The cloudUPDRS app: A Medical Device for the Clinical Assessment of Parkinson's Disease

C. Stamate[a], G.D. Magoulas[a], S. Kueppers[a,b], E. Nomikou[c,d], I. Daskalopoulos[a], A. Jha[d], J.S. Pons[e], J. Rothwell[d], M.U. Luchini[b], T. Moussouri[c,d], M. Iannone[e], G. Roussos[a]

[a] *Birkbeck College, University of London*
[b] *Benchmark Performance Ltd*
[c] *Audience Focus Ltd*
[d] *University College London*
[e] *Re:technica Ltd*

Abstract

Parkinson's Disease is a neurological condition distinguished by characteristic motor symptoms including tremor and slowness of movement. To enable the frequent assessment of PD patients, this paper introduces the cloudUPDRS app, a Class I medical device that is an active transient non-invasive instrument, certified by the Medicines and Healthcare products Regulatory Agency in the UK. The app follows closely Part III of the Unified Parkinson's Disease Rating Scale which is the most commonly used protocol in the clinical study of PD; can be used by patients and their carers at home or in the community unsupervised; and, requires the user to perform a sequence of iterated movements which are recorded by the phone sensors. The cloudUPDRS system addresses two key challenges towards meeting essential consistency and efficiency requirements, namely: (i) How to ensure high-quality data collection especially considering the unsupervised nature of the test, in particular, how to achieve firm user adherence to the prescribed movements; and (ii) How to reduce test duration from approximately 25 minutes typically required by an experienced patient, to below 4 minutes, a threshold identified as critical to obtain significant improvements in clinical compliance. To address the former, we combine a bespoke design of the user experience tailored so as to constrain context, with a deep learning approach based on Recurrent Convolutional Neural Networks, to identify failures to follow the movement protocol. We address the latter by developing a machine learning approach to personalize assessments by selecting those elements of the test that most closely match individual symptom profiles and thus offer the highest inferential power, hence closely estimating the patent's overall score.

1. Introduction

Parkinson's Disease (PD) is a degenerative neurological condition associated with a wide spectrum of motor and cognitive symptoms including tremor, slowness of movement and freezing, muscular stiffness, poor postural stability, sleep-related difficulties, depression and psychosis [24]. The underlying cause of PD is the degeneration of the so-called *dopaminergic neurons*, that is, a small group of neurons located in the mid-brain that are the main source of dopamine in the human central nervous system [11]. Dopaminergic neurons play a crucial role in the control of many brain functions including voluntary movement, mood, reward, addiction, stress and in particular, in the reward system that controls learning. Although the cause for the loss of these neurons is unknown, their selective degeneration leads to PD and its distinctive presentation.

Care for patients with PD involves the management of both motor and non-motor symptoms as well as palliative care. Since there is no cure, symptom management is a life-long process that affects not only the patients but also their families and carers. Clinical care pathways include pharmacological treatment corresponding to the exact stage of the disease, physiotherapy, and surgery [45]. As a result of medication with L-Dopa, a key element of the typical pharmacological regime for PD, patients are expected to develop side effects such as dyskinesias [55]. Since symptoms vary greatly independently of treatment and PD progresses at different rates in different individuals, treatment requires regular clinical monitoring and medication adjustment.

There are over 130,000 people with Parkinson's in the UK and as many as one million in the US, each individual seen by a specialist doctor or nurse only once or twice a year, allowing only brief and intermittent assessment of the wide range of their motor and non-motor symptoms [48]. This is due to the fact that the application of clinical measures of disease progression is laborious and as a consequence costly, because they require the direct involvement of a member of the clinical team. Moreover, although clinical measure protocols are detailed and formally structured they nevertheless represent assesement at relatively coarse-grain granularity, typically not involving the use of specialised measurement instrumentation. Despite generally good internal consistency in the application of these measures, they still depend on subjective estimations of patient performance by the clinician. Collectively, these factors restrict opportunities to precisely quantify PD progression and the effectiveness of patient stratification: the limited availability of data concerning individual variability and actual symptom trends limits opportunities to adapt care to the needs of a particular individual at a specific time.

To address this challenge we developed cloudUPDRS, the first smartphone app to achieve certification as a Class I medical device by the Medicines & Healthcare products Regulatory Agency in the UK for the clinical assessment of the motor symptoms of Parkinson's. cloudUPDRS augments standard clinical care pathways by enabling daily assessments which lead to (i) more consistent and reliable care, (ii) early identification of problems such as medication side-effects, thus enabling earlier intervention, (iii) monitoring of individualised patient trends leading to more effective patient stratification, and (v) enabling patients to take ownership of their own care through non-pharmacological measures such as improved nutrition and physical therapy.

The cloudUPDRS system is based on the Universal Parkinson's Disease Rating Scale [18] and the Parkinson's Disease Questionnaire [25], and incorporates a cloud-based Big Data management and analytics service to generate objective and reliable assessments of motor performance. Patients use the app at home to record sensor measurements while performing a series of simple actions with each limb, such as tapping the screen to assess bradykinesia and holding the phone on their knee to assess resting tremor. The data captured by the phone is then used to calculate the clinical UPDRS score through the application of a biomedical signal processing pipeline. Additional longitudinal analytics are performed subsequently to enable trend analysis and patient stratification.

In this paper, we extend the work discussed in [57] to present two distinctive features of cloudUPDRS developed to address specific challenges related with care for PD patients, namely:

- A deep learning technique employing Recurrent Convolutional Neural Networks applied to sensor observations so as to assess compliance with the actions dictated by the UPDRS protocol. Combined with a bespoke user experience facilitated by the app, this technique can replace expert supervision while maintaining high-quality data collection.
- Personalised tests reducing the time required to carry out an assessment to less than 4 minutes. These so-called *quick tests* are created using machine learning to select a subset of UPDRS observations that closely estimate the motor performance of a particular patient.

In the following section we review research related to this work and in Section 3 summarise the state-of-the-art in current clinical practice for the assesmenet of PD. We proceed to report on key factors for patient compliance identified through user research in Section 4. We present the cloudUPDRS system in Section 5 and in Section 6 we report on how the process of certification as a medical device affetcs software development. We then present the details of the two techniques identified above in Sections 7 and 8 correspondingly.

2. Related Work

During tha past two years, several smartphone apps have been developed to address the different needs of PD patients including: the mPower app (`http://parkinsonmpower.org/`) developed for iOS by Apple (`http://researchkit.org/`) and Sage Bionetworks in the US [53]; the uMotif app developed with NHS SBRI Healthcare funding in the UK; the Wearable Companion app developed by the M.J. Fox Foundation and Intel; the mHP app for Parkinsons developed by myHealthPal; PD Dr by the Muhammad Ali Parkinson Center at the Barrow Neurological Institute inthe US [47]; the Verily app in collaboration with ParkinsonNet in Holland; and several others. In this list we do not include apps that rely solely on self-reporting using diaries as they offer no direct way to conduct objective measurements of performance. In contrast to standard clinical practice, these apps follow a High-Frequency pattern of assessment; this terminology reflects the fact that the apps are able to carry out measurements of most elements of motor and cognitive performance of PD patients multiple times per day or even continuously when wearables are used in addition to a smartphone [31]. Nevertheless, none of the above apps has been certified as a medical device and their stated purpose is to assist research into PD or support self-quantification as a lifestyle choice for patients. Data collected by these apps cannot be used for clinical purposes.

In particular, the two major projects by the M. J. Fox Foundation listed above explore the diversity of PD motor symptoms within a large population sample. The first employs the mPower app and aims to develop a large data set of motor performance observations, which can be used as a benchmark for the experimental evaluation of algorithms providing PD diagnosis. Clearly, long-term research in PD necessitates the development of such open data sets however the approach adopted depends solely on self-reporting to discriminate between patient and non-patient data and confirm compliance with the prescribed data collection protocol, and as such it is limited by the fact that these cannot be verified objectively.

The second M.J. Fox Foundation project is carried out in collaboration with Intel and The Grove Foundation, and employs wearables to provide 24×7 monitoring of PD patients. Specifically, a Pebble watch (`https://www.pebble.com/`) is provided to participants to measure wrist tremor relayed via an Android app to a Cloudera-based back end for storage and analysis. The stated goal of this study is the development of a deep longitudinal data set capturing in minute detail the second-by-second variations of motor symptoms from a population of tenths of thousands of volunteers. However, battery longevity and data transmission issues have set considerable challenges in attempts to capture complete traces and the project has explored alternative strategies. Moreover, progress towards the automatic identification of PD-related tremor episodes has been limited by practical difficulties, such as the problem of accurately interpreting raw acceleration data captured from a single body location especially when lacking contextual information. Moreover, the ambitious target of harvesting a complete high-resolution data stream from the wearable sets very considerable challenges for battery longevity as devices have fewer opportunities to switch to low-power mode.

In [27], we demonstrated the feasibility of using smartphones as a means to assess commonly occurring motor symptoms of PD in a clinical setting. Specifically, we design, develop and validate in a clinical study a prototype app on Android implementing Part III of the MDS-UPDRS [18]. Using the accelerometer and touch screen sensors commonly available in modern smartphones, we are able to carry out hand and leg tremor measurements, as well as gait and bradykinesia assessments using finger tapping tasks to replicate the majority of these tests (cf. Table 1 and Section 5.2). Other research groups have followed a similar approach focusing on specific symptoms. Most commonly tremor measurements are considered for example [12], [29], [34] and [35] all provide proof-of-concept implementations of upper limb tremor estimation.

Recently, the suitability of machine learning has been investigated for the assessment of PD. Voice samples are processed using standard machine learning algorithms in [4] to correlate individual performance and MDS-UPDRS score. A deep learning approach is adopted in [20] to identify patients in ON and OFF states using Restricted Boltzmann Machines to analyse accelerometre data; and in [13] for the detection of bradikynesia. Both projects report encouraging results which merit further investigation but current performance limitations prevent these techniques from becoming an effective clinical tool.

3. Motor Performance Assessment in PD

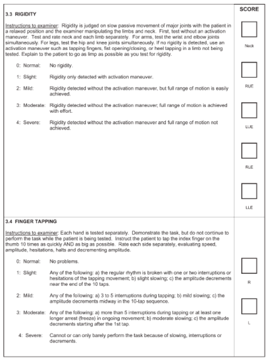
Currently there is no definitive test for the diagnosis of PD and only a post-mortem examination can confirm that reported Parkinsonian symptoms were actually caused by PD. Instead, specialists look for common signs of PD and offer a diagnosis only after other conditions with similar presentation have been excluded. The basis for the diagnosis is the patient's medical history and an examination which typically explores motor performance during various tasks. Often, a diagnosis is not confirmed until after medication for PD is prescribed and an improvement is recorded following a few weeks or months of administration. Often, patients suspect of sufferring from PD will undergo MRI, CT or SPECT brain scans, however these are used to identify and exclude other syndromes with similar presentation rather for the diagnosis of PD as such.

During an examination the specialist would typically ask the patient to perform various tasks to assess the agility of arms and legs, muscle tone, gait and balance and record the results into a table, namely the Unified Parkinsons Disease Rating Scale (UPDRS). UPDRS is a universal scale of PD symptoms and used to comprehensively assess and document the exam. The purpose of this record is so that clinicians be able to compare it with the patient's future follow up visits, or to communicate about the progression of PD symptoms in each patient with other neurologists. UPDRS was initially introduced in 1987 and significantly revised in 2008 as a result of extensive consultation coordinated by the Movement Disorder Society (MDS) [18]. Although the latter is more accurately referred to as the MDS-UPDRS, the term UPDRS is often used to refer to either. In this paper we are only considering the MDS-UPDRS variant and for this reason will use both terms interchangeably.

