

Article

Application of ML Techniques for Assessing Gross Motor Function in Adults Suffering from Cerebral Palsy

Jesus Gomez ¹ , Jose Luis Jodra ^{2,*} , Raquel Martinez ³ , Asier Salazar-Ramirez ³ ,
Maria-Arantzazu Ruescas-Nicolau ⁴  and M. Luz Sánchez-Sánchez ⁴ 

¹ Department of Physiotherapy, Faculty of Physiotherapy, University of Valencia, 46010 Valencia, Spain; jegope@alumni.uv.es

² Department of Computer Architecture and Technology, University of the Basque Country (UPV/EHU), 20018 Donostia, Spain

³ Department of System Engineering and Automation, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain; raquel.martinez@ehu.eus (R.M.); asier.salazar@ehu.eus (A.S.-R.)

⁴ Physiotherapy in Motion, Multispeciality Research Group (PTinMOTION), Department of Physiotherapy, Faculty of Physiotherapy, University of Valencia, 46010 Valencia, Spain; m.luz.sanchez@uv.es (M.L.S.-S.)

* Correspondence: joseluis.jodra@ehu.eus

Abstract

Cerebral palsy (CP) is a neuromotor disorder that affects movement and posture, causing functional limitations and musculoskeletal deformities that persist into adulthood. Variability in motor expression makes identifying the functional level a clinical challenge; achieving greater accuracy in this assessment enables detection of risk factors for postural deterioration. This study analysed a sample of 56 adults with CP to evaluate the ability of different machine learning (ML) models to discriminate between levels IV and V of the Gross Motor Function Classification System (GMFCS), using a total of 78 clinical variables such as spasticity, range of motion, deformities, and postural asymmetries. Different supervised learning models were compared, and a relabelling procedure was applied to improve classification consistency. The results indicate that there is valuable information in the collected standardised variables for classifying levels IV and V. The best performances were subsequently obtained by the neural network and the linear logistic regression, achieving the latter, which has explanatory properties interesting for clinicians, accuracy, F1-score and AUC metrics of 92.83%, 93.44% and 99.35%, respectively. These findings suggest that ML could emerge as a useful and explanatory tool for functional assessment of CP in adults and for the design of personalised rehabilitation strategies.

Keywords: cerebral palsy; machine learning; Gross Motor Function Classification System; explainability



Received: 15 November 2025

Revised: 12 December 2025

Accepted: 23 December 2025

Published: 25 December 2025

Copyright: © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Cerebral palsy (CP) is a heterogeneous group of developmental disorders that affect movement, muscle coordination and postural control, resulting in varying degrees of functional impairment [1,2]. It is a permanent condition caused by lesions or abnormalities in the developing brain, which alter the maturation of motor circuits and lead to lifelong limitations in activity and social participation [3].

CP is linked to a variety of factors that can occur during the prenatal, perinatal and postnatal stages. These include congenital malformations, intrauterine growth restriction, multiple pregnancies, infections during the foetal or neonatal period, asphyxia at birth,

premature birth, untreated maternal hypothyroidism, perinatal strokes, and thrombophilias. Among all these factors, premature birth is considered to be the most common and has the greatest impact on the risk of developing this condition [4,5].

The symptomatology of CP is very diverse and includes motor, sensory, cognitive and speech impairments, epilepsy, and musculoskeletal problems [3,6]. From a motor perspective, the spastic syndromes are the most common (frequently unilateral or bilateral) and can affect postural control and body alignment [7]. Over time, these impairments often lead to postural asymmetries, musculoskeletal contractures and deformities that negatively impact mobility, comfort, independence and quality of life in adulthood [8–11].

The functional assessment of people with CP is usually performed according to the Gross Motor Function Classification System (GMFCS) [12]. This system consists of a widely accepted scale that classifies gross motor function into five levels: the higher the level, the greater the impact on the person's mobility. The first level, level I, includes those people who can walk without limitations. Then, level II refers to those who walk with speed or balance restrictions. Level III includes those who require technical aids or a wheelchair to traverse long distances. The next level, IV, includes people with very limited mobility who need assistance or devices to stand or move. Finally, level V corresponds to those who are completely dependent on help for antigravity posture and mobility.

Although the literature about CP in children is extensive, research focusing on adults with CP remains limited. As life expectancy improves, adults with CP face new clinical challenges: progressive loss of motor function, increasing skeletal deformities, chronic pain and reduced social participation [13]. Nevertheless, studies exploring the determinants of posture and body alignment in adulthood are still scarce. Research such as that by Rodby-Bousquet et al. [14,15] described the prevalence of postural asymmetries in adults with CP and their strong association with elevated GMFCS levels. They observed that as GMFCS levels increased, so did the likelihood of bodily asymmetries and the inability to maintain or change posture [15].

In adults with CP, postural control represents a multifactorial and dynamic phenomenon influenced by neuromuscular variables (such as spasticity and range of motion), biomechanical variables, postural variables, the use of assistive devices, and the person's sitting, lying down or standing positioning habits. Traditionally, these relationships have been analysed using conventional statistical methods, such as correlations or logistic regressions [15,16].

In this context, machine learning (ML) techniques are emerging as promising tools in rehabilitation and clinical research. These technologies permit analysing clinical and functional data, modelling multidimensional relationships and identifying patterns in data. Although there have been significant advances in different applications within the field of CP [17–19], its implementation in the adult population remains limited. However, the results obtained by studies focusing on this population suggest that ML techniques have promising potential for improving functional assessment, customising therapeutic intervention and optimising clinical decision-making processes.

More precisely, for performing the functional assessment of adults with CP, it is necessary to take into account data belonging to different natures, such as age and sex, GMFCS level, degree of spasticity, range of motion, the use of assistive devices, and the time that the patient could spend in certain postures, among others. In this sense, ML algorithms can help integrate multiple sources of information for predicting complex clinical outcomes, such as postural quality or the presence of musculoskeletal deformities. Accordingly, these models can emerge as clinical support for assessing postural risk and, in this way, help to promote the personalisation of interventions in adults with CP.

This work provides a novel contribution to the field of CP through the application of ML to standardised clinical variables to discriminate between GMFCS levels IV and V. The study analyses the feasibility of addressing this classification problem using variables from different clinical domains related to postural control: spasticity, range of motion, musculoskeletal deformities, and postural asymmetries. Moreover, the study focuses on the understudied population of adults suffering from CP and reinforces the literature on this topic. In addition, the paper presents a relabelling procedure that facilitates the identification of borderline cases that could have been mislabelled due to subjective observation and improves the consistency of functional classification. Finally, the work analyses the explainability of the trained ML models and highlights the most influential variables in the classifier's decisions, providing useful insights for clinical decision-making and the personalisation of rehabilitation interventions.

The rest of the article is organised as follows: Section 2 presents the dataset used for the experiment, providing a detailed explanation of all variables, as well as the filtering and preprocessing applied to obtain the dataset version used in the article. Next, Section 3 describes the feature selection process and the experimentation to determine the GMFCS level of the patients in the dataset. In addition, it also explains the relabelling process proposed for some doubtful cases and presents the classification results obtained from the relabelled dataset. Later, the section also analyses the explainability of the best explainable classifier's decisions, seeking to give professionals a reason on which they can base their assessments. Finally, the results of the previous section and the conclusions drawn from the article will be discussed in Section 4.

2. Materials and Methods

This section presents the dataset used for this work's experimentation. First, Section 2.1 will give a detailed explanation of the variables in the dataset, describing what they represent, how they were scored, and the system used for naming them. Then, Section 2.2 will briefly describe the filtering applied to the dataset to remove some variables, as well as the preprocessing to prepare it for the ML algorithms.

2.1. Description of the Dataset

As mentioned in the introduction, this study focuses on the functional assessment of adults suffering from CP. To do so, the study uses data collected by four physiotherapists of the Valencian Association of Cerebral Palsy (AVAPACE) centres in Valencia (Spain), with between 5 and 20 years of experience in the field of treating patients suffering from CP.

Concerning the sampled population, originally the data were collected from 58 adult users during functional and positional assessment tests. Nevertheless, it was impossible for the team to fully collect two of the registers. Thus, the data of those two users were discarded, and the dataset ended up having records of 56 users, who were aged between 19 and 74 (mean: 39.04 years; standard deviation: 14.11 years), 24 of them being female and the other 32 male. Before taking any test, all participants received an explanation of the research and gave their informed consent for using their data in this study, as approved by the Ethics Committee of Research in Humans of the Ethics Commission in Experimental Research of the University of Valencia (protocol code FIS-2837811).

To ensure consistent inter-rater assessments of the participants, before beginning with the study and aiming to reduce possible inter-rater biases, the researchers shared a manual with the physiotherapists in which the assessment protocol was explained. Moreover, physiotherapists were also given a seminar about how to apply the protocol. Finally, to further ensure inter-rater consistency, the assessments of the first data collection day were performed along with the four physiotherapists together, so that a consensus was reached

for the evaluation criteria. In addition, concerning the consistency and coherence in the assessment of the GMFCS level, the scale itself has been reported to have a high degree of inter-rater agreement in the literature ([20,21]).

The dataset collected by the physiotherapists originally consisted of 94 independent variables from six domains: clinical record, postural quality, postural ability, the range of passive movement, the presence of certain bodily asymmetries and the muscle tone or degree of spasticity. Concerning the completeness of the dataset, no imputation method had to be applied, as it had been possible to fully collect the 94 variables for all the 56 participants. Apart from those independent variables, the dataset also contains each participant's level in the GMFCS according to the empirical observations made by the physiotherapists, which is independent of all other variables in the dataset that do not directly measure gross motor function. This last variable corresponds to the class variable to identify with the ML analysis proposed in this work. The detailed explanation of the variables (both independent and class variables) will be given in the following paragraphs.

