An ordinal CNN approach for the assessment of neurological damage in Parkinson's disease patients

Javier Barbero-Gómez [a,*], Pedro-Antonio Gutiérrez [a], Víctor-Manuel Vargas [a], Juan-Antonio Vallejo-Casas [b], César Hervás-Martínez [a]

[a] Departamento de Informática y Análisis Numérico, Universidad de Córdoba, Campus de Rabanales, Edificio Albert Einstein, 14014, Córdoba, Spain
[b] UGC Medicina Nuclear, Hospital Universitario Reina Sofía, Universidad de Córdoba, IMIBIC, 14004 Córdoba, Spain

ARTICLE INFO

ABSTRACT

3D image scans are an assessment tool for neurological damage in Parkinson's disease (PD) patients. This diagnosis process can be automatized to help medical staff through Decision Support Systems (DSSs), and Convolutional Neural Networks (CNNs) are good candidates, because they are effective when applied to spatial data. This paper proposes a 3D CNN ordinal model for assessing the level or neurological damage in PD patients. Given that CNNs need large datasets to achieve acceptable performance, a data augmentation method is adapted to work with spatial data. We consider the Ordinal Graph-based Oversampling via Shortest Paths (OGO-SP) method, which applies a gamma probability distribution for inter-class data generation. A modification of OGO-SP is proposed, the OGO-SP-β algorithm, which applies the beta distribution for generating synthetic samples in the inter-class region, a better suited distribution when compared to gamma. The evaluation of the different methods is based on a novel 3D image dataset provided by the Hospital Universitario 'Reina Sofía' (Córdoba, Spain). We show how the ordinal methodology improves the performance with respect to the nominal one, and how OGO-SP-β yields better performance than OGO-SP.

1. Introduction

In the medical field, plenty of data is obtained from patients in order to give an accurate diagnosis and decide on a beneficial treatment. From simple scalar to fully structured spatiotemporal data, like electrocardiograms, projectional radiographs, and computed tomography scans, all of these sources aid medical staff look for abnormalities and signs of degradation or damage. Nonetheless, these examinations are time-consuming and costly for hospitals and other medical administrations. A great deal of expertise is needed in order to interpret all this data into useful information.

1.1. Machine learning

Machine Learning (ML) techniques can automatize some of these tasks by leveraging on existing data. They can serve as the core to a Decision Support System (DSS) in order to help medical staff make a better judgment or give an additional opinion.

In the case of structured data, Convolutional Neural Networks (CNNs) have shown excellent performance in tasks such as classification, segmentation or regression. These models rely on large amounts of labeled data in order to learn. Examples of such tasks can be diagnosis (classification) (El-Dahshan, Mohsen, Revett, & Salem, 2014; Mazaheri & Khodadadi, 2020), automatic identification of anatomical regions (segmentation) (Peng, Xiong, Peng, & Lu, 2020) or landmark location (regression) (Payer, Štern, Bischof, & Urschler, 2019).

* Corresponding author.
E-mail addresses: jbarbero@uco.es (J. Barbero-Gómez), pagutierrez@uco.es (P.-A. Gutiérrez), vvargas@uco.es (V.-M. Vargas), jantonio.vallejo.sspa@juntadeandalucia.es (J.-A. Vallejo-Casas), chervas@uco.es (C. Hervás-Martínez).

1.2. Parkinson's disease

One of many medical applications of Machine Learning (ML) explored in the literature is Parkinson's disease (PD) diagnosis. PD is a degenerative nervous system disorder affecting the brain whose symptoms are primarily motor-related: shaking, gait disturbances, slowness and difficulty walking. Other symptoms are related to sleeping, emotional or sensory problems. The cost on society of this disease grows as the symptoms worsen, as the greatest component of cost is in patient care and nursing home costs. Just in the UK, the total cost has been estimated to be between £445 million and £3.3 billion annually (Findley, 2007).

Assessing the severity of the neurological damage of PD patients is a crucial part for the correct treatment, as an unnecessarily high dose of levodopa (the most common medication for PD) may worsen symptoms in the long-term (Tomlinson et al., 2010). To evaluate this, doctors rely both on observations of motor capabilities (Marino et al., 2012) as well as imaging techniques like Magnetic Resonance Imaging (MRI) (Marino et al., 2012) and nuclear tomography like Single-Photon Emission Computerized Tomography (SPECT) or Positron Emission Tomography (PET) (Arbizu et al., 2014). Ioflupane (^{123}I) (known commercially as DaTscan) is a neuro-imaging drug often used to evaluate the dopaminergic activity in the nigrostriatal dopaminergic pathway when the disease may be in the early stages (Darcourt et al., 2010). It is injected into the patient's bloodstream before taking a SPECT image, so that the brain's dopaminergic activity can be inspected visually. The evaluation of this damage may require a great deal of experience.

In recent years there has been a growing interest on the application of ML techniques to this kind of images, which require no previous assumptions on the regions of interest or relevant areas for the given task. Instead, these methods are able to discover for themselves where and what to look for in images, based solely on data previously labeled by doctors. Popular methods tackle binary (healthy control or disease) (Vinícius dos Santos Ferreira et al., 2018) or nominal (healthy control, disease A, disease B, …) classification (Akyol, 2020).

Some datasets related to PD are available online for research use, such as the Parkinson's Progression Markers Initiative (PPMI)[1] or the LRRK2 Cohort Consortium (LCC).[2]

1.3. Ordinal classification

In the last decade, the concept of ordinal classification (sometimes referred as ordinal regression) has been popularized as a way to exploit extra information in a classification problem where a natural ordering of the classes is present (Gutiérrez, Pérez-Ortiz, Sánchez-Monedero, Fernández-Navarro, & Hervás-Martínez, 2016; Ben-David, 2008). This approach has been proven to outperform the classic nominal perspective in medical applications such as melanoma diagnosis (Sánchez-Monedero, Pérez-Ortiz, Sáez, Gutiérrez, & Hervás-Martínez, 2018) and liver transplantation (Dorado-Moreno et al., 2017). Up until now, research into this methodology applied to PD diagnosis has not been addressed, because existing scales, such as the Hoehn and Yahr scale (Hoehn & Yahr, 1967) or the Unified Parkinson's Disease Rating Scale (UPDRS), require subjective evaluation from clinicians of tremor, rigidity or movement (Ramaker, Marinus, Stiggelbout, & Van Hilten, 2002) and are difficult to quantify in a single measure.

A complete taxonomy of ordinal classification methods can be found in Gutiérrez et al. (2016). The most naive methods perform a simple regression using the class labels and then round the values when predicting (Kramer, Widmer, Pfahringer, & de Groeve, 2010) or simply apply a label distance cost penalty to a classical nominal classification

method (Kotsiantis & Pintelas, 2004). These basic approaches lack an understanding of the underlying label distance, as different misclassifications may represent a different error cost.

Threshold Models are another popular approach for this task. An underlying latent continuous variable is assumed to exist, from which the different ranks arise by assigning certain thresholds. Thus, in this framework, both the value of the latent variable and the thresholds need to be learned from the data. Some approaches, like the classical Proportional Odds Model (POM) (McCullagh, 1980) or the more recent `gologit` model (Williams, 2006) fall into the Cumulative Link Model (CLM) framework, a probabilistic method for predicting probabilities of groups of contiguous categories, taking the ordinal scale into account.

Other ordinal approaches consist on decomposing the ordinal problem into a set of binary problems (called Ordinal Binary Decomposition (OBD)). Sometimes these decompositions are solved by a set of different models, like in the cascade linear utility model (Wu et al., 2003). In other cases they are modeled by several outputs of the same underlying model (Cheng, Wang, & Pollastri, 2008). All OBDs present the same challenge: combining the results of all decompositions into a single final classification. The simplest approaches assign the first class to reach a certain threshold (Wu et al., 2003), but this can lead to an imbalance as the last classes are more difficult to select. Error Correcting Codes (ECOCs) are better suited to this task, as this approach considers all outputs equally in the final decision (Allwein, Schapire, & Singer, 2001).

The available PD datasets mentioned previously only provide label information of binary, nominal or continuous nature. Our study presents a novel dataset, collected by the Clinical Management Unit of Nuclear Medicine of the Hospital Universitario 'Reina Sofía' (Córdoba, Spain), containing SPECT Ioflupane (^{123}I) images of PD patients classified in 4 distinct ordinal labels, according to their stage of presynaptic dopamine binding damage ('no alteration', 'slight alteration', 'more advanced alteration' and 'severe alteration'). Ordinal methods are perfect candidates to tackle the task of predicting these labels, which serve as a better indicator for the medical decision process.

1.4. Data augmentation

Classification tasks may suffer from unbalanced data, especially in the medical field, as healthy patients are much more common than sick patients. Also, data gathering and proper labeling is expensive and time consuming. In this situation, data augmentation techniques are required in order to boost the performance of ML models.

The most basic strategy for augmenting spatial data, like medical images, is image translation, rotation, flipping and cropping (Perez & Wang, 2017). Several of these can be used depending on the specific task to learn. For example, object detection tasks, such as anomaly or lesion detection, can be accelerated by using cropped Region of Interest (ROI) as samples (Rey, Arcay, & Castro, 2020).

Classic techniques like Synthetic Minority Oversampling (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Rivera & Xanthopoulos, 2016) perform well on low dimensional data. Some techniques, such as Autoencoders (Hinton & Zemel, 1993) or Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) are able to leverage convolutional operations, which improve the performance and efficiency over spatial data. Those techniques, however, require a vast amount of training examples in order to avoid pitfalls like mode collapse.

In order to meet this data volume requirement, more sophisticated data augmentation methods are being applied to medical data as of recently. As an example, Salazar, Vergara, and Safont (2021) combine GANs with Markov Random Field models to augment 3D functional MRI multi-subject data and enhance nominal classification performance.

There also exist data augmentation methodologies that make use of ordinal information in the class labels to improve the synthetic data generation process. A family of such methods, presented by Pérez-Ortiz, Gutiérrez, Hervás-Martínez, and Yao (2015), are the Ordinal Graph-based Oversampling (OGO) methods. These consist on computing a

---

[1] http://www.ppmi-info.org/.
[2] https://www.neurodegenerationresearch.eu/cohort/lrrk2-cohort-consort ium/.

graph estimating the latent manifold structure in the data by exploiting ordinal information in the labels, and then using the edges on that graph to generate samples, similar to SMOTE.

