, analyze geometric components, and extend the analysis to multiple stages of the disease. The next subsections describe the results of these considerations.

4.4.1. Results from Early fusion. Firstly, we evaluated the convolutional (conv3D) representation to form SPD descriptors. Table 13 summarizes the experiments using 1 and 3

conv3D layers both considering 1 BiRe layer. The results consistently demonstrate that using three layers yields better performance for Ocular and Gait Bradykinesia, Wrong posture, Bilateral impairment and Gait autonomy (average accuracy and F1-score gain of 6%, 5%, respectively). In contrast, Freezing shows gain of 3% in F1-s with experiment of one 3D conv layer.

Early fusion 3D conv layers		Cardinal Symptoms		Posture and balance symptoms		Mobility symptoms	
Layers	Metrics	Bradykinesia Ocular	Bradykinesia in Gait	Wrong Posture	Bilateral Afectation (H&Y)	Freezing of Gait	Autonomy in Gait
1 layer 3D conv	Prec	0.85 ± 0.07	0.86 ± 0.17	0.85 ± 0.13	0.82 ± 0.07	0.85 ± 0.19	0.81 ± 0.07
	Sen	0.86 ± 0.14	0.89 ± 0.08	0.92 ± 0.06	0.94 ± 0.06	0.90 ± 0.11	0.94 ± 0.06
	Spec	0.73 ± 0.18	0.75 ± 0.32	0.66 ± 0.37	0.60 ± 0.33	0.76 ± 0.38	0.60 ± 0.33
	Acc	0.81 ± 0.12	0.83 ± 0.13	0.83 ± 0.11	0.84 ± 0.07	0.83 ± 0.17	0.83 ± 0.07
	F1-s	0.85 ± 0.09	0.85 ± 0.08	0.87 ± 0.06	0.87 ± 0.05	$\mathbf{0.85 \pm 0.10}$	0.87 ± 0.05
3 layers 3D conv	Prec	0.90 ± 0.13	0.88 ± 0.14	0.91 ± 0.10	0.90 ± 0.13	0.85 ± 0.17	0.90 ± 0.13
	Sen	0.96 ± 0.06	0.87 ± 0.07	0.95 ± 0.06	0.96 ± 0.06	0.82 ± 0.15	0.95 ± 0.07
	Spec	0.83 ± 0.21	0.86 ± 0.17	0.86 ± 0.16	0.82 ± 0.20	0.87 ± 0.15	0.84 ± 0.20
	Acc	$\mathbf{0.90 \pm 0.10}$	$\mathbf{0.86 \pm 0.11}$	$\mathbf{0.91 \pm 0.10}$	$\mathbf{0.90 \pm 0.10}$	0.84 ± 0.12	$\mathbf{0.89 \pm 0.11}$
	F1-s	$\mathbf{0.92 \pm 0.08}$	$\mathbf{0.87 \pm 0.10}$	$\mathbf{0.93 \pm 0.08}$	$\mathbf{0.92 \pm 0.07}$	0.82 ± 0.12	$\mathbf{0.92 \pm 0.08}$

Table 13. Performance metrics for the two best experiments associated with *Early* Fusion models (1 or 3 3D convolutional layers).

Secondly, we evaluated the geometric learning components that provide SPD patterns in the multiclassification report. Table 14 summarizes the experiments using 1 and 2 BiRe layers, being consistently better the use of two BiRe layers for Ocular and Gait Bradykinesia, Bilateral impairment, and Gait autonomy (with average accuracy and F1-score gain of 4% and 3%, respectively). Freezing shows similar metrics across both methods. In contrast, Wrong posture shows gain of 1% in F1-s on the experiment with 1 BiRe layer. In such a sense, the configuration the configuration with 3 3D CNN layers and 2 BiRe blocks results the most consistent representation in early fusion.

Early fusion BiRe layers		Cardinal Symptoms		Posture and balance symptoms		Gait and mobility symptoms	
Layers	Metrics	Bradykinesia Ocular	Bradykinesia in Gait	Wrong Posture	Bilateral Afectation (H&Y)	Freezing of Gait	Autonomy in Gait
2 layers BiRe 3 layers 3D conv	Prec	0.92 ± 0.14	0.89 ± 0.13	0.90 ± 0.12	0.93 ± 0.14	0.84 ± 0.20	0.93 ± 0.14
	Sen	0.97 ± 0.05	0.94 ± 0.06	0.96 ± 0.04	0.98 ± 0.04	0.82 ± 0.14	0.97 ± 0.05
	Spec	0.89 ± 0.21	0.86 ± 0.17	0.85 ± 0.18	0.89 ± 0.21	0.85 ± 0.17	0.89 ± 0.21
	Acc	$\mathbf{0.93 \pm 0.10}$	$\mathbf{0.90 \pm 0.08}$	0.91 ± 0.10	$\mathbf{0.93 \pm 0.10}$	0.83 ± 0.13	$\mathbf{0.93 \pm 0.10}$
	F1-s	$\mathbf{0.94 \pm 0.08}$	$\mathbf{0.91 \pm 0.07}$	0.92 ± 0.08	$\mathbf{0.95 \pm 0.08}$	0.81 ± 0.13	$\mathbf{0.94 \pm 0.08}$
1 layer BiRe 3 layers 3D conv	Prec	0.90 ± 0.13	0.88 ± 0.14	0.91 ± 0.10	0.90 ± 0.13	0.85 ± 0.17	0.90 ± 0.13
	Sen	0.96 ± 0.06	0.87 ± 0.07	0.95 ± 0.06	0.96 ± 0.06	0.82 ± 0.15	0.95 ± 0.07
	Spec	0.83 ± 0.21	0.86 ± 0.17	0.86 ± 0.16	0.82 ± 0.20	0.87 ± 0.15	0.84 ± 0.20
	Acc	0.90 ± 0.10	0.86 ± 0.11	0.91 ± 0.10	0.90 ± 0.10	0.84 ± 0.12	0.89 ± 0.11
	F1-s	0.92 ± 0.08	0.87 ± 0.10	$\mathbf{0.93 \pm 0.08}$	0.92 ± 0.07	0.82 ± 0.12	0.92 ± 0.08

Table 14. Performance metrics for the two experiments associated with *Early* Fusion models (1 or 2 BiRe layers with three 3D convolutional layers).

4.4.2. Results from Intermediate fusion. For intermediate fusion representation, we also first evaluated the convolutional representation to form SPD descriptors. Table 15 summarized the achieved results at different items, evidencing a superior performance for representation with three 3D convolutional layers for Gait and Ocular Bradykinesia, Wrong posture, Bilateral impairment and Freezing (average accuracy and F1-score gain of 4%, 3%, respectively). In contrast, one 3D convolutional layer was more consistent for the Wrong posture (gain of 2% and 5% in accuracy and F1-score). All experiments included 1 BiRe layer.

Intermediate fusion Number of 3D conv layers		Cardinal Symptoms		Posture and balance symptoms		Gait and mobility symptoms	
Layers	**Metrics**	**Bradykinesia Ocular**	**Bradykinesia in Gait**	**Wrong Posture**	**Bilateral Afectation (H&Y)**	**Freezing of Gait**	**Autonomy in Gait**
2 layers 3D conv	**Prec**	0.84 ± 0.13	0.83 ± 0.20	0.83 ± 0.17	0.80 ± 0.13	0.80 ± 0.24	0.84 ± 0.13
	Sen	0.91 ± 0.16	0.84 ± 0.11	0.95 ± 0.10	0.89 ± 0.19	0.71 ± 0.27	0.90 ± 0.16
	Spec	0.71 ± 0.23	0.75 ± 0.3	0.68 ± 0.32	0.62 ± 0.26	0.86 ± 0.17	0.72 ± 0.23
	Acc	0.82 ± 0.15	0.79 ± 0.15	$\mathbf{0.83 \pm 0.18}$	0.77 ± 0.18	0.77 ± 0.16	0.82 ± 0.14
	F1-s	0.86 ± 0.12	0.81 ± 0.12	$\mathbf{0.88 \pm 0.13}$	0.82 ± 0.14	0.72 ± 0.16	0.86 ± 0.11
3 layers 3D conv	**Prec**	0.89 ± 0.14	0.86 ± 0.16	0.84 ± 0.17	0.88 ± 0.14	0.81 ± 0.22	0.89 ± 0.14
	Sen	0.90 ± 0.17	0.83 ± 0.12	0.83 ± 0.21	0.88 ± 0.17	0.78 ± 0.24	0.89 ± 0.18
	Spec	0.81 ± 0.23	0.84 ± 0.21	0.80 ± 0.21	0.80 ± 0.23	0.86 ± 0.17	0.81 ± 0.23
	Acc	0.84 ± 0.15	$\mathbf{0.84 \pm 0.13}$	0.81 ± 0.20	$\mathbf{0.83 \pm 0.15}$	$\mathbf{0.80 \pm 0.16}$	0.84 ± 0.15
	F1-s	0.87 ± 0.12	$\mathbf{0.84 \pm 0.12}$	0.83 ± 0.19	$\mathbf{0.86 \pm 0.12}$	$\mathbf{0.77 \pm 0.21}$	0.87 ± 0.12

Table 15. Performance metrics for the two best experiments associated with *Intermediate* Fusion models (2 or 3 3D convolutional layers).

Regarding geometrical learning components, Table 16 summarizes the achieved results in the intermediate representation. Interestingly, adding one more geometric layer shows an average gain of 5% in accuracy and F1-score for all impairments. In the context of intermediate fusion the best configuration was achieved with three 3D convolutional layers and two BiRe blocks.

Intermediate fusion BiRe layers		Cardinal Symptoms		Posture and balance symptoms		Gait and mobility symptoms	
Layers	**Metrics**	**Bradykinesia Ocular**	**Bradykinesia in Gait**	**Wrong Posture**	**Bilateral Afectation (H&Y)**	**Freezing of Gait**	**Autonomy in Gait**
2 layer BiRe 3 layers 3D conv	**Prec**	0.90 ± 0.13	0.88 ± 0.15	0.92 ± 0.10	0.91 ± 0.13	0.84 ± 0.19	0.90 ± 0.12
	Sen	0.96 ± 0.08	0.88 ± 0.15	0.91 ± 0.11	0.95 ± 0.08	0.79 ± 0.15	0.90 ± 0.12
	Spec	0.83 ± 0.20	0.85 ± 0.18	0.88 ± 0.15	0.85 ± 0.19	0.86 ± 0.17	0.83 ± 0.19
	Acc	$\mathbf{0.89 \pm 0.10}$	$\mathbf{0.87 \pm 0.12}$	$\mathbf{0.90 \pm 0.11}$	$\mathbf{0.90 \pm 0.09}$	$\mathbf{0.82 \pm 0.12}$	$\mathbf{0.90 \pm 0.09}$
	F1-s	$\mathbf{0.92 \pm 0.08}$	$\mathbf{0.88 \pm 0.11}$	$\mathbf{0.91 \pm 0.09}$	$\mathbf{0.93 \pm 0.07}$	$\mathbf{0.80 \pm 0.12}$	$\mathbf{0.92 \pm 0.07}$
1 layer BiRe 3 layers 3D conv	**Prec**	0.89 ± 0.14	0.86 ± 0.16	0.84 ± 0.17	0.88 ± 0.14	0.81 ± 0.22	0.89 ± 0.14
	Sen	0.90 ± 0.17	0.83 ± 0.12	0.83 ± 0.21	0.88 ± 0.17	0.78 ± 0.24	0.89 ± 0.18
	Spec	0.81 ± 0.23	0.84 ± 0.21	0.80 ± 0.21	0.80 ± 0.23	0.86 ± 0.17	0.81 ± 0.23
	Acc	0.84 ± 0.15	0.84 ± 0.13	0.81 ± 0.20	0.83 ± 0.15	0.80 ± 0.16	0.84 ± 0.15
	F1-s	0.87 ± 0.12	0.84 ± 0.12	0.83 ± 0.19	0.86 ± 0.12	0.77 ± 0.21	0.87 ± 0.12

Table 16. Performance metrics for the two experiments associated with *Intermediate* Fusion models (1 or 2 BiRe layers with three 3D convolutional layers).

4.4.3. Multimodal and geometric contribution. During validation, the proposed multimodal approach was also compared regarding unimodal version using three 3D convolutional layers with one or two BiRe block. In unimodal approaches, a label prediction is made for eye movement (ocular bradykinesia) and five labels are predicted based on the gait modality. As reported in Table 17, standing out scores were achieved for the sensitivity of ocular movement as well as for predicting bilateral impairment (85%) and gait autonomy (83%). However, the other metrics report a remarked reduction as observed in the F1-score. For the unimodal gait model, there is not significative improvements to include additional BiRe blocks, a fact associated to information with low variance from only one modality, which limit the learning of more complex patterns.

Contrary, fusing the modalities yields improvements over the best unimodal results. For instance, in the intermediate fusion, the F1 score for the motor impairments shows an average improvement of 13%, while accuracy increases by 12%. Early fusion presents slightly higher gains, with improvements of 15% and 13%, respectively demonstrating that fusing in the Riemannian manifold yields the best results. In Table 17, it is also observed that the difference in performance between unimodal and multimodal models is more pronounced for mobility symptoms such as Autonomy in Gait and Freezing of Gait, where multimodal models outperform unimodal ones. The F1-scores across table indicate consistent improvements with multimodal fusion strategies, particularly Early Fusion, which demonstrates the strongest overall performance for predicting motor impairments.

In addition, we compared the proposed approach with a 3D convolutional architecture. The outputs of the two branches are then fused through concatenation and a dense layer for the subsequent prediction of the six items. Table 18 summarizes the results obtained using this baseline model, compared to the Riemannian early fusion approach. The contribution of the geometric phase is evident, increasing most of the various metrics by 40% to 60%. However, the purely 3D convolutional fusion also achieves outstanding sensitivity for some items, such as Ocular Bradykinesia and Walking autonomy 80%, and highlights specificity in

Modalities		Cardinal Symptoms		Posture and balance symptoms		Gait and mobility symptoms	
Configuration	Metrics	Bradykinesia Ocular	Bradykinesia in Gait	Wrong Posture	Bilateral Afectation (H&Y)	Freezing of Gait	Autonomy in Gait
Unimodal 3 3D Conv-1 BiRe	Prec	0.78 ± 0.13	0.81 ± 0.22	0.78 ± 0.20	0.78 ± 0.14	0.78 ± 0.26	0.78 ± 0.16
	Sen	$\mathbf{1}$	0.76 ± 0.14	0.81 ± 0.19	0.85 ± 0.18	0.72 ± 0.26	0.83 ± 0.17
	Spec	0.65 ± 0.15	0.74 ± 0.32	0.6 ± 0.36	0.59 ± 0.33	0.80 ± 0.27	0.59 ± 0.33
	Acc	0.85 ± 0.12	0.75 ± 0.18	0.72 ± 0.19	0.73 ± 0.16	0.74 ± 0.19	0.72 ± 0.16
	F1-s	0.87 ± 0.08	0.76 ± 0.15	0.77 ± 0.15	0.79 ± 0.11	0.70 ± 0.21	0.78 ± 0.11
Unimodal 3 3D Conv-2 BiRe	Prec	0.81 ± 0.15	0.83 ± 0.15	0.77 ± 0.17	0.82 ± 0.15	0.76 ± 0.26	0.83 ± 0.15
	Sen	0.82 ± 0.23	0.81 ± 0.15	0.80 ± 0.25	0.83 ± 0.21	0.71 ± 0.27	0.82 ± 0.22
	Spec	0.64 ± 0.30	0.76 ± 0.24	0.63 ± 0.29	0.67 ± 0.30	0.77 ± 0.27	0.68 ± 0.31
	Acc	0.73 ± 0.15	0.79 ± 0.10	0.73 ± 0.18	0.75 ± 0.13	0.72 ± 0.18	0.74 ± 0.14
	F1-s	0.77 ± 0.12	0.80 ± 0.09	0.76 ± 0.17	0.79 ± 0.10	0.69 ± 0.22	0.78 ± 0.11
Intermediate Fusion 3 3D conv-2 BiRes	Prec	0.90 ± 0.13	0.88 ± 0.15	0.92 ± 0.10	0.91 ± 0.13	0.84 ± 0.19	0.90 ± 0.12
	Sen	0.96 ± 0.08	0.88 ± 0.15	0.91 ± 0.11	0.95 ± 0.08	0.79 ± 0.15	0.90 ± 0.12
	Spec	0.83 ± 0.20	0.85 ± 0.18	0.88 ± 0.15	0.85 ± 0.19	0.86 ± 0.17	0.83 ± 0.19
	Acc	0.89 ± 0.10	0.87 ± 0.12	0.90 ± 0.11	0.90 ± 0.09	0.82 ± 0.12	0.90 ± 0.09
	F1-s	0.92 ± 0.08	0.88 ± 0.11	0.91 ± 0.09	0.93 ± 0.07	0.80 ± 0.12	0.92 ± 0.07
Early Fusion 3 3D Conv-2 BiRes	Prec	0.92 ± 0.14	0.89 ± 0.13	0.90 ± 0.12	0.93 ± 0.14	0.84 ± 0.20	0.93 ± 0.14
	Sen	0.97 ± 0.05	0.94 ± 0.06	0.96 ± 0.04	0.98 ± 0.04	0.82 ± 0.14	0.97 ± 0.05
	Spec	0.89 ± 0.21	0.86 ± 0.17	0.85 ± 0.18	0.89 ± 0.21	0.85 ± 0.17	0.89 ± 0.21
	Acc	0.93 ± 0.10	0.90 ± 0.08	0.91 ± 0.10	0.93 ± 0.10	0.83 ± 0.13	0.93 ± 0.10
	F1-s	$\mathbf{0.94 \pm 0.08}$	$\mathbf{0.91 \pm 0.07}$	$\mathbf{0.92 \pm 0.08}$	$\mathbf{0.95 \pm 0.08}$	$\mathbf{0.81 \pm 0.13}$	$\mathbf{0.94 \pm 0.08}$

Table 17. Comparison between Unimodal and Multimodal approaches for different item predictions

the Freezing of gait 80%. Additionally, 95% confidence intervals for sensitivity and specificity were calculated using the Clopper–Pearson method within the early fusion approach for the predictions of each motor item. Firstly, the narrowest sensitivity intervals were observed for ocular bradykinesia and gait impairment ([0.82, 1]), followed by slightly wider intervals for gait bradykinesia, wrong posture, and autonomy of gait ([0.74, 1]), and the widest interval was associated with freezing of gait ([0.60, 0.97]). Secondly, the narrowest specificity intervals across all motor items were observed ([0.64, 1]), except for freezing of gait, which presented a broader interval ([0.55, 1]). These results suggest the importance of using larger datasets in future studies.

The proposed approach was also compared with classical machine learning approaches (including the proposed approach in chapter 2) that have been previously used for Parkinson's disease classification [47, 106, 107]. To this end, pretrained deep features were extracted, and the Riemannian mean was computed for each video descriptor [47]. Subsequently, the ocular

[106] Shaohua WAN et al. "Deep multi-layer perceptron classifier for behavior analysis to estimate Parkinson's disease severity using smartphones". In: *IEEE Access* 6 (2018), pp. 36825–36833.

[107] Wee Shin LIM et al. "An integrated biometric voice and facial features for early detection of Parkinson's disease". In: *npj Parkinson's Disease* 8.1 (2022), p. 145.

movement descriptor was concatenated with the gait descriptor, and the classification task was performed for the six motor impairments using Logistic Regression (LG), Random Forest (RF, our method of chapter 2), Support Vector Machine (SVM), and Multilayer Perceptron (MP). The results are presented in Table 19. Overall, the proposed multimodal geometric method outperforms classical approaches across all motor impairments. In particular, it achieves superior performance in predicting ocular bradykinesia (F1-score of 94%), freezing of gait (F1-score of 81%), and gait independence (F1-score of 94%). While all metrics are higher for the geometric approach, Random Forest (RF) yielded the best results among the classical methods [47, 107]. For the remaining classical approaches, the best performance was associated with the classification of bilateral gait impairment, whereas the lowest results were observed in the classification of freezing of gait, making it the most challenging motor impairment to predict across all methods, including ours.