To begin with, the dataset's first three variables correspond to the subject's clinical record: the sex, the primary medical diagnosis (named DX_P in the dataset) and the subtype of CP (named CP_SUB in the dataset). The sex variable was considered binary, using 0 for female and 1 for male subjects. Then variable DX_P was ranged from 0 to 5, with the following meanings: 0 stands for cerebral palsy, 1 for infant CP, 2 for diparesis due to congenital encephalopathy, 3 for perinatal CP, 4 for intrauterine encephalopathy, and 5 for encephalopathy. The third and last variable, CP_SUB, included the subtype of CP that was diagnosed by the doctors (or, in the case of not having a subtype, it was assessed by the physiotherapists). This variable ranged from 0 to 6, values which, in increasing order, stand for spastic hemiplegia, spastic diplegia, spastic tetraplegia, dyskinetic or athetoid CP, ataxic CP, hypotonic athetoid CP and mixed CP.

The dataset's second set of variables corresponds to postural quality, which was assessed using the Posture and Postural Ability Scale (PPAS) [22]. This scale is a tool to measure both body alignment and the ability to maintain a stable position. In this study, the scale was applied to four functional positions (standing, sitting, supine and prone) assessed in both the frontal and sagittal planes. Depending on the functional position analysed, the postural assessment considered different bodily items: head, trunk, pelvis, hips, knees, legs, feet and arms. Another item assessed whether the person's weight was evenly distributed for each functional position. For example, in the sitting position, this item would check whether both glutei and legs were evenly resting on the chair, supporting the weight equally on both sides. Finally, the last item assessed the overall quality of the posture for a position, understood as the sum of all the bodily items considered for that functional position (including the item for the weight distribution).

Nevertheless, not all items were evaluated in each position; the specific selection of items varied depending on the position adopted and the plane of observation. A total selection of six body and weight items, plus a seventh corresponding to the overall score for the selected position, was always maintained. Each item scored either 0 or 1, where 0 indicates the presence of asymmetry or deviation from the midline of the body, and 1 reflects symmetrical and adequate alignment. Concerning each position's overall score, they could vary between 0 (complete asymmetry) and 6 (optimal postural alignment).

Thus, according to the previous explanation, the dataset contained a total of 56 postural quality-related PPAS features (named PPASPQ_#, with the # being an identifier numeral). The following table (Table 1) gives a summarised explanation of the bodily items and functional positions represented by each of the PPASPQ_# features.

Table 1. Summary of the features included in the dataset related to the Positional Quality of the Posture and Postural Ability Scale (PPASPQ).

Feature	Functional Position	Plane and Items
PPASPQ_1-7	standing	frontal ¹
PPASPQ_8-14	standing	sagittal ²
PPASPQ_15-21	sitting	frontal ¹
PPASPQ_22-28	sitting	sagittal ²
PPASPQ_29-35	supine	frontal ¹
PPASPQ_36-42	supine	sagittal ³
PPASPQ_43-49	prone	frontal ¹
PPASPQ_50-56	prone	sagittal ⁴

¹ Items: head, trunk, pelvis, legs, arms, weight and overall position score (sum). ² Items: head, trunk, pelvis, hips, knees, feet and overall position score (sum). ³ Items: head, trunk, pelvis, legs, feet, weight and overall position score (sum). ⁴ Items: head, pelvis, hips, knees, arms, weight and overall position score (sum).

In addition to the features associated with postural quality, the PPAS also incorporates the “Postural Ability Level” (PAL) component. This second component constitutes the third block of characteristics in the dataset and assesses the individual’s ability to maintain or modify their posture. This component is scored on a scale from 1 to 7, where 1 represents a total inability and 7 indicates complete independence. To be consistent with the naming methodology used for the PPASPQ features, the features related to the PAL have been named PPASPAL_position (see Table 2).

Table 2. Summary of the Postural Ability Level PPAS (PPASPAL) features included in the dataset.

Feature	Postural Ability
PPASPAL_STAND	standing
PPASPAL_SIT	sitting
PPASPAL_SUP	supine
PPASPAL_PRON	prone

The fourth group of variables identifies contractures and articulation limitations associated with postural asymmetries. To this end, the physiotherapists measured the subjects’ passive range of motion (ROM) using a universal goniometer according to the standards of the American Academy of Orthopaedic Surgeons [23,24]. The shoulder, elbow, hip, knee and ankle joints’ ROMs were evaluated bilaterally, recording the maximum extension achieved without causing pain. The ROM was considered limited when the hip’s, knee’s, or elbow’s extension was less than 0°, or when ankle dorsiflexion did not reach the neutral position. Thus, these variables were scored either 1 or 0 for the cases with or without any joint limitation, respectively.

In addition to those variables, and also linked to the ROM, the physiotherapists assessed pelvic tilt. This variable analyses the alignment of the anterior superior iliac spines in the frontal plane. It was considered that there was a pelvic asymmetry when one anterior superior iliac spine was visibly higher or lower than the contralateral one, which indicates a deviation from the midline and an alteration in pelvic position. This variable was included in the dataset under the name INPED and coded binary (1 = pelvic tilt, 0 = aligned pelvis).

In total, the dataset contained 11 features linked to ROM measurements. 10 of them correspond to joint limitations and were named according to the LIMIT_xy system, where *x* referred to the joint selected for the measurement and *y* referred to the body side. For example, following the guide of Table 3, the variable measuring the ROM limitation of the left elbow would be named as LIMIT_EL. The remaining variable, INPED, has a postural

character and does not follow the aforementioned naming system because it assesses pelvic symmetry in the frontal plane rather than a specific joint limitation.

Table 3. Summary of the features related to the passive range of motion (ROM) limitations included in the dataset.

Feature	Definition
LIMIT_ <i>xy</i>	<i>x</i> = Shoulder, Elbow, Hip, Knee, Ankle <i>y</i> = Left, Right
INPED	Pelvic tilt

Following with the next set of variables, it relates to postural asymmetries of the different bodily segments (see Table 4). These asymmetries were determined through direct clinical observation of the alignment of those segments in the three positions. Specifically, the physiotherapists analysed the presence of dorsal hyperkyphosis or lumbar hyperlordosis and the relative position of the pelvis, lower limbs, and upper limbs with respect to the body's midline. The presence or absence of asymmetry was recorded as 1 or 0 for each segment, complementing the quantitative information obtained with the PPAS. In addition, scoliosis was recorded based on medical history, taking into account the Cobb angle [25] and, when necessary, radiographic review of the medical history. Scoliosis was considered present (value 1) when lateral curvature of the spine with visible vertebral rotation was observed or when the patient had undergone spinal arthrodesis for this reason.

Table 4. Summary of the features related to bodily asymmetries included in the dataset.

Feature	Definition
ESCOL	Scoliosis
CIF	Dorsal hyperkyphosis
LORD	Lumbar hyperlordosis
ROT_P	Pelvis rotation
ANT_RET	Anteversion/retroversion alterations

The last and sixth group of variables corresponds to muscle tone, which was assessed using the Modified Ashworth Scale (MAS) [26]. The MAS is a tool that quantifies the increase in resistance to passive movement as an indicator of the degree of spasticity. Measurements were taken in a relaxed position and at a constant speed, covering different muscle groups: psoas, glutei, adductors, quadriceps, hamstrings, tibialis and calves. The scale assigns scores from 0 to 4, with an intermediate category (1+), where 1 represents a slight increase with minimal resistance at the end of the range of motion, 1+ indicates slight resistance at the beginning and during less than half of the range, 2 reflects a more marked increase in most of the range but with ease of movement, 3 corresponds to a considerable increase that hinders passive movement, and 4 indicates complete stiffness in flexion or extension. A score of 0 is interpreted as no spastic response. However, as the variable was to be used with ML, its values had to be discrete numbers. As 1+ is not a number, value 2 was used to represent the scale's 1+ level and the scale scores from then on were represented with one more point with respect to their original scale value (i.e., value 2 in the scale is represented with 3 in the variable). Similarly to the ROM limit features, these features have been named according to the MAS_*xy* system (see Table 5), where *x* refers to the muscle group and *y* to the body side.

Finally, the class variable (named as GMFCS_LEV) corresponds to the level of gross motor function, which was classified using the expanded and revised version of the GMFCS, the GMFCS-E&R [27]. Although this scale was originally designed for children up to

18 years of age, this study applied it to an adult population, exclusively focusing on users classified as levels IV and V. Of these, 27 were at level IV and 29 at level V. On the one hand, level IV was assessed when the person had the functional ability to stand actively, requiring little assistance. On the other hand, for level V, the standing posture is achieved passively with the help of two people who provide physical support, with minimal or residual active participation by the user. Although there are five scale levels, the authors decided to focus on these two classes because of their physical similarities and because they are the most limiting and because they have the greatest impact on the patients themselves, their relatives, the healthcare system and society in general.

Table 5. Summary of the Modified Ashworth Scale (MAS) features included in the dataset.

Feature	Definition
MAS _{xy}	x = Psoas, Gluteus, Adductor, Quadriceps, Hamstring, Tibialis, Calf y = Left, Right

2.2. Preprocessing the Dataset

After identifying all variables, the next step was to preprocess the dataset and prepare it for the classification stage, in which the GMFCS level of the users would be determined using ML techniques.

As explained, the original dataset contained 94 variables. Nevertheless, not all variables could be used as features with the ML algorithms. This work focuses on the problem of discriminating between GMFCS levels IV and V, which, respectively, represent users able or not to stand on their own or with very little help. This way, it would not make sense to consider the PPASPQ variables collected in the standing position, as they implicitly carry information about the user's class. Therefore, the first 14 PPASPQ variables were filtered out from the database. Also, the PPASPAL_STAND variable was discarded for the same reason.

Then, the researchers standardised the remaining independent variables using z-score normalisation. With this, variables with bigger ranges are prevented from dominating the decisions of the ML, as all are transformed to have 0 and 1 mean and standard deviation values, respectively. The formula for applying the standardisation is presented in Equation (1), where Z_i is the standardised value for the variable sample, X_i is the variable sample, μ is the mean value of the variable and σ is the standard deviation of the variable.

$$Z_i = \frac{X_i - \mu}{\sigma} \quad (1)$$

In summary, after the aforementioned steps, the dataset consisted of 78 independent variables plus the class variable. This way, the dataset was ready for addressing the proposed classification problem using ML algorithms (this version of the dataset is available as Supplementary Material attached to this article).

