ORIGINAL ARTICLE

Parkinson's disease diagnosis using convolutional neural networks and figure-copying tasks

Mohamad Alissa[1] · Michael A. Lones[2] · Jeremy Cosgrove[3] · Jane E. Alty[4] · Stuart Jamieson[3] · Stephen L. Smith[5] · Marta Vallejo[6]

Abstract

Parkinson's disease (PD) is a progressive neurodegenerative disorder that causes abnormal movements and an array of other symptoms. An accurate PD diagnosis can be a challenging task as the signs and symptoms, particularly at an early stage, can be similar to other medical conditions or the physiological changes of normal ageing. This work aims to contribute to the PD diagnosis process by using a convolutional neural network, a type of deep neural network architecture, to differentiate between healthy controls and PD patients. Our approach focuses on discovering deviations in patient's movements with the use of drawing tasks. In addition, this work explores which of two drawing tasks, wire cube or spiral pentagon, are more effective in the discrimination process. With 93.5% accuracy, our convolutional classifier, trained with images of the pentagon drawing task and augmentation techniques, can be used as an objective method to discriminate PD from healthy controls. Our compact model has the potential to be developed into an offline real-time automated single-task diagnostic tool, which can be easily deployed within a clinical setting.

Keywords Convolutional neural networks · Parkinson's disease · Drawing tasks · Deep learning classifier · Diagnosis

1 Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder characterised histologically by the death of dopaminergic neurons in the substantia nigra pars compacta (SNpc) and the presence of Lewy bodies in various parts of the brain [17]. The SNpc is a compact structure in the midbrain that plays a vital role in motor coordination and movement control by producing a chemical substance called dopamine, which is integral for controlling the initiation, velocity, and fluidity of voluntary movement sequences [83]. The causes of most cases of PD (known as

✉ Marta Vallejo
   m.vallejo@hw.ac.uk

   Mohamad Alissa
   m.alissa@napier.ac.uk

   Michael A. Lones
   m.lones@hw.ac.uk

   Jeremy Cosgrove
   jezcos@doctors.net.uk

   Jane E. Alty
   jane.alty@utas.edu.au

   Stuart Jamieson
   stuart.jamieson1@nhs.net

   Stephen L. Smith
   stephen.smith@york.ac.uk

[1] School of Computing, Napier University, Edinburgh, UK

[2] School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK

[3] Leeds Teaching Hospitals NHS Trust, Leeds, UK

[4] Wicking Dementia Centre and School of Medicine, University of Tasmania, Hobart, Australia

[5] Department of Electronic Engineering, University of York, York, UK

[6] School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, UK

'sporadic' or 'idiopathic' PD) are still unknown, but involve complex interactions between genetic and environmental factors [46].

PD is the second most common neurodegenerative disorder after Alzheimer's disease, affecting 1% of the population over the age of 60 and reaching approximately 5% at 85 [69]. The prevalence is rising due to ageing populations. According to the Parkinson Disease Foundation [63], about 10 million people worldwide have PD, one million of them in the USA, 1.2 million in Europe [59], and two million projected in China by 2030 [19]. One out of 500 individuals in the UK is affected, and it is expected that this number will rise threefold in the next 50 years [61]. There is currently no proven disease-modifying therapy [24]. The diagnosis of PD requires the presence of bradykinesia (slowness of movements) in addition to muscle rigidity or tremor or postural instability [62]. Approximately 20% of patients do not develop a tremor [37]. The manifestations of PD are not limited to motor impairments.

Prompt diagnosis of PD is important in order to provide patients with appropriate treatment and information on prognosis. However, an accurate early diagnosis can be challenging because the movement symptoms can overlap with other conditions [72]. Doctors make the diagnosis of PD based on clinical evaluation, interpreting information gained predominantly through history-taking and examination of the patient. Sometimes brain imaging may be requested to help support the clinical diagnosis, but there are currently no tests that are wholly sensitive or specific for Parkinson's. The rate of misdiagnosis of PD is approximately 10–25% [38], and the average time required to achieve 90% accuracy is 2.9 years [36]. Autopsy is still the gold standard for the confirmation of the disease.

There remains a need for quick and non-invasive tests to provide objective results to support a clinician's diagnosis. We address this in our work, with the aim of developing a medical device that can assist with early diagnosis of PD, focusing on the primary care context where the rate of misdiagnosis is particularly high [38]. Patients with suspected PD could then be forwarded for expert assessment by movement disorder specialists. The approach is based around a graphics tablet on which a patient traces or copies a cognitive assessment figure; this has the benefit of collecting a lot of information about the patient's movements and cognitive processes in a short period of time using an inexpensive device. The system then uses a deep learning model to detect whether the patient's drawings shows signs of Parkinson's disease.

In this paper, we describe the training and selection of the deep learning model. Unlike earlier work in this area (see Sect. 2.2), we focus on developing a model that can diagnose Parkinson's disease from a single drawing. This is important, because elderly patients fatigue quickly,

meaning that it is not practical within a primary care context to ask them to carry out multiple drawing tasks. In particular, we show that the use of dynamic movement data (rather than static images) combined with data augmentation techniques allows us to build a highly predictive model without having to integrate information from multiple drawings. Also of importance from a clinical perspective, we show that PD can be diagnosed using an intentionally simple CNN model. Simple models are more likely to generalise beyond their training data and hence are considered more trustworthy for medical diagnosis.

1.1 Figure-drawing tasks for assessing Parkinson's disease

Due to the lack of accepted definitive biomarkers [53] and specific neuroimaging findings [51], the diagnosis of PD is typically based on patient history, observations, judgements on clinical examination criteria and specific symptom questionnaires. These test outcomes are highly examiner-dependent (based on training and experience), with variability among different groups of observers [68]. The necessity of systematic kinematic tests to aid for clinical decision making led to the development of independent and objective quantitative assessments, more suitable for statistical analysis and data processing. Some of these tools, such as the systematic analysis of data from the finger-tapping test [5], the use of handwriting [20] and sketching abilities [73], have already been proposed to evaluate motor and cognitive function in the clinical setting to assess and diagnose PD.

Kinematic aspects of handwriting movements such as size, speed, acceleration and stroke length are affected in PD from its early stages [82]. As PD progresses, changes in handwriting occur with reductions in writing size (micrographia) [16] and decreased ability to write in general (dysgraphia) [47]. These deficits can be used to diagnose and monitor PD. Research to date has investigated signature writing [67] and the writing of short phrases [41]. The disadvantage of selecting handwriting abilities for PD diagnosis is that this skill is correlated with culture and penmanship, along with the level of literacy and education of the individual [22]. On the contrary, the execution of drawing tasks is considered an education-independent measure and may be more sensitive in detecting early signs of PD [80]. They are also fast, non-invasive and relatively easy to perform. There are different graphonometric methods used as tests, where patients have to draw figures of different levels of complexity like a spiral [73], cube [8], pentagon [6], interlocking pentagons [4], meander [64], star [78], the Bender–Gestalt test [54] and more complex figures like the clock [9], the Benson or the Rey–Osterrieth figure copy test [76]. Each test can be applied to

particular aspects of PD. For instance, the pentagon task has been used for the analysis of cognitive decline [40], to assess at the same time both motor and cognitive levels [85] and to compare PD with other neurodegenerative diseases [15].

The analysis of drawings provides significant motor function data as a result of the force, speed, time, tightness and uniformity generated by the patient for a period of time. However, it is not straightforward for clinicians to diagnose PD based on a simple visual inspection and requires detailed analysis. Although tremor may be visually apparent, tremor manifestations are not a symptomatology requirement in PD. Some 30% of patients do not develop this sign, and it is even less predominant at the early stages of the disease. However, this information can be used as the input for a computational model designed to support the diagnosis of PD. Computational models have been effectively applied to classification problems in the area of health care for a long time [88]. One successfully and widely used complex model with a multi-layer structure is the deep neural network (DNN). The learning methods that support multi-layer models are generally categorised as deep learning (DL). DL is a multi-level feature learning method that can deal with multimodal data and high-dimensional search spaces [31, 44]. Its performance and versatility are two reasons why this technology has been extended to a variety of different domains, including image classification [33], speech recognition [34], among many others.

The goal of this work is to use DL to analyse the information collected from patients' drawings in the form of images as a basis for discriminating PD patients from healthy controls. The architecture selected for this work is a convolutional neural network (CNN), a form of DNN that is known to work well with image data. Specifically, we aim to develop DNN models to achieve the following objectives:

- Selecting the most suitable model structure for our CNN classifier to automatically learn significant features from drawing assessments in order to differentiate between PD and healthy controls.
- Developing a reliable set of tests to investigate which data representation is the most informative option for training predictive models.
- Comparing two different drawing tasks (pentagon and cube drawing) to examine which one is more informative for discriminating PD as input for a CNN classifier.
- Analysing the effect of applying augmentation techniques on the classification performance and its level of stability (variance).

The remainder of this paper is organised as follows: Sect. 2 introduces DL as a tool to support learning in DNN models, presents a general overview of the CNN topology and illustrates the way in which other studies have applied these techniques to medical diagnosis. Section 3 outlines the datasets and the methods employed in this work, the description of the experiments performed and the procedure used to validate our results. Section 4 shows the set of experiments conducted and the results obtained from the analysis of the multiple classification scenarios. Section 5 comments on the experimental results in detail. Finally, Sect. 6 summarises this paper and lays out directions for future work.

2 Deep neural networks

DNNs are advanced multi-layer network models that are able to deal with complex, nonlinear and unstructured data such as audio, video, image and text by transforming them into a hierarchical structure of features with multiple levels of abstraction [44]. A crucial advantage of such models is that the transformation is performed without the intervention of human expertise and without the need to perform any feature extraction and data preprocessing. The feature extraction is, instead, automatic [31].

2.1 Convolutional neural network topology

The way in which the multiple layers of a DNN are linked and arranged characterises its topology, also called architecture. A CNN is a deep feed-forward DNN that was inspired by the structure of the cat's visual cortex. Using only the local connectivity of the nodes arranged in adjacent layers, the CNN specialises in processing grid-like data such as images [32] and performs this learning by extracting features from raw data automatically [12]. The CNN architecture has shown remarkable performance on hard classification problems [33]. A typical CNN topology consists of a combination of several convolution layers that can extract features from input data based on the local underlying spatial patterns, allowing for learning features with a higher level of abstraction [44]. Each layer is composed of three cardinal stages: (1) convolution, (2) activation function (nonlinear transformation) and (3) pooling (nonlinear down-sampling). By stacking these layers together, the network is able to extract progressively more abstract patterns, reducing the number of connections of the network. Afterwards, the extracted features are transformed to a one-dimensional vector using a flattening layer, and finally, the CNN combines these convolutional layers with traditional dense layers to produce the output of the classifier.

2.2 Deep learning for medical diagnosis

DL has been successfully applied in the broad area of medical diagnosis [48], including medical imaging [87]. For image-related problems, CNNs and its variants have been widely used in this field due to their extraordinary ability to exploit image data [43].

The use of drawing data and DL techniques was first proposed by Pereira et al. [64]. The research group investigated the use of a five-layer CNN to aid PD discrimination using 264 scanned images of 256×256 pixels showing meanders and spiral tasks gathered from 35 individuals as input. The authors achieved higher recognition ability measured by the accuracy per class metric processing spiral images (90.38%) than meander figures (83.11%). Another, more recent work using scanned data was conducted by Seedat et al. [74]. The most important contribution of these authors is the size of the dataset, which is significantly larger than the rest included in other works, with data from 370 PD subjects and 357 controls. However, paper-based tests imply that only X, Y coordinates and pressure as changes in terms of shades of intensity were collected. Despite that, authors reported accuracies of over 98% using a pretrained hyperparameter optimised CNN approach with data augmentation.