Classic proven techniques such as SMOTE and OGO can be adapted to work on spatial data: a CNN can first be trained so that a projection from high-dimensional data is learned, and then a traditional data augmentation method can be applied to the resulting low-dimensional data.

1.5. Limitations of OGO

Pérez-Ortiz et al. (2015) propose the use of a gamma distribution for the generation of samples in the inter-class space of the latent manifold. This distribution is skewed towards the part of the graph adjacent to the class to be augmented but permits the inclusion of features in the frontier of both classes.

This, however, presents some limitations. The gamma distribution is not suited for the generation of values in a closed domain (in this case, $\delta \in [0,1]$), as its original domain is in the interval $[0,\infty]$. This is not only a theoretical overlook, but it also hinders the ability to tune its parameters in a meaningful way, subject to the dataset to which it is applied.

In this paper, an alternative to the gamma distribution is proposed: using the better suited beta distribution, which is bounded in the $[0,1]$ interval and has easier-to-tune parameters, enabling a higher degree of flexibility which can help achieve better performance.

1.6. Goals

To the best of our knowledge, there are no previous works using ML methods for the assessment of the severity of brain damage from a patient's brain SPECT 3D image. These methods could help doctors in the diagnosis and treatment of PD and other parkinsonisms through DSSs and contribute towards the relief of the public health cost of this disease.

Thus, the goals of the present work are:

1. Exploring the potential classification performance improvement in using ordinal label information.
2. Adapting the use of classical data augmentation and class balancing techniques to spatial three-dimensional data.
3. Analyzing the developed methodologies in points 1 and 2 using a novel and extensive database of SPECT images from Hospital 'Reina Sofía' (Córdoba, Spain).
4. Studying a potential improvement to the data augmentation methodology presented by Pérez-Ortiz et al. (2015) by applying a better suited probability distribution for generating synthetic samples in the class frontiers.

The rest of this paper is organized as follows: In Section 2 we describe the data to be used and we propose a fully 3D Deep CNN model for the evaluation of presynaptic deficit in SPECT Ioflupane (^{123}I) images. We design two versions of the same base model, one nominal and the other ordinal, differing on the output layer shape and the activation function as well as the loss function. We propose a novel approach for data augmentation using the beta probability distribution with efficiently estimated parameters. In Section 3, we describe the design of the experiments and metrics for the evaluation of the proposed methods. We show that our proposal performs better than the nominal methodology as well as the previous ordinal method. Finally, in Section 4, we discuss the results and propose future work.

2. Materials and methods

In this section, we present the dataset used for the experimentation, as well as the architecture of the proposed models and the novel data augmentation proposal.

These models have been designed to tackle the task of assessing the alteration of dopaminergic activity in the brain of PD patients by examining 3D SPECT scans of the brain.

2.1. Data description

The dataset consists of 508 3D images provided by the Clinical Management Unit of Nuclear Medicine of the Hospital Universitario 'Reina Sofía' (Córdoba, Spain). They are obtained by first administering the patients with Ioflupane (^{123}I), a radiopharmaceutical which binds to the presynaptic dopamine transporters in the brain. Some time later, a SPECT scan is performed, so as to inspect the dopaminergic activity in the nigrostriatal dopaminergic pathway, which is one of the neuropathological characteristics of PD.

Of these images, 314 (61.8%) are of healthy patients (class 0), 42 (8.3%) show slight alteration (class 1), 52 (10.2%) show more advanced alteration (class 2) and 100 (19.7%) show severe alteration (class 3). It is common for medical diagnosis datasets like this one to have a severe imbalance problem. In our case, more than 60% of samples are of healthy patients, and less than 10% belong to class 1.

The doctors' diagnosis is attached as a class label to each of the images. Because of the gradual nature of these classes, the task of recognizing which of the four categories an image belongs to can be posed as an ordinal classification problem and, thus, specific techniques can be employed to exploit the order information.

Automatic linear image registration has been performed on all images using the FMRIB's Linear Image Registration Tool (FLIRT) from the FMRIB Software Library (FSL) (Smith et al., 2004) considering the T2 version of the MNI152 2 mm standard space SPECT template (Evans, Janke, Collins, & Baillet, 2012). Thus, all images have a final resolution of $91 \times 109 \times 91$ voxels. Also, during training, the symmetrical nature of the images is exploited: with 50% probability, the images are flipped on the frontal axis (left to right) each time they are used.

2.2. Global architecture

The overall architecture of the CNN model considered in this paper consists on convolutional blocks of repeating layers reducing the size of the image while increasing the number of feature maps. Afterwards, the output of the convolutional part of the network is fed into a fully connected layer of neurons before computing the output decision of the model.

Each convolutional block consists of a 3D convolutional layer followed by a batch normalization layer. The kernel size and the stride of the convolution are parameters to be cross-validated during the training process. The output of each block is then fed into a Leaky Rectified Linear Unit (LReLU) activation function, which has been proven to have good convergence properties such as scale-invariance and 1-Lipschitz continuity (Suzuki, 2018).

The low resolution feature maps which are the output of the convolutional blocks are then the input of a densely connected neuron layer (the number of neurons of this layer is also cross-validated during training) using, again, LReLU as the activation function. A final output layer computes the final classification given by the model. In the training phase, the model weights are updated using the Adam optimization algorithm (Kingma & Ba, 2017) so that the outputs align with the annotated labels.

We test two different architectures: a classic architecture based on nominal classification and an ordinal architecture, considering the ordering of the class labels. In Fig. 1, it can be noted that both architectures share the same structure for their convolutional part, but they are different in the way the final output is computed. Moreover, two different class balancing methods are applied in each case, as explained in Section 2.5.
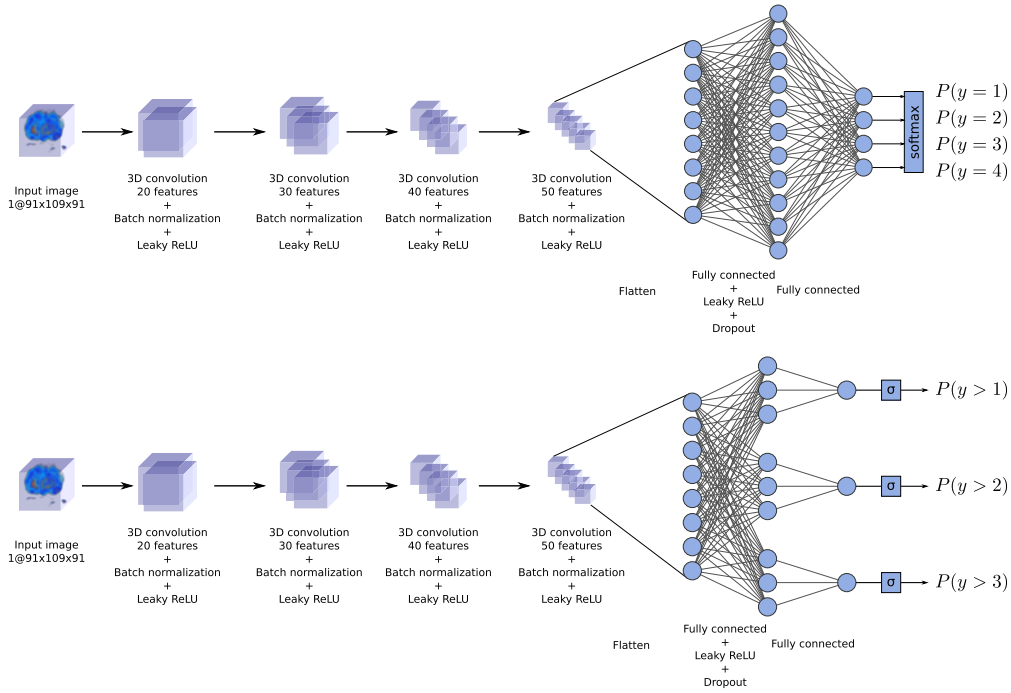
Fig. 1. The two network architectures: nominal (above) and ordinal (below).

2.3. Classic nominal classification architecture

As is the case with classic CNN architectures, the output of the convolutional part of the network is then fed to a single fully connected layer. This is then again fully connected with the output layer, which has as many outputs as classes to decide. Then, a softmax activation function maps the output of the network into a set of probabilities o_q of belonging to each class $\mathscr{C}_q : o_q = P(y = \mathscr{C}_q)$.

The optimizing criterion is the categorical cross-entropy loss, described as:

$$\ell(x_i) = -\sum_{q=1}^{Q} 1\{y_i = \mathscr{C}_q\}\log(P(y_i = \mathscr{C}_q|\mathbf{x}_i)), \tag{1}$$

where Q is the number of classes, $1\{y_i = \mathscr{C}_q\}$ is the indicator function that is equal to 1 when $y_i = \mathscr{C}_q$ and 0 otherwise and $P(y = \mathscr{C}_q|\mathbf{x}_i)$ is the predicted probability of $\mathbf{x}_i$ belonging to class $\mathscr{C}_q$. To evaluate the effectiveness of the trained model, a new unseen given sample $\mathbf{x}$ is classified as belonging to the class with maximum predicted probability

$$\widehat{y} = argmax_{1\leqslant q\leqslant Q}P(y = \mathscr{C}_q|\mathbf{x}).$$

2.4. Proposed ordinal classification architecture

When the classes are ordered, instead of dealing with the complete problem as previously mentioned, OBD can be applied: the problem is decomposed into $Q-1$ binary decision problems. Each problem q consists on deciding if $y \succ \mathscr{C}_q$ conditioned to $\mathbf{x}$ ($1\leqslant q < Q$).