3. Results

Having so far presented both the dataset and the proposed classification problem, the work in this paper went through five different stages. First, the researchers carried out a preliminary analysis to see the discriminant capacity of the dataset's variables for classifying the dependent variable and ranked them with the intention of having a first impression on whether the collected variables could be used for discriminating between GMFCS levels IV and V. Then, in a second stage, the researchers swept different subspaces to build classification systems for an initial solution to the problem proposed in this paper.

The third stage consisted in analysing the possible borderline cases to detect whether there was any possible mislabelled or ambiguous case. In order to obtain a consistent and trustworthy result, this was achieved using both statistical methods (principal component analysis, PCA) and unsupervised learning methods.

In the fourth stage, the dataset was automatically relabelled according to the previously mentioned analysis, and, after taking the relabelled cases to the experts for approval, the researchers generated new classification models that obtained better results compared to those achieved with the dataset without being relabelled.

Finally, in clinical fields, knowing why a decision has been taken is almost as important as taking the correct decision. That is why the fifth stage consisted in analysing the explainability of the decisions taken in the previous stage by calculating the Shapley values and the coefficients of the best explainable algorithm used in this work: the linear logistic regression.

The following Sections 3.1–3.5 will, respectively, explain in more detail all five stages, presenting the results obtained in each one.

3.1. Preliminary Analysis of the Relevance of the Features

First, the researchers decided to analyse the convenience of reducing the original dataset to a smaller subset of features. To do so, three different methods were used to measure the relation between the features and the GMFCS_LEV class variable. Whereas two of these methods correspond to automatic feature selection techniques (chi-square and minimum redundancy maximum relevance, Mrmr, both in their respective *fscchi* and *fscmr* Matlab implementations), the third belongs to Kendall's univariate correlation analysis.

The values obtained from the analysis are shown in Table 6, which lists the 20 features that were either selected or more strongly correlated with the class variable depending on the method used. The table indicates that the variables more strongly linked to the class variable were gathered in a rather stable subset of features for the three methods. In this sense, using the features of the chi-square column as the reference, the cells highlighted in grey indicate that the feature has also been marked among the most relevant ones by either of the other (or both) two methods.

Table 6. Ranked results of the most discriminant variables (feature selection methods and correlation) that point out the convenience of applying feature selection.

Chi-Square	Mrmr	Kendall
PPASPQ_53	PPASPQ_35	PPASPAL_SIT
PPASPQ_47	PPASPQ_19	PPASPQ_53
PPASPQ_52	LIMIT_EL	PPASPQ_47
PPASPAL_SIT	LIMIT_AL	PPASPQ_52
PPASPQ_15	PPASPAL_SIT	PPASPQ_56
PPASPQ_39	ESCOL	PPASPQ_15
PPASPQ_30	PPASPQ_31	PPASPQ_35
PPASPQ_43	PPASPQ_47	PPASPQ_39
PPASPQ_29	PPASPQ_16	PPASPQ_49
PPASPQ_54	PPASPQ_53	PPASPQ_42
PPASPQ_34	PPASPQ_30	PPASPQ_29
PPASPQ_38	PPASPQ_43	PPASPQ_52
PPASPQ_46	PPASPQ_15	PPASPAL_PRON
PPASPQ_56	PPASPQ_36	PPASPQ_29
PPASPQ_21	MAS_GR	PPASPQ_54
PPASPAL_PRON	PPASPQ_23	PPASPQ_34
PPASPQ_32	PPASPQ_39	PPASPQ_38
PPASPQ_50	PPASPQ_30	PPASPQ_46
PPASPQ_31	PPASPQ_46	PPASPQ_28
PPASPQ_35	LORD	PPASPQ_21

Seeing the consistency of the results of this initial feature selection analysis, the researchers decided to carry out a feature sweep to find out how many variables should be considered for the classification.

3.2. Classification

As mentioned before, the next step of the work consisted in sweeping the dataset to see how classifiers performed in different dimensional feature subspaces. More precisely, the preprocessed version of the dataset consisted of 78 features for a total of 56 instances (24 female and 32 male), 27 of them belonging to GMFCS level IV and the remaining 29 to level V.

For the experimentation presented in this section, the researchers opted for training nine classifiers belonging to different paradigms of machine learning [28,29]: discriminant analysis (Discr), decision trees (Tree), k-nearest neighbours (KNN), boosting (Boost), bagging (Bagg), feed-forward fully connected neural network (Net), support vector machines (Svm), linear logistic regression (Lin) and Gaussian kernel classifiers using vector machines (Kern). All algorithms were implemented with Matlab's default settings using a 10-run 3-fold cross-validation methodology [30] and the chi-square feature selection method to select the most relevant features in each fold.

The following Tables 7–9 respectively present the average accuracy, F1 and AUC algorithm performance metrics [31] obtained after 10 runs for the subspace sweeping process. It must be noted that although the experimentation was conducted using all possible dimension values, for the sake of clarity, Tables 7–9 only present the values between 5 and 78 dimensions, with a stride of five variables. Concerning the meaningfulness of the values in these tables, they have been highlighted differently depending on the meaning: values in bold refer to the best performance obtained by each classifier, the light grey cell corresponds to the best performance obtained by the best explainable algorithm (Lin, in this case), and the dark grey cell corresponds to the best performance in the whole table. Finally, seeking to reduce the number of features needed to obtain a good classification result, underlined values represent the best performance that can be obtained with fewer features if the best performance of each classifier is reduced by 2%, which would approximately correspond to missing one instance out of the 56.

To begin with the analysis of the results of Tables 7–9, a quick overview shows that the overall performances score around 70% and 75%, with certain performance peaks and troughs close to 80% and 60%, respectively. These results indicate that the features of the dataset contain useful information for approaching the GMFCS level IV and V classification problem.

Table 7. Accuracies for dimensions between 5 and 78, with a stride of 5 dimensions using the chi-square method for ordering the most relevant features from the original dataset.

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	78
Discr	73.37	69.09	66.79	66.99	64.69	63.50	63.29	62.39	61.72	60.31	59.82	60.65	61.73	60.65	63.17	62.24
Tree	71.57	67.87	66.97	66.39	66.23	65.19	65.41	65.40	65.73	65.89	66.08	65.91	65.91	64.80	64.80	66.78
KNN	<u>68.75</u>	68.23	74.10	73.03	74.68	75.59	<u>77.13</u>	76.26	75.92	77.67	76.77	76.98	75.70	74.64	72.31	78.72
Boost	69.44	68.34	67.30	66.96	70.51	68.33	<u>69.48</u>	67.14	68.73	70.53	67.89	67.70	67.39	68.79	67.90	68.97
Bagg	<u>70.87</u>	70.88	71.23	69.93	71.71	69.76	71.21	71.39	72.68	73.37	72.64	71.56	70.68	70.33	71.19	72.47
Net	<u>69.84</u>	69.82	68.44	70.29	<u>70.70</u>	69.30	70.51	69.82	71.40	69.99	68.55	69.64	70.20	70.35	68.53	69.58
SVM	73.20	69.81	67.47	67.51	64.62	64.62	65.70	63.75	66.41	67.62	66.75	66.20	65.38	64.81	64.82	64.80
Lin	74.26	70.89	68.76	68.74	68.18	66.97	69.11	66.93	70.00	69.43	68.00	67.99	68.60	68.04	68.38	68.35
Kern	70.54	68.91	66.42	62.25	59.08	53.62	53.95	56.04	51.41	55.34	49.05	50.53	52.37	52.81	50.03	49.15

Values in bold refer to the best performance obtained by each classifier, the light grey cell corresponds to the best performance obtained by the best explainable algorithm, the dark grey cell corresponds to the best performance in the whole table and underlined values represent the best performance that can be obtained with fewer features if the best performance of each classifier is reduced by 2%.

Table 8. F1 metrics for dimensions between 5 and 78, with a stride of 5 dimensions using the chi-square method for ordering the most relevant features from the original dataset.

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	78
Discr	74.24	69.80	68.41	67.67	65.86	65.43	64.10	62.54	63.68	61.31	61.44	61.83	63.06	61.40	64.24	63.08
Tree	71.78	67.77	66.96	66.45	66.58	67.18	66.31	66.05	66.81	66.87	67.16	66.80	66.43	65.22	65.09	66.74
KNN	68.18	68.43	74.11	72.88	75.41	76.01	77.45	76.18	75.71	<u>77.57</u>	77.14	77.75	76.43	75.91	73.83	79.52
Boost	70.54	68.29	68.36	67.68	71.05	68.68	69.81	66.70	69.01	71.35	68.79	69.08	68.28	69.56	68.80	68.23
Bagg	<u>71.97</u>	70.66	71.54	70.17	71.72	69.44	70.18	70.02	71.34	71.96	71.30	71.06	70.01	69.97	70.94	71.44
Net	71.65	69.52	69.01	70.88	71.44	69.40	71.16	69.81	71.39	70.41	68.54	69.26	71.17	70.98	68.81	68.45
SVM	73.66	70.03	68.51	67.71	66.79	66.19	66.86	64.94	67.56	68.18	67.22	66.81	66.58	64.92	65.66	65.45
Lin	75.09	71.51	69.22	69.66	69.35	67.73	70.72	68.09	71.22	70.10	68.72	68.80	69.55	68.91	69.30	69.50
Kern	70.42	66.09	62.55	59.79	57.03	53.72	55.82	61.01	57.71	62.00	57.46	57.25	59.94	62.19	59.58	58.57

Values in bold refer to the best performance obtained by each classifier, the light grey cell corresponds to the best performance obtained by the best explainable algorithm, the dark grey cell corresponds to the best performance in the whole table and underlined values represent the best performance that can be obtained with fewer features if the best performance of each classifier is reduced by 2%.

Table 9. AUC metrics for dimensions between 5 and 78, with a stride of 5 dimensions using the chi-square method for ordering the most relevant features from the original dataset.