In [66], the group of Pereira explored the use of different well-known CNN architectures to analyse a set of 308 images gathered from 35 individuals performing the same type of tests. The HandPD dataset, gathered initially as a time series from a biometric pen, was initially transformed into a set of vectors composed by six signal channels. For each time step, these vectors were stacked together to form an image. The approach achieved a performance level of 87.14% for the meander images and 80.19% for the spirals using a CaffeeNet topology. Pereira et al. [65] extended their work using the same sensors, a larger dataset, called NewHandPD, with information from 92 individuals and a time series-based image pattern representation. The paper covered the comparison of three different CNN architectures (CaffeNet, CIFAR-10_quick and LeNet), three baselines and a combination of six different tests that were linked in a fusion approach to reach an average accuracy of 95.74% for 128×128 pixel size images with the CaffeNet architecture.

Recurrence plots were applied by members of the same research group led, this time, by Afonso et al. [2], to map the signals gathered from the NewHandPD dataset onto the image domain. These images were further used as input of the previous three CNN topologies. The experiments compared also the same two image resolutions (64×64 and 128×128), achieving the best results (88.05%) with the meander 64×64 pixel-size figure and the CaffeNet architecture.

Two similar unsupervised clustering approaches using a deep optimum-path forest (OPF) model were then proposed by Afonso et al. in [1, 3], using the NewHandPD dataset. In both works, the OPF was used as a feature extractor for three traditional machine learning algorithms, namely Bayesian classifier, supervised OPF and support vector machine (SVM). In [3], accuracies from meander and spiral tests were rather similar, with values around 81%; meanwhile in [1], the accuracy from the meander dataset outperformed the spiral by over 2%, reaching almost 84%. Linked to this research is the work of De Souza et al. [79], where a fuzzy OPF is used, merging HandPD and New-HandPD datasets, and using restricted Boltzmann machines as feature extractors, reaching 79.57% and 77.94% accuracies for meander and spiral, respectively.

Four recent papers approached the diagnosis of PD using deep recurrent neural networks (RNN). A bidirectional gated recurrent unit network, along with an attention mechanism, was investigated using the NewHandPD dataset [70], achieving superior results with the meander figures (92.24%) compared to the spiral (89.48%) and outperforming previous works on this dataset. Gallicchio et al. [27] proposed another type of deep RNN architecture, a 10-layered deep echo state network (ESN) and a different significantly imbalanced public dataset called ParkinsonHW with 61 PD patients and 15 controls, reaching accuracies of up to 89.3%. This dataset contains information about pen position (x and y components), pressure and grip angle. Szumilas et al. [81] suggested also the use of an ESN-ensemble model to quantify kinetic tremor in PD by drawing circles on a digitising tablet, using, in this case, a dataset of 64 PD patients. Finally, in [75], the authors compared an ESN with a long short-term memory model using our dataset and reaching accuracies of 91% for the LSTM and 93.7% for the ESN.

Considering the same ParkinsonHW dataset, Canturk [11] employed a CNN-based approach, selecting the pretrained AlexNet and GoogleNet models as feature extractors to achieve an accuracy of 94%. In this case, the author applied a fuzzy recurrence plot to convert time-series signals into greyscale texture images and K-Nearest Neighbour (KNN) and SVM as final classifiers, reporting the superiority of SVM over KNN by only 1%. In [29], this accuracy was increased to 96.5% with the same AlexNet approach, but using spectrum points as input data, since PD symptomatology is better reflected in the frequency domain. Another similar, but inferior work in terms of final accuracy (88%) was also published by Khatamino et al. [42], inspired by the time-series image representation of [65].

Moetesum et al. [55] used a set of eight pre-trained CNNs (AlexNet) as a feature extractor system to be used by a SVM classifier. The networks were trained with the PaHaW dataset [20] that comprises 72 subjects (37 controls and 38 PD patients) performing eight different tests, one of them being a spiral drawing. Afterwards, using fusion techniques, the eight outputs were combined to provide a final single metric. Information was collected as sequential data by a digitised pen and transformed into images using X, Y coordinates and zero-pressure information, achieving 83% in overall accuracy and 62% for the spiral data.

Using the same dataset, Diaz et al. [18] integrated together the features extracted from three parallel VGG16 CNNs, which shared the same 16-layer architecture, but trained with different data representations and transfer learning. As a result, the extracted features were given as the input to a combination of traditional ML models (SVM, random forest and AdaBoost). This work reported a maximum accuracy of 86.67%, gathered by the ML ensemble, using a majority voting scheme.

The next work that continues experimenting with the PaHaW dataset is the study conducted by Naseer et al. [57]. In this case, authors proposed a deep 25-layer CNN classifier (AlexNet), with transfer learning and data augmentation, achieving an outstanding accuracy of 98.28%. Authors used the ImageNet and MNIST fine-tuning-based approach over the spiral data of the PaHaW dataset and reported that the AlexNet-ImageNet approach outperformed the MNIST pre-trained version by over 3%.

In the work of Vasquez et al. [86], data collected from speech, handwriting and gait were used together as a multimodal ensemble mechanism to distinguish between PD patients and healthy controls. Handwriting data consisted of 14 tasks, including circle, cube, rectangle and spiral drawings gathered from a total of 84 subjects, 44 PD patients and 40 controls, as a time series data. From that, a feature extraction step collected the transitions in handwriting. A one-dimensional CNN with four layers was designed to extract spatial features from these transitions and sent them as input to a SVM model. The approach achieved high accuracy (97.6%) when information from speech, handwriting and gait were combined. However, using the handwriting data as a single classifier was not very effective, resulting in only a 67.1% accuracy.

Much of the existing work in this area has been done using a small number of publicly-available datasets, containing relatively few data points. In addition, the focus has been on using increasingly complex predictive models to raise accuracy rates, with the best accuracies achieved using deep architectures and ensemble models. All of these factors contribute to the likelihood of overfitting. The use of small datasets to train and test deep neural architectures is particularly concerning, since this will likely lead to many model parameters being under-specified. However, large datasets are very difficult to acquire. Hence, going against this trend, our work focuses on using shallower CNNs, where the number of trainable parameters is much smaller, and hence the generality is likely to be greater when trained on small datasets. Rather than focusing on more complex models, we instead investigate the features within the data that are most significant for accurate classification, and tailor the representation of the data to emphasise these.

Another important consideration that has not really been addressed by the existing literature is the burden placed upon patients when collecting data within a clinical setting. The most accurate existing models have been achieved by forming ensembles from multiple data modalities. This, in turn, requires patients to undergo a corresponding number of data collection exercises, something that may be difficult to achieve in practice with elderly and physically infirm patients. In our work, we focus on training models that require only a single drawing as their input, hence minimising the burden placed upon patients in the clinic, and providing a more practical predictive model for use in a primary care setting.

A summary of the related work introduced in this section can be seen in Table 1, in chronological order. An extended comparison of these studies can be found in the Sect. 5, in Table 16.

3 Methodology

The methodology used in this paper is illustrated in Fig. 1.

3.1 Data acquisition

For this study, the data were collected by clinicians at Leeds Teaching Hospitals NHS Trust. The dataset comprises information acquired from 87 subjects (58 patients and 29 aged-matched healthy controls). Patients were recruited from neurology clinics and had been diagnosed by PD specialist consultants according to the Queen Square Brain Bank Criteria [28]. Controls were the spouses or friends of patients and were included if they had no neurological disorder. The study was conducted in accordance with the corresponding institutional review board. Every subject provided written informed consent before the tests.

All the subjects were asked to copy the wire cube from a sample image and draw the Archimedean spiral pentagon on top of a template image, using an inking stylus and a digitising, pressure-sensitive Wacom tablet (Wacom Technology Corporation) of size 20.3 cm × 32.5 cm. In the cube task, each subject performed one drawing with the dominant hand, whilst in the pentagon task, they carried

Table 1 List of works included in the literature review

Year	Reference	Authors	Dataset	Tests
2016	[64]	Pereira et al.	Former HandPD (paper), 32 subjects	Meander, Spiral
2016	[66]	Pereira et al.	HandPD (digital pen), 32 subjects	Meander, Spiral
2017	[3]	Afonso et al.	HandPD (digital pen), 32 subjects	Meander, Spiral
2018	[65]	Pereira et al.	HandPD, 92 subjects, time-series image representation	Meander, Spiral
2018	[27]	Gallicchio et al.	ParkinsonHW, 77 subjects	Spiral
2018	[42]	Khatamino et al.	ParkinsonHW, 77 subjects, time-series image representation	Spiral
2018	[55]	Moetesum et al.	PaHaW, 72 subjects	Spiral, 7 writing tasks
2018	[86]	Vásquez-Correa et al.	Custom (speech, handwriting and gait), 84 subjects	Spiral, circle, cube... (8 tasks)
2019	[2]	Afonso et al.	HandPD, 35 subjects	Meander, Spiral
2019	[70]	Ribeiro et al.	HandPD, 35 subjects	Meander, Spiral
2019	[29]	Gil-Martín et al.	ParkinsonHW, 77 subjects	Spiral
2019	[18]	Diaz et al.	PaHaW, 72 subjects	Spiral, 7 writing tasks
2020	[1]	Afonso et al.	HandPD, 35 subjects	Meander, Spiral
2020	[74]	Seedat et al.	Custom (paper), 707 subjects	Spiral
2020	[81]	Szumilas et al.	Custom, 64 patients	Circles
2020	[11]	Canturk	ParkinsonHW, 40 subjects	Spiral (dynamic and static)
2020	[57]	Naseer et al.	PaHaW, 75 subjects	Spiral, 7 writing tasks
2021	[79]	De Souza et al.	Merged HandPD and NewHandPD, only final number of samples,	Meander, Spiral
2021	[75]	Shenoy et al.	Same as this paper, 87 subjects	Only Spiral pentagon
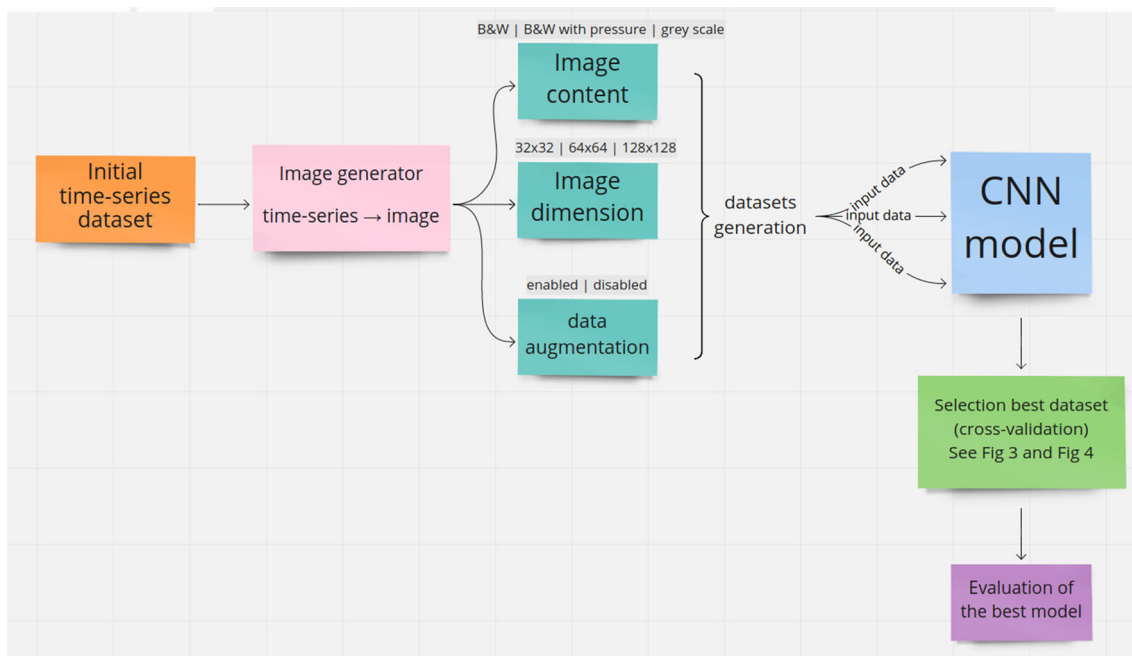
Fig. 1 Workflow followed in this study

out four drawings, two with each hand. The instructions indicated that the figures should be drawn as accurately and as fast as possible. Figure 2a shows the spiral pentagon template that subjects were asked to follow. Figure 2b–d are examples of pentagon and cube drawings.

