(I123)FP-CIT reporting: Machine Learning, Effectiveness and Clinical Integration

Jonathan Christopher Taylor

Medical Physics Group
Mathematical Modelling in Medicine
Infection Immunity and Cardiovascular Disease
Faculty of Medicine, Dentistry and Health
The University of Sheffield

Thesis submitted for the degree of PhD
January 2019

Abstract

(I123)FP-CIT imaging is used for differential diagnosis of clinically uncertain Parkinsonian Syndromes. Conventional reporting relies on visual interpretation of images and analysis of semi-quantification results. However, this form of reporting is associated with variable diagnostic accuracy results. The first half of this thesis clarifies whether machine learning classification algorithms, used as computer aided diagnosis (CADx) tool, can offer improved performance.

Candidate machine learning classification algorithms were developed and compared to a range of semi-quantitative methods, which showed the superiority of machine learning tools in terms of binary classification performance. The best of the machine learning algorithms, based on 5 principal components and a linear Support Vector Machine classifier, was then integrated into clinical software for a reporting exercise (pilot and main study).

Results demonstrated that the CADx software had a consistently high standalone accuracy. In general, CADx caused reporters to give more consistent decisions and resulted in improved diagnostic accuracy when viewing images with unfamiliar appearances.

However, although these results were undoubtedly impressive, it was also clear that a number of additional, significant hurdles remained, that needed to be overcome before widespread clinical adoption could be achieved.

Consequently, the second half of this thesis focuses on addressing one particular aspect of the remaining translation gap for (I123)FP-CIT classification software, namely heterogeneity of the clinical environment. Introduction of new technology, such as machine learning, may require new metrics, which in this work were informed through novel methods (such as the use of innovative phantoms) and strategies, enabling sensitivity testing to be developed, applied and evaluated.

The pathway to acceptance of novel and progressive technology in the clinic is a tortuous one, and this thesis emphasises the importance of many factors in addition to the core technology that need to be addressed if such tools are ever to achieve clinical adoption.

Acknowledgements

I would like to thank my supervisor, John Fenner, for always being supportive through the many challenges faced during this PhD. He was always willing to set aside time to look through my work or discuss the project, no matter how busy.

I would also like to thank the National Institute for Health Research for providing me with the opportunity to undertake a PhD, and the Nuclear Medicine department at Sheffield Teaching Hospitals for allowing me to take time out from my clinical work.

Finally, I would like to thank my partner Cat for her continual encouragement and support over the course of my fellowship, and for providing a welcome distraction from work when it was needed.

Publications

The work presented in this thesis contributed to a series of open-access, peer-reviewed publications (1–5). These are listed below along with the relevant chapters from which they originated. The license associated with each publication is also shown.

1) Taylor JC, Fenner JW. Comparison of machine learning and semi-quantification algorithms for (I123)FP-CIT classification: the beginning of the end for semi-quantification? EJNMMI Phys. 2017 Dec;4(1):29. (license CC-BY 4.0)
   CHAPTER 2 and 3

2) Taylor JC, Romanowski C, Lorenz E, Lo C, Bandmann O, Fenner J. Computer-aided diagnosis for (123I)FP-CIT imaging: impact on clinical reporting. EJNMMI Res. 2018 May 8;8(1):36. (license CC-BY 4.0)
   CHAPTER 4

3) Taylor JC, Vennart N, Negus I, Holmes R, Bandmann O, Lo C, et al. The subresolution DaTSCAN phantom: a cost-effective, flexible alternative to traditional phantom technology. Nucl Med Commun. 2018 Mar;39(3):268–75. (license CC-BY 4.0)
   CHAPTER 6

4) Taylor J, Fenner J. The challenge of clinical adoption—the insurmountable obstacle that will stop machine learning? BJR Open 2019; 1: 20180017. (license CC-BY 4.0)
   CHAPTER 5 and 8

5) Taylor JC, Fenner JW. Clinical Adoption of CAD: Exploration of the Barriers to Translation through an Example Application. Procedia Computer Science 2016 90:93-98. (license CC BY-NC-ND 4.0)
   CHAPTER 5 and 8

Statement of contribution

The author declares that the work presented in this thesis is his own, with the exception of the following:

- The 3D printed phantom described in chapter 6, and the method for transforming the anatomical template to the dimensions of the phantom, was developed by colleagues at University Hospital Bristol.

Glossary of terms

SPECT	Single Photon Emission Computed Tomography
PS	Parkinsonian Syndrome
PD	Parkinson's Disease
ET	Essential Tremor
(123I)FP-CIT	((123)I-N-omega-fluoropropyl-2beta-carbomethoxy-3beta-(4-iodophenyl)nortropane
PSP	Progressive Supranuclear Palsy
MSA	Multiple system Atrophy
DLB	Dementia with Lewy Bodies
CBD	Corticobasal Degeneration
VaP	Vascular Parkinsonism
DIP	Drug Induced Parkinsonism
AD	Alzheimer's Disease
HC	Healthy Controls
SWEDD	Scans Without Evidence of Dopaminergic Deficit
PDD	Presynaptic Dopaminergic Deficit
DaT	Dopamine Active Transporters
EANM	European Association of Nuclear Medicine
SNM	Society of Nuclear Medicine
SDDD	Striatal Dopaminergic Deficit Disorder
BNMS	British Nuclear Medicine Society
NHS	National Health Service
SBR	Striatal Binding Ratio
PPMI	Parkinson's Progression Markers Initiative
STH	Sheffield Teaching Hospitals
LEHR	Low Energy High Resolution
LEUHR	Low Energy Ultra High Resolution
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
CNN	Convolutional Neural Network
ICA	Independent Component Analysis
SVM	Support Vector Machine
RBF	Radial Basis Function

PNN	Probabilistic Neural Network
CV	Cross Validation
SVD	Singular Value Decomposition
SD	Standard Deviation
CI	Confidence Interval
DSC	Dice Similarity Coefficient
ShIRT	Sheffield Image Registration Toolkit
ROC	Receiver Operator Curve
ICC	Intraclass correlation coefficient
CAD(x)	Computer Aided Diagnosis
CADe	Computer Aided Detection
NICE	National Institute for Health and Care Excellence
IRAS	Integrated Research Application System
SSP	Subresolution Sandwich Phantom
FDM	Fused Deposition Modelling
PLA	Polylactic Acid
AAL	Automated Anatomical Atlas
MNI	Montreal Neurological Institute
ADNI	Alzheimer's Disease Neuroimaging Initiative
ENCDAT	European Database of [123I]FP-CIT (DaTSCAN) SPECT scans of healthy controls
MTEP	Medical Technologies Evaluation Programme
DAP	Diagnostic Assessment Programme
MICCAI	Medical Image Computing and Computer Assisted Intervention
MRI	Magnetic Resonance Imaging
FDA	Food and Drug Administration

Table of Contents

List of Figures

List of Tables

1 Introduction

This work aims to assess the effectiveness of automated classification algorithms for assisted radiological reporting of clinical (I123)FP-CIT nuclear medicine scans. The major focus is on evaluation of algorithms in a clinical reporting context. The following sections set out the clinical and technical background to (I123)FP-CIT imaging and associated disease processes. An overview of current standard of care image analysis techniques (semi-quantification) is necessarily provided, along with a summary of machine learning classification algorithms and their history of application in (123I)FP-CIT imaging. This is the basis for a series of investigations which seek to establish the performance of developed classification algorithms, as compared to semi-quantification, and for evaluating the impact of such software tools on human reporter performance.

1.1 (123I)FP-CIT imaging

(123I)FP-CIT (DaTSCAN) imaging is a Single Photon Emission Computed Tomography (SPECT) brain scan technique used for differential diagnosis of patients with clinically uncertain Parkinsonian Syndrome (PS). In a clinical context it is used to detect the loss of dopaminergic neuron terminals associated with idiopathic Parkinson's Disease (PD), Multiple System Atrophy (MSA) and Progressive Supranuclear Palsy (PSP). It is also used to help distinguish between Dementia with Lewy Bodies (DLB) and other forms of dementia and to differentiate patients with presynaptic Parkinsonism from those with other forms of Parkinsonism (6). In a research context (123I)FP-CIT is increasingly used for monitoring progression of disease in patients suffering from PS.

1.1.1 Parkinsonian Syndromes

Parkinsonian Syndrome refers to a collection of movement disorders with similar clinical features but different pathologies. It includes, in addition to rarer causes of Parkinsonism:

- Parkinson's Disease (PD)
- Multiple System Atrophy (MSA)
- Progressive Supranuclear Palsy (PSP)
- Corticobasal degeneration (CBD)
- Vascular Parkinsonism (VaP)

- Drug induced Parkinsonism (DIP)

The most significant clinical symptom exhibited by all PS patients is bradykinesia, which is defined as ‘‘slowness of initiation of voluntary movement with progressive reduction in speed and amplitude of repetitive action” (7). PS is also typically associated with rest tremors, extrapyramidal rigidity and postural instability (8).

The most common form of PS is PD, affecting approximately 1% of people over the age of 65 (9). Diagnosis is predominately guided by clinical features. A number of guidelines have been published to assist with diagnosis, in particular the UK Parkinson’s Disease Society Brain Bank Diagnostic Criteria (10). Other forms of Parkinsonism may display subtly different features, which can guide differential diagnosis. For example, MSA is often associated with early, progressive autonomic dysfunction whilst PSP patients will typically present with eye movement problems (8).

However, differentiating between different forms of PS remains challenging. In addition, the progressive motor deficits typically displayed by PS patients are similar to those experienced by patients with essential tremor, which is a condition associated with involuntary limb or head movement. Other diseases such as multiple sclerosis and Huntington’s disease may also present as movement disorders.

In the UK and elsewhere patients presenting with motor deficits may be referred to experts in movement disorders or to clinicians with more general expertise, such as general neurologists. Given the subtle differences in the features of different PS sub-types it is perhaps unsurprising that clinical diagnosis of PD is often associated with disappointing low accuracy figures. A recent systematic review and meta-analysis by Rizzo and colleagues (11) identified 11 studies comparing clinical diagnosis of Parkinson’s Disease with pathologic diagnosis post-mortem (the gold standard). Clinical diagnosis by ‘non-experts’ (such as general neurologists) was associated with an accuracy of only 74% and a specificity of just 49% as compared to neuropathology results (experts in movement disorders achieved a higher accuracy of approximately 80%).

Similar accuracy results for general neurologists were found in a large Finnish study with 1362 patients (12). A study conducted in a specialised centre for movement disorders, where one might expect relatively high diagnostic accuracy, showed that 36% of patients were reclassified within a mean time window of 3.4 years following initial clinical diagnosis,

suggesting that clinical features may not be a reliable indicator of disease in the early stages (13). Indeed, it appears that there is a tendency to over diagnose PD in early disease stages (14). Similarly, studies have demonstrated that clinical diagnosis of DLB is associated with low sensitivity (15). Although the accuracy of diagnosis by clinical features alone is likely to increase over time as the disease progresses and symptoms evolve (13,16,17), these findings suggest that there is a need for other tests, particularly in the early stages of disease, which can highlight differences in disease pathology and increase diagnostic accuracy and certainty.

Getting diagnosis correct early on can be extremely important for decisions on patient management. For example, patients with PD will typically be prescribed with Levodopa to reduce symptoms. However, the drug is associated with significant side effects such as nausea and vomiting. In patients with essential tremor Levodopa offers no benefit but may reduce quality of life.

1.1.2 Dementia with Lewy Bodies

DLB is a progressive brain disease associated with the presence of cortical Lewy bodies (abnormal aggregates of protein) inside nerve cells. DLB presents with similar symptoms to both Alzheimer's Disease (AD) and PD. There is still uncertainty in regards to prevalence, largely due to difficulties in diagnosing the condition. However, a recent systematic review has estimated that DLB accounts for approximately 1 in 25 dementia cases diagnosed in the community and 1 in 13 cases diagnosed in secondary care (18). DLB patients suffer from typical dementia symptoms, including attention deficits and substantial memory impairment. However, damage to cells in the substantia nigra can also give rise to movement deficits, similar to those seen in PS.

Differential diagnosis between DLB and other forms of dementia is important as patient management is different for each disease. In particular, DLB patients often have severe sensitivity to neuroleptics, but these antipsychotic drugs are commonly prescribed for Alzheimer's patients to reduce disruptive behaviour.

1.1.3 Tracer uptake and differential diagnosis

The radioactive tracer (I123)FP-CIT targets just one protein involved in the nigrostriatal dopaminergic pathway, one of the four main dopamine pathways of the brain. In order to understand how imaging relates to disease it is necessary to appreciate the basic steps

involved in neurotransmission, whereby signals are passed from one neuron to the next. As shown by Figure 1-1 the neurotransmitter (dopamine) is held within vesicles in the pre-synaptic neuron. Action potentials passing along this neuron cause vesicles to release their dopamine into the synaptic cleft. Some of the released dopamine binds to and activates the receptors of the post-synaptic neuron, causing an action potential to be generated and thus allowing the signal to be passed from one neuron to the next. Unbound dopamine is reabsorbed from the synapse back into the pre-synaptic neuron by Dopamine active transporters (DaT).

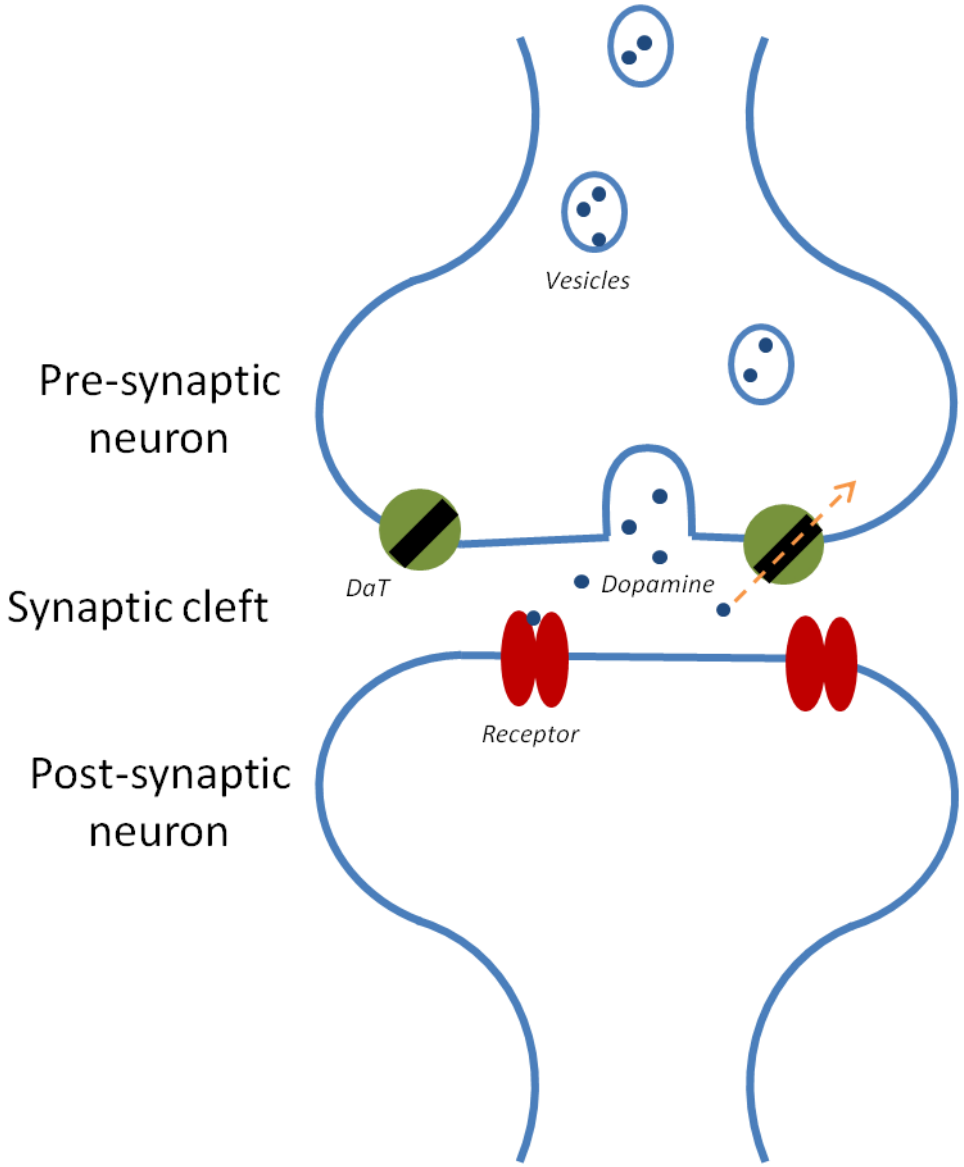


Figure 1-1 Schematic of two dopaminergic neurons and their synapse. The neurotransmitter dopamine is released from vesicles into the synaptic cleft

It has been shown in-vitro that FP-CIT binds reversibly to DaT (19). Human brain slices recovered post-mortem from individuals without a Parkinsonian Syndrome, following FP-CIT administration, have shown high concentrations of the tracer within the putamen and caudate (19,20). This area of the brain, known as the striatum, contains a large number of nigrostriatal axon terminals. Thus, by radiolabelling FP-CIT with I123, and administering to the patient, it is possible to image DaT function in-vivo within the striatum.

Some movement disorders cause dopaminergic neurons to die. The concentration of DaT in the striatum reduces as the number of healthy dopaminergic neurons decreases. Therefore, by examining the density of administered (123I)FP-CIT within the striatum, through SPECT imaging, an assessment can be made of the function of the pre-synaptic dopaminergic pathway and whether disease processes associated with certain conditions are present. Table 1-1 summarises which of the previously described disorders is associated with reduced DaT density, and which causes no significant change. For patients where differential diagnosis between these diseases is unclear, a positive or negative (123I)FP-CIT test can therefore help to identify the likely cause of a patient's symptoms.

Disease	**Impact on pre-synaptic DaT density**
Idiopathic Parkinson's Disease	Reduced (21)
Multiple System Atrophy	Reduced (21)
Progressive Supranuclear Palsy	Reduced (21)
Essential Tremor	Not affected (21,22)
Alzheimer's Disease	Not affected (21,23)
Drug-Induced Parkinsonism	Not affected (24)
Vascular Parkinsonism	Reduced or unaffected (25,26)
Corticobasal Degeneration	Reduced or unaffected (27)
Dementia with Lewy Bodies	Reduced in most cases (28–31)

Table 1-1 Summary of the effects of different brain disorders on pre-synaptic DaT density, according to current research

(123I)FP-CIT is just one of a range of radiopharmaceuticals that have been developed to enable imaging of both the pre- and post-synaptic dopaminergic pathway. However, this work focuses on clinical imaging. In the UK the majority of dopaminergic scans conducted in

the NHS are carried out with (123I)FP-CIT, and so all other radiopharmaceuticals are considered out of scope.

1.1.4 Clinical SPECT imaging

(123I)FP-CIT imaging is carried out using SPECT, where multiple planar projections are acquired from a gamma camera, which are then reconstructed into a 3D volume. Recommended clinical image acquisition parameters are set out in the "information for the physician" leaflet supplied by the manufacturer of (123I)FP-CIT. In addition, both the European Association of Nuclear Medicine (EANM) and Society of Nuclear Medicine (SNM) have produced guidelines for this test (6,32).

Typically, reporters will view reconstructed axial slices from the centre of the brain, encompassing the whole striatal area. In some cases a summed image will also be examined, created by adding together voxel intensities from consecutive slices.

1.1.5 Accuracy and variability of unaided visual analysis

In this study it is hypothesised that automated classification algorithms are a useful diagnostic aid, providing an objective, independent assessment of image appearances, which the clinician may use as part of his / her considerations in image reporting. Before developing tools for this purpose it is important to understand the performance of human observers alone in detecting abnormal tracer uptake patterns.

A number of studies have considered the accuracy of (I123)FP-CIT imaging. However, these have been conducted in different settings with different groups of patients (at different stages of disease), using different gold standard methods with different acquisition and reconstruction parameters. Results must therefore be interpreted in light of any potential biases.

Many modern studies consider the accuracy of (I123)FP-CIT imaging in conjunction with semi-quantification[a], and these are excluded from the following discussion in order to focus purely on the reporter's ability to visually interpret an image. In addition, the following

[a] Semi-quantification: relative quantitative measures of uptake within a region of interest. See section 1.2

discussion mainly focuses on larger scale studies. Results from studies with small or limited datasets are largely excluded.

Although post-mortem pathological examination of the brain is generally considered the gold standard method for diagnosing parkinsonisms, this is not practical in many studies. Clinical diagnosis with long term follow-up (covering disease progression, results of other tests and assessment of response to treatment) is the most commonly used reference standard, despite the possible limitations this places on interpretation of results. One of the main limitations of clinical follow-up, particularly for retrospective studies is that the results of (I123)FP-CIT imaging itself may have a significant impact on the final diagnosis.

Use of clinical diagnosis as a gold standard may appear counter-intuitive at first sight given that (I123)FP-CIT imaging was primarily introduced to clinic in order to overcome limitations associated with diagnosis by clinical features alone. However, (I123)FP-CIT imaging is mostly used in early stages of disease where clinical data is limited and uncertain. To emphasise the time-limited justification for carrying out a (I123)FP-CIT scan, de-La Fuente-Fernandez measured the diagnostic accuracy of clinical diagnosis using SPECT data as the reference standard for 322 PS and non-PS patients (33). He showed that accuracy was 84% in the early stages of disease but 98% for patients with established clinical diagnosis (i.e. the two tests were identical in latter stages of disease).

Table 1-2 summarises the main research articles focusing on measurement of the diagnostic performance of (I123)FP-CIT imaging (with interpretation through visual analysis only)

Summary of the evidence for diagnostic performance of (I123)FP-CIT imaging (visual image analysis only)		
Source	**Method**	**Results**
O'Brien et al. (21)	Pooled analysis of three phase three and one phase four prospective clinical trials, covering 928 participants. Visual interpretation was conducted by both on-site reporters and a panel of experts. Data was based on a gold standard diagnosis provided by clinical follow-up.	In the differentiation of patients with a striatal dopaminergic deficit disorder (SDDD), such as a PS or DLB, from patients without a SDDD, overall sensitivity was 91.9% and specificity 83.6% when interpretation was performed locally. The expert panel achieved a sensitivity of 88.7% and specificity of 91.2%. Inter-reporter agreement was generally good between members of the expert panel (Cohen's kappa varied from 0.81 to 1.00). However, greater variability was seen between the expert panel and local on-site reporters
O'Brien et al. (23)	Visual analysis of 164 scans by 5 reporters (consensus reporting), as compared to clinical diagnosis	Sensitivity of 78% and specificity 85% in classification of DLB vs. AD. Kappa values on inter-reporter agreement varied from 0.91 to 0.94
Benamer et al. (34)	Multi-centre study. 158 patient scans were read by local reporters and then by a central panel of 5 experts.	Local reporters achieved an accuracy of 98% in the binary diagnostic task of distinguishing between Parkinsonisms and ET / healthy volunteers, whilst the expert panel gave a correct interpretation in 95% of cases.
Marshall et al. (14)	Visual analysis of 99 SPECT images, as compared to clinical interpretation (via video recording) at 3 year follow up	(I123)FP-CIT had a sensitivity of only 78% in diagnosing degenerative Parkinsonism from non-degenerative tremor, but 97% specificity. Here, Inter-reporter agreement on SPECT image interpretation was high

		(kappa statistic varied from 0.94-0.97).
Tolosa et al. (35)	Follow-up study of 85 patients with clinically uncertain PS. SPECT findings were compared to clinical diagnosis established over 2 subsequent years.	(I123)FP-CIT findings agreed with a conclusive clinical diagnosis in 90% of cases
Kemp et al. (36)	Retrospective study of 80 patients, comparing visual analysis by a single observer against clinical diagnosis 12-24 months after SPECT imaging was completed.	(I123)FP-CIT imaging findings were in agreement with clinical diagnosis in 95% of cases
Thomas et al. (31)	Retrospective study of 55 research patients. Diagnosis confirmed by autopsy. Accuracy of (I123)FP-CIT imaging determined through consensus reporting	Accuracy of (I123)FP-CIT was 86% in differentiating DLB from Alzheimer's disease, which was greater than the accuracy measured from clinical diagnosis (79%)

Table 1-2 Summary of research articles measuring the ability of (I123)FP-CIT scans, with interpretation by human reporters, to distinguish between patient groups. Studies are listed in descending order according to the number of patient datasets included in the analysis

Most of the studies listed in Table 1-2 have relied upon data from large hospitals, with relatively high patient throughput. However, (I123)FP-CIT is a routine test that is often carried out in smaller institutions where the level of experience and expertise may be lower. A recent audit conducted by the British Nuclear Medicine Society (BNMS) provides some insight into the level of reporting performance on a wider scale in the NHS (37). It was shown that, for 86 different UK centres (each contributing 6 anonymised scans), independent reviewers agreed the original image report in 88% cases. In the remainder there were discordant findings, which suggests that visual analysis in the wider clinical community is perhaps more significant than that suggested by most research articles.

Overall, it appears that visual interpretation of (I123)FP-CIT images is associated with variable but relatively high accuracy, sensitivity and specificity figures (in the region of 80-90%) for differentiation of dopaminergic deficit disorders from those without such conditions. These relatively impressive performance figures are perhaps unsurprising given the size of impact that PS has on dopaminergic function. As shown in previous research on post-mortem brains, early stage dopaminergic disease is associated with a 70-80% reduction of dopamine in the striatum (38). Given that patients are only referred for (I123)FP-CIT imaging when clinical features have become apparent, it can be inferred that the classification task for visual analysis is to distinguish between two very different functional states.

The available data on inter-reporter agreement indicate that there are generally only small differences between performances of interpreting clinicians. However, importantly, there was a greater level of variability seen between reports by assigned ‘experts’ and locally performed visual analysis. Although there are a number of differences in the studies examined that may have affected results and may limit applicability of findings (particularly in terms of the case mix, reconstruction method and reference standard used), there does appear to be some potential for improving diagnostic accuracy of (I123)FP-CIT tests and for reducing inter-reporter variability. It is this that provides justification for new techniques (for example, machine learning) as described in this thesis.

1.1.6 Conclusion

Parkinsonisms affect a relatively large proportion of the population and although clinical diagnosis remains the dominant diagnostic method, the approach is associated with

somewhat disappointing accuracy figures, particularly in the early stages of disease. (123I)FP-CIT is a SPECT imaging test that enables clinicians to evaluate function in the pre-synaptic nigrostriatal dopaminergic pathway. Visual analysis of these images can help to distinguish between PS and other conditions such as essential tremor, with relatively high accuracy.

Although (123I)FP-CIT imaging appears to be a useful diagnostic tool within the appropriate clinical context, there is some variability in reported accuracy figures and there is evidence of differences in performance between human reporters. Consequently, there may be scope for improving upon the accuracy of (123I)FP-CIT imaging with assistive software based on machine learning.

A form of assistive software is already in use in many clinical departments for (123I)FP-CIT imaging, namely semi-quantification. This is recommended by EANM guidelines for routine image reporting (6) and is therefore a potential competitor to any machine learning tools developed during this work. The following section describes semi-quantification and considers its advantages and disadvantages to establish whether there is scope for machine learning to further increase clinical diagnostic performance.

1.2 Semi-quantification

Semi-quantification enables an objective assessment of an image to be performed, which is designed to help clinicians better and more consistently assess nigrostriatal dopaminergic function. Numerous commercial software solutions are available, including DaTQUANT (GE Healthcare) and BRASS (Hermes Medical).

Semi-quantification involves measurement of tracer uptake within regions of interest, placed over organs that are key to differential diagnosis (i.e. the striatum or subsections of the striatum such as the putamen and caudate, see Figure 1-2 for a typical example). The average voxel intensity (and hence tracer uptake concentration) within these regions is usually compared to another region of the brain, with low uptake, which represents non-specific uptake of the tracer. The ratio of the two values gives the specific to non-specific uptake ratio or striatal binding ratio (SBR). In this thesis SBR is calculated according to:

$$SBR = \frac{C_S - C_B}{C_B} \qquad \textit{Eq 1.1}$$

Where C_S refers to the mean count level within a striatal region (or sub-region), which may be defined on a full 3D volume or summed 2D slices, and C_B refers to the mean count level within a background region, such as the occipital lobe. In addition, other ratios are often calculated as part of semi-quantitative analysis, such as left to right asymmetry ratios and caudate to putamen ratios. The regions of interest used to define the boundaries of striatal uptake are often small and are often defined on a chosen template image. Each test image is then usually registered to the template in order that regions of interest can be applied automatically. Alternative methods have also been proposed. For example, the Southampton method (39) applies a wide region of interest around the individual striata, using manual placement. Background, non-specific uptake is estimated from the remainder of the brain.

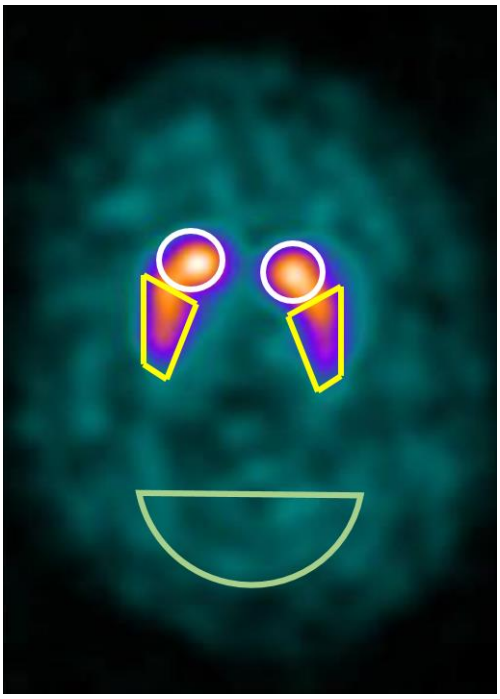


Figure 1-2 Example of the regions of interest used in the calculation of SBR. Caudate regions are shown in white, putamen regions in yellow and the region covering the occipital lobe in green

Whichever particular method is used to define and place regions of interest, the calculated SBRs (and other ratios) are usually provided to the clinician alongside data on expected

values for ‘normal’ (and possibly ‘abnormal’) patients, where ‘normal’ refers to either healthy controls or patients without dopaminergic deficit and ‘abnormal’ covers any patients with pre-synaptic dopaminergic deficit. This gives some context to the SBR figures.

One of the major reasons why interpretation of (I123)FP-CIT images can be difficult through visual analysis alone, and why semi-quantification is recommended, is that normal striatal tracer uptake is known to decline naturally with increasing patient age (40). It is difficult for a human to visualise precisely how images appearances should change with patient age and so it can be challenging to appreciate how the tolerances on normal appearances should be adjusted for each patient. For this reason normal ranges reported with SBR results are often age-matched, for example only considering SBRs from reference patients that are within +/- 5 years of the test patient.

Another justification often presented for the use of semi-quantification software in clinic is that in a minority of cases nigrostriatal deficit can manifest as balanced loss of DaT throughout the striatum, as mentioned previously, maintaining comma-shaped striatal appearances on reconstructed images even at advanced stages of disease. In these cases reporters must examine the contrast between voxel intensities within striatal structures, as compared to non-specific uptake in the rest of the brain, in order to identify that disease is present. Appreciating the exact intensity threshold (and hence display colour) of background tissues that indicates abnormality can be difficult. The fact that striatal tissues maintain a classic normal shape could be sufficient to distract the reporter from making the correct interpretation. Semi-quantification is easily able to highlight these ‘balanced loss’ cases as SBRs are simply a ratio of counts within striatal regions as compared to non-specific uptake regions.

1.2.1 Impact on clinical performance

A number of studies have previously sought to estimate the added value that semi-quantification brings. This data gives a useful indication as to the level of performance gain that may be possible with image analysis tools, and may provide some justification for pursuit of more sophisticated machine learning solutions.

Albert and colleagues (41) examined 62 historical patient datasets, where SPECT imaging had originally been reported as inconclusive. Reference diagnosis was established from clinical follow-up. Following re-reconstruction with different parameters each image was

reported visually by 2 reporters and then semi-quantification was performed using BRASS. Any study where SBR figures were less than 2 standard deviations from the mean of an age-matched normal comparison set was considered abnormal. The accuracy of visual analysis alone was found to be 89%, in line with many of the studies highlighted in section 1.1.5. Accuracy from semi-quantification alone was 85%. Where semi-quantification and visual analysis were in concordance the accuracy was 94%, evidence that, if in agreement, semi-quantification may add confidence to visual analysis.

Along similar lines, Ueda and colleagues (42) and Suarez-pinera and colleagues (43) examined retrospective clinical data to compare the performance of semi-quantitative software with that of visual analysis alone, and then examined results from the two approaches combined. Ueda found that visual analysis had a higher sensitivity but equal specificity to semi-quantification, and that a combined approach (where results agreed) gave an even higher sensitivity (96.7%) than either in isolation (42). Suarez-pinera found no significant difference between semi-quantification and visual analysis, and found no added performance benefit from combining the two approaches (43). However, the dataset used in this case was small (32 cases), limiting the chances of measuring significant differences between approaches. In both of these studies, the optimum cut-off for the semi-quantification classification was defined from the same data to which it was applied to measure classification performance. Therefore, performance figures are likely to represent an overestimate.

Focusing on studies where reporters were exposed to semi-quantitative output there is again a collection of relatively small scale investigations in the literature. The largest such study included 304 cases from previous clinical trials, using clinical diagnosis as the reference standard. Each case was read by 5 reporters with limited clinical experience, first using visual analysis alone and then repeated with semi-quantification results available (44). It was found that sensitivity was almost identical between the two approaches and that the introduction of semi-quantification increased mean specificity slightly (87.9% vs 89.9%). Interestingly, the mean confidence score of the reporters increased significantly when the semi-quantification results were available as compared to when performing visual analysis alone, apparently an advantage of semi-quantification may be in decreasing diagnostic uncertainty.

Two other studies of semi-quantification performance were carried out based on similar assumptions. Soderlund and colleagues (45) and Pencharz and colleagues (46) examined

the variability in reporting both with and without the assistance of semi-quantification software. Soderlund, using a dataset of 54 historical cases, found that mean inter-reporter variability was kappa = 0.8 for visual analysis alone. This is similar to the variability results found in 1.1.5. When reporters were given access to SBR results kappa increased to 0.86. When both SBR results and caudate-to-putamen ratios were provided to reporters the variability between them reduced further (kappa = 0.95) (45). Pencharz, using 109 historical patient cases, found that there was no difference in accuracy between visual analysis and visual + semi-quantification combined. However, they also found that the mean number of cases per reporter that were reported as equivocal reduced from 10.6 to 3.6 after introduction of semi-quantification results (46).

These results, taken together, confirm that semi-quantification offers some benefit in clinical practice (its usefulness in clinical trials is not considered). There is no compelling evidence of a significant increase in sensitivity or specificity as a result of introducing semi-quantification to the reporting process. However, it does appear that when semi-quantification and visual analysis agree, the diagnostic accuracy of the combined results is likely to be very high. When used by image reporters, semi-quantification seems to increase confidence in image reports and there is evidence that inter-observer variability reduces as a result. These findings may partly explain why semi-quantification continues to be in routine clinical use, particularly in Europe. Conversely, the relatively modest gains achievable with semi-quantification may explain why SNM guidelines suggest that semi-quantification is not an absolute necessity (32).

Semi-quantification is an imperfect tool for assisted image reporting. Firstly, due to the small, tight regions of interest that are often used, results usually rely on accurate registration of the test image to a template. Small errors in registration can cause big differences in the quantities measured. Secondly, semi-quantification results are usually provided to clinicians in the form of multiple SBR results (and possibly other ratio figures), each with an associated normal range or suggested normal / abnormal cut-off value. The clinician must interpret each of the SBR scores in light of normal ranges to come to an overall decision on patient diagnosis. Therefore, there is still a significant amount of interpretation required by the reporter after image analysis. Thirdly, semi-quantification is a relatively crude classification tool. It takes no account of the shape of striatal uptake or the distribution of voxel values, or any other image features which could be affected by disease processes. Finally, it is well known that semi-quantification is highly sensitive to differences in gamma camera equipment, scanning protocols and reconstruction methods (47–50). This is likely to be more

pronounced than the effects on visual analysis (as humans are less likely to be distracted by a slight difference in noise, for example). This dictates that individual hospitals may need to define their own normal ranges for SBR figures.

For these reasons there are benefits to be obtained from improved (I123)FP-CIT reporting software. Machine learning algorithms may be able to overcome some or all of the limitations associated with conventional semi-quantification methods and, given the industrious activity in this area, it is hypothesised that established machine learning technology is already sufficiently mature to offer improved performance in clinic. This work focuses on selection, implementation and evaluation of machine learning software to establish whether such systems offer effective diagnostic support to reporters. To this end, the following sections give an overview of machine learning algorithms along with a summary of the techniques applied to (I123)FP-CIT SPECT imaging in the recent literature, before setting out the aims of this work. Although the focus of much of the following section is on machine learning, there is no aspiration to develop a completely new algorithm, the main goal is to critically evaluate existing techniques in a clinical reporting scenario.

1.3 Machine Learning

Machine learning is a wide, rapidly evolving field. It is increasingly used in a variety of practical applications, from controlling driverless cars to computer game development. In research, machine learning is often applied to large datasets in order to identify complex patterns, which can then be used to inform future decisions. In this thesis machine learning is used as a tool for developing a whole-image automated classification system, to perform a reporting task in a similar manner to a radiologist. Specifically, the goal is to implement and evaluate a system, which when presented with a previously unseen image, is able to classify it as belonging to one of two patient groups (dopaminergic deficit and non-dopaminergic deficit groups). The intention is not to replace the radiologist but to provide an independent check of the likely differential diagnosis associated with an image. This independent reading will be presented to the clinician with the aim of improving his/her reporting performance. Software performing this task is often referred to as Computer Aided Diagnosis (CADx) software. This is very similar to but distinct from Computer Aided Detection (CADe) software, which identifies the locations of possible abnormalities within an image to a clinician.

Given that the focus of the thesis is not on development of a new machine learning algorithm per se, the following section provides only a high level introduction to machine learning

theory. It is not intended to provide an in depth technical review of all aspects of machine learning technology. This introductory section is followed by details of specific recent examples in the literature related to (I123)FP-CIT imaging, which have previously produced promising results. In chapters 2 and 3 a selection of these tools will be adapted and critically evaluated to identify a candidate algorithm for use in a reporting exercise. These later sections offer a focused insight into the chosen machine learning theory.

1.3.1 Overview

Fundamentally there are two main types of machine learning algorithm, supervised and unsupervised. Supervised algorithms use databases of labelled training data in order to define a mapping from the features of the training data to their pre-defined label. Thus, the chosen algorithm learns to associate a particular grouping of training data with a particular set of feature values. This is the form of machine learning most often applied in medical imaging and is typically used for regression and classification problems. For example, based on historical data, a supervised algorithm could be trained to predict organ size given a particular age value. Alternatively, machine learning could be used to create a model that learns to differentiate between tumours and healthy tissue based on the pixel intensity values. The general concept of supervised learning is depicted in Figure 1-3.

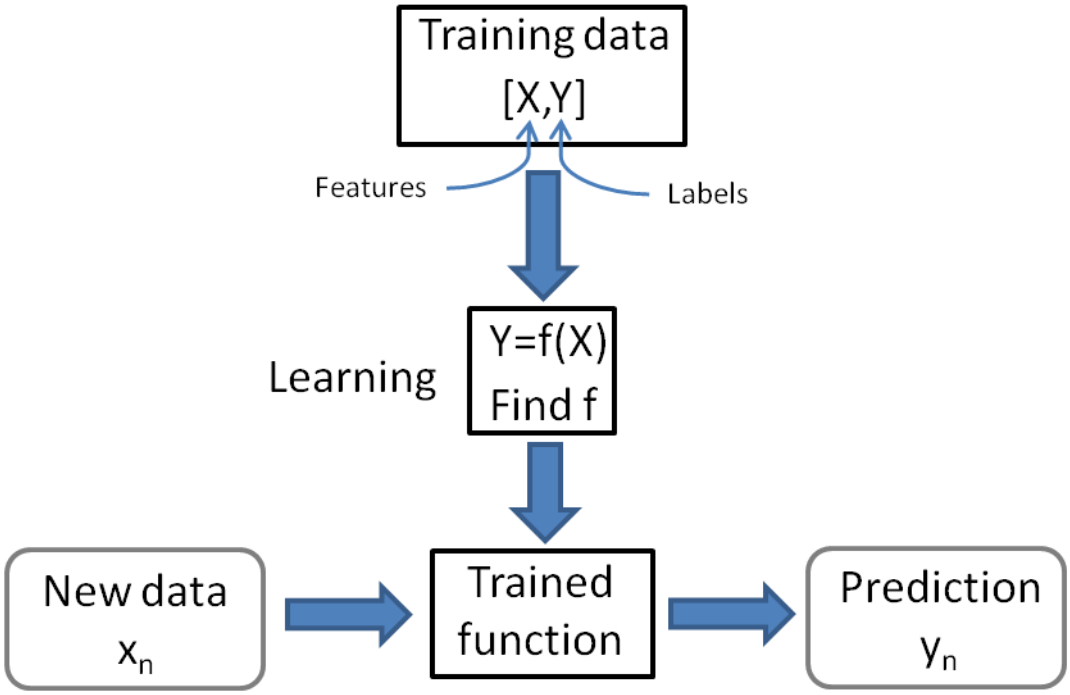


Figure 1-3 Supervised learning concept represented as a workflow

Conversely, unsupervised learning is applied to data where no pre-defined label (or 'ground truth' is available). The goal here is to model the underlying structure of the data. Typical applications include clustering, where the algorithm tries to find inherent groupings in the training set. Another potential application is in dimensionality reduction, where the goal is to find a more compact representation of the same data.