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	78
Discr	79.05	71.50	68.33	68.06	64.46	64.06	63.35	63.67	62.76	60.64	61.30	64.47	64.37	65.98	66.53	67.33
Tree	73.57	71.47	69.87	69.07	67.51	69.33	69.17	68.64	69.17	68.45	69.09	68.82	68.69	67.74	67.83	69.38
KNN	68.78	68.15	74.17	73.04	74.57	75.52	77.13	76.28	76.00	77.78	76.83	76.98	75.70	74.54	72.19	78.57
Boost	71.50	75.32	75.35	74.93	<u>77.35</u>	78.02	<u>77.29</u>	77.40	78.00	79.03	78.48	78.39	77.97	78.35	78.72	77.70
Bagg	77.37	77.84	<u>79.42</u>	78.38	<u>78.90</u>	79.73	79.78	80.18	80.83	80.81	80.24	79.94	80.09	79.98	79.82	80.54
Net	68.38	71.28	<u>72.61</u>	74.25	77.02	77.19	77.73	79.26	78.94	77.15	75.91	75.66	74.19	74.47	73.82	77.69
SVM	77.01	74.58	71.75	69.42	69.84	71.11	<u>71.01</u>	70.81	73.45	73.07	72.73	72.21	71.42	71.24	71.07	72.52
Lin	80.74	77.79	76.94	75.16	76.52	76.47	76.25	76.26	78.61	77.64	76.01	75.38	74.13	73.68	74.22	76.89
Kern	77.26	79.58	76.39	70.07	65.15	61.18	61.56	59.18	52.19	56.29	47.96	47.72	52.35	52.73	51.33	50.82

Values in bold refer to the best performance obtained by each classifier, the light grey cell corresponds to the best performance obtained by the best explainable algorithm, the dark grey cell corresponds to the best performance in the whole table and underlined values represent the best performance that can be obtained with fewer features if the best performance of each classifier is reduced by 2%.

Then, a deeper analysis of the tables shows that whereas the KNN was the best-performing algorithm for accuracy and F1 metrics (78.72% in Table 7 and 79.52% in Table 8, respectively), the bagging obtained the best AUC (80.83%, Table 9) in the scenario using the original dataset. Also, it can be noted that, even if it needed all the features to obtain its best performance, it would be possible to obtain a reduction to 35 or 50 features (depending on whether a loss of 2% is assumed for the accuracy and AUC or the F1, respectively) and still achieve over 77% for the three performance metrics. Looking at the tables from the perspective of receiving an explanation about the decisions made by the algorithms, both the Tree and Lin obtained their best performances around the five-feature subspace, where Lin was the best among them.

Finally, should the reader desire to consult other performance metrics, Tables A1–A3 presenting the precision, recall and specificity scores for this scenario are available in Appendix A.

3.3. Relabelling

After obtaining promising results from the initial approach, and considering that the physiotherapists label the patients based on observations with a certain degree of subjectivity for the borderline cases, the team decided to search for cases located in an

ambiguous region between classes IV and V. In this sense, if such cases existed, detecting and correcting them would positively impact the classifiers' results.

Accordingly, the team decided to do a principal component analysis (PCA) of all instances to obtain a graphical representation of their location in a scatter plot using the first two principal components of the classification space as the chart axes. This representation would give the team a different view of how the subjects are distributed from the perspective of the classification space, which relies on the variables listed in Section 2 and differs from the observations made by the physiotherapists for assessing the GMFCS level. The result of this analysis is depicted in Figure 1, where the blue and red dots correspond to the instances belonging to classes IV and V, respectively. In addition, to make the interpretation of the graph easier, the centroids of both classes have been marked with a rhombus of the same colour.

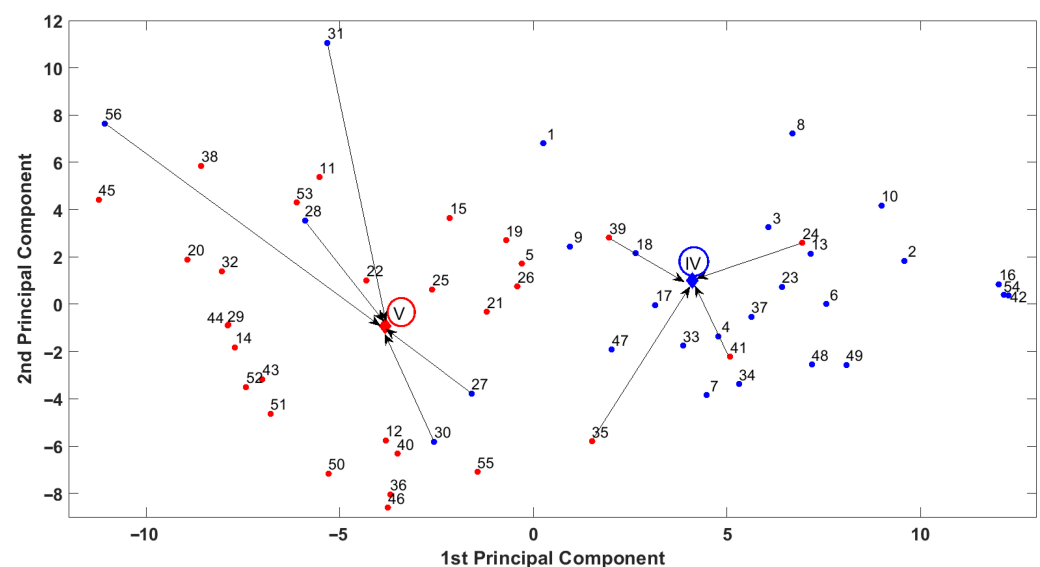


Figure 1. Scatter plot of all the instances according to the first two principal components. Blue and red colours correspond to classes IV and V respectively. The arrows point from the instances to the centroid which is closer to them for those cases that fall closer to the opposite class.

After examining Figure 1, the team realised that there were 9 subjects located closer to the centroid of the opposite class. In this sense, from the perspective of the PCA, it would make more sense if instances 27, 28, 30, 31 and 56 belonged to class V instead of IV. Likewise, instances 24, 35, 39 and 41 would better fit into class IV. Thus, the team thought that these nine cases could be good candidates for being taken back to the experts so that they could further analyse the given label and, if they considered them borderline cases, whether those labels could belong to the other class.

However, before asking the experts to reevaluate those cases, the team proceeded with a deeper analysis of the situation to ensure the stability of their relabelling proposal. In this sense, two approaches were followed to determine whether it was worth asking the experts. On the one hand, the first approach consisted in grouping all the features according to their nature and then calculating and comparing for each feature group the centroids of the borderline cases and the whole set of each class. On the other hand, the team used unsupervised learning algorithms to validate the relabelling proposal. More precisely, the team used two algorithms coming from two different paradigms: K-Means and Hierarchical clustering. In this case, the parameter corresponding to the number of groups (K) was set to 2 for both algorithms, i.e., the number of categories handled by the class variable GMFCS_LEV.

The following table (Table 10) relates to the first approach and presents the aggregated centroids for each feature group (MAS, ESCOL . . . , LIMIT, PPASPAL and PPASPQ) along with the labels. Looking at the table's rows, those identified as TOT correspond to all the instances of each class. Then, the other two rows correspond to the patterns subject to a possible class change (from class IV to V and vice versa). Looking at the table, it becomes clear that the centroids of the second row's five cases, originally labelled as class IV, are much closer to the third row's centroids (class V). In the same manner, the centroids of the last row's four cases (class V) are much closer to the ones of the first row (class IV). This phenomenon takes place for all feature groups, and therefore, the idea that it might be worth reconsidering the class of those nine borderline cases is reinforced.

Table 10. Aggregated centroids and class for each feature group of different nature (MAS_{xy}, asymmetries, LIMIT_{xy}, PPASPAL_{position} and PPASPQ_#).

Subject	MAS _{xy}	Asymmetries (ESCOL, CIF, ...)	LIMIT _{xy}	PPASPAL _{position}	PPASPQ _# ²	GMFCS_LEV
TOT (IV)	17	1.67	6	15	27	IV
5 cases ¹	39	2.20	8	7	10	IV
TOT (V)	23	2.66	7	7	8	V
4 cases ¹	13	2.00	5	14	21	V

¹ Cases subject to a possible change in class. ² Only overall position scores were considered, with # numerals: 21, 28, 35, 42, 49 and 56.

Regarding the clustering approach, the columns of Table 11 present the clusters generated by both the Hierarchical clustering and K-Means algorithms. In particular, the table shows the cases of the minority class in each cluster, which can be understood as the cases that each clustering algorithm would suggest labelling as belonging to their opposite class. In addition, the cells highlighted in bold correspond to cases that both algorithms considered more similar to the other class they had previously been labelled as (the intersection cases). For instance, whilst subject 19 was only included in Cluster 1 by the hierarchical algorithm, the two algorithms decided that subject 24 would better fit in Cluster 1 (whose majority class is class IV).

Table 11. Subjects proposed for a class change by the unsupervised learning algorithms. The cases marked in bold correspond to the intersection cases, i.e., the cases that both algorithms agreed that would better fit into the other class.

Hierarchical Clustering (10 Class Change Proposals)	K-Means (13 Class Change Proposals)
Cluster 1 (IV:22—V:5)	Cluster 1 (IV:18—V:4)
19	-
24	24
35	35
39	39
41	41
Cluster 2 (IV:5—V:24)	Cluster 2 (IV:9—V:25)
-	1
-	9
-	18
27	27
28	28
30	30
31	31
-	47
56	56

Bearing in mind the results of Tables 10 and 11, the authors saw that both approaches were concordantly suggesting that nine cases could better fit in the opposite class they had been labelled into. Moreover, those nine cases are consistent with the instances proposed by the PCA scatter plot of Figure 1.