The collected data were stored to assess the performed movements during the drawing process offline in the form of time series data. The tablet recorded data with a constant sample ratio of 200 Hz. In each sample, information of the time starting at zero, coordinates X and Y of each pen

location, the angles in which the pen is used with respect to the X and Y plane and the relative pressure exerted against the tablet were stored as a multivariate time series dataset.

Coordinates X and Y and pressure values are represented in the range of [0, 1] and pen angles in the range [− 1, 1], and timestamp entries are monotonic integer values starting from zero. We can interpret, when zero-pressure values are collected, that the pen at this location was not in contact with the tablet.

Along with the time series dataset, other general information of the subjects was gathered including whether they were patient or control, age, gender, hand used in the test and handedness. Baseline diagnosis and movement severity were assessed by clinicians using the Movement Disorder Society Unified Parkinson's Disease Rating Score (MDS-UPDRS) part 3 [56] for motor-related skills and the Montreal Cognitive Assessment (MoCA) score [58] for measuring cognition levels. All the information was stored in files for further analysis. A summary of the age, gender, disease duration, MDS-UPDRS score, MoCA score and Levodopa Equivalent Daily Dose (LEDD) is shown in Table 2.

There is a small, yet significant, difference ($p = 0.09$) between the mean ages of the control and PD groups. PD is more common in males and the gender gap has been exaggerated by the fact that the control subjects were the spouses or friends of the PD participants. The mean scores for both UPDRS and MoCA differ significantly ($p < 0.001$ for both) between the control and PD groups.

After a preliminary inspection of the dataset, it was seen that the complete set of samples was imbalanced, with the number of patients significantly higher than the number of control subjects. This factor has significant implications for the training of the classifier. In addition, for the pentagon dataset, we only used the collected data of the first and the second repetition of the subjects' dominant hand. The rationale for this decision was that the ability to complete the non-dominant hand tasks varied greatly between individuals, presumably related to their degree of ambidexterity, and was not felt to reliably reflect motor control.

Table 2 Participants information

Characteristics	PD group ($n = 58$)	Controls ($n = 29$)
Age, years	69.2 (8.4, 44–85)	66.1 (7.6, 50–79)
Gender, M:F	38:20	5:24
Handedness, R:L	51:7	22:7
Disease duration, years	6.2 (4.7, 0.5–20)	
MDS-UPDRS Part 3	28.8 (11.5, 3–56)	1.9 (2.3, 0–8)
MoCA score	23.1 (4.1)	26.3 (3.0)
LEDD, mg/day	662.7 (560.9)	

3.2 Data preprocessing

Following the preliminary inspection of the data, all incomplete drawings (2 patient and 3 control) were removed, and the image-based dataset was then created by representing the time series of each subject as a two-dimensional image, connecting the coordinates of the trajectory described by the pen [55]. Other alternatives have also been investigated. In Camps et al. [10], the data gathered by an IMU wearable device (accelerometer, gyroscope and magnetometer sensors) were formatted as a grid structure using a spectral window stacking procedure and transformed into images. In Pereira et al. [66], a five-column dataset gathered by a digitalised pen was transformed into an image to be the input of a CNN. The pen sensors included a microphone, finger grip, axial pressure of ink refill, tilt and acceleration in X, Y and Z directions.

In the present work, we investigate different data representations for the transformed set of images. Specifically, we cover three data representations with increasing levels of complexity. The first and most simple approach extracts the X and Y coordinate data, discarding zero-pressure values and angles. Afterwards, it transforms this information into a two-dimensional black and white image. The next version adds zero-pressure information (coordinates where the pen passed without touching the tablet) as grey strokes to the black and white image. We include this information
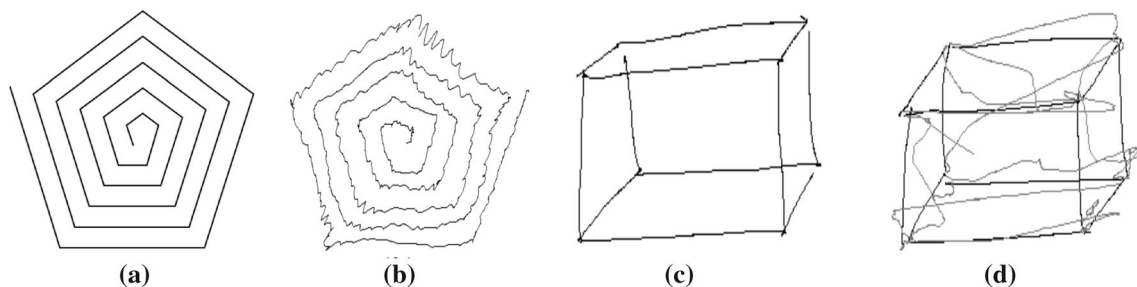

Fig. 2 **a** The spiral pentagon template, **b** a pentagon drawing from a patient, **c** a cube without zero-pressure information and **d** a cube with zero-pressure information

following the findings of Drotár et al. [20], who highlighted the importance of in-air trajectories in handwriting tasks for PD patients.

Finally, as our third approach, we are interested in introducing the whole range of pressure values in the image since it is known that pressure decreases with the progression of PD [84]. Here, we extend the black and white representation to a greyscale image, where the grey information has been generated by scaling the pressure values from [0, 1] to [0–254]. We are also interested in differentiating between areas where the pen did not pass and areas where the pen passed, but without touching the tablet. Based on this, we created the new images by using zero values (black) to represent minimum pressure, 254 values to draw points with the maximum pressure that the subject can exert over the tablet and 255 values (white) to depict non touching points.

After trimming away outer edges around the drawing (white space) that are not of interest for the classification, these images were resized and normalised by creating a zero-mean normalised version with a unit standard deviation. Data were finally formatted appropriately to be used as input to a CNN. The resize step created three different versions for each image of sizes 32×32, 64×64 and 128×128 pixels to study how resolution influences the classification. Afterwards, additional images were produced using augmentation techniques [12].

When the amount of labelled data is limited, which is often the case in the medical field, data augmentation is a critical preprocessing step for training CNNs to teach the network the desired invariance, provide robustness [71] and avoid the performance deterioration linked with class imbalance in the training data [50]. The process of augmentation involves the transformation of the existing images to create new ones. Choosing a strategy for augmentation is not trivial and could be even more crucial than the selection of the architecture [31]. Suitability of each technique can only be tested using trial and error methods since there is not a single strategy that is superior to the rest [45]. Advanced techniques require significant expert knowledge, such as texture transfer, selective blending, kernel filtering and directional lightning addition, and can also be computationally expensive like the use of generative adversarial networks [49]. On the contrary, traditional geometric transformations are fast, reproducible and easy to implement [52]. Flipping and rotation have proven useful on datasets such as CIFAR-10 and ImageNet. For some datasets, the use of rotation transformation can be heavily influenced by the rotation degree, e.g. in [77], where rotations greater than 20 degrees were found to be problematic.

In this work, new copies were generated by applying random rotation, random zoom with a certain value and random horizontal flip. In our case, we did not find that rotation misclassified drawings and we implemented this feature with a random rotation degree of up to 40 degrees. The amount and distribution of the new set of images are defined as follows: for each control image (cube or pentagon drawing), 23 perturbed copies were created and 11 for each patient (cube or pentagon drawing). Table 3 shows the initial and final numbers for each type (cube-pentagon and control-patients).

3.3 Architecture and training

The CNN architecture consists of two convolutional layers with 32 filters followed by two convolution layers with 64 filters and another two convolutional layers with 128 filters, three max-pooling layers of size (2×2), six dropout layers, three dense layers and one flattened layer. All the activation functions are ReLU (rectified linear unit), except for the last dense layer, where a sigmoid activation function was selected to map the binary output. ReLU is the most used activation function for CNN [43]. In each convolutional layer, we used the same padding mechanism to maintain the size of the layers after applying a series of convolutional operations. Finally, we use a stride of size (3×3). Figure 3 shows the CNN architecture.

For performing the image classification between subjects, the CNN model was trained using backpropagation on the images that were produced from the time-series datasets and through the application of augmentation techniques. After the training, the model was tested as a classifier to differentiate between healthy subjects and patients using a test set of previously unseen images.

DNN models, when training in supervised mode, use different datasets for the training and testing procedures. Following that, we employed 90% and 10% of the samples, extracted from the main datasets, for training and validating and testing purposes, respectively. The samples contained in each group were randomly selected. This procedure allows us to evaluate the accuracy of our framework. We conducted a tenfold cross-validation [23].

The CNN has the layers initialised using the Xavier/Glorot initialisation schema [30]. Other common hyperparameter values are 0.003 as the initial learning rate, $1e^{-6}$

Table 3 Images created over the original datasets (number of samples in brackets) by applying augmentation techniques

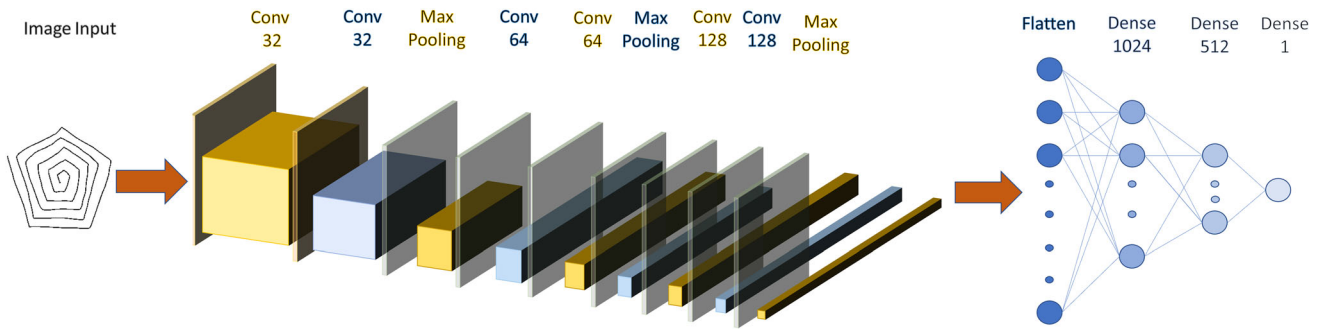|	Cube images	Pentagon images
Control	(56) 598	(51) 1173
Patient	(26) 616	(112) 1232
Total	(82) 1214	(163) 2405

Fig. 3 CNN architecture with a pentagon image as input (left), the convolutional and max-pooling layers (middle) and the schematic representation of the feature reduction that occurs from the flattened to the output layer (right)

as a decay function and momentum equal to 0.9. The training algorithm aims at minimising a binary cross-entropy loss function between the predicted and the real diagnosis. The optimisation algorithm uses mini-batch learning with a batch size equal to 16 to speed up the learning. The training uses an early stopping mechanism as a regularisation technique to avoid overfitting with a two-fold stopping condition: a maximum number of epochs equal to 150 and stopping after 25 epochs without improvement in the validation set.