The use of a 3D convolutional neural network and Riemannian-based geometric modeling is motivated by the spatiotemporal nature of the input data and the need to preserve the non-Euclidean structure of the features extracted from both ocular and gait videos. Traditional machine learning models, although competitive, rely on flattened representations that overlook the intrinsic temporal dynamics and geometry of the video descriptors. In contrast, the 3D CNN is capable of directly capturing spatiotemporal patterns, which are critical for identifying subtle impairments such as ocular bradykinesia and freezing of gait. Furthermore, the geometric component enables a principled treatment of feature distributions via symmetric positive definite (SPD) matrices, enhancing the expressiveness of the representations. The superior results obtained with our method confirm that this architectural design is well suited to model the complex motor patterns involved in Parkinson's disease.

4.4.4. Stratification of the disease. Initially, the normality of the probability distributions for the Parkinson and Control groups was assessed for each motor item. Based on the results of the Shapiro-Wilk test, p-values of ($\rho < 1 \times 10^{-9}$) were observed for the Control group

and p-values of ($\rho < 1 \times 10^{-7}$) for the Parkinson groups across the different motor items. Consequently, we examined the ability of early fusion to distinguish between various stages of the disease. The proposed method was developed and trained for binary PD/Control classification, but provides a PD probability output, allowing us to analyze the behavior of samples labeled with different PD stages. The value of 0.5 corresponds to the default threshold used for classification (probabilities above indicate stronger evidence toward the disease, while probabilities below indicate weaker evidence), facilitating the direct interpretation of the predictions without introducing threshold selection bias. Consequently, we are interested in evaluating stratification effectiveness on the basis of the patients' stage labels. Figure 16 displays the output probabilities for the 6 predictions. a) ocular bradykinesia, b) bradykinesia in gait, c) wrong posture, d) bilateral affectation of gait (H&Y scale), e) freezing, and f) Autonomy. To assess significant differences, we utilized the Kolmogorov–Smirnov (KS) test to validate the effectiveness of the proposed method in distinguishing between classes. The KS test is a non-parametric technique that evaluates the alignment between two distributions. It allows us to check whether the distributions of two samples are significantly different or not.
We analyzed the probability distribution of cardinal symptoms, such as bradykinesia at the ocular level and during gait. The prediction based on ocular bradykinesia was significantly different ($\rho < 10^{-44}$) between the control group and the PD group, as shown in the graph (a) of figure 16. The geometrical model clearly separates both groups, with most control predictions falling below the 0.5 threshold (indicated by the blue dashed line), while predictions for PD patients remain above. For bradykinesia during gait (graph b), significant differences are observed between the normal group (without problems) and all other stages (slight, mild, and moderate) of the disease. However, there were no significant differences between the mild and moderate stages ($\rho > 0.2$).
Postural impairment (graph (c) of the figure 16) presents significant differences across the various stages in the dataset: normal, slight and mild, all with ρ values less than 0.05. For the prediction based on bilateral gait impairment (item (d) from figure 16; modified H&Y scale),

Computational Approaches		Cardinal Symptoms		Posture and balance symptoms		Gait and mobility symptoms	
Configuration	Metrics	Bradykinesia Ocular	Bradykinesia in Gait	Wrong Posture	Bilateral Afectation (H&Y)	Freezing of Gait	Autonomy in Gait
Early Fusion 3 3D conv-2 BiRes	Prec	0.92 ± 0.14	0.89 ± 0.13	0.90 ± 0.12	0.93 ± 0.14	0.84 ± 0.20	0.93 ± 0.14
	Sen	0.97 ± 0.05	0.94 ± 0.06	0.96 ± 0.04	0.98 ± 0.04	0.82 ± 0.14	0.97 ± 0.05
	Spec	0.89 ± 0.21	0.86 ± 0.17	0.85 ± 0.18	0.89 ± 0.21	0.85 ± 0.17	0.89 ± 0.21
	Acc	0.93 ± 0.10	0.90 ± 0.08	0.91 ± 0.10	0.93 ± 0.10	0.83 ± 0.13	0.93 ± 0.10
	F1-s	$\mathbf{0.94 \pm 0.08}$	$\mathbf{0.91 \pm 0.07}$	$\mathbf{0.92 \pm 0.08}$	$\mathbf{0.95 \pm 0.08}$	$\mathbf{0.81 \pm 0.13}$	$\mathbf{0.94 \pm 0.08}$
3 3D conv layers	Prec	0.48 ± 0.26	0.30 ± 0.24	0.33 ± 0.27	0.23 ± 0.29	0.10 ± 0.2	0.48 ± 0.26
	Sen	0.80 ± 0.4	0.60 ± 0.48	0.60 ± 0.48	0.40 ± 0.48	0.20 ± 0.4	0.80 ± 0.40
	Spec	0.20 ± 0.40	0.40 ± 0.48	0.40 ± 0.48	0.60 ± 0.48	0.80 ± 0.40	0.20 ± 0.40
	Acc	0.58 ± 0.10	0.46 ± 0.06	0.50 ± 0.09	0.48 ± 0.13	0.50 ± 0.10	0.58 ± 0.10
	F1-s	0.59 ± 0.30	0.40 ± 0.32	0.42 ± 0.35	0.29 ± 0.3	0.13 ± 0.26	0.59 ± 0.30

Table 18. Geometrical early fusion architecture compared with 3D CNN based fusion, using 3 convolutional layers

there were significant differences ($\rho < 0.05$) between all stages included in the dataset (0, 1, 1.5, 2, and 3). With exception of stages 1.5 and 2 ($\rho > 0.5$), or between 2 and 3 ($\rho > 0.2$). Moreover, the H&Y scale utilizes the retropulsion test to classify patients at stage 3 by assessing postural stability through a controlled push, but this test was omitted from data acquisition to minimize fall risk, relying instead on gait and posture assessments to infer postural stability. Interestingly, freezing of gait (graph (e) of Figure 16) shows significant differences across the various stages in the dataset (all ρ values less than 0.05): normal, slight, mild, and moderate. In graph (f) of Figure 16, which shows the prediction of gait autonomy (gait item, MDS-UPDRS Part III scale), significant differences are observed between each of the Parkinson's stages in the dataset (normal,slight, mild, moderate). However, there are no statistical differences between the slight and mild stages ($\rho > 0.2$). All patients were recorded under the effects of medication to ensure their safety during data acquisition; however, this introduces a limitation in result interpretation, as medication can mitigate certain motor symptoms, making it more challenging to distinguish between different disease stages. Additionally, the progression of Parkinson's disease is heterogeneous, and differentiating between adjacent stages, such as mild and moderate bradykinesia, can be difficult.

4.5. DISCUSSION AND CONCLUDING REMARKS

Today, the evaluation of patients with PD is carried out via scales such as modified H&Y scale and MDS-UPDRS, but depending on the expertise of specialists [6, 8]. This work introduces

Computational Approaches		Cardinal Symptoms		Posture and balance Symptoms		Gait and mobility symptoms	
Config	**Metr**	**Bradykinesia Ocular**	**Bradykinesia in Gait**	**Wrong Posture**	**Bilateral Afectation (H&Y)**	**Freezing of Gait**	**Autonomy of Gait**
Early Fusion 3 3D conv 2 Bires	**Prec**	0.92 ± 0.14	0.89 ± 0.13	0.90 ± 0.12	0.93 ± 0.14	0.84 ± 0.20	0.93 ± 0.14
	Sen	0.97 ± 0.05	0.94 ± 0.06	0.96 ± 0.04	0.98 ± 0.04	0.82 ± 0.14	0.97 ± 0.05
	Spec	0.89 ± 0.21	0.86 ± 0.17	0.85 ± 0.18	0.89 ± 0.21	0.85 ± 0.17	0.89 ± 0.21
	Acc	0.93 ± 0.10	0.90 ± 0.08	0.91 ± 0.10	0.93 ± 0.10	0.83 ± 0.13	0.93 ± 0.10
	F1-s	$\mathbf{0.94 \pm 0.08}$	$\mathbf{0.91 \pm 0.07}$	$\mathbf{0.92 \pm 0.08}$	$\mathbf{0.95 \pm 0.08}$	$\mathbf{0.81 \pm 0.13}$	$\mathbf{0.94 \pm 0.08}$
Logistic Regression (LR)	**Prec**	0.66 ± 0.07	0.77 ± 0.17	0.76 ± 0.07	0.87 ± 0.13	0.60 ± 0.14	0.79 ± 0.22
	Sen	0.69 ± 0.26	0.71 ± 0.17	0.72 ± 0.21	0.69 ± 0.17	0.50 ± 0.22	0.68 ± 0.25
	Spec	0.67 ± 0.17	0.76 ± 0.17	0.74 ± 0.09	0.85 ± 0.14	0.72 ± 0.06	0.81 ± 0.18
	Acc	0.66 ± 0.08	0.74 ± 0.14	0.73 ± 0.10	0.76 ± 0.07	0.62 ± 0.10	0.75 ± 0.16
	F1-s	0.64 ± 0.11	0.73 ± 0.15	0.73 ± 0.13	0.75 ± 0.10	0.54 ± 0.19	0.72 ± 0.21
Random Forest (RF) (Our Method in chapter 2)	**Prec**	0.71 ± 0.28	0.81 ± 0.17	0.79 ± 0.14	0.85 ± 0.10	0.75 ± 0.20	0.80 ± 0.18
	Sen	0.74 ± 0.25	0.85 ± 0.13	0.93 ± 0.07	0.95 ± 0.08	0.74 ± 0.16	0.75 ± 0.22
	Spec	0.66 ± 0.34	0.74 ± 0.28	0.68 ± 0.25	0.75 ± 0.20	0.70 ± 0.26	0.73 ± 0.28
	Acc	0.67 ± 0.20	0.79 ± 0.10	0.81 ± 0.13	0.87 ± 0.10	0.72 ± 0.17	0.74 ± 0.10
	F1-s	0.68 ± 0.17	$\mathbf{0.81 \pm 0.06}$	$\mathbf{0.85 \pm 0.09}$	$\mathbf{0.89 \pm 0.08}$	0.73 ± 0.14	0.74 ± 0.12
Support Vector Machine (SVM)	**Prec**	0.54 ± 0.33	0.55 ± 0.17	0.66 ± 0.23	0.72 ± 0.21	0.59 ± 0.20	0.68 ± 0.22
	Sen	0.26 ± 0.20	0.61 ± 0.18	0.62 ± 0.24	0.68 ± 0.23	0.43 ± 0.09	0.66 ± 0.25
	Spec	0.61 ± 0.34	0.50 ± 0.15	0.61 ± 0.23	0.65 ± 0.19	0.67 ± 0.22	0.67 ± 0.23
	Acc	0.41 ± 0.09	0.55 ± 0.16	0.62 ± 0.19	0.67 ± 0.23	0.56 ± 0.12	0.66 ± 0.21
	F1-s	0.25 ± 0.16	0.58 ± 0.17	0.63 ± 0.20	0.69 ± 0.20	0.49 ± 0.11	0.66 ± 0.22
Multilayer Perceptron (MP)	**Prec**	0.66 ± 0.17	0.66 ± 0.07	0.77 ± 0.17	0.77 ± 0.09	0.64 ± 0.13	0.70 ± 0.11
	Sen	0.84 ± 0.23	0.76 ± 0.23	0.75 ± 0.22	0.63 ± 0.14	0.61 ± 0.09	0.79 ± 0.16
	Spec	0.55 ± 0.28	0.57 ± 0.17	0.69 ± 0.26	0.70 ± 0.16	0.66 ± 0.18	0.57 ± 0.33
	Acc	0.66 ± 0.12	0.66 ± 0.07	0.73 ± 0.16	0.67 ± 0.03	0.63 ± 0.09	0.70 ± 0.09
	F1-s	0.69 ± 0.11	0.69 ± 0.08	0.74 ± 0.15	0.68 ± 0.08	0.62 ± 0.09	0.73 ± 0.07

Table 19. Comparison of the geometrical early fusion architecture with classical machine learning methods.

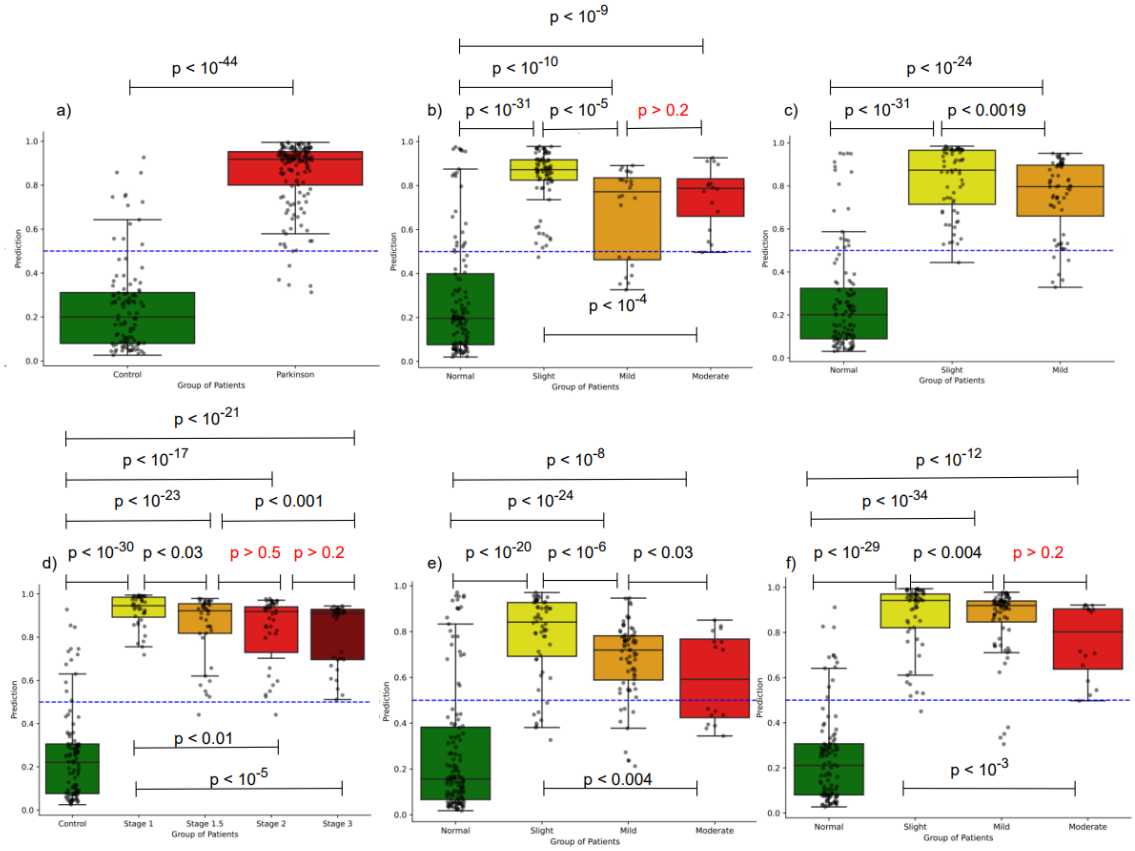

Figure 16. Distributions of Parkinson probability outputs by the early fusion binary classifier, for the different stage groups. The PD patients were categorized into different stages of the disease on the basis of a) ocular bradykinesia, b) bradykinesia in gait, c) wrong posture, d) bilateral affectation of gait (H&Y scale), e) freezing of gait, f) autonomy of gait. The values of ρ that appear in red correspond to insignificant differences between the two distributions.

an end-to-end geometric multimodal approach that combines gait and eye movement patterns to support multiple-scale motor impairments classification. For instance, bradykinesia (both ocular and during gait), but also to symptoms related to posture and balance, such as incorrect posture and bilateral gait impairment, as well symptoms associated with mobility impairments, such as freezing of gait and gait autonomy. The proposed approach explored the early and intermediate covariate fusion of 3D convolutional inputs, showing the robustness of the geometrical approach to predict scale items of the modified Hoehn and Yahr

scale, four items from the MDS-UPDRS Part III, and an observation by a specialist related to-eye-movement. Particularly, early fusion configuration shows greater robustness, improving by approximately 3% in F1-score and accuracy for cardinal symptoms such as ocular and gait bradykinesia, as well as bilateral impairment and gait autonomy. The intermediate fusion approach considering three 3D conv layers and 2 BiRes shows gains between 11% and 12% in F1 score compared with the best unimodal approaches. Alternatively, fusing in the Riemannian manifold (early fusion) considering three 3D conv layers and 2 BiRes, reports average gains from 15% in F1 score and 13% in accuracy compared to the best unimodal results.

Thereafter, we discuss in detail how each considered impairment can be support from the proposed approach concerning the analysis of PD. For instance, for body bradykinesia, where individuals exhibit slower movement initiation and execution [108], we observed that early fusion approach significantly differed between groups of subjects without bradykinesia (normal), slight, and mild. Nonetheless, there is not statistical difference between the mild and moderate stages (Figure 16, graph b), probably because patients were recorded under the effect of medication (mainly levodopa). Regarding posture, as observed in Figure 16, graph c), the output PD probabilities three groups are significantly different: normal, slight, and mild. The values closest to one correspond to the slight stage, with all patients being under the effect of medication. Also, the probability distributions for bilateral gait disturbances [109] have limitations to differentiate among intermediate stages (1.5 to 2 and 2 to 3), but it is compelling between stages (0 to 1.5). These results may be justified because observation of stage 1.5 involves unilateral and axial impairment. However, the video recording is only in the sagittal plane, making it difficult to identify axial impairment. Additionally, the evaluation of

---

[108] Damian M HERZ and Peter BROWN. “Moving, fast and slow: behavioural insights into bradykinesia in Parkinson’s disease”. In: *Brain* 146.9 (2023), pp. 3576–3586.

[109] Anat MIRELMAN et al. “Gait impairments in Parkinson’s disease”. In: *The Lancet Neurology* 18.7 (2019), pp. 697–708.

Stage 3 involves a retropulsion test, which is not recorded and included in the dataset, making it challenging for the computational method to differentiate between Stage 2 and Stage 3.

In addition to classification accuracy, the results provide insights into the ability of the proposed architecture to stratify disease severity. The use of the Kolmogorov–Smirnov (KS) test confirms that the output probability distributions are significantly different across various stages of Parkinson's disease for multiple symptoms. This analysis was included to place special emphasis on the statistical significance of the results, ensuring that the observed differences between groups are not due to random variability. In particular, the model demonstrates strong discriminative power for ocular bradykinesia, postural impairment, and freezing of gait. These findings support the clinical relevance of the learned representations and their potential application for disease monitoring. The statistical significance of these results (p-values < 0.05 in most contrasts) reinforces the robustness of the approach and its applicability beyond binary classification. To enhance multiple output PD analysis, the proposoed approach analyzed and supported freezing gait episodes [110]. In Figure 16, graph e), the probability distributions for Freezing of gait, reporting significant p-values in all evaluated stages: normal, slight, mild, and moderate. Aditionally, this work quantify the ability to walk independently and safely[111]. In Figure 16, graph f), the probability distributions are showing significant p-values between the stages, except between the mild and moderate stages. To even improve standard scales, with patterns reported at early stages, in this study was also considered the ocular bradykinesia related with a delay in initiating movement, as well as a

---

[110] Fengting ZHANG et al. "Clinical features and related factors of freezing of gait in patients with Parkinson's disease". In: *Brain and behavior* 11.11 (2021), e2359.

[111] Diego SANTOS GARCÍA et al. "Predictors of loss of functional independence in Parkinson's disease: results from the coppadis cohort at 2-year follow-up and comparison with a control group". In: *Diagnostics* 11.10 (2021), p. 1801.

reduction in the amplitude and speed of movement (ocular bradykinesia)[112]. In Figure 16, graph (a), significant distributions with p-values close to zero can be observed, differentiating between the control group and the Parkinson's group. Interestingly, through the early fusion approach, outstanding accuracy results of 92% are obtained.