Accordingly, after seeing what the analysis based on automated techniques suggested, the team queried the experts whether it would make sense to swap the class of subjects 27, 28, 30, 31 and 56 from IV to V according to the criteria they used for labelling these subjects. Similarly, they proposed changing subjects 24, 35, 39 and 41 from class V to IV. In both cases, the experts answered that those subjects fell within the limit of belonging to one class or the other and that reconsidering the class of those subjects would be pertinent.

3.4. Classification After the Automatic Relabelling

Having received a positive answer to their proposal, the researchers proceeded to relabel the aforementioned nine cases and repeated the algorithm training experimentation with the modified dataset. The accuracy, F1 and AUC performances achieved by the same nine algorithms and using the same methodology as before are respectively shown in the following tables (Tables 12–14). Again, the values of the tables have been highlighted as mentioned in Section 3.2.

Starting with a general analysis of Tables 12–14, it is possible to conclude that using the relabelled dataset had an overall positive impact on the classifiers' performance for all three performance metrics. This phenomenon might have its roots in the noise that the mislabelled instances were introducing into the dataset. This way, the overall performance of the classifiers increased around 10% to values surrounding 80% and 85%, with peak performances over 90%. Thus, these good results reinforce the hypothesis that it is possible to discriminate between GMFCS levels IV and V using the standardised clinical variables collected in this dataset.

Table 12. Accuracies for dimensions between 5 and 78, with a stride of 5 dimensions using the chi-square method for ordering the most relevant features from the relabelled dataset.

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	78
Discr	82.01	80.60	80.01	75.68	73.22	71.64	68.56	67.68	69.61	74.47	78.04	81.58	84.63	87.32	86.26	85.51
Tree	82.17	82.13	85.54	86.99	86.11	86.29	86.63	85.71	84.99	85.37	85.57	84.50	84.12	<u>83.94</u>	83.60	83.03
KNN	81.64	85.78	83.44	85.91	<u>87.70</u>	87.51	87.20	88.43	87.17	86.63	86.44	87.51	86.98	86.27	88.06	87.70
Boost	81.46	84.12	86.06	89.08	88.40	87.88	89.83	87.89	86.77	87.00	87.02	85.23	85.39	85.58	84.87	83.58
Bagg	84.69	85.06	87.85	<u>90.91</u>	90.71	91.05	90.69	89.99	89.45	89.80	89.26	89.07	88.55	88.90	89.25	88.90
Net	82.35	83.95	83.25	85.90	86.44	88.63	90.92	91.28	91.23	91.42	90.91	<u>92.69</u>	92.47	92.11	92.88	93.68
SVM	82.73	82.04	81.81	84.14	86.95	88.95	88.57	88.59	89.47	89.84	90.36	<u>90.18</u>	89.99	89.42	89.46	91.06
Lin	82.36	84.90	85.42	86.27	87.70	89.13	88.80	90.04	90.20	<u>91.27</u>	90.72	91.24	91.24	91.03	91.57	92.83
Kern	82.70	77.13	69.07	66.25	61.92	59.06	57.45	49.60	49.67	53.20	53.25	52.53	50.78	52.37	54.67	51.07

Values in bold refer to the best performance obtained by each classifier, the light grey cell corresponds to the best performance obtained by the best explainable algorithm, the dark grey cell corresponds to the best performance in the whole table and underlined values represent the best performance that can be obtained with fewer features if the best performance of each classifier is reduced by 2%.

If the tables are analysed in more detail, the reader will realise that when using the relabelled dataset, it was the Net which obtained the best performances for the three metrics: 93.68% for accuracy, 94.14% for F1 and 99.39% for AUC. These optimal results were obtained using the entire dataset or a subspace of 75 features.

However, bearing in mind that clinicians often desire a reason for taking their decisions and that the Net is not an explainable classifier, it is interesting to take a look at the good performances achieved by the Lin classifier (with explanatory properties). The performances scored by this classifier are less than 1% away from those obtained by the Net (92.83%, 93.44% and 99.35% for accuracy, F1 and AUC values, respectively). Furthermore,

taking a loss of 2% in Lin's AUC as reference (97.58%), it is possible to achieve over 89% scores for the three metrics using a much smaller subspace of 30 features. Similarly, being more conservative with overall loss of performance but still trying to compensate for the possible curse of dimensionality, the Lin classifier would achieve scores over 91% for the three metrics when using a reduced subspace of 50 features.

Table 13. F1 metrics for dimensions between 5 and 78, with a stride of 5 dimensions using the chi-square method for ordering the most relevant features from the relabelled dataset.

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	78
Discr	84.04	82.82	81.82	77.63	75.52	73.40	71.01	69.75	72.79	76.44	79.92	83.91	86.09	88.25	87.42	86.67
Tree	83.12	82.55	85.89	87.39	86.59	86.73	87.05	85.90	85.18	85.47	85.77	85.08	84.73	<u>84.53</u>	84.39	84.24
KNN	83.03	86.79	<u>85.17</u>	87.14	88.99	88.67	88.47	89.62	88.51	87.87	87.81	88.46	88.22	87.76	89.48	89.11
Boost	83.05	84.67	86.98	89.13	<u>88.74</u>	88.36	90.20	88.40	87.28	87.48	87.50	86.28	86.51	86.61	86.07	84.98
Bagg	86.08	86.27	88.70	91.37	91.02	91.35	90.93	90.35	89.95	90.47	89.90	89.86	89.26	89.65	90.03	89.82
Net	84.01	84.83	84.34	<u>86.55</u>	87.36	89.08	91.21	91.73	91.86	91.92	91.55	93.16	93.09	92.55	93.45	94.14
SVM	84.49	83.54	83.22	85.36	88.01	89.74	89.21	89.40	<u>90.26</u>	90.80	91.22	<u>91.01</u>	90.99	90.43	90.55	91.96
Lin	84.13	86.23	86.70	87.40	88.75	89.83	89.54	90.78	<u>91.07</u>	91.91	91.52	91.98	92.03	91.82	92.21	93.44
Kern	83.48	75.46	66.55	63.57	61.38	62.19	63.09	61.02	61.76	<u>65.09</u>	63.95	63.74	64.41	66.22	67.85	64.28

Values in bold refer to the best performance obtained by each classifier, the light grey cell corresponds to the best performance obtained by the best explainable algorithm, the dark grey cell corresponds to the best performance in the whole table and underlined values represent the best performance that can be obtained with fewer features if the best performance of each classifier is reduced by 2%.

Table 14. AUC metrics for dimensions between 5 and 78, with a stride of 5 dimensions using the chi-square method for ordering the most relevant features from the relabelled dataset.

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	78
Discr	89.88	87.43	84.13	79.55	78.01	74.68	69.43	68.03	72.48	79.64	83.61	89.96	92.22	94.78	94.21	94.19
Tree	82.79	83.38	<u>86.50</u>	87.60	87.02	87.14	87.34	86.22	85.67	86.08	86.60	85.47	85.00	<u>84.83</u>	84.44	84.25
KNN	81.52	85.71	<u>83.19</u>	85.75	87.37	87.33	86.92	88.14	86.88	86.38	86.06	87.30	86.69	85.91	87.67	87.38
Boost	87.95	92.08	94.17	<u>96.62</u>	<u>95.34</u>	95.81	96.69	95.34	95.25	94.91	94.33	93.41	93.47	93.19	92.85	91.15
Bagg	90.59	94.26	95.99	97.26	97.22	97.39	97.52	97.34	97.04	97.29	97.35	97.35	97.20	97.07	96.97	97.05
Net	84.50	90.83	<u>93.63</u>	95.64	96.74	<u>97.45</u>	97.69	97.92	98.19	98.44	98.06	98.84	99.12	98.97	99.39	99.09
SVM	87.12	89.85	91.83	93.86	95.57	<u>97.20</u>	97.14	97.55	97.70	98.14	98.10	98.18	98.48	98.55	98.13	98.49
Lin	91.74	93.23	95.09	95.77	96.74	<u>97.58</u>	97.78	98.08	98.36	98.59	98.63	98.89	99.08	99.19	98.85	99.35
Kern	88.52	86.07	81.32	77.78	73.48	<u>66.86</u>	63.62	49.18	46.00	47.69	51.37	46.14	47.29	51.68	51.98	51.28

Values in bold refer to the best performance obtained by each classifier, the light grey cell corresponds to the best performance obtained by the best explainable algorithm, the dark grey cell corresponds to the best performance in the whole table and underlined values represent the best performance that can be obtained with fewer features if the best performance of each classifier is reduced by 2%.

To conclude with this subsection, the authors want to note that, as they did with the results of Section 3.2, it is possible to find the precision, recall and specificity scores for this second scenario in Tables A1–A3 included in Appendix A.

3.5. Explainability Analysis

It is clear that the results obtained using the relabelled dataset are better than those obtained with the original. Nevertheless, this improvement comes at the cost of expanding the problem, needing to use more variables to obtain the best performances, just as mentioned in the previous Section 3.4. For this reason, and with the intention of supporting the physiotherapists in their work, the authors chose to extract explanatory information about which variables are the most important, from a discriminative point of view, in this new scenario after the relabelling.

With this in mind, the analysis presented in this section will focus on the Lin classifier, which obtained the second-best performance with the relabelled dataset and which, in contrast to the Net, has explanatory properties. To do such an analysis of the explainability of the classification, the researchers used two different sources of information to verify the stability of the most significant variables. In order to make maximum use of the information

in the dataset, all subjects were used for these analyses, since the aim was not to estimate the performance of the system in operation, as in the previous tables.

On the one hand, given that the Lin classifier induces a coefficient per variable to perform the classification, the team analysed the features with the highest absolute coefficient values, as they have the greatest influence on the classification result. Depending on whether the feature is directly or inversely proportional to an increase in the pathology, the sign of the coefficient will be positive or negative. Table 15 shows the 10 variables with the highest coefficients in the model.

Table 15. The 10 most significant coefficients of the linear logistic regression (Lin) classifier.