This would normally require $Q-1$ different models, each solving one of these binary problems. This approach, originally presented by Frank and Hall (2001), would imply to first compute every probability $p_q = P(y = \mathscr{C}_q)$ based on the obtained models and then select the highest probability class. The individual probabilities are computed as a function of the cumulative probabilities, $P(y \succ \mathscr{C}_q)$, estimated by the binary models:

$$\begin{aligned} p_1 &= P(y = \mathscr{C}_1) = 1 - P(y \succ \mathscr{C}_1)\\ p_q &= P(y = \mathscr{C}_q) = P(y \succ \mathscr{C}_{q-1}) - P(y \succ \mathscr{C}_q) \quad \forall 1 < q < Q\\ p_Q &= P(y = \mathscr{C}_1) = P(y \succ \mathscr{C}_{Q-1}). \end{aligned}$$

However, different problems are associated to this approach: because the outputs of the different decompositions are not combined in the same training process, the basic probability assumptions (that is, $P(y \succ \mathscr{C}_q) \geqslant P(y \succ \mathscr{C}_{q+1}), p_q \geqslant 0$ and $\sum_q p_q = 1$) are not necessarily satisfied, which can lead to inconsistencies. Moreover, when computing the individual probabilities, at most only two of the model outputs are considered, instead of all of them. Finally, when obtaining the decomposition, the ratio of positive to negative samples becomes very unbalanced in the extreme classes. In the case of an already unbalanced dataset, this procedure can become unrealistic.

To circumvent this limitation, a compromise is proposed: a single convolutional model can be trained simultaneously to then be fed into multiple fully connected blocks, each one solving an individual binary classification subproblem. This way, the imbalance of the training data can be less acute, and training can be done in parallel. The output of each of the $Q-1$ fully connected blocks has a sigmoid activation function representing the probability $o_k = P(y \succ \mathscr{C}_k|\mathbf{x}) \in (0,1)$.

Moreover, for obtaining the final probabilities, we use a more stable option based on the decision function of the ECOC framework (Allwein et al., 2001). The correct ideal output code for each class is considered as the coordinates of a vertex of a hypercube in $Q-1$ dimensions, e.g. for a 4 class ordinal problem classes $\mathscr{C}_1, \mathscr{C}_2, \mathscr{C}_3$ and $\mathscr{C}_4$ would be associated to the codes $(0,0,0), (1,0,0), (1,1,0)$ and $(1,1,1)$, respectively. This way, all the model outputs are considered for classification. To decide which class a sample $\mathbf{x}$ belongs to, the class with the nearest code according to some distance d is selected:

$$\widehat{y} = argmin_{1\leqslant q\leqslant Q} d(\mathbf{o}, \mathbf{c}_q),$$

where $\mathbf{o} = (o_1, o_2, ..., o_{Q-1})$ is the vector of output values, and $\mathbf{c}_q$ is the code vector associated with class $\mathscr{C}_q$. It can be noted that the aforementioned probability assumptions do not need to hold in order to assign a class in this framework.

Instead of individually training the different binary subproblems (as

proposed for ECOC) and to be consistent with the decision criterion, the global optimization criterion of the network is set to be the mean squared error (*MSE*) loss:

$$\ell(\mathbf{x}_i) = \frac{1}{Q-1} \sum_{k=1}^{Q-1} \left(1\{y_i \succ \mathscr{C}_k\} - P(y_i \succ \mathscr{C}_k \,|\, \mathbf{x}_i)\right)^2.$$

While classifying new samples, the L_2 norm is used as the distance metric d, as it aligns with the *MSE* optimization criterion:

$$\widehat{y} = \text{argmin}_{1 \leqslant q \leqslant Q} \, ||\mathbf{o} - \mathbf{c}_q||_2.$$

2.5. Class balancing

For problems of imbalanced nature, such as medical diagnosis, where samples of specific classes are a minority compared to the rest, some considerations need to be made in the training process to avoid biases in the final model that can undermine its generalization capabilities.

Different techniques such as class balancing can help with this problem. These generally consist on sampling or generating techniques that balance the ratio of training samples of each class presented to the classifier during training.

A popular approach is SMOTE: in order to generate new samples of a minority class, several existing samples of said class from the training set are randomly selected and a random weighted sum of their features is used as a new synthetic sample of that class. SMOTE can work well when the number of features is not too large and all features are continuous.

In order to account for the ordinal information in the sample synthesis process, OGO techniques based on the idea of SMOTE have been previously proposed, such as OGO via shortest paths using a probability function for the inter-class edges (abbreviated to OGO-SP) (Pérez-Ortiz et al., 2015). This algorithm defines a way to construct a graph which captures the neighbouring relations between samples of the dataset, considering the ordering information provided by the class labels.

2.6. OGO-SP algorithm

Consider a dataset $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, where each $\mathbf{x}_i \in \mathbb{R}^d$ is a training sample and $y_i \in \{\mathscr{C}_1, \mathscr{C}_2, \ldots, \mathscr{C}_Q\}$ its corresponding class label, which follows a ranking structure ($\forall i < j, \; \mathscr{C}_i \prec \mathscr{C}_j$).

From this, an *N*-vertex undirected graph $G = (V, E)$ will be constructed, where V corresponds to the vertices representing the N training samples and $E \subseteq V^2$ corresponds to the edges representing the neighbouring relations between samples:

$$V = \{v_1, v_2, \ldots, v_N\} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\},$$
$$E = \{e_{i,j}\} = \{(v_i, v_j)\} = \{(\mathbf{x}_i, \mathbf{x}_j)\}, \; 1 \leqslant i < j \leqslant N.$$

If q is the index of the class to be over-sampled, a graph G'_q of this form will be constructed based on three subgraphs:

$$G_{q-1,q} = (V_{q-1,q}, E_{q-1,q}),$$
$$G_{q,q} = (V_{q,q}, E_{q,q}),$$
$$G_{q,q+1} = (V_{q,q+1}, E_{q,q+1}),$$
$$G'_q = (V'_q, E'_q) = (V_{q-1,q} \cup V_{q,q} \cup V_{q,q+1}, E_{q-1,q} \cup E_{q,q} \cup E_{q,q+1}).$$

The edges of graph $G_{q-1,q}$ are determined by the intersection of two different sets obtained from a neighbourhood analysis based on the distance relation d:

$$E_{q-1,q} = \mathscr{N}_d(X_{q-1}, X_q, k) \cap \mathscr{N}_d(X_q, X_{q-1}, k),$$
$$\mathscr{N}_d(X_1, X_2, k) = \{e_{i,j} \mid (\mathbf{x}_i \in X_1) \wedge (\mathbf{x}_j \in X_2) \wedge (\mathbf{x}_j \in nn_d(\mathbf{x}_i, X_2, k))\},$$

where X_c is the subset of all samples with label $y = \mathscr{C}_c$, $\mathscr{N}_d(X_1, X_2, k)$ is the k-neighbourhood of X_1 with respect to X_2 according to some distance d and $nn_d(\mathbf{x}, X, k)$ is the set containing the k nearest neighbours of $\mathbf{x}$ from set X. The vertices $V_{q-1,q}$ are all those appearing on $E_{q-1,q}$:

$$V_{q-1,q} = \{\mathbf{x}_i \mid ((\mathbf{x}_i, \mathbf{x}) \in E_{q-1,q}) \vee ((\mathbf{x}, \mathbf{x}_i) \in E_{q-1,q})\}.$$

Using only the intersection of both neighbourhoods ensures that only the connecting regions of each class are considered. Parameter k will control how broad is the region to consider.

Graph $G_{q,q+1}$ is defined in an analogous way:

$$E_{q,q+1} = \mathscr{N}_d(X_q, X_{q+1}, k) \cap \mathscr{N}_d(X_{q+1}, X_q, k),$$
$$V_{q,q+1} = \{\mathbf{x}_i \mid ((\mathbf{x}_i, \mathbf{x}) \in E_{q,q+1}) \vee ((\mathbf{x}, \mathbf{x}_i) \in E_{q,q+1})\}.$$

Finally, G_q is simply defined as:

$$E_{q,q} = \mathscr{N}_d(X_q, X_q, k),$$
$$V_{q,q} = \{\mathbf{x}_i \mid ((\mathbf{x}_i, \mathbf{x}) \in E_{q,q}) \vee ((\mathbf{x}, \mathbf{x}_i) \in E_{q,q})\}.$$

For the case of the extreme classes ($\mathscr{C}_1$ and $\mathscr{C}_Q$), one of $G_{q-1,q}$ or $G_{q,q+1}$ may be empty and only the connecting region to the one adjacent class will be considered.

Based on the ordinal classification hypothesis that the distance to adjacent classes is lower than the distance to non-adjacent classes, the final graph $G_q = (V_q, E_q)$ will be constructed based on the previously constructed G'_q. Ideally, a distance-based manifold exists in the class label, such that X_q lies in the space between X_{q-1} and X_{q+1}. In reality, some outliers may exist in X_q that are not desirable in the over-sampling procedure. In order to identify the key samples which lie between the adjacent classes, the shortest paths between the samples of X_{q-1} and X_{q+1} are used to decide the edges present in the final G_q.

A path between two vertices v_1 and v_z of the graph is defined as the sequence $P = (v_1, v_2, \ldots, v_z) \in V^z$ such that $e_{i,i+1} = (v_i, v_{i+1}) \in E$. If a cost function $f : E \to \mathbb{R}$ assigning a cost to every edge is defined, the shortest path $P_{1,z}$ is that which minimizes the total sum of the costs of the edges $\sum_{i=1}^{z-1} f(e_{i,i+1})$. In our implementation, the cost function selected is the same as distance d used for $\mathscr{N}_d$, which is the L_2 norm or euclidean distance:

$$f(e_{i,j}) = d(\mathbf{x}_i, \mathbf{x}_j) = ||\mathbf{x}_i - \mathbf{x}_j||_2.$$

In order to find those patterns in X_q lying in the latent ordinal manifold, all the shortest paths between all the vertices in $V_{q-1,q}$ and all in $V_{q,q+1}$ will be computed using Dijkstra's algorithm (Dijkstra, 1959), and only the edges contained in one or more of these paths will be included in E_q:

$$E_q = \left\{ e_{i,j} \mid \exists a \in V_{q-1,q}, b \in V_{q,q+1} \right.$$
$$\left. \text{s.t. } (v_i, v_j \in P_{a,b}) \vee (v_j, v_i \in P_{b,a}) \right\}$$

$$V_q = \{v_i \mid e_{i,j} \in E_q\}.$$

Note that, if q is any of the extreme classes, $V_{q,q}$ will have to be used instead of $V_{q-1,q}$ or $V_{q,q+1}$, depending on the case.

An example of the computed graph (V_q, E_q) is shown in Fig. 2.