In this work only supervised algorithms are considered, in line with the majority of medical imaging research. As previously mentioned, the task is to classify (I123)FP-CIT images as belonging to one of two different patient groups. In classical machine learning theory the way to approach this problem would be to first define the likely image features that would offer the best chance of accurate classification. This may simply be the raw voxel intensity values, or could be derived features obtained from analysis of the shape of segmented structures, for example. Often, a large number of image features are first presented to the classifier and those features contributing little to the overall performance are removed. In recent years however, 'hand-crafting' features in this way has become less popular, largely due to the recent dominance of deep learning algorithms, such as convolutional neural networks (CNNs) in image analysis research (51). With CNNs a hierarchy of features is derived by the neural network itself as part of the overall training process, the advantage being that the classification algorithm finds the best features for the task and particular set of data it is presented with.

In the following section some of the techniques previously applied to (I123)FP-CIT classification will be highlighted, to give an indication of the tools that are currently available for use in CADx systems. A subset of these techniques will be adopted in this thesis for initial evaluation. Clinical tests of CADx software, in a simulated reporting scenario, will be based on the best performing algorithm from these initial results.

1.3.2 Automated classification for (I123)FP-CIT: a summary

Automated classification tools for (I123)FP-CIT imaging, based on machine learning methods, have been investigated by numerous authors. Details on the techniques applied in the available literature since 2010, to distinguish between patients with and without pre-synaptic dopaminergic deficit, are summarised in Table 1-3. Also included are details of the image features extracted, reported performance metrics, details of the data upon which the performance figures were derived and information related to the chosen cross-validation

technique. The table includes algorithms where training data is based on SPECT images only, multimodality inputs are excluded.

Summary of automated classification research for (I123)FP-CIT imaging since 2010 (articles ordered according to accuracy)				
Authors	**Image features (if applicable)**	**Classifier**	**Validation data + method**	**Results**
Augimeri, Cherubini, Cascini et al., 2016 (52)	Mean ellipsoid uptake, dysmorphic index (ellipsoid orientation)	Support Vector Machine (SVM)	43 local images (12 normal, 31 Parkinson's Disease (PD)), no cross validation mentioned	Up to 100% accuracy, specificity and sensitivity
Bhalchandra, Prashanth, Roy et al., 2015 (53)	Analysis of 42nd slice only. Striatal binding ratios in both caudates and putamena, radial features and gradient features. Features are tested for statistical significance (wilcoxon rank) before use in the classifier	SVM and SVM with Radial Basis Function (RBF) kernel, Linear Discriminant Analysis (LDA)	350 images from Parkinson's Progression Markers Initiative (PPMI) database (187 healthy controls (HC), 163 PD). 5 fold cross-validation (CV), repeated 100 times	Linear SVM. Maximum of: Accuracy = 99.4% RBF kernel. Maximum of: Accuracy = 99.4% LDA. Maximum of: Accuracy = 99.4%
Choi, Ha, Im et al., 2017 (54)	All voxels within the image	CNN – PD net	701 images from the PPMI database (431 PD, 193 HC, 77 scans without evidence of dopaminergic deficit (SWEDD)). 82 local images (72 PD, 10 non-parkinsonian)	Maximum of: Accuracy = 98.8% Sensitivity = 98.6% Specificity = 100.0%
Oliveira, Faria, Costa et al., 2017 (55)	Binding ratios in the putamen, caudate and striatum, striatal volume and length in both brain	SVM, k-nearest neighbour (k-NN), logistic regression	652 images from the PPMI database (209 HC, 443 PD). Leave-one-out CV	Maximum of: Accuracy = 97.9% Sensitivity = 98.0%

	hemispheres			Specificity = 97.6%
Oliveira, Castelo-Branco, 2015 (56)	Image voxels within striatal region of interest	SVM	654 images from PPMI database (209 HC, 445 PD). Leave-one-out CV	Maximum of: Accuracy = 97.9% Sensitivity = 97.8% Specificity = 98.1%
Prashanth, Dutta Roy, Mandal et al., 2017 (57)	16 shape and 14 surface fitting features of selected slices, following thresholding. Striatal binding ratios of both caudates and putamena and asymmetry indices were also considered. Features are tested for statistical significance (wilcoxon rank) before use in the classifier	SVM with RBF kernel, boosted trees, random forests, naive bayes	715 images from PPMI database (208 HC, 427 PD, 80 SWEDD). 10 fold CV, repeated 100 times. Hyperparameters for SVM chosen through 10 fold CV	SVM: Accuracy = 97.3 ± 0.1% Sensitivity = 97.4 ± 0.1% Specificity = 97.2 ± 0.2% Boosted trees: Accuracy = 96.8 ± 0.2% Sensitivity = 97.1 ± 0.3% Specificity = 96.3 ± 0.4% Random forests: Accuracy = 96.9 ± 0.2% Sensitivity = 97.2 ± 0.2% Specificity = 96.5 ± 0.3% Naive Bayes: Accuracy = 96.9 ± 0.1% Sensitivity = 96.4 ± 0.1% Specificity = 96.5 ± 0.2%
Tagare, DeLorenzo,	Voxel intensities within a region of	Logistic lasso	658 images from PPMI	Maximum of:

Chelikani et al., 2017 (58)	interest		database (210 HC, 448 PD). 3 fold CV for performance assessment. Parameters chosen through 10 fold CV (nested within outer 3 fold CV).	Accuracy = 96.5 ± 1.3%
Palumbo, Fravolini, Buresta et al., 2014 (59)	Striatal binding ratios for both caudates and putamena (and a subset of these 4 features), patient age	SVM with RBF kernel	90 local images from patients with 'mild' symptoms (34 non-PD, 56 PD). Leave-one-out and 5 fold CV	Maximum of: Accuracy = 96.4%
Prashanth, Dutta Roy, Mandal et al., 2014 (60)	Striatal binding ratio for both caudates and putamena	SVM, linear and with RBF kernel.	493 images from PPMI database (181 HC, 369 early PD), 10 fold CV, no repeats	RBF kernel: Accuracy = 96.1%, Sensitivity = 96.6%, Specificity = 95.0% Linear SVM: Accuracy = 92.3%, Sensitivity = 95.3%, Specificity = 84.0%
Martinez-Murcia, Gorriz, Ramirez et al., 2013 (61)	12 Haralick texture features within a brain region of interest	SVM	'Whole' PPMI database. Leave-one-out CV	Maximum of: Accuracy = 95.9%, Sensitivity = 97.3%, Specificity = 94.9%
Zhang, Kagen, 2016 (62)	Voxel intensities from a single axial	Single layer	1513 images from PPMI	Maximum of:

	slice, repeated for 3 different slices	Neural network	database (baseline and follow-up, 1171 PD, 211 HC, 131 SWEDD). 1189 images for training, 108 for validation, 216 for testing. 10 fold CV	Accuracy = 95.6 ± 1.5% Sensitivity = 97.4 ± 4.3% Specificity = 93.1 ± 3.6%
Rojas, Gorriz, Ramirez et al., 2013 (63)	Voxel intensities, independent component analysis (ICA) & principal component analysis (PCA) decomposition of voxel data (after applying empirical mode decomposition) within regions of interest	SVM	80 local images (39 non-pre-synaptic dopaminergic deficit (non-PDD), 41 PDD). Leave-one-out CV	Raw voxels: Accuracy = 87.5%, Sensitivity = 90.2%, Specificity = 84.6% ICA features. Maximum of: Accuracy = 91.2%, Sensitivity = 91.8%, Specificity = 92.9% PCA features. Maximum of: Accuracy = 95.0%, Sensitivity = 95.1%, Specificity = 94.9%
Martinez-Murcia, Gorriz, Ramirez et al., 2018 (64)	Downsampled voxel intensities	CNNs – modified versions of ALEXNET and LENET5	642 images from PPMI database (194 HC, 448 PD). 10 fold stratified CV	LENET5. Maximum of: Accuracy = 94.9 ± 2.5% Sensitivity = 94.0 ± 4.6% Specificity = 96.9 ± 5.1%

				ALEXNET. Maximum of: Accuracy = 94.1 ± 4.5% Sensitivity = 96.7 ± 2.9% Specificity = 96.9 ± 7.2%
Towey, Bain, Nijran, 2011 (65)	PCA decomposition of voxels within striatal region of interest	Naïve-Bayes, Group prototype	116 local images (37 non-PDD, 79 PDD). Leave-one-out CV	Naïve-Bayes: Accuracy = 94.8%, Sensitivity = 93.7%, Specificity = 97.3% Group prototype: Accuracy = 94.0%, Sensitivity = 93.7%, Specificity = 94.6%
Segovia, Gorriz, Alvarez, 2012 (66)	Partial least squares decomposition of voxels within striatal regions	SVM applied to hemispheres separately. RBF kernel	189 local images (94 non-PDD, 95 PDD). Leave-one-out CV	Features varied from 1 to 20. Maximum of: Accuracy = 94.7%, Sensitivity = 93.2%, Specificity = 93.6%
Martinez-Murcia, Gorriz, Ramirez et al., 2014 (67)	ICA decomposition of selected voxels	SVM, linear and with RBF kernel	208 local images (100 non-PDD, 108 PDD), 289 images from PPMI database (114 normal, 175 PD). 30 fold CV	RBF kernel. Maximum of: Accuracy = 94.7% Sensitivity = 98.1% Specificity = 92.0% Linear SVM. Maximum of: Accuracy = 92.8%

				Sensitivity = 98.2% Specificity = 93.0%
Kim, Wit, Thurston, 2018 (68)	Image voxel intensities in a single axial slice	CNN – Inception v3 network	108 local images for training, 45 for hold out testing	Maximum of: Accuracy = 84.4% Sensitivity = 96.3% Specificity = 66.7%
Illan, Gorriz, Ramirez et al., 2012 (69)	Image voxel intensities & image voxels within striatal region of interest	Nearest mean, linear SVM	208 local images (108 non-PDD, 108 PDD). 30 random permutations CV, with 1/3 data held out for testing	SVM. Maximum of: Sensitivity = 89.0%, Specificity = 93.2% Nearest mean. Maximum of: Sensitivity = 90.7%, Specificity = 84.0% k-NN. Maximum of: Sensitivity = 88.6%, Specificity = 86.9%
Palumbo, Fravolini, Nuvoli et al., 2010 (70)	Striatal binding ratios for caudates and putamena on 3 slices	Probablistic Neural network (PNN), Classification tree (CT)	216 local images (89 non-PDD, 127 PD). Two fold CV, repeated 1000 times	PNN: For patients with essential tremor mean probability of correct classification = 96.6 ± 2.6% CT:

				For patients with Essential tremor mean probability of correct classification = 93.5 ± 3.4%

Table 1-3 Summary of machine learning algorithms applied to (I123)FP-CIT image classification in the literature since 2010, including reported performance figures. Articles are listed in order of accuracy. Where accuracy values are not available these are grouped towards the bottom of the table. Table is adapted from (1)

A number of trends are immediately apparent from examination of Table 1-3. Firstly, the reported performance figures are universally high. Most accuracy values are greater than 90%, with some authors reporting almost perfect performance. This contrasts with accuracy figures previously summarised for visual image analysis (see section 1.1.5), and for semi-quantification (see section 1.2.1), which were typically in the 80-90% range. These results clearly show that established machine learning algorithms are a promising technology for creating CADx software. As in previous discussions however, these figures should be treated with a degree of caution. Not only is performance likely to be strongly related to the particular case mix in the database but the method of cross validation can also have a significant impact on results (71–73).

The Parkinson's Progression Markers Initiative (PPMI) database of SPECT data (www.ppmi-info.org/data) is cited by most authors as a source of validation data. This is perhaps unsurprising as the data is freely available to researchers, without the need to apply for ethical approval or to go through other lengthy governance processes. As patients were recruited prospectively, following a battery of other tests and screening stages, the diagnostic coding is likely to be relatively reliable. The other advantage of using the PPMI data is that it allows greater comparability between research studies. However, this research database is unlikely to reflect the patient cohorts seen in clinical nuclear medicine. The patient groups included are healthy controls, Parkinson's Disease and scans without evidence of dopaminergic deficit (SWEDD). In clinic, a range of atypical Parkinsonisms are seen, as well as DLB and other diseases which do not affect nigrostriatal pathways. Furthermore, patients were only included in the PD group if their SPECT scan showed DaT deficit (74), which may have excluded any patients for which signs of disease were subtle. The strict controls on imaging protocols, camera calibration steps and image reconstruction (75,76) also do not reflect clinical reality.

The range of classifiers used by researchers is wide, although support vector machines (SVM), either in conventional linear form, or with a radial basis function (RBF) kernel, appear to dominate. This is likely to be because SVM was considered as a 'state-of-the-art' algorithm up until relatively recently and had been successful in numerous classification problems. The image features extracted and used as input to the classifiers are varied. However, in most cases relatively simple features are chosen (such as raw voxel intensities and SBRs). This suggests that complex pre-processing is not required to achieve good classification performance. In general the most recent articles gave the highest accuracy figures, with some exceptions. This may be because authors have built upon the findings of

previous research work and sought to address limitations that were previously identified. The two-class classification paradigm dominates recent research, where the classifier is trained to separate out two different groups of data. Alternatively, the problem could be considered as a one-class system, where the classifier is trained to find the boundaries of one class within feature space, without explicit reference data from other diagnostic classes.

Overall, analysis of previous literature on automated binary classification of (I123)FP-CIT images confirms that existing machine learning algorithms are associated with high accuracy, which is generally in excess of accuracy figures reported for human observers alone, and human observers assisted by semi-quantification software. However, given the differences in patient datasets, acquisition protocols and analysis methods direct comparison between these different approaches to diagnosis is associated with significant uncertainty.

To date there has not been a direct, comprehensive comparison between semi-quantification methods and machine learning in terms of accuracy or any other performance metrics. Towey (65) did provide a comparison between two automated classifiers and a limited number of commercial semi-quantification tools. However, the dataset used was relatively small and there was a fundamental bias in the findings in that results for the semi-quantification approaches were reported from the training data rather than from an independent test set. Furthermore, no machine learning algorithm has yet been tested in the clinic under realistic reporting conditions (e.g. in support of a human reporter).

If CADx systems based on machine learning algorithms are to be used to benefit patient care these gaps in knowledge need to be filled, which is the main focus of this work.

1.4 Discussion and objectives

The introductory sections have laid out the clinical and technical background to Parkinsonian Syndromes and (I123)FP-CIT SPECT imaging. It was shown that image results can help to differentiate between patients with nigrostriatal dopaminergic deficit and those without, which is particularly useful in the early stages of disease. It is apparent that visual analysis of SPECT images is associated with relatively high but variable accuracy, sensitivity and specificity and that there is evidence of some discrepancies between reporters in some studies. In recent years the use of semi-quantification software as a diagnostic aid has become routine in clinical practice. There are numerous different ways of conducting semi-quantification analysis and there are several commercial products available. The evidence

for clinical impact of semi-quantification is relatively limited, with no significant increase in reporting accuracy seen. However, results from the literature do suggest that greater concordance between reporters and increased confidence in diagnosis is possible when these tools are adopted in the clinic.

There is evidence that machine learning algorithms are already sufficiently mature to offer high accuracy in the binary classification of (I123)FP-CIT images. One of the main advantages of such tools over semi-quantification is that the entire image can be distilled into a single classification metric, rather than a series of SBR figures and normal ranges, which, in an assisted reporting context, greatly simplifies the decision-making process for the reporting clinician and may lead to increased diagnostic performance. However, no direct comparison has yet been conducted against semi-quantification approaches and no tests have yet been conducted to assess the likely impact of existing machine learning algorithms on clinical reporting in a CADx scenario.

Consequently, the main research question for this thesis is:

How effective is a CADx tool, based on established machine learning algorithms, for assisted (I123)FP-CIT image reporting?

Effectiveness will be measured in terms of independent classification accuracy and in terms of the impact upon human reporter accuracy, sensitivity, specificity and inter-reporter reliability. Studies will be conducted utilising realistic clinical data where possible and considering standard-of-care competing technologies (semi-quantification).

In order to meet this research question a number of key objectives are proposed. These are summarised below. Figure 1-4 demonstrates how these objectives fit within the overall thesis workflow.

Objectives:

1) **Select and implement machine learning classification tools.** A limited number of promising machine learning algorithms will be selected and implemented in software for further evaluation
2) **Collect a database of (I123)FP-CIT images**. Data will be extracted from the archives at Sheffield Teaching Hospitals NHS Foundation Trust following ethical

approval. All patient identifiable information will be removed. Gold standard diagnosis will be established from patient records. This “local” database will be supplemented by data from the PPMI repository. All images will be pre-processed to enable further analysis

3) **Compare the performance of machine learning algorithms with semi-quantification.** A comprehensive range of semi-quantification methods will be selected and implemented. Cross validation will be carried out on the Sheffield data and the PPMI database to quantify the standalone effectiveness of machine learning-based classification algorithms as compared to semi-quantification methods. This will provide an indication as to whether machine learning offers added benefits over existing assistive reporting technology, and will help to identify a single algorithm for use in the reporting exercise
4) **Develop software for testing of human reporters.** Software will be created to enable measurement of human observer performance in reporting (I123)FP-CIT images. The software will mimic the interface used clinically for reporting patient data.
5) **Assess the impact of an automated classification tool, implemented as a CADx system, on reporting.** After selecting the best performing machine learning algorithm from cross-validation results, studies will be conducted to assess the magnitude of impact of a CADx system on reporter performance, both quantitatively and qualitatively. This will be carried out via an initial, smaller-scale pilot study, followed by a larger scale clinical evaluation.

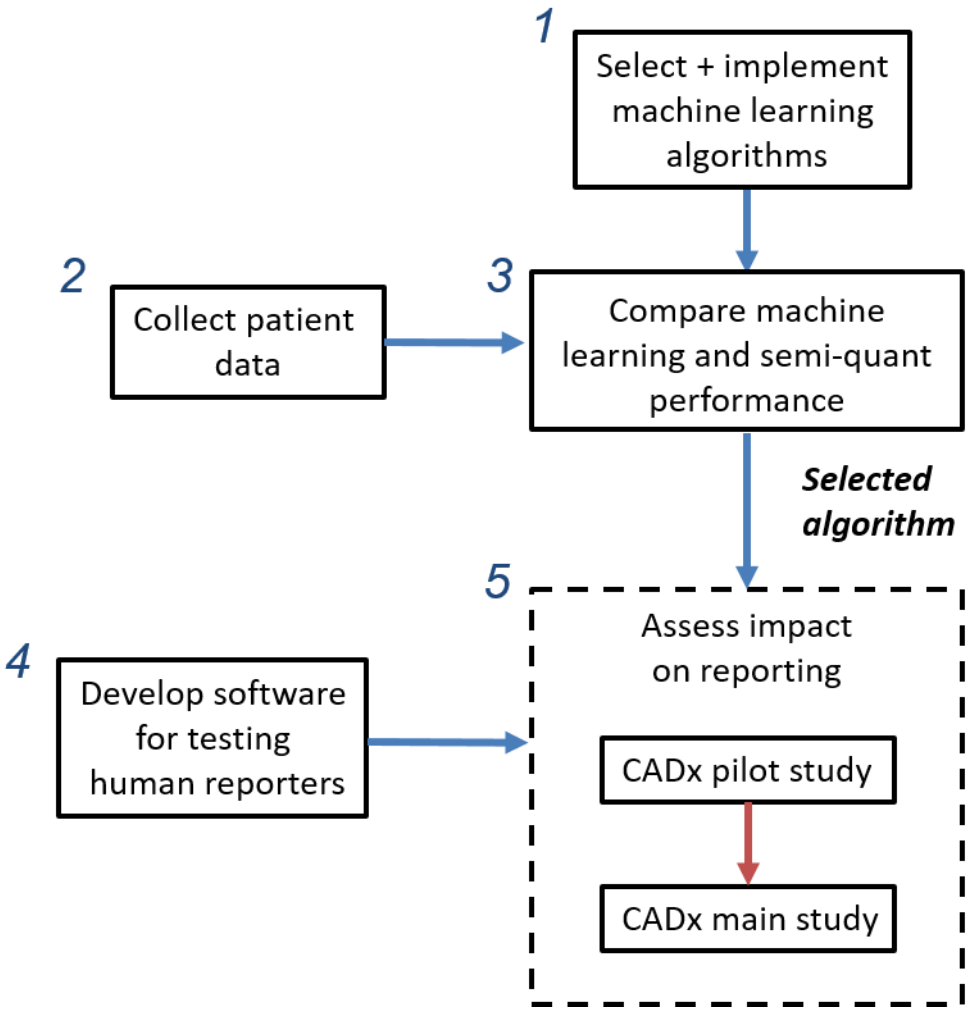


Figure 1-4 Thesis workflow

In line with objectives 1 and 2, the next chapter in this thesis focuses on implementing algorithms and gathering data, laying the groundwork for chapter 3, where the standalone performance of machine learning algorithms and semi-quantification methods is compared. Chapter 4 then addresses objectives 4 and 5 through a reporting exercise, using a selected machine learning tool implemented as a CADx system.

2 Algorithms and data

2.1 Machine learning algorithms

Objectives addressed by this section (in black, bold):
1) Select and implement machine learning classification tools
2) Collect a database of (I123)FP-CIT images
3) Compare the performance of machine learning algorithms with semi-quantification
4) Develop software for testing of human reporters
5) Assess the impact of an automated classification tool, implemented as a CADx system, on reporting

Table 2-1 Objectives addressed in section 2.1

This section focuses on identifying machine learning approaches from the literature that are likely to give the highest classification performance, and implementing them in software in preparation for a direct comparison with semi-quantification methods and a reporting study. As shown previously, classical machine learning algorithms require selection of both image features and a classifier.

From Table 1-3 it is clear that SVMs (with and without RBF kernel) are the most prevalent classifiers from recent research and are associated with the highest classification scores. These classifiers were therefore chosen for implementation and evaluation. Further theoretical justification for selecting SVMs is provided in the following background sections.

A number of different image features have been used as inputs to SVMs in previous work and there is no observable trend highlighting the particular suitability of one feature over another. This is largely due to the fact that most previous studies have involved selection and comparison of relatively few different feature types.

Therefore, the features extracted for this thesis had to be selected according to different criteria. Features are chosen with the aim of reducing the image pre-processing required, maximising the potential for automation and minimising algorithm complexity. This will reduce the risk of unforeseen errors occurring in the reporting exercise and will reduce the

potential for increased uncertainties in pre-processing. If, for example, a more complex derived feature related to shape were chosen, image segmentation may be required, which is an imperfect process that can itself be affected by issues such as image noise.

From the list of previously used features the following were chosen for initial evaluation: image voxels in the reconstructed image, SBRs and principal components of image voxels. These features have the added advantage that they have been investigated by multiple authors using different approaches and have all been associated with promising results. It was decided that features would not be combined but would be considered separately, as is largely the case in the recent literature.

Diagnostic input features from other modalities were not included, as development of a CADx tool that does not rely on results of other tests being available, is likely to be more amenable to clinical translation. However, due to the known age dependency of tracer uptake, patient age was used as an added input to the classification algorithm in all cases. This ensured that the classifier modelled the relationship between age and other features.

A summary of the different features and classifiers chosen for implementation and evaluation is shown in Figure 2-1.

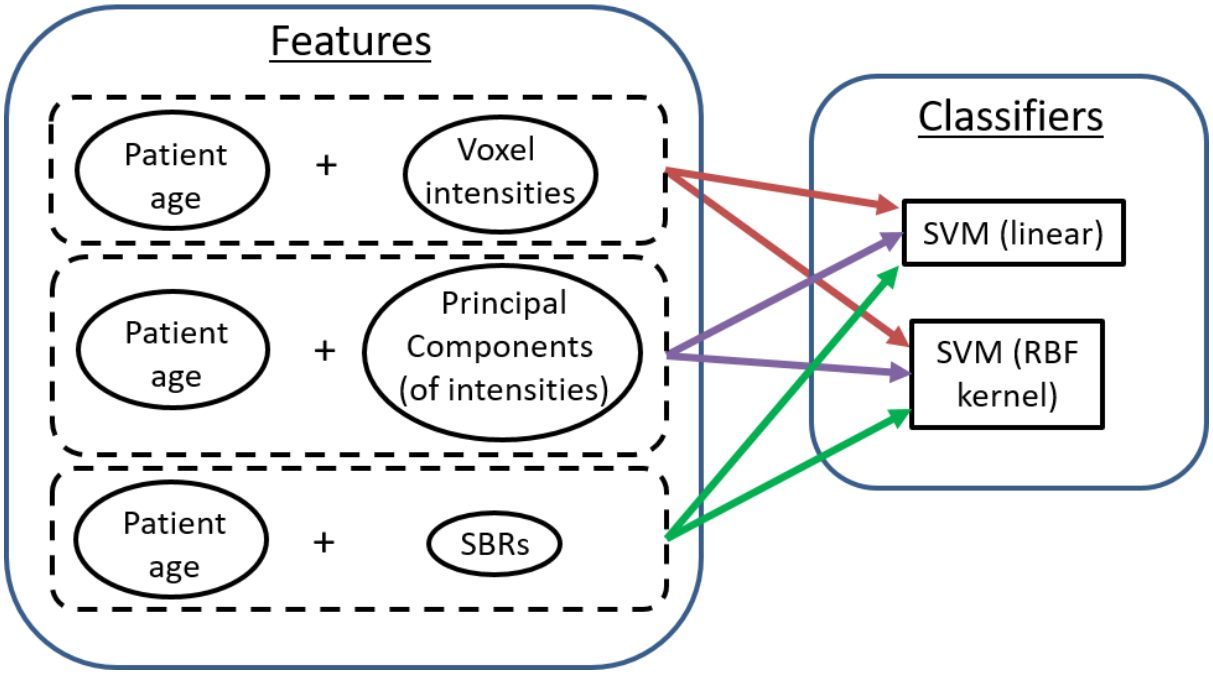


Figure 2-1 Summary of the feature – classifier combinations selected for implementation and evaluation

The following two sub-sections give a brief technical summary of Principal Component Analysis and Support Vector Machines respectively, two of the main methodologies chosen to form automated classification tools. The goal here is not to provide an exhaustive critique or in-depth mathematical analysis but to give enough information that the major advantages (and disadvantages) of the selected technologies used in this work can be demonstrated. These background sections are followed by details of the implemented machine learning pipelines.

2.1.1 Technical background – Principal Component Analysis

PCA is primarily used for dimensionality reduction, enabling representation of data with very large numbers of variables by a much smaller number of common components. One of the main benefits of PCA is that it can dramatically reduce computational time for classifier training. It is a technique that was established in the early 1900s and has been applied to numerous varied applications in the intervening years (77).

PCA takes a set of observations (i.e. images) and projects them to a new subspace whose axes are orthonormal (78). PCA attempts to maximise the variance of the data along each projected axis. Thus, the majority of the variance in the original dataset can be described by the first few axes or principal components of the new space, thereby achieving a significant reduction in the number of dimensions required to adequately reconstruct the data. The magnitude of the linear components (which when combined reconstruct the object) uniquely characterises the object in PCA space. A simple example of PCA is demonstrated graphically in Figure 2-2. Here, PCA is applied to a set of two dimensional points. The first principal component (PC1) is placed along the line of highest variance. As can be seen, if the data points were to be described only by their position along PC1 they would be relatively well represented in the space, with little residual error. Thus, the same data can be well characterised using half the number of original variables. Given that PCA does not consider the labels of a particular class of data it is an unsupervised technique.

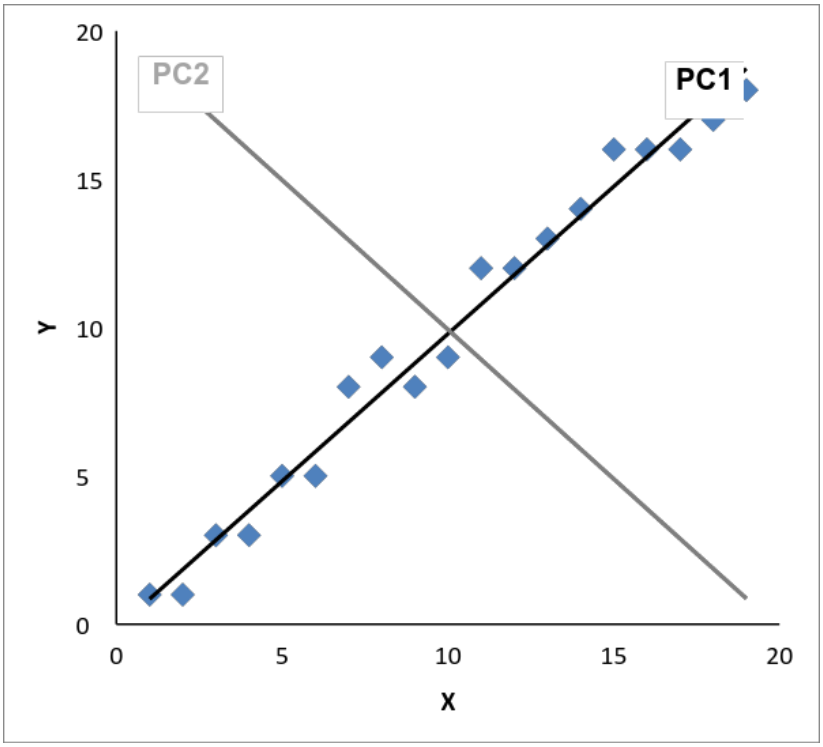


Figure 2-2 Example of PCA applied to a two dimensional dataset

In mathematical terms PCA is a linear transformation of the form:

$$\boldsymbol{Y} = \boldsymbol{PX} \qquad \textit{Eq 2.1}$$

Where $\boldsymbol{X}$ is the input sample matrix (whose rows represent variables and columns represent observations), $\boldsymbol{P}$ is the projection matrix and $\boldsymbol{Y}$ is the transformed output. In image processing research the input matrix is often composed of separate vectors, each of which is a collection of pixel values from separate training images that have been concatenated into a 1 dimensional form.

The variance-covariance matrix of $\boldsymbol{Y}$ defines the extent to which each of the variables within $\boldsymbol{Y}$ are linearly associated with each other (i.e. covariance), and also the spread along each axis (i.e. variance). Variances occupy the diagonal matrix positions and co-variances occupy all other off-diagonal positions. For a sample of data taken from a larger population, covariance and variance can be calculated according to:

$$cov(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum \frac{(\alpha_i - \bar{\alpha})(\beta_i - \bar{\beta})}{n-1} = \frac{1}{n-1}[(\boldsymbol{\alpha} - \bar{\alpha})][(\boldsymbol{\beta} - \bar{\beta})]^T \qquad \textit{Eq 2.2}$$

Where α_i, β_i represent individual values associated with two different variables within the sample, $\boldsymbol{\alpha}$,$\boldsymbol{\beta}$ are vectors of all values associated with the two variables, and $\bar{\alpha}$, $\bar{\beta}$ represent the variable means. The number of data points in the sample is defined by n. From the above it can be shown that the covariance of data with itself, i.e. $cov(\boldsymbol{\alpha},\boldsymbol{\alpha})$ or $cov(\boldsymbol{\beta},\boldsymbol{\beta})$, reduces to a statistic which is simply the variance of that data. It also follows that $cov(\boldsymbol{\alpha},\boldsymbol{\beta})$ = $cov(\boldsymbol{\beta},\boldsymbol{\alpha})$.

If $\boldsymbol{Y}$ is a two dimensional system containing 4 samples, i.e. $\boldsymbol{Y} = \begin{pmatrix}\alpha_1 & \alpha_2 & \alpha_3 & \alpha_4\\ \beta_1 & \beta_2 & \beta_3 & \beta_4\end{pmatrix}$, the variance-covariance matrix, $\boldsymbol{C_Y}$, can be written as:

$$\boldsymbol{C_Y} = \begin{pmatrix}cov(\boldsymbol{\alpha},\boldsymbol{\alpha}) & cov(\boldsymbol{\alpha},\boldsymbol{\beta})\\ cov(\boldsymbol{\beta},\boldsymbol{\alpha}) & cov(\boldsymbol{\beta},\boldsymbol{\beta})\end{pmatrix} = \frac{1}{n-1}\begin{pmatrix}[(\boldsymbol{\alpha}-\bar{\alpha})][(\boldsymbol{\alpha}-\bar{\alpha})]^T & [(\boldsymbol{\alpha}-\bar{\alpha})][(\boldsymbol{\beta}-\bar{\beta})]^T\\ [(\boldsymbol{\beta}-\bar{\beta})][(\boldsymbol{\alpha}-\bar{\alpha})]^T & [(\boldsymbol{\beta}-\bar{\beta})][(\boldsymbol{\beta}-\bar{\beta})]^T\end{pmatrix} = \begin{pmatrix}var(\boldsymbol{\alpha}) & cov(\boldsymbol{\alpha},\boldsymbol{\beta})\\ cov(\boldsymbol{\beta},\boldsymbol{\alpha}) & var(\boldsymbol{\beta})\end{pmatrix} \quad \textit{Eq 2.3}$$

Where $var(\boldsymbol{\alpha})$ and $var(\boldsymbol{\beta})$ are the variances of the two variables. If all the variables are mean centred (i.e. mean = 0) then the variance-covariance matrix simplifies to:

$$\boldsymbol{C_Y} = \frac{1}{n-1}\begin{pmatrix}\boldsymbol{\alpha}\boldsymbol{\alpha}^T & \boldsymbol{\alpha}\boldsymbol{\beta}^T\\ \boldsymbol{\beta}\boldsymbol{\alpha}^T & \boldsymbol{\beta}\boldsymbol{\beta}^T\end{pmatrix} = \frac{1}{n-1}\boldsymbol{Y}\boldsymbol{Y}^T \quad \textit{Eq 2.4}$$

In most practical applications, such as in medical image processing, the matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ are likely to contain many more than 2 variables and 4 samples. Substituting the output matrix $\boldsymbol{Y}$ in Eq 2.4 for the input matrix $\boldsymbol{X}$ (from Eq 2.1), and expanding for an unspecified number of variables, the following relationship is obtained:

$$\boldsymbol{C_Y} = \frac{1}{n-1}\boldsymbol{P}\left(\boldsymbol{X}\boldsymbol{X}^T\right)\boldsymbol{P}^T = \frac{1}{n-1}\boldsymbol{P}\begin{pmatrix}\boldsymbol{x_1}\boldsymbol{x_1^T} & \boldsymbol{x_1}\boldsymbol{x_2^T}\ldots & \boldsymbol{x_1}\boldsymbol{x_m^T}\\ \boldsymbol{x_2}\boldsymbol{x_1^T} & \boldsymbol{x_2}\boldsymbol{x_2^T}\ldots & \boldsymbol{x_2}\boldsymbol{x_m^T}\\ \vdots & \vdots & \vdots\\ \boldsymbol{x_m}\boldsymbol{x_1^T} & \boldsymbol{x_m}\boldsymbol{x_2^T}\ldots & \boldsymbol{x_m}\boldsymbol{x_m^T}\end{pmatrix}\boldsymbol{P}^T \quad \textit{Eq 2.5}$$

where each x_i is a vector containing all the values or observations for one particular input variable (i.e. each row of $\boldsymbol{X}$). PCA attempts to maximise the variance of the data along each projected axis (and consequently to minimise the co-variance between axes). The role of the projection matrix $\boldsymbol{P}$ is therefore to ensure that the co-variances of the transformed matrix $\boldsymbol{Y}$ are as close to zero as possible and the variances as large as possible. In effect PCA attempts to diagonalise $\boldsymbol{C_Y}$. One method for achieving matrix diagonalization is to perform eigen-decomposition, which for the square, symmetric matrix $\boldsymbol{XX^T}$ is defined as:

$$\boldsymbol{XX^T E = ED}$$
$$\boldsymbol{XX^T = EDE^T} \qquad \textit{Eq 2.6}$$

Where $\boldsymbol{E}$ is an orthonormal matrix containing the eigenvectors of $\boldsymbol{XX^T}$ and $\boldsymbol{D}$ is a diagonal matrix containing the (real) eigenvalues. See Figure 2-3 for an illustrative example of this eigen-decomposition concept.

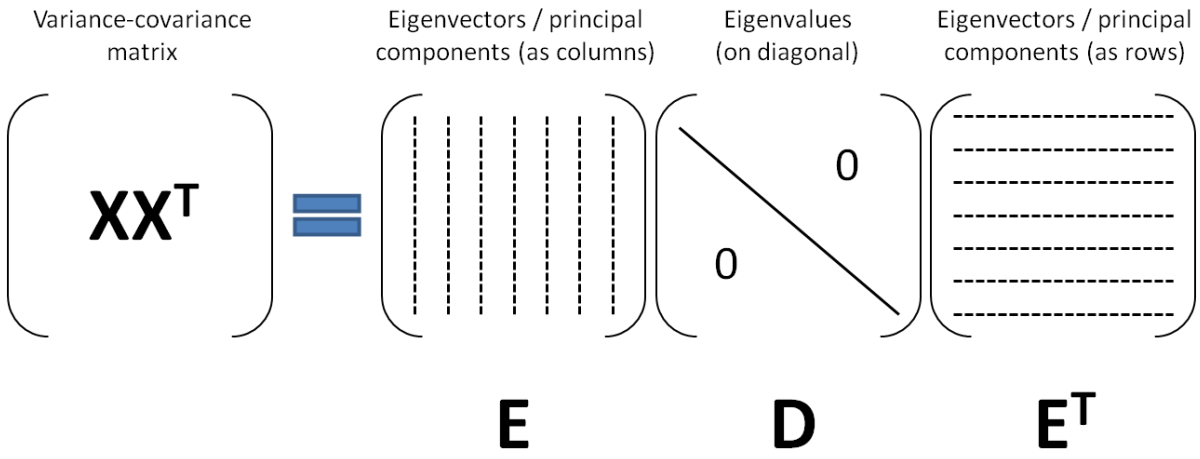


Figure 2-3 Illustration of the eigen-decomposition concept as applied to the variance-covariance matrix $\boldsymbol{XX^T}$

If the rows of the projection matrix $\boldsymbol{P}$ are chosen to equal the eigenvectors (columns of $\boldsymbol{E}$), such that $\boldsymbol{P = E^T}$, then the equation for the variance-covariance matrix becomes:

$$\boldsymbol{C_Y} = \frac{1}{n-1}\boldsymbol{E^T}\left(\boldsymbol{EDE^T}\right)\boldsymbol{E} = \frac{1}{n-1}\boldsymbol{D} \qquad \textit{Eq 2.7}$$

Thus, the variance-covariance matrix of the output has been reduced to a diagonal matrix (of eigenvalues), which was the original goal of PCA. The principal components are the rows of $\boldsymbol{P}$ and are equal to the eigenvectors of $\boldsymbol{XX^T}$. There are as many principal components (with non-zero eigenvalues) as there are observations in the training data. In image processing research this is likely to be far fewer than the number of variables (i.e. number of voxels). The principal components are usually stated in order of reducing variance such that the first component describes the largest amount of variance in the data.

In image processing and classification it is common to only use a small selection of derived principal components, often by choosing those which together represent a certain percentage of the overall variance in the data (77). This is because lower components with lower variance are often more likely to be made up of noise. In SPECT, where image noise is significant in comparison to other modalities, using only the first few principal components can help to remove the confounding effects of noise from algorithm training processes (78), which may help to improve the accuracy and robustness of the trained machine learning tool.