The MDS-UPDRS is a comprehensive 50-question assessment of both motor and non-motor symptoms associated with PD. It features four different sections, referred to as Parts I to IV, that focus on:

I Non-motor experiences of daily living

II Motor experiences of daily living



SCORE

3.3 RIGIDITY

Instructions to examiner: Rigidity is judged on slow passive movement of major joints with the patient in a relaxed position and the examiner manipulating the limbs and neck. First, test without an activation maneuver. Test and rate neck and each limb separately. For arms, test the wrist and elbow joints simultaneously. For legs, test the hip and knee joints simultaneously. If no rigidity is detected, use an activation maneuver such as tapping fingers, fist opening/closing, or heel tapping in a limb not being tested. Explain to the patient to go as limp as possible as you test for rigidity.

0: Normal: No rigidity.

1: Slight: Rigidity only detected with activation maneuver.

2: Mild: Rigidity detected without the activation maneuver, but full range of motion is easily achieved.

3: Moderate: Rigidity detected without the activation maneuver; full range of motion is achieved with effort.

4: Severe: Rigidity detected without the activation maneuver and full range of motion not achieved.

□ Neck

□ RUE

□ LUE

□ RLE

□ LLE

3.4 FINGER TAPPING

Instructions to examiner: Each hand is tested separately. Demonstrate the task, but do not continue to perform the task while the patient is being tested. Instruct the patient to tap the index finger on the thumb 10 times as quickly AND as big as possible. Rate each side separately, evaluating speed, amplitude, hesitations, halts and decrementing amplitude.

0: Normal: No problems.

1: Slight: Any of the following: a) the regular rhythm is broken with one or two interruptions or hesitations of the tapping movement; b) slight slowing; c) the amplitude decrements near the end of the 10 taps.

2: Mild: Any of the following: a) 3 to 5 interruptions during tapping; b) mild slowing; c) the amplitude decrements midway in the 10-tap sequence.

3: Moderate: Any of the following: a) more than 5 interruptions during tapping or at least one longer arrest (freeze) in ongoing movement; b) moderate slowing; c) the amplitude decrements starting after the 1st tap.

4: Severe: Cannot or can only barely perform the task because of slowing, interruptions or decrements.

□ R

□ L

Figure 1: Sample questions and scoring sheet for Part III of the MDS-UPDRS.

III Motor examination

IV Motor complications

The protocol also includes the specification of a decision tree process that clinicians are required to follow in order to assign a particular score to each question (each assessment must often be carried out twice, considering the left and right side separately) after exploring each question during a brief discussion with the patient or their carer (cf. Figure 1). The purpose of this provision is to ensure the internal consistency of the rating scale and limit the effects of subjective judgments by the person performing the assessment. The questions are not presented directly to the patient as they use medical terminology which may not be clear to them. The MDS holds the copyright of the scale and its use requires a ratings scales permissions request form to be completed and submitted to the MDS as well as the payment of licensing fees where applicable.

Of particular significance is Part III of the MDS-UPDRS as it is considered the most objective and thus reliable part of the scale. Notably, the European Medicines Agency (EMA) recognizes Part III as accepted scales to measure the efficacy of a drug for PD. Specifically, the disease progression marker employed in clinical trials measures an improvement of PD under a new drug as the observed Change from Baseline of the Part III score. The score ranges between 0, corresponding to no symptoms, to 56, corresponding to maximum effects typically representing full immobility. Although there is no generally accepted score which in isolation would be adequate to lead to a PD diagnosis, different patient groups representing different levels of symptom severity would often be organized along the following boundaries [52]:

1. Mild PD: Part III UPDRS score 20 or below.
2. Moderate PD: Part III UPDRS score from 21 to 35.
3. Severe PD: Part III UPDRS score greater than 35.

In addition to symptoms caused by PD, chronic L-Dopa use often eventually leads to a brittle response to the medication - sometimes the medication fails to work and the patient remains frozen and unable to move (the so-called OFF state), sometimes the effect of the mediation is well-balanced leading to the so-called ON state, and sometimes the medication effect is quickly overpowering, causing excessive movements called L-Dopa induced dyskinesias (LID). It is very common for a patient with mid-stage PD to fluctuate wildly between these extremes throughout the day. Multiple treatment strategies are available for these complications, including changes to medication, subcutaneous administration of apomorphine, intra-jejunal administration of L-Dopa or deep brain stimulation (DBS), a continuous electrical stimulation of a surgically implanted electrode in the brain.

While Part III of the UPDRS provides a numeric score based on the examination by a member of the clinical teeam, the 39-item Parkinson's Disease Questionnaire (PDQ39) [25] is a self-reported measure of health status and quality of life. The questionnaire assesses how often people affected by PD experience difficulties across eight dimensions of daily living namely mobility, activities of daily living, emotional well-being, stigma, social support, cognition, communications and bodily discomfort, as well as specific dimensions of functioning and well-being. Similar to the MDS-UPDRS the PDQ39 is under copyright and its use requires a license from Oxford University Innovations.

Further, Hoehn & Yahr (H&Y) is a clinical rating scale introduced in 1967 that defines categories of motor function in PD, ranging from minimal or no functional disability at level 0 to confinement to bed unless aided at Level 5. However, the H&Y scale has several problems including the fact that it is not linear so that its modern use is mainly as a means to describe patients groups rather than quantify disease progression especially in an epidemiological setting. For example, the severity of PD symptoms progress rapidly in diagnosed patients with the median time taken to transit between H&Y stage 2 to 3 and 3 to 5 being 87 and 50 months respectively. Older age at diagnosis and higher MDS-UPDRS motor scores at baseline (both increasingly prevalent as the result of an aging population) are associated with faster PD progression.

Finally, we note that cost of care increases manifold as PD progresses with annual medical costs for the NHS for H&Y Stage 5 patients estimated at £30,000 per person per year at inflation adjusted prices, with average cost of £9,500 across all categories. This represents a total cost for PD of over £1.25 Billion per annum today, projected to increase to over £1.6 Billion by 2020 (in 2014 prices). Furthermore, these costs relate only to direct medical care which represents only 7% of the total costs of PD, with the remainder 93% split between direct non-medical professional care (50%) and indirect informal care (43%).

4. Understanding Patients with PD

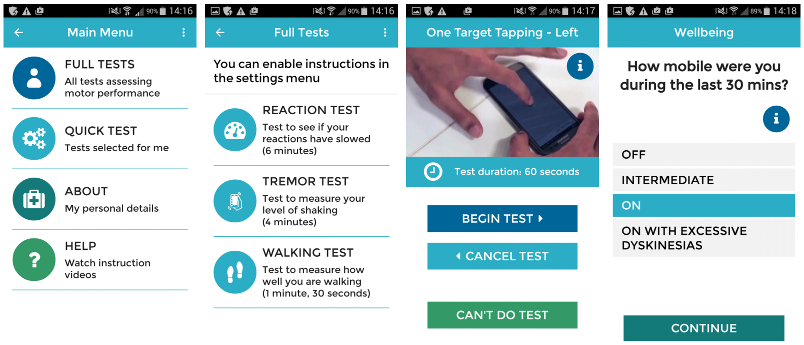
The wider adoption of cloudUPDRS by patient communities necessitates that tests are incorporated as part of their daily routine. To understand how to best facilitate this we carried out extensive interviews with clinicians, technologists, patients, carers and patient advocates (22 individuals in total); a web survey with participants from the research volunteer pool of Parkinson's UK receiving 166 unique submissions; and, three audience panels (16 participants in total). Across all studies we recruited participants with a confirmed diagnosis of PD and excluded individuals with generic symptoms of Parkinsonism. Patient participants represent all H&Y levels except for the audience panels in which participation was limited to Level 3, due to the practicalities of access to the venue.

The potentially transformative role of smartphone apps for PD was widely acknowledged in interviews. The expectation of positive outcomes was closely related to recent trends enabling the direct involvement of patients in establishing research priorities, the use of patient expertise in research, and towards greater transparency. This perspective was often related to opportunities for patient empowerment as expressed for example in online communities such as PatientsLikeMe [10], suggesting that evidence-based care must cater for the translation of evidence into practice in a manner directly accessible to and understandable by patients.

We employed the web survey to explore current phone usage patterns specifically among patients and to identify potential constraints that may place barriers for the adoption of the cloudUPDRS app. Responses received were primarily from mobile phone users (96%) with 77% coming from those with a smartphone. The majority of smartphone owners (87%) use it daily with only 14% reporting significant difficulties. A relative small proportion of those with smartphones (20%) use apps to track their symptoms or manage medication. The vast majority (86%) expects to make regular use of the cloudUPDRS app with 64% expressing a preference for the test session lasting a maximum of 5 minutes, 27% accepting a test duration of 10 minutes, and 5% even longer. The majority (68%) expect to make use of the app at least once per day to assess their symptoms.

One aspect of the app design investigated in the survey is the provision of feedback, especially considering the fact that results would inevitably indicate a decline in performance over time. Nevertheless, over 87% of respondents considered receiving direct feedback a key advantage of using the app despite the expectation of a negative trend. A small number of respondents suggested that emphasis should be on positive outcomes instead: "I don't want my decline to be the focus rather I'd prefer to have something that promotes my wellbeing! I use the Speech Tool to remind me to speak louder and clearer" (Female, 45-54).

Audience panels combined elements of user experience evaluation and a wider exploration of perceived costs and benefits of the cloudUPDRS app, which was


Main Menu
FULL TESTS
All tests assessing motor performance
QUICK TEST
Tests selected for me
ABOUT
My personal details
HELP
Watch instruction videos
Full Tests

You can enable instructions in the settings menu

REACTION TEST
Test to see if your reactions have slowed (6 minutes)
TREMOR TEST
Test to measure your level of shaking (4 minutes)
WALKING TEST
Test to measure how well you are walking (1 minute, 30 seconds)
One Target Tapping - Left
Test duration: 60 seconds
BEGIN TEST
CANCEL TEST
CAN'T DO TEST
Wellbeing

How mobile were you during the last 30 mins?

OFF
INTERMEDIATE
ON
ON WITH EXCESSIVE DYSKINESIAS
CONTINUE

Figure 2: Views of the user interface of the cloudUPDRS app showing session management, tremor recording and finger tapping activities..

demonstrated during the sessions. Panelists identified specific problems with the version presented, for example the potential effects of involuntary movements common to specific patient profiles and suggested improvements. As relates to the utility of the app and their motivation for regular use, the opportunity to manage symptoms was an unequivocal benefit for the majority of participants and strongly motivated their involvement.

However, access to detailed performance data was considered less important compared against the sense of understanding offered by the experience of using the app and the associated sense of control over the disease which this experience afforded. In particular, the set of recorded data was seen as a reminder of the changes in the patients' life caused by PD which they viewed as the basis for the development of an externally validated personal narrative. An exception to this is a small group of patients who appear committed to self-quantification and already collected and organised self-tracking data prior to their involvement with this study. This group valued the ability of the app to make accurate observations higher than its role as an aide-de-mémoire. All participants identified with the strong desire to make a contribution towards defeating the disease, and considered their contribution of personal data and their open availability for research as a means to achieve this goal. As a consequence, no privacy concerns were expressed.