3.4 Experimental set-up

In this subsection, we explain the experimental set-up and how the different test sets were defined and grouped. We used Python 3 to run our experiments and analyse the results. We worked under the Keras deep-learning framework [13] to take advantage of the straightforward configuration of DL pipelines. We also used several libraries specialised in DL such as NumPy and Pandas that help us to process the datasets and Sklearn to extract the results from the models. The experiments were grouped based on four factors:

- Experiments with black and white images with and without zero-pressure information to investigate whether keeping zero-pressure information is crucial in the discrimination process.
- Experiments with greyscale images with zero-pressure information.
- Experiments with balanced and imbalanced datasets to investigate the impact of the class distribution on classification performance and stability.
- Experiments with a variety of image resolutions including 32×32, 64×64 and 128×128 pixels.

In total, we completed 36 different experiments on both pentagon and cube datasets.

3.5 CNN assessment

The CNN models were assessed as follows:

1. *Phase 1:* Evaluating the results of the ten runs performed for each configuration described in the previous sections. The topology and configuration that achieve the best performance are then selected for further analysis.
2. *Phase 2:* Using the previous top-performing configuration, we select the best of the ten different models (set of weights) produced by the application of cross-validation when training. This model will be further evaluated and reported as the final performance output of this paper.

In the first phase, we analysed the results of the experiments using several nonparametric statistical tests including Mann–Whitney U test two-tailed, Kruskal–Wallis test and Tukey's honest significant difference test as a post hoc test based on the studentised range distribution. These tests had a level of statistical significance at $p < 0.05$.

We used Kappa [14] as our primary performance metric. Kappa is a statistical measurement of the agreement between two rankers. It is a robust metric, simple to compute, and with an output range between $[-1, 1]$. Kappa values K are calculated as follows:

$$K = \frac{p_0 - p_c}{1 - p_c} \tag{1}$$

where p_0 is the total agreement probability among rankers and p_c is the agreement probability due to chance. In our case, the rankers are the original class (ground truth) and the predicted class generated using the trained classifier. There is no standard method to interpret Kappa values, but Fleiss et al. [25] considered that a Kappa value > 0.75 is excellent, 0.4–0.75 is fair to good, and < 0.4 is a poor agreement. A Kappa value could be negative, but it is unlikely in practice.

The reason behind the use of Kappa instead of the traditional accuracy measure of classification performance is

that for a significant part of the experiments our datasets are imbalanced. This characteristic implies that using traditional metrics to calculate the classification accuracy can be misleading [35]. For example, consider a test set of three controls and six patients. If the classifier predicts all samples as patients, then the classification accuracy will be about 66%. Meanwhile, the Kappa statistic for the same configuration will be 0. In this case, it can be seen that Kappa gives a stronger indication than the traditional accuracy metric for classification.

The comparison procedure of phase one starts by evaluating the multiple configurations listed in the previous section, grouped as tuples. A summary of the process is illustrated in Figs. 4 and 5. The assessments are done level by level until reaching a winner. Figure 4 represents the different configurations tested for the cube and pentagon datasets, and Fig. 5 summarises the last comparison level and the network option finally selected as our best approach. For simplification purposes, notice that each box includes experiments with and without zero-pressure.

To analyse the results, we used boxplots and descriptive statistics (five number summary) to illustrate the distribution differences for our best six balanced configurations (Fig. 6), the three best balanced against the three best imbalanced configurations (Fig. 7), and comment on the stability of their performances.

Once this step concludes, we focus on our best configuration to further determine its performance and analyse its efficiency. The selected traditional assessments include the accuracy as a measurement to evaluate how well the predictor classifies both classes, the confusion matrix (actual vs predicted classification), specificity and sensitivity/recall

(recognition rate per each class, respectively), precision (positive predictive value), f1-score (harmonic mean between precision and recall) and the average precision score.

4 Results and evaluation

This section presents and analyses the results generated by the two validation phases. Afterwards, the major findings are discussed. The most accurate results in the tables below, based on the nonparametric statistical tests described in the previous section, are highlighted in bold.

4.1 Evaluating the experimental results

Using the comparison approach described in Sect. 3.5, we performed the experiments designed and summarised in Fig. 4. Results are shown in Tables 4, 5, and 6. Table 4 shows the results for the CNN classifier on the pentagon and cube datasets, using a black and white representation without zero-pressure information over the validation set. The table shows the averaged Kappa values, considering imbalanced and balanced cases and a variety of image resolutions.

In the next set of experiments, we continue with the black and white representation, but including areas where the pen was not in touch with the tablet. Table 5 shows the averaged Kappa values over the validation set, considering imbalanced and balanced cases and a variety of image resolutions.
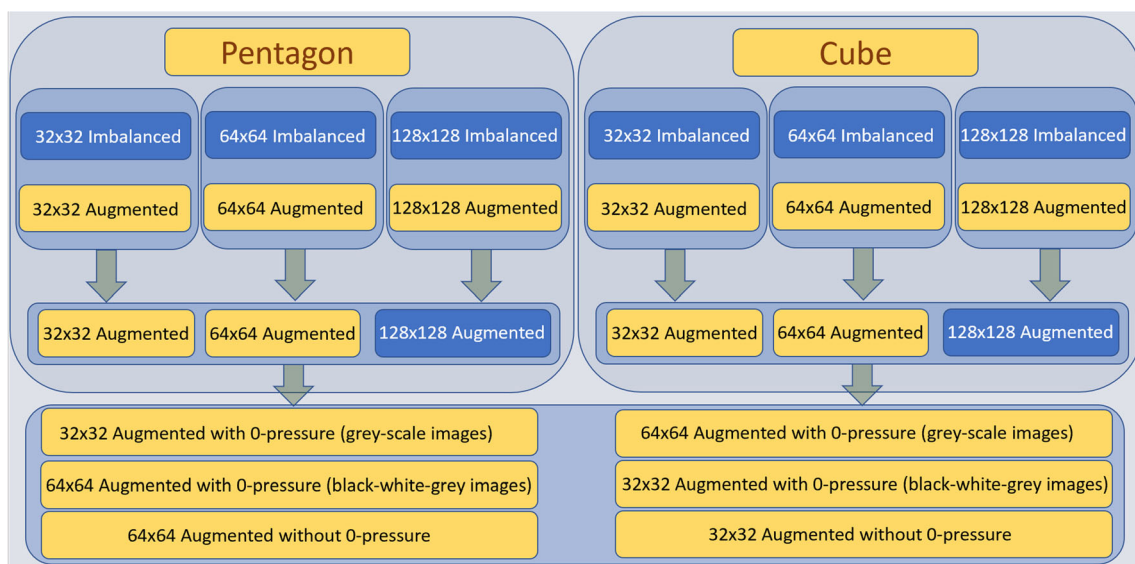

Fig. 4 Visual representation of the set of configurations being compared. Green arrows represent the order of the comparisons, blue boxes inferior configurations, and yellow options are the winning counterparts

Fig. 5 Final test configurations (left) and the option selected as the best approach (right)
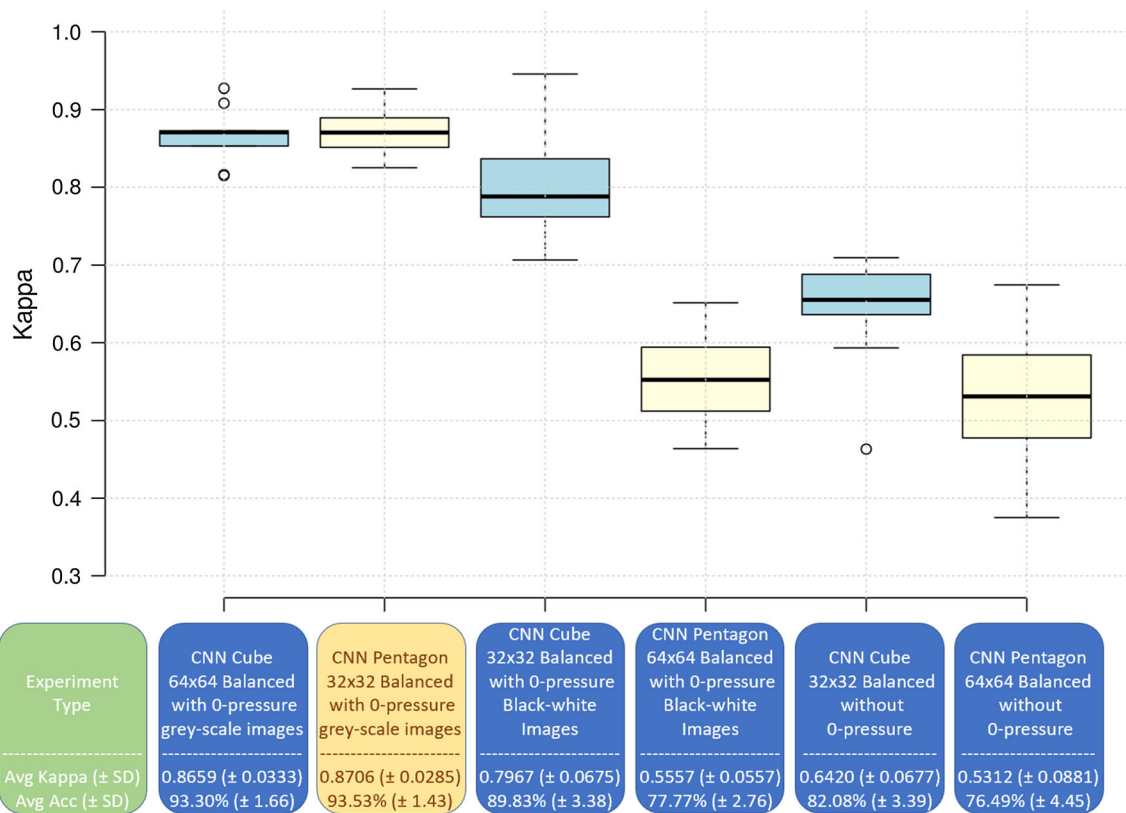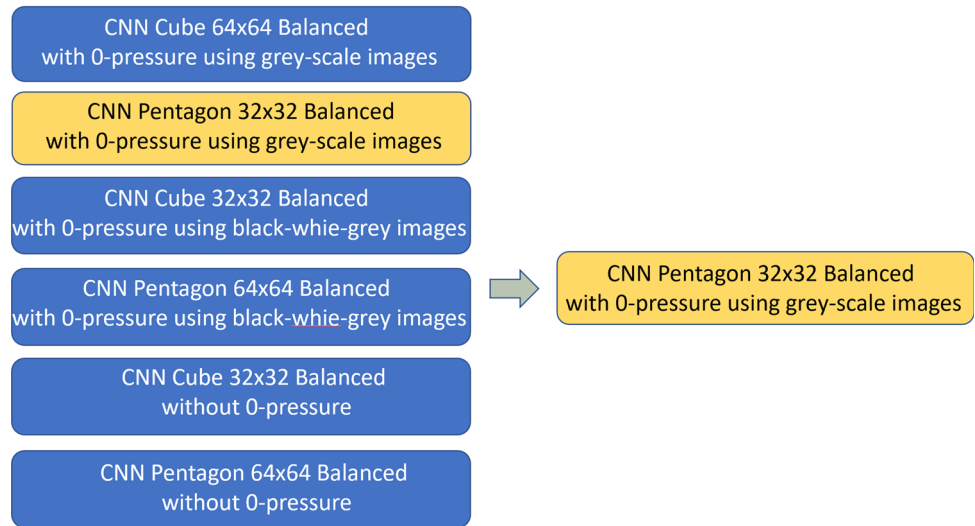

Fig. 6 Distribution of the Kappa values of the six best configurations detailed in Fig. 5

In addition to the inclusion of the in-air information, Table 6 incorporates the whole range of pressure values in the representation of the image using a greyscale representation.