The above eigen-decomposition is a common method for deriving principal components. However, other techniques are available. Singular value decomposition (SVD) is a computationally efficient method for deriving components that has been used in many applications (77). SVD theory states that the matrix $\boldsymbol{X}$ can be decomposed as follows:

$$\boldsymbol{X} = \boldsymbol{U}\sum\boldsymbol{V^T} \qquad \textit{Eq 2.8}$$

Where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal matrices and $\sum$ is a diagonal matrix containing the singular values of $\boldsymbol{X}$.

If the variance-covariance matrix, $\boldsymbol{XX^T}$, is expanded according to the SVD definition, the following equation is obtained:

$$\boldsymbol{XX^T} = \left(\boldsymbol{U}\sum\boldsymbol{V^T}\right)\left(\boldsymbol{U}\sum\boldsymbol{V^T}\right)^T = \boldsymbol{U}\sum^{\boldsymbol{2}}\boldsymbol{U^T} \qquad \textit{Eq 2.9}$$

Comparison of Eq 2.9 with the eigen-decomposition in Eq 2.6 demonstrates that the (non-zero) singular values of $\boldsymbol{XX^T}$ are equivalent to the square roots of the (non-zero) eigenvalues of $\boldsymbol{XX^T}$. It is also apparent that $\boldsymbol{U}$ is a matrix containing the eigenvectors of $\boldsymbol{XX^T}$. Thus, by

calculating the singular value decomposition of $\boldsymbol{X}$ it is possible to derive the principal components (and eigenvalues).

In the following chapters SVD is used to compute principal components from training data. Test data are projected on to these components by matrix multiplication. For example, a test image, $\boldsymbol{f}$ (in vector form), can be projected on to the basis represented by $\boldsymbol{U}$ through the following equation: $\boldsymbol{c} = \boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{f}$, where $\boldsymbol{c}$ is a vector of coefficients describing the position of the image in the principal component (PC) subspace. The distribution of test image coefficients is used as an input to SVM classification in the following investigations. Unless otherwise stated it is always assumed that training and test data are mean centred (i.e. a mean image is subtracted from each case).

PCA is a linear technique and as such its ability to represent systems where the underlying interactions between variables (or features) are non-linear is limited. Thus, more recently kernel PCA was introduced, whereby PCA is applied in a modified space dictated by a kernel function (79). Conventional PCA as a precursor to classification is also limited by its unsupervised nature. Choosing components according to the largest variance is not necessarily the best method for choosing a basis on to which to apply a classifier. However, due to its simplicity and wide-ranging, successful application in previous tasks (including for the task at hand) only conventional, linear PCA is applied as a dimensionality reduction method in this work.

2.1.2 Technical background – Support Vector Machines

A conventional SVM classifier is a supervised learning approach which attempts to draw a discrimination boundary between two classes of data. The boundary is created in such a way as to maximise the width of the margin between the samples in each class (a maximum-margin approach). SVM has been successfully used in numerous diverse applications. For example, recent reviews provide an insight into its contribution to the fields of computational biology, remote sensing, bioinformatics and hydrology (80–83). It is perhaps unsurprising therefore, that SVM has become a popular choice for classification of (I123)FP-CIT images (see section 1.3.2).

SVM is a type of linear classifier. This group of functions can be described as follows:

$$f(x) = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} + b \qquad \textit{Eq 2.10}$$

Where ***w*** represents a vector of weightings, ***x*** represents the (multidimensional) inputs, *b* is a bias term and *f(x)* is the algorithm output or the 'decision' of the classifier. The goal of the classifier is to learn the model parameters (i.e. ***w*** and *b*) that are most appropriate for separating groups in the space defined by the inputs. SVM achieves this by focusing on the samples that are closest to the opposite class, on the edge of each group. These are the 'support vectors', as represented graphically in Figure 2-4.

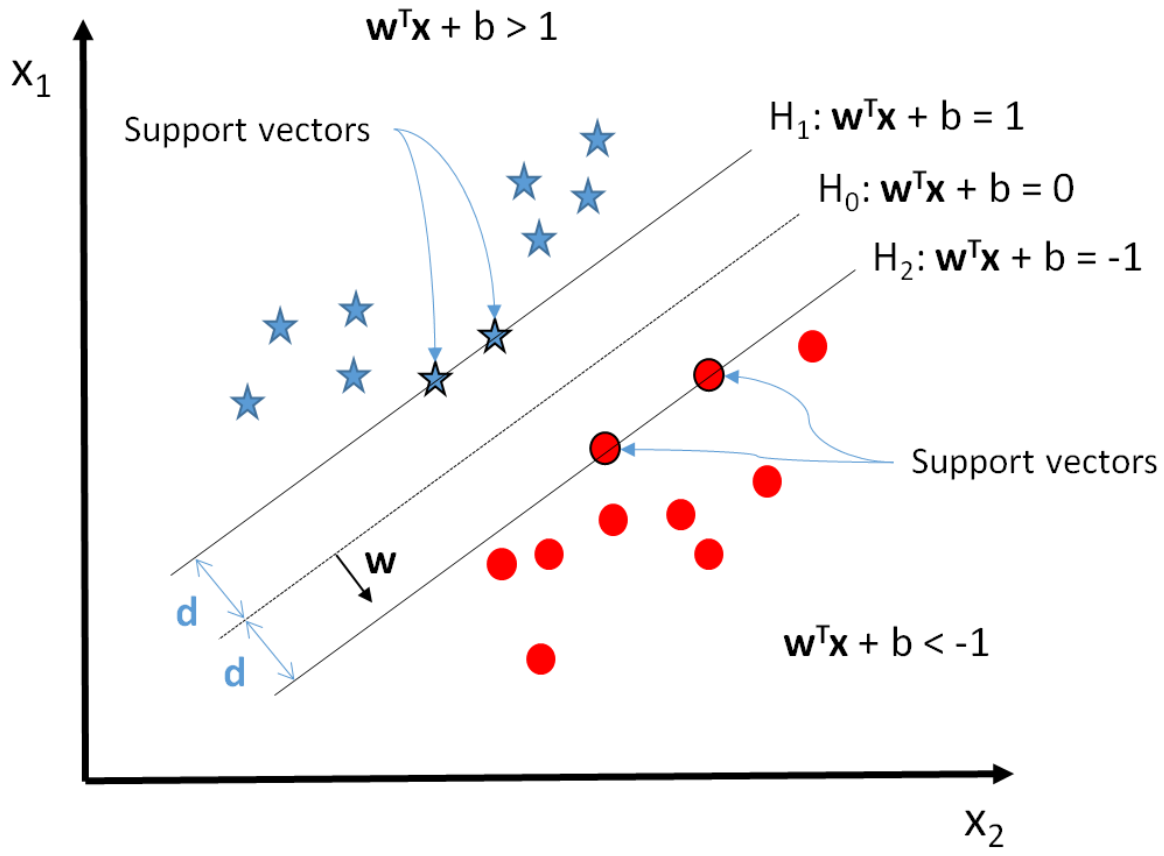


Figure 2-4 Graphical representation of classical SVM theory, where the goal is to define a maximal separating margin between class one (blue stars) and class 2 (red circles)

In SVM $\boldsymbol{w^T x} + b = 0$ defines the separating plane between classes (H_0). Since $c(\boldsymbol{w^T x} + b) = 0$ defines the same plane there is a free normalisation parameter, c, which can be selected. SVM normalises the linear equation such that there are 2 separate equations for the support vectors of each class:

$$\boldsymbol{w^T x} + b = 1 \quad where\ y = 1\ (H_1) \qquad \textit{Eq 2.11}$$

$$\boldsymbol{w^T x} + b = -1 \quad where\ y = -1\ (H_2) \qquad \textit{Eq 2.12}$$

Here, y indicates the binary class label. For linearly separable data any samples which are not support vectors will give $f(x)$ values of less than -1 or greater than 1, depending on class membership (see Figure 2-4). Noting that for any two arbitrary points along H_0, $\boldsymbol{a_1}$ and $\boldsymbol{a_2}$:

$$\boldsymbol{w^T}(\boldsymbol{a_1} - \boldsymbol{a_2}) = 0 \qquad \textit{Eq 2.13}$$

It is clear that the vector $\boldsymbol{w}$ is always normal to the surface of H_0 (and to H_1 *and* H_2, which share the same gradient). The separation between the support vector planes H_1 and H_2 is the margin between the classes, which SVM attempts to maximise. The width of this margin can be calculated by recalling that the perpendicular distance (d) between a point (p_0,q_0) and a line ($kx_1 + lx_2 + m = 0$) is:

$$d = \frac{|kp_0 + lq_0 + m|}{\sqrt{k^2 + l^2}} \qquad \textit{Eq 2.14}$$

Therefore, taking an arbitrary point on H_0, the perpendicular distance to H_1, can be calculated from:

$$d = \frac{|\boldsymbol{w}\boldsymbol{x} + b|}{\sqrt{w_1^2 + w_2^2}} = \frac{1}{\|\boldsymbol{w}\|} \qquad \textit{Eq 2.15}$$

Where w_1 and w_2 are the individual components of the vector $\boldsymbol{w}$ and $\|\boldsymbol{w}\|$ is the magnitude or norm of $\boldsymbol{w}$. The total separation between H_1 *and* H_2 is double this length. The goal of SVM is therefore to maximise $\frac{2}{\|\boldsymbol{w}\|}$. Inverting this expression, the objective function of SVM becomes the minimisation of $\|\boldsymbol{w}\|$.

This optimisation problem is subject to the following constraints, derived from Eq 2.11 and Eq 2.12:

$$y_i\left(\boldsymbol{w^T x_i} + b\right) \geq 1 \ \text{ for } \ i = 1, \ldots, N \qquad \textit{Eq 2.16}$$

Unfortunately, real world classification problems often involve data that are not linearly separable as shown in the simple example in Figure 2-4. However, SVM classifiers can still be useful in these circumstances if the objective function is modified to include additional terms. By introducing a slack variable (ε) to the constrained optimisation problem (see Eq 2.17) samples are permitted to lie beyond the support vector margin of their particular class (see Figure 2-5).

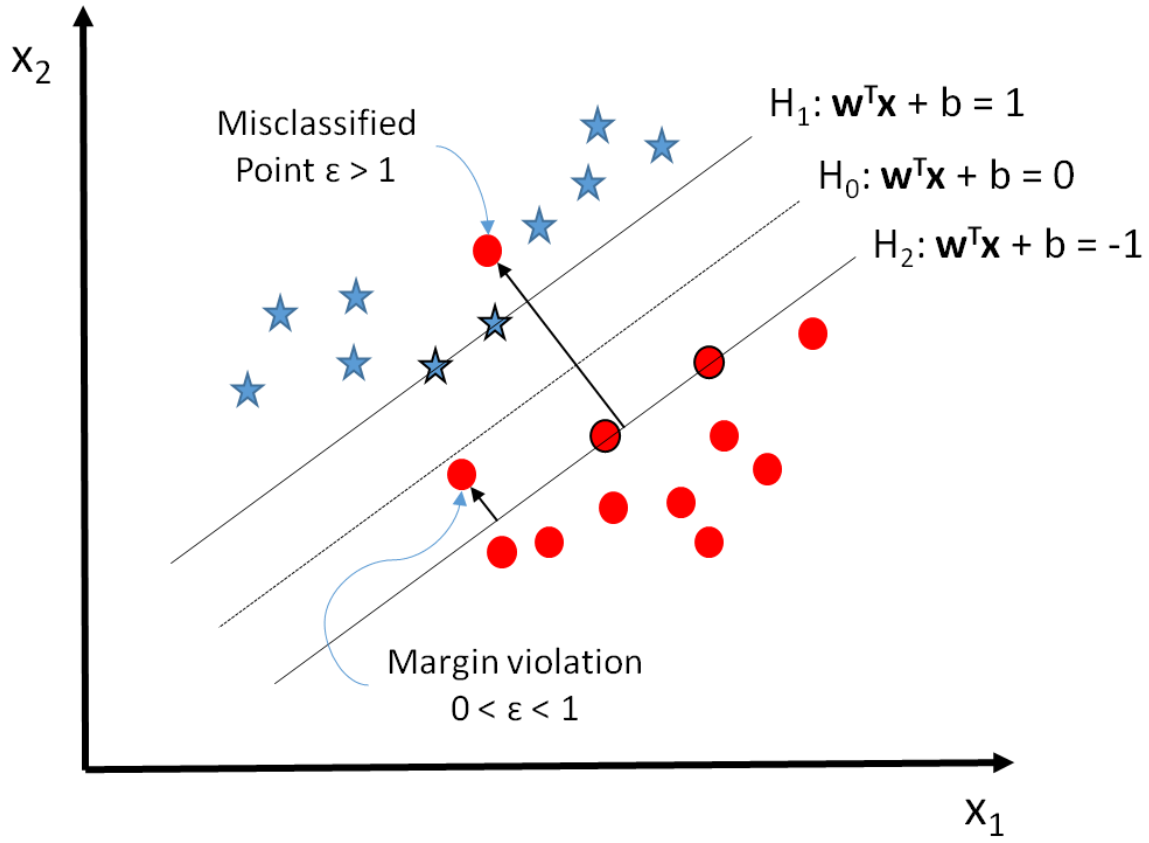


Figure 2-5 Graphical representations of soft-margin SVM, where complete linear separation between classes is not possible

$$\min_{w,b,\varepsilon} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i}^{N}\varepsilon_i \qquad \textit{Eq 2.17}$$

Subject to: $y_i\left(\boldsymbol{w}^T\boldsymbol{x_i} + b\right) \geq 1 - \varepsilon_i, \quad \varepsilon_i \geq 0 \ \text{ for } i = 1, \ldots, N$

The regularisation parameter C determines the relative contribution of the slack variable to the optimisation and balances the need for a large margin with that of maintaining the

constraints. In effect, the penalty for each margin violation is equal to $C\varepsilon$. A small value for C allows the constraints to be easily ignored such that many data points can violate the margin between classes, with outliers exerting less influence on the decision boundary. A large value for C ensures that the optimisation considers the constraints as more of a hard limit such that the optimal solution is likely to be a smaller margin hyperplane, which minimises the number of samples violating the margin.

SVM with slack variables is the 'soft margin' version of SVM (84), and is the approach used in this thesis as it provides a more robust approach to classification.

One further alteration can be applied to this problem formulation to widen its potential scope of application. Input data may not be linearly separable in the input space and classification performance may be poor, even with the addition of slack variables. However, if data can be mapped to a higher dimensional space then linear separation may become feasible. This idea is highlighted in the simple example shown in Figure 2-6. In the left image two variables are displayed, taken from two different populations (described by either blue circles or red crosses). In this case linear discrimination will be ineffective. On the right is the same data but with the addition of a z-axis which is the sum of squares of the original variables. This graph shows a clear separation between the groups, with linear separation possible in the z-axis.

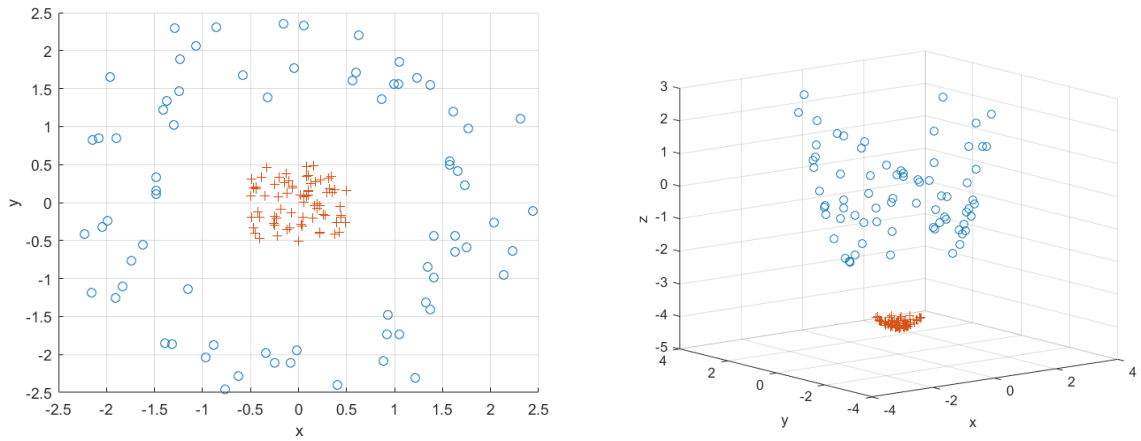


Figure 2-6 Illustration of the how mapping to a higher dimensional space can enable linear separation

Mapping all points to a higher dimensional space before performing the classifier optimisation can be very computationally expensive. Therefore, researchers often employ

the ‘kernel trick’ such that this mapping does not need to be performed directly. This involves taking the Lagrangian of the optimisation problem (Eq 2.17), examining the dual formulation and recognising that a kernel (similarity function) can be used in place of an inner product of data points that have been transformed to a higher dimensional space. The mathematics of these operations are not reproduced here as such steps are relatively unimportant in terms of demonstrating the broad concepts behind SVMs.

There are a number of different kernels that can be used when training SVMs. Their main benefit is that the derived separating plane between classes, when projected onto the original data axes, can be non-linear. This greatly extends the scope of application of SVMs and is likely to be one of the main reasons why they have been so popular in the literature.

One of the most commonly used kernels is the Gaussian kernel or Radial Basis Function (RBF) kernel (see Eq 2.18), the use of which is equivalent to implicitly applying the SVM in an infinitely large dimensional space.

$$K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) \qquad \textit{Eq 2.18}$$

Where σ is a hyperparameter which controls the width of the Gaussian function. Smaller values of σ tend to make the SVM decision boundary (in the input feature space) more flexible with greater curvature, which means that the algorithm can be trained on highly non-linear data but at the increasing risk of overfitting (i.e. high variance, lower bias). Larger values of σ lead to a smoother decision boundary with reduced curvature, which is less prone to overfitting (i.e. low variance, higher bias).

The main disadvantage of introducing kernels into SVM algorithms is that there are additional hyperparameters that need to be selected, which if not chosen carefully could produce a classification function that is too highly tuned to training data (with subsequent lower performance on independent test data). It is not clear from the literature whether linear or non-linear forms of SVM are likely to be the most successful for (I123)FP-CIT classification, and so both approaches are implemented.

2.1.3 Algorithm pipelines

For this work the well-established libSVM package (version 3.18, https://www.csie.ntu.edu.tw/~cjlin/libsvm/) (85) was adopted for training and implementing the SVM algorithms, utilised from within Matlab scripts. Prior to training each SVM algorithm all variables were normalised by subtracting the mean value and dividing by the standard deviation (available from training data). This was performed such that all variables are treated with equal weight during training.

When using principal components as input features to SVM, images were first examined to establish which side of the brain had the lowest uptake within the striatum. This was achieved through examination of SBR figures (see 2.2.5 for details on how these were extracted from the data). The images were mirrored about the central axis, if necessary, in order to ensure that the striatum of lowest tracer uptake was always on the left side. This approach, as implemented by Towey (65), was performed in order that the effects of unilateral disease are not ameliorated in the projection on to principal component axes. The image reorientation process was also conducted when raw voxel intensities were used as features, in order that abnormal data had more similar appearances and would be clustered closer together in the classification space.

When using image voxels or principal components as algorithm inputs, only the central portion of the image was of interest, containing the striata. The majority of the remaining brain, the skull and image background were not diagnostically significant. Thus, a loose region of interest was applied to all images. Voxel intensities outside of this region were masked out. Three different sized masks were investigated for each set of features and for each SVM model as it was unclear a-priori what size / volume would give the best performance. Different mask sizes were achieved by dilating the original mask different numbers of times. Dilating the mask once was equivalent to expanding the boundary of the mask by one voxel in all directions. Three different numbers of mask dilations were considered: 0, 2 and 4. As previously stated, patient age was also used as a separate input to the SVM classifiers, to force the algorithm to model the effects of this confounding variable on the classification result.

An overall summary of the different machine learning algorithms that were implemented (including different input features) is shown in Figure 2-7.

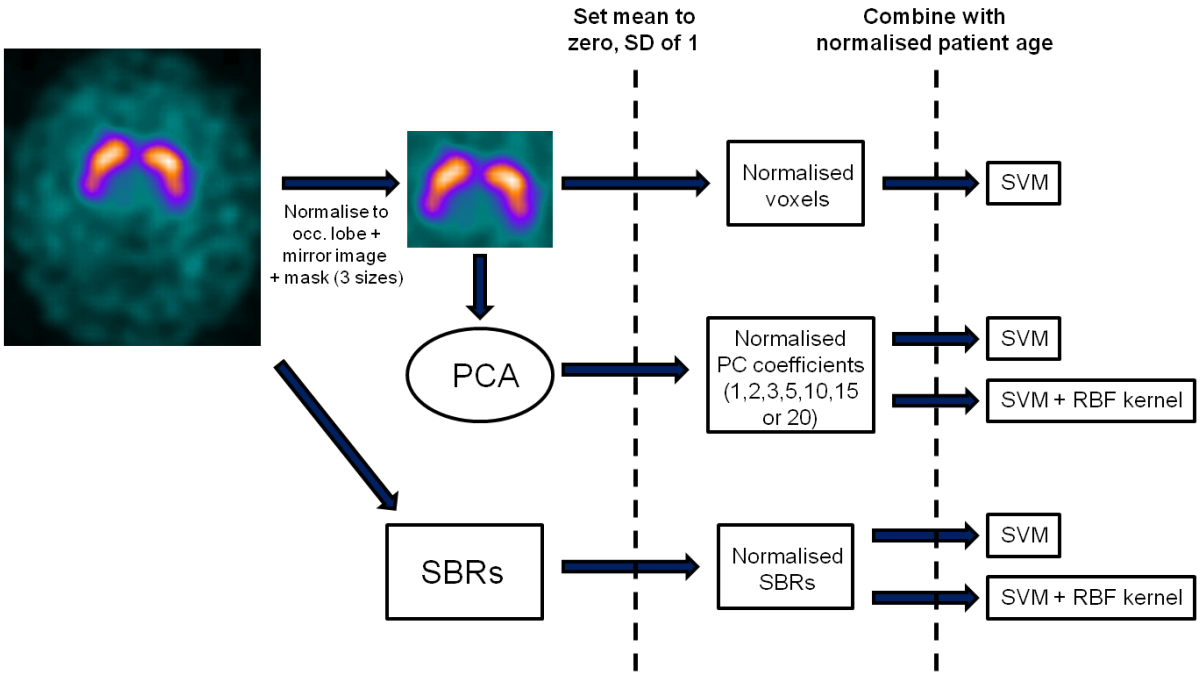


Figure 2-7 Summary of the different machine learning algorithms that were implemented (adapted from (1))

In total, considering all the different input features and both linear and non-linear SVM classifiers, there were 47 distinct machine learning approaches that were implemented in software. These are summarised in Table 2-2.

Machine learning algorithm	**Input feature**	**No. of PCs**	**Dilate (times)**	**SVM Kernel**
ML 1, 2, 3	PCs	1	0, 2, 4	Linear
ML 4, 5, 6	PCs	2	0, 2, 4	Linear
ML 7, 8, 9	PCs	3	0, 2, 4	Linear
ML 10, 11, 12	PCs	5	0, 2, 4	Linear
ML 13, 14, 15	PCs	10	0, 2, 4	Linear
ML 16, 17, 18	PCs	15	0, 2, 4	Linear
ML 19, 20, 21	PCs	20	0, 2, 4	Linear
ML 22, 23, 24	PCs	1	0, 2, 4	RBF
ML 25, 26, 27	PCs	2	0, 2, 4	RBF
ML 28, 29, 30	PCs	3	0, 2, 4	RBF

ML 31, 32, 33	PCs	5	0, 2, 4	RBF
ML 34, 35, 36	PCs	10	0, 2, 4	RBF
ML 37, 38, 39	PCs	15	0, 2, 4	RBF
ML 40, 41, 42	PCs	20	0, 2, 4	RBF
ML 43, 44, 45	Pixels	-	0, 2, 4	Linear
ML 46	SBR	-	-	Linear
ML 47	SBR	-	-	RBF

Table 2-2 List of the distinct machine learning algorithms developed and implemented in Matlab software. For each algorithm patient age was used as an additional input feature to the classifier (adapted from (1))

2.2 Patient data

Objectives addressed by this section (in black, bold):
1) Select and implement machine learning classification tools
2) Collect a database of (I123)FP-CIT images
3) Compare the performance of machine learning algorithms with semi-quantification
4) Develop software for testing of human reporters
5) Assess the impact of an automated classification tool, implemented as a CADx system, on reporting

Table 2-3 Objectives addressed in section 2.2

Two types of patient data were collected and used throughout this work, clinically-acquired retrospective data and prospective research data acquired from a clinical trial (the PPMI database). The former is key for ensuring that findings from all the investigations in this study are relevant to the clinic. The latter is likely to be associated with fewer confounding variables, as all data were acquired according to a tight research protocol, which will add certainty to any trends identified in clinical data. Furthermore, inclusion of the PPMI data enables results to be compared with those of other authors, giving context to the methods employed here.

The following sub-sections provide more details on the characteristics of these two datasets, and the additional image processing that was carried out in preparation for performance

tests of machine learning algorithms and semi-quantification methods. Many of the details of these datasets are also discussed in a peer-reviewed publication (1).

2.2.1 Clinical images and reference classification ("local data")

Approval was sought and granted by City and East Research Ethics Committee (15/LO/0736) to extract all (I123)FP-CIT images from the archives at Sheffield Teaching Hospital NHS Foundation Trust (acquired up until June 2015). All these images were acquired under standard conditions (see Table 2-4), other than the stopping conditions which were set a constant 30s per projection, rather than acquiring according to counts. Four scanners were used to acquire the data, namely three GE Infinia cameras and one GE Millenium scanner (all manufactured by GE Healthcare). No specific inter-scanner calibration was conducted so it is likely that there may be small systematic differences in semi-quantification results between systems. Reconstruction was conducted on the same system in all cases with the same settings: GE Xeleris v2.1 with 2 iterations, 10 subsets and Butterworth post-filter (order 10 cut-off 0.7).

Parameter	Value
Administered activity	167-185 MBq
Injection-to-scan delay	3-6 hours
Acquisition time	30 minutes
Acquisition pixel size	3.68 mm
Number of projections	60 per head (over 180^{o})
Energy window	159 keV ± 10%
Acquisition matrix size	128 x 128

Table 2-4 Summary of clinical data acquisition parameters

Cases where image quality was very poor, as highlighted by the image report, were excluded from the database. In addition, cases where significant previous vascular disease had been highlighted in the image report were excluded. This ensured that the pattern of dopaminergic function was unaffected by infarcts at or near the striatum.

Once extracted, each image required a 'gold-standard' diagnostic classification. It was intended that this information would be derived from examination of clinical notes over an extended period of time following SPECT imaging by expert neurologists. This would bring

the study in to line with other literature examining the accuracy of (I123)FP-CIT imaging. However, due to resource constraints, only a subset of the available images could be classified in this way (subset A). As an additional, alternative method (and consistent with subset B), the image reports for each patient were examined. The overall conclusion of the reporting radiologists was used to provide a 3 level classification, using a scoring system from 1-5 to reflect the level of confidence that the radiologists had in their findings. The scoring system was as follows:

- 1 = definitely abnormal appearances
- 2 = more likely abnormal than normal
- 3 = equivocal
- 4 = more likely normal than abnormal
- 5 = definitely normal appearances

At Sheffield (I123)FP-CIT images are all reported by two radiologists together, with additional contribution from a trained Clinical Scientist. In recent years semi-quantification results have also been provided to the reporting team, although very few patients in the cohort were reported with this additional information. Reporting radiologists have access to the patient's previous imaging results and relevant clinical information (such as presenting symptoms). This comprehensive approach to reporting should ensure that results are reflective of the best performance achievable from visual image analysis. Patients for whom only an image report was available for generating the reference diagnosis are considered to be part of subset B.

Results

In total 389 images were extracted from the archives. There were 55 cases where clinical follow-up by 2 neurologists had established a diagnosis with high confidence (subset A). The mean time of follow-up post SPECT imaging was 31 months, with a minimum of 15 months and maximum 51 months. There were 34 male and 21 female patients in this subset. At the time of scanning their mean age was 66 years (SD = 11 years) with a maximum of 80 and minimum 29 years. The patient characteristics of subset A are highlighted in Table 2-5.

Subset A		
Diagnosis	**Classification group**	**Number of patients**
Parkinson's Disease	Patients with pre-synaptic dopaminergic deficit (abnormal appearances)	29
Dementia with Lewy Bodies		4
Drug induced Parkinsonism	Patients without pre-synaptic dopaminergic deficit (normal appearances)	5
Hydrocephalus		1
Multiple Sclerosis		1
Essential tremor		10
Dystonia		3
Alzheimer's Disease		2

Table 2-5 Diagnostic categories and patient numbers for the 55 patients where diagnosis could be confirmed through long term follow-up, with high confidence (subset A)

The other 306 images were classified into 'normal appearances', 'abnormal appearances' and 'equivocal' groups using the image report only (subset B). Of these, the majority were reported with high confidence. As shown by Table 2-6, only one patient had an equivocal report, dictating that the classification was essentially binary. For this larger subset the mean age was 69 years (SD = 13 years), with a maximum of 92 years and minimum of 18 years. There were 194 males and 112 females.

Subset B		
Score	**Classification group**	**Number of patients**
1	Patients with pre-synaptic dopaminergic deficit (abnormal appearances)	174
2		17
3	equivocal	1
4	Patients without pre-synaptic dopaminergic deficit (normal appearances)	29
5		84

Table 2-6 Patient numbers and classification grouping for patients with no clinical diagnosis (subset B)

To assess the likely discrepancy between the two methods of classification, the clinical diagnosis of the 55 patients in subset A was compared to their image reporting scores in terms of accuracy. For the 31 patients with a reporting score of 1, there was 1 discrepancy with clinical follow-up results. For the 2 patients with a score of 2 there was also 1 discrepancy. For patients with a score of 4 or 5 there was complete agreement with the clinical follow-up results. Thus, if patients with scores of 1 and 2 are lumped together into an abnormal group, and those with scores of 4 and 5 lumped together into a normal group, the overall error in binary classification is only 3.6% between conventional image reporting and clinical follow-up. The specificity is 100% and sensitivity 94%. This suggests that the current clinical reporting system is relatively cautious, keeping the false negative rate low at the expense of slightly reduced sensitivity. However, overall, results provide a level of reassurance as to the reliability of radiologists' reports. Moreover, this level of error is smaller than that seen in the previous literature review on the accuracy of visual analysis (where error was generally 10-20%, see section 1.1.5). This may be due to the influence of other imaging data and clinical feature information that was available to radiologists, but which is generally excluded from previous studies on (I123)FP-CIT accuracy. It could also be because this particular cohort of patients was relatively easy to classify. Whatever the underlying causes, these results provide evidence that for the larger patient group (subset B), the impact of non-gold standard classification is likely to be relatively small.

Discussion

Given the two-tiered nature of the clinical data acquired from Sheffield, careful consideration was given to how each subset should be used. Subset A, with the more dependable diagnostic information, was kept aside for the most critical investigations in this study, where reliability of results was of highest importance. For this thesis, these investigations were considered to be the examinations of impact on radiologist performance, as evidence of impact in a real reporting scenario is likely to be key to judging the overall success of machine learning algorithms.

It was decided that subset B (without the singular equivocal case) would be used for cross-validation investigations, where the goal was to compare the standalone performance of semi-quantification and machine learning algorithms. It would also be used for algorithm training in later clinical studies. The justification for this was that algorithms trained to achieve the level of performance of an expert reporting team are still likely to be clinically useful. Furthermore, comparisons of cross validation metrics are unlikely to be significantly biased by a slightly increased level of uncertainty in the reference classification. The relative performance of each algorithm is key here, not the overall level of performance.

A possible solution for reducing discrepancies between subsets A and B would be to exclude all data from subset B with a score of 2, 3 or 4, only keeping data that was either definitely normal or definitely abnormal. However, this would bias the trained algorithms and cross-validation results towards a situation that did not reflect clinical reality, where images are sometimes difficult to classify. This approach was therefore rejected.

SBR values were calculated for each patient in each subset following further image pre-processing, which is summarised in later sections.

2.2.2 Research data – PPMI database

The PPMI dataset is a large online repository of diagnostic data from patients with PD and healthy controls, funded by the Michael J Fox foundation for Parkinson's Research. Different forms of Parkinsonism are explicitly excluded from the study, in contrast to the Sheffield data. As discussed previously, data were acquired prospectively from recruited patients. The full study protocol can be downloaded from the website (http://www.ppmi-info.org/). In summary, a battery of tests was applied to each recruit in order to assign them to a particular

diagnostic group. This methodology could dictate that the reference diagnosis is associated with reduced uncertainty as compared to the Sheffield data, particularly in comparison to the patients in subset B. For (I123)FP-CIT imaging specifically, the scanning protocol was largely similar to that used in Sheffield (see Table 2-7), other than imaging time which was set to a narrower window of 4 ± 0.5 hours post injection. In addition, Co^{57} markers were attached to each patient's head to enable correct orientation in subsequent processing. Specific scan parameters related to the collimators used and acquisition mode were set for each site and each scanner following initial assessment of phantom scans (75).

Parameter	PPMI database
Administered activity	111-185 MBq
Injection-to-scan delay	3.5-4.5 hours
Acquisition time	30-45 minutes
Acquisition pixel size	Variable (scanner dependent)
Number of projections	120 per head (over 360^{o})
Energy window	159 keV ± 10% and 122 keV ± 10%
Acquisition matrix size	128 x 128

Table 2-7 Summary of PPMI data acquisition parameters

Reconstruction was performed by a core lab, using Hermes HOSEM software (Hermes Medical), with 8 iterations, 8 subsets and a 6 mm Gaussian post-filter. This is quite different to the parameters used clinically for reconstruction at Sheffield. Furthermore, the reconstructed data available to download from the PPMI website is already non-linearly registered to a template (although the exact methodology is not clear). Images are also supplied with attenuation correction (through Chang's method (86)) applied.

Following reconstruction, the PPMI core lab calculated SBR values for all the SPECT datasets. These were derived using PMOD software (PMOD Technologies LLC), by taking the 8 axial slices with greatest striatal uptake and summing together to create a compressed 2D slice, to which regions of interest were applied (76).

Results

All screening SPECT images from the PPMI repository were downloaded. This included images from 209 healthy controls and 448 PD patients i.e. 32% of the data was from patients with normal image appearances. These proportions were similar to that of subset A (40% normal appearances) and subset B (37% normal appearances). The mean age of the PPMI dataset was 61 years (SD of 10 years), which is similar but slightly lower than that seen in the Sheffield data. For comparison, the mean age of the healthy controls alone was also 61 years. Maximum and minimum ages for the combined PPMI dataset were 85 years and 31 years respectively. Thus, the age range covered is similarly wide to that of the Sheffield data. There were 232 females and 425 males in the PPMI database. In addition to the imaging data, SBR values calculated by the core lab were also downloaded from the PPMI website.

Discussion

There are a number of differences between the local and PPMI data that could impact on results. Firstly, the higher number of expectation maximisation equivalent iterations used for the PPMI data is likely to produce images with higher noise but greater contrast, which may improve contrast between striatum and background, particularly for borderline classification cases (47). This could possibly make the binary classification task simpler for machine learning / semi-quantification software.

The use of non-linear registration for PPMI data may have caused some warping of striatal shape, which may impact upon semi-quantification and machine learning algorithms. The attenuation correction applied can help to reduce inter-patient differences in striatal appearances from variations in head geometry between subjects. This was not applied to the Sheffield data and so, again, this may cause the PPMI data to be incrementally easier to classify through a machine learning tool.

There are other key differences between the PPMI dataset and the Sheffield dataset that should be kept in mind in the following investigations. Firstly, each site involved in the PPMI study was required to scan a phantom prior to each patient. This provided calibration data, which were applied to reconstructed patient images in order to remove systematic inter-site and inter-scanner differences. This procedure was not performed for the Sheffield data (and is not mandatory according to clinical guidelines), which suggests that the PPMI data may be

associated with reduced systematic error between acquisition equipment. Furthermore, as highlighted previously, patients were only included in the PD group if visual analysis of their SPECT data showed reduced nigrostriatal dopaminergic function. By explicitly excluding patients with PD symptoms but normal SPECT appearances, the dataset is likely to be biased towards more favourable classification accuracies when image analysis techniques are applied.

Overall, the differences between local data and the PPMI database favour increased classification performance for the PPMI data. The different databases are therefore kept separate in the following investigations.

The local image data and PPMI data required pre-processing, to different extents, before SVM algorithm training. SBR figures also needed to be extracted from local data. The developed methods for carrying out these steps are described briefly in the following three sub-sections. In each of these sections theoretical details are only discussed briefly. The reader is reminded that the focus of this work is on the implementation and evaluation of existing technology, not on development of fundamentally new classification concepts.

2.2.3 Image pre-processing: spatial normalisation (local data only)

Registration or spatial normalisation is crucial for maximising performance for the chosen machine learning approach. Each voxel in each image is effectively considered to be at the same geometric location in the patient's body, for all of the features considered. Significant variability in patient positioning can therefore cause a significant shift in where each particular input variable lies in the feature space, leading the SVM algorithm to define a separating hyperplane between classes in an inappropriate position.

The PPMI images were downloaded having already been registered to a template by the trial core lab team and so required no further spatial normalisation. In contrast, the images extracted from the local hospital archives were orientated in the original patient positions on the scanner bed.

Application of a fully non-rigid registration to local data would be likely to provide a mapping that gave the closest fit between each image and a template, thereby minimising geometric differences. Indeed, the PPMI data were all processed using a non-rigid translation step by the co-ordinating core lab. However, it was decided that registration of local data should be

restricted to an affine transformation (which only permits rotation, scaling, translation and shearing in all dimensions). Such rigid registration ensures that the shape of the striata cannot be significantly warped after registration. A non-rigid registration that was poorly constrained could cause an apparently abnormal image, with ‘full stop’ striatal appearances, to be stretched into the shape of a comma if the template image was a normal dataset (which is often the case).

Image registration is itself a wide and established scientific field with a variety of approaches described in the literature (87), many of which share the same basic concepts as machine learning (i.e. most methods are based on minimisation of an objective function). For this study a reliable, proven technology was required to perform image registration. Finding a technique that is likely to work consistently across a range of patient studies is of utmost importance for a tool designed for clinical use and so this was the major consideration. Searching out the best performing algorithm for the relatively limited datasets considered in this work was a secondary consideration. The Sheffield Image Registration Toolkit (ShIRT (88)) is an established technology that has previously shown good results in the registration of nuclear medicine data (89,90). It is based on minimisation of the squared differences between intensity values in pixels in corresponding positions on two images, which is a commonly used cost function in many registration algorithms. Importantly, ShIRT has been used successfully in numerous clinical applications within Sheffield Teaching Hospitals for over 10 years. It was therefore an ideal candidate for the registration task in this study.

ShIRT, as with many other registration tools, requires parameters to be set by the user. In addition, it is common to apply registration in stages, to iteratively bring images into alignment. In order to evaluate the success of different parameter choices and different combinations of processing stages, suitable metrics are required. In the following investigation a combination of qualitative and quantitative assessments were used, visual scoring and Dice Similarity Coefficients (DSCs) (91,92). These are two of the most commonly applied strategies for optimising registration procedures. Although each is associated with well-known limitations (subjectivity in the case of visual scoring and sensitivity to segmentation inaccuracies in the case of DSC), they are simple to implement and in combination should enable the development of a well optimised registration approach.

Method

Spatial normalisation first required the creation of a suitable template image to provide a fixed target geometry. This was created from non-linear registration (again using ShIRT) of 10 patient cases without dopaminergic deficit to a single dataset, followed by averaging of all voxels in all images. The combined template was then manually re-aligned, to ensure that the long axis of the head was along the middle of the image and that there was no right to left rotation. Finally, the left half of the template brain was reflected about the centreline to produce a template image with identical striatal structures on each side. This procedure is summarised by the flowchart in Figure 2-8.