5. The cloudUPDRS System and app

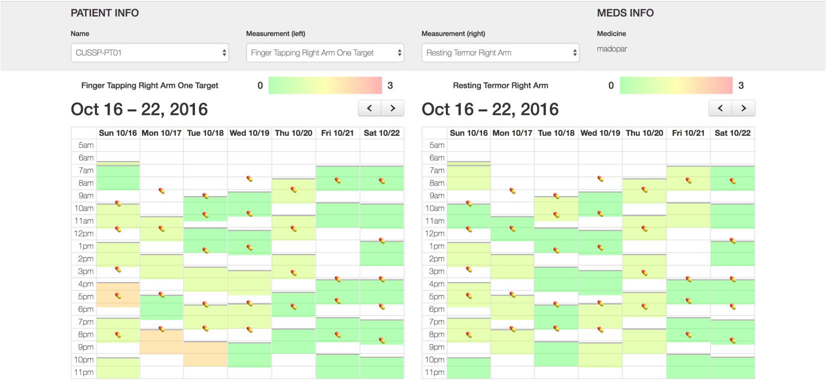
The cloudUPDRS app implements a comprehensive workflow (partially depicted in Figure 2) that provides audio, video and textual media to guide patients and their carers to conduct the tests at home and in the community with no requirement for supervision by a member of the clinical team. To provide full functionality the app communicates with a data management and processing back-end that enables aggregation and longitudinal analytics. Overall, the complete cloudUPDRS system consists of:

1. A smartphone app for Android that enables patients to carry out motor performance tests and complete a wellness self-assessment; conduct session management; and securely submit data to the cloudUPDRS service.
2. Cloud-based scalable data collection service that ingests data from patients' smartphones; ensures secure data management; and applies the signal processing pipeline.
3. Data-mining toolkit for medical intelligence incorporating quantitative and semi-structured data, and longitudinal analyses, clustering and classification; and a clinical user interface incorporating visualisation.

5.1. Presentation and Service Platfrom

The cloudUPDRS app implements a bespoke user interaction design to ensure that the data recorded capture the actual motor performance features as required for the successful application of MDS-UPDRS. Specifically, patients are guided through a carefully orchestrated sequence of actions while the app records sensor measurements. By requiring the execution of specific action sequences the app restricts the degrees of freedom of individual movement and thus imposes structure and disambiguates user context by limiting the range of observed behaviours.[1] As a result, the recorded signal can be interpreted accurately using a small set of heuristics rather than require the use of a full context model

[1] The action sequences can be seen in video demonstrations available at `http://www.updrs.net/help`.



PATIENT INFO

Name: CUSSP-PT01

Measurement (left): Finger Tapping Right Arm One Target

Measurement (right): Resting Termor Right Arm

MEDS INFO

Medicine: madopar

Figure 3: cloudUPDRS clinician dashboard: The specific view of the dashboard allows the comparison of motor performance measurements computer directly by cloudUPDRS (left) against self-reported assessments using PDQ39, which clearly show a significant divergence. The overlaid red dots on both sides of the dashboard represent the reported times of medication administration.

and reasoning approach [5]. Finally, the app automatically adapts to match the specifications of its host device and incorporates a delay tolerant service to manage data upload.

The full test administered by the cloudUPDRS app consists of 17 individual observations, specifically kinetic, postural and resting tremor for the left and right hand; left and right leg agility and resting tremor; single and double target finger tapping on both sides; and, gait. During each observation period lasting 60 seconds, the patient is required to assume a specific position and perform the prescribed movement as described in the previous paragraph. Following the recording of these observations the patient is presented with a questionnaire incorporating selected questions from PDQ39 and recording the time of the most recent medication intake.

One of the key diagnostic benefits of the cloudUPDRS app is the ability to conduct longitudinal studies. To this end, we have developed a clinician dashboard depicted in Figure 3 which allows the investigation of aggregated data over time for individual patients as well as over selected patient groups. For example, a heatmap visualisation is used to show the hour-by-hour changes in motor performance over a period of a week with a view to provide an overall assesement of disease progression.

The cloudUPDRS service is engineered to facilitate scalable performance by adopting the microservices architecture [44]. This approach is set in contrast to traditional monolithic web applications and aims to maximise opportunities for vertical decomposition and scaling-out, which are critical for high performance and service resilience in data intensive situations. cloudUPDRS microservices are implemented as composite Docker containers, are loosely coupled and employ lightweight communication and coordination mechanisms such as the Consumer-Driven Contract pattern. System componentization is enforced via versioning of published RESTful interfaces and sandboxed instances of the service can be deployed automatically to cater for data isolation between distinct regulatory domains. cloudUPDRS microservices are deployed as docker containers (cf. `https://www.docker.com/`) although internal implementation details vary to match the specific preferences and expertise of project partners responsible for their implementation and their suitability for the task in hand. For example, while the data collection and signal processing APIs are implemented using python and django REST within an nginx/gunicorn container, semi-structured longitudinal analytics are implemented as Ruby bundles. The overall service architecture has been designed for scalability so that real-time streams captured for example during concurrent patient consultations can be integrated on the fly with archival information from the longitudinal datastore service. To facilitate this *modus operandi*, we provide structured workflows implemented through microservices following the lambda architecture [39], which facilitates the efficient fusion of real-time and archival data on the fly.

5.2. Bio-signal Processing

Precise assessment of tremor, bradykinesia and gait is typically carried out using laboratory equipment for example tailor-made biomedical data acquisition systems incorporating transducers such as high-frequency/high-accuracy accelerometers and gyroscopes, signal amplifiers and filters and high-performance

Table 1: Analytics toolbox signal processing functions and correspondence to the sections of the MDS-UPDRS.

Analytic Function	MDS-UPDRS Section
Rest Tremor	3.17 (rest tremor amplitude)
Postural Tremor	3.15 (postural tremor of the hands)
Action Tremor	3.16 (kinetic tremor of the hands)
Pronation—supination Movements	3.6 (pronation—supination movements of the hands)
Leg agility	3.8 (leg agility)
Finger tapping	3.3 (rigidity) & 3.4 (finger tapping)
Gait	3.10 (gait) & 3.11 (freezing of gait)

analog-to-digital converters. The captured signal is analysed subsequently by specialist commercial software such as Spike 2 by Cambridge Electronic Design Ltd with the total cost of a complete system rising to tenths of thousands.

Laboratory based clinical rating however is constrained by the requirement that the patient is present in the clinic, and in practice can only be carried out as a "snap-shot" assessment. In [27] we show that the sensor, clock and data acquisition hardware of a low-end smartphone capture data with sufficient accuracy to precisely quantify the magnitude of PD motor symptoms across the majority of the tests included in Part III of the MDS-UPDRS by comparing its performance against results obtained using a biomedical analytics system by CED. In cloudUPDRS we automate the methodology presented in [27] as a bespoke cloud-based data analytics service [15]. For completeness of presentation, we briefly summarise the main features of this system here.

5.2.1. Tremor

Tremor measurements are recorded for both hands at rest, at posture and in action as listed in Table 1. For rest tremor measurements, users are asked to relax their hands on their lap in a supine position while the phone is lying in their palm. For the postural tremor measurements patients are guided to keep their arm outstretched directly on their front while holding the smartphone. Finally, for action tremor measurements they are required to hold the phone and move it between the chest and the fully outstretched position on their front. In all cases, acceleration is recorded along three axes in $\mathrm{m/s^2}$ at the maximum supported sampling rate (at least 50 Hz) and timestamped at maximum resolution (typically microseconds). Tremor is calculated as the cumulative magnitude of the scalar sum acceleration across three axes for all frequencies between 2 Hz and 10 Hz. To obtain this power spectrum the signal is first filtered with a Butterworth high-pass second order filter at 2 Hz and the Fast Fourier Transform (FFT) is subsequently applied to the filtered waveform data.

5.2.2. Bradykinesia

MDS-UPDRS assess bradykinesia, or else the slowness of movement, through three different factors: (i) pronation-supination movements, (ii) leg agility, and (iii) finger tapping. In the first test patients are asked to hold the phone and perform alternating pronation-supination movements, that is rotating the palm of the hand toward the inside so that it is facing downward and then toward the outside so that the palm is facing upward, as fast and as fully as possible. Leg agility measurements require the phone to be placed on the thigh of the patient while seated, holding the phone lightly with the ipsilateral hand, while raising and stomping the foot on the ground as high and as fast as possible. During both tests the phone is recording acceleration data in a manner similar to the tremor tests. The assessment of the pronation-supination movements and leg agility tests requires the estimation of the frequency and power of movement. To obtain these, the toolkit first removes DC offset and applies a Butterworth low-pass second order filter at 4 Hz in order to exclude most of the tremor. Subsequently, the power of the movement is calculated as the total amplitude between 0 Hz and 4 Hz and the frequency derived from the power spectrum.

Finger tapping performance is assessed in two tests using single and dual targets presented on the screen of the phone at set locations with patients attempting to tap them as fast and as accurately as possible (alternating between targets in the dual-target case). When tapping accidentally occurs outside the screen area the test is repeated. The touch-sensitive screen of the smartphone is used to collect the information used for performance calculations, specifically the timing of each touch event, its duration, the direction of movement (upwards or downwards), the coordinates on the phone screen, and the amount of pressure applied are recorded. For the two-target variant it is necessary that the distance between targets be at a specific distance irrespective of the size of the screen or of the device. To estimate finger tapping performance the analytical functions first identify all touch events and employ the associated timestamps to estimate tap frequency (taps per second), the mean hand movement time between taps (in milliseconds), and the actual movement distance between alternative tapings in the dual-target case (in centimetres).

5.2.3. Gait

Mds-updrs assesses gait by considering multiple behaviours including stride amplitude and speed, height of foot lift and heel strike, and turning and arm swing [63]. The cloudUPDRS variant of this test requires the patient to walk along a straight line for five meters, turn around and return to the point of departure, while the smartphone is positioned either in their belt or trousers pocket. Since it is only possible to measure acceleration data from a single point at the waistline we employ the techniques in [36, 37] to estimate stride frequency and length, velocity and turning time.

However, in comparisons against assessment by an experience clinician presented in [27] individual metrics were only weakly correlated to the corresponding Section 3.10 mds-updrs score. Since it is not possible to capture detailed information about the movement of the leg relative to the foot and the arms, the metrics calculated simply identify characterises of body types. Although perhaps not as useful in the context of calculating the mds-updrs score this data is still of interest for the exploration of new digital biomarkers related for example to freezing that can be useful for the development of disease progression indicators and thus we consider this feature to merit further investigation.

6. Certification

As noted in Section 2, there are several research, wellness and self-tracking apps for PD available on both major smartphone platforms and many more released selectively for research. The vast majority of these apps do not conform to the safety, quality, performance and regulatory requirements set for medical devices and as such can only be employed either to encourage a healthy lifestyle or for research purposes correspondingly — but are not tools that can be used to support medical diagnosis. This fact is often explicitly reflected in their terms and conditions of use for example, quoting from a popular PD app "we cannot, and thus we do not, guarantee or promise that you will personally receive any direct benefits."

Medical devices are regulated and must conform to rules enforced by regional legislation. Within the European Union, harmonisation of regulations across member countries is facilitated by the Medical Devices Directive (MDD), which provide the blueprint for country–specific legislation. Although the MDD considers situations when software would be treated as a medical device it does not explicitly examine smartphone apps and so its provisions are open to interpretation, an issue that we address in this section. Further, the MDD requires that each member state establishes a Competent Authority to provide guidance and enforce regulation of medical devices and in the UK this responsibility lies with the Medicines and Healthcare products Regulatory Agency (MHRA).