Finally, new synthetic samples can be generated from G_q: in order to generate sample $(\mathbf{x}', \mathscr{C}_q)$, a random edge $e = (\mathbf{x}_i, \mathbf{x}_j) \in E_q$ is selected so that $\mathbf{x}'$ lies in the segment between $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$\mathbf{x}' = (1 - \delta)\mathbf{x}_i + \delta \mathbf{x}_j,$$

where δ is a random variable in the range $[0, 1]$. The distribution from where δ is sampled will depend on the selected edge e:

- If both $y_i = \mathscr{C}_q$ and $y_j = \mathscr{C}_q$ (i.e. e is an intra-class edge), then δ is sampled from a uniform distribution $U(0, 1)$ just like SMOTE.
- If $y_i = \mathscr{C}_q$ but $y_j \neq \mathscr{C}_q$ (i.e. e is an inter-class edge), then δ is sampled from an asymmetrical distribution so that the new synthetic sample favours the augmented class but is able to capture the class transition phase. In the original OGO-SP paper (Pérez-Ortiz et al., 2015), $\delta \sim \text{Gamma}(k = 2, \theta = 0.15)$. While this has the previously mentioned properties, the gamma distribution is not bounded, $\delta \in [0,$
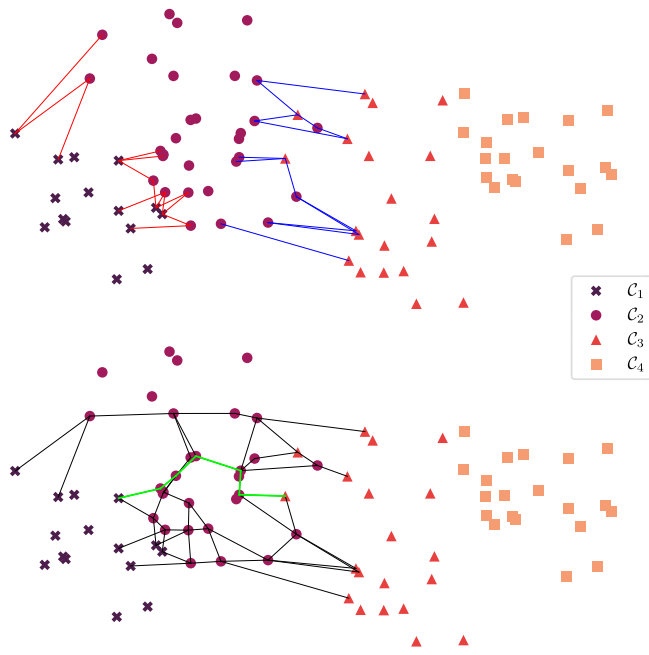
Fig. 2. Example of the OGOSP graph construction procedure. The markers represent samples of the dataset. The graph (V_2, E_2) corresponding to $class_2$ is constructed. The top diagram shows $E_{1,2}$ (in red) and $E_{2,3}$ (in blue). The bottom diagram shows the shortest path between two vertices in $V_{1,2}$ and $V_{2,3}$ (in green) and the edges of the final constructed graph E_2 (in black).

∞), which means that $P(\delta > 1) > 0$, and there is a risk that $\mathbf{x}'$ lies in a different region of the manifold than $\mathscr{C}_q$ and is therefore incorrectly labelled.

In order to overcome this problem of the original algorithm, we propose the following modification: instead of weighting on a parameter $\delta \sim \text{Gamma}(k, \theta)$, the better suited beta distribution is used $\delta \sim \text{Beta}(\alpha, \beta)$, where the parameters $\alpha > 0$ and $\beta > 0$ control the shape of the distribution (Keeping, 1995). We hypothesize that the beta distribution is better suited than the gamma distribution for this purpose, because it is bounded in the interval $[0, 1]$, it has lower variance and its parametrization allows the probability density to be skewed towards a specific extreme, depending on the values of α and β.

The beta distribution has been applied to model the behaviour of random variables limited to intervals of finite lengths in a wide variety of disciplines. This distribution, in its standard form, is a continuous distribution with probability density function $f(x)$ given by:

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

for $0 < x < 1$ and $\alpha > 0, \beta > 0$, where B is the beta function. Depending on the parameter values, the following properties of the distribution can be outlined:

- If $\alpha > 1$, then $f(0) = 0$. Similarly, if $\beta > 1$, then $f(1) = 0$.
- If both $\alpha > 1$ and $\beta > 1$ (see upper right plot in Fig. 3), it has a unique mode at $\frac{\alpha-1}{\alpha+\beta-2}$.
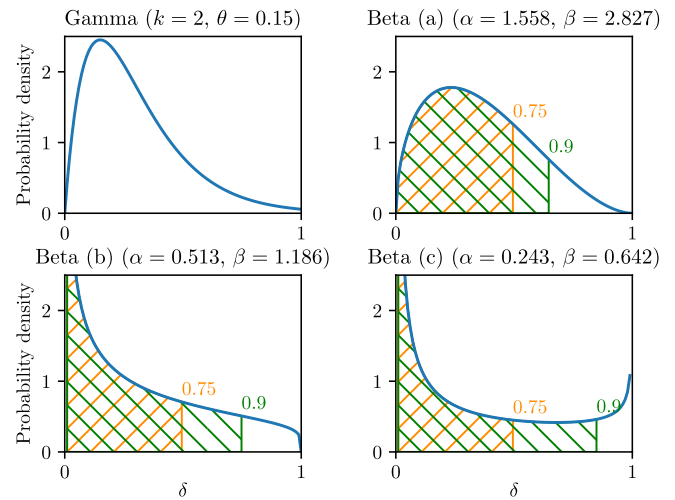

Fig. 3. Shape of the probability density functions for the four different distributions of δ, the original gamma distribution Pérez-Ortiz et al., 2015 and the three proposed configurations of the beta distribution (see Section 2.6).

- If $\alpha = \beta$, it is symmetric. If $\alpha = \beta = 1$ it becomes the uniform distribution.

We refer to this new method from now on as OGO-SP with beta frontier distribution (OGO-SP-β).

Based on the two different possibilities of the two endpoints $f(0)$ and $f(1)$, four different asymmetric shapes can be obtained for this distribution. One of these shapes ($\alpha > 1$ and $\beta < 1$) will not be considered, as this would put more probability mass in the neighbouring class side of the distribution. To ensure this, we use the quantile constraint $P(\delta < 0.5) = 0.75$, so that the majority of the probability mass is in the augmented class side.

Van Dorp and Mazzuchi (2000) prove that two quantile constraints are sufficient to parametrize the beta distribution and provide a numerical method to obtain the values of α and β that satisfy these constraints. We therefore choose three other quantile constraints that, in combination with the previous one, yield the three different shapes, and, using the aforementioned method, we compute the values of α and β:

(a) Beta distribution where $P(\delta < 0.5) = 0.75$ and $P(\delta < 0.65) = 0.9 : \delta \sim \text{Beta}(\alpha = 1.558, \beta = 2.827)$
(b) Beta distribution where $P(\delta < 0.5) = 0.75$ and $P(\delta < 0.75) = 0.9 : \delta \sim \text{Beta}(\alpha = 0.513, \beta = 1.186)$
(4) Beta distribution where $P(\delta < 0.5) = 0.75$ and $P(\delta < 0.85) = 0.9 : \delta \sim \text{Beta}(\alpha = 0.243, \beta = 0.642)$

All three configurations were tested and compared to the original OGO-SP.

Our hypothesis is that the beta distribution will be a better candidate for synthetic sample generation for certain datasets, like the one that will be studied in this paper (Section 2.1). While configuration (a) imitates the original gamma distribution just for comparison, configuration (b) and (c) of OGO-SP-β exploit the versatility of the beta distribution.

A graphical visualization of the shape of the probability density function for all distributions can be seen in Fig. 3. From there, it can be noted that when $\alpha < 1$ the value of the PDF for $\delta = 0$ tends to infinity, as more probability mass is directed to that extreme. The same thing happens for $\beta < 1$ and $\delta = 1$. This way, the probability of generating

samples in the inter-class region can be controlled, while favouring the generation of samples in the augmented class region with respect to the neighbouring class.

The detailed process is finally specified in Algorithm 1.

Algorithm 1: OGO-SP algorithm

Input D: The original dataset
Input q: The class to augment
Input n: How many instances of class q to generate
Input k: The neighbourhood size

1: **begin** OGO-SP D, q, n, k
2:     Construct graph $G_q = (V_q, E_q)$ as described in Section 2.6
3:     $D' \leftarrow \varnothing$
4:     **For** $i = 1$ to n **do**
5:       **repeat**
6:         Select a random $e = (\mathbf{x}_i, \mathbf{x}_j)$ from E_q, with uniform probability
7:       **until** $y_i = \mathscr{C}_q \vee y_j = \mathscr{C}_q$
8:       **if** $y_i = \mathscr{C}_q \wedge y_j = \mathscr{C}_q$ **then**   ▷class-interior edge case
9:         Sample δ from an uniform distribution $U(0, 1)$
10:      **else**   class-frontier edge case
11:        **if** $y_i \neq \mathscr{C}_q$ **then**   ▷Swap so that $\mathbf{x}_i$ is the one with label $\mathscr{C}_q$
12:          $\mathbf{x}_i, \mathbf{x}_j \leftarrow \mathbf{x}_j, \mathbf{x}_i$
13:        **end if**
14:        Sample δ from an asymmetrical distribution (see Section 2.6)
15:       **end if**
16:       $\mathbf{x}' \leftarrow (1 - \delta)\mathbf{x}_i + \delta\mathbf{x}_j$
17:       $D' \leftarrow D' \cup \{\mathbf{x}'\}$
18:     **end for**
19:     **return** D'
20: **end**

In our case, the data described in Section 2.1 is severely imbalanced, so data augmentation is crucial for the performance of the classification model. We argue that the OGO-SP-β algorithm is well suited here due to the ordinal nature of the problem: the alteration of dopaminergic activity is a gradual process, so it is expected to see a more severe damage in a later stage of the disease and vice versa. Moreover, because the intermediate classes are the minority ones, the beta distribution will favour the generation of samples in their regions.

2.7. Application to spatial data

Applying techniques like SMOTE or OGO-SP to spatial data, such as images or 3D scans is not appropriate. They are unable to capture the variability of the position of different objects in a scene or anatomical elements on a CT scan. Applying these techniques in the original space that the images are sampled on yields completely unnatural and inappropriate synthetic samples that detract from the generalization capabilities of the resulting model.