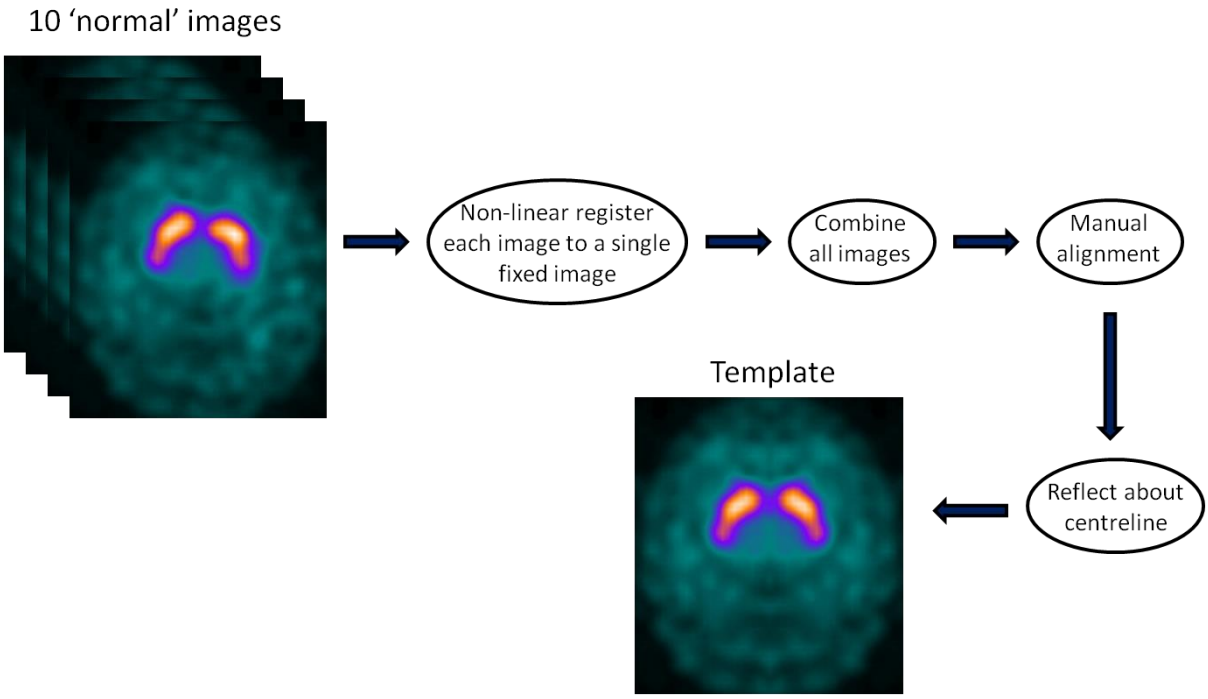


Figure 2-8 Flow diagram depicting the process for creating a registration template

The goal of registration was to automatically bring the striata in each local patient image into alignment with that of the template. Registration accuracy in the caudate and putamen was therefore of most interest. Relative alignment of other brain tissues was considered to be relatively unimportant. This prioritisation is reflected in the evaluation procedure. After each iterative registration step, where a different combination of registration parameters and processing procedures were tried out, each registered image was overlaid on the

corresponding template, one at a time. Visual analysis focused on the discrepancy in striatal boundary between each database image and the template. DSC was calculated for images from the non-PDD class through a simple segmentation process, whereby volumes of interest for the test image and template were defined via the application of a threshold at 20% of the maximum voxel intensity value in the image. The volumetric overlap of these segmentations was then calculated to give a DSC figure (see Figure 2-9 for a formal definition of DSC). DSC values for images from patients with PDD were not calculated as the striatal shapes in these images were expected to be very different to that of the template and thus DSC would be low whether an appropriate registration was achieved or not. For these images only qualitative analysis was used.

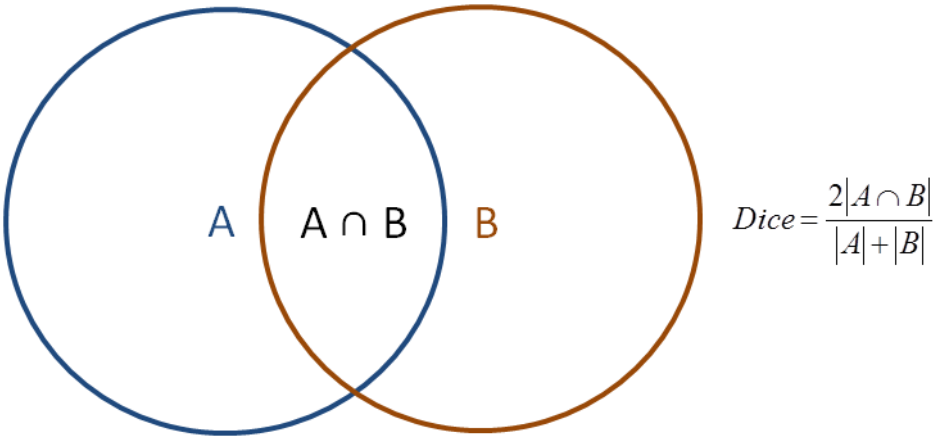


Figure 2-9 Dice similarity coefficient definition

Since the goal was to develop a tool that was robust to a range of image appearances, there was more of an emphasis on finding a technique that registered striata well in all images, rather than one that worked exceptionally well with just a few images. The optimal image registration procedure that was derived is summarised in Figure 2-10. In line with many other medical registration processes, the first stage involves a coarse registration of the test image to the template. This was followed by finer registration stages considering the left and right sides of the brain separately. In each registration step a loose registration mask was defined over the striatal region in the template image, such that registration focused on achieving spatial concordance in the striatal area. The DSC scores for the non-PDD images following this fully automated technique are shown in Table 2-8.

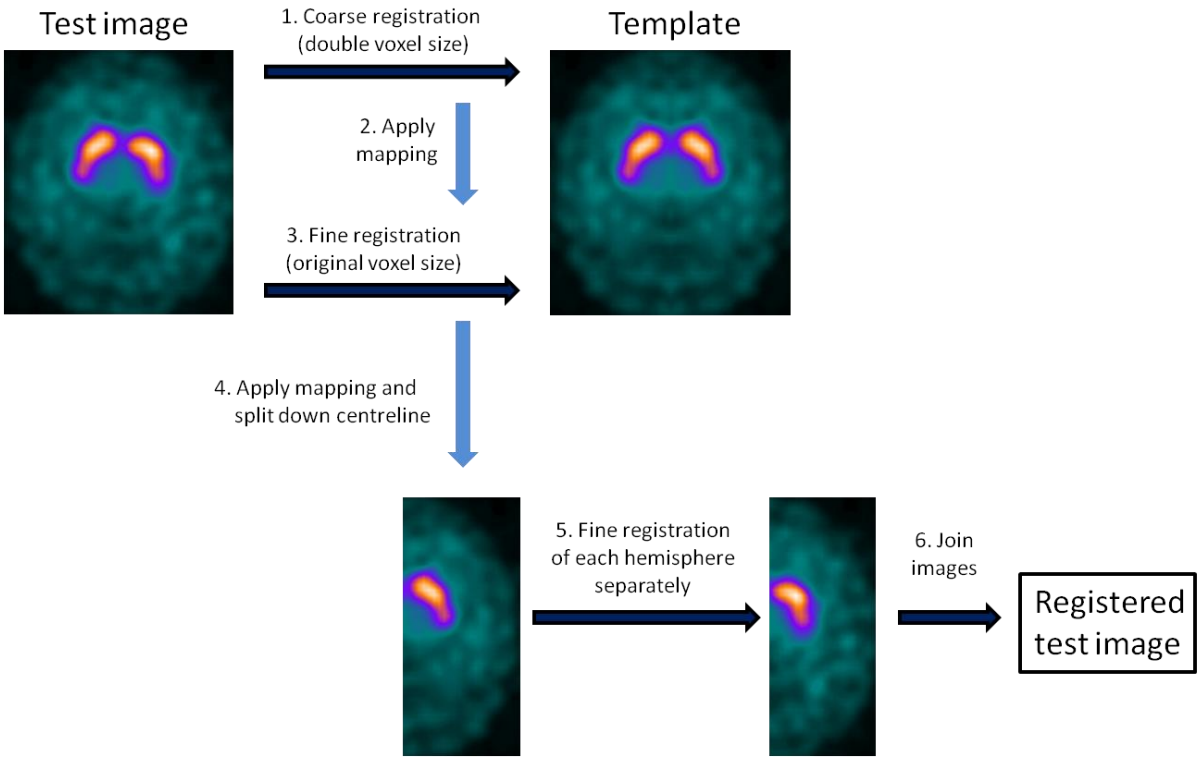


Figure 2-10 Flow diagram of the optimal registration procedure

Results

The similarity coefficients for datasets B and A were generally very similar and high (see Table 2-8), although the minimum reported value was lower for subset B than subset A

Dice Similarity Coefficient	Local data (subset A)	Local data (subset B)
Max	0.89	0.90
Median	0.84	0.84
Min	0.77	0.70

Table 2-8 DSC results for data with a non-PDD classification, following registration optimisation

Discussion

The DSC results indicate that the registration method was similarly successful for both subset A and subset B, with a similar spread of values in each case. This implies that the method is likely to be successful for a wide variety of patient images. Given that registration

was constrained to an affine transformation, concordance between the striata of test images and the templates was unlikely to be complete. Furthermore, the simple thresholding method used for quantitative evaluation is likely to have been an imperfect method for striatal segmentation. Thus, the fact that DSC results fell as low as 0.70 was perhaps unsurprising. On visual inspection, after applying the optimum registration technique, there were no datasets where registration errors between test images and the template were felt to be significant.

2.2.4 Image pre-processing: intensity normalisation (both PPMI and local data)

Voxel intensities within a SPECT image can vary depending on a variety of patient and technical factors, including time between tracer administration and imaging, biological clearance rate, image acquisition time and camera sensitivity. Thus, some form of voxel intensity normalisation is usually required to reduce this confounding inter-subject variability before training a classifier, particularly when the voxels themselves are used as features. In this study, an approach was adopted that has previously been cited by many other researchers, namely normalisation of all image counts to the occipital lobe. In most patients this area of the brain is a region of relatively uniform, non-specific uptake. Dividing all voxel intensities in the image by the mean intensity within the occipital lobe is, in effect, a very similar calculation procedure to that required for calculating SBRs, where the mean counts in striatal regions are divided by a mean non-specific uptake value. This procedure causes voxel intensities to become a measure of contrast, which is independent of many of the patient and technical factors previously cited.

The mean uptake in the occipital lobe was defined for each image in the PPMI and Sheffield datasets by applying a database-specific volume of interest. For the local data this volume was manually defined on the template image and transferred to each test image after the first coarse registration stage (see Figure 2-10). Normalisation was then applied. For the PPMI data the volume was manually defined on a single case and then propagated to all other (pre-registered) images in the dataset.

There are a number of additional pre-processing options that could be considered before semi-quantification or machine learning is applied. For instance, additional image smoothing could be applied to reduce noise. The reconstruction parameters could be altered in order to change the contrast between the striata and background. However, in this study the default reconstruction parameters and image filtering steps of each database were left unaltered. No

optimisation of these factors was carried out as each set of data had already been optimised to some extent. For instance, the processing applied to the Sheffield data was designed to enable the best visual classification performance in the clinic. Given that more distinct image appearances for each classification group are also likely to give a greater contrast in extracted image features, it was likely that the existing image processing steps were already suitable for achieving good classification performance.

If clinical performance of machine learning algorithms was found to be insufficient using the default reconstruction and filtering parameters then these would be re-evaluated as part of a whole processing pipeline re-assessment.

2.2.5 Extracting SBRs from local data

In order to compute the performance of the different semi-quantification approaches, and to provide SBR values for input to SVM algorithms, a method for extracting SBR figures from the local clinical images was required (PPMI data is downloaded with SBR figures already available). Having already registered the images to the same spatial location, a single series of regions needed to be defined on the registration template image, which could then be applied to all patients.

The boundaries of the tissues of interest were defined by adapting the template from an established commercial semi-quantification tool (MIM Neuro analysis v6.6, MIM software). The regions of interest defined for the MIM Neuro analysis template were warped to the space of the registration template used in this study, using non-linear registration (again implemented with ShIRT). Using these transformed regions, SBRs were calculated for every local clinical image by finding the ratio of uptake within caudate and putamenal regions as compared to the occipital lobe according to Eq 1.1 in chapter 1.

2.2.6 Conclusion

This chapter has demonstrated production of training and test datasets and methods for extracting SBR figures. These are used in the following chapters, first for comparing the performance of semi-quantification and machine learning tools, and then for evaluating the impact of CADx on reporter performance.

3 Comparison of semi-quantification and machine learning

Objectives addressed by this section (in black, bold):
1) Select and implement machine learning classification tools
2) Collect a database of (I123)FP-CIT images
3) Compare the performance of machine learning algorithms with semi-quantification
4) Develop software for testing of human reporters
5) Assess the impact of an automated classification tool, implemented as a CADx system, on reporting

Table 3-1 Objectives addressed in section 3

This chapter compares the standalone performance of established machine learning algorithms and a range of semi-quantification methods. A comprehensive, fair comparison between semi-quantification and machine learning, using the same data and validation methods for both approaches, has not yet been conducted in the literature. However, this is a vital step in understanding whether machine learning tools are effective classifiers for (I123)FP-CIT and whether they are likely to offer benefits above and beyond existing clinical tools. Furthermore, by conducting a wide comparison exercise, it will be possible to identify the most promising machine learning tool for use in the subsequent CADx reporting investigation.

A summary of this chapter was written for publication in a peer reviewed journal (1). Many of the methods, results and discussion are reported in this document.

3.1 Semi-quantification

In order that the potential diagnostic performance of semi-quantification is fairly represented, a range of semi-quantification methods needed to be defined. One approach would have been to use all the available commercial software for (I123)FP-CIT imaging. However, obtaining access to all examples of such (expensive) software was impractical. Furthermore, by only relying on the tools that are currently commercially available, the range of calculation

methods is likely to be limited. In this work a number of methods were defined within a single software analysis platform (Matlab), covering the majority of approaches cited by both commercial tools and those discussed in the literature. These are described below

3.1.1 Selected methods

One of the main differences between semi-quantification approaches described in the literature is the number and type of SBRs and uptake ratios that are output by the software. As described previously, multiple quantities and associated normal ranges may be displayed to the clinician. If just one of these quantities falls outside its associated normal range a clinician may classify the whole image as abnormal. Furthermore, the greater the number of quantitative figures displayed, the greater the chances that one of these will fall outside its normal range by chance (i.e. the greater the chances of type I statistical error). Therefore, in this investigation, only SBRs from the putamen and caudate were extracted, to avoid overly pessimistic performance results. These were used for classification in two different ways, considering the putamen only and the caudate and putamen together.

Semi-quantitative methods also differ in how they are compared to normal ranges and how they take account of the known correlation between age and SBR. One common approach is to establish the mean from healthy patients within a 10 year age window of the test sample, for each SBR result. The suggested cut-off may be established, for example, from a value that is a number of standard deviations from the healthy control mean (typically between 1 and 2 standard deviations). A similar alternative would be to establish a cut-off from the minimum of age matched controls. A third approach would be to perform a linear regression (of SBR value against age) on available training data from normal patients. The predicted SBR for the test sample can then be derived from the fitted line and a cut-off set according a number of standard errors on the regression coefficients. All three of these methods are implemented in the following investigation.

Another difference between semi-quantification approaches relates to the nature of the training data used to define the cut off in SBR values between the normal and abnormal class. So far it has been assumed that only data from healthy (or non-Parkinsonian) patients is used for learning the value of such cut-offs. However, this ‘one class’ approach is fundamentally limited since, without knowledge of where the abnormal group lies in the classification space, it is less likely that an optimum cut-off will be found. Therefore, in this investigation a two class semi-quantification approach is also tested, whereby analysis of the

ROC curve derived from SBR values of normal and abnormal training data is used to find the cut-off which achieves the highest accuracy in binary separation of the two classes. This is then applied to the test sample.

All of these different approaches to semi-quantification were implemented for direct comparison with the previously described machine learning algorithms. Table 3-2 summarises the key characteristics of all the different semi-quantification methods utilised in this work.

Semi-quantification method	**Comparison data**	**SBRs considered**	**SBR cut-offs defined by**
SQ 1	Age-matched normals	Left and right putamen	Mean – 2SD
SQ 2	Age-matched normals	Left and right putamen and caudate	Mean – 2SD
SQ 3	Age-matched normals	Left and right putamen only	Mean – 1.5SD
SQ 4	Age-matched normals	Left and right putamen and caudate	Mean – 1.5SD
SQ 5	Age-matched normals	Left and right putamen	Mean – 1SD
SQ 6	Age-matched normals	Left and right putamen and caudate	Mean – 1SD
SQ 7	Age-matched normals	Left and right putamen	Minimum
SQ 8	Age-matched normals	Left and right putamen and caudate	Minimum
SQ 9	All normals	Left and right putamen	Linear regression - 2SE
SQ 10	All normals	Left and right putamen and	Linear regression - 2SE

		caudate	
SQ 11	All normals	Left and right putamen	Linear regression - 1.5SE
SQ 12	All normals	Left and right putamen and caudate	Linear regression - 1.5SE
SQ 13	All normals	Left and right putamen	Linear regression - 1SE
SQ 14	All normals	Left and right putamen and caudate	Linear regression - 1SE
SQ 15	All normals and abnormals	Lowest putamen	Optimal point on ROC curve
SQ 16	All normals and abnormals	Lowest putamen and lowest caudate	Optimal point on ROC curve
SQ 17	Age matched normals and abnormals	Lowest putamen	Optimal point on ROC curve
SQ 18	Age matched normals and abnormals	Lowest putamen and lowest caudate	Optimal point on ROC curve

Table 3-2 Summary of the semi-quantification methods implemented for classification performance comparison (adapted from (1))

3.2 Cross-validation

Fair comparison between the standalone performance of machine learning algorithms and semi-quantification methods is vital for assessing whether machine learning offers any potential benefit over existing clinical decision support technology. However, unfortunately, there are many examples in the literature of poor evaluation methodology being applied to image classification or regression algorithms, dictating that results are unreliable (93,94). This work is different in that clinical translation is seen as the ideal end goal, and evaluation techniques are deliberately adopted to ensure, as far as possible, that results reflect likely clinical performance.

In the following investigations it is assumed that the optimal form of an SVM algorithm (including optimal choice of hyperparameters such as the C value) is unknown a-priori, and needs to be derived as part of the training process. In these circumstances data used for training, choosing hyper-parameters and estimating performance should be fully independent to avoid over-optimistic bias.

For the following investigation a repeated 10-fold cross-validation methodology was selected, with nesting and class stratification where appropriate. This approach has the advantage of providing estimates of uncertainty in performance results and should help to ensure a reasonable balance between bias and variance in the model (71–73,95,96). The model selection phase for each machine learning algorithm (within the nested loops) was performed using a grid search methodology, whereby each possible combination of parameters was exhaustively searched to find those which gave the highest mean F1 score in cross-validation. The F1 score is a commonly used metric for selecting classification algorithms and is defined as:

$$F1 = \frac{2TP}{2TP + FN + FP} \qquad \textit{Eq 3.1}$$

Where TP refers to the number of true positives, FN the number of false negatives and FP the number of false positives. This form of parameter searching is more computationally expensive than alternative approaches, such as a random parameter search (where parameter combinations are randomly selected) (97). However, grid search is straightforward to implement and guarantees that all parameter values deemed to be viable options are selected and tested.

In addition, tests were also carried out to generate evidence of the relative bias-variance trade-off in the trained machine learning models through the creation of learning curves. Such evidence is usually not reported in the machine learning literature, but providing this information is considered as important for understanding whether an algorithm will generalise well to the clinic (92). Learning curves are produced by training the algorithms with increasing numbers of training data and classification performance is compared between that achieved with the training data and that measured from an independent test set. For an algorithm with optimal bias and variance characteristics it is expected that the performance figures for both training and test data would reach a similarly high value. For an algorithm with high bias (i.e. a model that under fits the data) performance figures for training

and test data are expected to be well matched but low. Finally, for an algorithm suffering from over-fitting (i.e. high variance) there will be a gap in performance figures, with training data performance being higher than that of the test data.

3.2.1 Method

All machine learning algorithms and semi-quantification methods were evaluated according to the same stratified, 10-fold cross-validation procedure, repeated 10 times. However, nesting and grid search were only necessary for the machine learning algorithms. The overall process for comparing semi-quantification and machine learning algorithms is summarised in Figure 3-1. For the semi-quantification tools it was necessary to use normal limits as a hard cut-off in that any SBR result that fell below the specified normal limit would be classified as a positive test.

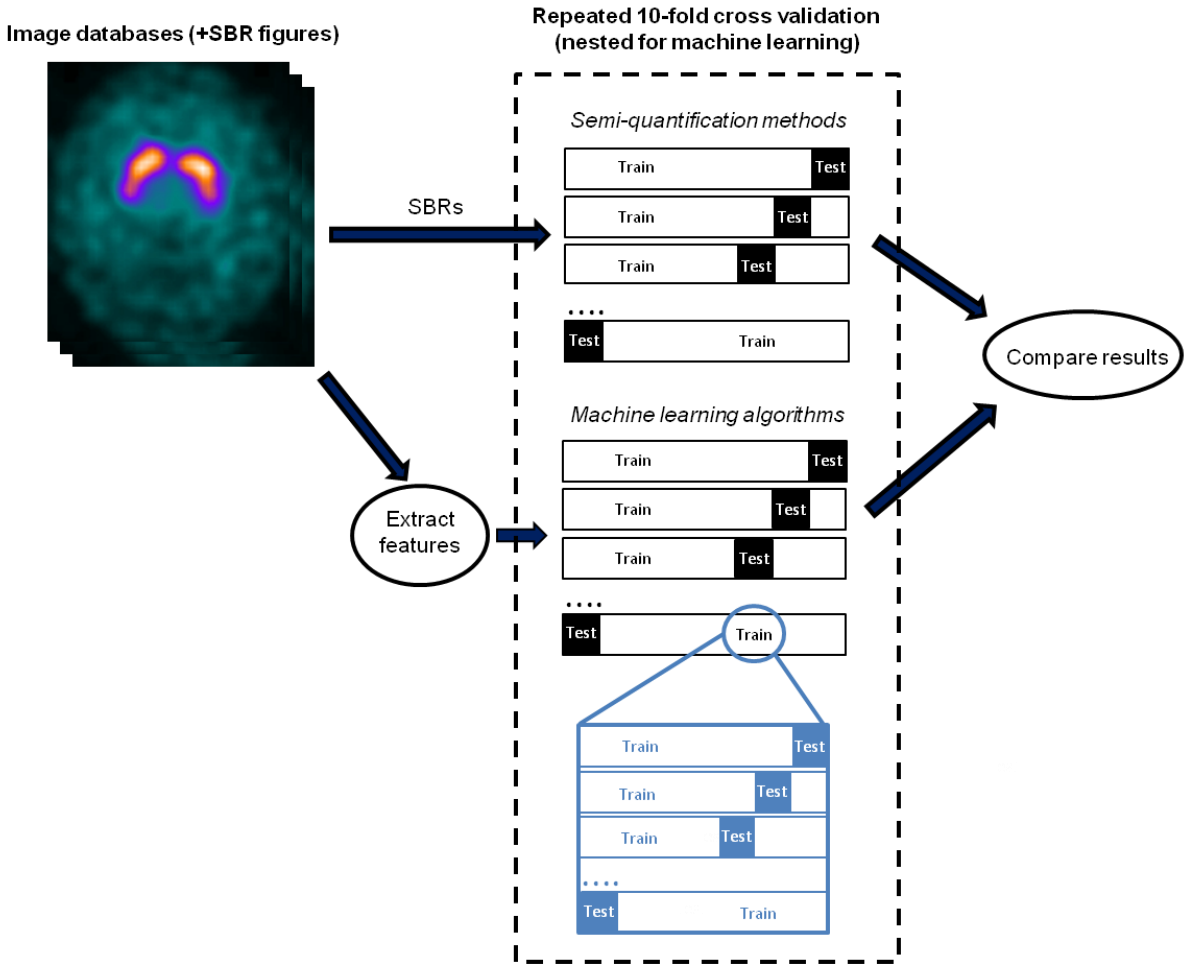


Figure 3-1 Summary of the semi-quantification / machine-learning comparison methodology (adapted from (1))

The parameter values available in the grid search are shown in Table 3-3. This range of values was set after preliminary work, using a wider array of possible C and gamma values, examining which figures are typically chosen for each of the models considered.

	Hyperparameters searched	
Algorithm	**C values**	**Gamma values**
Voxel intensities input to linear SVM	2^(-3,-2,-1,0,1,2,3,4,5,6,7,8)	
PCs input to linear SVM	2^(-3,-2,-1,0,1,2,3,4,5,6,7,8)	
PCs input to SVM with RBF kernel	2^(-3,-2,-1,0,1,2,3,4,5,6,7,8)	2^(-8,-7,-6,-5,-4,-3,-2,-1,0,1,2,3)
SBRs input to linear SVM	2^(-3,-2,-1,0,1,2,3,4,5,6,7,8)	
SBRs input to SVM with RBF kernel	2^(-3,-2,-1,0,1,2,3,4,5,6,7,8)	2^(-8,-7,-6,-5,-4,-3,-2,-1,0,1,2,3)

Table 3-3 Parameters selected during exhaustive grid search

Cross validation was completed for both the largest local database (subset B) and the full PPMI database, in both cases using Matlab scripts. Results were summarised using a number of different performance metrics including diagnostic accuracy, sensitivity and specificity, and their respective standard deviations.

Learning curves were generated for three different machine learning algorithms, each making use of one of the three different types of image feature: one utilising principal components as features (ML 10), one using raw voxel intensities (ML 43) and one utilising SBRs (ML 46). In each case the algorithms were trained with incrementally larger proportions of the local training dataset (subset B), with 50 cases set aside for testing only (containing 18 non-PDD and 32 PDD patients). After selection of hyperparameters in a cross-validation procedure, classification accuracy was measured once for the training data, and once for the 50 independent test cases.

3.2.2 Results

Table 3-4 and Table 3-5 show performance metrics for the machine learning algorithms using local data and PPMI data, respectively.
Table 3-6 and Table 3-7 show equivalent metrics for all the semi-quantitative methods. These results show that varying the size of the image mask had little effect on results. Accuracy results for all semi-quantification methods and machine learning algorithms (using the smallest mask size) are summarised graphically in Figure 3-2 and Figure 3-3, illustrating slightly improved performance with machine learning compared to semi-quantification for both local and PPMI datasets.

Algorithm	Feature	No. PCs	Dilate (times)	Kernel	Mean accuracy	SD	Mean Sensitivity	SD	Mean Specificity	SD
ML 1	PCs	1	0	Linear	0.86	0.06	0.90	0.06	0.79	0.13
ML 2	PCs	1	2	Linear	0.85	0.06	0.90	0.07	0.78	0.11
ML 3	PCs	1	4	Linear	0.85	0.06	0.90	0.07	0.78	0.14
ML 4	PCs	2	0	Linear	0.89	0.06	0.91	0.06	0.85	0.11
ML 5	PCs	2	2	Linear	0.89	0.05	0.92	0.06	0.84	0.10
ML 6	PCs	2	4	Linear	0.90	0.06	0.93	0.06	0.85	0.10
ML 7	PCs	3	0	Linear	0.91	0.05	0.93	0.05	0.88	0.10
ML 8	PCs	3	2	Linear	0.91	0.05	0.93	0.06	0.88	0.09
ML 9	PCs	3	4	Linear	0.91	0.05	0.93	0.06	0.87	0.09
ML 10	PCs	5	0	Linear	0.92	0.05	0.94	0.06	0.88	0.10
ML 11	PCs	5	2	Linear	0.91	0.05	0.93	0.05	0.87	0.11
ML 12	PCs	5	4	Linear	0.91	0.05	0.93	0.06	0.88	0.09
ML 13	PCs	10	0	Linear	0.91	0.05	0.93	0.06	0.86	0.10
ML 14	PCs	10	2	Linear	0.90	0.06	0.93	0.06	0.85	0.11
ML 15	PCs	10	4	Linear	0.91	0.05	0.94	0.05	0.87	0.10
ML 16	PCs	15	0	Linear	0.89	0.05	0.92	0.06	0.83	0.11
ML 17	PCs	15	2	Linear	0.89	0.06	0.92	0.06	0.83	0.11
ML 18	PCs	15	4	Linear	0.89	0.05	0.93	0.06	0.83	0.11
ML 19	PCs	20	0	Linear	0.89	0.05	0.92	0.07	0.83	0.12
ML 20	PCs	20	2	Linear	0.89	0.05	0.92	0.06	0.83	0.10
ML 21	PCs	20	4	Linear	0.89	0.05	0.93	0.05	0.84	0.10

ML 22	PCs	1	0	RBF	0.86	0.06	0.91	0.07	0.78	0.12
ML 23	PCs	1	2	RBF	0.85	0.07	0.90	0.07	0.76	0.13
ML 24	PCs	1	4	RBF	0.85	0.07	0.90	0.07	0.76	0.12
ML 25	PCs	2	0	RBF	0.91	0.05	0.91	0.06	0.90	0.10
ML 26	PCs	2	2	RBF	0.89	0.05	0.91	0.06	0.86	0.11
ML 27	PCs	2	4	RBF	0.90	0.05	0.92	0.06	0.88	0.09
ML 28	PCs	3	0	RBF	0.91	0.05	0.91	0.07	0.89	0.09
ML 29	PCs	3	2	RBF	0.91	0.05	0.92	0.06	0.90	0.08
ML 30	PCs	3	4	RBF	0.91	0.05	0.92	0.06	0.89	0.09
ML 31	PCs	5	0	RBF	0.91	0.06	0.92	0.06	0.89	0.10
ML 32	PCs	5	2	RBF	0.91	0.05	0.92	0.06	0.89	0.09
ML 33	PCs	5	4	RBF	0.91	0.04	0.92	0.05	0.89	0.10
ML 34	PCs	10	0	RBF	0.90	0.05	0.91	0.07	0.88	0.09
ML 35	PCs	10	2	RBF	0.91	0.05	0.92	0.06	0.89	0.10
ML 36	PCs	10	4	RBF	0.91	0.05	0.92	0.06	0.89	0.09
ML 37	PCs	15	0	RBF	0.89	0.05	0.91	0.07	0.87	0.10
ML 38	PCs	15	2	RBF	0.90	0.05	0.91	0.06	0.87	0.10
ML 39	PCs	15	4	RBF	0.90	0.05	0.92	0.07	0.88	0.10
ML 40	PCs	20	0	RBF	0.90	0.05	0.90	0.07	0.89	0.10
ML 41	PCs	20	2	RBF	0.90	0.06	0.91	0.07	0.89	0.10
ML 42	PCs	20	4	RBF	0.90	0.05	0.91	0.07	0.90	0.09
ML 43	Pixels		0	Linear	0.88	0.05	0.91	0.06	0.84	0.11
ML 44	Pixels		2	Linear	0.89	0.05	0.92	0.05	0.84	0.12
ML 45	Pixels		4	Linear	0.89	0.06	0.92	0.07	0.84	0.12
ML 46	SBR			Linear	0.89	0.05	0.92	0.06	0.82	0.10
ML 47	SBR			RBF	0.89	0.06	0.91	0.07	0.85	0.10

Table 3-4 Machine learning cross validation results for the local database (subset B, adapted from (1))

Algorithm	**Feature**	**No. PCs**	**Dilate (times)**	**Kernel**	**Mean accuracy**	**SD**	**Mean Sensitivity**	**SD**	**Mean Specificity**	**SD**

ML 1	PCs	1	0	Linear	0.87	0.03	0.92	0.04	0.75	0.08
ML 2	PCs	1	2	Linear	0.86	0.04	0.92	0.04	0.75	0.10
ML 3	PCs	1	4	Linear	0.86	0.04	0.92	0.04	0.74	0.10
ML 4	PCs	2	0	Linear	0.96	0.02	0.97	0.03	0.93	0.05
ML 5	PCs	2	2	Linear	0.94	0.03	0.95	0.03	0.90	0.07
ML 6	PCs	2	4	Linear	0.93	0.03	0.95	0.03	0.89	0.07
ML 7	PCs	3	0	Linear	0.97	0.02	0.98	0.02	0.96	0.04
ML 8	PCs	3	2	Linear	0.97	0.02	0.98	0.02	0.96	0.04
ML 9	PCs	3	4	Linear	0.96	0.03	0.97	0.03	0.93	0.06
ML 10	PCs	5	0	Linear	0.97	0.02	0.98	0.02	0.96	0.05
ML 11	PCs	5	2	Linear	0.97	0.02	0.98	0.02	0.96	0.05
ML 12	PCs	5	4	Linear	0.97	0.02	0.98	0.02	0.96	0.04
ML 13	PCs	10	0	Linear	0.97	0.02	0.98	0.02	0.96	0.04
ML 14	PCs	10	2	Linear	0.97	0.02	0.98	0.02	0.96	0.04
ML 15	PCs	10	4	Linear	0.97	0.02	0.98	0.02	0.96	0.04
ML 16	PCs	15	0	Linear	0.97	0.02	0.97	0.02	0.95	0.04
ML 17	PCs	15	2	Linear	0.97	0.02	0.98	0.02	0.95	0.04
ML 18	PCs	15	4	Linear	0.97	0.02	0.98	0.02	0.96	0.04
ML 19	PCs	20	0	Linear	0.97	0.02	0.98	0.02	0.96	0.05
ML 20	PCs	20	2	Linear	0.97	0.02	0.98	0.02	0.95	0.05
ML 21	PCs	20	4	Linear	0.97	0.02	0.97	0.02	0.96	0.04
ML 22	PCs	1	0	RBF	0.87	0.04	0.91	0.04	0.79	0.09
ML 23	PCs	1	2	RBF	0.86	0.04	0.91	0.04	0.76	0.08
ML 24	PCs	1	4	RBF	0.86	0.04	0.91	0.04	0.75	0.10
ML 25	PCs	2	0	RBF	0.95	0.02	0.96	0.03	0.94	0.06
ML 26	PCs	2	2	RBF	0.94	0.03	0.94	0.04	0.93	0.06
ML 27	PCs	2	4	RBF	0.93	0.03	0.94	0.03	0.91	0.06
ML 28	PCs	3	0	RBF	0.97	0.02	0.98	0.02	0.97	0.04
ML 29	PCs	3	2	RBF	0.97	0.02	0.97	0.02	0.97	0.04
ML 30	PCs	3	4	RBF	0.96	0.03	0.97	0.03	0.95	0.04
ML 31	PCs	5	0	RBF	0.97	0.02	0.97	0.02	0.97	0.03
ML 32	PCs	5	2	RBF	0.97	0.02	0.97	0.03	0.97	0.03

ML 33	PCs	5	4	RBF	0.97	0.02	0.97	0.02	0.97	0.04
ML 34	PCs	10	0	RBF	0.97	0.02	0.97	0.02	0.97	0.04
ML 35	PCs	10	2	RBF	0.97	0.02	0.97	0.02	0.97	0.04
ML 36	PCs	10	4	RBF	0.97	0.02	0.97	0.02	0.97	0.03
ML 37	PCs	15	0	RBF	0.97	0.02	0.97	0.02	0.97	0.04
ML 38	PCs	15	2	RBF	0.97	0.02	0.97	0.02	0.97	0.04
ML 39	PCs	15	4	RBF	0.97	0.02	0.97	0.03	0.97	0.04
ML 40	PCs	20	0	RBF	0.97	0.02	0.97	0.02	0.97	0.04
ML 41	PCs	20	2	RBF	0.97	0.02	0.97	0.03	0.97	0.04
ML 42	PCs	20	4	RBF	0.97	0.02	0.97	0.02	0.97	0.04
ML 43	Pixels		0	Linear	0.95	0.02	0.97	0.03	0.92	0.06
ML 44	Pixels		2	Linear	0.95	0.02	0.97	0.03	0.92	0.06
ML 45	Pixels		4	Linear	0.96	0.02	0.97	0.03	0.93	0.06
ML 46	SBR			Linear	0.95	0.03	0.97	0.03	0.91	0.06
ML 47	SBR			RBF	0.95	0.02	0.96	0.03	0.93	0.06

Table 3-5 Machine learning cross-validation results for the PPMI database (adapted from (1))

Summary of semi-quantification performance results for the local database								
Method number	**Cut-offs defined by**	**SBRs**	**Accuracy**	**SD**	**Sensitivity**	**SD**	**Specificity**	**SD**
SQ 1	mean -2SD	L+R putamen	0.79	0.08	0.68	0.12	0.97	0.05
SQ 2	mean -2SD	L+R putamen, L+R caudate	0.78	0.08	0.68	0.11	0.96	0.06
SQ 3	mean -1.5SD	L+R putamen	0.85	0.06	0.82	0.09	0.90	0.10
SQ 4	mean -1.5SD	L+R putamen, L+R caudate	0.85	0.06	0.83	0.08	0.88	0.11
SQ 5	mean -1SD	L+R putamen	0.86	0.06	0.91	0.06	0.77	0.12
SQ 6	mean -1SD	L+R putamen, L+R caudate	0.86	0.05	0.92	0.06	0.75	0.13
SQ 7	min	L+R putamen	0.83	0.06	0.78	0.08	0.92	0.08
SQ 8	min	L+R putamen, L+R caudate	0.84	0.07	0.81	0.09	0.89	0.10
SQ 9	regress -2SE	L+R putamen	0.82	0.07	0.72	0.11	0.99	0.03
SQ 10	regress -2SE	L+R putamen, L+R caudate	0.82	0.06	0.72	0.10	0.98	0.04
SQ 11	regress -1.5SE	L+R putamen	0.86	0.06	0.82	0.09	0.93	0.09
SQ 12	regress -1.5SE	L+R putamen, L+R caudate	0.86	0.06	0.83	0.08	0.91	0.10
SQ 13	regress -1SE	L+R putamen	0.87	0.06	0.92	0.06	0.78	0.12
SQ 14	regress -1SE	L+R putamen, L+R caudate	0.87	0.06	0.93	0.06	0.77	0.12
SQ 15	ROC age matched	lowest putamen	0.87	0.05	0.89	0.06	0.83	0.11
SQ 16	ROC age matched	lowest putamen, lowest caudate	0.83	0.07	0.92	0.07	0.67	0.16
SQ 17	ROC	lowest putamen	0.86	0.06	0.86	0.08	0.86	0.13

SQ 18	ROC	lowest putamen, lowest caudate	0.84	0.06	0.90	0.07	0.74	0.14

Table 3-6 Semi-quantification cross-validation results for the local database (adapted from (1))

Summary of semi-quantification performance results for the PPMI database								
Method number	**Method**	**SBRs**	**Accuracy**	**SD**	**Sensitivity**	**SD**	**Specificity**	**SD**
SQ 1	mean -2SD	L+R putamen	0.93	0.03	0.92	0.04	0.97	0.04
SQ 2	mean -2SD	L+R putamen, L+R caudate	0.93	0.03	0.92	0.04	0.96	0.04
SQ 3	mean -1.5SD	L+R putamen	0.94	0.03	0.95	0.03	0.92	0.06
SQ 4	mean -1.5SD	L+R putamen, L+R caudate	0.94	0.03	0.95	0.03	0.90	0.07
SQ 5	mean -1SD	L+R putamen	0.92	0.03	0.98	0.02	0.78	0.09
SQ 6	mean -1SD	L+R putamen, L+R caudate	0.89	0.04	0.98	0.02	0.71	0.11
SQ 7	min	L+R putamen	0.90	0.04	0.87	0.05	0.96	0.04
SQ 8	min	L+R putamen, L+R caudate	0.90	0.03	0.88	0.05	0.94	0.05
SQ 9	regress -2SE	L+R putamen	0.93	0.03	0.91	0.04	0.97	0.04
SQ 10	regress -2SE	L+R putamen, L+R caudate	0.93	0.03	0.91	0.04	0.97	0.04
SQ 11	regress -1.5SE	L+R putamen	0.94	0.03	0.95	0.03	0.92	0.05
SQ 12	regress -1.5SE	L+R putamen, L+R caudate	0.94	0.03	0.95	0.03	0.90	0.07
SQ 13	regress -1SE	L+R putamen	0.92	0.03	0.98	0.02	0.80	0.08
SQ 14	regress -1SE	L+R putamen, L+R caudate	0.89	0.04	0.98	0.02	0.71	0.11

SQ 15	ROC age matched	lowest putamen	0.94	0.03	0.96	0.03	0.91	0.07
SQ 16	ROC age matched	lowest putamen, lowest caudate	0.89	0.03	0.97	0.03	0.73	0.09
SQ 17	ROC	lowest putamen	0.95	0.03	0.96	0.03	0.92	0.06
SQ 18	ROC	lowest putamen, lowest caudate	0.89	0.03	0.97	0.03	0.71	0.10

Table 3-7 Semi-quantification cross-validation results for the PPMI database (adapted from (1))

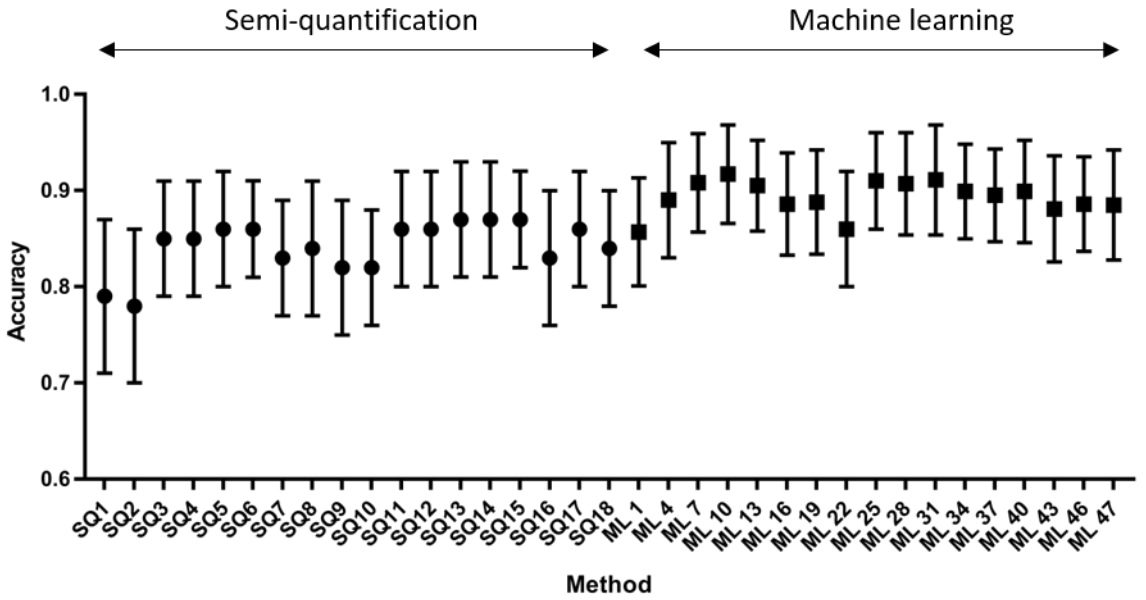


Figure 3-2 Accuracy results for all semi-quantification and machine learning methods (with 0 additional dilation) applied to local data. Semi-quantification results are grouped to the left of the graph (circular markers) and machine learning algorithms to the right (square markers). Whiskers represent one standard deviation. Taken from (1)

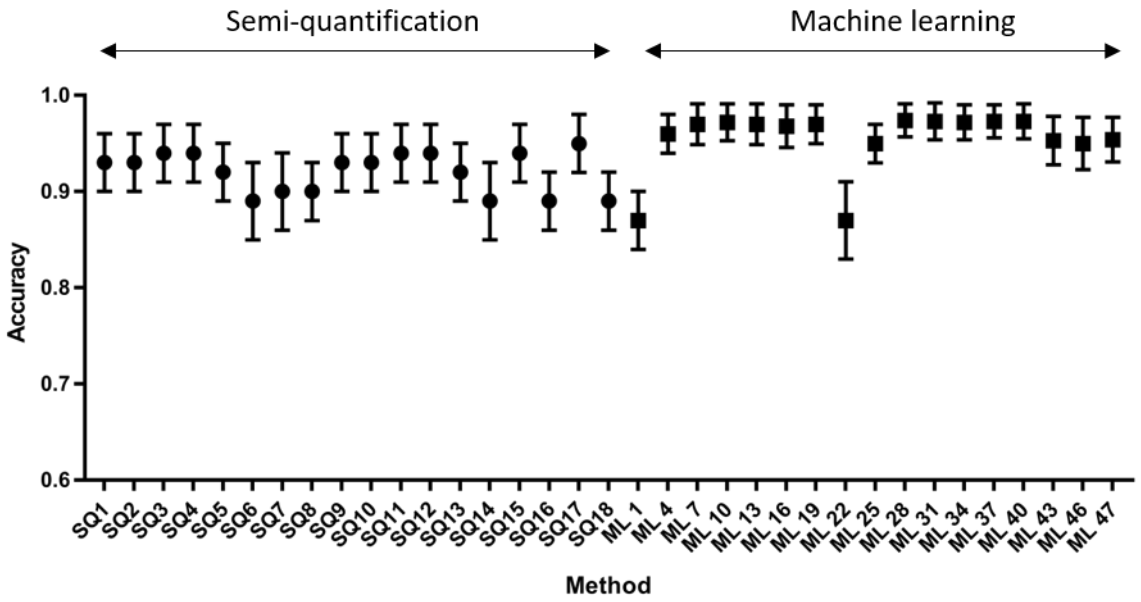


Figure 3-3 Accuracy results for all semi-quantification and machine learning methods (with 0 additional dilation) applied to PPMI data. Semi-quantification results are grouped to the left of the graph (circular markers) and machine learning algorithms to the right (square markers). Whiskers represent one standard deviation. Taken from (1)

Learning curve results from algorithms ML 10, ML 43 and ML 46 are shown in Figure 3-4, illustrating mismatched train-test performance figures for voxel intensity features, as compared to algorithms taking PCs and SBRs as input features, where train-test performance figures are more consistent with each other.