Under Article 1 Clause 2(a) of the MDD a medical device is defined as "any instrument, apparatus, appliance, software, material or other article, whether used alone or in combination, including the software intended by its manufacturer to be used specifically for diagnostic and/or therapeutic purposes." The current interpretation of this definition by the MHRA as relating to apps implies that "if the [mobile] application is intended to carry out further calculations, enhancements or interpretations of entered/captured patient data, [$\cdots$] it will be a Medical Device. If it carries out complex calculations, which replaces the clinician's own calculation and which will therefore be relied upon, then it will certainly be considered a Medical Device." Hence, the features of the cloudUPDRS app clearly place it within the provisions of the MDD. For certification purposes, the named publisher of the app on the selected platform store is considered its manufacturer as defined by the MDD, and thus the party obliged to ensure conformity with the provisions of the directive.

Hence, according to the MDD the cloudUPDRS app is a Class 1 medical device that is, an active transient non-invasive instrument. Class 1 devices are considered lower risk and as such as less closely regulated. In this case, certification requires that the app meets the Essential Requirements outlined in Annex I of the MDD including:

i evidence of software development in compliance with ISO/IEC 62304,

ii comprehensive documentation ensuring that it can be used safely and appropriately by patients, and

iii implementation of quality management processes which ensure that the device can be safely marketed as a consumer product.

Moreover, any software developed under the provisions of the MDD should also comply with European privacy regulations, which in this case in particular necessitates the conduct of a Privacy Impact Assessment (PIA) according to the provisions of the code of practice detailed in the PIA Handbook published by the Information Commissioner's Office (ICO). Clearly, these requirements add considerable complexity, cost and overhead to the development process and in particular require that when a new version of the app is published a full set of conformity checks and software tests must be carried out, and updated documentation produced however small the changes. Certainly, they specifically exclude testing with users outside a strictly controlled setting for example, as common with lean development approaches. As a consequence, successful certification requires the investment of considerable additional effort and resources and although we have not precisely audited these effects, in the case of cloudUPDRS has clearly resulted in a longer software development period by several months.

It is interesting to consider each of these requirements in more detail. ISO/IEC 62304 is an international standard which specifies the life-cycle requirements for the development of medical software in general and software incorporated in medical devices in particular. The standard describes provisions according to the potential of the software to create a hazard that can potentially result in injury. cloudUPDRS is classified as Safety Class A, which represents a lower level of risk to health. Nevertheless, despite the relatively less stringent provisions for this class, the standard still requires a structured software development process with distinct phases prescribed for planning, requirements analysis, design and unit testing, which is often at odds with modern agile software development processes commonly employed for mobile apps.

The key enabler for cloudUPDRS to satisfy the ISO requirements while not sacrificing the considerable advantages of agile methods for mobile, is to employ software development tools that enforce structured workflows. For example, the permissions and change control elements of our development process enforced a specific sequence of steps to be taken so that patches are only applied after they have passed checks successfully. Further, we employed popular software development tools in such a way so as to automatically generate the required documentation at every iteration. The overall goal of this practice is to ensure traceability and transparency throughout the life-cycle of the software especially in dealing with software faults and before the incorporation of new features. Although this leads to a more cumbersome process from the developer point of view, this objective can be achieved through the use of collaborative design tools, issue trackers, reports, change control and testing workflows enforced through software automation. In cloudUPDRS we implemented such control, review, approval and documentation mechanisms using a variety of common industry-standard tools including Atlassian Jira, Pivotal Tracker, github and Circle CI which together can provide the required level of auditing and automation.

The second requirement for certification however cannot be automated as it relates to the implementation of a risk management process to determine the safety of the medical device and must be carried out by the manufacturer throughout the product life-cycle. In the context of the MDD, risk management would commonly be carried out according to the provisions of ISO 14971. From the software development perspective, compliance requires regular, typically weekly, meetings of a risk assessment panel incorporating technical, medical and administrative representatives. The role of this panel is to consider the implications of current and planned developments, develop a mitigation strategy and monitor acceptable hazards that cannot be completely eradicated. The documentation produced automatically from the software development process described above provides significant input to the discussions among panel members but additional issues raised by individual members with responsibility for specific risk areas are also considered. The work of the panel results into a risk assessment document that identifies individual risks, the likelihood that they might occur and an assessment of their potential impact. Measures to alleviate these risks must also be considered and implemented as appropriate. The final output of this process is the maintenance of the so-called *risk management file* for cloudUPDRS.

Finally, the third requirement for certification is typically interpreted as meeting the provisions of ISO 13485 which specifies the quality management system that a manufacturer should meet. Although in the case of cloudUPDRS this added significantly to the overhead of certification due to the fact that for the manufacturer this was the first medical device registered, this requirements does not refer directly to the individual device but rather sets requirements at the organizational level, specifying specific processes to be in place, identifying quality assurance roles and commitments.

Last but not least, conducting the PIA has several technical implications relating to architecture design decisions. In particular, the PIA must clearly detail the information flows within the system and consider each step from the point of view of privacy and design mechanisms that address these risks. This requires the careful design of security provisions for example from the beginning of the project and their frequent review to follow changing project needs. In fact, this requirement represents best practice in software development and the PIA should simply provide an explicit reminder and record of the decisions made. Where design decisions do not fully address patient needs a mitigation plan should be developed. In the case of cloudUPDRS, conducting the PIA upfront paid dividends later in the project as it represents a key document required to obtain approval for clinical studies. Having said that, in May 2018 the new General Data Protection Regulation (GDPR) will come into force in the EU, which has considerable long-term implications for cloudUPDRS. While a full discussion of the architectural, operational and organizational modifications required to become GDPR-compliant is beyond the score of this paper, it is nevertheless appropriate to point out that it has far reaching technical implications for example to support the right to be forgotten and especially data portability.

cloudUPDRS received medical device status in the UK, and thus in the EU, in May 2016.

7. Learning Test Movements

As noted in Section 3, disease progression assessments in PD are typically carried out bi-annually under the supervision of a qualified member of clinical staff who ensures that patients follow closely the actions dictated by Part III of the MDS-UPDRS protocol. This is also the case in clinical studies where in addition to providing supervision, clinical and research staff would also operate the equipment used to carry out performance measurements and subsequently process the data using specialist software. In Section 5, we discussed how the cloudUPDRS app extends this practice enabling a patient to carry out precise measurements of motor performance unsupervised using a smartphone. Indeed, when conducting self-assessments at home using cloudUPDRS supervision by an expert is neither readily available nor desirable as it would nullify the cost benefits of this approach. Nevertheless, in order to ensure accurate symptom

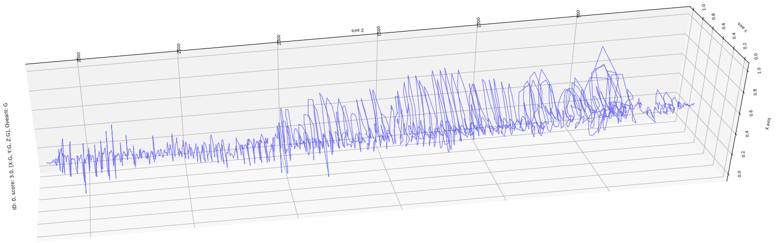


Figure 4: Typical tremor measurement trace representing a high quality observation.

assessment with cloudUPDRS it is necessary to establish whether the prescribed movements have been followed closely during data recording.

To address this lack of expert supervision and ascertain the fact that data has been captured under the appropriate circumstances, cloudUPDRS combines two methods that operate in tandem. Initially, the patient is guided by the user experience design presented in Section 5, which provides support in performing the actions accurately and steers the user through the process. While this approach has produced positive results, full compliance with the prescribed actions still cannot be guaranteed or confirmed. Hence, cloudUPDRS supplements this user interaction design with the development of a mechanism used to verify the quality of the data collected. Specifically, we introduce a deep learning methodology which aims to replace human supervision by providing a means to confirm that the recordings submitted have been captured while the patient performs the required actions correctly[2]. Failure to do so would produce bio-signal measurements that are not representative of the intended tremor type and are likely to result in erroneous scoring.

To achieve this, we adopt a deep learning methodology [19], in this case Re-

[2] In the case of the one- and two-finger tapping tests it is relatively straightforward to identify when the process has been followed accurately directly from the output of the bio-signal processing of Section 5.2.

current Convolutional Neural Networks [1], to enable the cloudUPDRS system to learn movement features associated with a high quality signal (cf. Figure 4 for a visual representation of the patterns of acceleration typically observed), and alert the user when an observation has not been captured under satisfactory conditions. Enabled by recent advances in general-purpose computing using graphics processing units and related algorithmic developments, this methodological approach employs multiple hidden layers to obtain notable results permitting neural networks to identify preferred features directly from the raw signals. This aspect of the selected methodology appears especially well-suited to the data quality issue under consideration.

The data set used to investigate the performance of this approach is taken from the first cohort of patients enrolled in the cloudUPDRS trials (8 male and 4 female). Specifically, we consider 227 distinct test sessions conducted over a period of three months (June to August 2016). Signals were collected from 9 different phone models providing acceleration measurements with a minimum sampling rate of 50 Hz, implemented using the data collection code base of the cloudUPDRS app (other source code elements not affecting data collection were modified during this period). Results are reported specifically for pronation-supination observations of the right hand, without loss of generality for the purposes of this paper. Data captured by the app are normalised but no other pre-processing is performed at this stage.

7.1. Rationale and Overview

To formulate an algorithmic solution, we re-frame the problem of captured data verification as one of binary classification. Specifically, the goal of the verification task is to discriminate between high-quality observations/signals and lower-quality sensor recordings captured during movements that do not closely adhere to the guidance of the MDS-UPDRS protocol. To this end, we employ a training data set of observations representing both acceptable and unsuitable cases with known data quality characteristics, guaranteed by the fact that they are collected by the app under controlled conditions or inspected manually.

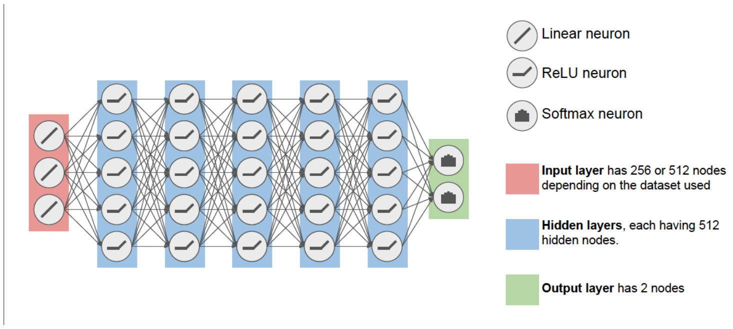

Linear neuron
ReLU neuron
Softmax neuron
Input layer has 256 or 512 nodes depending on the dataset used
Hidden layers, each having 512 hidden nodes.
Output layer has 2 nodes

Figure 5: DMLP neural network architecture employed by [57].

From this data set, features that are distinct within each class are identified algorithmically. Subsequently, the obtained representations are employed to test new observation data submitted by patients via the app: submissions classified as offering adequate quality are accepted and forwarded to the appropriate microservices for data ingestion and signal processing; otherwise they are rejected and excluded from further consideration.