On the other hand, the convolutional part of a CNN model tries to achieve a projection from the original space to a small number of features that can separate the samples correctly according to the classification problem at hand. The space of this projection should be better suited for interpolation and, consequently, for the application of SMOTE and derived techniques. Thus, in this paper, we propose a two-step training process for the application of class balancing:

1. First the whole network (convolutional + fully connected parts) is trained on the original dataset D.
2. Once the stopping criterion is reached, the convolutional part g is used to project the original dataset D into a new space with reduced dimensionality in order to obtain D':

$$D' = \{(g(\mathbf{x}_i), y_i) \mid (\mathbf{x}_i, y_i) \in D\}, g : \mathbb{R}^d \to \mathbb{R}^{d'},$$

where d is the original dimensionality of the data and d' is the new reduced dimensionality.
3. New synthetic samples for each class q are generated by using SMOTE in the nominal case and OGO-SP/OGO-SP-β in the ordinal

case. If n_q is defined as the number of samples labelled as $\mathscr{C}_q$ in D', then D'_{+q} are the generated samples of class q. The number of samples to generate for each class is chosen so as to equalize the number of training samples of all classes. Note that this means that no synthetic sample will be generated for the majority class:

$$\left| D'_{+q} \right| = \left(\max_{1 \leqslant k \leqslant Q} k \right) - n_q.$$

4. Synthetic samples are then merged with dataset D' to produce the augmented dataset D'_+:

$$D'_+ = D' \cup \left(\bigcup_{q=1}^{Q} D'_{+q} \right).$$

5. Finally, only the fully connected part of the original model is trained again using D'_+, with the same stopping criterion.

3. Experimentation

The five models previously described (one nominal and four ordinal for the different distributions for δ) will be compared against each other, in order to evaluate the effect of using the ordinal information in the learning process.

3.1. Experimental design

A stratified 5-fold cross-validation over the complete dataset is performed. The dataset is split into 5 (approximately) equal size subsets in a stratified fashion, so that the class distribution is maintained for each fold. For each step, one subset is used for testing and the rest are used as training samples.

For each of these 5-fold steps, in the first phase, a model selection process is performed, i.e. the hyperparameters of the algorithm are tuned. For this, three 90/10 holdout splits are performed and all the possible combinations of the following parameter values are considered:

- Learning rate (η): {10⁻³, 10⁻⁴}.
- Hidden layer size (H): {2048, 4096}[3].
- Convolution kernel size (k): {3, 5}.
- Neighbourhood size for the data augmentation method: {3, 5}.

The mean *MAE* score across the three splits is used to rank the parameter combinations, and the best combination for each fold is then used for evaluation.

Once the hyperparameter selection phase is completed, the optimal parameter combination is used in the second phase for final evaluation. The model is initialized and trained 30 times with different 90/10 train/validation splits, as well as a different random seed for initialization of the weights of the network and data augmentation.

The same scheme is repeated for each of the data folds. An illustration of this process can be seen in Fig. 4.

In every training instance, the validation samples are used for early stopping of the training process: when the loss over the validation set does not improve for 50 epochs ('patience' parameter), training is stopped, and the best performing weights are restored.

A batch size of 32 samples is used in all cases.

The code used to perform the experiments can be accessed through

---

[3] Note that in the case of the ordinal model, each output uses a $(Q-1)$th of this number of nodes for each binary output, so the number of parameters between models is comparable
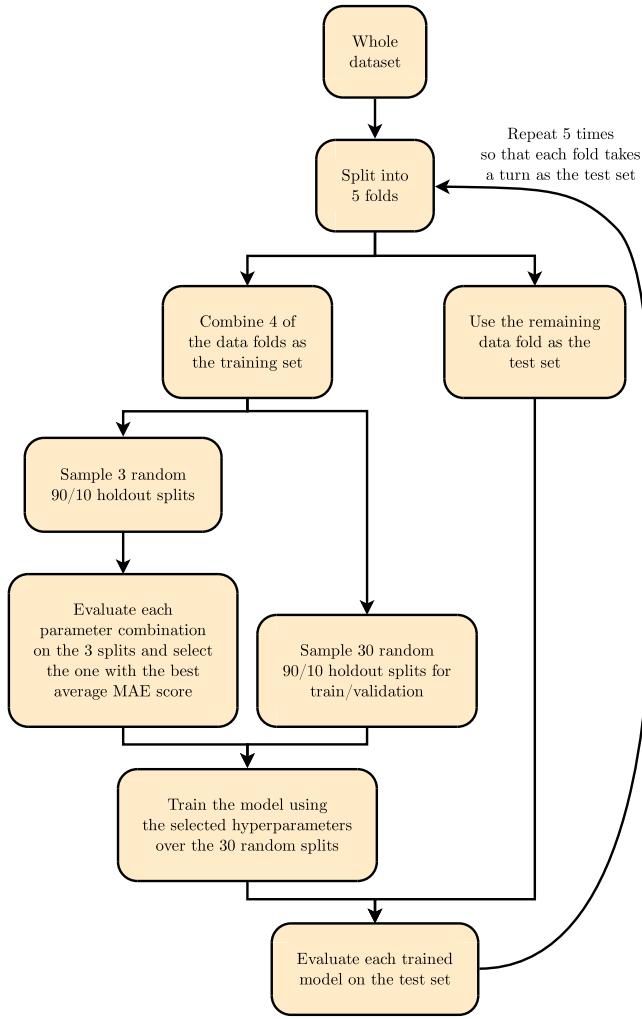
Fig. 4. The cross-validation scheme used for the validation of hyperparameters and evaluation of the models.

GitHub.[4]

3.2. Evaluation metrics

Multi-class nominal and ordinal metrics.

While the Correct Classification Rate (*CCR*) is usually the most important criterion in classification tasks, in the case of high class imbalance, it is less relevant (Provost & Fawcett, 1997). In scenarios like this, a model that disregards the minority classes can still obtain a high *CCR* score, which is not desirable, and per-class sensitivity needs to be addressed (Sánchez-Monedero, Gutiérrez, Fernández-Navarro, & Hervás-Martínez, 2011). Still, while optimizing sensitivity, specificity may suffer a performance hit, so it also needs to be monitored.

In addition to this, *CCR* does not consider how much each prediction deviates from the ground truth, as it is designed primarily for nominal classification problems (where all the mistakes are equally penalised). For ordinal classification problems, a classification error of only 1 class is more desirable over an error of 2 classes. For this reason, rank difference metrics like the Mean Absolute Error (*MAE*), Spearman's rank correlation coefficient (r_s), Kendall's rank correlation coefficient (τ_b) (Cardoso & Sousa, 2011) and Weighted Cohen's Kappa (κ) (Ben-David, 2008) are better suited to evaluate the performance of the model.

---
[4] https://github.com/ayrna/ordinal-cnn-parkinsons.

Moreover, *MAE* may also be deceiving in high class imbalance scenarios, so per-class *MAE* is also useful to consider.

When monitoring per-class metrics, it is useful to look at the minimum in order to ensure that performance does not increase at the expense of ignoring some of the classes.

In this study, the following metrics will be used. Metrics to be maximized are marked with (↑) and metrics to be minimized with (↓). In the following equations, N denotes the number of test samples, Q denotes the number of classes, y_i is the class label for sample $\mathbf{x}_i$ and $\widehat{y}_i$ is the estimated label for sample $\mathbf{x}_i$.

- *CCR* (↑): $CCR = \frac{1}{N}\sum_{i=1}^{N} 1\{\widehat{y}_i = y_i\}$.
- Geometric mean of the sensitivities and minimum sensitivity (↑):

$$GMS = \sqrt[Q]{\prod_{q=1}^{Q} S_q},$$
$$MS = \min_{1 \leqslant q \leqslant Q} S_q,$$

where per-class sensitivity (S_q) is:

$$S_q = \frac{\sum_{i=1}^{N} 1\{y_i = \widehat{y}_i = q\}}{\sum_{i=1}^{N} 1\{y_i = q\}}.$$

- Geometric mean of the specificities and minimum specificity (↑):

$$GMSp = \sqrt[Q]{\prod_{q=1}^{Q} Sp_q},$$
$$MSp = \min_{1 \leqslant q \leqslant Q} p_q,$$

where per-class specificity (Sp_q) is:

$$Sp_q = \frac{\sum_{i=1}^{N} 1\{y_i \neq q \wedge \widehat{y}_i \neq q\}}{\sum_{i=1}^{N} 1\{y_i \neq q\}}.$$

- Mean Absolute Error (*MAE*) (↓): $MAE = \frac{1}{N}\sum_{i=1}^{N} |\widehat{y}_i - y_i|$.
- Average and maximum *MAE* (↓):

$$AMAE = \frac{1}{Q} \sum_{q=1}^{Q} MAE_q,$$
$$MMAE = \max_{1 \leqslant q \leqslant Q} MAE_q,$$

where per-class *MAE* (MAE_q) is:

$$MAE_q = \frac{\sum_{i=1}^{N} 1\{y_i = q\}|\widehat{y}_i - y_i|}{\sum_{i=1}^{N} 1\{y_i = q\}}.$$

- Weighted Cohen's Kappa (↑): $\kappa = 1 - \frac{\sum_{i=1}^{Q}\sum_{j=1}^{Q} w_{ij}p_{ij}}{\sum_{i=1}^{Q}\sum_{j=1}^{Q} w_{ij}e_{ij}}$. where w_{ij} is the disagreement cost when $y = \mathscr{C}_i$ and $\widehat{y} = \mathscr{C}_j$ ($w_{ij} = |i - j|$), p_{ij} is the observed agreement and e_{ij} is the expected agreement due to chance.
- Kendall rank correlation coefficient (↑): $\tau_b = \frac{n_c - n_d}{\sqrt{(n_c+n_d+n_1)(n_c+n_d+n_2)}}$, where n_c, n_d, n_1 and n_2 are computed from every pair $\{(y_i, \widehat{y}_i), (y_j, \widehat{y}_j)\}$. n_c is the number of concordant pairs, n_d is the number of discordant

pairs, and n_1 and n_2 is the number of tied pairs only for y and only for $\widehat{y}$, respectively:

$$n_c = \sum_{i=1}^{N}\sum_{j=i+1}^{N} 1\left\{ \left[\left(y_i < y_j\right) \wedge \left(\widehat{y}_i < \widehat{y}_j\right) \right] \vee \left[\left(y_i > y_j\right) \wedge \left(\widehat{y}_i > \widehat{y}_j\right) \right] \right\},$$

$$n_d = \sum_{i=1}^{N}\sum_{j=i+1}^{N} 1\left\{ \left[\left(y_i < y_j\right) \wedge \left(\widehat{y}_i > \widehat{y}_j\right) \right] \vee \left[\left(y_i > y_j\right) \wedge \left(\widehat{y}_i < \widehat{y}_j\right) \right] \right\},$$

$$n_1 = \sum_{i=1}^{N}\sum_{j=i+1}^{N} 1\left\{ \left(y_i = y_j\right) \wedge \left(\widehat{y}_i \neq \widehat{y}_j\right) \right\},$$

$$n_2 = \sum_{i=1}^{N}\sum_{j=i+1}^{N} 1\left\{ \left(y_i \neq y_j\right) \wedge \left(\widehat{y}_i = \widehat{y}_j\right) \right\}.$$

- Spearman rank correlation coefficient (↑): $r_s = \frac{Cov(y,\widehat{y})}{\sigma_y \sigma_{\widehat{y}}}$, where $Cov(y, \widehat{y})$ is the covariance between the ground truth labels and the predicted labels and σ_y and $\sigma_{\widehat{y}}$ is their standard deviation.