From the results reported in these tables, we can study and comment on the effects of the different configurations with respect to the final performance of the classification.

4.1.1 The effect of applying augmentation

In all the configurations, the datasets with augmentation (balanced) led to better results than the imbalanced options. If we averaged the results of the balanced datasets on one side and imbalanced on the other among the different resolutions and we calculate the difference between them, it can be observed in Table 7 that applying augmentation
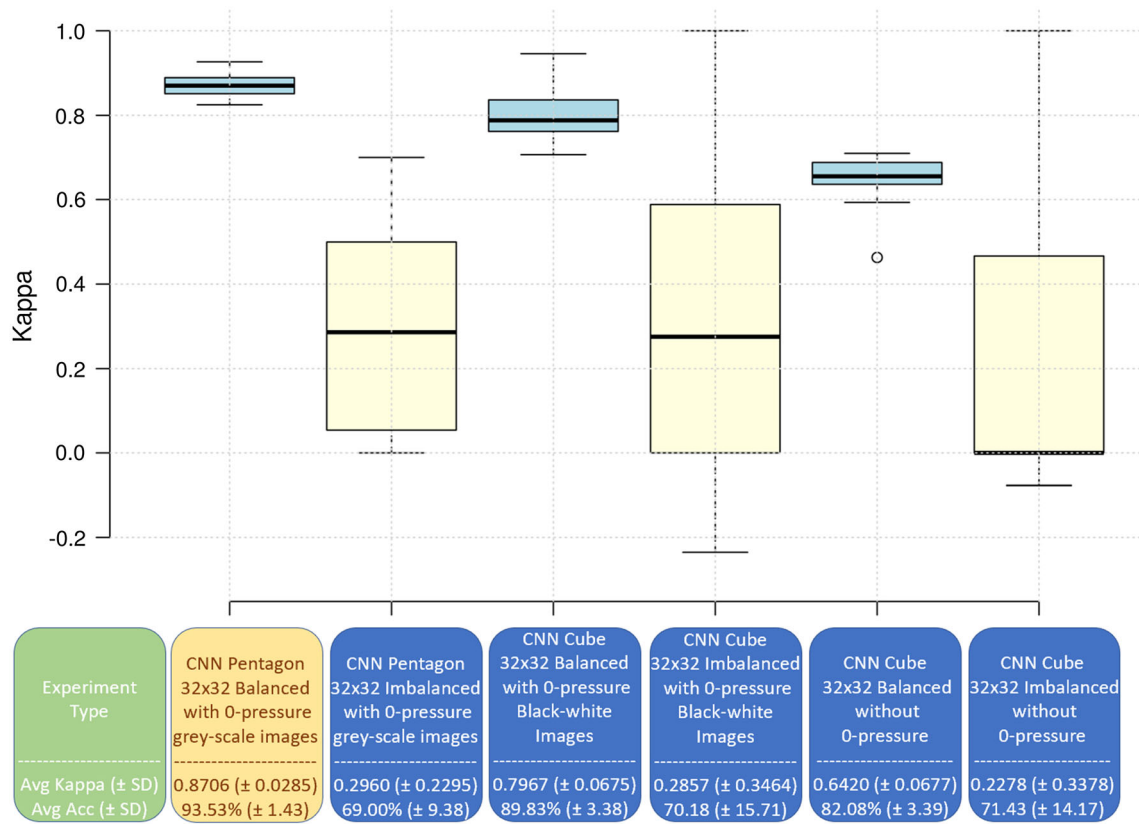
Fig. 7 Distribution of the Kappa values of the best three configurations for balanced and imbalanced datasets

Table 4 Summary of the average Kappa values over ten runs using only the coordinates of the pen when the pressure was bigger than zero over the validation set

Black and white without zero-pressure				
Drawing	Distribution	32×32	64×64	128×128	Mean
Pentagon	Imbalanced	0.082	0.086	0.055	0.075
	Balanced	0.48	**0.531**	0.461	0.491
Cube	Imbalanced	0.227	0.123	0.085	0.145
	Balanced	**0.642**	0.604	0.581	0.609
Mean		0.337	0.336	0.296	0.323

Table 5 Average kappa values for the black-white representation with zero-pressure over the validation set

Black and white with zero-pressure				
Drawing	Distribution	32×32	64×64	128×128	Mean
Pentagon	Imbalanced	0.084	0.066	0.041	0.064
	Balanced	0.552	**0.555**	0.499	0.535
Cube	Imbalanced	0.285	0.27	0.179	0.245
	Balanced	**0.796**	0.792	0.697	0.762
Mean		0.429	0.421	0.354	0.401

techniques, the CNN classifier is able to outperform the imbalanced versions.

The improvement in Kappa values among the three data representations shows that the importance of this technique increases in function of the complexity of the data representation. This tendency is depicted for both datasets, especially for the cube drawings.

4.1.2 The effect of adding zero-pressure information on the input data

To analyse the consequences of adding zero-pressure values on the representation of the images, attention should be focused on Tables 4 and 5, which correspond to performances with and without this particular information. If we concentrate on the total average performance for each table, we can see that the non-pressure information affects positively the overall performance by adding almost $0.08(0.401 - 0.323)$ over the averaged Kappa values for each configuration.

Table 6 Summary of the average Kappa values over ten runs using a greyscale representation with zero-pressure information over the validation set

Greyscale with zero-pressure					
Drawing	Distribution	32 × 32	64 × 64	128 × 128	Mean
Pentagon	Imbalanced	0.296	0.283	0.291	0.29
	Balanced	**0.87**	0.841	0.75	0.82
Cube	Imbalanced	0.195	0.202	0.222	0.206
	Balanced	0.844	**0.865**	0.787	0.832
Mean		0.551	0.555	0.512	0.54

Table 8 Improvement achieved by adding zero pressure information over the black and white representation

Drawing task	No pressure		Pressure	
Pentagon	Imbalanced	Balanced	Imbalanced	Balanced
	0.075	0.491	0.064	0.535
	Mean 0.282		Mean 0.3	
	Improvement 0.017			
Cube	Imbalanced	Balanced	Imbalanced	Balanced
	0.145	0.609	0.245	0.762
	Mean 0.377		Mean 0.503	
	Improvement 0.126			

If, instead, we take into account the consequences of this addition for each dataset, the improvements are summarised in Table 8.

The results suggest that adding zero-pressure information increases the averaged Kappa values for both types of tests. We can see, however, that the cube performance gets higher benefits from adding this extra information to the image representation (≈ 0.13) than the pentagon dataset (≈ 0.02).

4.1.3 The effect of adding the range of pressure values

Using the same procedure with Tables 5 and 6, the global averaged Kappa value over all the configurations for the black and white representation with zero-pressure information is 0.401 and for the greyscale with zero-pressure is 0.54. Then, the general improvement achieved by the addition is 0.138.

If, instead of analysing the performances globally, we consider the performances for the pentagon and cube tasks independently, it can be observed in Table 9 that the pentagon task highly benefits from incorporating the greyscale representation (≈ 0.25) in comparison with the cube task (≈ 0.02).

4.1.4 The effect of the image resolution on the CNN architecture

The image size that generates the best Kappa results differs between the pentagon and cube datasets, and it is linked with the use of augmentation techniques, in-air information and pressure values. Variance values are outlined for balanced and imbalanced datasets in Table 10.

The effect of the image resolution is not homogeneous from the point of view of the use of balanced and imbalanced datasets. For imbalanced images:

- Pentagon images do not improve significantly with the use of different image sizes, independently of the representation used. The variance of performance between image sizes is minimal for the three types of image representation.
- The best absolute performance of the pentagon task in each configuration is rather poor (0.086, 0.084, 0.296), achieving in general higher performance values with small image sizes (32 × 32).
- Variance values decrease as the complexity of the data representation rises. This can be interpreted as the size

Table 7 Effect of applying augmentation

Data representation	Pentagon		Cube	
Black and White	Imbalanced	Balanced	Imbalanced	Balanced
	0.075	0.491	0.145	0.609
	Dif: 0.416		Dif: 0.463	
Black and white pressure	Imbalanced	Balanced	Imbalanced	Balanced
	0.064	0.535	0.245	0.762
	Dif: 0.471		Dif: 0.517	
Grey scale pressure	Imbalanced	Balanced	Imbalanced	Balanced
	0.29	0.82	0.206	0.832
	Dif: 0.53		Dif: 0.625	

Each value represents differences in the averaged performance for all the configurations using augmentation or not

Table 9 Improvement achieved by adding the full range of pressure values over the black and white with zero-pressure information

Pressure option	Pentagon		Cube	
B&W—zero pressure	Imbalanced	Balanced	Imbalanced	Balanced
	0.064	0.535	0.245	0.762
	Mean 0.3		Mean 0.503	
Greyscale	Imbalanced	Balanced	Imbalanced	Balanced
	0.29	0.82	0.206	0.832
	Mean 0.555		Mean 0.555	
	Improvement 0.255		Improvement 0.015	

Table 10 Variance values of the performances gathered among image resolutions, grouped into imbalanced and balanced datasets for each data representation

Data representation	Pentagon		Cube	
	Imbalanced	Balanced	Imbalanced	Balanced
B&W—no pressure	0.027	0.129	0.535	0.094
B&W—pressure	0.046	0.097	0.331	0.313
Greyscale	0.004	0.393	0.018	0.165

of the image is a very important factor for the black and white representation and insignificant for the greyscale images.

- For the first two configurations, the cube dataset achieves better results with a small image size (32×32). For the greyscale images, the tendency changes for a larger configuration (even when other two configurations are rather close), being the only configuration where the 128×128 pixel-size option achieves the best performance.

Focusing on the results gathered by applying augmentation techniques, the use of the three different resolutions shows the following:

- Performance values for pentagon and cube vary among resolutions, with also heterogeneous behaviours for both tests in the different configurations. Variance values for pentagon show that the performance of the grey images is more prone to variability across resolutions. In the case of the cube task, the variance indicates that the black and white with nonzero pressure information is the approach that changes more between resolutions.
- In terms of preferred sizes, pentagon images are more successful in medium size images (64×64), except for the grey-image scenario, where it is surpassed by the 32×32, but only by ≈ 0.03. Cube images tend to achieve better performance with smaller images (32×32), with again the exception of the greyscale representation, where 64×64 achieves slightly higher performance by only ≈ 0.02.