Our initial experiments with this methodology were presented in [57] where we employed Deep Multilayer Perceptrons (DMLP), as the one shown in Figure 5, using the middle segment of the signal (Figure 6) to solve the classification problem. The work presented in this paper, extends [57] in two significant ways. First, the software has been re-implemented from first principles using TensorFlow [2] so that it can be fully incorporated into the app running on the mobile device rather than be applied at the service back-end. This is possible due to the fact that TesorFlow provides strong support for mobile platforms and because the classification process has two distinct stages: an initial model training phase representing the most computationally intensive task followed by a sample assessment phase which is relatively lightweight for modern smartphone hard-

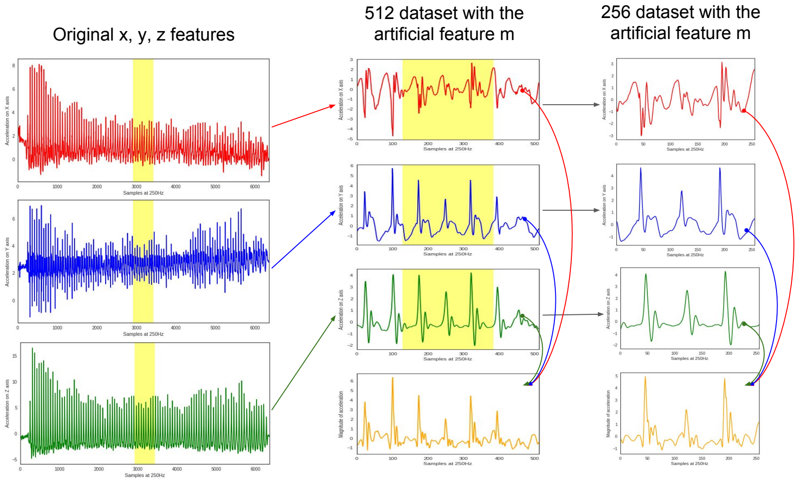


Figure 6: Inputs to the DMLP neural network architectures employed by [57] consist of signal segments along the x, y, z acceleration axes together with the acceleration magnitude m.

ware. As such, using TensorFlow the model can be constructed off-line using archival observation data for training and later incorporated in the app, which can conduct real-time quality assessments at the time of data recording and interactively request the repeat of specific individual observations as appropriate to ensure that all submitted tests are usable.

Second, although we obtained good performance using the approach presented in [57], the specific neural network architecture employed examines only a segment around the mid-section of the recorded signal (cf. Figure 6). Considering the significant changes observed in time during a complete observation trace as detailed in Section 8, notably the considerable drop in the power of the dominant tremor frequency for example, as depicted in Figure 11, it is clearly preferable to process the full trace for the duration of the test. To this end and to take into account the temporal aspect of the input signal, we introduce in this paper a novel deep learning architecture using Recurrent Convolutional Neural Networks (RCNN) [41].

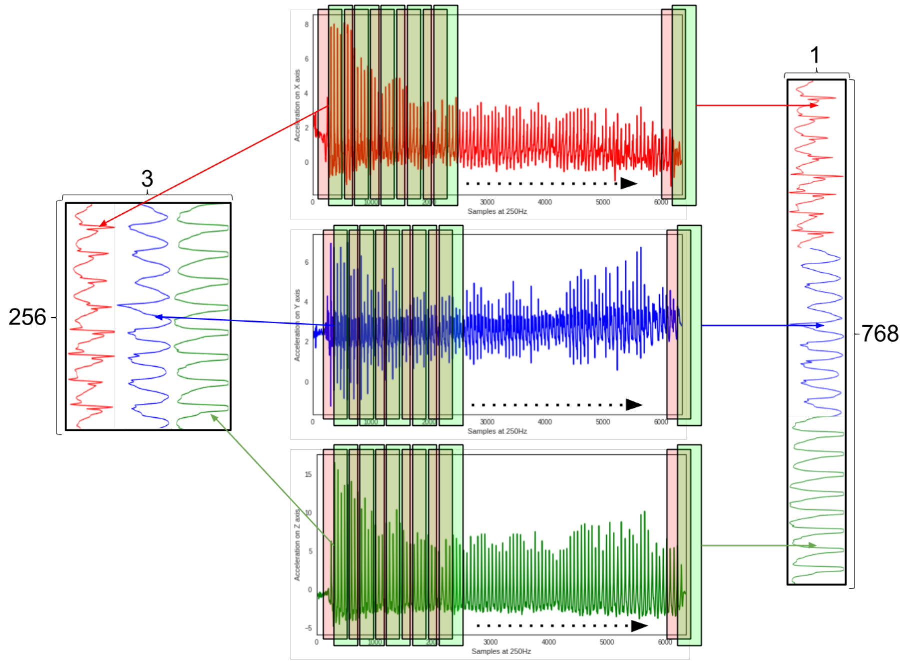


Figure 7: Sliding window of length 256 with 128-overlap applied on the recorded signals along the three acceleration axes to create the dataset used in this paper. Left side shows the dataset in a 256 × 3 matrix form to feed the RCNN. Right side shows the same data in a 768 × 1 vector form as required to feed the rest of the architectures.

7.2. Deep Learning Architecture

There are two main observations that guided the development of a new architecture to replace the work presented in [57]. First, in [57] each individual axis of acceleration is considered separately; that is four separate DMLPs are generated and trained for each of the x, y and z coordinates as well as the derived feature m representing the magnitude of acceleration (cf. Figure 6). Second, the input vector for each of the four neural networks constructed is a 256 or 512 long segment taken from the middle section of each signal trace rather than sampling the full signal (cf. Figure 6). As a consequence, the temporal features of the captured signal are not taken into consideration. At the time, both were reasonable options and in practice able to provide good performance.

In this paper we adopt an alternative approach, whereby we examine all three signal elements at the same time so as to unlock the full spatiotemporal characteristics of the complete signal. The practical implication of this is that the generated data takes the form of a matrix rather than vector, which is the typical formulation to feed a classifier. To reflect this change, we opt to replace the four DMLP architectures employed previously with one based on Convolutional Neural Networks (CNN) [68, 19], because they provide better performance with multidimensional (grid like) inputs. Moreover, to account for the temporal properties of the signal we employ Recurrent Neural Networks (RNN) [1], which are eminently suitable for processing temporal sequences. By using both convolution operations and recurrence relations we aim to exploit the whole range of information that the data has to offer, including local and temporal characteristics.

The synergy of CNN and RNN also provides another advantage over Multilayer Perceptrons (MLP), such as the DMLP used in [57]. It relates to the capability of MLP related architectures to deal with variations in the input space. For example, translation invariance relates to small changes in the input space which, however, require a lot of computation from MLPs to learn. This is because networks, like MLP and DMLP, have no shared weights or shared parameters, so small variations in a multidimensional input imply that a DMLP would have to learn all local features from scratch. In contrast, architectures that share weights and parameters, such as CNNs, are more robust to such variations in the input space.

7.2.1. Convolutional Neural Networks (CNN)

CNNs can detect and extract local informative patterns with fewer parameters than MLPs by iteratively traversing a smaller set of weights, called kernels, on n-dimensional input grids. This means that CNNs really shine when we have a grid or volume like input, and the features in this input space are locally correlated. CNNs are a good fit for the binary classification problem we are considering in this paper due to the fact that they exploit three distinct concepts, namely *sparse interactions*, *parameter sharing* and *equivariant representations.*

Sparse connectivity, or sparse weights, refers to the fact that the kernel used in convolution operations is much smaller than the input, thus making the weights w.r.t. the input quite sparse. For example, using a 3×3 kernel on a grid of inputs with size 256×3 requires only 9 connections for the detection of patterns and other meaningful features, as opposed to an order of magnitudes more connections for fully connected layers. By storing less parameters (weights) for our feature representation not only we optimize storage but we also improve the efficiency of the representation [19].

Parameter sharing, or tied weights, refers to the fact that CNNs are learning only one set of parameters for every location in the input, as opposed to MLPs where one parameter is required for each input part. Parameter sharing brings considerable improvements in storage as it is only necessary to store a few parameters corresponding to the weights, instead of a full matrix of parameters for a fully connected layer. As a consequence of parameter sharing, CNNs achieve equivariance to translation, essentially a time-line that highlights when specific features are present in the input. For example, when the same feature appears at different times in the input, the kernel produces the same representation in the output at the corresponding times.

7.2.2. Recurrent Neural Networks (RNNs)

RNNs are generally considered the best choice when dealing with sequences of symbolic, non-symbolic or mixed data, as happens for example in natural language processing, handwriting recognition and speech processing [1]. RNNs are more successful than fully connected or convolutional alternatives, because they operate efficiently over sequences and thus are not constrained by input size or fixed number of computational steps. Similar to CNNs that scan the multidimensional grid input and share weights locally, RNNs scan the input sequences and share the weights in time. Sharing weights across time makes RNNs sensitive not only to specific input patterns but also specific input sequences. Although many RNN variants have been explored in the literature [1], in this paper we draw inspiration from Recurrent Convolutional Neural Networks (RCNN) [41] to develop our particular architecture.

We follow [41] in our implementation, however our architecture has some distinct features. First, there is the extra addition of a fully connected layer before the network output, as depicted in Figure 8. Second, in contrast to [41] we do not use pooling layers or dropout as our investigation suggests that tremor appears to be sensitive to this type of stochastic noise injection. Furthermore, we opt for a modular topological approach, as suggested in [59, 21], which groups layers into modules so that they can be easily replicated across the whole network. We have used batch normalization after every convolutional layer, as seen in Figures 9 and 10, because it has the property to stabilize the gradients; thus, alleviating both the exploding and the vanishing gradients problems [1]. Lastly, recurrence happens in the convolutional layers according to Figure 10, where state is shared across convolutional steps and also the original input of the shared states is added to each of the steps. We have named this recurrence module Recurrent Convolutional Layer (RCL) and can be identified in Figure 10. Convolution BatchNorm ReLU (CBR) is also a key ingredient of the RCNN and can be seen in Figure 9.

To take advantage of the synergy between convolutional and recurrent networks we section the input space into equally sized chunks so that the CNN can fully exploit the local features. Time–order of these chunks is maintained so that the RNN can exploit the temporal aspect of the slices.

To get our data into the required form, a sliding window of length 256 with 128-overlap is employed on each individual session of observations out of 227. For the benchmark algorithms we have used a flattened version of the 256×3 input, as can be seen in Figure 7, in order to form an input vector. Previous experimentation on cloudUPDRS suggested that this choice outperforms other options, such as a length of 512 as reported in [57].

Moreover, our experiments suggest that the RCNN architecture that performs the best has the following structure:

- One input layer of size 256×3, followed by

- One CBR layer, followed by
- Two recurrent convolutional layers (RCLs) with a kernel size of 9×3 and 32 and 64 filters respectively, followed by
- One flattening layer where the layer is vectorized, followed by
- One fully connected layer of size 512 with Batch Normalisation and ReLU, followed by
- One softmax layer of size 2

Each RCL has 3 time steps with kernels of the same size as their input, as can be seen in Figure 10. All layers use the same activation function namely the Rectified Linear Unit (ReLU), which takes any input and produces the maximum value between the input itself and zero. This form of activation function is very popular because it helps alleviating the well known vanishing gradients problem and also creates sparse connections in the hidden layers of the network, limiting values to zero.