Feature	Coefficient Value
PPASPQ_33	−0.538
PPASPAL_PRON	−0.368
PPASPAL_SUP	−0.352
LIMIT_SR	0.324
PPASPQ_52	−0.320
MAS_PL	0.296
PPASPAL_SIT	−0.292
LIMIT_ER	0.292
MAS_PR	0.291
PPASPQ_19	−0.286

As seen in Table 15, the variables in the PPASPQ domain have negative coefficients. This makes sense, as if subjects are able to maintain these postures (value 1 means that they can), they are classified as class IV (less severe pathology). Inversely, the contribution of variables involving spasticity (MAS) and range of movement limitations (ROM) points to class V with positive values. As shown in the table, the three sources of information provide discriminating power to the system, so they can be taken into account by specialists when assessing gross motor function.

On the other hand, as a second means of analysing the explanation, the team opted to calculate the Shapley values ([32], SHapley Additive exPlanations) for this problem. Specifically, we applied the Matlab implementation [33] to extract the Shapley values. The values for the 10 features that contribute most to the classification for the Lin classifier are presented in Figure 2. The horizontal axis of the figure represents the contribution of each feature to a classification towards class V if it is positive (right side) or to class IV if it is negative (left side). The colour of the dots reflects whether the contribution is with high feature values (yellow) or with low feature values (blue). Whilst some variables are binary (0/1, which is why they appear only in yellow and blue), others have several categories (hence the scale of values).

As can be seen in Figure 2, the 10 most relevant features belong to three different natures: postural ability and quality (PPASPQ and PPASPAL), ROM range of motion (LIMIT_xy), and spasticity (Modified Ashworth Scale, MAS). Features PPASPQ_33, PPASPAL_PRON, PPASPQ_52, PPASPAL_SUP, PPASPQ_19, PPASPQ_47, and PPASPQ_39 (see Tables 1 and 2) contribute in such a way that lower values (i.e., they cannot hold certain postures in a stable manner) bring them closer to class V (positive horizontal axis). Likewise, their higher values (in yellow) point to class IV (negative horizontal axis), that is, their greater functional autonomy allows them to hold these postures in a stable manner. On the contrary, variables related to ROM and spasticity have a symmetrical behaviour. Specifically, LIMIT_SR and LIMIT_ER point towards class V with high values (yellow), as they imply a greater level of limitation. The same tendency applies to variable MAS_PL, which behaves in the same way.

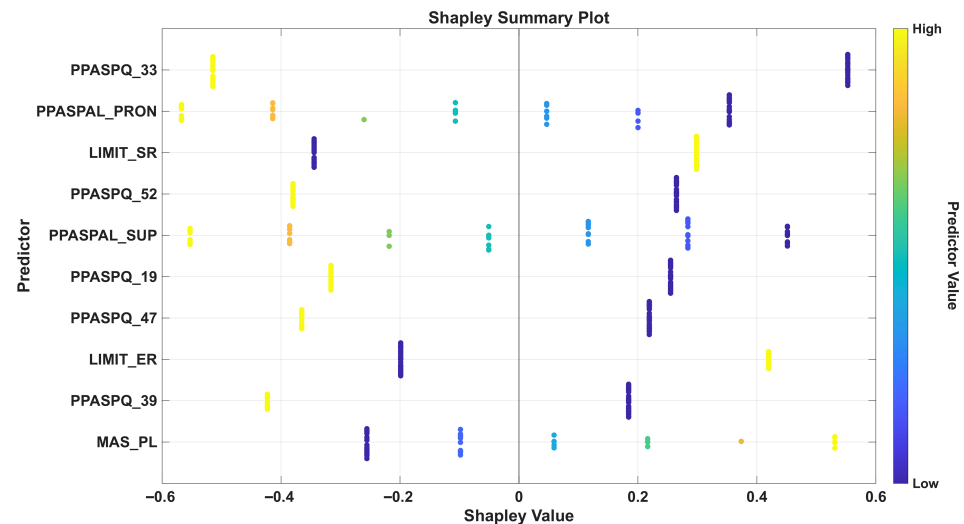


Figure 2. Shapley plot for the 10 features with the highest contribution to the classification according to the chi-square method for the logistic linear regression (Lin) classifier.

Having performed both types of analysis and comparing the coefficients of the Lin classifier at its best performance with the most discriminating according to the Shapley values, it can be seen that the results are consistent with an 80% intersection (variables that coincide in both analyses are marked in the grey table (Table 15)). Likewise, it could be concluded that relabelling the nine subjects made it possible to avoid inconsistencies in the values of certain features that were interfering with the classification. Thanks to this, the performance of the Lin classifier improved several points, and also allowed other types of features to emerge as discriminants for solving the problem (coming from ROM and MAS domains).

4. Discussion

The application of ML in CP has traditionally been focused on the paediatric population and the use of instrumental or self-reported data. On the one hand, studies such as that by Ahmadi et al. [19] used supervised models, including random forest and support vector machines, to recognise patterns of physical activity in children based on accelerometer signals. Additionally, Schwartz and Munger [34] used random forest to predict GMFCS levels based on functional questionnaires. Although both studies showed artificial intelligence's potential for automating functional assessment, their models were limited to young subjects and relied on indirect measures. On the other hand, recent reviews by Nahar et al. [18] and Balgude et al. [17] have highlighted the scarcity of studies made with adults and the lack of integration of objective clinical variables.

From the perspective of physiotherapy, even if there are studies that link asymmetries and range of motion limitations to the GMFCS level [14], the present study has added to the analysis variables that quantify spasticity alongside variables that determine asymmetries. In this line, the explainability analysis of the present paper has pointed out the discriminating power of three variables of those natures (LIMIT_SR, LIMIT_ER, MAS_PL) and thus reinforces what was proposed in the aforementioned research.

Concerning deformities such as wind-swept hips and scoliosis, other studies observed how these appear at high GMFCS levels [35]. In addition to those mentioned above, the study presented in this paper also included other variables that cause misalignment due to spinal deformities: dorsal hyperkyphosis, lumbar hyperlordosis, pelvic tilt, and anteversion/retroversion. However, during the present work, none of these variables was found among the subgroup that most influenced the classification of the GMFCS level.

In terms of the ability to predict variables that can determine the GMFCS level, the authors have not found any research seeking this association using automated systems.

In this work, the team found a subset of variables linked to bodily asymmetries to be highly correlated and stable in determining the GMFCS class. This idea suggests that characteristics referring to postural asymmetries are very decisive in determining the GMFCS level, more than variables linked to muscle tone or deformities. In this sense, previous studies already linked asymmetries to high GMFCS levels [14], which reinforces the results presented in this paper.

Taking an overall look at the results presented in this paper, it can be concluded that there is a relationship between the functional assessment of class IV and V patients and the information that can be gathered from the three areas mentioned before (postural quality and ability, ROM and MAS). In addition, some of the variables appear to be more discriminating than others and may provide experts with information to perform a more accurate assessment. In addition, these conclusions have made the authors think that the use of ML algorithms could emerge as a useful tool to make an early prognosis of the patient's possible evolution. Nevertheless, this study used cross-sectional data from subjects at a single time point. This is a limitation of the work that prevents the authors from analysing the prognosis capability of the solution. However, this limitation opens the line that proposes collecting data from the same subjects in the future to analyse the predictive capability of the approach presented in this article.

With regard to the relabelling process proposed in the paper, it suggests the possibility of evaluating possible errors in the GMFCS assessment. As this scale seeks to classify individuals based on their gross motor function, some subjects might fall on the borderline between one score and another due to aspects of functionality that a broad scale does not capture so finely. Although previous studies have shown that the GMFCS has good concordance between measurements [36], the relabelling procedure proposes a new way to better classify those subjects who fall within a range close to two possible scores. Anyway, due to the limited size of the studied sample (a problem that happens more often than desired when working with clinical data), this study was forced to test the efficiency of the relabelling with the whole relabelled dataset and not with an independent partition. This is a weakness that is going to be addressed by the team in the future by collecting a bigger sample that will permit testing the relabelling approach with an independent data partition. Furthermore, having more data will also open the line of optimising the classifiers' hyperparameters, an approach that had a limited scope due to the scarcity of data at the current state of the research.

In this line, the improved results obtained after the relabelling process support the idea that subjects evaluated in borderline areas could be assigned to different scores from the one assigned in a first evaluation. This is particularly relevant for improving the good levels of concordance presented by the scale and allowing for clearer assessment that results in better experimental practice. It is important to bear in mind that health professionals often use the GMFCS score to characterise the functional capabilities of a patient, and, based on that score, they plan different rehabilitation plans, propose realistic and feasible objectives, and assign specific resources to that patient. A clear example of this can be found in wheelchairs, where those designed for GMFCS classes IV and V are completely different in terms of offered functionalities and costs.

In summary, this study expands the application of ML techniques to real standardised clinical and functional data (spasticity, range of motion, deformities, and postural asymmetries) obtained from adults with CP, using a relabelling procedure that improves the accuracy and consistency of GMFCS classification. This approach introduces a novel and clinically relevant perspective by transferring artificial intelligence from the paediatric experimental setting to functional and postural assessment in adulthood, providing a potentially useful tool for the customisation of therapeutic interventions.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/app16010226/s1>. The preprocessed dataset is attached to this article as a supplementary file and can be downloaded from the journal's website.

Author Contributions: Conceptualization, J.G. and R.M.; methodology, J.L.J. and M.-A.R.-N.; validation, M.-A.R.-N. and M.L.S.-S.; formal analysis, A.S.-R.; investigation, J.G. and J.L.J.; resources, J.G., M.-A.R.-N. and M.L.S.-S.; data curation, R.M., J.L.J. and A.S.-R.; writing—original draft preparation, A.S.-R.; writing—review and editing, A.S.-R.; visualization, J.G. and R.M.; supervision, J.L.J., A.S.-R. and M.L.S.-S.; project administration, M.-A.R.-N. and M.L.S.-S.; funding acquisition, R.M., J.L.J. and A.S.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially funded by grant PID2021-123087OB-I00 funded by MICIU/AEI/10.13039/501100011033 and ERDF, a way of making Europe; and by the Department of Education, Universities and Research of the Basque Government (ADIAN, IT-1437-22).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Research in Humans of the Ethics Commission in Experimental Research of the University of Valencia (protocol code FIS-2837811, approved on the 14th of September of 2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors would like to thank AVAPACE for their help with the data collection, as well as the participants in the experiment and their relatives for their accessibility and patience.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CP	Cerebral Palsy
GMFCS	Gross Motor Function Classification System
ML	Machine Learning
PPAS	Posture and Postural Ability Scale
PAL	Postural Ability Level
ROM	Range Of Motion
MAS	Modified Asworth Scale
MRMR	Minimum Redundancy Maximum Relevance
PCA	Principal Component Analysis
AUC	Area Under the Curve

Appendix A

This appendix respectively presents the performance precision, recall and specificity metrics achieved by the classifiers using both the original dataset (Tables A1–A3) and the relabelled dataset (Tables A4–A6).