Of the previous metrics, *CCR*, *MS*, *MAE*, *AMAE*, *MMAE*, κ, τ_b and r_s are completely defined by Cruz-Ramírez, Hervás-Martínez, Sánchez-Monedero, and Gutiérrez (2014) and *GMS* by Pérez-Ortiz et al. (2015). *CCR*, *GMS*, *MS*, *GMSp* and *MSp* are all purely nominal metrics, while the rest apply only to ordinal class labels and are generally more relevant for this kind of classification problems.

Also, *GMS*, *MS*, *MSp*, *GMSp*, *AMAE* and *MMAE* were chosen as class imbalance-sensitive metrics, to asses the effectiveness in these scenarios.

Lastly, based on the output scores of the models, the area under the Receiver Operating Characteristic (ROC) curve (*AUC*) will be computed. In order to obtain this value, as many curves as the number of classes are obtained, where each curve is based on a binary one-vs-rest labelling (OvR). Then, the average *AUC* is obtained from the Q curves.

Binary metrics. Given that no previous literature exists on the diagnosis of the stage of presynaptic damage from PD, some binary metrics will be extracted from the results with the only purpose of comparing it to previous works, which deal with the diagnosis problem of discerning PD patients from Healthy Controls (HC). Directly comparing *CCR* or any other of the previously mentioned metrics against the case of a binary classifier would be incorrect and unfair, as the ordinal metric is quite more demanding, having to classify into four different classes instead of only two.

For this purpose, all class 0 labels will be considered as 'negative class' or HC and the rest (class 1 through 3) will be considered as 'positive class' or PD. From this, a confusion matrix can be extracted. Using the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) the following metrics can be obtained:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}.$$

3.3. Experimental results

Table 1 includes a summary of all the results from the 150 executions of the different methodologies, based on the different performance metrics introduced in the previous subsection. A graphical summary of the results can be found in Fig. 5.

From these results, it can be noted that some configuration of OGO-SP-β always shows better average performance over all other classifiers in all metrics, with the exception of *CCR* for which the nominal

Table 1
Summary of evaluation results. Best mean results are highlighted in bold.

	CCR (↑)		GMS (↑)		MS (↑)	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Nominal	**0.7448**	0.0412	0.1927	0.2015	0.0679	0.0774
OGO-SP	0.7121	0.0422	0.4256	0.1847	0.2426	0.1611
OGO-SP-β (a)	0.7048	0.0487	0.4134	0.1891	0.2297	0.1575
OGO-SP-β (b)	0.7108	0.0644	0.4239	0.1734	0.2235	0.1475
OGO-SP-β (c)	0.7255	0.0423	**0.4403**	0.1830	**0.2460**	0.1451

	MSp (↑)		GMS (↑)		AUC (↑)	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Nominal	0.7691	0.0569	0.8898	0.0179	0.8486	0.0344
OGO-SP	0.8276	0.0407	0.8979	0.0138	0.8588	0.0345
OGO-SP-β (a)	0.8223	0.0406	0.8957	0.0146	0.8553	0.0385
OGO-SP-β (b)	0.8294	0.0601	0.8977	0.0229	0.8567	0.0331
OGO-SP-β (c)	**0.8312**	0.0512	**0.9009**	0.0170	**0.8596**	0.0366

	MAE (↓)		AMAE (↓)		MMAE (↓)	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Nominal	0.3826	0.0737	0.6803	0.1127	1.1427	0.1328
OGO-SP	0.3738	0.0586	0.5671	0.1131	0.9043	0.2148
OGO-SP-β (a)	0.3791	0.0643	0.5668	0.1070	**0.9021**	0.2005
OGO-SP-β (b)	0.3729	0.0751	0.5631	0.1038	0.9187	0.1800
OGO-SP-β (c)	**0.3639**	0.0649	**0.5594**	0.1186	0.9113	0.1896

	τ_b (↑)		κ (↑)		r_s (↑)	
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Nominal	0.7119	0.0565	0.6816	0.0645	0.7702	0.0600
OGO-SP	0.7323	0.0442	0.6822	0.0548	0.7980	0.0516
OGO-SP-β (a)	0.7282	0.0496	0.6776	0.0580	0.7968	0.0527
OGO-SP-β(b)	0.7343	0.0467	0.6844	0.0654	**0.8011**	0.0512
OGO-SP-β (c)	**0.7389**	0.0481	**0.6898**	0.0596	0.7986	0.0548

methodology obtains the best results, with OGO-SP-β (c) in second place. By looking at the *GMS* and *MS* metrics, it is very clear that the nominal methodology fails to address the present class imbalance and ignores the minority classes, arranging all test samples into a couple of the majority classes, which drops the score to zero in most of the evaluation splits.

Comparing OGO-SP-β to the original OGO-SP, OGO-SP-β generally obtains better average performance over all splits, specially according to *MAE*.

Both ordinal methodologies clearly outperform the nominal one. Fig. 6 shows the corresponding ROCs for the different classes of the problem. Both ordinal methodologies obtain a noticeable advantage in the intermediate minority classes, specially for class 1, the class with the least number of samples.

We used a Wilcoxon signed-rank test to detect significant differences between the metrics. This version of the test was used to measure the performance of the classifiers, because the variables come from the same sample (Wilcoxon, 1945). Purely nominal methodology was compared with the overall best performing ordinal methodology using OGO-SP-β (configuration (c)). We also used the test to compare the ordinal methodology using OGO-SP with the best performing ordinal methodology using OGO-SP-β (configuration (c)).

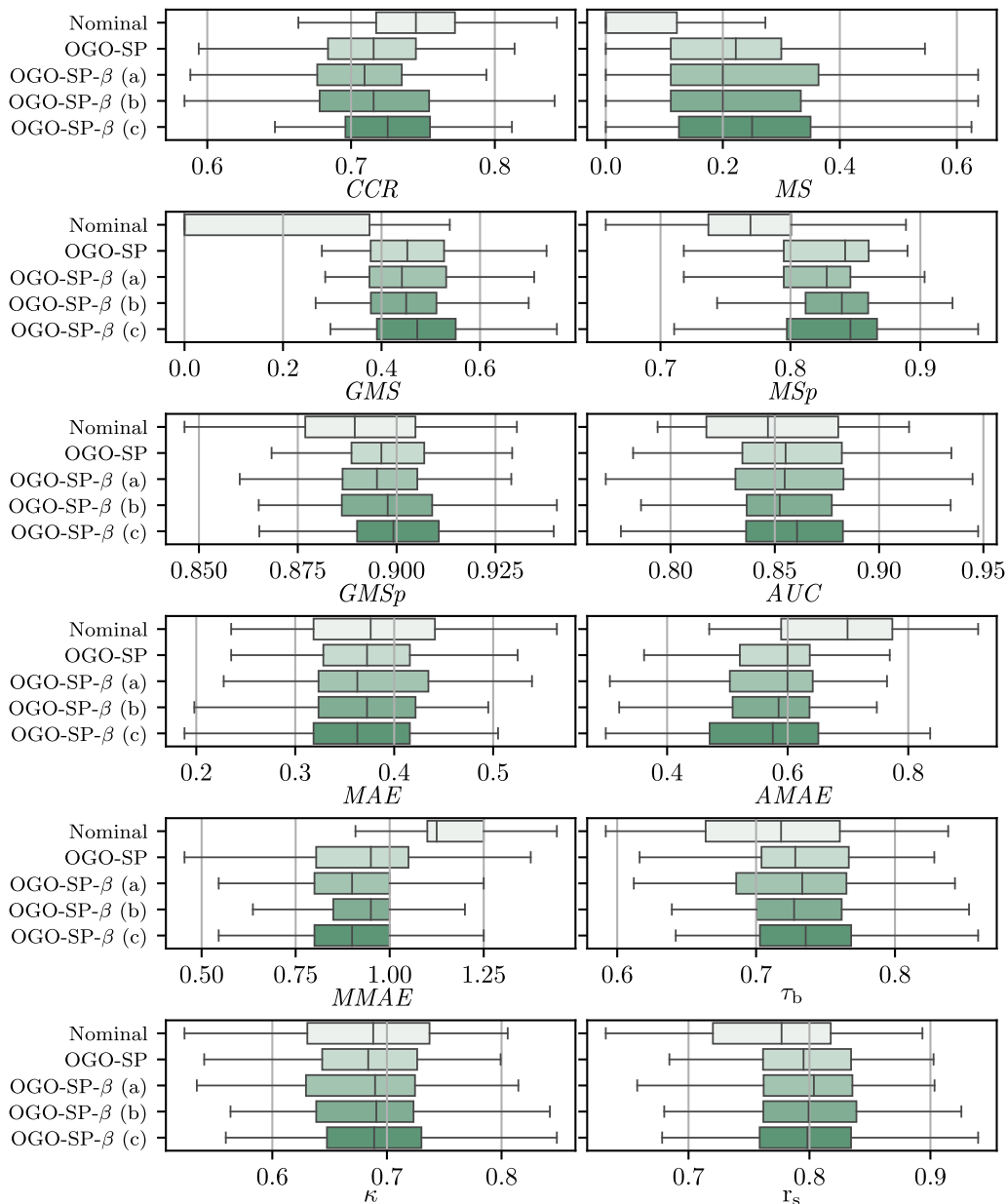The critical values (*p*-values), for a bilateral test of each metric, are

Fig. 5. Graphical summary of the experimentation results as boxplots.

provided in Table 2.