Regarding PD observations, state-of-the-art approaches principally report unimodal models to predict local impairments, but with restrictions to broad the multisystemic nature of the disease [113, 114]. Alternative a multimodal approach has included finger tapping, hand movements, and rapid alternating movements, via inertial sensors, predicting and quantifying bradykinesia, using kinematic features [115]. Besides, triaxial gyroscopes and triaxial accelerometers of smartwatches have used to record hand movements, and gait to support quantification of bradykinesia [116]. These approaches however are limited to only detect bradykinesia alterations, losing complementarity with other key motor alterations in PD patients. Other multimodal approaches have used gait, hand movements, and limb tapping movements to quantify rigidity and postural stability in Parkinson's patients [117]. In contrast, the proposed approach not only predicts bradykinesia and postural instability, also, predicts other motor impairments observed in clinical routines: bilateral impairment, freezing of gait,

---

[112] Yue-meng SUN et al. "Digital biomarkers for precision diagnosis and monitoring in Parkinson's disease". In: *NPJ Digital Medicine* 7.1 (2024), p. 218.

[113] Seyed-Mohammad FERESHTEHNEJAD et al. "New clinical subtypes of Parkinson disease and their longitudinal progression: a prospective cohort comparison with other phenotypes". In: *JAMA neurology* 72.8 (2015), pp. 863–873.

[114] Bastiaan R BLOEM et al. "M easurement instruments to assess posture, gait, and balance in P arkinson's disease: Critique and recommendations". In: *Movement Disorders* 31.9 (2016), pp. 1342–1355.

[115] Dong Jun PARK et al. "Evaluation for Parkinsonian Bradykinesia by deep learning modeling of kinematic parameters". In: *Journal of Neural Transmission* 128 (2021), pp. 181–189.

[116] Luis SIGCHA et al. "Bradykinesia detection in Parkinson's disease using smartwatches' inertial sensors and deep learning methods". In: *Electronics* 11.23 (2022), p. 3879.

[117] Ling-Yan MA et al. "Remote scoring models of rigidity and postural stability of Parkinson's disease based on indirect motions and a low-cost RGB algorithm". In: *Frontiers in Aging Neuroscience* 15 (2023), p. 1034376.

and gait autonomy. Additionally, it can predict ocular bradykinesia, which, although not observed in clinical standards, has been shown in numerous studies to be sensitive in detecting Parkinson's disease, particularly in its early stages. The proposed approach achieves an average F1-score range for the six predicted impairments from a minimum of 81% (freezing of gait) to a maximum of 95% (Bilateral affectation) through early fusion. Intermediate fusion yields an average F1-score ranging from a minimum of 80% (freezing of gait) to a maximum of 93% (bilateral affectation). As a contribution of this work, the analysis of the binary probability distributions reveals that the predicted motor impairments that significantly differ across all stages observed by the expert specialist are: ocular bradykinesia, postural instability, and freezing of gait. However, a few distributions related to bradykinesia in gait, bilateral gait impairment, and gait autonomy did not reach statistical significance. Recording additional tests, such as the push test from the modified H&Y scale or frontal recordings during gait, could address this issue. Furthermore, the highest predictions are associated with the early stages of various motor impairments. This is probably related to the fact that all patients were recorded under the effects of medication, with patients in more advanced stages receiving higher doses, which complicates the differentiation between stages.

The proposed method is capable of discriminating and quantifying motor impairments between Parkinson's patients and control subjects using spatiotemporal features and multimodal geometric configurations. The descriptors provided by the covariance representation are highly compact, with an input size of 256×256 and an output size of 64×64 (with 2 BiRe blocks) in the LogEig layer during early fusion, and input size of 128×128 with an output size of 32×32 (with 2 BiRe blocks) in the LogEig layer for each branch modality during intermediate fusion, enabling efficient learning. Increasing the depth of geometric layers in multimodal approaches enables learning more compact and discriminative representations for distinguishing between patients and control subjects.

The quantification of motor impairments through multimodal geometric strategies enhances diagnostic accuracy by capturing complementary information that includes diverse symp-

toms, such as cardinal symptoms and those related to balance and mobility, providing a comprehensive assessment of the patient and offering a detailed profile of motor status. This is particularly useful for cases in which certain symptoms may not be evident or are subtle, such as ocular bradykinesia. With this proposed approach, the physician can make informed decisions regarding treatment or follow-up based on objective data. Potentially, this approach can facilitates the evaluation of therapeutic interventions' effectiveness, adjustment according to changes in the patient's symptom profile, and evaluation on disease progression. Furthermore, the proposed approach opens new perspectives in supporting the identification of Parkinson's phenotypes (such as the akinetic-rigid phenotype and the postural instability and gait difficulty phenotype), helping to personalize treatment and improve diagnosis, as to verify how some phenotypes respond better to specific types of intervention. By automating the evaluation of complex motor symptoms and generating reports, the proposed approach can supports the specialist's observations.

Now the proposed approach needs to be evaluated in a larger cohort of patients to determine statistical significance within the affected population. Additionally, integrating other modalities, such as hand movements, would allow the assessment of additional items from the MDS-UPDRS, providing the quantification of other key impairments, such as postural or resting tremor. Furthermore, future work should evaluate the performance of the models in states where the patient is not under the effects of medication. In this regard, the assessment of ON-OFF effects would allow for models that are more adaptable to the patients' everyday situations.

5. CONCLUSIONS AND PERSPECTIVES

In this doctoral thesis, the problem of quantifying, characterizing, and classifying Parkinsonian signs was addressed, focusing primarily on gait and eye movement through video sequences without the need for markers. This problem posed a classification challenge due to the disease's multifactorial nature, reflected in the varying intensities and frequencies of motor impairments among patients. Moreover, diagnostic errors based solely on clinical expertise range from 6% to 25%, highlighting the need for computational tools to support specialists by enhancing diagnosis, quantification, and characterization. The development of such tools has the potential to enable personalized treatment, facilitate early diagnosis, and improve disease monitoring.

A first contribution of this work was the coding of compact covariance descriptors at the video- and frame-level, integrating handcrafted kinematic features, pre-trained representations, and end-to-end learned features. These descriptors efficiently combined multiple modalities, achieving an accuracy up 90% in binary classification, when integrating gait and eye movement, and 70% accuracy when incorporating voice and facial expression features. Their discriminative power was further validated for Parkinson's disease stage classification using an online fusion strategy (Appendix A) that integrated gait and eye movement patterns, encoding covariances of dense optical flow features to capture postural configurations and motion velocity patterns. These representations, processed via LSTM networks, achieved 78% accuracy in global video classification. These findings confirm that covariance descriptors provide a compact, flexible, and robust approach for motor impairment quantification, preserving geometric properties on the Riemannian manifold while effectively capturing spatial and temporal information relevant to Parkinson's characterization.

In a second contribution, in this thesis was introduced deep geometric architectures that learn compact feature representations from covariance matrices, preserving their intrinsic mathematical properties. These architectures effectively capture complex multimodal relationships,

improving generalization, especially in limited datasets. The geometric network exploits invariance and stability properties, ensuring robust learning despite high inter-subject variability. Experimental results demonstrated that: Early and intermediate fusion approaches based on covariance matrices achieved up to 96% accuracy. Late fusion (combining probability outputs) reached 93% accuracy. A key innovation was the ability of geometric networks to operate directly on Riemannian representations, eliminating the need for intermediate projections or normalization steps, which can lead to loss of critical information.

Furthermore, a third central contribution consist on a multimodal, geometric end-to-end learning framework that integrated gait patterns and eye movements to classify multiple motor impairments assessed by clinical scales. By leveraging early and intermediate fusion strategies using covariance representations from 3D convolutional features, the methodology achieved: A 15% increase in F1-score and 13% accuracy gain with early fusion on the Riemannian manifold, compared to unimodal approaches. These results indicate that quantifying motor impairments through multimodal geometric learning significantly improves diagnostic precision, capturing complementary sources of motor dysfunction. This approach is particularly valuable for identifying subtle symptoms, such as ocular bradykinesia, which may be overlooked in traditional assessments. By providing objective, data-driven insights, this framework supports clinicians in decision-making, optimizing treatment strategies and patient monitoring.

The proposed models were trained and validated using a proprietary dataset, which incrementally incorporated new participants, enabling the validation of both diagnostic classification hypotheses and motor pattern prediction as characterized by a specialist neurologist. This dataset represents a significant effort in Parkinson's research, and an alternative to test alternative methodologies around the characterization of Parkinson disease.

Perspectives of research

The development of this research opens new horizons in the study of Parkinson's disease. Some perspectives are detailed below:

Reproducibility and real-world applicability of the recording system.

A key limitation of our study lies in the use of a controlled experimental environment, in which the positions, distances, and orientations between the cameras and the participants were fixed. This methodological choice ensured consistency and minimized variability during data collection, which was essential for the development and validation of our method. However, it also raises questions about the reproducibility of the system in real clinical settings, where such controlled conditions may not be feasible. Translating the method into practical applications will require a careful assessment of how variations in camera placement, lighting, or space availability might affect model performance. Future work should explore the robustness of the proposed approach under different acquisition conditions and investigate domain adaptation techniques or normalization procedures to mitigate potential performance degradation in less standardized environments.

Self and weakly supervised Parkinsonian representations.

Currently motor deficit predictions were modelled from supervised architectures, using labels generated by a specialist. These labels may introduce subjectivity, associated with interobserver variability and diagnostic uncertainty. This dependence on manual labeling limited the models' ability to learn generalizable representations due to variability in clinical annotations. As a future research direction, the incorporation of non supervised and weakly supervised representations may enhance motor deficit representations. Since motor features can be represented using covariance matrices, which reside in the manifold of symmetric pos-

itive definite (SPD) matrices, this approach will leverage the intrinsic geometry of the data, avoiding the limitations imposed by traditional Euclidean spaces. Besides, the adaptation of geometric encoders and autoencoders will enable efficient compression while maintaining the geometric structure. From a clinical perspective, this approach will allow for a more interpretable capture of spatial and temporal relationships among anatomical landmarks. By preserving the geometric structure of the data and reducing reliance on manual labels, this approach presents a promising pathway toward the development of more interpretable and clinically relevant models in medical practice.

Attention blocks in Riemannian manifold.

In my current study, these modalities were combined using early and intermediate fusion strategies based on covariance representations in the Riemannian manifold, enabling an end-to-end learning approach for geometric feature extraction. However, this methodology does not optimally distinguish which temporal and spatial regions are most relevant for diagnosis. Besides, in some experiments, the use of consecutive geometric blocks collapses the representation. So, a future research direction may include the exploration of attention blocks into multimodal models to enhance the fusion of information from gait and eye movements for Parkinson's classification. This improvement may increase classification accuracy by capturing subtle motor deterioration patterns that might currently be diluted in the feature fusion process. Another key benefit will be model explainability, as attention mechanisms within the geometry of the manifold will provide interpretable visualizations of the video regions most influential in classification, facilitating clinical validation and real-world applicability. Consequently, incorporating attention blocks into multimodal models within the Riemannian manifold will represent a significant advancement over current methods, enabling a more robust, interpretable, and precise integration of gait and eye movement data for Parkinson's characterization.

Predicting continuous severity stages.

Current models persist on limitations about to measure the progression and severity of each condition. A key future research direction will be to explore alternatives to predict continuous severity values, providing a more detailed quantification of each motor impairment. To achieve this, the current loss function would need to be modified, replacing cross-entropy with appropriate regression metrics for learning in the Riemannian manifold. From a clinical standpoint, the adaptation of these models allow for the correlation of continuous severity values with traditional specialist-assessed scales, improving the model's interpretability and applicability in real-world settings.

Handling missing modalities.

In multimodal models used for predicting motor deficits in Parkinson's disease, the absence of certain modalities in some subjects poses a significant challenge. To address this issue, future works should include strategies to carry out a geometric interpolation in the manifold of symmetric positive definite (SPD) matrices, leveraging the Riemannian structure of the data. This interpolation may ensure that the estimates preserve the geometric structure of the data, avoiding imputation biases associated with Euclidean spaces. Additionally, techniques such as Riemannian autoencoders will be explored to refine the interpolated representation and enhance its coherence with the available modalities. From a clinical perspective, this methodology will contribute to enhancing the interpretability and applicability of the models in real-world settings, reducing reliance on fully annotated data and maximizing the utility of partial records in the characterization of motor deficits.

Integrating the modalities of tremor, facial expression, and voice.

From the achieved results of multimodal strategies, we consider that the integration of new modalities may impact in the global characterization of the disease. For instance, the integration of tremor, facial expression, and voice analysis as combined modalities is highly beneficial to broadening the range of observable symptoms in Parkinson's patients. Tremor is one of the most characteristic and debilitating symptoms of the disease, and its inclusion could significantly enhance diagnostic accuracy and the assessment of disease progression. Additionally, facial expression and voice are important aspects that can reflect neurological and emotional changes in patients. Facial expression can provide insights into facial rigidity and bradykinesia, while voice analysis can detect alterations in prosody, articulation, and speech fluency, which are common in Parkinson's disease. By combining data on tremor, facial expression, voice, gait, and eye movements, a more comprehensive view of the motor and non-motor states of patients could be obtained. This would enable a more holistic and personalized approach to treatment and disease management, facilitating more effective and tailored interventions. Furthermore, multimodal analysis could uncover complex interactions between different motor and non-motor symptoms, providing a deeper understanding of Parkinson's disease pathophysiology.

Extracting relevant information through landmarks.

In the approaches developed during my doctoral thesis, the use of the original video in both gait and facial expression modalities introduced irrelevant information, such as the background and external elements unrelated to the patient's movement, which may have affected the specificity of the extracted representations. As a future research direction, an approach based on the extraction and analysis of modality-specific landmarks may be useful to obtain kinematic representations encoded in covariance matrices, enabling a more compact and robust characterization of spatial and temporal relationships within each modality. From a computational perspective, the use of landmarks reduces the dimensionality of the input data and mitigates the influence of factors unrelated to movement, thereby improving the

discriminability of motor patterns associated with Parkinson's disease. Furthermore, the combination of these representations with geometric analysis methods on Riemannian manifolds will allow for a more structured multimodal fusion, preserving the intrinsic structure of the data and facilitating the identification of motor phenotypes with greater precision.

Longitudinal studies on disease progression.

Conducting longitudinal studies using the same video protocols taken over different years or periods would allow for a more detailed analysis of Parkinson's disease progression. This approach could help identify patterns of motor and ocular deterioration over time, providing valuable data to predict disease progression and adjust treatments more effectively. Additionally, it could contribute to the development of biomarkers that indicate the speed of disease progression, which would be crucial for planning therapeutic interventions. Longitudinal studies could also reveal how different factors, such as age, gender, and comorbidities, affect disease progression, providing a solid basis for developing personalized treatments.

A. APPENDIX: A RECURRENT APPROACH FOR PREDICTING PARKINSON STAGE FROM MULTIMODAL VIDEOS

A.1. ABSTRACT

Parkinson's disease is a neurodegenerative disease that affects more than 6.1 million people worldwide. In the clinical routine, the main tool to diagnose and monitor disease progression is based on motor impairments, such as postural instability, bradykinesia, tremor, among others. Besides, new biomarkers based on motion patterns have emerged to describe disease findings. Nonetheless, this motor characterization has low sensitivity, especially at early stages, and is largely expert-dependent, because protocols are mainly based on visual observations. However, most of these analyses require complex and some invasive systems that additionally only bring global information of complete recordings. This work introduces a multimodal approach that integrates gait and eye motion videos to quantify and predict patient stage on-the-fly. This method starts by computing dense apparent velocity maps that represent the local displacement of the person seen from the gait in a sagittal plane and as micro-movements during the fixation experiment. Then, each frame is described as a covariance descriptor of deep feature activation maps computed over the motion field at each video time. Then, the covariance video manifold is mapped to a recurrent LSTM network to learn higher non-local dependencies and quantify a motion descriptor. Also, an end-to-end scheme allows to lately fuse both modalities (gait and fixational eye) to obtain a more sensitive Parkinson disease descriptor. In a study with 25 subjects, the proposed approach reaches an average F1-score of 0.83 with an average recall of 0.78. In a temporal prediction analysis, the approach reports major correlations with the disease considering swing phase.

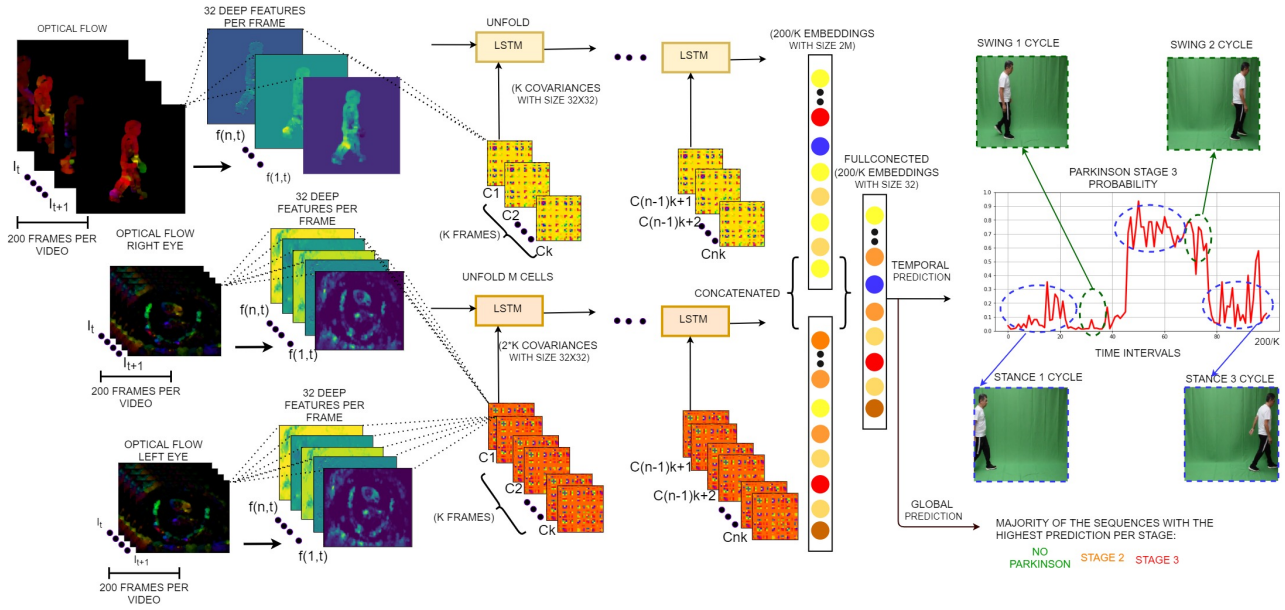

Figure 17. Proposed approach pipeline. A markerless strategy is proposed by computing a spatially dense representation for each frame along the video sequence.

The partial content of this appendix has been accepted and published in [118].

A.2. PROPOSED APPROACH

Using frame-level spatial covariance matrices to fed a recurrent neural network, the proposed approach exploit spatial and temporal motor information. Moreover, this work has the capability to integrate gait and eye fixation sequences to improve characterization and description of PD. The proposed strategy is illustrated in Figure 17. Firstly, videos that record each modality of interest, are represented as a set of frame-covariance matrices. Then, the temporal processing is performed using a Long Short Term Memory (LSTM) network. Subsequently, a latent temporal vector is learned on LSTM to capture motor patterns along

[118] John ARCHILA; Antoine MANZANERA, and Fabio MARTÍNEZ. "A recurrent approach for predicting Parkinson stage from multimodal videos". In: *17th International Symposium on Medical Information Processing and Analysis*. Vol. 12088. SPIE. 2021, pp. 37–45.

sequences. Finally, a softmax function is used to predict the probabilities of the different PD stages, for each video slice of every patient.

A.2.1. PD motion modalities. Parkinson disease causes different motor and non-motor impairments, at different stages, that impact different human activities, and can be recorded using different sensors. Quantifying these impairments allows to evaluate the disease progression or the effect of a particular treatment. In this study, we include the gait and eye fixational movements as observational descriptors of the disease. These motion modalities are described thereafter:

- **The gait** is a complex locomotion process that requires coordination between neuromotor commands and muscles to obtain optimal displacement. Regarding PD, the gait alteration is generally characterised by unbalanced postures, slow movements and rigidity, which gives a global perspective of kinematic behavior. In video analysis have been proposed strategies based on silhouettes to encode different poses of each movement, highlighting geometric patterns typical of PD [119]. Also, 2D pose patterns have allowed to measure gait cadence and limb flexion angles, but the analysis was limited to advanced stages of the disease [120].