4.1.5 The effect of the application of augmentation techniques in the variability of the performance among runs

It is considered that a classifier is stable if the variance between multiple trainings is low, which is accepted as a desired characteristic of any learning algorithm [7]. Stability is linked with the randomness of the system that comes from the sampling of the training set. To measure this, we look at the standard deviation (SD) between cross-validation folds, focusing our analysis on the best performance configurations, listed in Fig. 5. Figure 6 illustrates the shape of the cross-validation distributions for these six best configurations. Numerical values are shown in Table 11.

The boxplots show a limited range of variability for all the CNN configurations, with an average of 0.056 and a maximum value of 0.088. The most stable results, which are also the two best performers, correspond to the greyscale representation for the pentagon and cube datasets with an averaged Kappa SD value of 0.03. We can also observe that both configurations have similar shapes. The rest of the approaches achieve values with an SD of (≈ 0.06).

It is also interesting to notice similarities in the shape of the distributions generated from the black and white pentagon with and without zero-pressure information. On the contrary, the distributions of the cube with and without zero-pressure information depict very distinctive shapes, which reinforces the idea that the addition of the zero-pressure information in the cube task affects its performance noticeably.

To study the variability between balanced and imbalanced configurations, a similar plot is included. Figure 7

Table 11 Summary of the average Kappa and accuracy values and their corresponding SD for the best six configurations

CNN experiment	Avg Kappa ($\pm$SD)	Avg Acc ($\pm$SD)
Cube 64 $\times$ 64 0-pres greyscale	0.865 ($\pm$ 0.033)	93.30% ($\pm$ 1.66)
Pentagon 32 $\times$ 32 0-pres greyscale	0.87 ($\pm$ 0.028)	93.53% ($\pm$ 1.43)
Cube 32 $\times$ 32 0-pres B&W	0.796 ($\pm$ 0.067)	89.83% ($\pm$ 3.38)
Pentagon 64 $\times$ 64 0-pres B&W	0.555 ($\pm$ 0.055)	77.77% ($\pm$ 2.76)
Cube 32 $\times$ 32 no 0-pres B&W	0.642 ($\pm$ 0.067)	82.08% ($\pm$ 3.39)
Pentagon 64 $\times$ 64 no 0-pres B&W	0.531 ($\pm$ 0.088)	76.49% ($\pm$ 4.45)

shows the three best balanced configurations along with the three best imbalanced options. Numerical values of the three best imbalanced configurations are shown in Table 12. From a visual inspection of the Kappa SD values, we can see that all the balanced configurations provide more stable results than the best imbalanced counterparts.

In the table, it can be seen that all the balanced configurations have a Kappa SD value lower than 0.07, with an average of 0.054. On the contrary, the imbalanced models have SD values no lower than 0.22, with an average of 0.304.

4.2 Evaluating the final model

The best-performing configuration is the CNN architecture using 32 $\times$ 32 pixel-size images of the pentagon drawing task including zero-pressure information and greyscale representation. To get a better idea of its generality, we take the best model trained during cross-validation (as measured by the validation set), reevaluate it on the test set and consider various performance metrics. Table 13 shows Kappa values, classification accuracy, specificity and average precision for the validation and test sets.

The Kappa value and the accuracy for the best single-model over the validation set achieve 0.926 and 96.31% and for the test set these figures drop slightly to 0.9 and 95.02%, respectively. On top of that, Table 14 illustrates additional metrics gathered from the same model such as specificity, sensitivity, F1-score and support (number of samples in the test set).

Finally, Table 15 illustrates the confusion matrix of this model using the validation and the testing sets. The matrix shows the number of samples that the system classifies as

true positive (TP), true negative (TN), false positive (FP) and false negative (FN). We can see that this model successfully classified 116 out of 118 of control images and 113 out of 123 of patient images in the test set.

It is notable that the CNN correctly classifies patients who are in the early stages of the disease, with the misclassified patients all having had the disease for more than three years. This suggests that the model could be useful for the early detection of PD, something that is particularly challenging for clinicians. Furthermore, analysis of the misclassified images suggests that they were misclassified due to the patient not pressing sufficiently hard against the tablet; see, for instance, Fig. 8, where parts of the drawing are not visible due to the low pressure values, obscuring the movement signal from the CNN. This issue could be mitigated against using preprocessing, or potentially by using a colour gradient rather than greyscale.

5 Discussion

This paper has approached the automated diagnosis of PD using drawing tasks and DL techniques under multiple configurations. The factors analysed, such as the effect of applying augmentation techniques, the resolution of the images, and the data representation used to create the images, show a rich and complex performance profile. One of the most crucial factors is the analysis of the classifiers trained with balanced and imbalanced data. The augmentation process has a very significant effect, improving considerably the diagnostic performance of the classifier. This is especially true for the most complex representations and when the cube task is used. The equal contribution of

Table 12 Average kappa and accuracy values and their SD for the best three imbalanced configurations

Best imbalanced CNNs	Avg Kappa ($\pm$ SD)	Avg Acc ($\pm$ SD)
Pentagon 32 $\times$ 32 0-pres greyscale	0.296 ($\pm$ 0.229)	69.00% ($\pm$9.38%)
Cube 32 $\times$ 32 0-pres greyscale	0.285 ($\pm$ 0.346)	70.18% ($\pm$15.71%)
Cube 32 $\times$ 32 no 0-pres B&W	0.227 ($\pm$ 0.337)	71.43% ($\pm$14.17%)

Table 13 Different metrics applied to the best CNN model

Metric	Validation set	Test set
Kappa	0.926	0.9
Accuracy (%)	96.31	95.02
Specificity (%)	98.11	98.31
Average precision	100	98

Table 14 Results from multiple performance metrics calculated for the best single CNN configuration

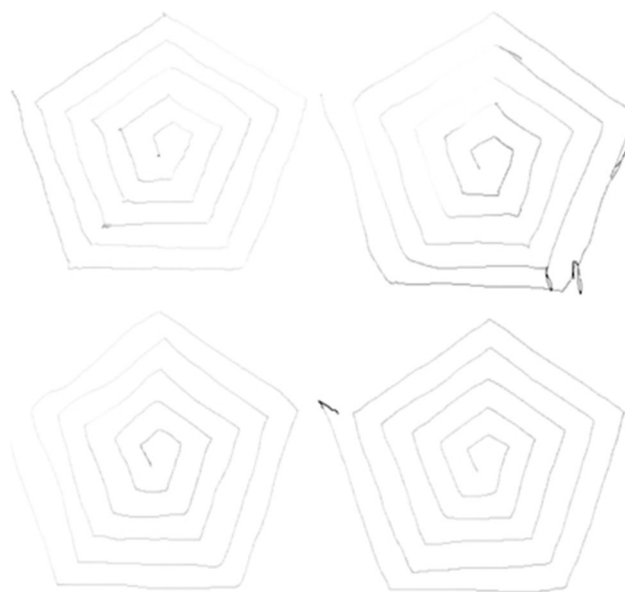|	Specificity	Sensitivity	F1-score	Support
Validation set				
Control	0.95	0.98	0.96	106
Patient	0.98	0.95	0.96	111
Avg/total	0.96	0.96	0.96	217
Test set				
Control	0.92	0.98	0.95	118
Patient	0.98	0.92	0.95	123
Avg/total	0.95	0.95	0.95	241

Table 15 Confusion matrix resulted from the best CNN model using images of 32×32 pixels of the pentagon task with zero-pressure information and greyscale representation

	Validation set		Test set	
	0 (Control)	1 (Patient)	0 (Control)	1(Patient)
0 (Control)	TN = 104	FP = 2	TN = 116	FP = 2
1 (Patient)	FN = 6	TP = 105	FN = 10	TP = 113

Fig. 8 Misclassified images: the top two were drawn by patients, and the bottom two were drawn by controls. Note that the lack of intensity reflects the absence of pressure when the subjects carried out the drawings

both classes in the learning process helps add robustness to the network. Subsequently, we agree with Pereira [66] that an imbalanced dataset negatively affects classification performance.

Augmentation also causes generality gain in DL models. In this context, numerous works did not report any mechanisms to increase the generality, such as augmentation techniques [57] or transfer learning [11, 18, 55, 57]. Due to the reduced size of all the datasets reviewed in this paper, if no hard measures against overfitting are implemented, high accuracy results can easily be a consequence of overfitting. Under these circumstances, this risk should be considered when comparing final reported results.

Regarding the size of the images, we observed that higher-resolution images (128×128) tend to reduce the

performance. However, this could be a direct consequence of the limited size of our CNN architecture. There is no dominating size with best results: images with 32×32 and 64×64 pixels showed approximately similar behaviour regarding PD discrimination. Their differences depend on other external factors like the type of drawing task. Pentagon and cube drawings could require different number of pixels to allocate the features required to perform an adequate PD classification. As a comparison, only two other papers investigated different resolutions. In [66] and later in [2], the same research team gathered metric values for 64×64 and 128×128 pixel-images and reported similar accuracy values, with slightly higher results for meander images of 64×64 in size and the opposite for the spiral, where 128×128 images outperformed a reduced 64×64 version [2]. The opposite behaviour was reported in [66], using both a 8-layer CNN architecture.

Our results indicate that the subject's movement signals, when the pen was in contact with the tablet, were insufficient to fully differentiate between PD patients and healthy controls. If we focus on the role of the non-pressure data in the classification, this information can be very effective to boost the performance, above all in the case of the cube task with black and white images. We consider that the planning and visual-spatial reasoning involved in constructing a three-dimensional cube might be significant factors to identify PD patients, which helps in reaching higher performance. This mechanism is not present in the pentagon dataset, which is a two-dimensional figure that is usually drawn without raising the pen. Adding the full

range of pressure values was important, especially for the pentagon task since it did not benefit much from including in-air information. For the cube task, pressure information contributes to its performance as much as the in-air information. Both characteristics together aid the cube task to reach a performance that is very close to the best pentagon configuration.

Previous works give a mixed view of which test is most discriminative for PD. Pereira et al. [64] attributed differences in performance between drawing tasks (meanders and spirals) to their complexity. They claimed that the hardest test (in their case the spiral) was more discriminative. However, the same authors, in their next work [66], drew the opposite conclusions, achieving better results with the meander task. Other authors such as [1, 70], also agreed on the superiority of the meander drawings using the same NewHandPD dataset. Regarding the PaHaW dataset, Moetesum et al. [55] also found the spiral task more effective than seven other handwritten tasks, the opposite to Drotár et al. [21] who also considered the same dataset. These authors also mentioned that results can be influenced by the features under consideration or how the data is represented.

From the analysis of our results, we conclude that our two tests have similar capabilities to distinguish PD patients from healthy controls. However, each of them needs different information included in the representation of the images: the pentagon drawing bases more its accuracy on pressure information and the cube on in-air movements.

Direct comparison of our results against previous studies poses some problems. The papers of Pereira [64–66] related to the use of a CNN for classifying PD, proposed an alternative accuracy metric to deal with imbalanced data [60], two drawing tasks (meanders and spiral), that differ from our selected tests and alternative sensors. In [64, 66], the images were extracted from scanned tests that include also the trace of the template and in [65], they used a very imbalanced dataset (18 controls–74 PD patients) with samples collected with a biometric pen. Moetesum et al. [55] applied a similar data acquisition method to the present work, creating images using only the X and Y coordinates and in-air information. However, it was not explicitly mentioned if the in-air information was represented differently than the areas where the pen did not pass. Apart from that, their dataset was balanced, using a traditional accuracy metric to measure the performance. Overall, the results from this approach can be more directly compared with the outcomes reported here. Finally, the ParkinsonHW dataset [39] used in [11, 27] can also be considered, in the same way, similar to our dataset but with a significantly more imbalanced number of samples and reduced size (62

PD patients and 15 healthy subjects) and without in-air information.