Table A1. Precision metrics for dimensions between 5 and 78, with a stride of 5 dimensions using the chi-square method for ordering the most relevant features from the original dataset.

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	78
Discr	74.38	70.42	67.34	69.02	66.86	64.69	65.05	64.40	63.25	61.66	60.65	62.16	63.05	60.64	64.09	63.30
Tree	74.12	70.00	70.03	67.96	68.09	65.85	67.86	67.85	68.09	68.04	67.81	67.79	68.62	67.57	67.61	70.41
KNN	72.25	70.58	77.27	75.51	75.69	76.98	79.21	79.10	79.86	82.47	81.05	78.68	77.61	75.03	72.63	78.67
Boost	70.53	71.05	68.59	68.12	72.51	70.39	71.68	69.40	70.14	72.03	70.52	69.19	70.07	70.67	68.70	71.44
Bagg	72.49	72.64	73.57	71.66	74.06	71.61	74.84	75.39	76.45	77.66	77.03	73.48	73.19	72.17	72.89	75.14
Net	70.49	72.53	69.73	71.94	72.37	72.95	72.79	72.16	73.43	71.69	69.98	73.10	71.98	71.46	70.79	72.99
SVM	74.10	71.24	68.04	69.89	65.01	65.41	67.28	64.96	67.68	68.48	68.21	67.98	66.95	66.56	66.16	65.76
Lin	75.26	71.74	70.43	69.45	69.16	68.86	69.35	67.20	70.30	71.09	68.92	69.52	70.25	69.34	69.26	69.43
Kern	73.36	75.25	73.53	65.90	63.35	56.10	54.84	57.07	52.77	55.52	51.07	51.88	53.17	53.25	51.14	51.25

Table A2. Recall metrics for dimensions between 5 and 78, with a stride of 5 dimensions using the chi-square method for ordering the most relevant features from the original dataset.

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	78
Discr	75.82	71.33	71.00	68.04	65.93	67.44	64.63	63.19	65.52	62.22	64.19	63.37	65.44	63.63	66.33	65.00
Tree	72.11	68.19	67.26	67.52	67.81	70.96	69.59	68.85	69.82	70.44	71.11	70.41	69.07	67.67	67.33	68.07
KNN	67.56	68.52	72.41	72.00	76.19	76.22	76.48	74.41	73.11	74.81	75.15	78.04	76.59	77.96	76.22	81.59
Boost	72.67	68.04	69.63	68.52	71.67	68.85	69.67	66.85	69.70	72.44	69.67	71.30	69.04	70.78	71.11	69.00
Bagg	73.67	70.89	72.00	70.48	71.52	69.15	69.15	68.11	69.56	69.56	68.85	71.19	69.52	70.52	71.52	70.56
Net	75.22	68.78	70.04	71.52	71.85	68.56	71.33	69.26	71.30	70.70	69.22	68.19	73.41	71.70	68.44	66.11
SVM	75.19	70.59	70.56	67.93	69.85	68.85	68.56	66.93	69.00	69.52	67.85	67.11	67.89	65.04	66.56	66.56
Lin	76.44	72.63	70.26	71.33	71.56	68.81	73.41	70.56	73.78	70.63	69.93	69.89	70.74	70.30	70.96	71.00
Kern	70.37	62.67	57.67	55.67	57.11	54.44	62.26	68.44	65.22	71.85	67.37	65.81	70.15	76.37	72.41	71.37

Table A3. Specificity metrics for dimensions between 5 and 78, with a stride of 5 dimensions using the chi-square method for ordering the most relevant features from the original dataset.

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	78
Discr	70.74	66.67	62.22	65.93	63.33	59.26	61.85	61.48	57.78	58.15	55.19	57.78	57.78	57.41	59.63	59.26
Tree	71.11	67.41	66.67	65.19	64.44	58.89	60.74	61.48	61.11	60.74	60.37	60.74	62.22	61.48	61.85	65.19
KNN	70.00	67.78	75.93	74.07	72.96	74.81	77.78	78.15	78.89	80.74	78.52	75.93	74.81	71.11	68.15	75.56
Boost	65.93	68.52	64.81	65.19	69.26	67.78	69.26	67.41	67.78	68.52	65.93	63.70	65.56	66.67	64.44	68.89
Bagg	67.78	70.74	70.37	69.26	71.85	70.37	73.33	74.81	75.93	77.41	76.67	71.85	71.85	70.00	70.74	74.44
Net	64.07	70.74	66.67	68.89	69.26	70.00	69.63	70.37	71.48	69.26	67.78	71.11	66.67	68.89	68.52	73.33
SVM	71.11	68.89	64.07	67.04	58.89	60.00	62.59	60.37	63.70	65.56	65.56	65.19	62.59	64.44	62.96	62.96
Lin	71.85	68.89	67.04	65.93	64.44	64.81	64.44	62.96	65.93	68.15	65.93	65.93	66.30	65.56	65.56	65.56
Kern	70.74	75.56	75.93	69.26	61.11	52.59	44.81	42.96	36.67	37.78	29.63	34.44	33.70	27.78	26.30	25.93

Table A4. Precision metrics for dimensions between 5 and 78, with a stride of 5 dimensions using the chi-square method for ordering the most relevant features from the relabelled dataset.

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	78
Discr	82.20	81.30	81.35	76.86	74.34	75.19	70.18	70.01	70.33	75.87	79.29	81.34	85.69	88.02	86.90	86.47
Tree	85.17	86.74	90.02	90.59	89.63	89.59	89.56	89.04	89.01	89.64	89.11	87.80	87.51	87.47	86.93	84.71
KNN	83.53	87.29	83.52	86.45	86.78	87.68	86.69	87.62	86.54	86.13	85.02	87.31	86.24	84.73	86.25	86.42
Boost	83.41	87.94	89.14	93.66	91.32	89.99	91.37	89.25	88.67	89.69	89.36	86.69	86.82	87.25	86.67	83.76
Bagg	85.86	87.10	90.48	93.33	93.65	93.13	93.57	92.40	91.66	91.00	90.65	90.21	89.90	89.90	90.26	89.58
Net	84.37	85.70	84.98	88.42	88.21	91.64	94.52	93.21	93.01	94.02	92.15	94.19	92.89	93.58	92.52	94.58
SVM	83.03	83.98	83.15	85.22	87.61	90.25	90.97	89.68	90.50	90.39	91.04	90.80	89.88	89.47	88.69	91.47
Lin	82.95	85.99	86.03	86.84	88.72	91.51	90.61	91.35	90.97	92.81	91.33	91.69	91.89	91.38	92.01	93.38
Kern	86.96	88.23	81.29	79.15	69.31	63.69	61.62	51.83	52.20	54.13	54.70	54.03	52.74	53.55	54.84	52.80

Table A5. Recall metrics for dimensions between 5 and 78, with a stride of 5 dimensions using the chi-square method for ordering the most relevant features from the relabelled dataset.

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	78
Discr	87.33	86.00	84.00	80.00	78.33	73.67	73.67	71.00	76.67	79.00	82.00	87.67	88.00	90.00	89.33	88.33
Tree	83.33	81.33	83.67	85.67	85.67	86.00	86.67	84.67	83.33	83.67	84.67	85.00	84.67	84.33	84.67	86.00
KNN	84.33	87.67	88.00	89.33	92.33	91.00	91.67	92.67	91.67	90.67	91.67	90.67	91.33	92.00	94.00	93.00
Boost	84.67	82.67	86.67	86.00	87.67	88.00	90.33	89.00	87.67	87.33	87.67	88.33	88.67	88.33	88.00	88.33
Bagg	87.67	86.67	88.67	90.33	89.67	91.00	89.67	89.67	89.67	91.00	90.33	90.67	90.00	90.67	91.00	91.33
Net	85.00	85.00	85.33	86.33	88.00	88.00	89.00	91.00	91.67	90.67	91.67	93.00	94.00	92.67	95.33	94.67
SVM	87.33	84.67	85.00	87.33	90.00	90.67	89.00	90.33	91.33	92.33	92.33	92.33	93.33	92.67	93.67	93.67
Lin	86.67	88.00	89.00	89.00	90.00	89.67	90.00	91.33	92.33	92.00	92.67	93.33	93.33	93.33	93.67	94.33
Kern	82.00	69.33	61.67	59.67	61.33	66.33	71.00	75.00	76.67	82.67	78.67	79.33	83.67	87.33	90.00	83.00

Table A6. Specificity metrics for dimensions between 5 and 78, with a stride of 5 dimensions using the chi-square method for ordering the most relevant features from the relabelled dataset.