- **Fixational eye patterns** have been identified as strongly correlated with dopamine deficit [18]. This analysis then can provide complementary information of locomotor patterns, being sensitive even in early stages. Most studies of oculomotor patterns use the video oculography technique (VOG). This method records two-dimensional movements

---

[119] Javier ORTELLS; María Trinidad HERRERO-EZQUERRO, and Ramón A MOLLINEDA. “Vision-based gait impairment analysis for aided diagnosis”. In: *Medical & biological engineering & computing* 56.9 (2018), pp. 1553–1564.

[120] Kenichiro SATO et al. “Quantifying normal and parkinsonian gait features from home movies: Practical application of a deep learning–based 2D pose estimator”. In: *PloS one* 14.11 (2019), e0223549.

and performs measurements such as velocity, latency, and other kinematics [121], [122]. As an example, the latency of divergence and convergence movements is greater for PD patients than for control subjects [123]. Nevertheless, the complexity of the acquisition system and its constant calibration makes the clinical routine of the specialist more difficult.

A.2.2. Frame covariance representation from deep motion features. The bradykinesia that is associated with the slowness of movement [124] and the tremor described as a periodic motion [125] are the predominant motor impairments for PD. Hence, velocity fields computed from video recordings are relevant to characterize and quantify PD. In fact, this hypothesis has been successfully implemented to analyze the disease in single modality, using either gait or tremor videos [125, 126]. The proposed strategy then starts by computing dynamic patterns from gait and eye video sequences based on apparent velocity fields computed on consecutive frames. The Farnebäck optical flow method was applied considering its good trade-off between speed and accuracy [66]. This method estimates a dense and regular optical flow using a quadratic polynomial approximation of each pixel's neighborhood and

---

[121] Andrew H CLARKE. "Laboratory testing of the vestibular system". In: *Current opinion in otolaryngology & head and neck surgery* 18.5 (2010), pp. 425–430.

[122] Ajit KHOSLA and Dongsoo KIM. *Optical Imaging Devices: New Technologies and Applications*. CRC Press, 2017.

[123] Jaromír HANUŠKA et al. "Fast vergence eye movements are disrupted in Parkinson's disease: a video-oculography study". In: *Parkinsonism & Related Disorders* 21.7 (2015), pp. 797–799.

[124] Csaba VÁRADI. "Clinical Features of Parkinson's Disease: The Evolution of Critical Symptoms". In: *Biology* 9.5 (2020). DOI: 10.3390/biology9050103.

[125] Mehmet Akif ALPER; John GOUDREAU, and Morris DANIEL. "Pose and Optical Flow Fusion (POFF) for accurate tremor detection and quantification". In: *Biocybernetics and Biomedical Engineering* 40.1 (2020), pp. 468–481.

[126] Khalid BASHIR et al. "Gait Representation Using Flow Fields." In: *BMVC*. 2009, pp. 1–11.

estimating the translation vector that locally affects each polynomial.

Then, the kinematic information provided by optical flow was projected to a bank of convolutional filters extracted from the layers of a trained deep convolutional network. This projection is carried out at each frame allowing to enrich description of motion patterns. These deep features have recently demonstrated great capability to represent complex motion patterns, by modelling non linear and large scale relations. The set of learned filters decompose kinematic information in a total of n features maps $F_t = \{f_{(1,t)}, \ldots, f_{(n,t)}; f_{(i,t)} \subset \mathbb{R}^{W \times H}\}$, enhancing nonlinear relationships. In this work, we select the n filters F_t from the first layer of a pre-trained convolutional net, as a set of learned low-level features. The deep representation was provided by an in-depth separable convolution architecture, that allows computational reduction, less redundancy in activation maps, and that requires fewer data for training [73, 74]. Hence, for each frame t, a spatial covariance matrix C_t relative to the set of feature maps F_t was computed to obtain a compact embedding description of postural and dynamic performance at each time. The covariance matrix is computed as: $C_t(i,j) = \mathbb{E}\left((f_{(i,t)} - \mathbb{E}(f_{(i,t)}))(f_{(j,t)} - \mathbb{E}(f_{(j,t)}))\right)$ where the expectation $\mathbb{E}$ is calculated over the $W \times H$ points of each feature map. The dimention of gait descriptor is $n \times n$ and ocular descriptor is $2 \times n \times n$ where n is the number of deep features per frame. The compact representation of covariance gives the correlation between poses. In this way, every patients has a signature per modality. This covariative signature show different trends between deep features motion maps, enhancing discordance of movements, arrhythmic patterns and tremor[127, 128].

---

[127] *Towards Data-Driven Modeling of Pathological Tremors*. Vol. Volume 2: 16th International Conference on Multibody Systems, Nonlinear Dynamics, and Control (MSNDC). International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. V002T02A030. Aug. 2020. eprint: https://asmedigitalcollection.asme.org/IDETC-CIE/proceedings-pdf/IDETC-CIE2020/83914/V002T02A030/6586091/v002t02a030-detc2020-22147.pdf.

[128] Xu LIU et al. “A dual-branch model for diagnosis of Parkinson’s disease based on the independent and joint features of the left and right gait”. In: *Applied Intelligence* (2021), pp. 1–12.

A.2.3. A continuous multimodal motion pattern quantification. During the video sequence, the proposed method quantifies temporal information of deep covariation maps, representing potential motor impairments during movement execution. Two motion modalities (gait and eye fixation) were analyzed over time by computing recurrent deep representation, quantifying the probability for each time but also recovering a global disease classification. We implemented a recurrent LSTM with the capability to encode short and long dynamic temporal deep representations. These recurrent modules have proven successful in many different tasks to temporal modelling of events, including the analysis of other neurodegenerative diseases [129, 130].

Formally, the set of frame covariances $\mathbf{F} = \{f_1, f_2 \dots f_k\}$, are sequentially propagated by a set of M recurrent units. In this layer, a resulting mid-level representation captures temporal dependencies, being m_t the internal state memory that is updated from the expression: $m_t = fg_t \otimes m_{t-1} + i_t \otimes \tilde{m}_t$. Then, the resulting hidden state h_t is computed as $h_t = o_t \otimes \tanh(m_t)$, where $\tanh$ the hyperbolic tangent function (valued in $[-1, +1]$) and hidden states are initialized to zero. The fg_t is a forget gate that uses a group of k covariances Ck_t to decide how much information omit. The expression for this function is herein calculated as: $fg_t = \sigma(W_f[x_t, h_{t-1}] + b_f)$, where σ is the sigmoid function (valued in $[0, 1]$), and $[.,.]$ is the concatenation operator. The trainable parameters here are the $M \times (N + M)$ weight matrix W_f, and the bias vectors b_f.

For this recurrent module, the input i_t is learned from a vector Ck_t made by a concatenated set of covariances, and the previous state h_{t-1}. This input vector is then expressed as $i_t = \sigma(W_i[Ck_t, h_{t-1}] + b_i)$, which thereafter will be used to weight the internal memory state m_t.

---

[129] Solale TABARESTANI et al. “Longitudinal prediction modeling of alzheimer disease using recurrent neural networks”. In: *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE. 2019, pp. 1–4.

[130] Aite ZHAO et al. “Dual channel LSTM based multi-feature extraction in gait for diagnosis of Neurodegenerative diseases”. In: *Knowledge-Based Systems* 145 (2018), pp. 91–97.

Then, the memory update $\tilde{m}_t = \tanh(W_m[Ck_t, h_{t-1}] + b_m)$ selects the information for the internal state. Finally, the output is computed as $o_t = \sigma(W_o[Ck_t, h_{t-1}] + b_o)$, which forms a compact descriptor of the spatio-temporal information of the covariance set. Likewise, the learned parameters are represented by the $M \times (N + M)$ weight matrices W_i, W_m, and W_o and the bias vectors b_i, b_m and b_o.

In summary, the temporal and recurrent motion description has the capability to detect and model time intervals correlated with multimodal parkinsonian patterns. For each motion modality was then built a recurrent deep model that learn complex temporal relationships from video manifold form by the set of frame-covariances. Despite the gait and eye fixational recordings are not synchronized, the proposed strategy take advantage of eye oscillatory behaviour ranging in short time from (2 ms to 20 ms) and on a long time scale (100 to 400 ms) [131]. Under this premise both modalities can easily aligned and the resultant recurrent embedding descriptors can be fussed to enhance PD description. In this way, the fused model temporally quantifies the probability that depends on the execution of the patient's movement in each modality, which potentially helps to identify postures prone to correction in physiotherapeutic therapies or in medical follow-up.

A.3. EXPERIMENTAL SETUP

A.3.1. Data. A total of 25 participants was included in this study: 13 control subjects (average age of 72.2 ± 6.1) and 12 PD patients (average age of 72.3 ± 7.4). The PD patients were diagnosed in second (5 patients) or third (7 patients) stage of the disease by a physician following the Hoehn-Yahr scale. This study was approved by the Ethics Committee of Universidad Industrial de Santander and a written informed consent was obtained. For eye fixational recording, the patients observed a fixed spotlight projected on a screen with

[131] Carl JJ HERRMANN; Ralf METZLER, and Ralf ENGBERT. "A self-avoiding walk with neural delays as a model of fixational eye movements". In: *Scientific reports* 7.1 (2017), pp. 1–17.

a dark background, with an average duration of 4 seconds. The eye region was manually cropped (210×140 pixels) to obtain the sequences of interest. Regarding walking, marker-less sagittal-plane videos were recorded with a spatial resolution of 1280×720 pixels and a temporal resolution of $60\,fps$. The locomotion was recorded during a 5 meter displacement, for an average duration of 4 seconds. For each participant, 6 videos for gait and 6 videos for each eye were recorded with a conventional camera, Nikon D3200 with spatial resolution of 1280×720, resulting in a total dataset of 450 videos.

A.3.2. Experimental configuration. A leave-one-patient-out cross-validation was carried out to evaluate the proposed approach, i.e. at each iteration on the evaluation, one patient is left out to test and the remaining ones are used for training. For the whole dataset of RGB videos was computed the Farnebäck optical flow (with 5 scales and a 3×3 pixels averaging window). Then, each frame of the videos was sent to the second layer of a pre-trained MobileNet V2 that counts with a total of 32 learned filters. The resulting deep features have a spatial size of (112×112). Then, for each frame a spatial covariance matrix is computed to summarize deep feature correlations, resulting in matrices with size of 32×32. Each video is herein represented by a total of 200 frame-level covariance matrices. The LSTM outputs for the two channels are concatenated and fully connected with a softmax layer to perform the frame-level prediction. The training was performed by following a categorical cross-entropy with Adam optimizer and learning rate that varies from 0.0001 to 0.001 with a step size of 0.0001, and including an early stopping strategy with categorical cross-entropy in 15 epochs.

A.4. RESULTS

A first evaluation was carried out to determine the best temporal interval to capture Parkinsonian motion patterns from bi-modal inputs. Figure 18 shows the performance achieved from different per-frame covariances. This plot also summarizes the performance using different hidden recurrent vector sizes of 16, 32 and 64, respectively. The window intervals

Table 20. Confusion matrices for each modality and for fusion approach.

	GAIT			OCULAR FIXATION			MERGE		
	C	**S2**	**S3**	**C**	**S2**	**S3**	**C**	**S2**	**S3**
C	**13(100%)**	0	0	**13(100%)**	0	0	**13(100%)**	0	0
S2	3(60%)	**2(40%)**	0	3 (60%)	**2(40%)**	0	0	**4(80%)**	1(20%)
S3	4(57%)	2(28%)	**1(14%)**	3(43%)	0	**4(57%)**	0	2(28%)	**5(71%)**

with relative small input vectors (number of covariance matrices k from 1 to 5) shows less variable results, achieving a proper integration of both modalities in very short time intervals (less than 100 ms). Particularly, the integration of gait and eye fixation with $k = 2$ and 32 cells obtained the best results with an average accuracy of 0.78. This result may be related to the fact that eye fixation movements at short time scales are more persistent [131]. The following experiments will then follow the validation approach with the best configuration, *i.e,* two frame-covariances with a recurrent cell of 32 units.

Secondly, we analyze the contribution of multimodal approach with respect to isolated modalities. To obtain a global classification score for each video, a majority voting is implemented from the per-frame predictions achieved in this proposed scheme. Table 20 summarizes the classification achieved from gait and eye isolated inputs, but also, by considering a late fusion of both modalities. As expected the fusion modality has a better performance to discriminate among control and Parkinson disease at level two and three, labelled according to Hoehn and Yahr scale. In other cases, the eye fixation has a better performance to discriminate third level of Parkinson, but the fusion of modalities overcomes the discrimination of patterns.

Table 21 reports a more detailed analysis including statistical metrics such as precision (prec), recall (rec), specificity (spec), F1-score (F1-s), and Mathews correlation coefficient (MCC). The multimodal fusion alternative achieves the best performance with a predominant F1-score (average of 0.83) and MCC (average of 0.74) with respect to single modalities. The eye fixation achieves a remarkable score to discriminate the Parkinsonian population (Recall = 1) but reports a low sensitivity to differentiate between considered Parkinson stages. Ac-

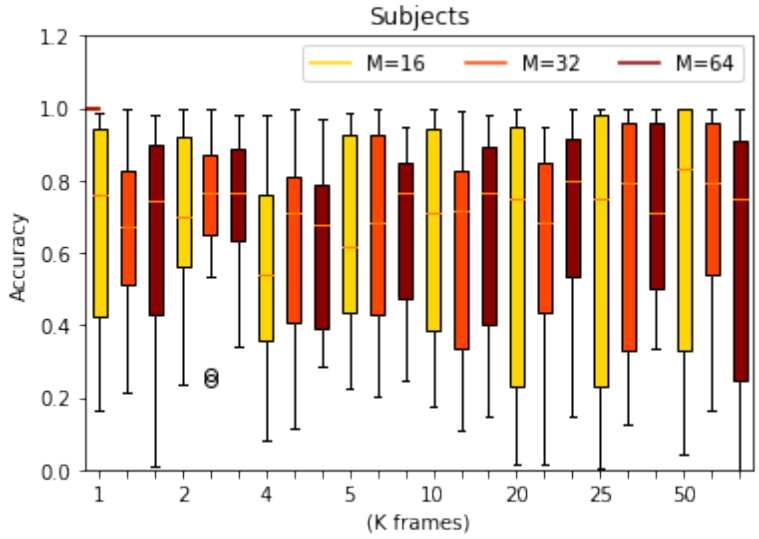

Figure 18. Mean classification accuracy as a function of the number k of covariance matrices, and the number M of cells in LSTM layer

cordingly, the Gait has a low classification rate between Parkinson stages and a low MCC, even for control population. In contrast, learning from both modalities the proposed approach exploits the best embedding correlations, which improves the classification performance.

Modalities		prec	rec	spec	F1-s	MCC
GAIT	**C**	**1**	0.65	**1**	0.79	0.52
	S2	0.40	0.50	0.86	0.44	0.33
	S3	0.40	**1**	0.75	0.24	0.33
OCULAR FIXATION	**C**	**1**	0.68	**1**	0.81	0.58
	S2	0.40	**1**	0.87	0.57	0.65
	S3	0.57	**1**	0.86	0.72	0.70
FUSION	**C**	**1**	**1**	**1**	**1**	**1**
	S2	0.80	0.67	0.95	0.73	0.57
	S3	**1**	0.68	**1**	0.76	0.66

Table 21. Scores for two modalities and fusion approach.

A major advantage of the proposed strategy is the capability to output classification scores over time, which may enrich evaluation and analysis of disease progression. Figure 19 sum-

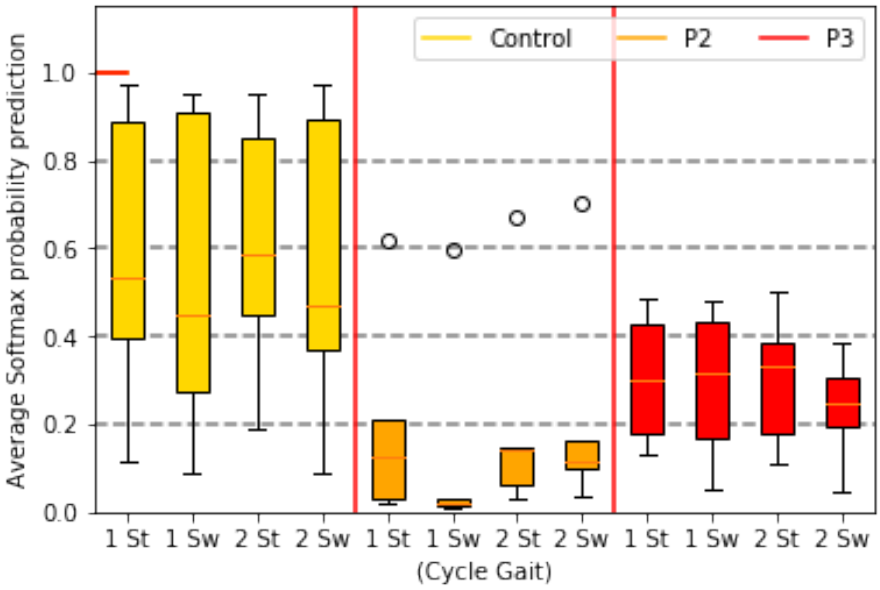

Figure 19. Average Softmax probability prediction for control reference, in function of the gait cycle.

marizes the temporal prediction achieved for the three different population classes (13 Control, 5 in PD stage 2, 7 in PD stage 3). In y-axis is plotted the control probability as reference of whole videos, while in x-axis is arranged information according to main gait phases, *i.e.,* first stance (1 St), second stance (2 St), first swing (1 Sw) and second swing (2 Sw). The stance phases (1,2 St) include the heel-to-toe contact sequence of the foot, while the swing phases (1,2 Sw) proceed with the foot suspended in the air[91].

Interestingly, the patients in stage two have lower probabilities to belong to control group, with a marked low variability in the reported predictions and reporting an outlier that corresponds to a patient with additional artifacts during the capture. It should be noted also that major discrimination in Parkinson between the two stages was in one of the swing phases, which may be related with major postural instability, a clear biomarker of the disease. For control there exist large variations for all phases, but with a clear tendency to larger probabilities. These temporal (partial) measures can be affected by a small number of observations during online predictions and therefore they should be contrasted with the final video classification.

Anyway this method has the potential to quantify and support disease diagnosis not only from a global video perspective but also discriminating among different gait phases, which are intrinsically combined with eye fixational behaviours. The identification of these poses and temporal alterations can provide tools to adapt the personalized physiotherapy technique according to the patient's needs.

A.5. DISCUSSION AND CONCLUDING REMARKS

This work introduces an online fusion strategy that integrates Gait and eye fixational patterns to classify Parkinson disease at two different stages and with respect to control population. The proposed approach uses a compact encoding of frame-level covariance of deep features calculated on dense optical flow images, that allows a rich primary representation of postural configurations, as well as, local motion velocity patterns. This statistical representation is then mapped to a recurrent LSTM network to learn motion temporal patterns, with the major advantage to produce online predictions of the disease. The gait patterns bring global descriptors while the eye fixational patterns can contribute to a major sensitivity to discriminate among disease stages. The results show a robust performance for global video prediction in the considered study of 25 patients and a total of 450 video sequences. The frame-level prediction shows remarkable correlation with the disease during swing phases. Nonetheless, some mistakes were reported on Parkinson stages, which may be associated to insufficient sensitivity of the descriptors, but also to expert annotation biases. The proposed approach results very compact, which brings opportunities to support observational analysis on-the-fly, allowing to suggest (spatial) regions, and (temporal) segments with major association to the disease. Reciprocally, the physiotherapist can design customized routines for the improvement of the most critical gait movements. Future works include a study that involves additional expert observations, larger number of patients and a more detailed analysis about time prediction. Also, in future works will be included additional description of side dependent (e.g. unilateral) symptom patterns, and any other conditions that may be explained into the

developed study.

B. APPENDIX: A MIXED AUDIO-VIDEO SPD NETWORK FOR ONLINE CLASSIFICATION OF PARKINSONIAN SPEECH PATTERNS

B.1. ABSTRACT

Parkinson's disease (PD) is a neurodegenerative disease that produces progressive motor impairments. Dysarthria (speech disorders) and hypomimia (face rigidity) are two major Parkinsonism patterns observed even at the early stages of the disease. Nonetheless, the clinical diagnosis is mainly observational and dependent on the specialists' expertise. Besides, the categorization of each of these patterns is isolated, which may lead to delayed diagnosis and misplanning of treatments. This work introduces a non-invasive multimodal strategy that integrates video and audio modalities into the online characterization of speech exercises. Subjects were invited to pronounce sustained vowels while video and audio were recorded. Then, a temporal window is run along the sequence to build online covariance matrices of synchronized face landmarks position and characteristic voice frequencies. From these temporal covariance matrices are learned Riemannian descriptors that allow to discriminate between Parkinson's and control subjects. From a study with 14 subjects, the proposed approach achieved a mean accuracy of 70% in sustained vowel pronunciation. Considering online predictions, the proposed approach evidenced a consistent accuracy of 0.77 during pronunciation of close vowels.