Our best performance result over tenfold cross-validation, 93.53%, calculated using a traditional accuracy metric, see Table 11, is almost as good as the best performance reported by Pereira et al. [65] (95.74%), using an ensemble classifier, and Canturk [11] (94%) with a more complex CNN architecture. It is significantly better ($\approx 10\%$ improvement) than the performance included in the work of Moetesum [55], using both different CNNs and fusion techniques, and the accuracy reported by Vasquez et al. [86] if the results for spiral data are only considered (67.1%). The same is the case for Afonso [1–3] and Diaz [18], with reported accuracies lower by 5–10%. However, the accuracy achieved in our work is less than [57], whose complex fine-tuned-ImageNet and AlexNet approach reached 98.28% accuracy. A comparative summary can be seen in Table 16. However, it should be noted that comparing approaches based on published accuracies is problematic, since it does not account for differences in the datasets, and differences in the ways in which models are assessed, both of which are likely to dominate over small numerical differences in the performance metrics.

It is arguable that many of the published methodologies are already sufficient in terms of accuracy, especially given the low diagnostic accuracies achieved by many human raters. Nevertheless, accuracy is only part of the picture and, for a model to be useful in practice, it must meet the broader requirements of clinical diagnosis. One of these is the burden placed on the subject. Many models reported in the literature require a patient to undergo multiple tests in order to generate the required data: for instance, works based on the PaHaW dataset, and other multimodal approaches, like [86]. Whilst the use of fusion techniques that integrate data from multiple tests for each patient may be advantageous in terms of accuracy, sourcing this data could be very difficult for patients with significant movement impairment. Our approach, by comparison, requires only a single drawing. A second advantage of our approach is the simple architecture used in our CNN. We transfer the complexity to the representation of the data instead of the DNN architecture, and consequently this requires less data and computational power to be properly trained. A further advantage of the relatively small size of our CNN is that it is more likely to generalise to unseen data than other DL models found in the literature. We also improve the robustness and generalisation of our results by implementing augmentation techniques like in [57], comparing multiple combinations of configurations, and studying the robustness of the results in terms of variance.

Table 16 Comparison of the major characteristics or the best results reported in the works reviewed in this paper

Year	Reference	DL Model	DA	TL	EN	Accuracy
2016	[64]	CNN (Caffe 5 layer)				90.38% (Spiral)
2016	[66]	CNN (ImageNet—5 layers)				87.14% (Meander)
2017	[3]	Deep-hierarchical optimum-path forest				83.79 (± 2.51)%
2018	[65]	CNN (Imagenet)			Yes	73.41($\pm$3.66)% (Spiral, 128 × 128) and 95.74($\pm$1.6)% (Ensemble, 128 × 128)
2018	[27]	10-layered deep ESN				89.3%
2018	[42]	CNN inspired by AlexNet (simplified version)				88%
2018	[55]	CNN to extract visual features used in a SVM, ensemble		Yes	Yes	83% (ensemble), 62% (Spiral)
2018	[86]	CNN (11 layers) for speech and gait, CNN (1 dimension, 4 layers) for handwriting			Yes	97.6% (ensemble), 67.1% (only handwriting—14 tasks)
2019	[2]	CaffeNet (8 layers—1-GPU version of AlexNet), CIFAR-10 (5 layer), LeNet (5 layers)				88.05% (Meander, 64 × 64)
2019	[70]	Bidirectional Gated Recurrent Units with an attention mechanism				92.24 ($\pm$2.65)% (Meander)
2019	[29]	CNN inspired by AlexNet (simplified version)—6 layers				96.5%
2019	[18]	Three CNNs (VGG16—16 layers) with different data representation as feature extractor		Yes	Yes	86.67% (ensemble, 150 × 150) and Spiral (75%, SVM)
2020	[1]	Deep optimum-path forest classifier				83.79 ($\pm$2.51)% (Meander)
2020	[74]	CNN—ResNet32	Yes	Yes		98.2%
2020	[81]	ESN				Pearson's correlations 0.839 to 0.89
2020	[11]	CNNs (25 and 144 layers) as a feature extractor for k-NN and SVM		Yes		94% (SVM)
2020	[57]	CNN (AlexNet)	Yes	Yes		98.28%
2021	[79]	A restricted Boltzmann machine for feature extractor to a fuzzy optimum-path forest				79.57 ($\pm$1.5)% (Meander, 128 × 128)
2021	[75]	ESN and LSTM	Yes			93.7% (ESN) (Pentagon)
2021	This work	CNN (19 layers)	Yes			93.53 ($\pm$1.43)% (Pentagon, 32 × 32)

Column DA shows papers where data augmentation has been reported, column TD where transfer learning (pre-trained networks) was used and EN where an ensemble classifier was implemented

6 Conclusions and future work

This work investigates the potential for using deep learning within the clinical assessment of PD. Wire cube and pentagon spiral drawing tasks, both designed to assess the motor and visuospatial capabilities of patients with neurodegenerative conditions, were performed by subjects with and without PD. Whilst they performed these tasks, their movements were digitised on a graphics tablet. The resulting dataset was used to train a CNN deep learning architecture, which achieved an accuracy of 93.53% when discriminating PD subjects from healthy controls on previously unseen data. Significantly, our method requires less data than most DL models used elsewhere in the literature, potentially reducing the burden on patients during the course of undergoing clinical assessment. It is also considerably simpler, meaning that it is more likely to

generalise to new data and is more amenable to behavioural analysis. In the course of this work, we have explored the effect of augmentation techniques, different data representations and different image resolutions on the performance of trained CNN models, finding all of these to have significant effects upon the discriminative ability of the deep learning system.

The limitations of this study include (1) its proof-of-concept nature; (2) the interpretability of the results typical of using a black-box optimisation approach; (3) the relative small size of the dataset and its imbalanced nature. Although the accuracy of the model is competitive against other approaches, and at a level that is likely to be clinically useful, there is likely further scope for improvement, for instance, a broader search for other, perhaps more innovative, CNN configurations, the use of more complex data representations that encode more information, the

implementation of transfer learning, and the use of more complex augmentation techniques.

In future work, we aim to investigate other DL models. Notably, deep RNNs are able to work directly on time series data, and could potentially be used to analyse dynamical aspects of a patient's drawing, like in [27]. However, there are certain obstacles that need to be overcome to use these approaches practically, including the development of suitable augmentation techniques. We also intend to examine whether we can extract useful knowledge from trained deep learning models, with the aim of understanding the basis of their discrimination by interpreting the features that the models use to classify PD patients. Additionally, we aim to investigate whether the developed models can give more information about disease staging (as done in [26]) and disease prognosis, for example whether they can differentiate between patients with and without cognitive impairment. These new experiments are expected to be supported by the gathering of new drawing data in the clinical environment.

Declarations

References

1. Afonso LC, Pereira CR, Weber SA, Hook C, Falcão AX, Papa JP (2020) Hierarchical learning using deep optimum-path forest. J Vis Commun Image Represent 71:102823

2. Afonso LC, Rosa GH, Pereira CR, Weber SA, Hook C, Albuquerque VHC, Papa JP (2019) A recurrence plot-based approach for Parkinson's disease identification. Future Gener Comput Syst 94:282–292

3. Afonso LCS, Pereira CR, Weber SAT, Hook C, Papa JP (2017) Parkinson's disease identification through deep optimum-path forest clustering. In: 30th conference on graphics, patterns and images (SIBGRAPI). IEEE, pp 163–169

4. Alty JE, Cosgrove J, Jamieson S, Smith SL, Possin KL (2015) Which figure copy test is more sensitive for cognitive impairment in Parkinson's disease: Wire cube or interlocking pentagons? Clin Neurol Neurosurg 139:244–246

5. Alty JE, Cosgrove J, Lones MA, Smith SL, Possin K, Schuff N, Jamieson S (2016) Clinically 'slight' bradykinesia in Parkinson's disease is accurately detected using evolutionary computation analysis of finger tapping. Mov Disord 31:S184–S184

6. Aly N, Playfer J, Smith S, Halliday D (2007) A novel computer-based technique for the assessment of tremor in Parkinson's disease. Age Ageing 36(4):395–399

7. Bousquet O, Elisseeff A (2002) Stability and generalization. J Mach Learn Res 2:499–526

8. Bu XY, Luo XG, Gao C, Feng Y, Yu HM, Ren Y, Shang H, He ZY (2013) Usefulness of cube copying in evaluating clinical profiles of patients with Parkinson disease. Cogn Behav Neurol 26(3):140–145

9. Cahn-Weiner DA, Williams K, Grace J, Tremont G, Westervelt H, Stern RA (2003) Discrimination of dementia with Lewy bodies from Alzheimer disease and Parkinson disease using the clock drawing test. Cogn Behav Neurol 16(2):85–92

10. Camps J, Sama A, Martin M, Rodriguez-Martin D, Perez-Lopez C, Arostegui JMM, Cabestany J, Catala A, Alcaine S, Mestre B et al (2018) Deep learning for freezing of gait detection in Parkinson's disease patients in their homes using a waist-worn inertial measurement unit. Knowl-Based Syst 139:119–131

11. Canturk I (2020) Fuzzy recurrence plot-based analysis of dynamic and static spiral tests of Parkinson's disease patients. Neural Comput Appl 33: 349-360

12. Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531

13. Chollet F et al (2015) Keras: Deep learning library for Theano and Tensorflow. 7(8) https://keras.io/k

14. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46

15. Cormack F, Aarsland D, Ballard C, Tovée M (2004) Pentagon drawing and neuropsychological performance in dementia with Lewy bodies, Alzheimer's disease, Parkinson's disease and Parkinson's disease with dementia. J Geriatr Psychiatry 19(4):371–377

16. Derkinderen P, Dupont S, Vidal JS, Chedru F, Vidailhet M (2002) Micrographia secondary to lenticular lesions. Mov Disord 17(4):835–837

17. Dexter DT, Jenner P (2013) Parkinson disease: from pathology to molecular disease mechanisms. Free Radical Biol Med 62:132–144

18. Diaz M, Ferrer MA, Impedovo D, Pirlo G, Vessio G (2019) Dynamically enhanced static handwriting representation for Parkinson's disease detection. Pattern Recogn Lett 128:204–210

19. Dorsey E, Constantinescu R, Thompson J, Biglan K, Holloway R, Kieburtz K, Marshall F, Ravina B, Schifitto G et al (2007) Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. Neurology 68(5):384–386

20. Drotár P, Mekyska J, Rektorová I, Masarová L, Smékal Z, Faundez-Zanuy M (2014) Analysis of in-air movement in handwriting: a novel marker for Parkinson's disease. Comput Methods Programs Biomed 117(3):405–411

21. Drotár P, Mekyska J, Rektorová I, Masarová L, Smékal Z, Faundez-Zanuy M (2016) Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease. Artif Intell Med 67:39–46

22. Duffy J, Keith R, Shane H, Podraza B (1976) Performance of normal (non-brain injured) adults on the porch index of communicative ability. In: Conference in clinical aphasiology, pp 32–42. BRK Publishers

23. Efron B (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 78(382):316–331

24. Fahn S, Sulzer D (2004) Neurodegeneration and neuroprotection in Parkinson disease. NeuroRX 1(1):139–154

25. Fleiss JL, Levin B, Paik MC (2013) Statistical methods for rates and proportions. Wiley, London

26. Frid A, Manevitz LM, Mosafi O (2018) Kohonen-based topological clustering as an amplifier for multi-class classification for Parkinson's disease. In: International conference on the science of electrical engineering in Israel (ICSEE), pp 1–5. IEEE