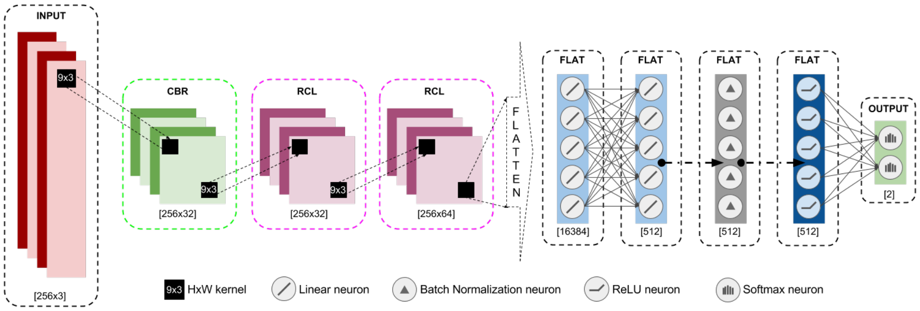


Figure 8: RCNN architecture with the size of each layer at the bottom in square brackets, where the notation $[n \times m]$ denotes the size of the matrix. The internal structure of the CBR and RCL modules is presented in Figures 9 and 10.

7.3. Classifier Training and Validation

Training the above architecture requires comparing the output $\hat{y}$ of the constructed deep network against the desired output y, which in the cloudUPDRS

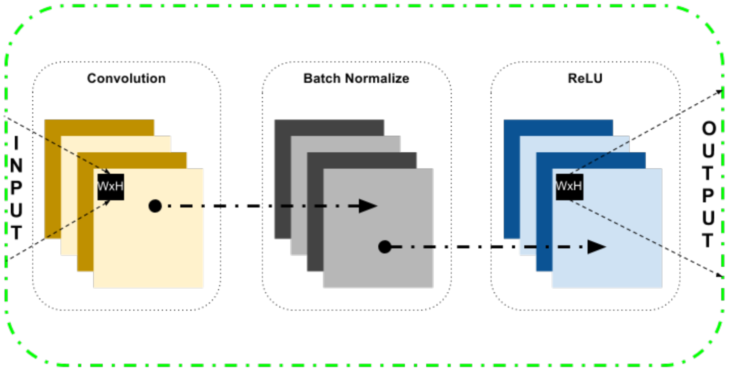


Figure 9: Convolution BatchNormalize and ReLU (CBR)

case represents the appropriate quality class label that the network should produce, i.e. accept or reject the signal window presented at the input. This information is used in the so-called cost, or objective, function that the RCNN aims to minimise. Here we adopt the *categorical cross-entropy* $\mathcal{L}$ as the objective function, defined as $\mathcal{L}(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$.

The final step in the process is the application of the backpropagation algorithm [62] which enables the network to learn the distribution that generated the training data. Backpropagation (BP) employs the chain rule to calculate the derivatives of the objective function with respect to each connection weight between neurons, and uses this information to update the weights. In this work we have used a variant of the standard BP: the Stochastic Gradient Descent (SGD) with momentum [58], as it has been shown by [65] that deep networks trained with this method are able to provide better generalization than other methods tested.

Training is carried out on the full signal across the three acceleration axes at the same time. As the three signal components may be of different lengths, as mentioned above, we have employed a sliding window of size 256 with 128 overlap ensuring that we have segments of the same length and also preserving the temporal aspect of the signal; thus making each the input a grid like structure (cf. Figure 7). As overlapping sliding windows are used, the order of the

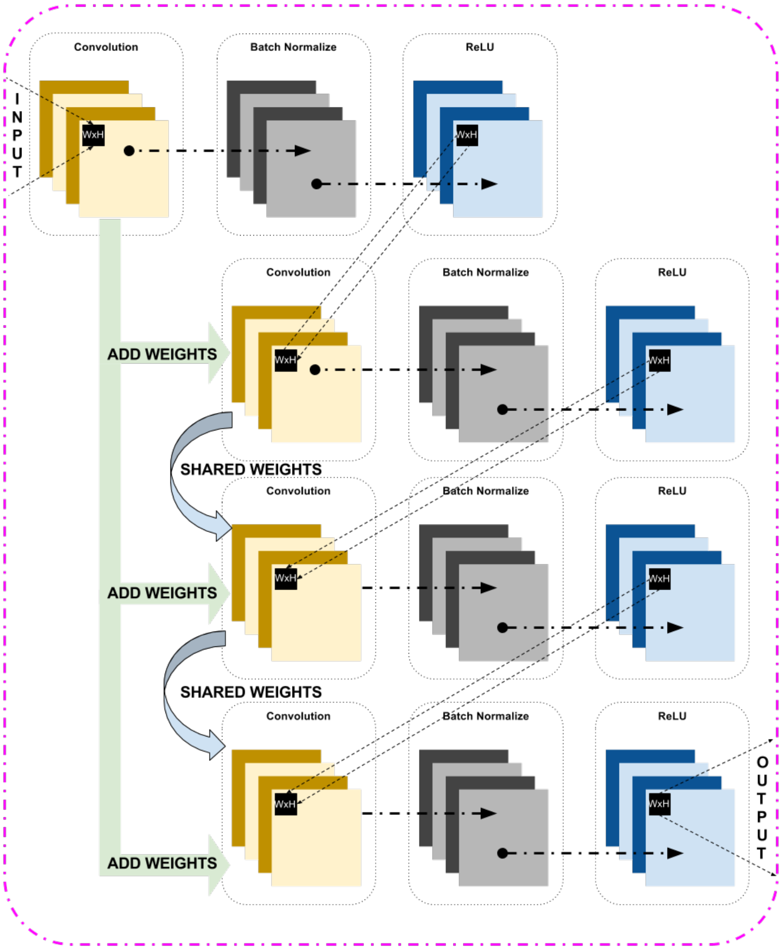


Figure 10: Recurrent Convolutional Layer (RCL)

windows is preserved when training so as to exploit the temporal features of the signal. Further, the so-called leave-one-out method [3] is combined with early stopping to asses the true predictive power each learning algorithm has on all of the data and to reduce the risk of overfitting. The choice of this approach reflects the fact that the data set under investigation was based on observations from 227 patient sessions, which is relatively low in this context.

Each iteration of leave-one-(session)-out process involves the exclusion of a single session from the data set, training the classifier on the remaining sessions and testing on performance on the omitted. Consequently, the RCNN of Section 7.2 is trained as many times as the sessions available in the data set, in this case 227 times. One limitation of this technique is that it can become biased on the weight initialization. To address this issue the process is repeated ten times using different initial random weights and the mean is used as the overall performance metric. Thus, the experiments summarized below are conducted using ten cycles of leave-one-(session)-out cross-validation, so that $2,270$ classifiers have been trained and averaged.

The early stopping heuristic applied ensures that the learning process is terminated when it reaches a certain predefined threshold. Specifically, we employ three criteria: (i) the categorical cross-entropy falls below 0.001; (ii) training classification success reaches 100%; or, (iii) the learning process has reached 200 iterations. The benefit of using early stopping is that it prevents the RCNN classifier from memorizing counter-productive characteristics discovered in certain samples, especially when these are spurious or irrelevant for the accurate determination of high versus low quality observations. This technique works well when used in conjunction to leave-one-out as it ensures that the RCNN is not over trained [42] on any part of the data set.

To validate the effectiveness of the cloudUPDRS approach we compare its performance against several well-established alternatives selected for their recent success in industrial systems or in highly-regarded competitions such as Kaggle. Full details are provided in Section 7.4 below.

7.4. Experiments and Results

The deep learning approach described in Sections 7.2 and 7.3 is implemented using the computational graph engine Tensorflow [2]. Training was carried out on an array of NVIDIA K40 GPUs achieving a 20-fold speedup against a standard multi-core CPU. To provide a baseline against which to evaluate our approach we compare its performance with the following classifiers implemented using the `scikit-learn` [49] machine learning library: (i) Gaussian Naive Bayes [46]; (ii) Bernoulli Naive Bayes [46]; (iii) Random Forest Classifier [9] which employs

Table 2: Classification reports with F1 score and Area Under the Precision-Recall curve (AUC).

Classifiers	Accuracy	F1-score	AUC
ExtraTrees	0.73	0.79	0.83
BernoulliNB	0.73	0.79	0.83
RandomForest	0.73	0.79	0.83
GradientBoosting	0.72	0.80	0.83
Bagging	0.72	0.78	0.83
AdaBoost	0.66	0.75	0.81
GaussianNB	0.69	0.75	0.83
DMLP	0.75	0.81	0.85
RCNN	**0.78**	**0.82**	**0.87**

an ensemble of random decision trees each selected from a sample drawn with replacement; (iv) Extra Trees Classifier [17] is a variation of random forest with thresholds randomly drawn for each candidate feature; (v) AdaBoost Classifier [70] is a meta-estimator which adjusts classifier weights so as to improve learning from difficult classes; (vi) Bagging Classifier [8] is also a meta-estimator which operates on random subsets of the training data to reach a final prediction by aggregating their results; and (vii) Gradient Boosting Classifier [16] which performs optimization of arbitrary differentiable loss functions. Furthermore we are benchmarking against the main algorithm from our previous work [57], the DMLP, which has been reproduced with the exact same settings as before, the only difference being the input layer, which is now 786 instead of 256 or 512.

As you can see from Table 2, the proposed RCNN–based approach outperforms all other methods on all the metrics used to assess overall classification success. The F1 score represents the harmonic mean between the precision and recall and AUC is the area under the precision-recall curve.

Further, the confusion scores from Table 3 shows that the proposed RCNN outperforms all the rest mostly because it is the only method that can actually

Table 3: Confusion matrix for the classifiers compared (averaged performance). Labels: T/F True/False and P/N Positive/Negative.

	TP	(%)	FN	(%)	TN	(%)	FP	(%)
ExtraTrees	141.52	93	8.98	6	13.36	17	63.14	82
BernoulliNB	146.23	**96**	4.27	**3**	6.92	8	69.58	91
RandomForest	138.39	91	12.11	8	16.19	20	60.31	79
GradientBoosting	146.02	**96**	4.48	**3**	8.12	10	68.38	89
Bagging	135.58	89	14.92	10	18.03	23	58.47	76
AdaBoost	128.0	84	22.5	15	17.34	22	59.16	77
GaussianNB	116.01	76	34.49	23	35.41	45	41.09	54
DMLP	135.73	89	15.77	10	28.19	37	49.31	63
RCNN	133.22	87	18.28	12	38.38	**50**	39.12	**50**

differentiate between positive and negative observations. We have to take into account that this is quite an unbalanced dataset, we have in total 5691 signal windows, out of which 3765 are positive and 1926 negative. If you look closely at the BernoulliNR and GradientBoosting, they have very good True Positive scores but at the cost of True Negatives. This means that they are very biased towards predicting positives, classifying low quality recordings as high quality useful signals.

In contrast, the RCNN produces more balanced performance, recognizing larger number of low quality recordings than other methods, and producing a lower value for the false positives compared to other classifiers. This is reflected in the overall RCNN performance, as shown in Table 2, in terms of its Accuracy, F1-score and AUC value.

8. The cloudUPDRS Quick Test

In this section, we turn our attention onto the development of methods that achieve significant reductions in test duration so as to enable patients to use cloudUPDRS on a daily basis. As suggested by the user studies summarised in

Section 4, the majority of patients identified a maximum of 5 minutes as the desirable duration for the test. However, even after the initial familiarisation period the full implementation of the procedure typically requires 25 minutes, an estimate that has been confirmed from system logs and independently through user feedback. The critical influence of test duration on user adoption rates was further confirmed during the initial three months of field testing. While the majority of participants carried out tests regularly during the first week following the commission of the app, compliance rates dropped sharply by the end of the third week, and only one out of the 12 participants continued to carry out tests at the end of the three-month testing period.