The partial content of this appendix has been accepted and published in [132].

---

[132] John ARCHILA; Antoine MANZANERA, and Fabio MARTINEZ CARRILLO. "A Mixed audio-video SPD network for online classification of Parkinsonian speech patterns". In: *IBERAMIA 2024. 18th Ibero-American Conference on Artificial Intelligence*. Montevideo, Uruguay, Nov. 2024.

B.2. PROPOSED APPROACH

This work introduces an online multimodal approach that fuses orofacial patterns, following an early fusion method based on covariance patterns. The covariance descriptor encodes both face landmarks trajectories and fundamental frequencies of the audio speech, aligned in intervals of time. Then, a geometrical representation is learned on the Riemannian manifold, to classify Parkinsonian patterns. The general pipeline of the proposed approach is illustrated in Figure 20.

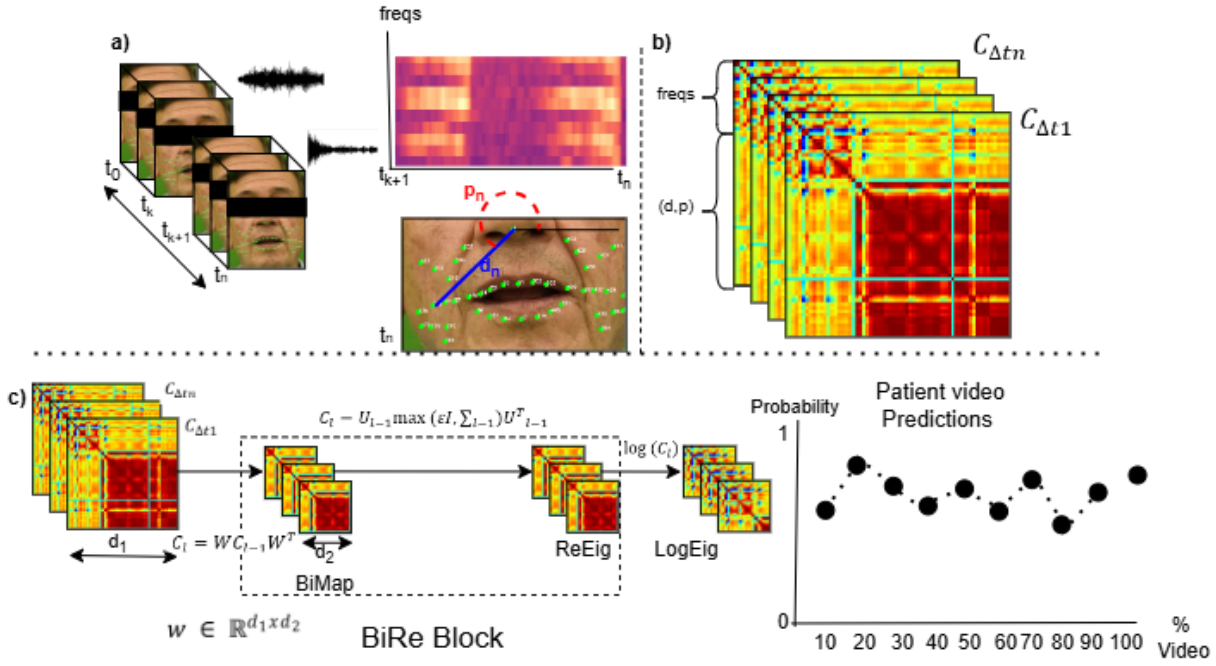

Figure 20. Multimodal Architecture: a) The position of each key-point in polar coordinates $d(t,k)$ and $p(t,k)$ where d is the distance between the nose to the landmark and p is the angle, is combined with short time spectrogram $\sigma(t,f)$ through b) covariance matrices in time intervals $C_{\Delta t}$. c) Then, the model learns new representations more compact for quantification of PD, with the capacity to output a prediction for each video slice. Riemannian geometry is form of BiRe blocks with a subsequent projection onto the tangent plane to carry out a classification. Thus, this approach characterizes the patient's pronunciation temporally, by predicting the probability of PD during the vocalization (bottom right plot).

B.2.1. Facial and Audio low-level features. In this work, we first computed low-level features, at each sequence time, to encode dysarthria and hypomimia disorders. For dysarthria, we computed short-time spectrograms $\sigma(t, f)$ as fundamental representations (Fig. 20(a), top right), capturing the essential frequency dynamics for frequencies f over sliding window at time t. Consequently, an audio sequence is represented by a spectrogram map with dimensions $N_f \times N_t$ where N_f is the number of frequencies and N_t is the number of time samples.

Regarding hypomimia, we computed the displacement of face key points in regions around the mouth because of the association with facial muscles involved in lip expression. The MediaPipe architecture was used to compute facial landmarks using only video information [133]. We selected 44 landmarks near the mouth and muscles involved in jaw movement during pronunciation. These landmarks allow summarizing the dynamics of the subject's face during various expressions and movements. Specifically, at each time synchronised with the audio spectrogram samples, we encode the position of each keypoint in polar coordinates, using as centre the tip of the nose (Fig. 20(a), bottom right), resulting in a sequence $\{d(t, k), p(t, k)\}$ of dimensions $2N_k \times N_t$, where d is the distance between the nose to the landmark and p is the angle. N_k is the number of keypoints and N_t the number of time samples. Using the nose as centre of coordinates allow to eliminate head movements and to focus on the motion of the mouth.

B.2.2. Temporal Covariance Computation. Now, for each time interval Δt, made of consecutive N_t time samples, we calculate the covariance matrix of the synchronised features $\Phi(t, i)$ composed of concatenated spectrogram frequencies $\sigma(t, f)$ and face keypoints $\{d(t, k), p(t, k)\}$:

$$C_{\Delta t}(i, j) = \mathbb{E}_{\Delta t}\left(\Phi(t, i)\Phi(t, j)\right) - \mathbb{E}_{\Delta t}\Phi(t, i)\mathbb{E}_{\Delta t}\Phi(t, j)$$

---

[133] Camillo LUGARESI et al. "Mediapipe: A framework for building perception pipelines". In: *arXiv preprint arXiv:1906.08172* (2019).

where $\mathbb{E}_{\Delta t}$ refers to the expectancy calculated over the N_t samples $t \in \Delta t$. This temporal covariance matrix, with dimension $(N_f + 2N_k)^2$ (Fig. 20(b)), encodes the dynamic relationships among integrated facial and speech features, providing a comprehensive description of their temporal dependencies. This representation helps with classification performance but also results self-explainable to support recognition of coordination patterns, which is crucial for unraveling the intricate temporal interplay between facial and voice features.

B.2.3. Covariance-based learning for temporal video predictions. Covariance matrices are Symmetric Positive Definite (SPD) matrices that lie in Riemannian manifolds with particular geometry, and need to be processed in a dedicated framework. For each temporal covariance $C_{\Delta t}$, we then learn a geometrical representation, capturing the inherent temporal dependencies between the different modalities. To do so, we first code a BiMap Layer following a bilinear mapping in each layer l, as: $\mathcal{C}_l = W_l \mathcal{C}_{l-1} W_l^T$, with $C_{l-1} \in \mathbb{R}_*^{d_{l-1} \times d_{l-1}}$ being the SPD matrix output of the layer $l-1$ and $W_l \in \mathbb{R}_*^{d_l \times d_{l-1}}$ the weight matrix transformation [89]. Hence, to ensure SPD property, an eigenvalue rectification layer is carried out, as: $\mathcal{C}_l = U_{l-1} \max(\varepsilon I, \Sigma_{l-1}) U_{l-1}^T$ where U_{l-1} and Σ_{l-1} are defined by the diagonal decomposition $\mathcal{C}_{l-1} = U_{l-1} \Sigma_{l-1} U_{l-1}^T$. Here, $\varepsilon > 0$ is a rectification threshold value, I is the identity matrix and Σ_{l-1} the diagonal matrix of the eigenvalues of $\mathcal{C}_{l-1}$. This operation adjusts the eigenvalues, avoiding negative values and improving discriminative performance. This specialized block facilitates the extraction of relevant information from the input data, contributing to the computation of effective covariation patterns.

Finally, to carry out the classification task, the learned matrix is projected onto a tangent plane (i.e. back to a Euclidean space), following a logarithm map $\mathbf{log}(\mathcal{C}) = U \log(\Sigma) U^T$. Then, classical dense layers are implemented to achieve the classification of the multimodal pronunciation exercise input.

B.3. EXPERIMENTAL SETUP.

This study involved 14 participants, consisting of 7 patients diagnosed with Parkinson's disease (PD) and 7 control patients. The PD group had an average age of 65 ± 4, while the control group had an average age of 61 ± 3. All PD patients were on (Levodopa) medication during data acquisition. Informed consent was obtained from each participant, and the study was approved by the ethics committee of the Universidad Industrial de Santander. The dataset captured synchronized audio and video modalities, with participants performing sustained vowel pronunciation used in the clinical routine. All recordings were conducted in the same environment using a Nikon D3500 digital camera with an integrated monaural microphone. Video was recorded at 1080p resolution and 60 fps, focusing on the face region, while audio was captured at a sampling rate of 48 kHz. Phonation patterns included the pronunciation of five vowels, each vowel being repeated three times, providing a comprehensive dataset for phonation and articulatory analysis. In the study, participants are asked to sustain the pronunciation of vowels for about 5 seconds. This exercise is incorporated into clinical routines to detect voice abnormalities and to observe the facial expressions of the individuals. The proposed approach was validated with the oral task of sustained vowels, which allows the identification of voice impairments such as dysarthria during PD diagnosis, but also to peculiar conditions such as strengthening vocal muscles and motor coordination during rehabilitation therapies. For validation was followed leave-one-patient-out cross-validation, where at each iteration, one patient is left out for testing and the remaining ones (13 subjects in our experiment) are used for training. To evaluate the performance of the multimodal prediction, the implemented model configurations were assessed for the sensitivity, specificity, accuracy, precision, and F1-score per video. A video was considered correctly predicted by majority vote of its temporal predictions. Specifically, table metrics were quantified by considering either 5, or 10 or 15 predictions during each video.

Table 22. Hypomimia video classification (facial expression alone) with different number of video slices and polar coordinates of landmarks.

Facial Features	Predictions per video	Ac	Pr	Sen	Spec	F1-s
Phase	5	0.5	0.5	0.69	0.3	0.58
	10	**0.65**	**0.62**	**0.78**	**0.52**	**0.69**
	15	0.4	0.4	0.41	0.39	0.41
Distance	5	0.59	0.58	0.64	0.54	0.61
	10	0.56	0.55	0.66	0.46	0.6
	15	0.55	0.55	0.56	0.53	0.56
Phase and Distance	5	0.58	0.58	0.56	0.6	0.57
	10	0.57	0.57	0.58	0.56	0.57
	15	0.41	0.4	0.39	0.43	0.4

B.4. RESULTS

A first validation was carried out to establish the best video representation to classify PD according to hypomimia-encoded patterns. In this experiment, the temporal covariance matrices were built from landmarks information using only phase (dimension of 44×44), only distance (dimension of 44×44), and integrating both variables (dimension of 88×88). These experiments were also evaluated in different temporal intervals, by evenly dividing the video in five, ten, and fifteen slices respectively. Table 22 summarizes the achieved results, reporting the best performance with the covariance descriptor using only phase information. These results highlight a high sensitivity of 78%, with an accuracy of 65%, evidencing a capability to capture motor coordination changes, especially with 10 slices per video.

In a second evaluation the audio branch was assessed concerning its capability to classify dysarthria patterns from temporal covariance matrices of spectrograms only, with 20 and 50 frequency bands. Each configuration was also evaluated with five, ten, and fifteen slices per video. Table 23 summarizes the achieved results, reporting a better score with the configuration of 20 frequencies and ten slices (sensitivity of 64%). The improvement in results with 20 frequencies in sustained vowel pronunciation could be attributed to a higher generalization capacity or efficiency in representing relevant features for detecting individuals with

Parkinson. It is possible that the learning covariance model can extract more discriminative information with fewer dimensions, facilitating the identification of distinctive patterns in the case of 20 frequencies.

Table 23. Dysarthria Audio classification with different frequencies and different number of video slices

Freqs	Predictions per video	Ac	Pr	Sen	Spec	f1-s
20	5	0.52	0.52	0.6	0.45	0.55
	10	**0.62**	**0.61**	**0.64**	**0.6**	**0.62**
	15	0.57	0.57	0.58	0.56	0.57
50	5	0.54	0.54	0.5	0.57	0.52
	10	0.55	0.55	0.51	0.58	0.53
	15	0.53	0.54	0.5	0.56	0.52

Table 24. Multimodal (audi-video) classification with 20 speech frequencies, phase and distance facial features

Fusion Features	Predictions per video	Ac	Pr	Sen	Spec	f1-s
20 freqs, phase	5	0.44	0.44	0.44	0.45	0.44
	10	0.66	0.65	0.65	0.65	0.66
	15	0.65	0.64	0.64	0.62	0.67
20 freqs, Distance	5	0.6	0.59	0.69	0.56	0.61
	10	0.58	0.61	0.61	0.58	0.59
	15	0.64	0.63	0.63	0.6	0.65
20 freqs, Distance, Phase	5	0.58	0.58	0.58	0.54	0.6
	10	**0.70**	**0.69**	**0.73**	**0.68**	**0.71**
	15	0.62	0.62	0.62	0.64	0.61

Then, in a third experiment, the proposed approach was evaluated by fusing vocal spectrogram frequencies with facial landmark phases and distances. In such cases, it was considered 20 frequency bands for audio, and whole facial configurations. Table 24 summarizes the achieved results with multimodal configurations, being the best performance achieved in the third experiment, where vocal frequencies were fused with both facial landmark phase and distance, improving accuracy to 70% (10 intervals). These results highlight the complemen-

tarity and synergy of features extracted from both modalities. Also, the temporal interval of ten frames shows an appropriate trade-off to capture pronunciation dynamics and avoiding excessive fragmentation of the task.

Besides, the probability for the multimodal approach was calculated for patients and control subjects, for each video percentage during the sustained vowel pronunciation (see Figure 21). The performance remains stable for both Parkinson's and control groups, suggesting that all vocalization phases can yield similar predictions. Figure 23 (resp. Figure 22) shows the probability predictions and accuracy for the pronunciation of open vowels, in Spanish: A, E and O (resp. closed vowels: I and U), for Control and Parkinson groups at each interval per video. Interestingly, this categorization is related to movement: Closed vowels are produced with minimal mouth cavity amplitude, while open vowels involve greater mouth cavity expansion with the tongue positioned low. The Figure 22 of the closed vowels shows greater consistency in the control groups, maintaining the average probability and its stable variability. As for the Parkinson group, higher and more variable results were observed during the initial pronunciation of closed vowels. Similarly, in Figure 23 of the open vowels, the Parkinson group presents greater variability in the initial intervals. But in contrast to the group of closed vowels for the Control group, the best-predicted values (closer to zero) are found in intermediate pronunciation stages. These results show different dynamics for each vowel group in patients and control subjects. The pronunciation is divided into three phases: initial, stabilization, and decay [134]. Figure 24 indicate in the initial phase (predictions at 10%, 20%, and 30%), there is significant effort, with pronounced facial muscle movements. The most discriminative predictions in this phase are 20% considering all vowels (blue line) with a mean accuracy of 72%. The stabilization phase (predictions from 40% to 70%) represents the maximum vocal production stability, with constant acoustic characteristics and minimal facial movement. The most discriminative intervals here are at 50% of videos with a mean

[134] Kumud TRIPATHI and K Sreenivasa RAO. "Robust vowel region detection method for multimode speech". In: *Multimedia Tools and Applications* 80.9 (2021), pp. 13615–13637.

accuracy of 70% (blue line). Finally, the decay phase (predictions at 80%, 90%, and 100%) shows a decline in vocal production and increased facial movement until the mouth closes. The most discriminative intervals in this phase are at 90% of video with a mean accuracy of 68% (blue line). The red and green line indicate that accuracy trends for both open and closed vowels remain relatively stable across video percentages, suggesting that prediction variability does not significantly change, indicating robustness in results. For control subjects, the most discriminative percentages are 20% for closed vowels (red line) with a mean accuracy of 76% (initial stage) and 60% for open vowels (green line) with a mean accuracy of 72% (stabilization stage). Future works will include the analysis of enriched representations with other input modalities, as well as an investigation toward an end-to-end processing of the complete information, since vowels are versatile and can combine with a variety of consonants to create a wide range of sounds and words. Also, this study will be extended to other voice instructions to explore the capabilities of the proposed approach.

B.5. DISCUSSION AND CONCLUSIVE REMARKS

This work introduced an online multimodal approximation to classify Parkinson disease from facial expression (hypomimia) and voice patterns (dysarthria). In the literature there exist ev-

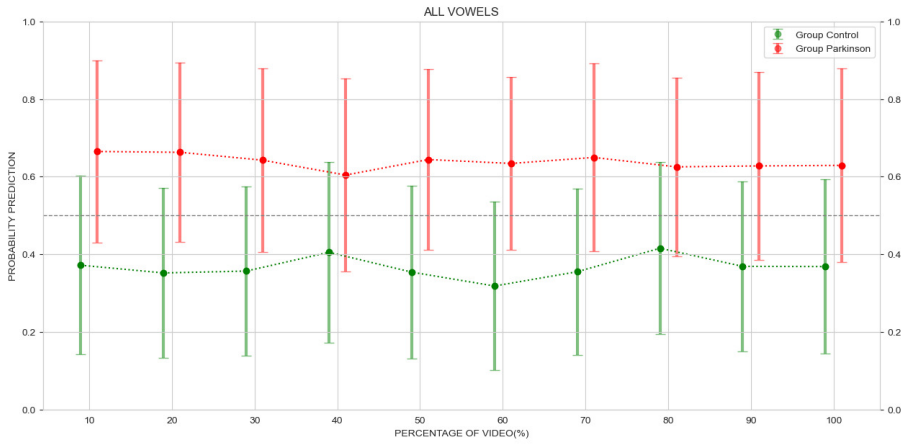

Figure 21. Probability prediction per interval of video (red line and green line), for all vowels

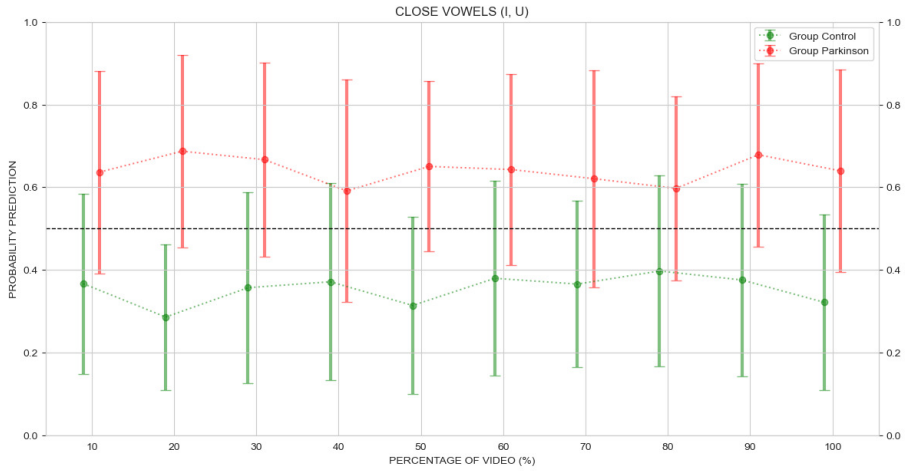

Figure 22. Probability per interval of video (red line and green line), for close vowels

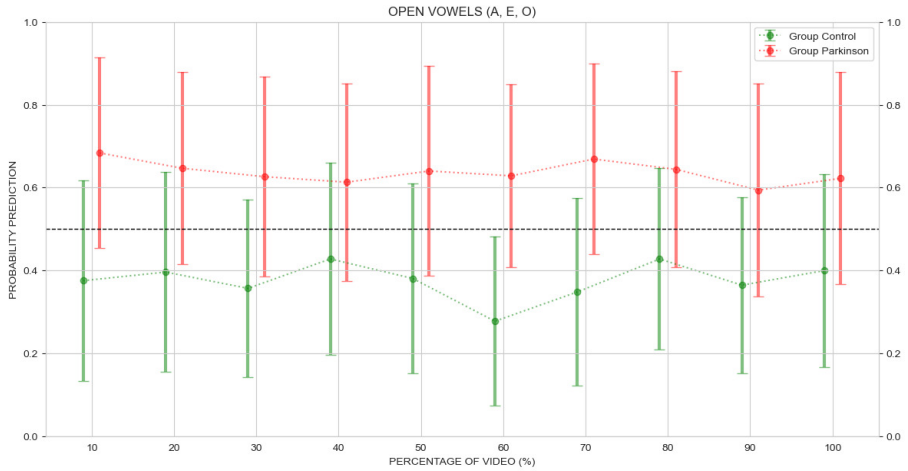

Figure 23. Probability per interval of video (red line and green line), for open vowels

idences that dysarthria, through the pronunciation of sustained vowels, can identify speech difficulties, associated with early Parkinson's disease [135]. Additionally, treatments have been

[135] Virginie ROLAND et al. "Vowel production: a potential speech biomarker for early detection of dysarthria in Parkinson's disease". In: *Frontiers in Psychology* 14 (2023), p. 1129830.