Significant differences in performance (with $\alpha = 0.05$) favouring the ordinal methodology over the nominal one are found for all metrics, but *CCR* and κ. As expected, the nominal methodology fixates on improving *CCR* instead of the rest of ordinal metrics, which are generally more desirable for ordinal type classification problems. Moreover, the nominal methodology tends to ignore minority classes, obtaining significantly worse results for *GMS*, *MS*, *MSp*, *GMSp* and *AUC*. Furthermore, OGO-SP-β performs significantly better in *CCR*, *MAE* and *GMSp* compared to the original OGO-SP using the gamma distribution.

Clearly, even when the tests do not find significant differences, OGO-SP-β shows better overall performance in the ordinal metrics. The inconclusiveness present in the imbalance-sensitive metrics (*GMS*, *MS*, *MSp* and *MMAE*) could be explained by their unstable nature, which results in a larger standard deviation, and thus it is more difficult for the test to draw a conclusion.

Lastly, the results for the binary metrics results can be found in Table 3. The results are compared against the following works:

- Rizzo et al. (2016): A meta-analysis of 20 different studies, all using different techniques, between 1988 and 2014.
- de la Fuente-Fernández (2012): An aggregation of 2 different studies both using SPECT imaging.
- Martinez-Murcia et al. (2017): A CNN approach for binary classification from SPECT imaging.
- Orozco-Arroyave et al. (2016): An application of radial base Support Vector Machines to running speech audio samples from patients.
- El Maachi, Bilodeau, and Bouachir (2020): An application of neural networks to gait sensor data from patients.

Note that the task considered in these works is different, as the different possibilities for positive labels (PD) are not differentiated, greatly reducing the complexity. However, even considering that the proposed models are more informative and taking into account that the experimental settings and the datasets are not same, we can conclude that the performance obtained by the different proposals is competitive, specially when trying to achieve a balance between the three binary
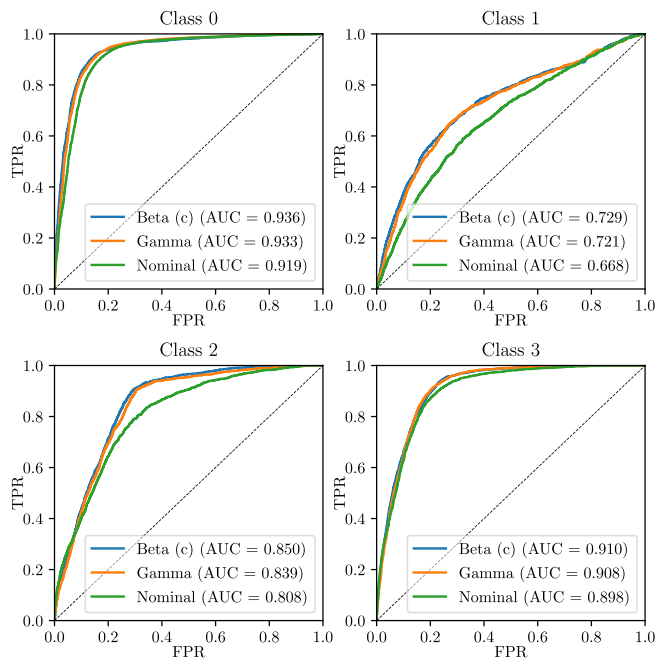
Fig. 6. ROC obtained for each of the four classes by three of the evaluated methodologies. The curves are obtained according to the output scores of each model using a one-vs-rest labelling (OVR) for each class.

Table 2

p-values for the two-tailed Wilcoxon signed rank test. N stands for the nominal methodology, γ stands for the methodology using the original OGO-SP algorithm and β stands for the methodology using the OGO-SP-β algorithm. Values less than $\alpha = 0.05$ have been highlighted in bold. For these tests, the (+) subscript denotes an advantage for OGO-SP-β and (−) otherwise.

	CCR	GMS	MS
N vs. β (c)	3.5×10^{-5} (−)	9.6×10^{-18} (+)	6.4×10^{-19} (+)
γ vs. β (c)	9.5×10^{-4} (+)	2.7×10^{-1}	7.6×10^{-1}
	MSp	GMSp	AUC
N vs. β (c)	3.0×10^{-17} (+)	2.4×10^{-12} (+)	7.7×10^{-6} (+)
γ vs. β (c)	2.5×10^{-1}	1.9×10^{-2} (+)	3.9×10^{-1}
	MAE	AMAE	MMAE
N vs. β (c)	7.5×10^{-4} (+)	5.1×10^{-23} (+)	1.3×10^{-18} (+)
γ vs. β (c)	3.8×10^{-2} (+)	1.7×10^{-1}	7.3×10^{-1}
	τ_b	κ	r_s
N vs. β (c)	9.2×10^{-9} (+)	9.3×10^{-2}	6.8×10^{-10} (+)
γ vs. β (c)	7.3×10^{-2}	8.0×10^{-2}	6.7×10^{-1}

metrics. That is, the extra information provided by the proposed multiclass classifiers is not obtained at the cost of losing performance in the binary task.

4. Conclusions

We have confirmed experimentally that the exploitation of ordinal information can improve the performance of a complex task such as the assessment of brain activity alteration in PD. This exploitation comes from aspects such as the model architecture, the optimization target and the data augmentation strategy. This expands on the current range of models capable of exploiting ordinal information, in this case, a fully 3D CNN.

This methodology is able to alleviate the class imbalance problem while improving ordinal performance metrics. This approach could be

Table 3

Binary metrics results of the five proposed methodologies along with three other works. Rows marked as type 'A' correspond to an automatic method while rows marked as type H correspond to studies about diagnosis by human experts. The Data column indicates the reference data used for diagnosis: imaging (I), speech (S), gait (G) or a refined clinical diagnosis (C).

Method/work	Type	Data	Accuracy	Sensitivity	Specificity
Nominal	A	I	87.62%	77.52%	93.84%
OGO-SP	A	I	87.30%	86.74%	87.64%
OGO-SP-β (a)	A	I	86.94%	86.67%	87.09%
OGO-SP-β (b)	A	I	87.14%	87.10%	87.16%
OGO-SP-β (c)	A	I	88.19%	86.21%	89.38%
Rizzo et al. (2016)[a]	H	C	86.4%	86.6%	84.7%
de la Fuente-Fernández (2012)[b]	H	I	84%-98%	98%	67%-94%
Martinez-Murcia et al. (2017)	A	I	95.5%	96.2%	–
Orozco-Arroyave et al. (2016)[c]	A	S	85%-97%	76.7%-98%	93.3%-96%
El Maachi et al. (2020)	A	G	98.7%	98.1%	100%

[a] Values are taken from the refined clinical diagnosis by experts.

[b] Ranges represent the performance between early and established diagnosis, when provided.

[c] Ranges represent the worst and best performance across all three tested languages.

applied to already existing ordinal classification tasks which suffer from this same problem, which is common among the medical field.

Furthermore, the proposed OGO-SP-β ordinal augmentation algorithm improves performance of ordinal and nominal metrics compared to the original OGO-SP.

From the more classical binary diagnosis point of view, a good performance is achieved, as can be noted in the comparison against expert diagnoses and other ML techniques.

Future work lines include further data acquisition or the application of this methodology to publicly available data. Current available data commonly lacks ordinal information, but ordinal target labels may be extracted from already available information. Larger datasets could help obtain more relevant and precise results. Also, 3D ordinal applications outside the medical can be explored to asses the performance over nominal approaches.

CRediT authorship contribution statement

Javier Barbero-Gómez: Writing - original draft, Conceptualization, Methodology, Software, Validation, Investigation. **Pedro-Antonio Gutiérrez:** Writing - review & editing, Supervision. **Víctor-Manuel Vargas:** Writing - review & editing, Software. **Juan-Antonio Vallejo-Casas:** Data curation. **César Hervás-Martínez:** Formal analysis, Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Akyol, K. (2020). Stacking ensemble based deep neural networks modeling for effective epileptic seizure detection. *Expert Systems with Applications, 148*, Article 113239. https://doi.org/10.1016/j.eswa.2020.113239

Allwein, E. L., Schapire, R. E., & Singer, Y. (2001). Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research, 1*, 113–141. https://doi.org/10.1162/15324430152733133

Arbizu, J., Luquin, M., Abella, J., de la Fuente-Fernández, R., Fernandez-Torrón, R., García-Solís, D., Garrastachu, P., Jiménez-Hoyuela, J., Llaneza, M., Lome na, F.,

Lorenzo-Bosquet, C., Martí, M., Martinez-Castrillo, J., Mir, P., Mitjavila, M., Ruiz-Martínez, J., & Vela, L. (2014). Functional neuroimaging in the diagnosis of patients with parkinsonism: Update and recommendations for clinical use. Rev. Esp. Med. Nucl. E Imagen Mol. (English Edition), 33, 215–226. https://doi.org/10.1016/j. remnie.2014.05.002.

Ben-David, A. (2008). Comparison of classification accuracy using Cohen's Weighted Kappa. *Expert Systems with Applications, 34*, 825–832. https://doi.org/10.1016/j. eswa.2006.10.022

Cardoso, J. S., & Sousa, R. (2011). Measuring the performance of ordinal classification. *International Journal of Pattern Recognition and Artificial Intelligence, 25*, 1173–1195. https://doi.org/10.1142/S0218001411009093

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953. arXiv:1106.1813

Cruz-Ramírez, M., Hervás-Martínez, C., Sánchez-Monedero, J., & Gutiérrez, P. A. (2014). Metrics to guide a multi-objective evolutionary algorithm for ordinal classification. *Neurocomputing, 135*, 21–31. https://doi.org/10.1016/j.neucom.2013.05.058

Darcourt, J., Booij, J., Tatsch, K., Varrone, A., Vander Borght, T., Kapucu, O. L., Någren, K., Nobili, F., Walker, Z., & Van Laere, K. (2010). EANM procedure guidelines for brain neurotransmission SPECT using (123)I-labelled dopamine transporter ligands, version 2. *European Journal of Nuclear Medicine and Molecular Imaging, 37*, 443–450. https://doi.org/10.1007/s00259-009-1267-x

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik, 1*, 269–271. https://doi.org/10.1007/BF01386390

Dorado-Moreno, M., Pérez-Ortiz, M., Gutiérrez, P.A., Ciria, R., Brice no, J., & Hervás-Martínez, C. (2017). Dynamically weighted evolutionary ordinal neural network for solving an imbalanced liver transplantation problem. Artificial Intelligence in Medicine, 77, 1–11. 10.1016/j.artmed.2017.02.004.