8.1. Test Duration and Characteristics

Recall from Section 5 that according to the MDS-UPDRS protocol each individual observation requires 60 seconds of recording and the full test consists of 17 observations, in addition to the medication and well-being questionnaire. Clearly, to reduce the overall duration of the test there are two main options namely to shorten the recording time for individual observations or to reduce the number of observations carried out by selecting a subgroup of the full 17-item set. The final questionnaire requires approximately 30 seconds and is always required because it is used to track medication.

First, consider the option to reduce the length of individual observations without loss of precision in the estimation of motor performance. Specifically, we investigate whether the 60 second observation period set by the MDS-UPDRS protocol is necessary or instead consistent scoring can be still maintained after significantly reducing its duration. To this end, we conduct observations of motor performance for alternative recording periods of 20 and 40 seconds and compare these against measurements carried out for the the full 60 seconds. Tremor and bradykinesia performance metrics were calculated for all observation types in our test data set consisting of 133 full tests carried out by 35 different individuals. In the remainder of this section we report scores calculated for tremor power at rest for the right hand, without loss of generality and so as to

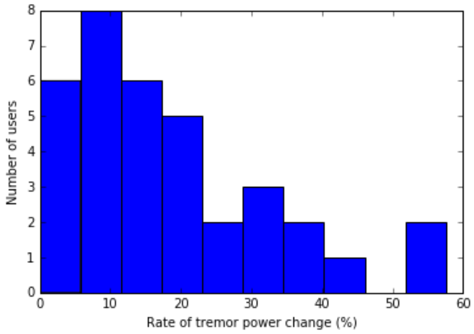


Figure 11: Change in recorded tremor power between tests of 60 and 20 seconds.

specifically quantify our findings.

Figure 11 summarises the results of this analysis and demonstrates that for the majority of patients a shorter observation period results in a significant change of their reported motor performance. Specifically, Figure 11 shows that when the recording period is reduced from 60 to 20 seconds the power of the tremor for 60% of the patients is reduced by more than 10%. Similar results are obtained when the observation length is reduced to 40 seconds with the same magnitude of change observed for 35% of the participants in this case. These changes in motor performance for shorter recording periods correspond to significant changes in the estimated MDS-UPDRS score for a single observation ranging between 1 and 2.5 points on the MDS-UPDRS scale. This difference in the actual clinical score corresponds to an average expected disease progression over a six- and twelve-month period respectively, thus representing a significant error in precisely assessing motor performance.

These results clearly imply that that it is necessary to maintain the full 60 second recording period for each individual observation. Relevant clinical literature considering the MDS-UPDRS does not appear to offer explicit justification for this performance. However, it seems that this is an observation readily confirmed by experienced clinicians such as those participating in focus groups conducted by cloudUPDRS (cf. Section 4). In particular, it was suggested that the longer duration is required in most cases to cause mild fatigue that reveals the true characteristics of motor performance. In any case, the option to develop the quick test by reducing the duration of individual observations does not appear viable and alternatives must be considered.

8.2. Identifying Clinically Distinct Factors

Clinical investigations of the MDS-UPDRS scale reported in the medical literature have identified a smaller group of clinically distinct factors, typically five to six, that provide high correlation to the overall score of the motor examination of Part III of the MDS-UPDRS [60, 61]. This observation corroborates the possibility to develop the quick test by reducing the number of individual observations to a much restricted group, which correlates well with the overall patient score. Furthermore, note that the MDS-UPDRS protocol was designed to explore exhaustively the full range of possible motor symptoms caused by PD, but a specific individual would typically present a smaller number of symptoms (especially in earlier stages of PD) that dominate their MDS-UPDRS score and that remain relatively stable over a time frame of a few months. Indeed, a common observation is that PD motor symptoms are asymmetric [7, 51] for example, for a particular patient one side can be significantly affected by tremor while the opposite side may not be affected at all thus contributing zero units towards their MDS-UPDRS score.

Motivated by this observation, we develop a methodology using standard machine learning methods that successfully identify the appropriate subgroup of observations for a specific patient which offer the highest predictive power of their overall motor performance. Upon enrolment in cloudUPDRS, patients are required to carry out the full test at least five times during the first week of monitoring. At the end of this calibration period we use the data of the full test to conduct a feature importance analysis. Specifically, following [17] we apply an ensemble of randomized decision trees on multiple sub-samples of the test data improving its predictive accuracy through averaging and over-fitting control. We then rank individual observations according to the relative importance of their corresponding features (two and three features per tremor and bradykinesia test respectively). Finally, we select the subgroup of top performing observations which account for at least 80% of the variance in the overall UPDRS score.

At the end of this process, the cloudUPDRS system is configured with an individual user profile detailing the subgroup of observations identifed for inclusion in the quick test. This profile is automatically communicated to the app at the next start up so that it is reconfigured to enable the quick test feature in its home screen (cf. Figure 2). The selected settings remain active for a period of six months after which a new set of full tests is required due to the likelihood of changes in motor symptoms over this time frame.

8.3. Results

To evaluate the effectiveness of this approach we employed the data set described in Section 8.1 selecting only patients for which at least five full test results are available. For each patient we apply the above methodology to create a personalised quick test profile. We discover that in all cases we are able to account for the target variance using features associated with only three or less observations. This result is consistent across all patients examined representing medium and progressed stages of the disease.

Figure 12 shows the results of this analysis for a typical patient from this cohort suggesting in this case that just three observations (from which seven features are calculated) are adequate to account for approximately 90% of the variation. Specifically, this patient's quick test profile consists of observations of left leg agility, right arm rest tremor and single tapping of the left hand which provide the adequate information to track their overall motor performance. System logs confirm that this patient was able to complete the quick tests consistently in less than 4 minutes over 50 times in the two months following the availability of their profile. Note that this patient is at an advanced stage of PD presenting significant mobility impairments.

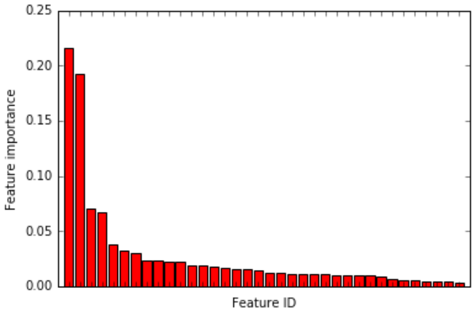


Figure 12: Predictive power of features associated with individual UPDRS observations.

9. Future Work and Conclusions

The results presented in this paper are promising but for the clinical community to accept cloudUPDRS as a standard tool for the assessment of PD, it is necessary to validate its performance against the higher standards required for medical research and practice. To this end, we are currently undertaking a full clinical study of the cloudUPDRS app.[3] CUSSP is a single-site, open label study carried out at University College Hospital in London, U.K. comparing the validity and usability of the cloudUPDRS score for home monitoring of symptoms and signs in PD. The primary objective is to validate in a clinical setting the UPDRS score computed automatically by cloudUPDRS against the clinically defined UPDRS score which is currently recognized as the gold standard. Participating patients with PD are assessed in the clinic under so-called L-Dopa challenge conditions, that is after an overnight period of no medication, and subsequently after medication administration thus observing motor performance in OFF and ON conditions. These assessments are video-recorded at a baseline visit by an unblinded rater, and the video is further assessed by two blinded raters, to produce three repeated scores of the clinical gold-standard UPDRS. During the same hospital visit, the patient performs the UPDRS assessment

[3] CUSSP: The CloudUPDRS Smartphone Software in Parkinson's Study cf. https://clinicaltrials.gov/ct2/show/NCT02937324.

in the ON and OFF conditions using the cloudUPDRS. The clinical and app scores are then compared, first using the Bland-Altman agreement between the score calculated by cloudUPDRS and the average clinical UPDRS rating score. This is in the form of a mean difference and 95% confidence interval of the limit of agreement. To assess whether the mean difference varies with UPDRS score, we also generate the Bland-Altman plot. Second, to determine how the cloudUPDRS score compares to the inter-rater agreement, we also assess agreement between different combinations of gold-standard raters. Finally, to assess the relative validity of different elements of the cloudUPDRS score, we present their intra-class correlation coefficients. CUSSP is expected to report its findings at the end of 2017.

Further to the consideration of questions of clinical validation and the longer-term improvement of treatment strategies for PD, cloudUPDRS also represents an attempt to address the challenges created by a global aging population. Indeed, according to the World Health Organization [67], ageing populations generate considerable economic effects, notably intensifying pressures on health-care systems which for many of the more economically developed countries already represent the largest area of expenditure. In the UK, the cost of caring for PD patients exceeds 1.25 billion British pounds annually and is rapidly increasing. In this socioeconomic situation, mobile health apps present a unique opportunity for the provision of cost effective care at population scale. Yet, to reach their full potential such apps must offer safety guarantees and facilitate a seamless user experience.

To this end, cloudUPDRS is the first smartphone app to receive certification as a medical device for the clinical assessment of Parkinson's Disease. The design and development of cloudUPDRS follows the structured process required to ensure the safety guarantees required by medical devices and at the same time support an efficient and effective patient experience that facilitates its wider adoption. A key ingredient to achieve the latter is the introduction of two novel features developed specifically for cloudUPDRS, supplemented by a bespoke patient expereince design that provides structure and guidance during use at home. First, a tailored deep learning approach was employed to replace expert human supervision during the administration of the common motor performance assessment protocol for PD. In particular, in this paper an improved neural network architecture was presented, implemented as two-stage approach where high-performance resources are used to train the model which is subsequently embedded in the cloudUPDRS app operation on the mobile phone to provide interactive oversight. Second, recognizing the need for frequent assessments especially in the period leading to a clinical appointment, a personalised quick test was developed lasting less than 5 minutes per application to accurately trace overall motor performance while considerably improving patient compliance. In our experiments both approaches performed reliably and produce promising results. We anticipate both techniques to be useful for a wider class of mobile health-care apps with similar requirements. Further experimentation with a larger patient population outside the PD context is of course necessary to fully assess the wider potential of the techniques presented in this paper.

Acknowledgments

Project cloudUPDRS: Big Data Analytics for Parkinson's Disease patient stratification is supported by Innovate UK (Project Number 102160). Additional support was provided by the Birkbeck Wellcome Trust Institutional Strategic Support Fund (ISSF) through research translation study CIC06. We gratefully acknowledge the support of NVIDIA Corporation with the donation of Tesla K40 GPUs used for this research. Project partners would also like to thank Parkinson's UK for providing access to their online forums and assisting with the recruitment of participants.

References

[1] J. Schmidhuber. "Deep learning in neural networks: An overview." *Neural Networks*, 61, 85–117, 2015.

[2] M. Abadi *e*t. al. *T*ensorFlow: Large-scale machine learning on heterogeneous systems, Whitepaper available at tensorflow.org, 2015.

[3] D. M. Allen. "The relationship between variable selection and data augmentation and a method for prediction." *Technometrics*, 16(1), 125-127, 1974.

[4] S. Arora, V. Venkataraman, A. Zhang, S. Donohue, K.M. Biglan, E.R. Dorsey and M.A. Little. "Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study." *Parkinsonism and Related Disorders*, 21(6), 650–653, 2015.

[5] C. Bettini, O. Brdiczka, K. Henricksen, J. Indulska, D. Nicklas, A. Ranganathan and D. Riboni. "A survey of context modelling and reasoning techniques." *Pervasive and Mobile Computing*, 6(2), 161–180, 2010.

[6] J. Bergstra, Y. Bengio. "Random Search for Hyper-parameter Optimization." *J. of Machine Learning Research*, 13, 281-305, 2012.