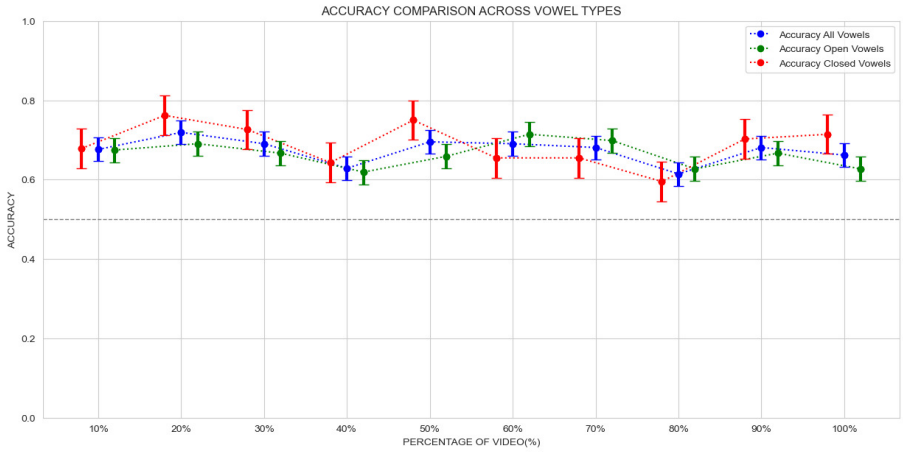

Figure 24. Accuracy per interval of video for close vowels (red line), open vowels (green line) and all vowels (blue line).

proposed that use vowel pronunciation in the attempt to improve these impairments [136]. Different studies have integrated voice with other modalities to achieve a broader range of motor impairments in patients and improve diagnostic prediction [137, 138]. For example, they have integrated voice, gait, and tremor to extract kinematic features and classify between Parkinson's and Control using machine learning techniques [137]. Alternatively, the voice modality has been integrated with videos of smile expression and finger tapping. In this approach, landmarks are used to identify key points of movement, and models are trained in independent branches. An intermediate fusion is performed using a convolutional architecture, followed by classification [138]. However, this method integrates unsynchronized modalities,

---

[136] Sheila WIGHT and Nick MILLER. "Lee Silverman Voice Treatment for people with Parkinson's: audit of outcomes in a routine clinic". In: *International journal of language & communication disorders* 50.2 (2015), pp. 215–225.

[137] Juan Rafael OROZCO-ARROYAVE et al. "Apkinson: the smartphone application for telemonitoring Parkinson's patients through speech, gait and hands movement". In: *Neurodegenerative Disease Management* 10.3 (2020), pp. 137–157.

[138] Md Saiful ISLAM et al. "Accessible, At-Home Detection of Parkinson's Disease via Multi-task Video Analysis". In: *arXiv preprint arXiv:2406.14856* (2024).

making it difficult to analyze and identify how the modalities and their associated symptoms are correlated. Considering that, this work reported a multimodal approach that integrates visual and audio information to recover hypomimia and dysarthria-associated patterns. For doing so, the proposed approach captured face landmarks in video, and coded spectrograms from audio, which are integrated into temporal covariance descriptors, allowing to obtain a representation of bimodal vocalization. Then, this temporal covariance embedding is projected to a geometrical deep architecture to obtain a refined second-order representation with the ability to distinguish Parkinson patterns from control signals. Thanks to the sliding nature of time covariance descriptors, the geometrical net can bring a prediction at each time interval, allowing to detect abnormal patterns associated to PD, during the exercise, in clinical routine. The proposed geometrical representation was validated with respect to isolated video and audio patterns, and also with the integration of both modalities. Using only videos, the proposed approach encodes temporal covariance matrices using only the correlation among face landmarks. In such case, the proposed approach achieved 65% of accuracy, a f1-score of 69%, and a total of 4 Parkinson and 5 Control subjects were correctly classified. The mistakes in classification may be partially associated to instability of landmarks and recording conditions, but also to the limitation of visual information alone to determine Parkinsonian patterns. Regarding, an audio geometrical net, trained using only spectrogram voice information, was obtained an accuracy of 62% and a f1-score of 62%. These scores were achieved from a configuration of 20 frequencies an 10 intervals per video. Then, we conducted multimodal experiments using a geometrical net, learning from covariance matrices encoding the two modalities. In such case, the multimodal approximation has a gain of 5% and 8% in accuracy, and a gain of 2% and 9% in f1-score. The proposed approach, however, needs to be examined in a larger cohort of patients to determine statistical significance within the affected population. Additionally, it is crucial to design mechanisms that output disease stages based on observational scales, enabling their use in tracking disease progression.

C. APPENDIX: DESCRIPTION OF THE DEVELOPED DATASET

This appendix describes the main characteristics considered during the construction of the dataset. Initially, a demographic and statistical analysis of the population and recorded patients were conducted. Subsequently, the protocol was developed with the assistance of a neurologist. Then, the recordings were carried out, and the neurologist's evaluations were included.

C.1. DEMOGRAPHIC AND STATISTICAL ANALYSIS OF THE PARKINSON'S DISEASE POPULATION.

C.1.1. National and International Demographic Analysis of Parkinson's Disease. In the global context, Parkinson's disease is now recognized as the fastest-growing neurological disorder in terms of prevalence, disability, and mortality. Data from the Global Burden of Disease Study (GBD) indicate that the number of individuals living with PD worldwide increased from approximately 2.5 million in 1990 to over 6.1 million by 2016, with estimates projecting more than 12 million cases by 2040, largely driven by increased life expectancy and population aging [139]. Although global prevalence are between 100 and 300 cases per 100 000 inhabitants, high-income countries often report higher rates due to more robust diagnostic capabilities, while underdiagnosis and lack of reporting persist in low and middle income countries, as reflected in the Colombian case (30.7 cases per 100 000 inhabitants [140]). The accelerated aging of the Colombian population, as projected by the National Administrative

---

[139] E DORSEY et al. "The emerging evidence of the Parkinson pandemic". In: *Journal of Parkinson's disease* 8.s1 (2018), S3–S8.

[140] Jorge Luis SÁNCHEZ et al. "Prevalence of Parkinson's disease and parkinsonism in a Colombian population using the capture-recapture method". In: *international Journal of Neuroscience* 114.2 (2004), pp. 175–182.

Department of Statistics (DANE), indicates that by 2030, more than 20% of Colombians will be over the age of 60 [141]. This demographic trend could potentially double the burden of PD in the coming decades. Such a scenario demands a comprehensive response from the healthcare system, which currently faces significant limitations in terms of timely diagnosis and access to specialized neurologists. In this context, alternatively, computer vision methods and the use of artificial intelligence could partially address these challenges by serving as diagnostic support tools, particularly in remote areas of the country.

C.1.2. Statistical sampling of patients with Parkinson's disease According to data reported in the Individual Registry of Health Service Provision (RIPS), a total of 148 224 individuals were treated with a diagnosis of Parkinson's disease [142]. The prevalence of the disease in Colombia varies by age group. For instance, the prevalence is 0.7% among individuals aged $70\breve{}74$, 1% in those aged $75\breve{}79$, and 1.4% in individuals over 80 years of age [143]. To obtain a more conservative estimate, the calculations will be based on a prevalence of 1.4% ($\rho = 0.014$). The following standard formula will be used to calculate the sample size: $\frac{(z^2\rho(1-\rho))\div e^2}{1+(z^2\rho(1-\rho))\div e^2 N)}$ where N corresponds to the number of patients with Parkinson's disease, (e) corresponds to the margin of error. Z corresponds to the Z-score, which depends on the desired confidence level. (ρ) represents the estimated proportion of the Parkinson's population that has a characteristic of interest.

---

[141] MINISTERIO DE SALUD Y PROTECCIÓN SOCIAL DE COLOMBIA. *Análisis de situación de salud (ASIS) Colombia 2020*. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/ED/PSP/asis-2020-colombia.pdf. Accedido el 21 de junio de 2025. 2020.

[142] MINISTERIO DE SALUD Y PROTECCIÓN SOCIAL DE COLOMBIA. *Día Mundial del Parkinson: Colombia se destaca en atención*. https://www.minsalud.gov.co/Paginas/Dia-Mundial-del-Parkinson-Colombia-se-destaca-en-atencion.aspx. Accedido el 21 de junio de 2025. 2022.

[143] Ángela G RINCON-MONTANA et al. "Prevalencia, características demográficas de la enfermedad de Parkinson y comorbilidades asociadas: un análisis del registro oficial del Ministerio de Salud de Colombia". In: *Acta Neurológica Colombiana* 41.1 (2025).

Table 25. Sample sizes for different confidence intervals and margins of error

Parkinson population	Confidence interval	Z-score	Margin of error (e)	Sample size of Parkinson
148 224 Parkinson prevalence (p=0.014)	0.85	1.44	0.05	11
			0.03	32
	0.90	**1.64**	**0.05**	**15**
			0.03	41
	0.95	1.96	0.05	21
			0.03	59
	0.99	2.57	0.05	37
			0.03	102

The dataset developed in this study comprises 32 participants, including 19 patients diagnosed with Parkinson's disease and 13 control subjects. The current sample size of 15 patients meets the minimum requirement for a 90% confidence level with a 5% margin of error. In addition, it may be made available to the scientific community, provided that prior written consent is obtained from the requesting researcher [88].

C.2. MULTIMODAL DATA

The study involved 13 control subjects (average age 72.2 ± 6.1 years) and 19 PD patients (average age 72.3 ± 7.4 years). The following section provides a detailed explanation of the acquisition protocol, the preprocessing of the captured data, and the motor characterization of the dataset as determined by an expert neurologist.

C.2.1. Protocol for the Acquisition and Preprocessing of Multimodal Data Regarding the data acquisition, each participant underwent 10 gait recordings (5 in which the individual moved from left to right and 5 in the opposite direction) and 10 smooth ocular motion recordings (5 for the right eye and 5 for the left eye) with a conventional camera Nikon D3200 with

spatial resolution of 1280×720. To ensure participant safety, recordings were only conducted if the patient was under the effect of medication. Before beginning, the patient was asked whether they had taken their medication in the morning, as the recordings were conducted in the early hours to minimize the risk of falls. Although the videos were generally of high quality, controlled lighting conditions were maintained throughout the study. To preserve the symmetry of laterality, 8 out of the 10 recordings per modality were selected. Some recordings were discarded when participants got out of focus during the gait or ocular motion task. During the study, the participants were instructed to adhere to the following movement instructions:

- **Gait.** The participants were invited to walk in a straight line while the camera captured them from a sagittal view. Locomotion was recorded along a 5-meter walk, with an average video duration of 5 seconds per recording. Each video had a spatial resolution of 520 × 520 pixels and a temporal resolution of 60 frames per second (fps). The recording environment was standardized in terms of lighting and flooring to reduce variability. Participants were asked to walk at a self-selected pace, and up to three trials were performed to ensure consistency. The camera was placed at a fixed height of 1 meter and a distance of 5 meters from the center of the walking path. Figure 25 illustrates the postural configuration during the gait exercise, realized in front of a uniform green background. In the Figure 26 is illustrated the gait locomotion for two PD-affected patients, in different stages of the disease. As observed, patients with unilateral impairment tend to exhibit more evident motor difficulties on one side of the body, whereas those with bilateral impairment experience a greater reduction in step amplitude and more pronounced postural alterations. These differences can influence model predictions, as motor patterns vary significantly depending on disease progression. These views illustrate the diversity of clinical manifestations captured in the dataset, facilitating the interpretation of the model's results.

- **Smooth ocular motion.** In this scenario, patients were instructed to maintain their gaze on a spotlight projected onto a screen with a dark background. The monitor's

height was adjusted to align the center of the screen with the center of the pupillary plane, as depicted in Figure 25 The motion of the spotlight was controlled both horizontally (from right to left and vice versa) and vertically (from top to bottom and vice versa). Recordings were performed with uniform ambient lighting to minimize reflections or pupil dilation effects. The recorded video was subsequently manually cropped to a region of interest around the eye, with dimensions of 210×140 pixels. In the Figure 27 is illustrated the ocular motion for a control and for a PD patient, synchronized according to the spotlight projected onto a screen. As observed, the PD patient shows a delay in tracking the point, evidencing an anomaly in the correct execution of the exercise.

Sixteen videos were obtained for each participant, encompassing both gait and ocular smooth motion exercises. For both modalities, the same camera, a conventional Nikon D3200, was utilized, providing a spatial resolution of 1280 × 720. Consequently, the entire dataset for this study comprises a total of 512 videos. The recording of this dataset was approved by the Ethics Committee for Scientific Research at the Universidad Industrial de Santander (CEINCI-UIS). Additionally, the study was classified as minimal risk, as it involved video recordings using conventional cameras without any invasive procedures. All participants signed an informed consent form, and the data were anonymized and coded, with their use restricted to academic purposes. Furthermore, participants authorized the use of their data for future studies, and the ethics committee conducted periodic reviews to ensure data integrity.

Preprocessing Characteristics: Before training, all videos were normalized to approximately 5 seconds by subsampling frames equidistantly throughout the video. The gait recordings captured the complete displacement of the patient. Regarding ocular motion, initial recordings included the entire facial expression, but the eye region was manually cropped while maintaining a spatial resolution of 210×140 pixels. The labeling process was con-

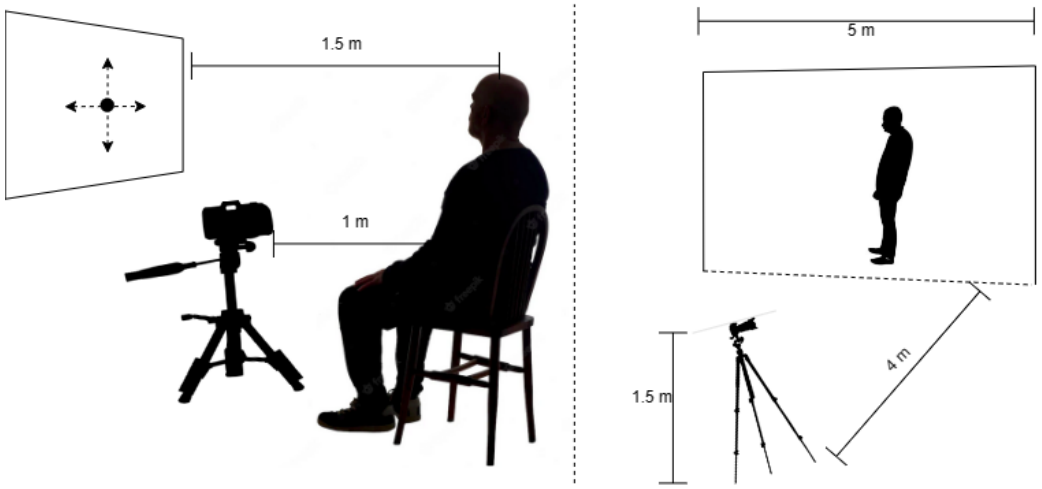

Figure 25. Gait and ocular smooth motion acquisition setups for our markerless video dataset. On the left is represented how oculomotor video sequences are recorded while the patient maintains his/her gaze on a spotlight projected onto a screen. On the right is illustrated how walking videos are recorded from a sagittal view.

ducted manually by a neurologist trained in standardized evaluation scales. Each ocular video received a single annotation, while each gait video was assigned five distinct labels corresponding to motor impairments: ocular bradykinesia, gait bradykinesia, bilateral gait impairment, postural instability, freezing of gait, and gait autonomy.

C.2.2. Confounding Variables. The study was limited to a single clinical center. Additionally, gender and age were evaluated as potential confounding variables, and adjusted odds ratios were calculated using a logistic regression procedure. In this model, the predictor variables x_i are linearly related to the log-odds of the outcome y (diagnosis). For three predictors, this is expressed as:

$$\ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

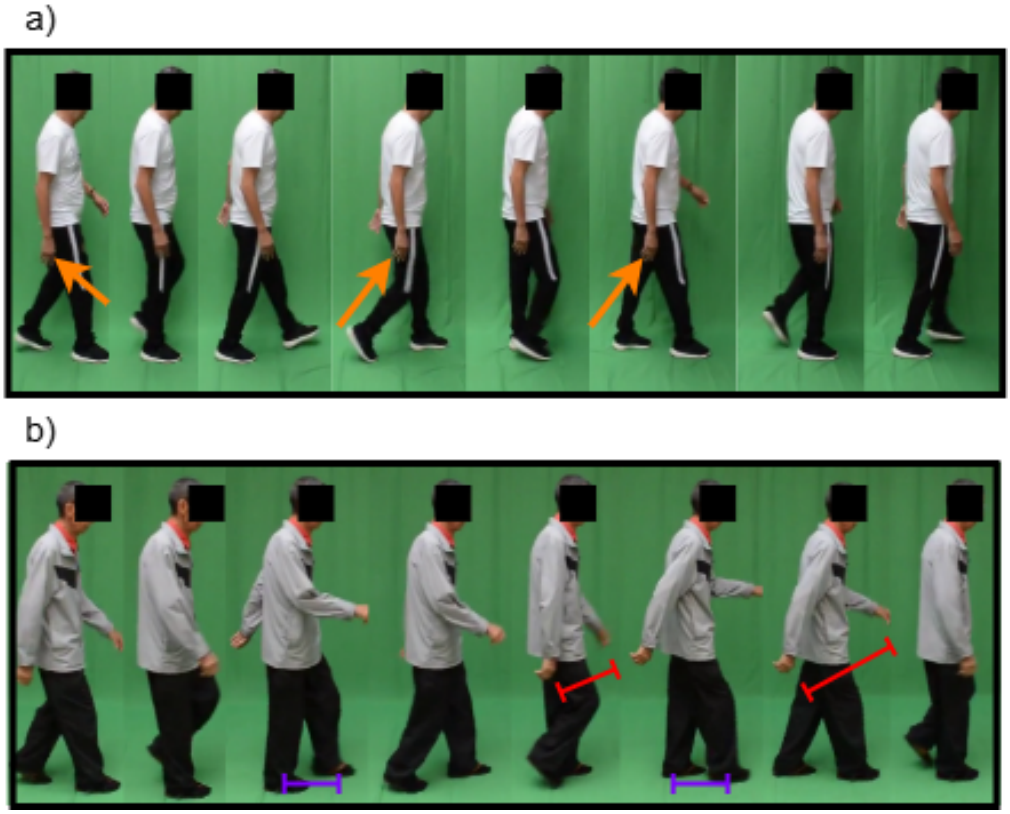

Figure 26. Two patients during gait recording. (a) Patient with unilateral impairment: the right arm exhibits reduced mobility (see orange arrows), and coordination with the left arm is impaired. (b) Patient with advanced impairment: arm swing amplitude is uncoordinated (see red lines), and there is also a noticeable reduction in step length relative to arm swing amplitude (see purple lines).

where β_1 is the coefficient corresponding to gender, and β_2 is the coefficient corresponding to age and β_0 is the intercept. The adjusted odds ratios (aOR) for gender and age were calculated as e^{β_i}, respectively (1.03 and 1.16). Corresponding p-values were estimated using a custom statistical approach to evaluate the significance of each predictor (0.65 and 0.84). These results indicate that neither of these variables had a statistically significant effect on the diagnosis, suggesting that the model was not substantially influenced by the bias in the gender and age distribution of the participants.