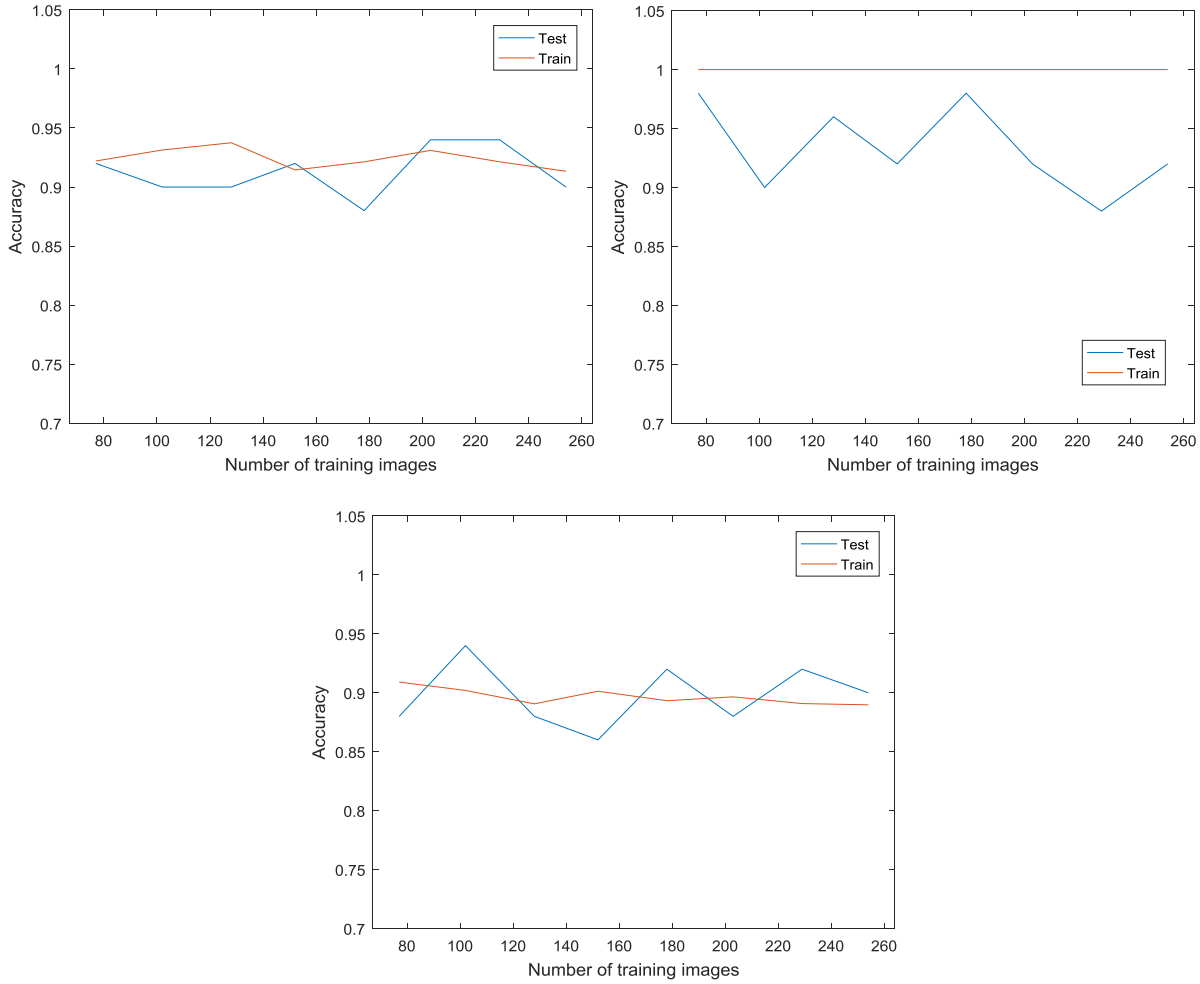


Figure 3-4 Learning curves for linear SVM algorithms using 5 PCs (top left), voxel intensities (top right) and SBRs (bottom) as input features (and no additional mask dilation, ML 10, 43 and 46)

3.2.3 Discussion

This investigation provided a detailed study of the differences in performance that might be expected when using machine learning algorithms for classification rather than semi-quantitative methods. Two contrasting databases of images were used, of different sizes, demonstrating how performance can change depending on the characteristics of the available data

Semi-quantification

For the semi-quantitative methods, performance was superior for the PPMI data as compared to local data, with higher mean values generated for the PPMI data and lower variance. This was as expected and highlights the differences in performing measurements on a research database, where screening procedures and imaging investigations are tightly controlled, in contrast to real clinical data where the reference diagnosis is less certain, the patient group more diverse and where inter-camera calibration is not routinely conducted. The measured performance of semi-quantitative methods for the local data was similar to that found by other researchers utilising a mixed clinical database, with established and commercial software tools (48). This provides added confidence that the methods developed in this study had similar discriminatory power to that of existing software that is used in clinic.

Clinically, multiple SBRs and other derived ratios may be provided by semi-quantitative software to guide diagnosis. Typically, SBRs from the whole striatum as well as individual caudates and putamena on the left and right side are given. In addition, the caudate to putamen ratio and the right to left ratio may also be displayed. If all these individual SBRs and their associated normal limits are treated as individual tests, the final semi-quantification classification is likely to be overly sensitive (increasing the risk of type I error) and may give a pessimistic view on current standard of care approaches. Therefore, in this study only SBRs from individual putamena (with or without caudate results) were considered and so it is more likely that results reflect the best achievable from semi-quantification rather than typical performance.

However, the way in which semi-quantification was evaluated is not completely reflective of its clinical function. The normal limits for each set of SBR results were used as hard cut-offs in that any value falling outside the boundary would lead to an overall abnormal image classification. In reality, semi-quantification is unlikely to be used in such a rigid manner. Semi-quantification results require some interpretation by clinicians and figures are usually treated as a whole, in light of other clinical information. It is therefore possible that a single SBR result which just falls outside the normal limits would be interpreted as an overall normal result, despite the classification being abnormal according to the rules of this study. Given this limitation, it should be kept in mind that measures of standalone performance do not reveal all of the advantages and disadvantages of a particular assistive reporting platform.

Semi-quantitative methods gave a relatively narrow range of accuracy scores across all the methods tested, with a wide range of sensitivities and specificities. It is interesting to note that two of the methods which treat classification as a two class problem, generating cut-offs from both normal and abnormal putamenal SBRs (i.e. methods SQ 15 and SQ 17), produced some of the highest accuracy figures, with lower variance and well balanced sensitivity and specificity values. This is perhaps unsurprising as all other semi-quantitative methods (which are more reflective of commercially available tools) define cut-offs from the normal population only, with no knowledge of the distribution or likely crossover of abnormal data.

In general, the addition of caudate data to semi-quantitative calculations caused a slight increase in sensitivity and slight reduction in specificity with little effect on accuracy, other than for methods based on ROC curve calculations, which saw a drop in performance. This suggests that the vast majority of diagnostically useful information can be gleaned from consideration of putamen uptake only. Again, this is unsurprising as image appearances often show more marked reduction in putamen uptake than in the caudate (34).

It should be noted that the Southampton semi-quantification method (39) was not investigated in this study. Recent research (48) suggests that the sensitivity of this approach is very low when calibration is not performed between different camera systems and is also significantly reduced when correction (including scatter correction) is not performed. Camera-specific calibration data was not available for the local database of images and scatter data were not accessible for the PPMI dataset and so the method was excluded.

Furthermore, only one method of image registration and SBR calculation was used in this investigation. Commercial semi-quantification solutions use different registration algorithms. Some software also performs quantification in two dimensions, by summing consecutive slices, utilising different regions of interest. Other image corrections that are also sometimes implemented (such as partial volume correction) were not considered in this investigation. Thus, results presented here cannot be representative of all semi-quantification techniques. However, it is unlikely that investigation of a wider range of methods would have led to significantly different performance as the same fundamental limitations apply to all semi-quantification techniques.

Machine Learning

The machine learning algorithms produced performance metrics which generally exceeded that of the semi-quantitative methods on the same data. Other than algorithms based on just one principal component (ML1 - ML3), the machine learning algorithms all gave accuracies as high as or higher than any of the semi-quantitative methods. Accuracy, sensitivity and specificity were generally high and well balanced for each machine learning tool, with small standard deviation values, providing evidence that these approaches are accurate and have low variability. The smaller standard deviation results as compared to semi-quantification is particularly important from a clinical perspective as this suggests that machine learning tools will be more robust when used more widely on new datasets.

Machine learning performance metrics for the PPMI data matched the best performing algorithms produced by other authors (see Table 1-3), with results that are comparable with current state-of-the-art. This provides some justification for the particular model selection (and grid search) processes used in this study. As with the semi-quantitative results, performance for the PPMI database was substantially higher than for the local data, reinforcing the assertion that classification of the PPMI dataset is an easier task than that seen in clinical reality.

For both databases, algorithms using principal components as features gave the highest accuracies (as high as 0.97 for the PPMI database), though the addition of larger numbers of principal components and the use of a non-linear RBF kernel appeared to have little additional impact on results. Greater dilation of the image mask, incorporating a greater proportion of the brain, also appeared to have a minimal impact on results in most cases. The fact that high accuracies were achieved with just 2 principal components (only slightly lower than that measured for 20 components), shows that separation between groups can be achieved with very limited numbers of variables. The lack of significant improvement in classification accuracy using greater than 2 principal components was also reported by Towey (65). 2 PCs accounted for over 80% of the total variance in the training data when applied to the local database, demonstrating that (I123)FP-CIT images have few modes of variation, even for this relatively diverse set of patients. Results for one principal component were the lowest of all machine learning algorithms and the contrast in performance as compared to the other algorithms was particularly striking for the PPMI data. This demonstrates that the second principal component contains significant diagnostic

information and that there is a lower limit on the degree of algorithm simplification that can be applied without adversely impacting upon classification performance.

Features based on raw voxel intensities gave slightly lower performance values in general, for example achieving an accuracy of 0.88 on the local database (algorithm ML43). Furthermore, there is evidence to suggest that such algorithms were associated with higher variance than the other types of algorithm investigated. The learning curve depicted in Figure 3-4 shows that for algorithm ML 43, accuracy on training data was consistently at 100% across the available training data subgroups. This is a strong indicator that the algorithm is fitting a model to the available data, rather than the underlying trend. Comparison to performance on the independent test data shows a variable but consistent gap in results across all different numbers of training data, which again suggests that the extent to which the algorithm will generalise to other data samples is sub-optimal. This may be because the algorithm was performing classification largely based on individual voxel values that separate out test data well, but which may not be in a spatial location that correlates with the presence of disease. It is likely that the training and test performance curves would move closer together if many more training images were available. However, collecting much larger numbers of datasets is not feasible in this study.

Conversely, learning curve results for the algorithms based on 5 principal components (ML 10) and SBRs (ML 46) showed training and test performance figures that were more closely matched, at a relatively high level, even for the smallest number of training images. This suggests that both variance and bias were relatively low and that many fewer training samples than the 306 available could be used to create machine learning tools of high accuracy. Importantly, results indicate that algorithms based on principal components or SBRs are more likely to give more consistent performance when applied to new test data, than algorithms based on raw voxel intensities.

Despite the favourable learning curve results, machine learning algorithms based on SBRs produced mean performance figures from the main cross-validation comparison that were slightly lower than that of PC-based algorithms. In this case reduced performance is likely to be a consequence of the limitations of using features based on ratios of mean intensity values inside regions of interest. In particular, SBRs do not contain information on the shape of striatal uptake patterns and they are reliant on accurate fitting of small regions of interest to the anatomical striatal outline. This reflects limitations that generally apply to semi-quantification methods.

Although in general the machine learning algorithms appeared to perform better than the semi-quantification tools, the level of absolute performance improvement as compared to the best performing semi-quantification techniques was relatively small in this study. It is difficult to determine whether differences were statistically significant due to the non-independence of training and test data in each fold. However, examination of the standard deviation on performance results (see Figure 3-2 and Figure 3-3) suggests that there is some crossover in accuracy of the machine learning and semi-quantitative methods, particularly for the local data. Although utilising different techniques and a different evaluation methodology, Towey also reported accuracy results from machine learning algorithms that were similar or better than that of selected (commercial) semi-quantification tools (65). This small gain in performance from machine learning tools should be kept in mind in the following investigations of impact on reporting.

Given that standalone semi-quantification accuracy is up to approximately 87% for clinical data (and 95% for research data), the margin available for performance gains from new machine learning algorithms is real but narrow. Even with the introduction of more advanced machine learning tools (such as convolutional neural networks) there cannot be a substantial gain in accuracy over the classical algorithms presented here, which suggests that developing more advanced software is of limited value. Therefore, the following sections will continue to use the machine learning algorithms defined in this chapter.

Comparisons of standalone accuracy are not, by themselves, an adequate test of clinical utility. As previously suggested, clinical investigations demonstrating the impact of such assistive reporting tools on clinical decision making are required to fully understand any potential benefits. One particular aspect not covered by the current study is that machine learning algorithms simplify the information that is shown to the clinician. Rather than having to examine and interpret multiple SBR results and other ratio data, along with their normal ranges, clinicians are presented with a single number representing the overall likelihood of abnormality. This less ambiguous software output may be a better, more effective way of influencing clinicians' decisions.

Overall, this investigation showed that there is a small gain in absolute standalone classification performance (and a corresponding reduction in variability) that can be gained from using effective machine learning algorithms rather than the best performing semi-quantification methods. The superiority of machine learning algorithms is more substantial if

only ‘one class’ semi-quantification methods are considered (which is the form of semi-quantification that is frequently used in commercial, clinical tools). Furthermore, in this investigation only a subset of possible striatal uptake ratios were included in performance metric calculations. Typically, a greater range of semi-quantitative values is presented to reporting radiologists in clinic. If greater numbers of ratios were included in the analysis, classification performance for semi-quantitative methods would have been lower, providing more compelling evidence of the benefits of machine-learning algorithms.

3.2.4 Conclusion

This study has compared a range of semi-quantification approaches with different machine learning tools (based on SVMs) in order to evidence whether classical machine learning techniques are a superior means of classifying (I123)FP-CIT data into normal and abnormal groups. A research and local clinical database were used for repeated 10-fold cross-validation.

Results showed that classification performance was lower for the local database than the research database for both semi-quantitative and machine learning algorithms.
However, for both databases, the majority of the machine learning methods generated high mean accuracies with low variability, and well balanced sensitivity and specificity. Results compared favourably with that of semi-quantification methods and are comparable with accuracies cited for clinician performance.

Learning curve results indicate that algorithms taking raw voxel intensities as inputs were associated with high variance. In contrast, algorithms based on 5 principal components or SBRs were well balanced with low bias and variance.

The increase in accuracy offered by machine learning algorithms as compared to the best performing semi-quantification methods was relatively small. However, the performance gap is likely to be an underestimate of that which might be seen in clinic with commercial, clinical semi-quantification packages. Furthermore, machine learning algorithms offer other benefits, such as the generation of just a single output, rather than multiple outputs, each of which must be interpreted by clinicians. Thus, the evidence suggests that machine learning algorithms may provide more effective and better assistance to reporters than established clinical reporting aids.

Further evidence is now required to establish whether machine learning can enhance reporter performance, since it is envisaged that the human reporter will continue to make the final diagnostic decision for (I123)FP-CIT tests.

4 Impact on reporting performance – pilot and main studies

Objectives addressed by this section (in black, bold):
1) Select and implement machine learning classification tools
2) Collect a database of (I123)FP-CIT images
3) Compare the performance of machine learning algorithms with semi-quantification
4) Develop software for testing of human reporters
5) Assess the impact of an automated classification tool, implemented as a CADx system, on reporting

Table 4-1 Objectives addressed in section 4

This chapter builds on the promising performance results of the previous chapter by measuring the impact of machine learning algorithms on reporter performance. Such tests are necessary for quantifying the effectiveness of CADx. However, performing clinical trials of medical devices is a costly and time consuming process. In order to minimise the risk that data are biased or uninformative a number of factors need to be considered, which in this case includes: the numbers of patient datasets to use, the number (and experience level) of radiologists to recruit, study design, the form of CADx output to adopt and the form of its display on the screen. In the absence of previous studies covering this specific topic, particular care should be taken when considering how the reporting investigation should be performed.

In this work a pilot investigation was undertaken on a subset of the available data. Results from the pilot study were used to inform a subsequent larger reporting study and the following sections describe preparations for the pilot and main study, in particular the derivation of optimal machine learning algorithm parameters and development of reporting software. Following this, a summary of the testing methodologies is described, with results and a discussion.

Of the available machine learning models that were trained and evaluated in previous sections, those based on principal components had been found to perform best in terms of mean classification metrics. These algorithms also appeared to have well balanced bias and

variance properties, in contrast to learning curve analysis of features based on raw voxel intensities, which showed signs of increased variance. Although there was much overlap between performance figures, the highest classification score was achieved by an algorithm based on 5 principal components, with a linear SVM model and no additional mask dilation (see section 2.1.3, algorithm ML10). Therefore, this was the algorithm chosen in the pilot and main study.

Sections of the following method, results and discussion, particularly in relation to the main study, were published in a peer-reviewed journal article (2).

4.1 Derivation of optimal SVM hyperparameter

Although the previous chapter provided an estimate of machine learning algorithm performance on independent data, using different models, it is not yet known what the optimal hyperparameter is for the available training data, i.e. which C value should be chosen for the SVM classifier. Previously, optimal hyperparameters were estimated from within nested cross validation loops using approximately 81% of the data for training in each pass and 9% for validation. This information was then passed to the outer loop to estimate performance on the final 10% of the data. However, when the goal of cross-validation is simply to find optimal parameters, the outer cross validation loops are not required. Thus, optimal hyperparameters can be chosen using more of the available data, in a standard 10-fold cross validation procedure i.e. using 90% of the data for training in each loop and 10% for testing. Utilising an incrementally greater proportion of the data to find the best hyperparameters may help to create an improved classification tool.

For the pilot study, the optimal value of the C hyperparameter was selected by running a 10 fold cross-validation procedure repeated 10 times. Within each set of cross validations the C parameter was changed according to a sparse grid search, as before (see section 3.2.1). The highest mean F1-score was used as the selection criteria. With this approach the optimal C value was found to be 2^{-2} or 0.25. The final classification algorithm was then derived by training on all the available training data from subset B, using this C value. This algorithm formed the basis for the CADx tool.

4.2 Software development

Measuring reporting performance using realistic software (i.e. as foreseen for the clinical situation) is a key consideration. Significant deviations from the setup used clinically introduce additional uncertainty to results. Therefore, for these reporting studies the current standard clinical image viewer, Jview, was adapted to show CADx results to clinicians alongside image data in a standard format. Jview is a platform based on java software. Additional software was written to augment the functionality already available, rather than re-writing the clinical software code.

Data for the study were stored within a remote MySQL database, other than the image data itself which was held within a separate, remote filestore. An applet was written in Java to automatically manage the display of data to radiologists and to send results to the database. A schematic depicting each piece of software used and how they linked to each other is displayed in Figure 4-1. This particular model has the advantage that all data and the software classes are kept on a remote server, which can be accessed by several workstations at once. Any updates to the software code or data are therefore immediately passed on to local users. Each workstation only requires java to be installed. The remote servers and workstations depicted in Figure 4-1 are all within the Sheffield Teaching Hospitals IT network. No external hardware was used for this study.

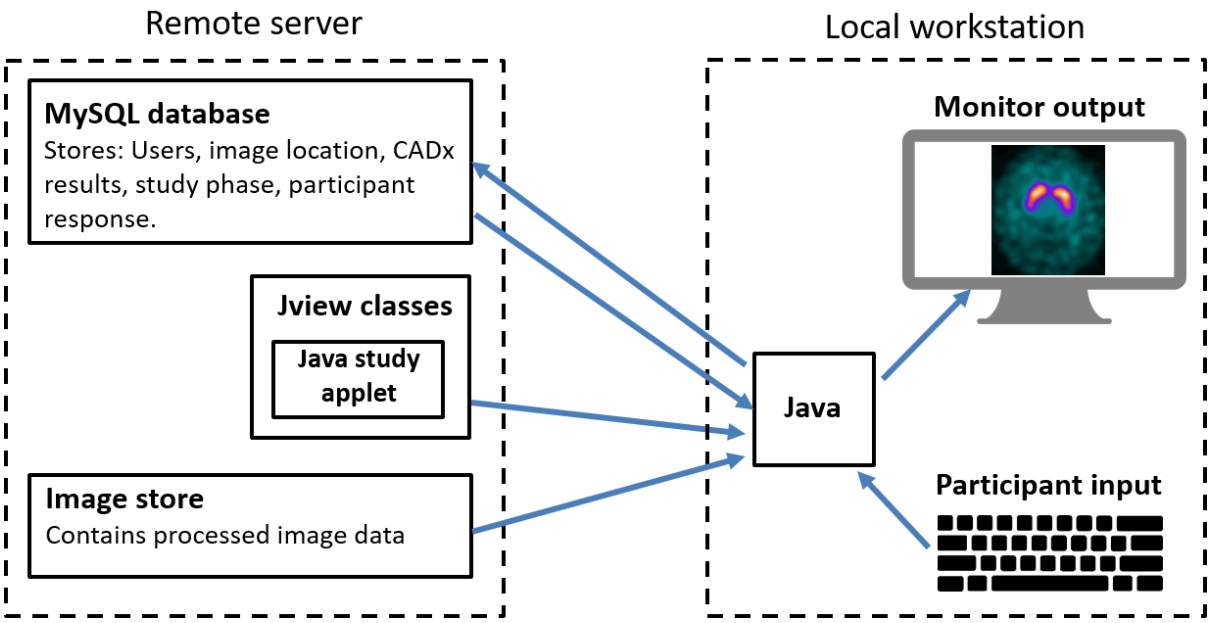


Figure 4-1 Schematic depicting the different elements of the data capture and display software used for the reporting study. Blue arrows represent data flows

It was hypothesised that the most helpful form of CADx output would be a probability value. This would give the radiologist an idea of the relative certainty of the machine learning tool, in addition to a decision on the binary classification. However, the output from a standard SVM function is a number which simply dictates on which side of the separating plane the test case lies. Values greater than zero are classified as a particular class and values less than zero classified as the alternative class. These are not probabilities. An SVM score needs to be viewed in the context of the typical values that would be expected for examples of either class in order to give the clinician an indication of relative confidence. Thus, a common step often applied to raw SVM scores is to convert them to probability values using techniques such as Platt scaling (98), whereby a logistic regression model is fitted to the available data. libSVM's inbuilt function for converting SVM scores to probabilities was adopted for this purpose [https://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf], which utilises cross-validation to fit the available data to the logistic function.

The probability of belonging to the abnormal class was calculated for all the patient cases and added to the MySQL database. These values were displayed to the left of the user's computer screen at an appropriate point during the study. At all other times the CADx output was hidden. Given that the trained algorithms were binary, for cases where $P >= 0.5$, the corresponding probability of belonging to the normal class was 1-P (i.e. less than 0.5). For these patients the CADx output value was displayed in red font. For patients where $P < 0.5$, the corresponding probability of belonging to the normal class was greater than 0.5 and a blue font was used in the display. Thus, in this scheme red font is associated with an abnormal diagnosis.

In the standard clinical protocol (I123)FP-CIT images are viewed following rigid registration to a template (to remove asymmetric appearances associated with head tilt). Four reconstructed slices are typically displayed to the reporter, from within the centre of the brain. A summed image, derived from axial slices throughout the central brain is also available. The java applet written for the study enforced this display format for every case. Additionally, a series of buttons were located in the left-hand pane to allow the user to move between cases and provide a classification decision. These were provided below the box which was used for display of the CADx output. Figure 4-2 provides an example of the software display that the reporters saw, in this case with the CADx result being visible.

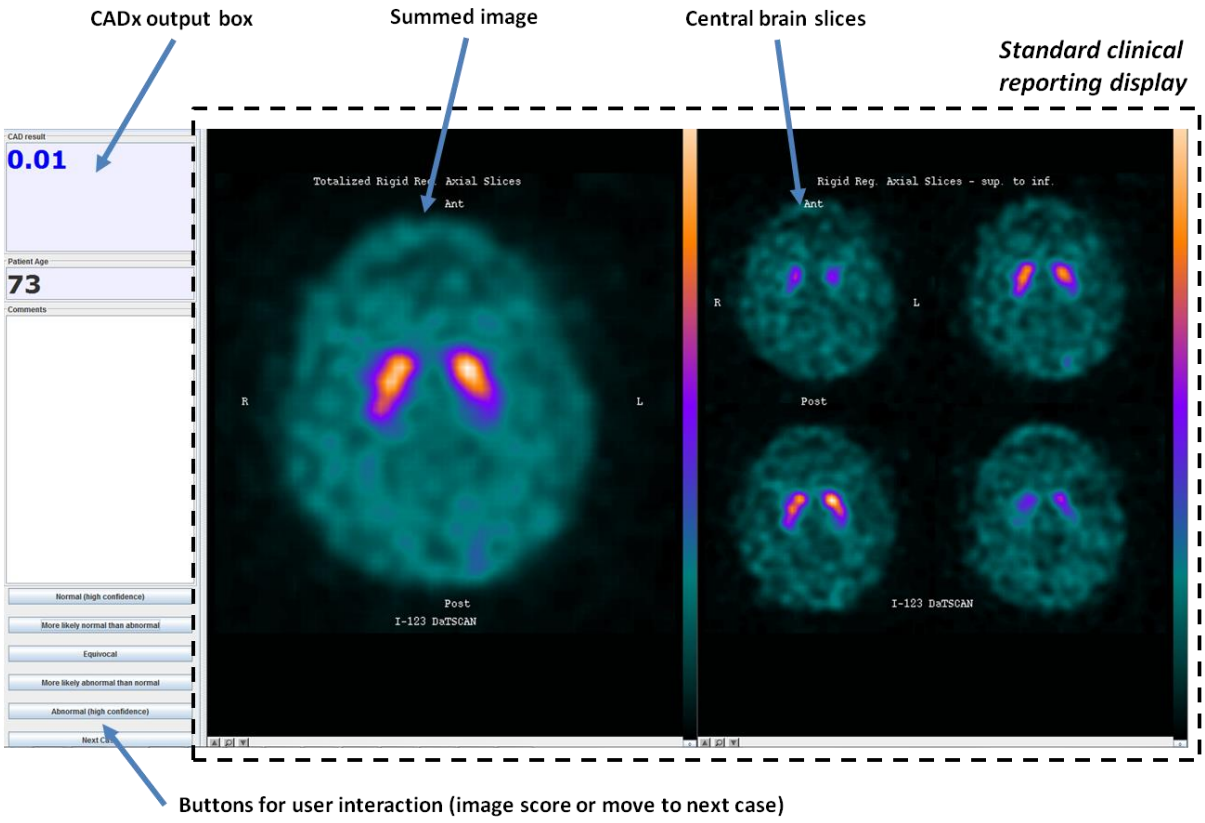


Figure 4-2 Example of the Jview software display provided to reporters (The CADx probability output is visible in the top left corner this case. The number below refers to the patient age on the day of the scan)

4.3 Pilot study

The pilot study had three main aims:

- Quantify the influence of CADx on reporting accuracy and reliability
- Obtain qualitative feedback on the current CADx design
- Document the effects of CADx on reporting behaviours

In order to gain an estimate of inter-reporter reliability multiple reporters were required. Due to limitations in the number of available staff with necessary expertise, the pilot study focused on 7 junior radiologists (specialist registrars) who had significant experience of reporting images, but not of reporting (I123)FP-CIT scans specifically.

4.3.1 Method

The methodology adopted for the pilot study and the full clinical study are largely dictated, where possible, by the recommendations set out by Eadie and colleagues (99) following a review and critique of existing literature on the impact of assistive reporting software (100). This aims to make the results as relevant to clinical practice as possible, within the particular constraints of the study, and that the metrics measured are of clinical relevance.

The overall approach involved reporters examining images three times and giving a diagnostic confidence score in each case on a scale from 1 to 5. A score of 1 was equivalent to having high confidence that the image showed abnormal dopaminergic function and a score of 5 was equivalent to having high confidence that the image was normal. Scores of 2 and 4 were assigned to images where reporters were less confident in their overall assessment, but still favoured one of the binary choices and a score of 3 was used for any equivocal cases.

An overview of the pilot study methodology is shown in Figure 4-3. A chronological description of the main steps involved is summarised below:

1) **Training**. An introductory lecture was delivered on tracer uptake processes, image acquisition and indications to provide context to the reporting exercise. A series of 10 images (separate to those of subset A) were shown to the group to demonstrate typical normal and abnormal appearances. Following this a further 10 datasets were displayed and the group was encouraged to give their opinions on classification, in order that training could be reinforced and checked. Finally, the radiologists were given an overview of the functionality of Jview along with a clear, concise summary of the information that would subsequently be displayed to them. A brief introduction to the concept of CADx was given, which included an explanation of the probabilistic output of the machine learning tool used in the study. It was explained to the radiologists that the algorithm had been trained to accurately distinguish between normal and abnormal (I123)FP-CIT scans and that initial tests on other data had suggested a binary accuracy of approximately 90%. The worksheets used as a guide to the pilot study are shown in the appendix (see Appendix 1).
2) **Read 1**. Reporters scored all 30 images, shown in a random order, through visual assessment.
3) **1 hour break** (to reduce recall bias)

4) **Read 2 and 3**. Reporters examined the images again, shown in a different random order (read 2). However, immediately after giving a diagnostic score from visual analysis the same image was presented alongside a probability value from the machine learning tool. The reporters then gave a score for a third time (read 3). Thus, comparison between the first and second visual reads provided an insight into intra-reporter reliability. Comparison of the second and third reads gave an indication of the impact of CADx (i.e. whether the reporter chooses to change his / her decision when supported by CADx software).
5) **Questionnaire**. Each reporter was given a questionnaire to fill in. Questions were designed to assess the influence of CADx software on clinician decision making. In addition, the questions solicited information on possible differences in the way that the CADx system could be designed or used. The questionnaire included a mix of open and closed questions with both restricted response categories (to allow for more straightforward analysis) and the opportunity for general comment. Where possible, questions were posed in a neutral manner in order not to overly influence the response.

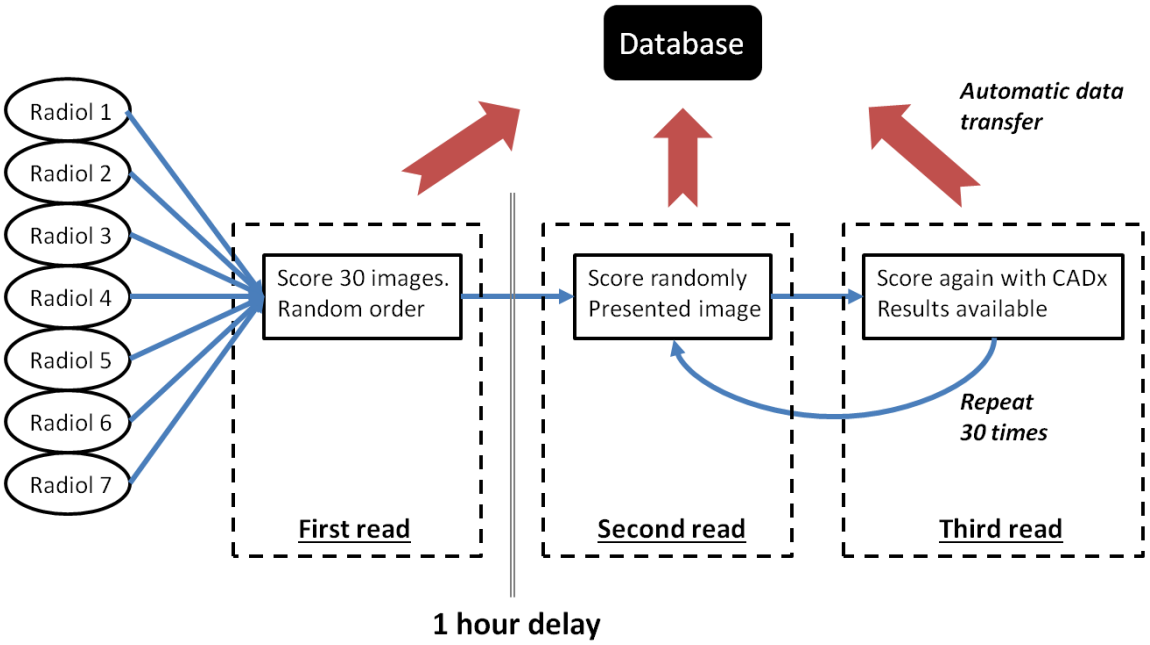


Figure 4-3 Overview of pilot study methodology

All radiologists were in the same room during the study and all used separate workstations, although these computers were not optimised for clinical reporting. An advisor was available

throughout the study to provide technical support if required. In this pilot investigation patient age was not revealed to the radiologists and the patients' clinical history was not available.

The standalone performance of the machine learning algorithm was also measured for the 30 cases. This was done to confirm that the assisted reporting tool, trained on cases where only the radiology report was available for ground truth classification, was sufficiently accurate when exposed to independent data with a clinical gold standard diagnosis.

The performance metrics selected for the pilot study were in line with those normally used in clinical investigations, namely sensitivity, specificity and diagnostic accuracy. These metrics were calculated by compressing the submitted confidence scores into 3 classification categories: with disease, without disease and equivocal.

In addition, intra-class correlation coefficient (ICC) was used for evaluating intra and inter-reporter reliability. Values of ICC can range from 0 to 1 where 1 represents perfect reliability with no measurement variability and zero is representative of no reliability. ICC is calculated from the ratio of variance between subjects (patients) as compared to the total variance (which includes between-subjects variance and error variance). In this study, the two-way random model was implemented for measuring inter-reporter reliability, with single measures (i.e. ICC(2,1)), and the one-way random model with single measures (i.e. ICC(1,1)) implemented for assessing intra-reporter reliability. These particular forms of ICC were selected based on the guides by Rankin (101) and Koo (102).

4.3.2 Results

Quantitative

The results presented below summarise the data transferred to the MySQL database during the pilot study. Unfortunately, a technical fault invalidated results from two of the radiologists on their final read. Therefore, the tables and figures below are not quite complete. Table 4-2 presents an overview of the quantitative results captured in the database whilst Table 4-3 provides a comparison between mean performance figures for read 2 and read 3 (for those radiologists with a complete set of results). Figure 4-4, Figure 4-5 and Figure 4-6 display the summary data in graphical form.