[7] O. Blin, A. M. Ferrandez and G. Serratrice. "Quantitative analysis of gait in Parkinson patients: increased variability of stride length." *J. Neurol. Sci.*, 98, 91–97, 1990.

[8] L. Breiman. "Bagging Predictors." *Mach. Learn.*, 24(2), 123–140, 1996.

[9] L. Breiman. "Random Forests." *Mach. Learn.*, 45(1), 5–32, 2001.

[10] J.R. Brubaker, C. Lustig and G.R. Hayes. "PatientsLikeMe: empowerment and representation in a patient-centred social network." *CSCW10 Workshop Research in Healthcare: Past, Present, and Future*, Savannah, GA, USA, 2010.

[11] S. J. Chinta and J. K. Andersen. "Dopaminergic neurons," *The International Journal of Biochemistry & Cell Biology*, 37(5), 942–946, 2005.

[12] J.F. Daneault *et al.*. "Using a smart phone as a standalone platform for detection and monitoring of pathological tremors." *Front Hum Neurosci*, 6, 357, 2008.

[13] B.M. Eskofier *et al.*. "Recent machine learning advancements in sensor-based mobility analysis: Deep learning for Parkinson's disease assessment," *Conf Proc IEEE Eng Med Biol Soc.*, 655–658, 2016.

[14] European Brain Council. *Parkinson's disease Fact Sheet*, 2011.

[15] N.F. Fragopanagos, S. Kueppers, P. Kassavetis, M.U. Luchini, and G. Roussos. "Towards Longitudinal Data Analytics in Parkinson's Disease." *Proc. 1st Int. Conf. on IoT and Big Data Technologies for HealthCare*, June 15-16, Budapest, Hungary, 2016.

[16] J. Friedman. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, 29(5), 1189–1232, 2001.

[17] P. Geurts, D. Ernst and L. Wehenkel. "Extremely Randomized Trees." *Mach. Learn.*, 63(1), 3–42, 2006.

[18] C.G. Goetz *et al.*. "Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale Presentation and Clinimetric Testing Results." *Movement Disorders*, 23(15), 2129–2170, 2008.

[19] I. Goodfellow, Y. Bengio and A. Courville. "Deep Learning." *MIT Press*, 330-372, 2016

[20] N. Y. Hammerla, J. Fisher, P. Andras, L. Rochester, R. Walker and T. Ploetz. ' 'PD Disease State Assessment in Naturalistic Environments Using Deep Learning." *AAAI Conf. on Artificial Intelligence*, 2015.

[21] K. He, X. Zhang, S. Ren and J. Sun. "Deep Residual Learning for Image Recognition." *CoRR.* `http://arxiv.org/abs/1512.03385`. 2015.

[22] S. Ioffe and C. Szegedy "Batch Normalization Accelerating Deep Network Training by Reducing Internal Covariate Shift." *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 448-456, 2015.

[23] S. Hochreiter and J. Schmidhuber. “Long Short-Term Memory.” *Neural Computation*, 9(8), 1735–1780, 1997.

[24] J. Jankovic. “Parkinson’s disease: clinical features and diagnosis.” *J. Neurology, Neurosurgery and Psychiatry*, 79(4), 368-76, 2008.

[25] C. Jenkinson, R. Fitzpatrick, V. Peto, R. Greenhall and N. Hyma. “The Parkinson’s Disease Questionnaire (PDQ39): development and validation of a Parkinson’s disease summary index score.” *Age Ageing*, 26(5), 353–357, 1997.

[26] A. Jha, P. Kassavetis, E Nomikou, J. Rothwell, K. Bhatia and G. Roussos. “The cloudUPDRS smartphone app: home monitoring for Parkinson’s Disease.” *The Future of Medicine Conference*, Royal Society of Medicine, May 19, London, UK, 2016.

[27] P. Kassavetis, T. A. Saifee, G. Roussos, L. Drougkas, M. Kojovic, J. C. Rothwell, M. J. Edwards, K. P. Bhatia. “Developing a tool for remote digital assessment of Parkinson’s disease.” *Movement Disorders - Clinical Practice*, 3(1), 2015.

[28] D. Kingma and J. Ba. “Adam: A method for stochastic optimization.” *arXiv preprint* arXiv:1412.6980, 2015.

[29] N. Kostikis *et al.*. “Towards remote evaluation of movement disorders via smartphones.” *Proc. IEEE Eng Med Biol Soc*, 5240–3, 2011.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks.” *Advances in Neural Information Processing Systems 25*, 1097–1105, 2012.

[31] S. Kueppers, I. Daskalopoulos, A. Jha, N.F. Fragopanagos, P. Kassavetis, E. Nomikou, T. Saifee, J.C. Rothwell, K. Bhatia, M.U. Luchini, M. Iannone, T. Moussouri, and G. Roussos. “From Wellness to Medical Diagnostic apps: The Parkinson’s Disease Case.” *Proc. Int. Conf. on Personal, Pervasive and mobile Health*, June 14-16, Budapest, 2016.

[32] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. "Backpropagation applied to handwritten zip code recognition." *Neural Computation*, 1(4):541-551, 1989.

[33] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. "Handwritten digit recognition with a back-propagation network." *Advances in Neural Information Processing Systems (NIPS)*, 396-404, 1990.

[34] R. Lemoyne *et al.* "Implementation of an iPhone for characterizing Parkinson's disease tremor through a wireless accelerometer application." *Pro.c IEEE Eng. Med. Biol. Soc.*, 4954–8, 2010.

[35] W. Maetzler, J. Domingos, K. Srulijes, J. J.Ferreira and B. R. Bloem. "Quantitative wearable sensors for objective assessment of Parkinson's disease,' *Movement Disorders*, 28(12), 1628–1637, 2013.

[36] E. Martin. "Novel method for stride length estimation with body area network accelerometers." *IEEE BioWireleSS*, 79–82, 2011.

[37] E. Martin, V. Shia and R. Bajcsy. "Determination of a Patient's Speed and Stride Length Minimizing Hardware Requirements." *Proc. Int. Conf. Body Sensor Networks*, 144–149, 2011.

[38] V. Marx. "Human phenotyping on a population scale." *Nature Methods*, 12, 711–714, 2015.

[39] N. Marz and J. Warren. *Big Data: Principles and best practices of scalable realtime data systems.* Manning Publications, 2013.

[40] P.M. Matthews, P. Edison, O.C. Geraghty and M.R. Johnson. "The emerging agenda of stratified medicine in neurology." *Nature Reviews*, 10, 15–27, 2014.

[41] M. Liang and X. Hu. "Recurrent convolutional neural network for object recognition." *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[42] J. Moody. "Prediction Risk and Architecture Selection for Neural Networks." *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*, 136, 147–165, 1994.

[43] V. Nair and G. E. Hinton. "Rectified linear units improve restricted Boltzmann machines." *Int. Conf. Machine Learning (ICML)*, 807–814, 2010.

[44] S. Newman. *Building Microservices: Designing Fine-Grained Systems.* O'Reilly Media, 2015.

[45] National Institute for Health and Clinical Excellence. *Parkinson's disease: diagnosis and management in primary and secondary care: National cost-impact report.* NICE clinical guideline no. 35, 2006.

[46] A. McCallum and K. Nigam. "A comparison of event models for naive Bayes text classification." *Proc. AAAI/ICML-98 Work. on Learning for Text Categorization*, 41–48, 1998.

[47] D. Pan, R. Dhall, A. Lieberman and D.B.Petitti. "A mobile cloud-based Parkinson's disease assessment system for home-based monitoring." *JMIR Mhealth Uhealth*, 3(1), e29, 2015.

[48] Parkinson's UK. *Parkinson's prevalence in the United Kingdom.* `http://www.parkinsons.org.uk/sites/default/files/parkinsonsprevalenceuk_0.pdf`, 2009.

[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine learning in Python." *J. of Machine Learning Research*, 12, 2825–2830, 2011.

[50] L. Y. Pratt. "Non-literal transfer of information among inductive learners." In R. J. Mammone and Y. Y. Zeevi (ed.) *Neural Networks: Theory and Applications II*, 1992.

[51] L. Ricciardi, D. Ricciardi, F. Lena *et al.*. “Working on asymmetry in Parkinson’s disease: randomized, controlled pilot study.” *Neurol. Sci.*, 36, 1337–1343, 2015.

[52] J.A. Robichaud, K.D. Pfann, S. Leurgans, D.E. Vaillancourt, C. L. Comella and D.M. Corcos. “Variability of EMG patterns: a potential neurophysiological marker of Parkinsons disease?” *Clinical neurophysiology: official journal of the International Federation of Clinical Neurophysiology*, 120(2), 390–397, 2006.

[53] R. Robinson. “Electronic Sensors Break New Ground in Neurology Practice and Research,” *Neurology Today*, 15(7), 20–26, 2015.

[54] A. Rodriguez-Molinero *et al.*. “Validation of a portable device for mapping motor and gait disturbances in Parkinson’s disease.” *JMIR Mhealth Uhealth*, 3(1), e9, 2015.

[55] A. H. V. Schapira, M. Emre, P. Jenner and W. Poewe. “Levodopa in the treatment of Parkinson’s disease.” *European Journal of Neurology*, 16, 982–989, 2009.

[56] K. Simonyan and A. Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” *International Conference on Learning Representations.* 2014

[57] C. Stamate, G. D. Magoulas, S. Kueppers, E. Nomikou, I. Daskalopoulos, M. U. Luchini, T. Moussouri and G. Roussos. “Deep learning Parkinson’s from smartphone data.” *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2017.

[58] I. Sutskever,J. Martens, G. Dahl and G. Hinton. “On the Importance of Initialization and Momentum in Deep Learning.” *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, III-1139–III-1147, 2013.

[59] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich. "Going Deeper with Convolutions." *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[60] G. T. Stebbins and C. G. Goetz. "Factor structure of the unified Parkinson's disease rating scale: Motor examination section." *Movement Disorders*, 13(4), 633–636, 1998.

[61] S. D. Vassar *et al.*. "Confirmatory Factor Analysis of the Motor Unified Parkinson's Disease Rating Scale." *Parkinson's Disease*, Article ID 719167, 2012.

[62] P. Werbos. *Beyond regression: new tools for prediction and analysis in the behavioral sciences.* Ph.D. Thesis, Harvard University, 1974.

[63] M. W. Whittle. *Gait Analysis: An introduction.* Butterworth-Heinemann, 2014.

[64] P. Wicks. *The patient of the future.* `http://parkinsonsmovement.com/the-patient-of-the-future/`, 2015.

[65] A. C .Wilson, R. Roelofs, M. Stern, N. Srebro and B. Recht. "The Marginal Value of Adaptive Gradient Methods in Machine Learning" *arXiv*, `https://arxiv.org/pdf/1705.08292v1.pdf`, 2017.

[66] D. H. Wolpert. "Stacked Generalization." *Neural Net.*, 5, 241–259, 1992.

[67] World Health Organization. *Current Status of the World Health Survey.* `http://www.who.int/healthinfo/survey/`. 2013.

[68] M. D. Zeiler and R. Fergus. "Visualizing and Understanding Convolutional Networks." *ECCV 2014: 13th European Conference*, 2014.

[69] G. P. Zhang. "Neural networks for classification: a survey." *IEEE Trans. Systems, Man, and Cybernetics, Part C*, 30(4), 451–462, 2000.

[70] J. Zhu, H. Zou, S. Rosset, T. Hastie. "Multi-class AdaBoost." *Stat. and Interface*, 2, 349–360, 2009.