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	78
Discr	76.02	74.44	75.51	70.83	67.31	69.31	62.69	63.84	61.62	69.31	73.61	74.58	80.88	84.44	82.87	82.45
Tree	81.06	83.29	87.92	88.70	86.85	86.85	86.81	87.08	87.08	87.50	86.67	84.07	83.66	83.66	82.55	79.72
KNN	78.70	83.75	78.38	82.18	82.41	83.66	82.18	83.61	82.08	82.08	80.46	83.94	82.04	79.81	81.34	81.76
Boost	78.01	85.97	85.60	92.73	89.31	87.82	89.35	86.67	85.83	86.76	86.44	81.81	81.76	82.59	81.48	78.29
Bagg	81.44	83.33	87.13	91.76	92.13	91.39	92.08	90.56	89.35	88.61	88.19	87.45	87.04	87.04	87.41	86.25
Net	79.44	82.87	81.02	85.46	84.77	89.44	93.15	91.67	90.74	92.31	90.09	92.45	90.74	91.53	90.09	92.59
SVM	77.55	79.17	78.33	80.65	83.66	87.13	88.19	86.71	87.41	87.08	88.15	87.73	86.16	85.69	84.68	88.15
Lin	77.55	81.48	81.48	83.19	85.19	88.70	87.59	88.70	87.87	90.56	88.61	88.94	88.94	88.47	89.21	91.16
Kern	83.61	86.44	78.15	74.49	63.38	51.16	42.13	20.32	18.66	19.35	24.12	21.67	13.10	12.08	13.98	14.35

References

- Rosenbaum, P.; Paneth, N.; Leviton, A.; Goldstein, M.; Bax, M.; Damiano, D.; Dan, B.; Jacobsson, B. A report: The definition and classification of cerebral palsy April 2006. *Dev. Med. Child Neurol. Suppl.* **2007**, *49*, 480.
- Oskoui, M.; Coutinho, F.; Dykeman, J.; Jetté, N.; Pringsheim, T. An update on the prevalence of cerebral palsy: A systematic review and meta-analysis. *Dev. Med. Child Neurol.* **2013**, *55*, 509–519. [[CrossRef](#)]
- Sadowska, M.; Sarecka-Hujar, B.; Kopyta, I. Cerebral Palsy: Current Opinions on Definition, Epidemiology, Risk Factors, Classification and Treatment Options. *Neuropsychiatr. Dis. Treat.* **2020**, *16*, 1505–1518. [[CrossRef](#)]
- Stavsky, M.; Mor, O.; Mastrolia, S.A.; Greenbaum, S.; Than, N.G.; Erez, O. Cerebral Palsy—Trends in Epidemiology and Recent Development in Prenatal Mechanisms of Disease, Treatment, and Prevention. *Front. Pediatr.* **2017**, *5*, 21. [[CrossRef](#)]
- Hankins, G. Defining the pathogenesis and pathophysiology of neonatal encephalopathy and cerebral palsy. *Obstet. Gynecol.* **2003**, *102*, 628–636. [[CrossRef](#)]
- Mathewson, M.A.; Lieber, R.L. Pathophysiology of Muscle Contractures in Cerebral Palsy. *Phys. Med. Rehabil. Clin. N. Am.* **2015**, *26*, 57–67. [[CrossRef](#)]
- Albright, A.L. Spasticity and movement disorders in cerebral palsy. *Child's Nerv. Syst.* **2023**, *39*, 2877–2886. [[CrossRef](#)]
- Holmes, C.; Brock, K.; Morgan, P. Postural asymmetry in non-ambulant adults with cerebral palsy: A scoping review. *Disabil. Rehabil.* **2018**, *41*, 1079–1088. [[CrossRef](#)]
- Fulford, G.E.; Brown, J.K. Position as a Cause of Deformity in Children with Cerebral Palsy. *Dev. Med. Child Neurol.* **1976**, *18*, 305–314. [[CrossRef](#)] [[PubMed](#)]
- Ágústsson, A.; Sveinsson, T.; Pope, P.; Rodby-Bousquet, E. Preferred posture in lying and its association with scoliosis and windswept hips in adults with cerebral palsy. *Disabil. Rehabil.* **2018**, *41*, 3198–3202. [[CrossRef](#)] [[PubMed](#)]
- Morrell, D.S.; Pearson, J.M.; Sauser, D.D. Progressive Bone and Joint Abnormalities of the Spine and Lower Extremities in Cerebral Palsy. *RadioGraphics* **2002**, *22*, 257–268. [[CrossRef](#)]
- Palisano, R.; Rosenbaum, P.; Walter, S.; Rossell, D.; Wood, E.; Galuppi, B. Development and reliability of a system to classify gross motor function in children with cerebral palsy. *Dev. Med. Child Neurol.* **1997**, *39*, 214–223. [[CrossRef](#)] [[PubMed](#)]
- Haak, P.; Lenski, M.; Hidecker, M.; Li, M.; Paneth, N. Cerebral palsy and aging. *Dev. Med. Child Neurol.* **2009**, *51*, 16–23. [[CrossRef](#)]
- Rodby-Bousquet, E.; Czuba, T.; Häggglund, G.; Westbom, L. Postural asymmetries in young adults with cerebral palsy. *Dev. Med. Child Neurol.* **2013**, *55*, 1009–1015. [[CrossRef](#)] [[PubMed](#)]
- Rodby-Bousquet, E.; Agustsson, A. Postural Asymmetries and Assistive Devices Used by Adults With Cerebral Palsy in Lying, Sitting, and Standing. *Front. Neurol.* **2021**, *12*, 758706. [[CrossRef](#)]

16. Cankurtaran, D.; Abidin, N.; Unlu Akyuz, E.; Tezel, N. An investigation into the factors which affect the sitting balance of non-ambulatory children with cerebral palsy. *Neurol. Asia* **2021**, *26*, 575–583. [\[CrossRef\]](#)
17. Balgude, S.D.; Gite, S.; Pradhan, B.; Lee, C.W. Artificial intelligence and machine learning approaches in cerebral palsy diagnosis, prognosis, and management: A comprehensive review. *PeerJ Comput. Sci.* **2024**, *10*, e2505. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Nahar, A.; Paul, S.; Saikia, M. A systematic review on machine learning approaches in cerebral palsy research. *PeerJ* **2024**, *12*, e18270. [\[CrossRef\]](#)
19. Ahmadi, M.; O'Neil, M.; Fragala-Pinkham, M.; Lennon, N.; Trost, S. Machine learning algorithms for activity recognition in ambulant children and adolescents with cerebral palsy. *J. Neuroeng. Rehabil.* **2018**, *15*, 105. [\[CrossRef\]](#)
20. McDowell, B.C.; Kerr, C.; Parkes, J. Interobserver agreement of the Gross Motor Function Classification System in an ambulant population of children with cerebral palsy. *Dev. Med. Child Neurol.* **2007**, *49*, 528–533. [\[CrossRef\]](#)
21. Benedict, R.E.; Patz, J.; Maenner, M.J.; Arneson, C.L.; Yeargin-Allsopp, M.; Doernberg, N.S.; Van Naarden Braun, K.; Kirby, R.S.; Durkin, M.S. Feasibility and reliability of classifying gross motor function among children with cerebral palsy using population-based record surveillance. *Paediatr. Perinat. Epidemiol.* **2010**, *25*, 88–96. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Rodby-Bousquet, E.; Ágústsson, A.; Jónsdóttir, G.; Czuba, T.; Johansson, A.; Hägglund, G. Interrater reliability and construct validity of the Posture and Postural Ability Scale in adults with cerebral palsy in supine, prone, sitting and standing positions. *Clin. Rehabil.* **2012**, *28*, 82–90. [\[CrossRef\]](#)
23. Moromizato, K.; Kimura, R.; Fukase, H.; Yamaguchi, K.; Ishida, H. Whole-body patterns of the range of joint motion in young adults: Masculine type and feminine type. *J. Physiol. Anthropol.* **2016**, *35*, 23. [\[CrossRef\]](#) [\[PubMed\]](#)
24. American Academy of Orthopaedic Surgeons; British Orthopaedic Association. *Joint Motion: Method of Measuring and Recording*; Edinburgh British Orthopaedic Association: Edinburgh, UK, 1965.
25. Langensiepen, S.; Semler, O.; Sobottke, R.; Fricke, O.; Franklin, J.; Schönauf, E.; Eysel, P. Measuring procedures to determine the Cobb angle in idiopathic scoliosis: A systematic review. *Eur. Spine J.* **2013**, *22*, 2360–2371. [\[CrossRef\]](#)
26. Harb, A.; Margetis, K.; Kishner, S. Modified Ashworth Scale. Internet, 2025. Available online: <https://www.ncbi.nlm.nih.gov/books/NBK554572/> (accessed on 6 November 2025).
27. Palisano, R.; Rosenbaum, P.; Bartlett, D.; Livingston, M. Content validity of the expanded and revised Gross Motor Function Classification System. *Dev. Med. Child Neurol.* **2008**, *50*, 744–750. [\[CrossRef\]](#)
28. Mitchell, T.M. *Machine Learning*, 1st ed.; McGraw-Hill, Inc.: New York, NY, USA, 1997.
29. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press: Cambridge, MA, USA, 2014. [\[CrossRef\]](#)
30. Purushotham, S.; Tripathy, B.K. Evaluation of Classifier Models Using Stratified Tenfold Cross Validation Techniques. In *Global Trends in Information Systems and Software Applications*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 680–690. [\[CrossRef\]](#)
31. Sammut, C.; Webb, G. (Eds.) *Encyclopedia of Machine Learning*; Springer: New York, NY, USA, 2010. [\[CrossRef\]](#)
32. Lundberg, S.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874. [\[CrossRef\]](#)
33. Matlab. Shapley—Shapley Values. Internet, 2025. Available online: <https://es.mathworks.com/help/stats/shapley.html> (accessed on 14 November 2025).
34. Schwartz, M.H.; Munger, M.E. Using machine learning to overcome challenges in GMFCS level assignment. *Gait Posture* **2018**, *63*, 290–295. [\[CrossRef\]](#)
35. Porter, D.; Michael, S.; Kirkwood, C. Patterns of postural deformity in non-ambulant people with cerebral palsy: What is the relationship between the direction of scoliosis, direction of pelvic obliquity, direction of windswept hip deformity and side of hip dislocation? *Clin. Rehabil.* **2007**, *21*, 1087–1096. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Jahnsen, R.; Aamodt, G.; Rosenbaum, P. Gross Motor Function Classification System used in adults with cerebral palsy: Agreement of self-reported versus professional rating. *Dev. Med. Child Neurol.* **2006**, *48*, 734. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.