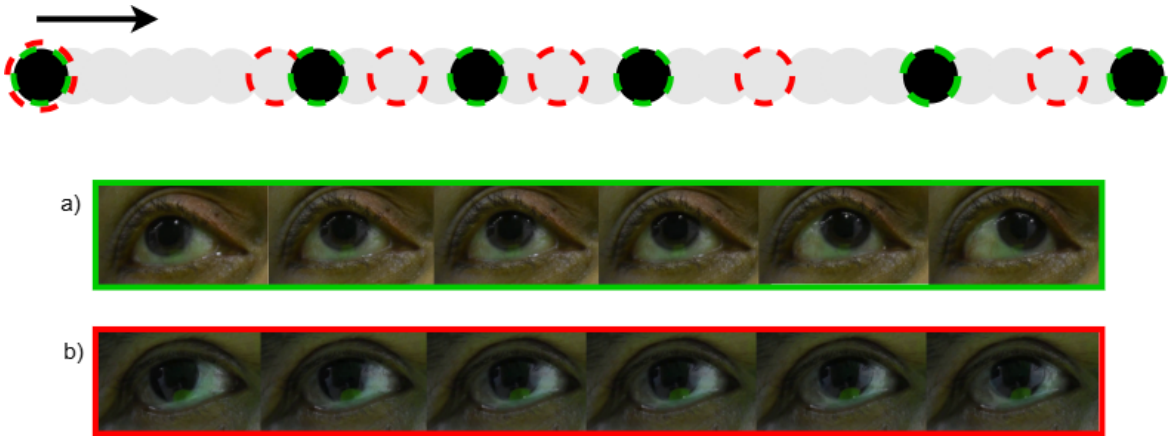

Figure 27. Example of smooth eye movement. In green: smooth eye movement of a control subject, who appropriately follows the point on the screen (see Figure a). In red: eye movement of a patient who has difficulty tracking the point on the screen. The patient's reflexes and eye movements are generally slower compared to those of the control subject (see Figure b).

C.2.3. Clinical Distribution of Expert-Labeled Motor Impairment Items. This subsection presents a descriptive analysis of the distribution of six motor items evaluated in this study. This visualization facilitates the clinical characterization of patients and the degree of observed impairment. In Figure 28, three different types of assessments are presented. The first, based on the modified Hoehn and Yahr (H&Y) scale, shows patients stratified into early stages (stage 1 and 1.5), intermediate stages (stage 2), and more advanced stages (stage 3). The second assessment is based on the MDS-UPDRS scale evaluating four predominant symptoms, categorized as: normal (0), slight (1), mild (2), and moderate (3). The third assessment is non-standardized but relies on identifying the presence or absence of ocular bradykinesia. In the figure can be observed that patients within the same H&Y stage exhibit varying levels of impairment in the key symptoms assessed by the MDS-UPDRS scale. These two scales provide complementary perspectives: the H&Y scale offers a general overview of disease severity, while the MDS-UPDRS focuses on specific and predominant impairments. The third evaluation highlights the presence of ocular bradykinesia across different stages of the disease, suggesting its potential as a clinical indicator. To the best

of our knowledge, this type of dataset characterization using multiple motor scales has not been previously explored in the state of the art, constituting one of the contributions of this study.

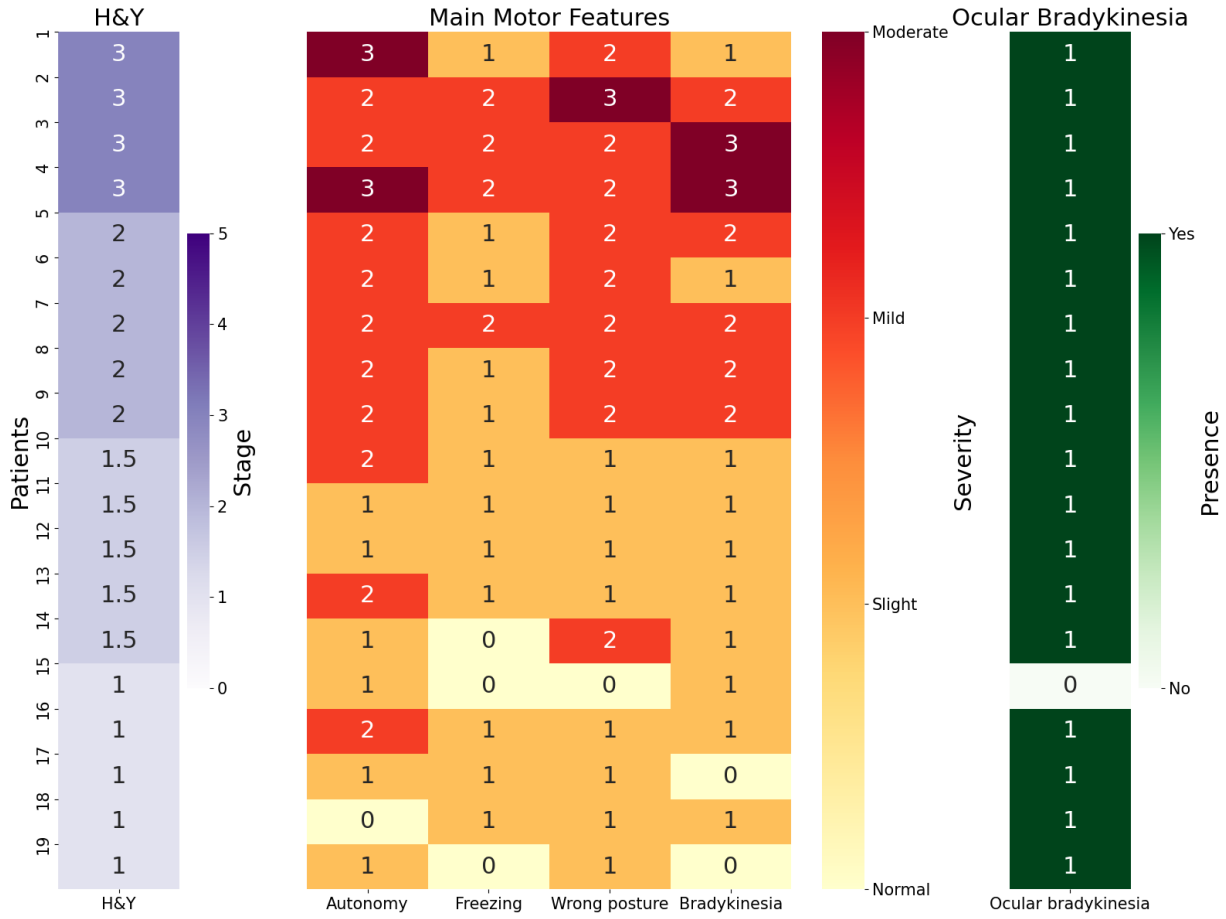

Figure 28. Characterization of patients according to the H&Y scale, four cardinal symptoms from the MDS-UPDRS scale, and a non-standardized observation of ocular bradykinesia

BIBLIOGRAPHY

ABROMAVIČIUS, Vytautas et al. “Two-stage monitoring of patients in intensive care unit for sepsis prediction using non-overfitted machine learning models”. In: *Electronics* 9.7 (2020), p. 1133 (cit. on p. 57).

AKODAD, Sara et al. “Ensemble learning approaches based on covariance pooling of CNN features for high resolution remote sensing scene classification”. In: *Remote Sensing* 12.20 (2020), p. 3292 (cit. on p. 31).

ALAM, Md Zahangir; RAHMAN, M Saifur, and RAHMAN, M Sohel. “A Random Forest based predictor for medical data classification using feature ranking”. In: *Informatics in Medicine Unlocked* 15 (2019), p. 100180 (cit. on p. 54).

ALI, Jehad et al. “Random forests and decision trees”. In: *International Journal of Computer Science Issues (IJCSI)* 9.5 (2012), p. 272 (cit. on p. 54).

ALPER, Mehmet Akif; GOUDREAU, John, and DANIEL, Morris. “Pose and Optical Flow Fusion (POFF) for accurate tremor detection and quantification”. In: *Biocybernetics and Biomedical Engineering* 40.1 (2020), pp. 468–481 (cit. on p. 130).

ARCHILA, John; MANZANERA, Antoine, and MARTINEZ, Fabio. “A multimodal Parkinson quantification by fusing eye and gait motion patterns, using covariance descriptors, from non-invasive computer vision”. In: *Computer Methods and Programs in Biomedicine* (2021), p. 106607 (cit. on pp. 29, 46, 94, 109, 110).

ARCHILA, John; MANZANERA, Antoine, and MARTÍNEZ, Fabio. “A recurrent approach for predicting Parkinson stage from multimodal videos”. In: *17th International Symposium on*

Medical Information Processing and Analysis. Vol. 12088. SPIE. 2021, pp. 37–45 (cit. on p. 128).

ARCHILA, John; MANZANERA, Antoine, and MARTÍNEZ, Fabio. “A Riemannian multimodal representation to classify parkinsonism-related patterns from noninvasive observations of gait and eye movements”. In: *Biomedical Engineering Letters* 15.1 (2025), pp. 81–93 (cit. on pp. 75, 154).

ARCHILA, John; MANZANERA, Antoine, and MARTINEZ CARRILLO, Fabio. “A Mixed audio-video SPD network for online classification of Parkinsonian speech patterns”. In: *IBERAMIA 2024. 18th Ibero-American Conference on Artificial Intelligence*. Montevideo, Uruguay, Nov. 2024 (cit. on p. 140).

ARCHILA, John et al. “A multimodal gait and ocular geometric representation to generate a Parkinson progression report”. In: *Engineering Applications of Artificial Intelligence* 160 (2025), p. 111834. DOI: https://doi.org/10.1016/j.engappai.2025.111834 (cit. on p. 97).

ARMSTRONG, RA. “Oculo-visual dysfunction in Parkinson’s disease”. In: *Journal of Parkinson’s disease* 5.4 (2015), pp. 715–726 (cit. on p. 27).

BAN, Rebecca et al. “Dynamic gait stability in people with mild to moderate Parkinson’s disease”. In: *Clinical Biomechanics* 118 (2024), p. 106316 (cit. on pp. 90, 91).

BARACHANT, Alexandre et al. “Riemannian geometry applied to BCI classification”. In: *International conference on latent variable analysis and signal separation*. Springer. 2010, pp. 629–636 (cit. on p. 31).

BASHIR, Khalid et al. “Gait Representation Using Flow Fields.” In: *BMVC*. 2009, pp. 1–11 (cit. on p. 130).

BELIĆ, Minja et al. “Artificial intelligence for assisting diagnostics and assessment of Parkinson’s disease—A review”. In: *Clinical neurology and neurosurgery* 184 (2019), p. 105442 (cit. on p. 33).

BIASE, Lazzaro di et al. “Parkinson’s disease wearable gait analysis: kinematic and dynamic markers for diagnosis”. In: *Sensors* 22.22 (2022), p. 8773 (cit. on p. 24).

BLOEM, Bastiaan R et al. “M easurement instruments to assess posture, gait, and balance in P arkinson’s disease: Critique and recommendations”. In: *Movement Disorders* 31.9 (2016), pp. 1342–1355 (cit. on p. 117).

BREDEMEYER, Oliver et al. “Oculomotor deficits in Parkinson’s disease: Increasing sensitivity using multivariate approaches”. In: *Frontiers in Digital Health* 4 (2022), p. 939677 (cit. on p. 27).

BREIMAN, Leo et al. *Classification and regression trees*. CRC press, 1984 (cit. on p. 54).

BRIEN, Donald C et al. “Classification and staging of Parkinson’s disease using video-based eye tracking”. In: *Parkinsonism & Related Disorders* 110 (2023), p. 105316 (cit. on p. 28).

BRONSTEIN, Michael M et al. “Geometric deep learning: going beyond Euclidean data”. In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 18–42 (cit. on pp. 78, 143).

BROOKS, Daniel et al. “Riemannian batch normalization for SPD neural networks”. In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on p. 79).

CAPRONI, Stefano and COLOSIMO, Carlo. “Diagnosis and differential diagnosis of Parkinson disease”. In: *clinics in geriatric medicine* 36.1 (2020), pp. 13–24 (cit. on p. 30).

CHERIET, Mohamed et al. “Multi-speed transformer network for neurodegenerative disease assessment and activity recognition”. In: *Computer Methods and Programs in Biomedicine* 230 (2023), p. 107344 (cit. on p. 25).

CHICCO, Davide and JURMAN, Giuseppe. “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation”. In: *BMC genomics* 21.1 (2020), p. 6 (cit. on p. 57).

CICIRELLI, Grazia et al. “Human Gait Analysis in Neurodegenerative Diseases: a Review”. In: *IEEE Journal of Biomedical and Health Informatics* (2021) (cit. on pp. 81, 137).

CLARKE, Andrew H. “Laboratory testing of the vestibular system”. In: *Current opinion in otolaryngology & head and neck surgery* 18.5 (2010), pp. 425–430 (cit. on p. 130).

CROUSE, Jacob J et al. “Postural instability and falls in Parkinson’s disease”. In: *Reviews in the Neurosciences* 27.5 (2016), pp. 549–555 (cit. on p. 21).

DAGAN, Moria et al. “The role of the prefrontal cortex in freezing of gait in Parkinson’s disease: insights from a deep repetitive transcranial magnetic stimulation exploratory study”. In: *Experimental brain research* 235 (2017), pp. 2463–2472 (cit. on p. 20).

DENG, Jia et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. on p. 34).

DORSEY, E et al. “The emerging evidence of the Parkinson pandemic”. In: *Journal of Parkinson’s disease* 8.s1 (2018), S3–S8 (cit. on p. 152).

EKKER, Merel S. et al. “Ocular and visual disorders in Parkinson’s disease: Common but frequently overlooked”. In: *Parkinsonism & Related Disorders* 40 (2017), pp. 1–10. DOI: 10.1016/j.parkreldis.2017.02.014 (cit. on pp. 24, 26, 129).

FARASHI, Sajjad. “Analysis of vertical eye movements in Parkinson’s disease and its potential for diagnosis”. In: *Applied intelligence* 51.11 (2021), pp. 8260–8270 (cit. on p. 92).

FARNEBÄCK, Gunnar. “Two-frame motion estimation based on polynomial expansion”. In: *Scandinavian conference on Image analysis*. Springer. 2003, pp. 363–370 (cit. on pp. 48, 130).

FEIGIN, Valery L; NICHOLS, Emma, and ALAM Tahiya, et al. “Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016”. In: *The Lancet Neurology* 18.5 (2019), pp. 459–480. DOI: 10.1016/S1474-4422(18)30499-X (cit. on pp. 19, 32).

FERESHTEHNEJAD, Seyed-Mohammad et al. “New clinical subtypes of Parkinson disease and their longitudinal progression: a prospective cohort comparison with other phenotypes”. In: *JAMA neurology* 72.8 (2015), pp. 863–873 (cit. on p. 117).

FLETCHER, P Thomas and JOSHI, Sarang. “Riemannian geometry for the statistical analysis of diffusion tensor data”. In: *Signal Processing* 87.2 (2007), pp. 250–262 (cit. on pp. 52, 54, 76).

FREI, Karen. “Abnormalities of smooth pursuit in Parkinson’s disease: A systematic review”. In: *Clinical parkinsonism & related disorders* 4 (2021), p. 100085 (cit. on p. 27).

FU, Rongrong et al. “Dynamical differential covariance based brain network for motor intent recognition”. In: *IEEE Sensors Journal* 24.5 (2024), pp. 6515–6522 (cit. on p. 31).

GITCHEL, George T; WETZEL, Paul A, and BARON, Mark S. “Pervasive ocular tremor in patients with Parkinson disease”. In: *Archives of neurology* 69.8 (2012), pp. 1011–1017 (cit. on pp. 27, 71).

GITCHEL, George T et al. “Experimental support that ocular tremor in Parkinson’s disease does not originate from head movement”. In: *Parkinsonism & related disorders* 20.7 (2014), pp. 743–747 (cit. on pp. 23, 26).

GOETZ, CG et al. “MDS-Unified Parkinson’s Disease Rating Scale (MDS-UPDRS)”. In: *Available from the International Parkinson and Movement Disorder Society website: https://www. movementdisorders. org/MDS/MDS-Rating-Scales/MDS-Unified-Parkinsons-Disease-Rating-Scale-MDS-UPDRS. htm* (2008) (cit. on pp. 20, 112).

GOETZ, Christopher G et al. “Movement Disorder Society-sponsored revision of the Unified Parkinson’s Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results”. In: *Movement disorders: official journal of the Movement Disorder Society* 23.15 (2008), pp. 2129–2170 (cit. on pp. 20, 21).

— “Movement Disorder Society-sponsored revision of the Unified Parkinson’s Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results”. In: *Movement disorders: official journal of the Movement Disorder Society* 23.15 (2008), pp. 2129–2170 (cit. on p. 89).

GOETZ, Christopher G 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”. In: *Movement disorders* 19.9 (2004), pp. 1020–1028 (cit. on pp. 20, 21, 112).

GUAYACÁN, Luis C and MARTÍNEZ, Fabio. “Visualising and quantifying relevant parkinsonian gait patterns using 3D convolutional network”. In: *Journal of biomedical informatics* 123 (2021), p. 103935 (cit. on pp. 25, 34, 90, 91).

GUAYACÁN, Luis Carlos; RANGEL, Edgar, and MARTÍNEZ, Fabio. “Towards understanding spatio-temporal parkinsonian patterns from salient regions of a 3D convolutional network”. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2020, pp. 3688–3691 (cit. on pp. 71, 72).

HANUŠKA, Jaromír et al. “Fast vergence eye movements are disrupted in Parkinson’s disease: a video-oculography study”. In: *Parkinsonism & Related Disorders* 21.7 (2015), pp. 797–799 (cit. on p. 130).

HAWKES, Christopher H; DEL TREDICI, Kelly, and BRAAK, Heiko. “A timeline for Parkinson’s disease”. In: *Parkinsonism & related disorders* 16.2 (2010), pp. 79–84 (cit. on pp. 19–21).

HENDRICKS, Renee M; KHASAWNEH, Mohammad T, et al. “An investigation into the use and meaning of Parkinson’s disease clinical scale scores”. In: *Parkinson’s Disease* 2021 (2021) (cit. on p. 32).

HERRMANN, Carl JJ; METZLER, Ralf, and ENGBERT, Ralf. “A self-avoiding walk with neural delays as a model of fixational eye movements”. In: *Scientific reports* 7.1 (2017), pp. 1–17 (cit. on pp. 133, 135).

HERZ, Damian M and BROWN, Peter. “Moving, fast and slow: behavioural insights into bradykinesia in Parkinson’s disease”. In: *Brain* 146.9 (2023), pp. 3576–3586 (cit. on p. 115).

HIJAZI, Samer; KUMAR, Rishi, and ROWEN, Chris. “Using convolutional neural networks for image recognition”. In: *Cadence Design Systems Inc.: San Jose, CA, USA* (2015), pp. 1–12 (cit. on p. 49).

HOEHN, Margaret M and YAHR, Melvin D. “Parkinsonism: onset, progression, and mortality”. In: *Neurology* 17.5 (1967), pp. 427–427 (cit. on p. 89).

HOWARD, Andrew G et al. “Mobilenets: Efficient convolutional neural networks for mobile vision applications”. In: *arXiv preprint arXiv:1704.04861* (2017) (cit. on p. 50).

HUANG, Jie-Ru; LIU, Mei-Chen, and ZHOU, Jin-Min. “Piano Soundboard Classification Based on Intelligent Neural Network and Multi-feature Fusion Algorithm”. In: (2025) (cit. on p. 31).