27. Gallicchio C, Micheli A, Pedrelli L (2018) Deep echo state networks for diagnosis of Parkinson's disease. In: 26th European symposium on artificial neural networks, pp 397–402

28. Gibb W, Lees A (1988) The relevance of the Lewy body to the pathogenesis of idiopathic Parkinsons disease. J Neurol Neurosurg Psychiatry 51(6):745–752

29. Gil-Martín M, Montero JM, San-Segundo R (2019) Parkinson's disease detection from drawing movements using convolutional neural networks. Electronics 8(8):907

30. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th international conference on artificial intelligence and statistics, pp 249–256

31. Goodfellow I, Bengio Y, Courville A, Bengio Y (2016) Deep learning, vol 1. MIT Press, Cambridge

32. Greenspan H, Van Ginneken B, Summers RM (2016) Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. Trans Med Imaging 35(5):1153–1159

33. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

34. Hinton G, Deng L, Yu D, Dahl G, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P et al (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag 29(6):82–97

35. Hollmén J, Skubacz M, Taniguchi M (2000) Input dependent misclassification costs for cost-sensitive classifiers. WIT Trans Inform Commun Technol https://doi.org/10.2495/DATA000481

36. Hughes AJ, Daniel SE, Ben-Shlomo Y, Lees AJ (2002) The accuracy of diagnosis of Parkinsonian syndromes in a specialist movement disorder service. Brain 125(4):861–870

37. Hughes AJ, Daniel SE, Blankson S, Lees AJ (1993) A clinico-pathologic study of 100 cases of Parkinson's disease. Arch Neurol 50(2):140–148

38. Hughes AJ, Daniel SE, Kilford L, Lees AJ (1992) Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. J Neurol Neurosurg Psychiatry 55(3):181–184

39. Isenkul M, Sakar B, Kursun O (2014) Improved spiral test using digitized graphics tablet for monitoring Parkinson's disease. In: International conference on e-health and telemedicine, pp 171–5

40. Kaul S, Elble R (2014) Impaired pentagon drawing is an early predictor of cognitive decline in Parkinson disease. Movem Disord 29(3):427

41. Kawa J, Bednorz A, Stpie P, Derejczyk J, Bugdol M (2017) Spatial and dynamical handwriting analysis in mild cognitive impairment. Comput Biol Med 82:21–28

42. Khatamino P, Cantürk İ, Özyılmaz L (2018) A deep learning-CNN based system for medical diagnosis: an application on Parkinson's disease handwriting drawings. In: 6th international conference on control engineering and information technology (CEIT). IEEE, pp 1–6

43. Krizhevsky A, Sutskever I, Hinton G (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105

44. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436

45. Lemley J, Bazrafkan S, Corcoran P (2017) Smart augmentation learning an optimal data augmentation strategy. IEEE Access 5:5858–5869

46. Lesage S, Brice A (2009) Parkinson's disease: from monogenic forms to genetic susceptibility factors. Hum Mol Genet 18(R1):R48–R59

47. Letanneux A, Danna J, Velay JL, Viallet F, Pinto S (2014) From micrographia to Parkinson's disease dysgraphia. Mov Disord 29(12):1467–1475

48. Mahmud M, Kaiser MS, McGinnity TM, Hussain A (2021) Deep learning in mining biological data. Cogn Comput 13(1):1–33

49. Mao X, Li Q, Xie H, Lau RY, Wang Z, Smolley SP (2018) On the effectiveness of least squares generative adversarial networks. IEEE Trans Pattern Anal Mach Intell 41(12):2947–2960

50. Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD (2008) Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. Neural Netw 21(2–3):427–436

51. Michaeli S, Öz G, Sorce DJ, Garwood M, Ugurbil K, Majestic S, Tuite P (2007) Assessment of brain iron and neuronal integrity in patients with Parkinson's disease using novel MRI contrasts. Move Disord 22(3):334–340

52. Mikołajczyk A, Grochowski M (2018) Data augmentation for improving deep learning in image classification problem. In: International interdisciplinary PhD workshop (IIPhDW), pp 117–122. IEEE

53. Miller DB, O'Callaghan JP (2015) Biomarkers of Parkinson's disease: present and future. Metabolism 64(3):S40–S46

54. Moetesum M, Siddiqi I, Ehsan S, Vincent N (2020) Deformation modeling and classification using deep convolutional neural networks for computerized analysis of neuropsychological drawings. Neural Comput Appl 32:1–25

55. Moetesum M, Siddiqi I, Vincent N, Cloppet F (2018) Assessing visual attributes of handwriting for prediction of neurological disorders—a case study on Parkinson's disease. Pattern Recogn Lett 121:19–27

56. Movement Disorder Society Task Force on Rating Scales for Parkinson's disease: the unified Parkinson's disease rating scale (UPDRS): status and recommendations. Move Disord 18(7):738–750 (2003)

57. Naseer A, Rani M, Naz S, Razzak MI, Imran M, Xu G (2020) Refining Parkinson's neurological disorder identification through deep transfer learning. Neural Comput Appl 32(3):839–854

58. Nasreddine Z, Phillips N, Bédirian V, Charbonneau S, Whitehead V, Collin I, Cummings J, Chertkow H (2005) Montreal cognitive assessment MoCA brief screening tool for mild cognitive impairment. J Am Geriatr Soc 53(4):695–699

59. Olesen J, Gustavsson A, Svensson M, Wittchen H, Jönsson B, Group CS, Council EB (2012) Economic cost of brain disorders in Europe. Eur J Neurol 19(1):155–162

60. Papa JP, Falcao AX, Suzuki CT (2009) Supervised pattern classification based on optimum-path forest. Int J Imaging Syst Technol 19(2):120–131

61. Parkinson Society: Website of the Parkinson's disease society. http://www.parkinsons.org.uk (2018). Accessed on 23-07-2021

62. Parkinson Study Group (2004) Levodopa and the progression of Parkinson's disease. N Engl J Med 351(24):2498–2508

63. Parkinson's Foundation: Statistics on Parkinson's: who has Parkinson's? https://www.parkinson.org/Understanding-Parkinsons/Statistics (2015). Accessed on 23-07-2021

64. Pereira C, Pereira D, Papa J, Rosa G, Yang X (2016) Convolutional neural networks applied for Parkinson's disease identification. In: Machine learning for health informatics. Springer, pp 377–390

65. Pereira C, Pereira D, Rosa G, Albuquerque V, Weber S, Hook C, Papa J (2018) Handwritten dynamics assessment through convolutional neural networks: an application to Parkinson's disease identification. Artif Intell Med 87:67–77

66. Pereira C, Weber S, Hook C, Rosa G, Papa J (2016) Pereira C, Weber S, Hook C, Rosa G, Papa J (2016) Deep learning-aided Parkinson's disease diagnosis from handwritten dynamics. In: Conference on graphics, patterns and ages. IEEE, pp 340–346

67. Pirlo G, Diaz M, Ferrer M, Impedovo D, Occhionero F, Zurlo U (2015) Early diagnosis of neurodegenerative diseases by handwritten signature analysis. In: Conference on image analysis and processing (ICIAP). Springer, pp 290–297

68. Post B, Merkus MP, de Bie RM, de Haan RJ, Speelman JD (2005) Unified Parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? Move Disord 20(12):1577–1584

69. Reeve A, Simcox E, Turnbull D (2014) Ageing and Parkinson's disease: why is advancing age the biggest risk factor? Ageing Res Rev 14:19–30

70. Ribeiro LC, Afonso LC, Papa JP (2019) Bag of samplings for computer-assisted Parkinson's disease diagnosis based on recurrent neural networks. Comput Biol Med 115:103477

71. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 234–241

72. Samii A, Nutt JG, Ransom BR (2004) Parkinson' disease. Lancet 363(9423):1783–1793

73. Saunders-Pullman R, Derby C, Stanley K, Floyd A, Bressman S, Lipton RB, Deligtisch A, Severt L, Yu Q, Kurtis M et al (2008) Validity of spiral analysis in early Parkinson's disease. Offic J Move Disord Soc 23(4):531–537

74. Seedat N, Aharonson V, Schlesinger I (2020) Automated machine vision enabled detection of movement disorders from hand drawn spirals. In: 2020 IEEE international conference on healthcare informatics (ICHI). IEEE, pp 1–5

75. Shenoy AA, Lones MA, Smith SL, Vallejo M (2021) Evaluation of recurrent neural network models for Parkinson's disease classification using drawing data. In: 43rd annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE

76. Shin MS, Park SY, Park SR, Seol SH, Kwon JS (2006) Clinical and empirical applications of the Rey–Osterrieth complex figure test. Nat Protoc 1(2):892

77. Shorten C, Khoshgoftaar T (2019) A survey on image data augmentation for deep learning. J Big Data 6(1):60

78. Smits EJ, Tolonen AJ, Cluitmans L, van Gils M, Conway BA, Zietsma RC, Leenders KL, Maurits NM (2014) Standardized handwriting to assess bradykinesia, micrographia and tremor in Parkinson's disease. PLoS ONE 9(5):e97614

79. de Souza RW, Silva DS, Passos LA, Roder M, Santana MC, Pinheiro PR, de Albuquerque VHC (2021) Computer-assisted Parkinson's disease diagnosis using fuzzy optimum-path forest and restricted Boltzmann machines. Comput Biol Med 131:104260

80. Stanley K, Hagenah J, Brüggemann N, Reetz K, Severt L, Klein C, Yu Q, Derby C, Pullman S, Saunders-Pullman R (2010) Digitized spiral analysis is a promising early motor marker for Parkinson disease. Parkin Rel Disord 16(3):233–234

81. Szumilas M, Lewenstein K, Ślubowska E, Szlufik S, Koziorowski D (2020) A multimodal approach to the quantification of kinetic tremor in Parkinson's disease. Sensors 20(1):184

82. Tucha O, Mecklinger L, Thome J, Reiter A, Alders G, Sartor H, Naumann M, Lange K (2006) Kinematic analysis of dopaminergic effects on skilled handwriting movements in Parkinson's disease. J Neural Transm 113(5):609–623

83. Turner RS, Desmurget M (2010) Basal ganglia contributions to motor control: a vigorous tutor. Curr Opin Neurobiol 20(6):704–716

84. Ünlü A, Brause R, Krakow K (2006) Handwriting analysis for diagnosis and prognosis of Parkinson's disease. In: International symposium on biological and medical data analysis. Springer, pp 441–450

85. Vallejo M, Jamieson S, Cosgrove J, Smith SL, Lones MA, Alty JE, Corne DW (2016) Exploring diagnostic models of Parkinson's disease with multi-objective regression. In: Symposium series on computational intelligence (SSCI). IEEE, pp 1–8

86. Vásquez-Correa JC, Arias-Vergara T, Orozco-Arroyave JR, Eskofier B, Klucken J, Nöth E (2018) Multimodal assessment of Parkinson's disease: a deep learning approach. J Biomed Health Inform 23(4):1618–1630

87. Wang Q, Hopgood JR, Finlayson N, Williams GO, Fernandes S, Williams E, Akram A, Dhaliwal K, Vallejo M (2020) Deep learning in ex-vivo lung cancer discrimination using fluorescence lifetime endomicroscopic images. In: 42nd annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 1891–1894

88. Wang SH, Phillips P, Sui Y, Liu B, Yang M, Cheng H (2018) Classification of Alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling. J Med Syst 42(5):85