	Radiologist						
Metric	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Sensitivity read 1	0.75	1.00	0.94	1.00	1.00	1.00	0.94
Sensitivity read 2	1.00	1.00	1.00	1.00	1.00	0.94	0.81
Sensitivity read 3	1.00		0.88	1.00	1.00		0.88
Specificity read 1	0.71	0.71	0.86	0.86	0.86	0.64	0.93
Specificity read 2	0.79	0.79	0.93	0.86	0.86	0.79	0.93
Specificity read 3	0.93		0.93	0.86	0.93		0.93
Accuracy read 1	0.73	0.87	0.90	0.93	0.93	0.83	0.93
Accuracy read 2	0.90	0.90	0.97	0.93	0.93	0.87	0.87
Accuracy read 3	0.97		0.90	0.93	0.97		0.90

Table 4-2 Summary of quantitative results for the pilot study

	Mean	**95% CI (lower)**	**95% CI (upper)**
Sensitivity read 2	0.96	0.80	1.13
Sensitivity read 3	0.95	0.82	1.08
Specificity read 2	0.87	0.75	0.99
Specificity read 3	0.91	0.85	0.98
Accuracy read 2	0.92	0.85	0.99
Accuracy read 3	0.93	0.87	1.00

Table 4-3 Mean performance figures for read 2 as compared to read 3 (for radiologists 1,3,4,5 and 7)

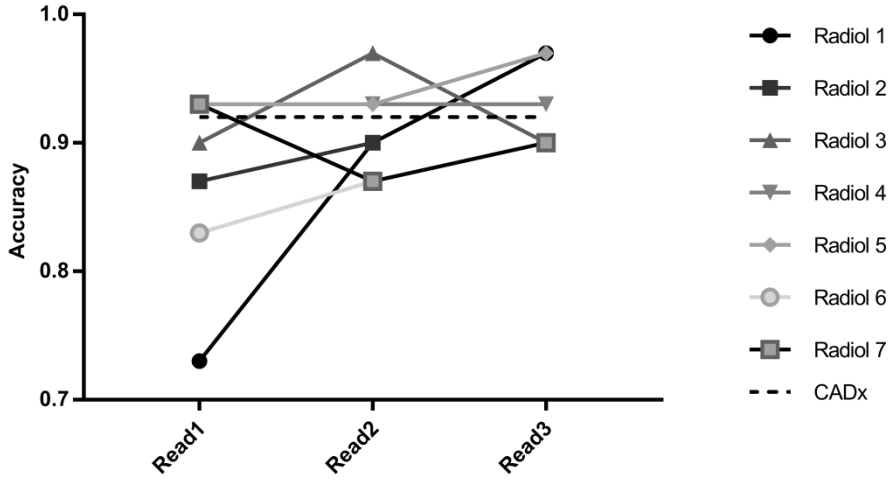


Figure 4-4 Diagnostic accuracy figures for the 3 image reads, as compared to standalone CADx performance

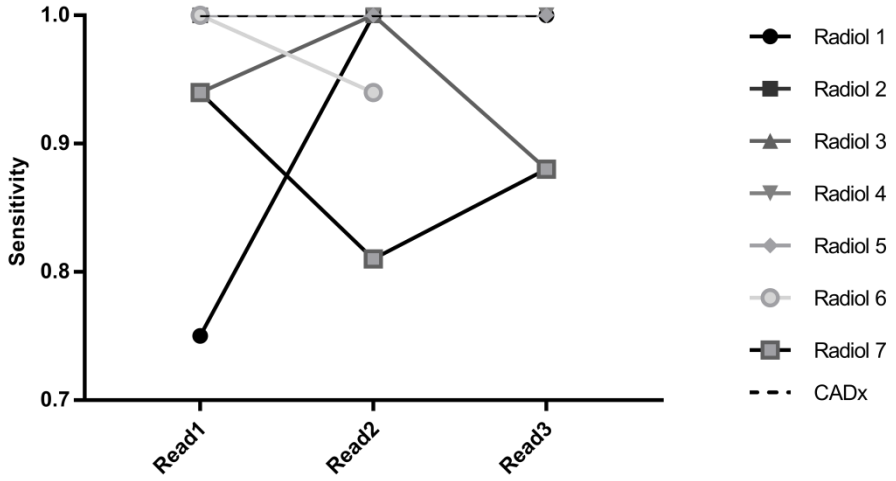


Figure 4-5 Sensitivity figures for the 3 image reads, as compared to standalone CADx performance

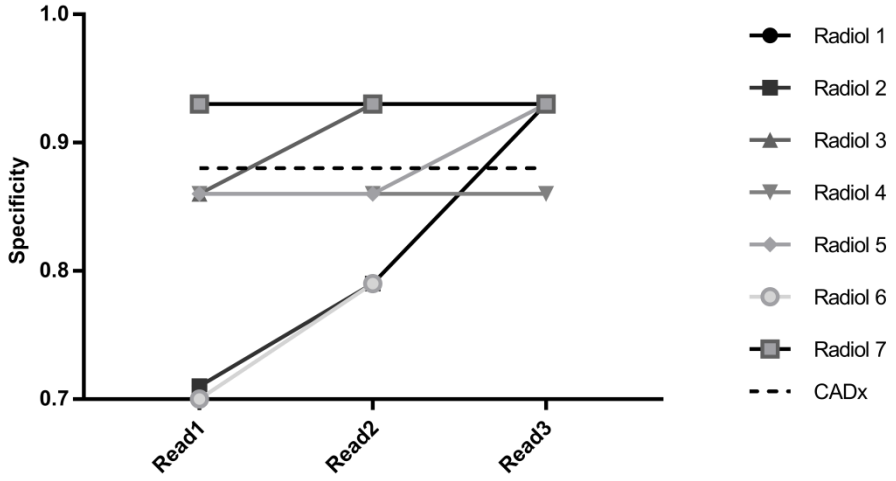


Figure 4-6 Specificity figures for the 3 image reads, as compared to standalone CADx performance

Examining the impact of CADx on a scan by scan basis, comparing read 2 and 3, there was a change in classification score in approximately 15% of cases taken across all radiologists where data was complete. Table 4-4 and Figure 4-7 summarise the intra- and inter-reporter reliability results, respectively

	Intra-reporter reliability		
Radiologist	**ICC**	**95% CI (lower)**	**95% CI (upper)**
1	0.65	0.39	0.82
2	0.82	0.71	0.93
3	0.93	0.87	0.97
4	0.91	0.83	0.96
5	1.00	0.99	1.00
6	0.72	0.49	0.86
7	0.84	0.70	0.92

Table 4-4 Intra-reporter reliability (ICC) results for all radiologists

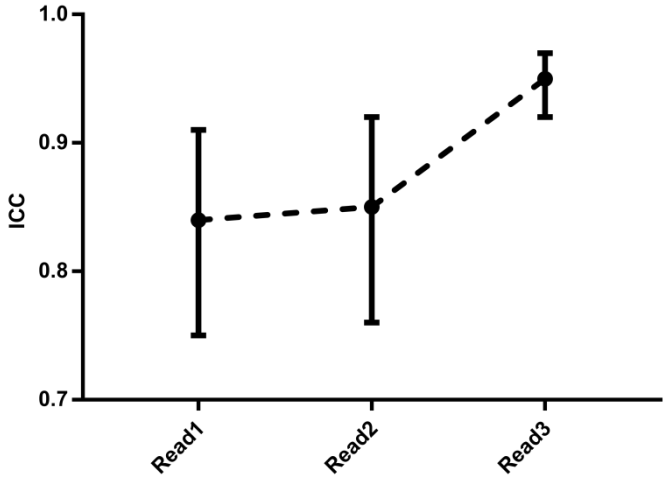


Figure 4-7 Inter-reporter reliability (ICC) results for each of 3 image reads (for radiologists 1,3,4,5 and 7). Whiskers represent 95% confidence intervals

Qualitative

The questionnaire revealed that 3 of the radiologists had used quantitative diagnostic tools before and so the idea of assisted reporting was not an entirely new concept to all the participants. However, as expected, none of the radiologists stated that they had previously reported (I123)FP-CIT images.

The following results summarise the responses given on the distributed questionnaire in relation to the CADx software tool. Each question is first stated and the proportion of the radiologists giving different answers is then shown. Any additional written comments which are of interest are also stated

Question 1: In general, how would you rate the impact of the CAD algorithm on your reporting decisions?
Responses: 0/7, - no impact, 3/7 – small impact, 3/7 – moderate impact, 1/7 - substantial impact
Comments:
“Gave more confidence if computer agreed with what I thought, disappointing when I wasn't sure and computer wasn't either”
“Would help identify borderline cases that need more scrutiny”
“Helped to re-look and decide”

“Good impact when CAD agreed with my opinion. Prompted me to view the image again”
“Made up my mind when I thought abnormal but unsure”

Question 2: To what extent did you trust the CAD algorithm results?
Responses: 0/7 – not at all, 0/7 – a little, 4/7 – moderately, 3/7 – a lot
Comments:
“I was more trusting when it agreed with me when I was confident”
“Still new to CAD therefore would need to 'trust' and prove it is accurate”
“If the CAD algorithm was in line with my thought then more likely to trust it”

Question 3: Would you prefer a binary CAD output as opposed to a probability value?
Responses: 0/7 – yes, 5/7 – no, 2/7 – not sure

Question 4: Would it benefit you if the CAD system also provided information on the location of image abnormalities?
Responses: 0/7 – no, 0/7 – yes (small benefit), 0/7 – yes (moderate benefit), 6/7 – yes (substantial benefit), 1/7 – not sure
Comments:
“Depends on accuracy and would it be able to identify several areas or 'globally' abnormal?”
“This will help to concentrate on that abnormal area and decide”
“It would help to highlight areas to review”

Question 5: To what extent would the CAD system be a useful training tool to improve DaTSCAN reporting performance for inexperienced clinicians?
Responses: 0/7 – no benefit, 0/7 – small benefit, 1/7 – moderate benefit, 5/7 – substantial benefit, 1/7 – not sure

4.3.3 Discussion

This pilot study represents the first test of machine learning classification tools for assisted (I123)FP-CIT reporting. There are no directly comparable studies in the literature and so findings must be largely considered on their own merit. Results are discussed below with respect to quantitative findings first, followed by qualitative analysis.

Quantitative

The quantitative data produced by the study indicates the potential value of the designed CADx system. One of the main findings from the data was that the performance of individual radiologists was variable both in terms of consistency in decision making and metrics of overall performance. For example, radiologist 1 had highly variable sensitivity and accuracy scores between read 1 and 2 (see Figure 4-4 and Figure 4-5) despite there being no change in reading conditions (with only a one hour break between reads). Their intra-reporter reliability score (ICC) was also low. Conversely, radiologist 5 had very consistent performance scores across the 3 image reads and had a perfect ICC score.

However, examination of Table 4-2 shows that for 6 of the 7 radiologists individual accuracy results were in the range 0.83-0.97 when reading images without CADx assistance (i.e. reads 1 and 2). This is in line with visual accuracy results previously reported by other authors, with more experienced radiologists (see section 1.1.5). This suggests that relatively accurate reporting of (I123)FP-CIT images, at least in terms of binary classification, can be achieved with little training. It also suggests that for the majority of radiologists recruited in this study there is no compelling evidence that performance is different to that of the more experienced radiologist population.

Despite the variability seen at the individual radiologist level, taken together the results show that there was a small change in overall accuracy between reads 2 and 3 (see Table 4-3), indicating that the overall impact of CADx in terms of a change in binary diagnosis was low. Interestingly, however, the combined data showed a larger increase in mean specificity (with a slight decrease in sensitivity). This is consistent with the CADx tool causing radiologists to change their overall image classification on a few occasions (i.e. moving between scores of 1-2, 3 and 4-5) but that the overall error rate only improved slightly as a result. Thus, the main influence of the machine learning algorithm output on binary decision making appears to be that it encouraged the radiologists to be slightly more cautious in reporting images as abnormal.

The most significant influence of CADx can be seen from the inter-reporter reliability results (see Figure 4-7). Taken across 5 radiologists there is a substantial difference in the ICC value from reads 1 and 2 as compared to read 3 (0.84 and 0.86 vs 0.95). Taking into account the 95% confidence intervals on the ICC scores, there is no overlap between the two different reporting scenarios, suggesting that this difference is significant. This implies

that although CADx did not cause a significant shift in overall binary classification accuracy, it may have had a greater effect in 'pushing' radiologists towards a common image score.

The standalone performance results of the machine learning algorithm were reassuringly high (higher than many of the reporters) and demonstrate that the tool was suitable for use as an assistant to radiologists. The results also show, in a relatively limited sample, that an algorithm trained with data that has an inferior ground truth diagnosis (visual analysis only) can achieve high diagnostic accuracy. This identifies that the selected machine learning model has the potential to be a clinically useful CADx tool.

In this study the delay between reads 1 and 2 was short and it is likely that recall bias would not have been completely eliminated. Furthermore, reading multiple images of the same type in a short space of time is not necessarily reflective of clinical workloads. In Sheffield, for example, there are typically only 2 new (I123)FP-CIT cases per week to report. These factors are likely to increase uncertainty in pilot study results.

The stark differences in performance between individual radiologists are difficult to account for. It may have been that some of the radiologists didn't fully understand the reporting task, hadn't fully appreciated the differences between normal and abnormal appearances or were unsure how to use the software (at first). As suggested by Eadie, training processes used to familiarise reporters with CADx software (100) can be vitally important. It is possible that the training provided was inadequate, and so results may have included learning curve effects, where the radiologist becomes more confident in using the software over time. A longer training period may have helped to reduce these issues.

Qualitative

The radiologist-CAD relationship is complex and cannot be fully reported in terms of diagnostic performance figures alone. Qualitative evaluation of the psychological aspects of computer assisted reporting can provide a much richer dataset, giving insights into software design steps that may improve overall performance. However, such investigations are rarely carried out (99). Thus, although relatively basic, the questionnaires provided to participating radiologists in the pilot study offer a useful, complimentary and novel insight into how CADx affects decision making for (I123)FP-CIT.

The responses to question 1 show that CADx mostly had a small or moderate impact on decision making processes. This is reflected in classification scores that were changed for 15% of patient cases. It should be emphasised however, that the junior radiologists tested in this study may be more open to influence by CADx software than more experienced radiologists.

An important consideration was the extent to which the accuracy of the machine learning algorithm was revealed. By disclosing that the standalone accuracy of the system was approximately 90%, i.e. at the level of a human expert, it is possible that participants may have been more trusting of the algorithm than if no performance data had been provided.

The comments received in relation to question 1 indicate that the tool was of most use and had the biggest impact when its output reflected the original opinion of the radiologist. Generally, if the radiologist and CADx classified the image in the same way, the radiologist was more confident in his/her diagnosis. In addition, comments also suggested that the algorithm output caused some of the radiologists to look again at the image, to scrutinise appearances in light of objective findings from the CADx. This is reassuring as it reflects the intended purpose of the CADx system.

However, one of the comments suggested that the machine learning algorithm was of less help in difficult cases, where the probability was also on the borderline between the two classes. Here, the CADx system appears to have a similar ability to that of the radiologists, providing more equivocal results when image appearances were difficult to classify. Although this may limit the usefulness of CADx (a high-probability, independent check may be more helpful in equivocal cases), it is perhaps unsurprising. If a human struggles to visually classify an image then it is likely that extracted features, based on the same data, will also not provide a clear classification in all cases.

The responses to question two reveal that in general the CADx software was trusted by the radiologists, which is vital if such a system is to achieve clinical acceptance. Interestingly, the submitted comments reveal that trust was intrinsically linked to the radiologists' experiences. If the tool agreed with their initial image read then trust increased. This might indicate that an assisted reporting tool should largely agree with reporters in order for it to be accepted in clinic. However, this potentially creates a problem. If a CAD tool always agrees with a radiologists' first impressions then its usefulness as a means of increasing performance is decreased. There needs to be some discrepancy in a minority of cases in

order to influence reporters to move away from an incorrect diagnosis. Conversely, if the disagreements between a radiologist and the CADx tool are too frequent, even if the radiologist is wrong, then trust in the tool may be decreased and the reporter may simply ignore it.

The responses to question 3 justify the choice of a probability value as an output from the classifier, as opposed to a binary discrimination. However, question 4 reveals that the current system design is less than perfect. The current classification approach takes the whole image as an input and so is unable to provide local information on the possible sites of any abnormalities. This is a potential disadvantage as compared to semi-quantification, which localises uptake quantities to striatal sub-regions (in some cases). The desire for a localisation mechanism also implies that reporters may benefit from gaining a better understanding as to which aspects of an image's appearance caused the CADx system to classify a patient in one way or another. However, it should again be emphasised that the radiologists recruited for the study were relatively inexperienced. For consultant radiologists, localisation information (and further details on why a classification decision was made) may be less of an asset.

Responses to the final question suggest that radiologists found the CADx tool to be a potentially useful training aid. Given that the standalone accuracy of the classification tool was high this is perhaps unsurprising. Allowing a reporter to analyse an image, form their own opinion, then compare to an independent 'expert', the overall experience is similar to that of being trained by a more experienced colleague. The major advantage of this form of training is that once setup, the costs of the software would be negligible. Utilising the machine learning tool in this way deviates from the original intended purpose for which it was developed, but this added application could offer another route for demonstrating effectiveness in the clinic.

4.3.4 Conclusion

The pilot study provided a useful insight into the effects of a machine learning algorithm on radiologist performance, when utilised as a CADx system. Although the radiologists recruited had no previous experience of reporting (I123)FP-CIT images their unaided reporting accuracy (following training) was in most cases at a similar level to that typically reported for more experienced reporters. For the limited set of 30 images used in the study there was a small change in overall performance, in terms of accuracy, after the introduction of the

machine learning algorithm to the reporting process. However, there was a substantial improvement in inter-reporter reliability. The standalone accuracy of the algorithm was found to be high, justifying its use as a CADx tool.

The results of the questionnaire demonstrate that the tool was well trusted and had a small/moderate impact on reporting decisions. The questionnaire also revealed that a probabilistic CADx output was preferred to a binary one but that localisation information would have made the assistive reporting tool more useful. The received opinions suggest that the machine learning tool may have an important role to play in future training of inexperienced radiologists, which could offer a new route to clinical acceptance.

4.3.5 Implications

Although limited, the pilot study was a useful precursor to a larger scale clinical evaluation. Results justify inclusion of a much larger number of clinical cases to measure a significant change in reporting accuracy after introduction of CADx, assuming that such a difference exists. Biasing the test patients in the main study towards more difficult cases may help to further expose the performance benefits available from CADx, particularly for more experienced radiologists, for whom reporting opinions are unlikely to change if visual analysis shows classical normal or abnormal appearances.

The effect on inter-reporter reliability and reporting confidence of the CADx system were encouraging, but results for more experienced radiologists may be less dramatic (i.e. they may trust their own judgement more and be less swayed by algorithm output). This again dictates that a larger number of cases should be included in a clinical trial to measure what may be a relatively small effect. It is difficult to predict the minimum number of cases required and so utilising as much of the available data as possible is likely to be the best way of ensuring that clinical impact is measured adequately.

The pilot study results also suggest that a localisation mechanism would be a useful addition. However, this would require a complete algorithm redesign. It was not the intended focus of this work to design a completely new tool and so this option was rejected for the larger trial. In general, the results from the questionnaire were insightful and added useful contextual information on the influence of CADx. The larger, main study described below will include an expanded list of questions for discussion, to explore these issues more deeply.

4.4 Main study – assessment of experienced reporters

The aim of the main clinical evaluation study was to generate in-depth, reliable evidence of the impact of a CADx tool in a clinical reporting scenario with experienced reporters, from a qualitative and quantitative perspective. This data adds to the evidence gathered so far on the effectiveness of CADx for (I123)FP-CIT reporting.

4.4.1 Introduction

The data from the local hospital that was available for clinical studies was limited. All of subset B had already been used for training the machine learning algorithm, and so could not be used for clinical testing. Subset A contains only 55 cases in total, which is unlikely to be sufficient for measuring the impact on reporting decisions with high confidence. Therefore, for the main clinical study the PPMI data was used in addition to the local data. As previously discussed the PPMI data is not necessarily a good representation of the images seen in clinic in the UK. However, this added data does enable a more comprehensive assessment of the possible benefits of CADx for (I123)FP-CIT reporting. In addition, the inclusion of PPMI data enables assessment of reporter performance with unfamiliar images (reflective of the situation when radiologists move to a new hospital, for example).

Given the very different acquisition conditions and processing parameters associated with the PPMI data, a separate classification algorithm was trained. Utilising the algorithm trained on local data, for classification of PPMI data, is likely to have led to reduced performance, giving a false impression of the potential for CADx.

The pilot study showed high performance figures for many of the junior radiologists when reporting local data unaided and there were relatively few cases where the CADx caused a change of opinion. Given that the PPMI data is associated with strict inclusion and exclusion criteria (for example PD patients without abnormal SPECT appearances are excluded) it is likely that unaided visual reporting performance and confidence may be even higher for this cohort, and the potential for CADx to influence reporters' decisions even lower. Thus, test data were skewed towards more difficult cases in order that the number of cases where reporters' visual impression was uncertain was maximised, increasing the potential for CADx system to influence diagnostic performance

To achieve this, the PPMI set was split in half, maintaining the same proportion of normal and abnormals in each sub-group. The first half, containing 328 images, was used for algorithm training. For the second half of the data SBR figures were examined to find the 40 healthy controls with the lowest putaminal uptake ratios and the 60 PD cases with the highest uptake ratios. This collection of 100 images, skewed towards more equivocal data (according to semi-quantification results), was used in the clinical evaluation. The remaining data, which was neither used for algorithm training nor for testing with radiologists, was excluded.

Method

Two radiologists and one clinical scientist were recruited for the study. All three had at least 5 years of experience of reporting on (I123)FP-CIT image appearances as part of a routine clinical service. By including reporters from two different specialisms it was possible to gain a wider perspective on potential differences in opinion as to the value of CADx in a clinical scenario. The study procedure was similar to that adopted for the pilot study, with three separate reads conducted for each image, two without CADx support and one with. Measurements of standalone performance on the test data were also conducted. However, there was no reporter training phase other than a brief demonstration of the software. Furthermore, the time gap between the first and second read was much longer in order to reduce uncertainties associated with recall bias (a minimum of 4 months). In contrast to the pilot study, each reporter worked through the test cases at their convenience, on their own (using standard clinical reporting hardware). The main differences between pilot and main studies are summarised in Table 4-5.

Calculation of performance metrics and inter/intra reporter variability was the same as for the pilot study, considering the PPMI and local data separately. The qualitative aspects of the study were expanded, with additional questions added to the questionnaire. In contrast to the pilot study where time was very limited, the volunteers were individually guided through the questionnaire after the 3 image reads had been completed. This enabled a greater exploration of any salient points that were raised. Due to the reporters' previous experience of standard reporting using semi-quantification, there was also scope to explore the perceived benefits (or disadvantages) of using CADx instead through specific questions.

	Pilot study	Main study
Time delay between reads 1 and 2	1 hour	4 months+
Total images reported	30	155
Number of reporters	7	3
Experience level of reporters	None	5 years+
Reporting environment	Single room, shared with all reporters, standard PCs	Separate, standard clinical workstation for each reporter
Reporting conditions	Time pressured, all cases for all 3 reads completed within one joint lab session	Reporters worked through cases at their convenience

Table 4-5 Summary of the differences in methodology between pilot and main CADx studies

4.4.2 Results

Quantitative

A different naming system is used in the following results to distinguish between each reporter (referred to as Rad1, Rad2 and CS1). This was done partly to emphasise the differences as compared to data measured from junior radiologists in the pilot study, and partly because the reporters now represented a mixed group, containing two radiologists and one clinical scientist. The delay between reads 1 and 2 ranged from 137 days to 356 days across the two datasets and 3 reporters, well in excess of 4 months.

Figure 4-8, Figure 4-9 and Figure 4-10 summarise performance metrics for each reporter for each for the 3 reads, with local data and PPMI data. These graphs also display the standalone performance of the CADx system, where appropriate.

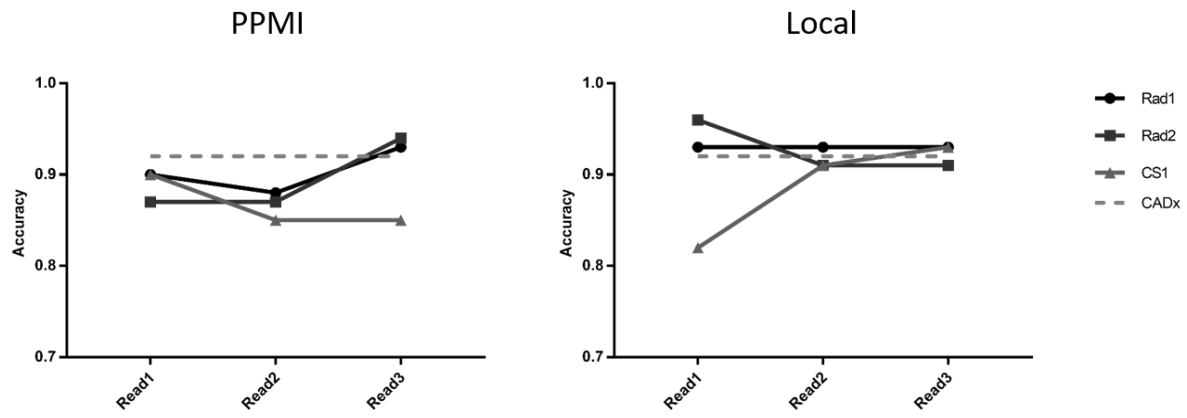


Figure 4-8 Diagnostic accuracy figures for the 3 image reads, for PPMI data (left) and local data (right). Standalone CADx performance is also shown, for comparison. Adapted from (2)

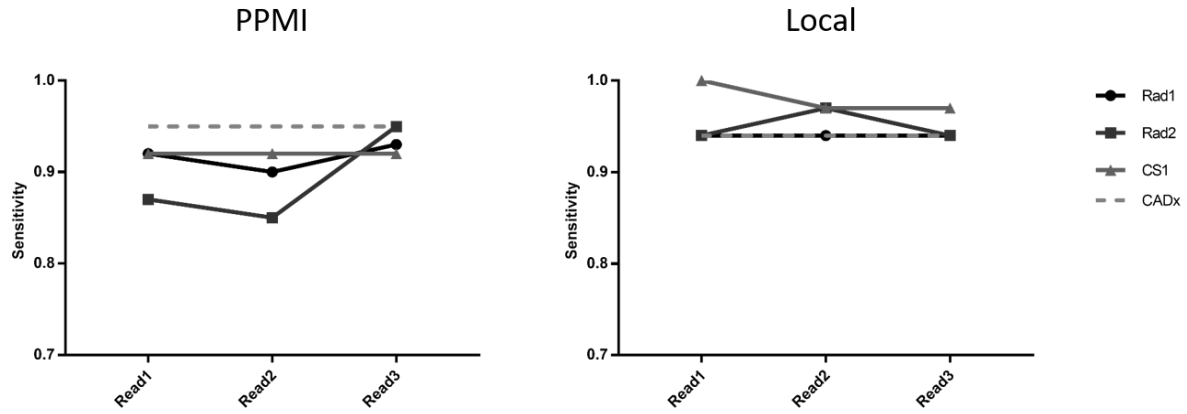


Figure 4-9 Sensitivity figures for the 3 image reads, for PPMI data (left) and local data (right). Standalone CADx performance is also shown, for comparison. Adapted from (2)

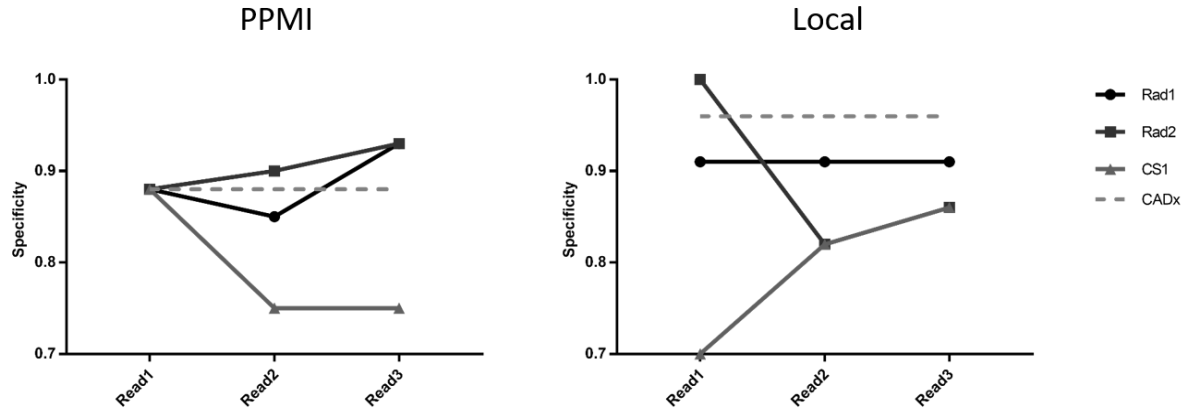


Figure 4-10 Specificity figures for the 3 image reads, for PPMI data (left) and local data (right). Standalone CADx performance is also shown, for comparison, Adapted from (2)

As a result of being exposed to the CAD software output the reporting score was changed in approximately 13% of cases for the local data, and in approximately 17% of cases for the PPMI data (similar to the 15% change rate seen in the pilot study). Intra and inter reporter reliability results are shown in Table 4-6 and Figure 4-11. Due to apparent differences in the ways that the radiologists and the clinical scientist responded to the CADx, separate inter-reporter reliability figures are displayed from all three reporters together and considering just the radiologists alone.

	Intra-reporter reliability					
	PPMI			**Local**		
Reporter	**ICC**	**95% CI (lower)**	**95% CI (upper)**	**ICC**	**95% CI (lower)**	**95% CI (upper)**
Rad1	0.87	0.82	0.91	0.89	0.82	0.93
Rad2	0.95	0.92	0.96	0.93	0.88	0.96
CS1	0.91	0.87	0.94	0.88	0.80	0.93

Table 4-6 Intra-reporter reliability (ICC) results for all reporters, for PPMI data and local data. Adapted from (2)

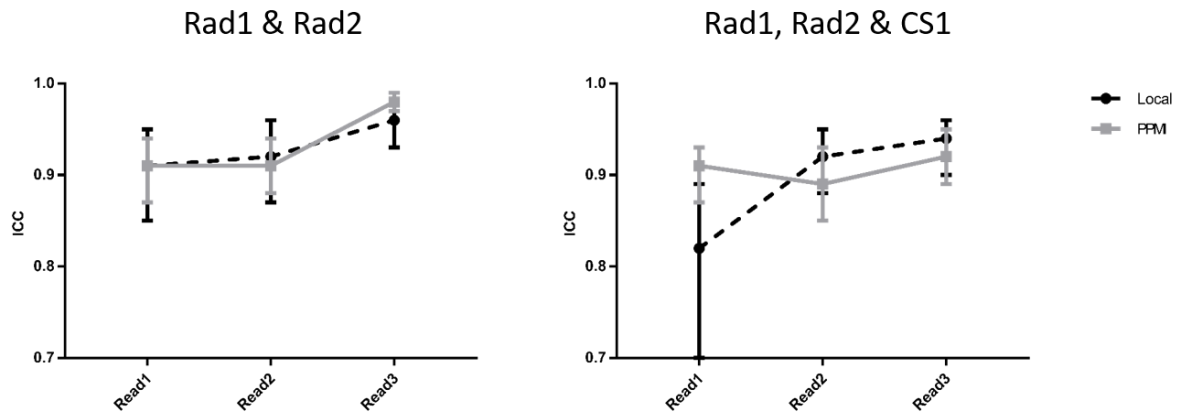


Figure 4-11 Inter-reporter reliability (ICC) results for each of the 3 image reads for PPMI data and local data. The graph on the left is derived from radiologist data only (Rad1 and Rad2), the graph on the right is from all reporters. Whiskers represent 95% confidence intervals. Adapted from (2)

Qualitative

As previously, the following results summarise responses received to specific questions. In addition, any salient comments that reporters made are highlighted. Given the differences shown in reporting trends between the radiologists and clinical scientist, the responses and comments are assigned to the particular reporter.

Question 1: In general, how well did your reporting decisions correlate with the CAD output?

	Responses				
Reporter	**Not at all**	**A little**	**Moderately**	**A lot**	**Comment**
Rad1				✓	
Rad2				✓	"Only one or two cases where I disagreed"
CS1				✓	
Total	*0*	*0*	*0*	*3*	

Table 4-7 Reporter responses to question 1

Question 2: In general, how would you rate the impact of the CAD algorithm on your reporting decisions?

	Responses				
Reporter	**No impact**	**Small impact**	**Moderate impact**	**Substantial impact**	**Comment**
Rad1		✓			"Depends on the case. I liked the fact that in most cases it confirmed my opinion as we only single read these scans. It generally added confidence"
Rad2			✓		"On the few occasions where my opinion differed significantly from the CAD, it had a substantial impact"

CS1		✓			
Total	*0*	*2*	*1*	*0*	

Table 4-8 Report responses to question 2

Question 3: To what extent did you trust the CAD algorithm results?

	Responses				
Reporter	**Not at all**	**A little**	**Moderately**	**A lot**	**Comment**
Rad1				✓	"My performance changed depended on the preceding cases [if several CAD scores agreed with my opinion in a row I grew more confident and trusting in the algorithm then gave CAD more weight]"
Rad2				✓	"Trust increased as I gained more experience"
CS1			✓		"After my experiences so far I'd give it 7 out of 10 in terms of how much I trust it"
Total	*0*	*0*	*1*	*2*	

Table 4-9 Reporter responses to question 3

Question 4: Would you prefer a binary CAD output as opposed to a probability value?

	Responses				
Reporter	**Yes**	**No**	**Other**	**Not sure**	**Comment**
Rad1			✓		"I would like to see both. I was swayed by the colour of the probability value as well as the actual number"
Rad2		✓			"Very much liked the scale"
CS1		✓			
Total	*0*	*2*	*1*	*0*	

Table 4-10 Reporter responses to question 4

Question 5: As compared to semi-quantification, how does CAD compare in terms of what it offers you as a reporting assistant?

	Responses
Reporter	**Comment**
Rad1	"I prefer CAD to semi-quantification. Semi-quantification presents too many numbers and can be confusing"
Rad2	**"**Bit of extra information. Difficult to know exactly how they differ"
CS1	"It would be nice to know why an image was classified as abnormal (which semi-quantification gives you)"

Table 4-11 Reporter responses to question 5

Question 6: Would you prefer to have CAD for assistive DaTSCAN reporting or semi-quantification? Or Both?

	Responses				
Reporter	**CAD**	**Semi-quantification**	**Both**	**Not sure**	**Comment**
Rad1			✓		"CAD as a first line then semi-quantification if I needed it"
Rad2			✓		
CS1			✓		"I wouldn't want anything too complex though"
Total	*0*	*0*	*3*	*0*	

Table 4-12 Reporter responses to question 6

Question 7: Would it benefit you if the CAD system also provided information on how it came to its decision (e.g. reduced putamen uptake, high background uptake etc.)

	Responses				
Reporter	**No**	**Yes (small**	**Yes (moderate**	**Yes (substantial**	**Comment**

		benefit)	benefit)	benefit)	
Rad1		✓			"I had a good idea why the algorithm decided what it did"
Rad2		✓			"I could usually see why it thought what it did"
CS1			✓		"It needs to give some understanding of how it reached its decision (if possible)"
Total	*0*	*2*	*1*	*0*	

Table 4-13 Reporter responses to question 7

Question 8: To what extent would the CAD system be a useful training tool to improve DaTSCAN reporting performance for inexperienced clinicians?

	Responses					
Reporter	**No benefit**	**Small benefit**	**Moderate benefit**	**Substantial benefit**	**Not sure**	**Comment**
Rad1		✓				
Rad2		✓				
CS1			✓			"Already have training sets for people to work through"
Total	*0*	*2*	*1*	*0*		

Table 4-14 Reporter responses to question 8

4.4.3 Discussion

As for the pilot study, the discussion is presented in separate sections, considering the quantitative data first, followed by the collected qualitative data.

Quantitative

In common with the pilot study, standalone diagnostic accuracy for both machine learning classification algorithms was in excess of 90%, which again shows that the chosen classification model performs at least as well as humans using visual analysis alone.

Analysis of Figure 4-8, Figure 4-9 and Figure 4-10 indicates that there was relatively high variation in per-reporter performance metrics between the first and second reads in some cases, for both sets of data. This identifies the degree of intra-reporter variability when analysing images visually, even for experienced reporters. This is backed to some degree by intra-reporter reliability (ICC) figures, which although generally higher than those seen for junior radiologists, were less than 0.9 for Rad1 and CS1 for the local data. Rad2 appeared more consistent. These findings were unexpected and may be exacerbated by the relatively long time gap between image reads, such that reporters' impressions of what constitutes a normal or abnormal image may have drifted. The variability seen may be an exaggeration of what is normally expected in the local clinical service, where a group reporting scenario is used routinely, with semi-quantitative results and patient notes available. This may help to ameliorate the effects of individuals' changing visual impression. Nonetheless, results do provide a reminder that human perception and understanding of medical images is not a constant. This again adds weight to arguments on the need for assistive software (or algorithms which take diagnostic decisions independently).

Comparing reads 2 and 3 (i.e. directly before and after the CADx was shown to the reporter) it does appear that there was some uplift in performance for the PPMI data, where every performance metric either stayed the same or increased for all reporters. Conversely, for the local data there was no clear change in performance as a result of the introduction of CADx. For the PPMI data it is interesting to note the contrasting results between the clinical scientist (CS1) and the two radiologists (Rad1 and Rad2). For both radiologists there was a substantial increase in accuracy after viewing the CADx results, with similar increases in specificity and sensitivity. However, for the clinical scientist there was no change in any of these figures. Indeed, analysis of the individual scan results indicates that CS1 only changed his / her diagnostic confidence score in 7% of cases for the PPMI data, as compared to 21% and 22% for Rad1 and Rad2 respectively. A similar but less marked trend was seen in the local data, where CS1 changed his score in 6% of cases as compared to 9% and 23% for Rad1 and Rad2. It appears that the radiologists relied more heavily on the CADx decision than the clinical scientist, particularly for the unfamiliar PPMI data.

The fact that the PPMI test data was skewed towards more borderline cases may have increased reliance on the CADx for the radiologists, which may have emphasised its benefits to a greater extent (as was intended). The performance gains seen for the radiologists in this half of the study are also likely to be related to the fact that the standalone performance of the CADx tool was generally higher than that of the volunteers during reads 1 and 2, i.e. by aligning their opinions more with the CADx tool, the radiologists' performance was pulled closer to that of the trained algorithm. This contrasts to trends seen with the local data where baseline reporter performance was generally higher, and standalone accuracy of the CADx tool was essentially the same as that of the reporters.

The inter-reporter reliability results echo previous findings in that responses from different reporters became more consistent after exposure to the machine learning output. Figure 4-11 demonstrates that for the radiologists at least, there was a noticeable increase in the intraclass correlation coefficient between reads 2 and 3. For the PPMI data the 95% confidence interval bounds suggest that this increase in reliability (and hence reduction in variability) was statistically significant. These trends are reinforced by percentage agreement figures: for the PPMI data the radiologists had complete agreement in confidence scores in 77 and 74% of cases for reads 1 and 2, rising to 87% agreement after introduction of CADx. However, as for the performance figures related to accuracy, sensitivity and specificity, these trends are less clear when the clinical scientist was included in the analysis. The apparent differences in performance between the two staff groups are explored further in the following qualitative analysis section.

Given the increased consistency between reporters during read 3 it is likely that the introduction of a CADx system would also have benefits in terms of reduced intra-reporter variability. However, estimation of such an effect would benefit from the reporting exercise with CADx assistance being repeated.

The increase in accuracy for the radiologists, as a result of exposure to CADx output, when scoring unfamiliar (PPMI) data, is perhaps clearer than the mixed performance trends seen in the pilot study. This may be due to increased noise in the pilot study data due to variable understanding of the task presented, and the much increased time pressure of the pilot study setup. However, both the pilot study and main study showed an overall increase in inter-reporter reliability after introduction of the CADx (though there were differences in the magnitude of change). The proportion of cases where reporters / radiologists changed their

scores was also similar between the studies. Apparently the assistive reporting tool can be as influential on very experienced reporters as those who are beginners.

It is difficult to directly compare these findings to those of wider studies evaluating the effects of semi-quantification on radiologists' performance, mainly due to differences in data used and methodology. However, the broad findings of this work – that CADx can improve accuracy if adopted by reporters with limited experience of the data, and that inter-reporter reliability may also improve as a result – are consistent with much of the previous work related to semi-quantification, where increased confidence and consistency were found to be the main benefits (see section 1.2.1).

This study was conducted under more realistic conditions than the pilot investigation. As well as the much longer time gap between reads (to reduce recall bias), the reporters used the same workstations as they would normally view clinical images on. This focus on more realistic testing conditions contrasts with much of the machine learning literature, where clinical validation is often not performed or is insufficient (100,103–106). Thus, findings provide a useful addition to current knowledge on the clinical potential for CADx.

However, there remains some limitations in the testing scenario, as listed below:

- Patients' clinical history was not available to reporters as it would have been in clinic. If such information were available the impact size of CADx may have been reduced.
- Although it was intended that patient age would be visible to reporters for all reads, it was only displayed to reporters on read 2 and 3. This may have caused additional intra-reporter variability.
- The reference diagnoses of all the images studied was binary (i.e. either with or without disease). However, the 5 point confidence scale used by reporters associated a score of 3 with an equivocal classification, giving users a choice of 3 different classifications. This mismatch dictated that accuracy, sensitivity and specificity were all negatively affected whenever a reporter submitted an equivocal confidence score. Although a score of 3 was selected in less than 3% of cases for the main study, diagnostic performance figures may have provided a more pessimistic outcome than might have been the case if only two classifications were available for users to select.