HUANG, Zhiwu and VAN GOOL, Luc. “A Riemannian Network for SPD Matrix Learning”. In: *Association for the Advancement of Artificial Intelligence (AAAI)*. 2017 (cit. on p. 99).

HUANG, Zhiwu et al. “A Riemannian network for SPD matrix learning”. In: *Thirty-first AAAI conference on artificial intelligence*. 2017 (cit. on p. 31).

ISLAM, Md Saiful et al. “Accessible, At-Home Detection of Parkinson’s Disease via Multi-task Video Analysis”. In: *arXiv preprint arXiv:2406.14856* (2024) (cit. on p. 150).

JONES RACHEL. “Biomarkers: casting the net wide”. In: *Nature* 466.7310 (2010), S11–S12. DOI: https://doi.org/10.1038/466S11a (cit. on pp. 22, 24).

KAUR, Rachneet et al. “A Vision-Based Framework for Predicting Multiple Sclerosis and Parkinson’s Disease Gait Dysfunctions—A Deep Learning Approach”. In: *IEEE Journal of Biomedical and Health Informatics* 27.1 (2022), pp. 190–201 (cit. on p. 25).

KAY, Will et al. “The kinetics human action video dataset”. In: *arXiv preprint arXiv:1705.06950* (2017) (cit. on p. 34).

KERSBERGEN, Jannis van et al. “Camera-based objective measures of Parkinson’s disease gait features”. In: *BMC research notes* 14 (2021), pp. 1–6 (cit. on pp. 90, 91).

KHOSLA, Ajit and KIM, Dongsoo. *Optical Imaging Devices: New Technologies and Applications*. CRC Press, 2017 (cit. on p. 130).

KOCH, Nils A et al. “Eye movement function captured via an electronic tablet informs on cognition and disease severity in Parkinson’s disease”. In: *Scientific Reports* 14.1 (2024), p. 9082 (cit. on p. 28).

LANG, Anthony E. “The progression of Parkinson disease: a hypothesis”. In: *Neurology* 68.12 (2007), pp. 948–952 (cit. on p. 19).

LARRAZABAL, A.J.; GARCÍA CENA, C.E., and MARTÍNEZ, C.E. “Video-oculography eye tracking towards clinical applications: A review”. In: *Computers in Biology and Medicine* 108 (2019), pp. 57–66. DOI: 10.1016/j.compbiomed.2019.03.025 (cit. on p. 27).

LI, Han et al. “Abnormal eye movements in Parkinson’s disease: From experimental study to clinical application”. In: *Parkinsonism & Related Disorders* (2023), p. 105791 (cit. on p. 26).

LI, Han et al. “Combined diagnosis for Parkinson’s disease via gait and eye movement disorders”. In: *Parkinsonism & Related Disorders* 123 (2024), p. 106979 (cit. on pp. 29, 94).

LI, Tianbai and LE, Weidong. “Biomarkers for Parkinson’s disease: how good are they?” In: *Neuroscience bulletin* 36.2 (2020), pp. 183–194 (cit. on p. 19).

LI, Tianpeng et al. “Automatic timed up-and-go sub-task segmentation for Parkinson’s disease patients using video-based activity classification”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26.11 (2018), pp. 2189–2199 (cit. on pp. 90, 91).

LIM, Shen-Yang and TAN, Ai Huey. “Historical perspective: the pros and cons of conventional outcome measures in Parkinson’s disease”. In: *Parkinsonism & related disorders* 46 (2018), S47–S52 (cit. on p. 22).

LIM, Wee Shin et al. “An integrated biometric voice and facial features for early detection of Parkinson’s disease”. In: *npj Parkinson’s Disease* 8.1 (2022), p. 145 (cit. on pp. 109, 110).

LIN, Chia-Hung et al. “Tremor Class Scaling for Parkinson Disease Patients using an Array X-Band Microwave Doppler based Upper Limb Movement Quantizer”. In: *IEEE Sensors Journal* (2021) (cit. on p. 71).

LIU, Jinxin; KANTARCI, Burak, and ADAMS, Carlisle. “Machine learning-driven intrusion detection for Contiki-NG-based IoT networks exposed to NSL-KDD dataset”. In: *Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning*. 2020, pp. 25–30 (cit. on p. 57).

LIU, Peipei et al. “Quantitative assessment of gait characteristics in patients with Parkinson’s disease using 2D video”. In: *Parkinsonism & Related Disorders* 101 (2022), pp. 49–56 (cit. on p. 25).

LIU, Xu et al. “A dual-branch model for diagnosis of Parkinson’s disease based on the independent and joint features of the left and right gait”. In: *Applied Intelligence* (2021), pp. 1–12 (cit. on p. 131).

LU, Mandy et al. “Vision-based estimation of MDS-UPDRS gait scores for assessing Parkinson’s disease motor severity”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer. 2020, pp. 637–647 (cit. on p. 26).

LUGARESI, Camillo et al. “Mediapipe: A framework for building perception pipelines”. In: *arXiv preprint arXiv:1906.08172* (2019) (cit. on p. 142).

MA, Ling-Yan et al. “Remote scoring models of rigidity and postural stability of Parkinson’s disease based on indirect motions and a low-cost RGB algorithm”. In: *Frontiers in Aging Neuroscience* 15 (2023), p. 1034376 (cit. on p. 117).

MARTÍNEZ-MARTÍN, Pablo et al. “The clinical impression of severity index for Parkinson’s disease: international validation study”. In: *Movement Disorders: Official Journal of the Movement Disorder Society* 24.2 (2009), pp. 211–217 (cit. on p. 89).

“Metric properties of nurses’ ratings of parkinsonian signs with a modified Unified Parkinson’s Disease Rating Scale”. In: *Neurology* 49.6 (1997), pp. 1580–1587. DOI: 10.1212/WNL.49.6.1580. eprint: https://n.neurology.org/content/49/6/1580.full.pdf (cit. on p. 21).

MINISTERIO DE SALUD Y PROTECCIÓN SOCIAL DE COLOMBIA. *Análisis de situación de salud (ASIS) Colombia 2020*. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/ED/PSP/asis-2020-colombia.pdf. Accedido el 21 de junio de 2025. 2020 (cit. on p. 153).

— *Día Mundial del Parkinson: Colombia se destaca en atención*. https://www.minsalud.gov.co/Paginas/Dia-Mundial-del-Parkinson-Colombia-se-destaca-en-atencion.aspx. Accedido el 21 de junio de 2025. 2022 (cit. on p. 153).

MIRELMAN, Anat et al. “Gait impairments in Parkinson’s disease”. In: *The Lancet Neurology* 18.7 (2019), pp. 697–708. DOI: 10.1016/S1474-4422(19)30044-4 (cit. on p. 24).

— “Gait impairments in Parkinson’s disease”. In: *The Lancet Neurology* 18.7 (2019), pp. 697–708 (cit. on p. 115).

OKUDA, Shiho et al. “Gait analysis of patients with Parkinson’s disease using a portable triaxial accelerometer”. In: *Neurology and Clinical Neuroscience* 4.3 (2016), pp. 93–97 (cit. on pp. 48, 49).

OROZCO-ARROYAVE, Juan Rafael et al. “Apkinson: the smartphone application for telemonitoring Parkinson’s patients through speech, gait and hands movement”. In: *Neurodegenerative Disease Management* 10.3 (2020), pp. 137–157 (cit. on p. 150).

ORTELLS, Javier; HERRERO-EZQUERRO, María Trinidad, and MOLLINEDA, Ramón A. “Vision-based gait impairment analysis for aided diagnosis”. In: *Medical & biological engineering & computing* 56.9 (2018), pp. 1553–1564 (cit. on p. 129).

OTERO-MILLAN, Jorge et al. “Saccades during attempted fixation in parkinsonian disorders and recessive ataxia: from microsaccades to square-wave jerks”. In: *PLoS One* 8.3 (2013), e58535 (cit. on p. 49).

PARK, Dong Jun et al. “Evaluation for Parkinsonian Bradykinesia by deep learning modeling of kinematic parameters”. In: *Journal of Neural Transmission* 128 (2021), pp. 181–189 (cit. on p. 117).

PHAM, H. N. et al. “Multimodal Detection of Parkinson Disease based on Vocal and Improved Spiral Test”. In: *2019 International Conference on System Science and Engineering (ICSSE)*. 2019, pp. 279–284. DOI: 10.1109/ICSSE.2019.8823309 (cit. on p. 70).

PHAM, Hung N et al. “Multimodal detection of Parkinson disease based on vocal and improved spiral test”. In: *2019 International Conference on System Science and Engineering (ICSSE)*. IEEE. 2019, pp. 279–284 (cit. on p. 29).

POST, Bart et al. “Unified Parkinson’s disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable?” In: *Movement disorders : official journal of the Movement Disorder Society* 20.12 (2005), 1577—1584. DOI: 10.1002/mds.20640 (cit. on pp. 21, 69).

PRINCE, John; ANDREOTTI, Fernando, and DE VOS, Maarten. “Multi-source ensemble learning for the remote prediction of Parkinson’s disease in the presence of source-wise missing data”. In: *IEEE Transactions on Biomedical Engineering* 66.5 (2018), pp. 1402–1411 (cit. on p. 29).

PRZYBYSZEWSKI, Andrzej W et al. “Multimodal learning and intelligent prediction of symptom development in individual Parkinson’s patients”. In: *Sensors* 16.9 (2016), p. 1498 (cit. on p. 26).

RAJ, K Deepa et al. “A Visibility Graph Approach for Multi-stage Classification of Parkinson’s Disease Using Multimodal Data”. In: *IEEE Access (2024)* (2024) (cit. on p. 30).

RASCOL, Olivier et al. “Abnormal ocular movements in Parkinson’s disease: evidence for involvement of dopaminergic systems”. In: *Brain* 112.5 (1989), pp. 1193–1214 (cit. on p. 92).

RASTEGARI, Elham; AZIZIAN, Sasan, and ALI, Hesham. “Machine Learning and Similarity Network Approaches to Support Automatic Classification of Parkinson’s Diseases Using Accelerometer-based Gait Analysis”. In: *Proceedings of the 52nd Hawaii International Conference on System Sciences*. 2019 (cit. on pp. 24, 49).

REINER, Johnathan et al. “Oculometric measures as a tool for assessment of clinical symptoms and severity of Parkinson’s disease”. In: *Journal of Neural Transmission* 130.10 (2023), pp. 1241–1248 (cit. on p. 28).

RIEDERER, P et al. “Lateralisation in Parkinson disease”. In: *Cell and tissue research* 373 (2018), pp. 297–312 (cit. on p. 20).

RINCON-MONTANA, Ángela G et al. “Prevalencia, características demográficas de la enfermedad de Parkinson y comorbilidades asociadas: un análisis del registro oficial del Ministerio de Salud de Colombia”. In: *Acta Neurológica Colombiana* 41.1 (2025) (cit. on p. 153).

ROLAND, Virginie et al. “Vowel production: a potential speech biomarker for early detection of dysarthria in Parkinson’s disease”. In: *Frontiers in Psychology* 14 (2023), p. 1129830 (cit. on p. 149).

RUBIANO-CRUZ, Ricardo Andres. “Detection of Parkinson’s Disease with Multimodal Deep-Learning”. In: (2024) (cit. on p. 93).

RUCCI, Michele and POLETTI, Martina. “Control and functions of fixational eye movements”. In: *Annual Review of Vision Science* 1 (2015), pp. 499–518 (cit. on p. 48).

RUPPRECHTER, Samuel et al. “A clinically interpretable computer-vision based method for quantifying gait in parkinson’s disease”. In: *Sensors* 21.16 (2021), p. 5437 (cit. on p. 26).

RUSSO, Michela et al. “Kinematic and Kinetic Gait Features Associated With Mild Cognitive Impairment in Parkinson’s Disease”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2024) (cit. on pp. 90, 91).

SAAD, Ali et al. “A preliminary study of the causality of freezing of gait for Parkinson’s disease patients: Bayesian belief network approach”. In: *International Journal of Computer Science Issues* 10.3 (2013), pp. 88–95 (cit. on p. 34).

SABO, Andrea et al. “Estimating parkinsonism severity in natural gait videos of older adults with dementia”. In: *IEEE journal of biomedical and health informatics* 26.5 (2022), pp. 2288–2298 (cit. on p. 26).

SALAZAR, Isail et al. “A convolutional oculomotor representation to model parkinsonian fixational patterns from magnified videos”. In: *Pattern Analysis and Applications* 24.2 (2021), pp. 445–457 (cit. on pp. 71, 92).

SALEH, Adel et al. “Exploiting the Kinematic of the Trajectories of the Local Descriptors to Improve Human Action Recognition.” In: *VISIGRAPP (3: VISAPP)*. 2016, pp. 182–187 (cit. on pp. 48, 49).

SÁNCHEZ, Jorge Luis et al. “Prevalence of Parkinson’s disease and parkinsonism in a Colombian population using the capture-recapture method”. In: *international Journal of Neuroscience* 114.2 (2004), pp. 175–182 (cit. on p. 152).

SANDLER, Mark et al. “Mobilenetv2: Inverted residuals and linear bottlenecks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520 (cit. on pp. 50, 131).

SANDOVAL, Edward; OLMOS, Juan, and MARTÍNEZ, Fabio. “RIEMAE: Riemannian Masked Autoencoder for Classifying Malignant Prostate Cancer Patterns”. In: *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2025, pp. 1–4 (cit. on p. 32).

SANTOS GARCÍA, Diego et al. “Predictors of loss of functional independence in Parkinson’s disease: results from the coppadis cohort at 2-year follow-up and comparison with a control group”. In: *Diagnostics* 11.10 (2021), p. 1801 (cit. on p. 116).

SATO, Kenichiro et al. “Quantifying normal and parkinsonian gait features from home movies: Practical application of a deep learning–based 2D pose estimator”. In: *PloS one* 14.11 (2019), e0223549 (cit. on p. 129).

SIFRE, Laurent and MALLAT, Stéphane. “Rigid-motion scattering for image classification”. In: *Ph. D. thesis* (2014) (cit. on pp. 50, 131).

SIGCHA, Luis et al. “Bradykinesia detection in Parkinson’s disease using smartwatches’ inertial sensors and deep learning methods”. In: *Electronics* 11.23 (2022), p. 3879 (cit. on p. 117).

SILVA, Ana Beatriz Ramalho Leite et al. “Premotor, Nonmotor And Motor Symptoms Of Parkinson’s Disease: A New Clinical State Of The Art”. In: *Ageing Research Reviews* (2022), p. 101834 (cit. on p. 24).

SUBASI, Abdulhamit and MIAN QAISAR, Saeed. “EEG-based emotion recognition using modified covariance and ensemble classifiers”. In: *Journal of Ambient Intelligence and Humanized Computing* 15.1 (2024), pp. 575–591 (cit. on p. 31).

SUN, Yue-meng et al. “Digital biomarkers for precision diagnosis and monitoring in Parkinson’s disease”. In: *NPJ Digital Medicine* 7.1 (2024), p. 218 (cit. on p. 117).

TABARESTANI, Solale et al. “Longitudinal prediction modeling of alzheimer disease using recurrent neural networks”. In: *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE. 2019, pp. 1–4 (cit. on p. 132).

TINELLI, Michela; KANAVOS, Panos, and GRIMACCIA, Federico. “The value of early diagnosis and treatment in Parkinson’s disease: a literature review of the potential clinical and socioeconomic impact of targeting unmet needs in Parkinson’s disease”. In: (2016) (cit. on p. 23).

TOLOSA, Eduardo et al. “Challenges in the diagnosis of Parkinson’s disease”. In: *The Lancet Neurology* 20.5 (2021), pp. 385–397 (cit. on p. 21).

TRABASSI, Dante et al. “Machine learning approach to support the detection of Parkinson’s disease in IMU-based gait analysis”. In: *Sensors* 22.10 (2022), p. 3700 (cit. on p. 25).

TRIPATHI, Kumud and RAO, K Sreenivasa. “Robust vowel region detection method for multimode speech”. In: *Multimedia Tools and Applications* 80.9 (2021), pp. 13615–13637 (cit. on p. 147).

TSITSI, Panagiota et al. “Fixation duration and pupil size as diagnostic tools in Parkinson’s disease”. In: *Journal of Parkinson’s Disease* 11.2 (2021), pp. 865–875 (cit. on pp. 24, 92).

TUZEL, Oncel; PORIKLI, Fatih, and MEER, Peter. “Region covariance: A fast descriptor for detection and classification”. In: *European conference on computer vision*. Springer. 2006, pp. 589–600 (cit. on p. 31).

VASQUEZ-CORREA, J. C. et al. “Comparison of User Models Based on GMM-UBM and I-Vectors for Speech, Handwriting, and Gait Assessment of Parkinson’s Disease Patients”. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 6544–6548. DOI: 10.1109/ICASSP40776.2020.9054348 (cit. on p. 70).

VÁSQUEZ-CORREA, Juan Camilo et al. “Comparison of user models based on GMM-UBM and i-vectors for speech, handwriting, and gait assessment of Parkinson’s disease patients”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 6544–6548 (cit. on p. 93).

VÁSQUEZ-CORREA, Juan Camilo et al. “Multimodal assessment of Parkinson’s disease: a deep learning approach”. In: *IEEE journal of biomedical and health informatics* 23.4 (2018), pp. 1618–1630 (cit. on pp. 30, 93).

VÁRADI, Csaba. “Clinical Features of Parkinson’s Disease: The Evolution of Critical Symptoms”. In: *Biology* 9.5 (2020). DOI: 10.3390/biology9050103 (cit. on p. 130).

VÁSQUEZ-CORREA, J. C. et al. “Multimodal Assessment of Parkinson’s Disease: A Deep Learning Approach”. In: *IEEE Journal of Biomedical and Health Informatics* 23.4 (2019), pp. 1618–1630. DOI: 10.1109/JBHI.2018.2866873 (cit. on p. 70).

WAN, Shaohua et al. “Deep multi-layer perceptron classifier for behavior analysis to estimate Parkinson’s disease severity using smartphones”. In: *IEEE Access* 6 (2018), pp. 36825–36833 (cit. on p. 109).

Towards Data-Driven Modeling of Pathological Tremors. Vol. Volume 2: 16th International Conference on Multibody Systems, Nonlinear Dynamics, and Control (MSNDC). International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. V002T02A030. Aug. 2020. eprint: https://asmedigitalcollection.asme.org/IDETC-CIE/proceedings-pdf/IDETC-CIE2020/83914/V002T02A030/6586091/v002t02a030-detc2020-22147.pdf (cit. on p. 131).

WANG, Qinghui; ZENG, Wei, and DAI, Xiangkun. “Gait classification for early detection and severity rating of Parkinson’s disease based on hybrid signal processing and machine learning methods”. In: *Cognitive Neurodynamics* (2022), pp. 1–24 (cit. on p. 25).

WANG, Ruiping et al. “Covariance discriminative learning: A natural and efficient approach to image set classification”. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 2496–2503 (cit. on p. 31).

WIGHT, Sheila and MILLER, Nick. “Lee Silverman Voice Treatment for people with Parkinson’s: audit of outcomes in a routine clinic”. In: *International journal of language & communication disorders* 50.2 (2015), pp. 215–225 (cit. on p. 150).

ZENG, Xianhua et al. “Deep hybrid manifold for image set classification”. In: *Image and Vision Computing* 143 (2024), p. 104935 (cit. on pp. 31, 32).

ZENG, Zheng et al. “A robust gaze estimation approach via exploring relevant electrooculogram features and optimal electrodes placements”. In: *IEEE Journal of Translational Engineering in Health and Medicine* (2023) (cit. on p. 27).

ZHANG, Fengting et al. “Clinical features and related factors of freezing of gait in patients with Parkinson’s disease”. In: *Brain and behavior* 11.11 (2021), e2359 (cit. on p. 116).

ZHANG, JianYuan et al. “Eye movement especially vertical oculomotor impairment as an aid to assess Parkinson’s disease”. In: *Neurological Sciences* 42 (2021), pp. 2337–2345 (cit. on p. 92).

ZHAO, Aite et al. “Dual channel LSTM based multi-feature extraction in gait for diagnosis of Neurodegenerative diseases”. In: *Knowledge-Based Systems* 145 (2018), pp. 91–97 (cit. on p. 132).