El-Dahshan, E.-S. A., Mohsen, H. M., Revett, K., & Salem, A.-B. M. (2014). Computer-aided diagnosis of human brain tumor through MRI: A survey and a new algorithm. *Expert Systems with Applications, 41*, 5526–5545. https://doi.org/10.1016/j. eswa.2014.01.021

El Maachi, I., Bilodeau, G.-A., & Bouachir, W. (2020). Deep 1D-Convnet for accurate Parkinson disease detection and severity prediction from gait. *Expert Systems with Applications, 143*, Article 113075. https://doi.org/10.1016/j.eswa.2019.113075

Evans, A. C., Janke, A. L., Collins, D. L., & Baillet, S. (2012). Brain templates and atlases. *NeuroImage, 62*, 911–922. https://doi.org/10.1016/j.neuroimage.2012.01.024

Findley, L. J. (2007). The economic impact of Parkinson's disease. *Parkinsonism & Related Disorders, 13*, S8–S12. https://doi.org/10.1016/j.parkreldis.2007.06.003

Frank, E., & Hall, M. (2001). In *A Simple Approach to Ordinal Classification In European Conference on Machine Learning* (pp. 145–156). Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-44795-4_13

de la Fuente-Fernández, R. (2012). Role of DaTSCAN and clinical diagnosis in Parkinson disease. *Neurology, 78*, 696–701. https://doi.org/10.1212/WNL.0b013e318248e520

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. arXiv:1406.2661.

Gutiérrez, P. A., Pérez-Ortiz, M., Sánchez-Monedero, J., Fernández-Navarro, F., & Hervás-Martínez, C. (2016). Ordinal Regression Methods: Survey and Experimental Study. *IEEE Transactions on Knowledge and Data Engineering, 28*, 127–146. https://doi.org/10.1109/TKDE.2015.2457911

Hinton, G. E., & Zemel, R. S. (1993). Autoencoders, minimum description length and Helmholtz free energy. In *6th International Conference on Neural Information Processing Systems NIPS'93* (pp. 3–10).

Hoehn, M. M., & Yahr, M. D. (1967). Parkinsonism: Onset, progression and mortality. *Neurology, 17*, 427–442. https://doi.org/10.1212/wnl.17.5.427

Cheng, Jianlin, Wang, Zheng, & Pollastri, G. (2008). A neural network approach to ordinal regression. In *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (pp. 1279–1284). https://doi.org/10.1109/IJCNN.2008.4633963

Keeping, E. S. (1995). *The Beta Distribution. In Introduction to Statistical Inference* (pp. 83–85). Dover Publications.

Kingma, D.P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Kotsiantis, S. B., & Pintelas, P. E. (2004). A Cost Sensitive Technique for Ordinal Classification Problems. In *Methods and Applications of Artificial Intelligence Lecture Notes in Computer Science* (pp. 220–229). https://doi.org/10.1007/978-3-540-24674-9_24

Kramer, S., Widmer, G., Pfahringer, B., & de Groeve, M. (2010). Prediction of Ordinal Classes Using Regression Trees. In *Foundations of Intelligent Systems Lecture Notes in Computer Science* (pp. 426–434). Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-39963-1_45.

Marino, S., Ciurleo, R., Di Lorenzo, G., Barresi, M., De Salvo, S., Giacoppo, S., Bramanti, A., Lanzafame, P., & Bramanti, P. (2012). Magnetic resonance imaging markers for early diagnosis of Parkinson's disease. *Neural Regeneration Research, 7*, 611–619. https://doi.org/10.3969/j.issn.1673-5374.2012.08.009

Martinez-Murcia, F. J., Ortiz, A., Górriz, J. M., Ramírez, J., Segovia, F., Salas-Gonzalez, D., Castillo-Barnes, D., & Illán, I. A. (2017). A 3D Convolutional Neural Network Approach for the Diagnosis of Parkinson's Disease. In *Natural and Artificial Computation for Biomedicine and Neuroscience Lecture Notes in Computer Science* (pp.

324–333). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-59740-9_32.

Mazaheri, V., & Khodadadi, H. (2020). Heart arrhythmia diagnosis based on the combination of morphological, frequency and nonlinear features of ECG signals and metaheuristic feature selection algorithm. *Expert Systems with Applications, 161*, Article 113697. https://doi.org/10.1016/j.eswa.2020.113697

McCullagh, P. (1980). Regression Models for Ordinal Data. J.R. *Journal of the Royal Statistical Society: Series B, 42*, 109–142. https://doi.org/10.1111/j.2517-6161.1980. tb01109.x

Orozco-Arroyave, J.R., Hönig, F., Arias-Londo no, J.D., Vargas-Bonilla, J.F., Daqrouq, K., Skodda, S., Rusz, J., & Nöth, E. (2016). Automatic detection of Parkinson's disease in running speech spoken in three different languages. Journal of the Acoustical Society of America 139, 481–500. https://doi.org/10.1121/1.4939739.

Payer, C., Štern, D., Bischof, H., & Urschler, M. (2019). Integrating spatial configuration into heatmap regression based CNNs for landmark localization. *Medical Image Analysis, 54*, 207–219. https://doi.org/10.1016/j.media.2019.03.007

Peng, D., Xiong, S., Peng, W., & Lu, J. (2020). LCP-Net: A local context-perception deep neural network for medical image segmentation. *Expert Systems with Applications*, 114234. https://doi.org/10.1016/j.eswa.2020.114234

Perez, L., & Wang, J. (2017). The Effectiveness of Data Augmentation in Image Classification using Deep Learning. arXiv:1712.04621.

Pérez-Ortiz, M., Gutiérrez, P. A., Hervás-Martínez, C., & Yao, X. (2015). Graph-Based Approaches for Over-Sampling in the Context of Ordinal Regression. *IEEE Transactions on Knowledge and Data Engineering, 27*, 1233–1245. https://doi.org/10.1109/TKDE.2014.2365780

Provost, F., & Fawcett, T. (1997). Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions. In *Proceedings of the 3rd International Conference on Knowledgediscovery and Data Mining* (pp. 43–48).

Ramaker, C., Marinus, J., Stiggelbout, A. M., & Van Hilten, B. J. (2002). Systematic evaluation of rating scales for impairment and disability in Parkinson's disease. *Movement Disorders, 17*, 867–876. https://doi.org/10.1002/mds.10248

Rey, A., Arcay, B., & Castro, A. (2020). A hybrid CAD system for lung nodule detection using CT studies based on soft computing. *Expert Systems with Applications*, 114259. https://doi.org/10.1016/j.eswa.2020.114259

Rivera, W. A., & Xanthopoulos, P. (2016). A priori synthetic over-sampling methods for increasing classification sensitivity in imbalanced data sets. *Expert Systems with Applications, 66*, 124–135. https://doi.org/10.1016/j.eswa.2016.09.010

Rizzo, G., Copetti, M., Arcuti, S., Martino, D., Fontana, A., & Logroscino, G. (2016). Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis. *Neurology, 86*, 566–576. https://doi.org/10.1212/WNL.0000000000002350

Salazar, A., Vergara, L., & Safont, G. (2021). Generative Adversarial Networks and Markov Random Fields for oversampling very small training sets. *Expert Systems with Applications, 163*, Article 113819. https://doi.org/10.1016/j.eswa.2020.113819

Sánchez-Monedero, J., Gutiérrez, P. A., Fernández-Navarro, F., & Hervás-Martínez, C. (2011). Weighting Efficient Accuracy and Minimum Sensitivity for Evolving Multi-Class Classifiers. *Neural Process Lett, 34*, 101. https://doi.org/10.1007/s11063-011-9186-9

Sánchez-Monedero, J., Pérez-Ortiz, M., Sáez, A., Gutiérrez, P. A., & Hervás-Martínez, C. (2018). Partial order label decomposition approaches for melanoma diagnosis. *Applied Soft Computing, 64*, 341–355. https://doi.org/10.1016/j.asoc.2017.11.042

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage, 23*, S208–219. https://doi.org/10.1016/j. neuroimage.2004.07.051

Suzuki, T. (2018). Fast generalization error bound of deep learning from a kernel perspective. In *In International Conference on Artificial Intelligence and Statistics* (pp. 1397–1406).

Tomlinson, C. L., Stowe, R., Patel, S., Rick, C., Gray, R., & Clarke, C. E. (2010). Systematic review of levodopa dose equivalency reporting in Parkinson's disease. *Movement Disorders, 25*, 2649–2653. https://doi.org/10.1002/mds.23429

Van Dorp, J. R., & Mazzuchi, T. A. (2000). Solving for the parameters of a beta distribution under two quantile constraints. *Journal of Statistical Computation and Simulation, 67*, 189–201. https://doi.org/10.1080/00949650008812041

dos Santos, Vinícius, Ferreira, M., de Carvalho, Oseas, Filho, A., Dalília de Sousa, A., Corrêa Silva, A., & Gattass, M. (2018). Convolutional neural network and texture descriptor-based automatic detection and diagnosis of glaucoma. *Expert Systems with Applications, 110*, 250–263. https://doi.org/10.1016/j.eswa.2018.06.010

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin, 1*, 80–83. https://doi.org/10.2307/3001968

Williams, R. (2006). Generalized Ordered Logit/Partial Proportional Odds Models for Ordinal Dependent Variables. *Stata Journal, 6*, 58–82. https://doi.org/10.1177/1536867X0600600104

Wu, H., Lu, H., & Ma, S. (2003). A practical SVM-based algorithm for ordinal regression in image retrieval. In *11th ACM International Conference on Multimedia MULTIMEDIA '03* (pp. 612–621). Association for Computing Machinery. https://doi.org/10.1145/957013.957144.