- Participating reporters understood that they were taking part in a research study and their decisions would not affect patient care. This may have caused them to be less cautious than would normally be the case (107).
- Re-reporting of the same images with and without the assistance of an automated classification system is an artificial process necessary for clinical evaluation, but which can lead to changes in reporter performance (108).

Although it is important to be aware of such uncertainties, these factors do not detract from the main, positive findings of the quantitative analysis, namely: standalone performance of the CADx tool was at least as high as that of experienced reporters, CADx improved performance of the reporters for the PPMI data and that CADx increased consistency between reporters across both datasets.

Qualitative

The qualitative findings indicate that the CADx tool generally agreed well with the reporters' classification decisions, with only a very limited number of disagreements. This reflects the quantitative findings which showed similar standalone performance between the CADx tool and reporters using visual analysis. Many of the responses to the questions are very similar to those recorded in the pilot study. For example, the CADx tool was generally felt to have a small impact on reporting decisions and was found to increase the confidence of reporters when both the reporter and the classification tool came to the same conclusion. Furthermore, as with the pilot study, the tool was generally well trusted (although the level of trust changed across the duration of the study) and in most cases the reporters preferred the probability output to a purely binary image score.

These observations provide evidence that the more experienced, established reporters viewed CADx similarly to the junior radiologists. However, there were some notable differences. For example, Rad1 and Rad2 felt that having an idea of why the CADx tool came to a particular decision was relatively unimportant. This contrasts with the junior radiologists who felt that having a localisation mechanism to identify where in an image the likely abnormality was, would be useful. These findings are perhaps unsurprising given that more experienced reporters will more easily recognise patterns of normal and abnormal uptake, requiring less prompting from the computer.

The questions asked of the more experienced reporters also offered additional insight. Of particular interest was the contrast between semi-quantification and CADx. Interestingly, all three reporters felt that having access to both CADx and semi-quantification was preferable to having access to one or other. This implies that the functionality of each was felt to be different but complementary. Perhaps a greater impact on reporting performance would be measured by performing a clinical study using a combined software algorithm giving SBRs and overall probabilities.

The qualitative results provide additional evidence that the approach and opinions of the two radiologists were close to each other, but differed with that of the clinical scientist. In general, the clinical scientist was less positive about the CADx tool, and more cautious about relying upon it. For example, CS1 gave a lower relative score for his / her level of trust in the CADx tool than the two radiologists. This is also reflected in the fact that CS1 selected the equivocal image score more times than Rad1 and Rad2 when using CADx. CS1 also felt it important that the CADx tool gave a reason for its classification decision, which contrasts with the radiologists who felt that they did not need this extra information. Indeed, Rad1 saw the very simple colour-coded CADx output as an advantage. Furthermore, the radiologists were generally positive about using the CADx tool as a training resource for inexperienced radiologists, whereas CS1 felt that existing data and methods were sufficient.

These differences in how the CADx tool was appreciated could, at least in part, be attributed to differences in the professional background of the two staff groups. Clinical scientists are taught to understand the technology that is associated with their area of expertise. Indeed the local semi-quantification tool used clinically in Sheffield was originally developed and tested by clinical scientists. Without providing the volunteers with information on how the CADx tool worked, it was effectively presented as a ‘black box’ with little scope for gaining intuition as to why certain classification decisions were reached. For radiologists there is much less focus on understanding imaging technology in their training and a greater emphasis on interpreting images using provided software. Thus, they may have been more at ease accepting the output from the CADx than the clinical scientist, who normally interprets images using technology which they understand in detail. Arguably, for centres where clinical scientists carry out reporting, more information on the technology behind the CADx tool may need to be provided in order to persuade them of its merits. Furthermore, it may help the case for adoption if the CADx tool could be adapted to provide some indication as to why a decision has been made.

The fact that radiologists were generally trusting of the CADx algorithm, and that their level of trust increased with experience of using the tool (in both this study and the pilot study), does perhaps present an added risk. There is a danger that individuals could come to rely more and more on the CADx to make the diagnostic decision for them, relying less on their own judgement. This is particularly relevant in the current healthcare environment where radiologists' workloads are becoming ever larger. It is feasible that in cases where the CADx tool made an incorrect classification, perhaps due to unusual image appearances not seen in training data, this could have undue influence on the final report, that might otherwise have given a different conclusion if the radiologist was working alone. Therefore, it is important that reporters are trained to understand that the technology is not always right. Striking a balance in reporters between scepticism (as displayed by the clinical scientist) and being open to influence is likely to be a challenge for this CADx application and for others.

Although the questions asked during the main study expanded upon those presented during the pilot study, there are perhaps still additional questions that could have been posed to gain additional insight. For example, given the similar (or higher) performance of the CADx tool as compared to the reporters there is potentially a role for such automated classification systems in auditing the reports produced in the clinical department. This is likely to become increasingly important as imaging centres (and healthcare professionals) seek ongoing accreditation. Exploring opinions on this aspect of the CADx software may have yielded useful insights.

4.4.4 Conclusion

Overall there were many similarities between the pilot study and main study. Standalone accuracy of the automated classification tool was at a similarly high level in both cases, and exposure to its output caused a similar proportion of changes in reporting decisions. Both sets of reporters also had a similarly high level of trust in the algorithm. The unaided, visual diagnostic performance of all the experienced reporters was more variable over time than was expected, which suggests that CADx could potentially have a role to play in reducing intra-reporter variability as well as inter-reporter variability.

Quantitative results demonstrated positive benefits of CADx in terms of increased accuracy for the two experienced radiologists, when viewing (unfamiliar) PPMI data. In addition, the introduction of CADx appeared to increase consistency between the two radiologists, for both the PPMI and local data. However, the clinical scientist was less affected by the CADx

tool, with less change in reporting performance between reads 2 and 3, for both sets of patient images. The more cautious approach of the clinical scientist is apparent in answers to the questionnaire, which suggested a lower level of trust than for the radiologists and a greater need to understand the mechanism behind the machine learning algorithm's output probability. Questionnaire results also indicated that clinical reporters would prefer to have access to both CADx and semi-quantification in clinic.

5 Dilemmas of clinical application

The clinical studies described in the previous chapter provide valuable insight into (I123)FP-CIT reporting and how a particular CADx tool may impact upon performance. These results, combined with previous tests, have addressed the original aims of this thesis.

The original research question for this work was: How effective is a CADx tool, based on established machine learning algorithms, for assisted (I123)FP-CIT image reporting? Effectiveness was defined in terms of independent classification accuracy and in terms of the impact upon human reporter accuracy, sensitivity, specificity and inter-reporter reliability. Overall, CADx was found to be highly effective in that it increased consistency between reporters and increased their diagnostic performance (particularly when viewing unfamiliar data). Furthermore, the standalone accuracy of machine learning tools was found to be in excess of semi-quantification tools, which are the current standard for clinical assistive software in (I123)FP-CIT imaging.

The work conducted so far represents a step change in comparison to previous research on machine learning for (I123)FP-CIT in that algorithms here have been considered in the clinical context. The direct comparison with competing clinical technologies (semi-quantification) and testing with reporters as part of a CADx workflow are novel aspects not yet investigated by other researchers. Indeed, given the high standalone performance of the developed algorithms, a case could be made for using the classification tools independently in clinic, as part of a different reporting paradigm. For example, the classification tool could perhaps be used to screen out images with high chance of being normal from the reporting list. Alternatively, the tool could be used as a training device, allowing junior radiologists to compare or audit their reporting decisions against software which performs at a similar level to that of an experienced reporter. Such an approach could reduce the supervisory burden on consultant radiologists (who may be difficult to access for junior staff, particularly in small hospitals).

However, although the results presented are undoubtedly persuasive, and add weight to the case for the routine adoption of machine learning tools for (I123)FP-CIT, the approach adopted so far has arguably been naïve. Firstly, clinical reporting has been considered as a single, isolated classification task involving binary classification of an image. In reality, reporting is a more complicated mental process, which takes into account multiple other

factors such as results from other tests and the patient's clinical history. Extrapolating findings to the clinic is therefore associated with a degree of uncertainty. Furthermore, there are many additional barriers that remain in the quest for widespread clinical adoption, both in relation to this application and for other machine learning classification software. Most starkly, these are: regulations, economics, heterogeneity of the clinical environment, data ownership and change management

These are considered in the list below, (extracts of which are also presented in peer reviewed journal articles (4,5)):

1) **Regulations**. In Europe, when software which is designed to have an impact on patient diagnosis or treatment is released, it is considered to be a medical device and the manufacturer must adhere to the Medical Device Directive (Medical Device Regulations from May 2020). Whatever the classification under the regulations, there are a minimum series of requirements that need to be met. Under the updated regulations, requirements related to risk management and quality management systems are prominent. Products need to be designed in such a way that patient or user safety is not compromised and that testing is carried out to ensure that the product performs as intended. This could require a substantial re-writing or repackaging of the original software code, as well as clinical trials. Importantly, ongoing surveillance is required in order to identify and fix any bugs associated with the software. Meeting the regulations requires significant financial resources and specialist knowledge not normally found in an academic environment. Costs are particularly big if the risk classification is high, which may be the case for (I123)FP-CIT classification software developed in this work.
2) **Health economics**. When deciding on whether to invest in particular medical products many healthcare systems around the world utilise economic analysis to inform their decision. In the UK for example, the National Institute for Health and Care Excellence (NICE, www.nice.org.uk) places strong emphasis on such data when generating guidance on medical technologies. Therefore, ensuring that developed products have a strong health economic case is important for promoting adoption. However, even for the most simplistic economic analysis methods, such as cost-consequence analysis, evidence is required to quantify resource implications of the technology, in addition to data on the likely clinical benefits. For CADx in particular, gathering convincing data on the implications for patient care, as compared to standard reporting, is likely to be complex and difficult.

3) **Heterogeneity of the clinical environment**. Machine learning tools cannot be implemented in isolation. Software needs to be integrated within hospital infrastructure such that it is easy to access and use. However, the available information technology resources and associated restrictions may vary considerably between hospitals. Furthermore, software outputs need to be adapted to the particular scanning equipment and imaging protocols used locally, such that associated differences in image appearance do not cause algorithm performance to be degraded (it is well known, for example, that SBRs can vary according to the gamma camera used (50)). Understanding and adapting to this heterogeneity is vital when considering software design. However, gathering such data is again resource intensive. Creating automated classification or CADx tools that are applicable to many different settings is difficult and, ultimately, it may not be possible to accommodate all the requirements of different hospital environments.
4) **Data ownership**. Using retrospective patient data for machine learning research, as in the case of this thesis, requires appropriate governance approvals to be in place, particularly with regards to ethics. These requirements are well established, with systems such as the Integrated Research Application System (IRAS) providing guidance. However, if the developed algorithms presented in this work were ultimately used in a clinical software package that was sold for profit, issues around data ownership and ethics can arise (109). This is another hurdle to development and may dissuade commercial partners from assisting with the push towards clinical adoption.
5) **Change management**. As highlighted by a recent Kings Fund report on adoption of innovation in the NHS, significant investment is usually needed to promote and support implementation of new technology. Simply generating evidence of impact, as in this thesis, is not enough to guarantee uptake (110). This conclusion is reflected in much of the literature on change management, which often highlights people's natural aversion to changing practice. For example, one of the most frequently cited models of change is that created by Kotter (111). This 8 stage model places emphasis on a guiding coalition leading and managing the change process (which in this case would be moving to reporting with CADx assistance, or automated computer screening of images). Without such a leadership team in place it is argued that (successful) changes do not happen. Therefore, if machine learning is to become a truly game-changing technology for (I123)FP-CIT imaging, and for the rest of radiology, support and leadership is likely to be needed from key professionals such as IT specialists, managers, radiographers as well as radiologists to ensure it is

properly integrated in clinic. Not only does this require protected time (and therefore increased financial support) but these individuals have to be persuaded of classification software's merits. Significant investment is therefore also required to promote the technology, to ensure that clinicians actively push the implementation. However, the perceived threat to radiologists' role from machine learning, which is often inflated by articles in the popular press, is likely to make it harder to persuade the clinical community of the need for change.

Issues 1, 2, 3 and 5 would be relevant to any new diagnostic technology being introduced into the health service on a wide scale. However, given the additional reliance on large databases of realistic clinical data, acquired according to appropriate governance procedures (issue 4), translation barriers are perhaps even greater for machine learning technology.

Although other authors have begun to recognise the enormity of the translation challenge, and have identified the inadequacy of validation and verification often performed in machine learning research (99,112,113), the translation issues described above are perhaps more wide ranging than has yet been identified in the literature. Barriers to translation are multi-factorial, going beyond technical and clinical considerations, covering psychology, economics, law and management. This dictates that a multi-disciplinary approach is needed. Clearly, the resources required to push machine learning into the clinic are considerable.

The scale of this translation challenge is demonstrated by the recently reported failure of IBM Watson for Oncology to achieve widespread clinical adoption (114), despite the use of advanced cutting-edge algorithms, and with backing from a major multi-national company.

The machine learning literature for medical imaging is substantial and results appear to be impressive. For instance, machines have already been shown to outperform radiologists in specific disease recognition tasks, such as diagnosis of pneumonia from chest x-rays (115). Even for the relatively niche application of classification in (I123)FP-CIT imaging there are a large number of articles reporting high accuracy results (see section 1.3.2), and this thesis has shown that even basic machine learning tools are highly capable. Thus, algorithm technology appears to be ready for clinical usage. However, current clinical uptake for any radiological application remains vanishingly small. Plainly, therefore, the prevailing approach to machine learning research and development in radiology requires a radical overhaul if the technology is to fulfil its potential in the clinic.

Without tackling any of the listed issues for (I123)FP-CIT imaging, the work presented so far is unlikely to be sufficient for ensuring that developed classification tools are used clinically on a wide scale. As for the vast majority of previous machine learning research in radiology, algorithms would most likely remain in the research arena, or at best, be used only locally (in Sheffield). This would be an unsatisfactory and wasteful outcome, perpetuating the limitations and lack of foresight that are common in machine learning research. A much greater focus on addressing the barriers to translation is required, and it is on this basis that the remainder of the thesis proceeds.

Given the enormity of the translation burden, and the limited remaining resources available in this research work, identifying a strategy for making meaningful progress is challenging. My approach is to consider which of the previously described translation issues needs to be addressed first in the pathway towards routine clinical usage, focusing solely on this area in the following chapters. Targeting one specific area in this way is likely to be more fruitful than dedicating small amounts of effort to each of the different translational hurdles.

Arguably, heterogeneity of the clinical environment is the most pressing consideration for (I123)FP-CIT classification tools. In particular, if the classification tools cannot demonstrate adequate performance outside of the specific equipment and scanning protocols used at Sheffield Teaching Hospitals, it is unlikely that a convincing case can be made for further investment to develop clinical software, and to overcome regulatory, economic and management barriers.

To address this issue, ideally multiple patients would be scanned according to a variety of different scanning conditions, using a variety of different camera equipment. Any changes in classifier performance associated with each combination of parameters in these sensitivity tests could then be measured and assessed. However, such an approach would be prohibitively expensive, logistically difficult and would require ethical approval. Repeatedly scanning realistic patient phantoms is a much more viable option, which is also being actively pursued in other CADx fields such as mammography (116).

With this in mind, the following chapters have three main objectives, to appreciate the influence of heterogeneity in the clinical environment:

- Objective A: Examine and develop phantom technology to provide a toolset that can be adapted to simulate a range of realistic (I123)FP-CIT image appearances.
- Objective B: Use the toolset to demonstrate the influence of heterogeneity by:
    1) Analysing and prioritising the individual imaging parameters that may affect classification software performance.
    2) Performing sensitivity tests to measure the impact of different imaging parameters on developed classification tools

In this way, the following work demonstrates how aspects of the translation gap, in relation to heterogeneity of the clinical environment, could be addressed for (I123)FP-CIT classification software. Unlike the investigations conducted in chapters 3 and 4, which were mostly specific to certain classification algorithms, and certain datasets, much of the following work is dedicated to creating generally applicable methodologies and phantom technology that may also be useful for other researchers. This is important because translation issues are universal to all classification / CADx systems and, given the resources required, meeting these challenges is much more achievable as part of a group endeavour.

6 Beyond reporter performance – new tools for a new diagnostic paradigm

New objectives addressed by this section (in black, bold):
A) Examine and develop phantom technology to provide a toolset that can be adapted to simulate a range of realistic (I123)FP-CIT image appearances.
B-1) Use the toolset to demonstrate the influence of heterogeneity by: Analysing and prioritising the individual imaging parameters that may affect classification software performance
B-2) Use the toolset to demonstrate the influence of heterogeneity by: Performing sensitivity tests to measure the impact of different imaging parameters on developed classification tools

Table 6-1 New objectives addressed in section 6

Having measured the performance of machine learning classification tools and CADx software for (I123)FP-CIT imaging, the results of which were found to be largely positive, the remaining barriers to translation are now considered in order that the developed tools remain on a pathway towards widespread clinical use. This chapter is dedicated to phantom technology, which is a key ingredient required for assessing the impact of different acquisition factors (i.e. for measuring the influence of heterogeneity of the clinical environment).

The following sections first consider the available, commercial phantom technology for (I123)FP-CIT imaging in the context of sensitivity testing for machine learning classification algorithms, the key requirement being that simulated images must be sufficiently similar to that of real patients under the same scanning conditions. Furthermore, the candidate technology must also be adaptable to different patient appearances. Due to a lack of “off-the-shelf” solutions, new phantom technology is proposed and developed.

6.1 Conventional technology – the need for a new type of phantom

Imaging phantoms are typically divided into two categories: physical phantoms and digital phantoms. The advantage of physical phantoms is that real acquisition equipment can be used, incorporating the imaging characteristics seen in clinic. Conversely, for digital phantoms the imaging physics must be approximated in software, adding uncertainty to results. The major advantage of digital phantoms is that multiple tests can be simulated and run very quickly on a computer, dramatically increasing the number of variables that can be investigated.

If phantoms are to be used to generate compelling evidence of the significance of different acquisition factors on classifier or CADx performance, uncertainties in the imaging process need to be minimised, particularly if results are to be used to justify clinical adoption. Therefore, physical phantoms are likely to the most appropriate choice.

For (I123)FP-CIT imaging there is one commercially available phantom that can be purchased. This is the Alderson striatal phantom (http://www.rsdphantoms.com/nm_striatal.htm). In the context of this work it has a number of significant disadvantages. Firstly, it is constructed from fixed plastic cavities. Therefore, there is no possibility of altering the anatomy to reflect a range of patient appearances. This is unlikely to be sufficient for comprehensively assessing the performance of classification tools.

Secondly, the design represents an oversimplification of tracer uptake patterns that are seen in patients. The putamen and caudate on both sides, and the remaining brain, are manufactured as single, separate cavities that must be filled with a single liquid. This dictates that more complex variation in uptake patterns, such as the reduced tracer levels often seen in the brain ventricles, cannot be replicated. The shape of the striatum is also not reflective of most patients. It extends further in the medial-lateral direction than is typically seen. Figure 6-1 demonstrates these contrasting image appearances using example data acquired at Sheffield Teaching Hospitals.

These features severely limit the usefulness of the phantom in evaluating machine learning classification tools, particularly those algorithms which take whole images as inputs, such as the PCA-based algorithms presented in previous chapters. In these cases phantoms are required which are able to reproduce the voxel intensity distribution of real patients.

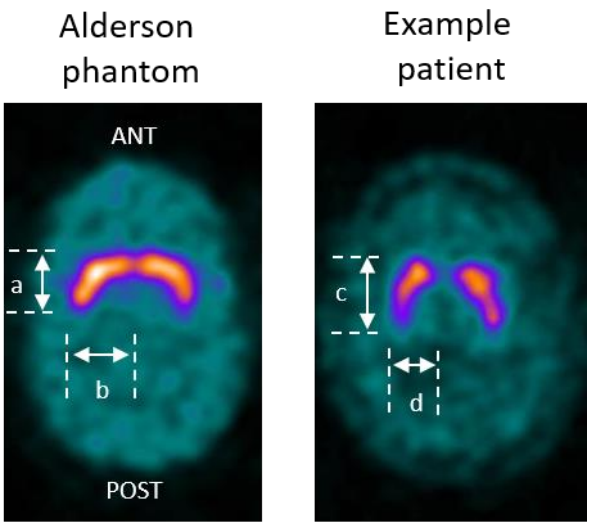


Figure 6-1 Reconstructed, central trans-axial slice from a typical normal patient (right) and from the Alderson phantom (left), demonstrating clear differences in striatal geometry (a/b < c/d). In this case the phantom was filled with an 8 to 1 striatum to reference brain activity concentration ratio. Each slice is scaled to its maximum pixel value. Adapted from (3)

For these reasons a different approach is needed. The next section considers the creation of a new type of (I123)FP-CIT phantom, based on sub-resolution sandwich phantom technology.

6.2 Development of a sub-resolution sandwich phantom for (I123)FP-CIT imaging

A promising physical phantom technique that could be adapted for creating a range of realistic image appearances is sub-resolution sandwich phantoms (SSPs). SSPs are created from inter-leaved layers of attenuating material and paper sheets with radioactive ink patterns on the surface. The ink patterns, reflective of patient uptake appearances, are typically created from an inkjet printer using cartridges containing both standard printer ink and aqueous radioactive solution. The greater the amount of ink printed per unit area, the higher the subsequent radioactive concentration.

SSP technology is highly flexible and has been successfully adapted for a number of applications, including simulation of SPECT brain perfusion scans (117–119) and Positron Emission Tomography (PET) scans (120). With this impressive history in closely related

fields, SSP technology represents a low risk choice in the search for cost-effective technology that can facilitate sensitivity tests of classification algorithms.

Although SSP technology appears capable of fulfilling the main requirements of objective A, the developed solution must also be practical, controllable and repeatable. Therefore, after setting out the phantom design concept, the following sub-sections examine each aspect of the phantom printing process and derive relevant metrics of performance. This data is used as a platform for creating a full head phantom, representative of a patient and useful for evaluating the impact of variations in acquisition on classifier / CADx performance.

The following investigations were conducted in collaboration with colleagues at University Hospitals Bristol. Specifically, the 3D printed head was loaned from Bristol, and the method for generating and warping the anatomical template was adapted from that previously used in brain perfusion studies (119). Extracts of the following investigations contributed to a peer-reviewed publication (3).

6.2.1 Sub-resolution sandwich phantom: design concept

SSPs consist of two separate parts that must be brought together when the final phantom is assembled – thin slabs of attenuating material and radioactive ink printed sheets. However, each of these constituent parts can be produced in several different ways. For the creation of (I123)FP-CIT phantoms, which is a previously unexplored application for SSP, the goal was to devise a production method that would lead to suitably realistic images for assessing classification / CADx systems. This is the motivating factor behind the following design choices.

In relation to the attenuating material, the conventional approach is to use stacked plastic layers cut to a simple shape (119,120). However, with the advent of inexpensive additive manufacturing devices, 3D printing has also started to be used (121). The major advantage of 3D printing using Fused Deposition Modelling (FDM) is that infill density can be adjusted and filament materials changed in order to better reflect the radiation attenuation properties of tissue. Furthermore, the shape of the print can be finely tuned to achieve a geometry that is reflective of a real patient. For these reasons a 3D printing approach was adopted in this case.

In this work a 3D printed head was loaned from Bristol Teaching Hospitals to use as a basis for creating SSPs (see Figure 6-2). The head was constructed from 1.9mm thick slabs, the geometry of which was defined from a patient's segmented, high-resolution CT scan. The slab thickness was several times smaller than SPECT resolution and therefore each non-radioactive slab, which was placed between radiation emitting paper, was indistinguishable on SPECT imaging. Each slab was printed with two different filament materials, conventional Polyactic Acid (PLA), at 85% infill density, and bronze-doped PLA at 100% density. The materials were designed to reflect the radiation attenuation properties of soft tissue and bone respectively. At a photon energy of 159 keV the linear attenuation coefficient of the PLA structure was approximately 0.16 cm^{-1}, for the bronze-doped material it was approximately 0.21 cm^{-1}.

Figure 6-2 Pictures of the 3D printed head loaned from Bristol, fully assembled (left) and with individual slices laid out separately (right)

The paper sheets that define the radiopharmaceutical uptake are printed from an inkjet printer, where the cartridge contains both conventional ink and radioactive solution. The particular printer model selected for all the following investigations was an HP 8100 Officejet pro. This particular printer was selected due to its low cost and most importantly, the ease with which liquids can be injected into its cartridges. The input to the printer is a set of images, representing the patient's uptake profile in separate 2D slices (the anatomical template). These can be created manually or may be derived from real patient scans. The latter option was chosen in this study, in order to maintain clinical realism as far as possible. Each pixel in the anatomical template is converted from a greyscale value to an ink and radioactivity density by the printer. The printer's ink profile curve defines the mapping

between the two. Once printed, each sheet is drilled and then interleaved between the attenuation slabs, using the guide rods to locate the paper in the right position. The proposed overall workflow for creating (I123)FP-CIT phantoms is depicted in Figure 6-3

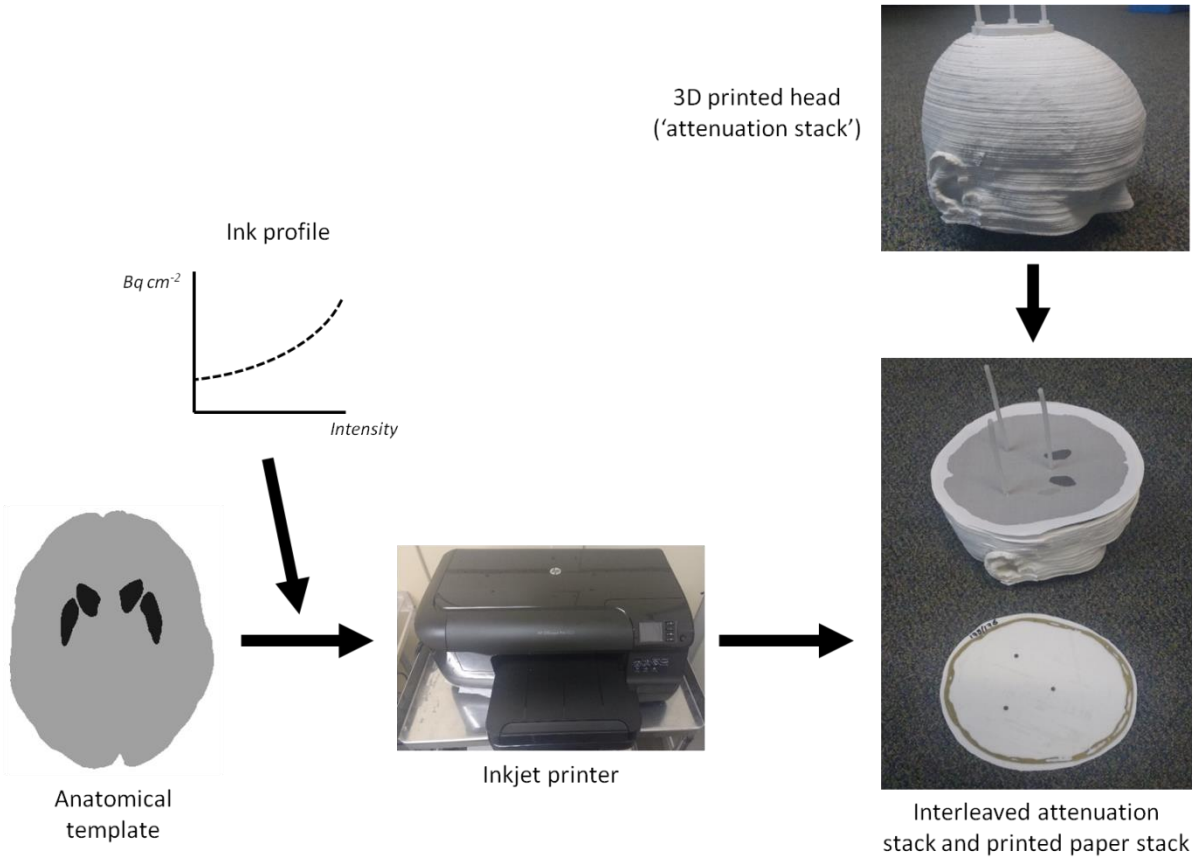


Figure 6-3 Workflow depicting the proposed manufacturing process for creating physical (I123)FP-CIT phantoms. Adapted from (3)

Further details on the anatomical template are set out in following sections in relation to the specific tests performed.

6.2.2 Ink profile curve derivation

The ink profile curve of the particular inkjet printer selected needs to be characterised in order to ensure correct activity distribution for the phantoms. The profile is defined in the printer software and the exact profile shape, which relates screen intensity to printed 'blackness' could theoretically be accessed by sending a request to the manufacturer. However, empirical testing was used in this case in order to derive a curve that was particular to the device in use. The ink (and radioactivity) density deposited by the printer

should be highly repeatable if several phantoms are to be created, with fixed design parameters. With this in mind the following investigation examined several repeated measurements of count density at different greyscales in order to characterise printer performance

Method

12 different greyscale levels were selected on a linear scale from 0.1. to 1.0, where 0.0 represents the colour white and 1.0 black. Intermediate values are different shades of grey. The selected values were: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.85, 0.9, 0.95 and 1.0. The increments in greyscale are smaller towards the top of the range to account for the likely exponential nature of the profile curve (i.e. small changes between darker shades of grey are likely to give bigger changes in ink density than the same magnitude changes at lower greyscale levels).

For this test a black ink cartridge was filled with black ink and $99Tc^{m}$ Pertechnetate in a 1:1 volume ratio. The overall radioactive concentration of the ink-radionuclide mixture was approximately 50MBq / ml. Although the radionuclide used in this case was different to that of (I123)FP-CIT, mainly due to the much reduced cost of $99Tc^{m}$ Pertechnetate, it is not anticipated that this will have any bearing on repeatability results.

Each grey level was printed on to paper in the shape of a small rectangle (2cm x 5cm), using standard office paper (density $80g/m^2$). Each was then cut out, rolled up and placed within a tube on a rack before counting on a PerkinElmer 2480 sample counter (PerkinElmer). To ensure consistent geometric efficiency each piece of printed paper was placed at the bottom of each tube. Counting proceeded for 6 minutes per tube using an energy window centred at 140 keV (±15%), with decay correction turned on. Total counts recorded in each case was greater than 400 kcts. The experiment was repeated 5 times (i.e. 60 measurements were made in total). This data was then used to create a continuous profile, mapping greyscale level to output printed radioactivity concentration.

Results

Figure 6-4 summarises the results of the sample counting experiments

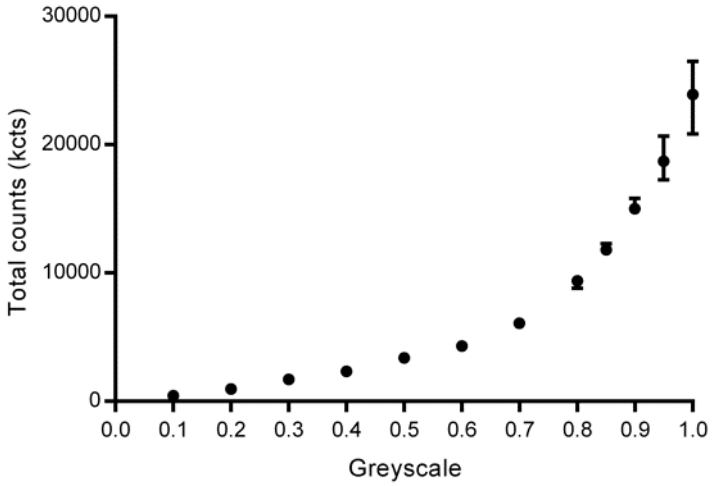


Figure 6-4 Graph of total measured counts (measured over 6 minutes) against input greyscale level. Error bars depict maximum and minimum values across the 5 experiments. Adapted from (3)

Discussion

The results demonstrated that the relationship between greyscale of the input and count density of the output print was non-linear, as expected. However, it also appeared that the repeatability was substantially degraded at higher greyscale levels as compared to lower greyscale levels. For example, the coefficient of variation for a greyscale of 0.95 and 1.0 was 7% and 9% but 3% or lower for greyscales of 0.9 or less. This random error was in most cases greater than that which would be expected from the random nature of radioactive decay (i.e. greater than the standard deviation expected from Poisson statistics). It is likely that the ink printing mechanism introduced additional uncertainties, which were more acute for higher levels of ink deposition. This could perhaps be due to reduced precision associated with printer jets at the highest output rates or saturation effects on the paper. Given these results, it is prudent to restrict all future investigations to a maximum greyscale of 0.9 where repeatability errors are acceptably small for creating highly controlled prints.

6.2.3 Printer uniformity

In addition to tests of ink deposition repeatability it is also important to assess the uniformity of the printed output. If the printer is unable to achieve consistent radioactive ink deposition across a page then fully assembled phantoms may contain artificially raised or lowered areas of radionuclide density as compared to the anatomical design template.

Method

The most convenient method for assessing the uniformity of radioactivity across a page is gamma camera imaging. Although a gamma camera's response is unlikely to be perfectly uniform across the detector face, if resultant uniformity metrics are similar to that which would be expected from a uniform flood source then it can be assumed that any non-uniformities present are insignificant in terms of SPECT imaging.

For this test a constant level of greyscale was printed across a whole A4 page, excluding margins, at greyscale values of 0.5 and 0.9. This size of print is likely to be bigger than the cross-sectional area of most brain slices and so results will represent the likely worst case scenario.

Following printing using the same printer and cartridge setup as for section 6.2.2, each sheet was placed within a thin plastic wallet and then placed flat on the face of a GE Infinia gamma camera (GE Healthcare), with LEHR collimators in place. The camera in question had already passed relevant quality control checks on that day, including a test of uniformity. Imaging was conducted over a long acquisition period (10 hours) due to the low activity level present in each single printed sheet, using an acquisition matrix size of size 256 x 256 (giving a smaller pixel size than that used during clinical SPECT acquisition). A standard 140 keV ± 10% energy window was used. Images were assessed both visually and by measuring summary statistics across the printed area (after reducing the matrix size to 64 x 64 to increase the counts per pixel).

Results

Figure 6-5 shows the raw acquired images from the uniform printed sheets at greyscale levels of 0.5 and 0.9. Table 6-2 shows summary descriptive statistics for the quantitative uniformity results. Here, parametric measures are used as it was assumed that variation in pixel values was largely due to random radioactive decay.

Figure 6-5 Screen captures depicting the raw images acquired from A4 sheets printed with greyscale values of 0.9 (left) and 0.5 (right). Images are colour scaled individually.

Figure 6-5 demonstrates that although images were relatively noisy there were no significant non-uniformities present in either of the printed sheets

	Pixel counts (kcts)				
Greyscale level	**Maximum**	**Minimum**	**Mean**	**Standard deviation**	**Coefficient of variation**
0.9	30.98	25.86	27.99	1.08	0.039
0.5	2.97	2.34	2.71	0.10	0.036

Table 6-2 Summary descriptive statistics for counts detected from A4 sheets printed with uniform greyscale levels

Discussion

The results of the uniformity test showed that the printer utilised in this study provided a relatively consistent ink output across a large area, with no areas of significant non-uniformity seen from visual analysis. Even though quantitative results included the effects of extrinsic camera uniformity errors, the coefficient of variation of pixel values across each sheet was less than 4%. This compares to an expected counting error (from Poisson statistics) of approximately 0.6% per pixel for the image of printed ink at a greyscale value of 0.9, and 2.0% for the image acquired from a sheet printed at a greyscale value of 0.5.

Previous tests of camera uniformity conducted during routine quality assurance investigations, utilising a uniformly filled flood phantom, with a similar count per pixel to that of the 0.5 greyscale image, also produced a coefficient of variation of 4%. This indicates that the printing process does not introduce significant additional uncertainties in uniformity on top of the uncertainty already present due to the imperfect detection system.

6.2.4 Printer input-output comparison

Having demonstrated that the printer produces consistent and controllable ink output across a wide area, one further set of validation tests comparing the input design template and output print was undertaken to confirm close correlation between the two. This validation step mainly investigates the printer's resolution and geometric accuracy, and the reliability of the printer head motors. It is the first test conducted so far that scrutinises the whole printing process.

Method

An anatomical template typical of a (I123)FP-CIT subject was generated such that results were relevant to the intended application. For this, a segmented brain scan was required which depicted the putamen and caudate as well as the remaining brain “background”. An assumed level of radiopharmaceutical uptake also had to be set in each of these different areas. In this case the well-established Montreal Neurological Institute (MNI)152 template was adopted (122). This is an MRI-based anatomical reference, created through the combination of scans from 152 healthy subjects, after non-linear registration to a common co-ordinate system. The automated anatomical atlas (AAL) is a freely available and frequently cited parcellation of the MNI template (123), providing regions of interest which encircle the different brain structures. This was used to denote the boundaries of the left and right striata of the MNI reference as well as the remaining brain (derived by combining all remaining brain regions).

As previously suggested, apparent increased uptake within the putamen on SPECT imaging is in reality a combination of tracer uptake in the putamen and globus pallidus, which cannot be resolved separately due to limitations in camera resolution. In this study the globus pallidus was kept as part of the larger, non-specific background region and thus was assumed to not have a strong affinity for the tracer.

Voxels within the striatal and brain background regions were set to intensity values that reflected uptake ratios from a healthy patient such that the striatum shape would appear prominently on subsequent imaging. Thus, all voxels within the left and right putamen and caudate were set to a greyscale level of 0.9, the maximum achievable for consistent printing using standard paper. Voxel intensities in the remaining background brain area were set to a greyscale value of 0.33, which through linear interpolation of the graph in Figure 6-4 represents a count density ratio of 8 to 1, which is reflective of normal striatal binding ratios typically seen for healthy individuals (40).

5 central slices from the created anatomical template were printed with radioactive ink, again using the same printer setup as for sections 6.2.2 and 6.2.3. Each was then inserted into separate plastic wallets and placed, one by one, on to the surface of a LEHR collimator of a GE Infinia gamma camera (GE Healthcare). Each was imaged for one hour using a 512 x 512 matrix. By imaging with a large matrix size, at the surface of the detector, the extrinsic

resolution of the system was maximised, enabling a more detailed inspection of radioactive ink patterns.

For analysis, the input anatomical template and imaged printed sheets were compared in terms of their relative overlap. This was achieved through rigid registration of each pair of images on a slice by slice basis, followed by a degree of smoothing (using a 5mm Gaussian filter). Segmentation of each gamma camera image was conducted by applying a whole image threshold, set at two different levels, in order to isolate the boundary of the striata and whole brain separately. Threshold levels were determined through measuring the mean count level at the centre of each structure and then dividing by 2. Each of these steps was performed using MIM software v6.6.

The segmented gamma camera images and the original anatomical templates (with regional boundaries defined by the AAL) were compared visually and through the measurement of Dice Similarity Coefficients (DSCs, see Figure 2-9 for a formal definition).

Results

Visual comparison demonstrated that the whole brain region of each gamma camera image closely fitted the whole brain region of the input anatomical template. However, the segmented striatal shape extracted from the gamma camera images deviated from the anatomical template to a greater degree than the whole brain region. This was particularly noticeable in the small area in between the putamen and caudate on both sides, and at the inferior pole of the putamen (see Figure 6-6 for an example).

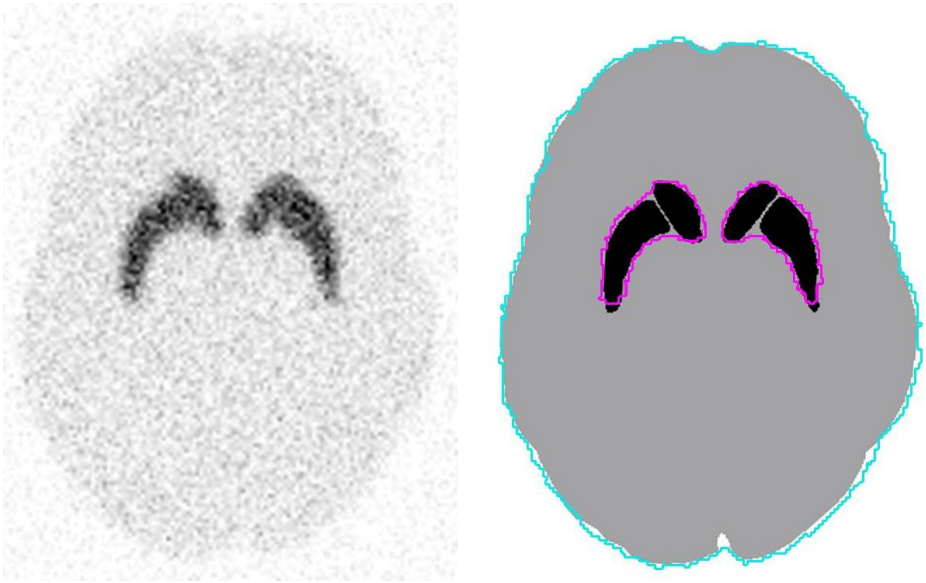


Figure 6-6 Example acquired image (left) alongside corresponding template image (right). Regions of interest generated from segmentation of the acquired data are shown, overlaid on the anatomical template

Table 6-3 shows overlap results from comparison between segmented output images and the regions of interest derived from the AAL, averaged over 5 images. Standard deviation of the results is shown in brackets.

Structure	**Dice Similarity Coefficient**
Whole brain	0.99 (0.00)
Striatum	0.90 (0.01)

Table 6-3 DSC scores of the relative overlap between imaged, segmented brain structures and regions in the anatomical template. Adapted from (3)

Discussion

The gamma camera images of the individual slices were largely as expected, showing consistent count levels through the striatum and the brain background. Despite the relatively high noise level in the images, the outline shape of the striata could be clearly seen. Overlaying of the gamma camera images on to the registered, corresponding anatomical template slices demonstrated that the two sets of data had a very similar shape. This is confirmed by Dice scores, which were high (with small standard deviation figures), particularly for the whole brain region. As might be expected from partial volume effects, the

scores were slightly lower for the striata, where the divide between putamen and caudate could not clearly be visualised on gamma camera acquisitions and where the thin ‘tail’ of the putamen was cut-off following segmentation procedures.

The analysis method chosen to quantify the relative overlap between input and output data was simplistic. The accuracy of this approach was limited to some extent by the high levels of noise. This explains the relatively jagged appearance of the brain region boundaries that were generated (see Figure 6-6). However, despite the limitations in the chosen methodology, results demonstrated that the printer output was a good representation of the input anatomical template in terms of the position of the boundaries of representative brain regions.

6.2.5 Assembly and validation of a full phantom

Having established that the technology is suitable for consistently producing radioactive ink patterns that accurately reflect a design template, and that ink output is predictable, tests of a fully assembled phantom may now be conducted. Resultant images will enable assessment of the complete SSP production and assembly process. In addition, usability aspects of phantom assembly and performance can also be assessed, in particular the time taken to fully build the phantom, the amount of ink required and resultant count rate from a full printed head.

Validation of the fully assembled phantom is based on measurements of striatal binding ratios and linear dimensions of the striata. These values are compared with those of real patients (from subset A). SBRs provide an insight into whether mean uptake within the striatum was set at a realistic level for the patient group being replicated, whilst measurements of its dimensions give an indication as to the suitability of the chosen striatal geometry and shape (the Alderson phantom being deficient in this respect, see section 7.1).

Method

The anatomical template from section 6.2.4, which was largely derived from the automated anatomical labelling atlas, was registered to the geometrical space of the 3D printed head such that the ink patterns would fit within the assembled phantom structure. This process first involved segmentation of the CT scan originally used for manufacture of the 3D printed head. This was carried out in SPM12 software

(http://www.fil.ion.ucl.ac.uk/spm/software/spm12/). The procedure included segmentation and spatial normalisation of the CT images to MNI space, enabling the creation of forward and inverse deformation fields (124). The inverse deformation field was applied to warp the anatomical template from MNI geometric space on to that of the CT scan (see flowchart in Figure 6-7). Following registration, the template was resampled to a 2mm slice thickness such that each printed sheet would be positioned correctly between the 1.9mm thickness attenuation slabs (3).

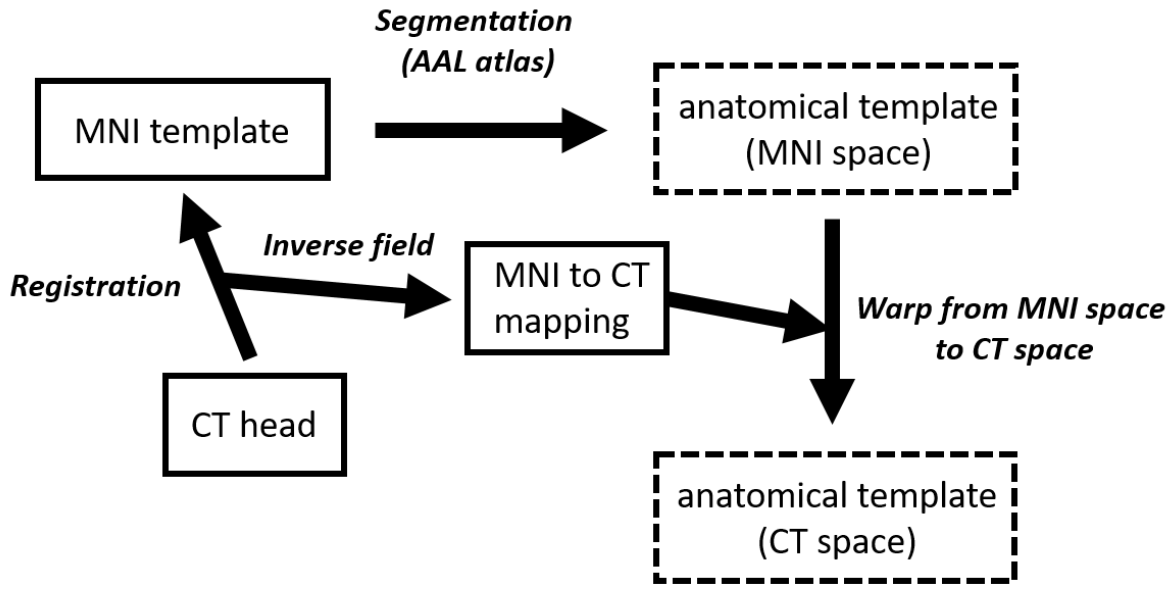


Figure 6-7 Workflow depicting the steps taken to create an anatomical (123)I-FP-CIT template fitted to the same geometry as the 3D printed head (adapted from (3))

In contrast to previous tests of the SSP technology, this investigation required that ^{123}I be used as the printed radionuclide rather than ^{99m}Tc in order that gamma photons detected were representative of those emitted by (I123)FP-CIT. The most inexpensive (and accessible) form of this radionuclide is I123 Iodide, provided to most UK hospitals at a concentration of 37MBq/ml, which is lower than the radioactivity concentration used previously in this work and in published data from other authors. In order to maximise the phantom count rate the ratio of black ink to radioactive solution injected into the ink cartridge should be minimised. However, if too little ink is used then the printer deposition method is likely to be adversely affected due to the associated change in viscosity. Thus, a volume ratio of 1:1 was maintained for this investigation.

Greyscale values in the modified anatomical template were set to give a count density ratio of 8 to 1 in the striatum as compared to the surrounding brain tissue, using the measured ink profile curve. Again, this is reflective of the uptake ratios expected in healthy controls. Each slice was sent to a HP8100 printer and printed using a black ink cartridge containing approximately 16ml of combined ink-radionuclide solution. The full stack of paper from a single template print was placed within a jig, drilled in 3 places and then the paper sheets were interleaved, one-by-one, within the 3D printed head as shown in Figure 6-3. Excess paper was cut from around the border of the head as necessary. Once the whole phantom had been tightened using nylon screws it was placed on the camera bed, in the head support routinely used for patient acquisitions.

The phantom was scanned on a GE Infinia camera with LEHR collimators in place. Acquisition parameters were the same as those used clinically (see section 2.2.1). Acquisition time was adjusted such that the total counts in the scan were approximately 1.5Mcts, which is the minimum level considered acceptable by SNM guidelines and similar to the mean count level recorded for clinical patients over a 35 minute scan.

Following acquisition, projections were reconstructed using the same, standard reconstruction protocol on Xeleris and the same method for calculating SBRs as applied in previous investigations (see section 2.2.1 and 2.2.5 for details). In addition, the medial-lateral extent and the anterior-posterior extent of the striata in the new SSP design were measured on a 2D slab that was created through summation of 10 central brain slices. Ratios of these two values gave an aspect ratio that was a simple measure of striatal shape.

This analysis was conducted in MIM software after manually aligning reconstructed data such that the trans-axial plane was parallel to the line connecting the anterior and posterior commissure, and then further adjusting alignment such that there was approximate symmetry between left and right hemispheres in both coronal and trans-axial views (see Figure 6-8 for a diagrammatic representation of these pre-processing steps). Linear measurements of left and right striata were carried out manually with the caliper tool.

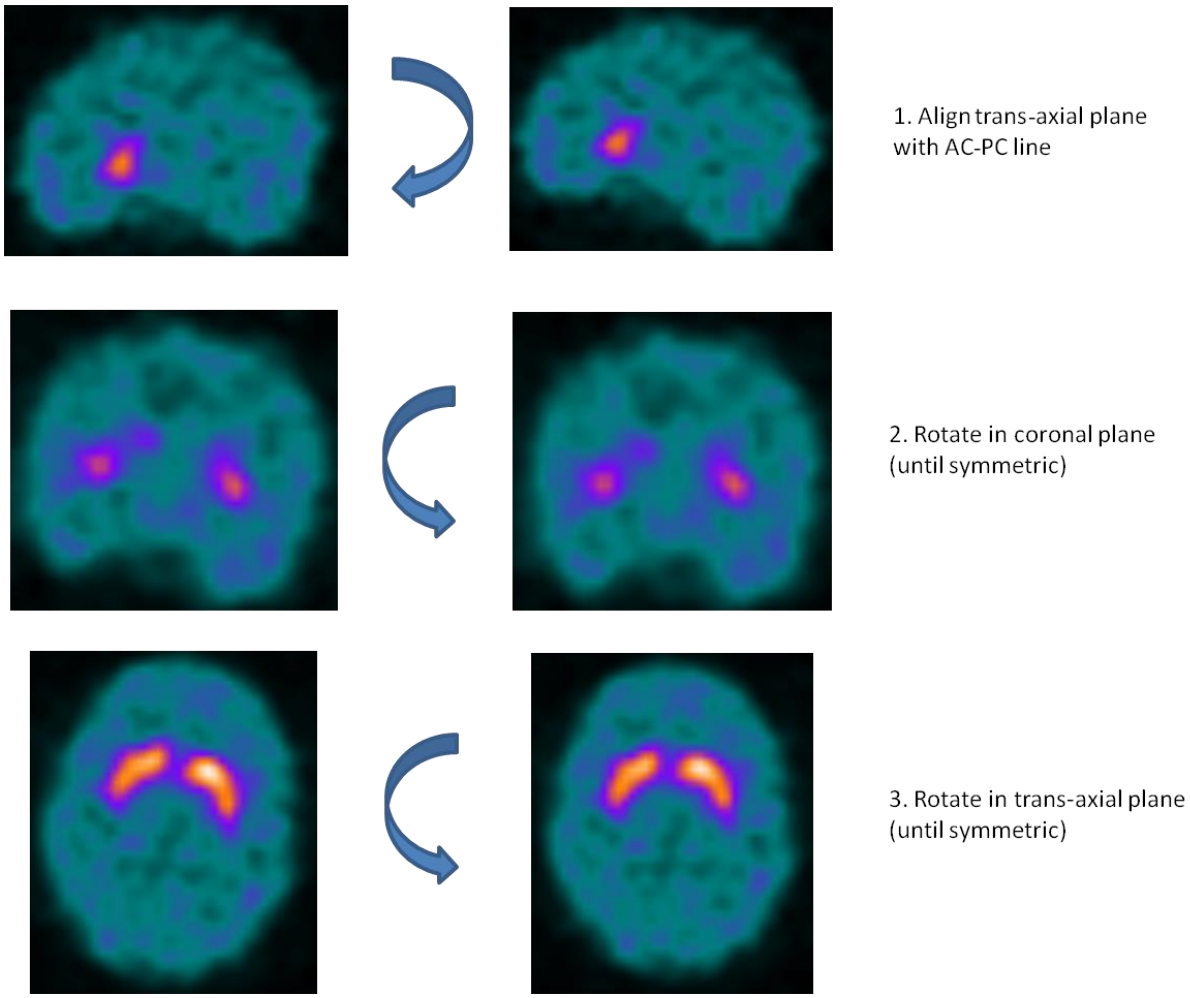


Figure 6-8 Diagrammatic representation of the image rotation steps carried out before summation of axial slices and measurement of striatal lengths

For comparison, analysis methods were also applied to 22 patient images from subset A where the probability of not having PDD was high (i.e. where uptake in the striatum was normal).

Results

Phantom printing and assembly took approximately 75 minutes to complete. Approximately 4ml of I123 Iodide-ink solution was required to print 56 slices of the (I123)FP-CIT template, covering the entire brain. In order to acquire 1.5Mcts over the course of the acquisition an imaging time of 30s per projection was required. Four central, transaxial slices taken from the reconstructed images of the new SSP design are displayed in Figure 6-9 (reconstructed slice thickness was 7.2mm).

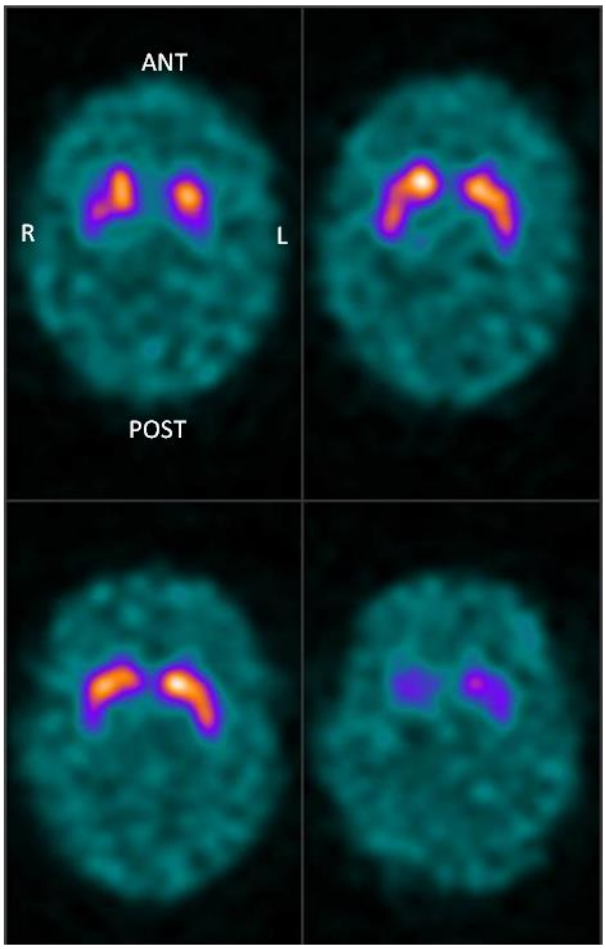


Figure 6-9 Four reconstructed slices from the centre of the new SSP design

Figure 6-10 shows striatal binding ratio results for the SSP and for 22 clinical patients without evidence of PDD (whiskers represent maximum and minimum values).

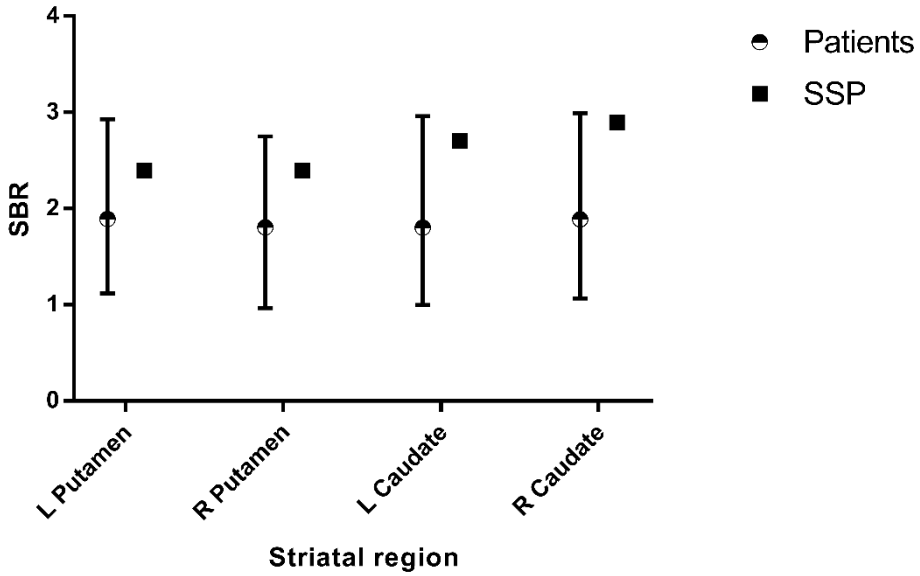


Figure 6-10 Striatal binding ratio results for the new SSP design and for 22 clinical patients from subset A without pre-synaptic dopaminergic degeneration. Whiskers represent maximum and minimum SBRs

Figure 6-11 and Figure 6-12 show linear measurements of striatal geometry and derived anterior-posterior / medial-lateral aspect ratio measurements for the new SSP design phantom and for the group of 22 patients.

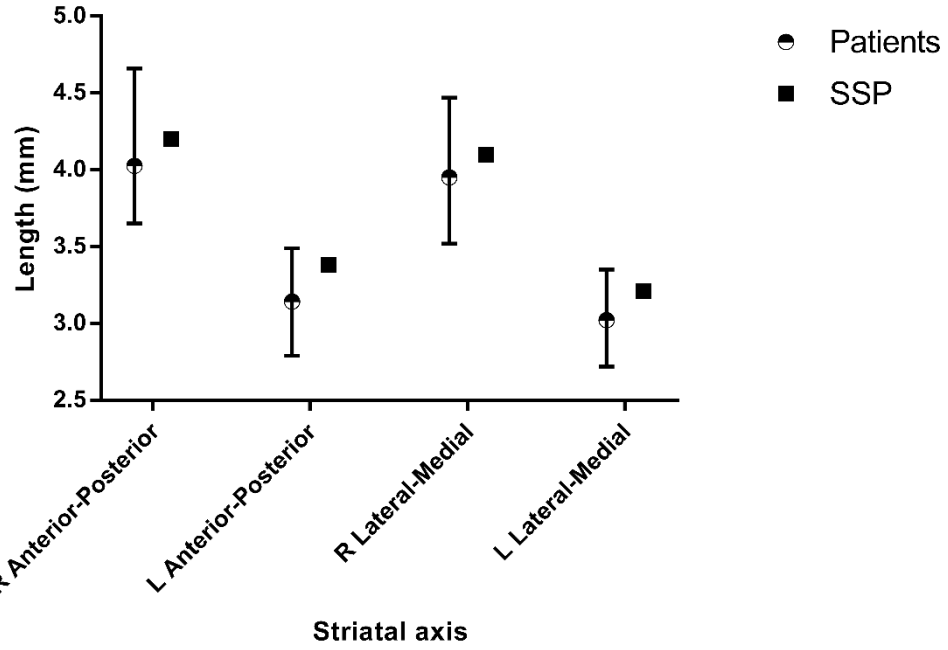


Figure 6-11 Linear measurements of the striatum in images acquired from the new SSP design and from a group of 22 patients without evidence of dopaminergic deficit. Whiskers represent maximum and minimum lengths. Taken from (3)

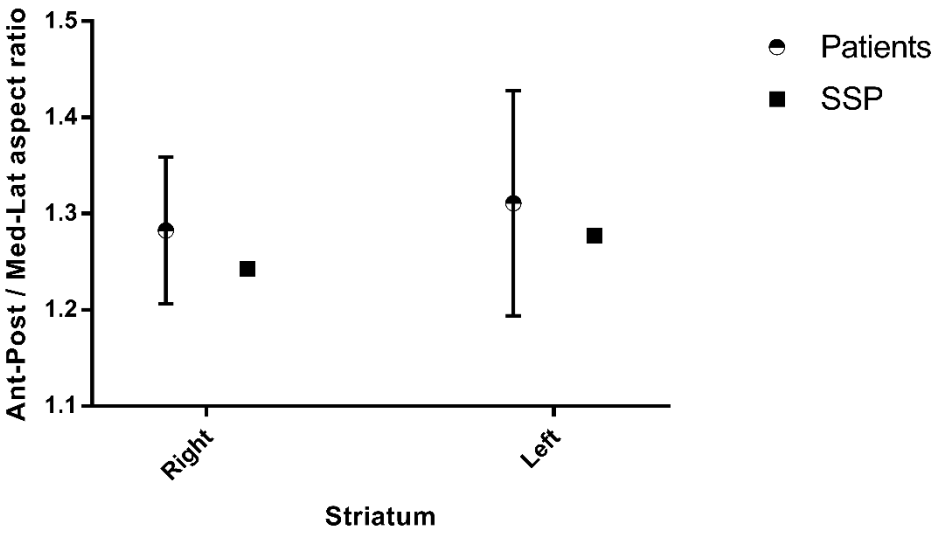


Figure 6-12 Anterior-posterior / medial-lateral aspect ratio measurements from the phantom and a group of 22 patients without evidence of dopaminergic deficit. Whiskers represent maximum and minimum aspect ratios. Taken from (3)

Discussion

This investigation presents a new application of SSP technology, printing phantoms with an ink mixture based on 123I Iodide. The volume of ink solution required to print one full phantom suggests that approximately 5 different radioactive ink patterns could be printed from a single standard-sized cartridge (if filled to a maximum capacity of 23ml). The total printing and assembly time was longer than the preparation time required for the Alderson phantom (approximately 1-1.5 hours for the SSP as compared to 30-45 minutes for the Alderson phantom). However, the SSP assembly process has not yet been optimised. There is potential for time savings by, for example, rounding off the ends of the nylon guides rods to enable quicker stacking of paper and plastic layers.

The imaging time required to reach typical clinical count levels was similar to that of patients, which suggests that even when using 123I Iodide at a relatively low radioactivity concentration (37 MBq/ml), SSP is a practical alternative technology to traditional fixed cavity phantoms. This is an important finding as 123I Iodide is widely (and cheaply) available to UK nuclear medicine departments.

Visual analysis of the reconstructed phantom images (see Figure 6-9) demonstrates uniform, high uptake throughout the striata on both sides with a wide area of relatively low, uniform uptake in remainder of the brain. This closely reflects the characteristics of the simplified anatomical template and provides some reassurance that there were no significant problems in phantom assembly. Visually, the reconstructed SSP images had similar appearances to that of real patients, although there was a noticeable lack of skin uptake (this could however be corrected in the design template). In addition, linear measurements of the shape and size of the striata were also within the range measured from real patients. This is reassuring in light of previous criticisms of the Alderson phantom.

Striatal binding ratio results (in the left and right putamen and caudate) for the SSP were above the mean but still within the range of results generated for a group of 22 non-Parkinsonian patients. This suggests that the SSP design was reflective of non-PDD patients with higher uptake ratios, and provided a degree of reassurance that the developed SSP method was capable of producing images reflective of a particular patient cohort.

However, it is also worth acknowledging methodological limitations:

- Patient age is a known co-variate for striatal binding ratio (40) and was ignored in the comparison exercise.
- SBR and striatal dimension measurements do not define every important feature of a (I123)FP-CIT dataset, and are therefore not a comprehensive assessment of the SSP's suitability for replicating particular scan appearances.
- The anatomical template used in this investigation is idealised. There are several areas where the template differs from real patient data. For example, patient SPECT images often show reduced uptake in the ventricles and increased uptake in extra-striatal tissues such as the pons and thalamus (125). Furthermore, the template only included increased uptake in the putamen and caudate, and not the smaller globus pallidus.

Although these limitations dictate that the developed phantom cannot yet be considered as providing an exact replica of real patient appearances, nonetheless the findings indicate suitability of the technique for creating 3 dimensional uptake patterns closely resembling certain (I123)FP-CIT scan features. Using the current anatomical template, image appearances are already more realistic than the available commercial phantom technology (i.e. the Alderson phantom).

Less idealized, more patient specific ink patterns could be created through suitable alteration of the anatomical template pattern. This could be achieved, for example, through comparison of reconstructed phantom and patient scans, followed by iterative adjustment and reprinting of the anatomical template, then repeated scan comparisons, as per Holmes et al. (119).

Thus, the results presented in this section show that SSP technology can be successfully adapted to (I123)FP-CIT imaging. The developed techniques lay the groundwork for the creation of realistic, patient specific phantoms.

6.3 Summary

Performing sensitivity analysis to examine the impact of different acquisition factors on machine learning algorithm performance in (I123)FP-CIT imaging requires repeated imaging of the same patient uptake pattern. Phantoms can enable this type of investigation without having to scan real patients. However, the only commercially available physical phantom, the Alderson phantom, is inadequate for replicating patient appearances to the level

demanded for investigations of classification software performance. Thus a new type of physical phantom was investigated and developed, based on sub-resolution sandwich phantom technology.

Tests were conducted to probe the practicality, controllability and repeatability of the chosen SSP production method for simulating (I123)FP-CIT appearances. Firstly, the radioactive ink printing process, which is central to SSP technology, was shown to be highly controllable and consistent. Ink deposition was predictable for a range of greyscale values (up to 0.9) and the printer produced sufficiently uniform printed sheets. Individual slices from an input anatomical template (based on the MNI dataset) closely matched the resultant output. All of these findings justify the use of the particular printer setup for creating phantoms for investigations of classification algorithm performance.

The final section of this chapter demonstrated successful application of SSP technology to creating a full (I123)FP-CIT phantom. Although the radionuclide used had a relatively low radioactivity concentration (approx. 18 MBq/ml when mixed with ink), resultant count rate was acceptable. Total assembly and printing time was slightly longer than the assembly time of the conventional Alderson phantom. However, unlike for the Alderson phantom, measures of SBR and linear striatal dimensions showed that the simple MNI-based anatomical template produced quantitative measures in line with clinical data.

Overall, this chapter demonstrated that the highly flexible SSP process is a practical, inexpensive and repeatable solution for creating controllable, 3D (I123)FP-CIT phantoms. This is a crucial first step for enabling comprehensive evaluation of automated classification tools, which isn't possible with existing Alderson phantoms. Through careful selection of the anatomical input template, the developed SSP technology can now be used to mimic a range of patient appearances and to evaluate any classification tool.

Having selected an appropriate physical phantom technology, the next chapter considers sensitivity analysis for (I123)FP-CIT classification algorithms.

7 Measuring the impact of clinical heterogeneity: sensitivity analysis

New objectives addressed by this section (in black, bold):
A) Examine and develop phantom technology to provide a toolset that can be adapted to simulate a range of realistic (I123)FP-CIT image appearances.
B-1) Use the toolset to demonstrate the influence of heterogeneity by: Analysing and prioritising the individual imaging parameters that may affect classification software performance
B-2) Use the toolset to demonstrate the influence of heterogeneity by: Performing sensitivity tests to measure the impact of different imaging parameters on developed classification tools

Table 7-1 New objectives addressed in section 7

Image acquisition procedures and protocols vary between hospitals, giving rise to different image appearances for the same tracer distribution. To minimise the effects from such confounding variables, clinical data in this work was taken from only a single institution where the acquisition procedures were consistent and the range of gamma cameras used was limited. If a mixed database of patients from multiple hospitals was used to train a classifier, with no algorithm adaptations to take account of the increased variability in the data, it is likely that accuracy would be reduced. Used in a CADx context, this could increase the likelihood of a clinician making an incorrect diagnosis, which could lead to inappropriate treatment.

However, even if the training database is restricted to data from a single institution, if the algorithm is applied on a wide scale throughout the NHS it is likely that at least some of the clinical cases would have been acquired in a different way to the training data, which would again reduce classification accuracy and increase the risk of incorrect patient care.

Consequently, the sensitivity of classification algorithms to different acquisition factors needs to be examined, and this is particularly pertinent if a fixed diagnostic threshold is involved. Once this important issue has been addressed, further effort can be expended in addressing other regulatory, economic and management aspects of the clinical translation challenge.

This chapter first seeks to develop a strategy for identifying and prioritising imaging parameters in terms of their potential impact on CADx / classification software. Following this, the developed SSP technology is used as an example, illustrating a sensitivity analysis of classification tools, according to the strategy.

7.1 Strategy for prioritising image acquisition factors

A strategy for sensitivity analysis relies on an understanding of the expected variability in image acquisition factors currently found in the clinic. In order to understand such variability, an audit should ideally be performed. However, although a national audit of Nuclear Medicine departments has recently been performed in the UK, revealing useful information with regards to the make/model of gamma cameras currently in use, and the type of collimator adopted (37), the data available is limited. A more comprehensive audit is required for this study but is not feasible within the time frame of this work.

Therefore, relevant clinical guidelines (from EANM, SNM and the leaflet supplied with vials of (I123)FP-CIT) are used instead. The range of values cited within these documents for different acquisition parameters is assumed to be representative of the breadth of acquisition conditions seen in different hospitals. This information is used as a basis for creating a focused strategy, dictating which types of tests should be carried out in order to understand the likely performance variability of classification software in clinic. Information from guidelines is supplemented by data from protocols used locally at Sheffield Teaching Hospitals, as well as relevant research, which may also inform practice.

Some of the differences in acquisition parameters that are implied by relevant guidelines are likely to be relatively unimportant in terms of impact on classification algorithm performance, or may be controlled to such an extent they are no longer of concern. Acknowledging that resources available for testing classification algorithms are likely to be limited, each acquisition factor is considered in turn to find those which are of highest priority (see Table 7-2). In each case the potential impact on classification algorithm performance is scored qualitatively considering the likely variability between centres. The potential for controlling variability is also scored. By combining these two outputs a priority rating is given. The highest scores are associated with acquisition factors judged to have a high potential impact on classification algorithm performance, and low potential for control.

In addition to image acquisition recommendations, clinical guidelines also provide guidance on patient factors, such as the influence of vascular lesions on image appearances and the effects of different types of drugs on tracer uptake levels. Although these are important, such patient selection and patient preparation considerations are excluded from the table below in order to provide a strategy that is focused purely on the mechanics of acquiring and processing data.

Image acquisition factors, ordered according to relative priority for investigation in sensitivity tests				
Image acquisition factor	**Explanation and potential variability**	**Potential impact on classification algorithm performance**	**Potential for control**	**Priority for investigation**
Camera-collimator design	Collimator geometry dictates image resolution, sensitivity and also image noise (through relative contribution from septal penetration). There are differences in guideline recommendations: the leaflet supplied with (I123)FP-CIT recommends “high resolution” collimators, SNM recommends LEHR or LEUHR collimators (32), EANM suggests that LEHR / LEUHR collimators are the most frequently used but specifically recommends fan-beam design over parallel hole design (6). This potential variability in selected collimator is exacerbated by the known large differences between the resolution and sensitivity of GE cameras with LEHR collimators and Siemens cameras with LEHR collimators (the two most commonly used systems (37)).	**HIGH**. Different camera systems (with different collimators) are known to give rise to different semi-quantitative values (50), and therefore different appearances	**LOW**. Many departments only have camera(s) from one manufacturer with limited collimator options	**HIGH**

Non-standard detector positioning	For patients that are not able to tolerate a standard acquisition due to, for example, claustrophobia or a physical deformity preventing scanning close to the head, gamma camera detectors can be positioned differently during acquisition to reduce chances of a failed scan. Although not specifically recommended by any guidelines, in Sheffield a single planar vertex view (acquired superiorly to the skull) is currently acquired for those that cannot tolerate a detector passing close to the face. Detectors are also set at an increased radius, incorporating shoulders, for patients with kyphosis. Recent studies suggest that diagnostically useful information could be collected from claustrophobic patients by acquiring data with detectors moving behind the patient only (126). Therefore, there are potentially many different non-standard acquisition procedures used in hospitals	**HIGH**. Planar vertex views cannot be processed by the developed algorithms. Acquiring data at a greater radius, or using an incomplete acquisition arc is known to cause significant changes in appearances as compared to a standard acquisition (126,127)	**LOW**. Claustrophobia is likely to be very difficult to control and physical deformities, such as kyphotic spine, cannot be scanned under a normal protocol using conventional gamma camera equipment	**HIGH**

Total acquisition counts	Radioactive decay is a random process and so the more counts detected from the patient, the greater the signal to noise ratio in the final image. The leaflet supplied with (I123)FP-CIT recommends total counts of > 0.5Mcts. SNM guidelines recommend > 1.5Mcts (32) and EANM guidelines recommend > 3Mcts (6). Thus, there is potentially wide variation between centres.	**HIGH**. Wide ranging image counts is likely to be associated with wide ranges in image noise and appearances	**HIGH**. Total counts could be much more strictly controlled if required (for example by estimating count rate from an initial static image and setting acquisition time accordingly).	**MEDIUM**
Radius of rotation	The radius of rotation dictates distance from the patient and therefore image resolution, which is an important consideration given that striatal tissues are small and relatively deep within the brain. The (I123)FP-CIT leaflet recommends a radius of 11-15cm (SNM guidelines suggest these values are “typical” (32)), EANM recommends the “smallest possible” radius (6).	**MEDIUM**. Imaging at 11cm or 15cm, as permitted by guidelines, has been shown to be associated with small changes in SBRs (127) (and therefore appearance)	**MEDIUM**. Departments could set stricter radius settings if tolerable by patients (if not, see non-standard acquisition protocols for alternatives)	**MEDIUM**
Reconstruction software and	Reconstruction method (whether analytical or iterative) and associated parameters have a substantial impact on contrast, noise and resolution.	**HIGH**. Reconstruction methods /	**HIGH**. Most departments have access to adjustable	**MEDIUM**

parameters	EANM and SNM both recommend either filtered back projection or iterative reconstruction with low pass filtering (6,32). Attenuation correction is also recommended by both. However, no guidance is given on number of iterations, subsets or other specific parameters and so clinical departments may have a wide range of different reconstruction settings.	parameters can have a substantial effect on measured SBRs (47,48) (and therefore appearances)	reconstruction software. It is possible to obtain similar reconstruction results between software from different vendors for example by measuring and matching the frequency response curves after applying different reconstruction parameters, as performed by Lawson et al. (128)	
Rotation step size	The rotation step size is one of the factors which determines SPECT image resolution EANM and SNM guidelines both recommend 3 degrees per step (6,32). The (I123)FP-CIT leaflet suggests this should be a minimum target. Thus, there is little difference in recommendations.	**LOW**.	**HIGH**. All modern gamma cameras allow fine control of step size	**LOW**
Energy	The energy window selected dictates the relative proportion of	**LOW**.	**HIGH**. All modern	**LOW**

window	primary and scattered photons detected by the system and is therefore linked to the signal to noise ratio and appearances of the image SNM guidelines and the (I123)FP-CIT leaflet both recommend a single window of 159 keV ± 10% (32). EANM gives no specific recommendations (6). Therefore, there is little difference in recommended values.		gamma cameras allow fine control of energy windows	
Pixel size (image zoom)	Image zoom (and pixel size) have an impact on image resolution and noise Both SNM guidelines and the (I23)FP-CIT leaflet recommend a pixel size of 3.5-4.5mm (32). EANM recommendations are very similar: a pixel size that is one third to one half of the system resolution (6). Therefore, there is little difference in recommended values.	**LOW**.	**HIGH**. Pixel size can be controlled to a high degree on modern camera systems	**LOW**

Table 7-2 Summary of acquisition factors with qualitative assessment of their potential impact on classification algorithm performance, and potential for control in clinic. Factors are ordered according to decreasing priority for investigation.

Table 7-2 provides a guide to machine learning researchers as to which acquisition parameters should be considered first when performing sensitivity analysis of classification algorithms for (I123)FP-CIT. It shows that there are two image acquisition factors which are judged to be of highest priority: camera-collimator design and non-standard detector positioning. Both of these factors are related to issues that cannot be controlled in clinic and which could potentially limit the scope of application of classification algorithms. Camera-collimator design is perhaps of most concern as the vast majority of (UK) patients are currently scanned with either a GE or Siemens system, equipped with LEHR collimators (37). If significant differences are found in terms of how a classification algorithm responds to these systems, a significant proportion of (I123)FP-CIT patients could immediately be excluded from the benefits of CADx. Should alternative detector positioning be contraindicated when applying CADx, the number of patients potentially benefitting is likely to be smaller.

The following section will consider these two highest priority acquisition factors in a set of practical investigations, putting into practice the developed phantom technology to assess changes in classification output as a result of varying acquisition conditions.

7.2 Method

Ideally, sensitivity tests would be conducted with a number of different, realistic SSP designs, to reflect the expected range of patient appearances, in order to comprehensively assess all the developed classification algorithms. However, the radionuclide and gamma camera resources available within this study for further developing anatomical template designs are limited. Therefore, the following sensitivity tests continue to use the simplified SSP anatomical template that is already established, in a limited number of imaging acquisitions.

The continued use of an idealised anatomical template dictates that acquired images are most useful for testing classification algorithms based on simple, derived image features such as SBRs, rather than those based on raw voxel intensities or principal components (which would require phantoms that were clinically realistic at the voxel level). Therefore, only the simplest of the algorithms developed in chapter 2 is analysed: the ML46 algorithm (see chapter 2.1.3), where 4 SBR values and patient age are input to a linear SVM. Algorithms based on principal components, such as that adopted for reporting tests in

chapter 4, could be tested in the same way in future once more realistic anatomical templates have been created and optimised.

7.2.1 Metrics

To assess the impact on classification algorithm performance of different acquisition settings, appropriate metrics are required. However, given that clinical heterogeneity is rarely considered in the machine learning literature, there is little guidance or consensus on which metrics may be most appropriate. For the following studies the raw SVM scores output from the algorithm are examined, in line with previous work by Abdulkadir et al. (129). To provide context, Figure 7-1 characterises the distribution of SVM scores measured for patients in subset A for the ML46 algorithm. This shows that the SVM output was more variable for the non-PDD, normal patients than for the diseased group. The gap in scores between the pre-synaptic dopaminergic degeneration group and the non-dopaminergic degeneration group (approximately 4 on average) provides a baseline against which changes in SVM score, as a result of patient and equipment factors, can be compared.

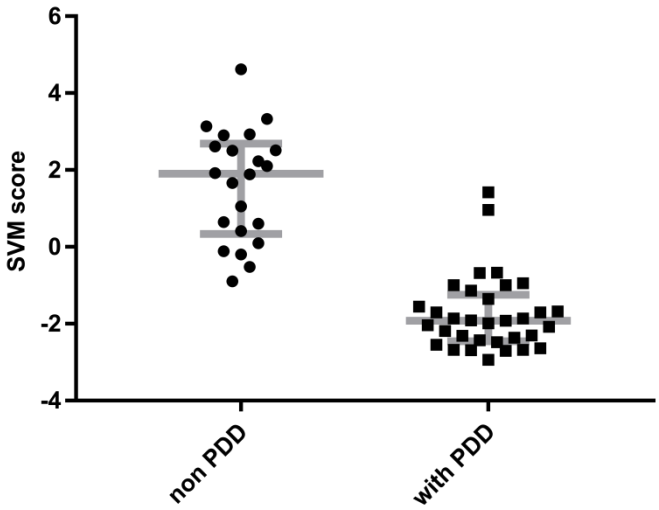


Figure 7-1 SVM score distribution for patients in subset A, using features based on SBRs. Median and inter-quartile ranges are shown in grey

7.2.2 Acquisition

The methodology for each investigation is summarised in Table 7-4. For each set of tests, unless otherwise stated, most camera acquisition parameters were consistent (see Table 7-3). These parameters were largely dictated by the available guidance on (I123)FP-CIT imaging (6,32) and the settings used in the local clinical service.

Acquisition parameter	Value
Total counts	1.5 Mcts
Matrix size	128 x 128
Radius of rotation	14cm
Zoom	1.2
Projections	60 per head over 180 degrees
Energy window	159keV ± 10%
Collimator	LEHR

Table 7-3 Camera acquisition parameters

Following acquisition, each scan was reconstructed using the standard reconstruction protocol on Xeleris. Data were then processed through the selected machine learning algorithm as previously described (see chapter 2). Patient age was assumed to be 60 years in all cases. All investigations in this chapter are based on algorithms trained with local clinical data (from subset B) only, as such algorithms are more likely to be of use in the clinic than algorithms trained with PPMI research data.

Sensitivity analysis: summary of investigation methods	
Factor	**Investigation method**
Camera-collimator design	Tests considered two different acquisition cameras: a GE Infinia and a Siemens Symbia. These are the most commonly used acquisition systems, according to a recent UK audit (37). Using the same template shape described in the previous chapter, two different phantom prints were designed with two different greyscale ratios. The first phantom had printed count densities of 8 to 1 in both striata as compared to the remainder of the brain ('normal' phantom) whilst the second phantom had a count density ratio of 5 to 1 ('borderline' phantom). 8 to 1 is representative of the true underlying tracer concentration ratio present in normal controls (130). 5 to 1 represents a more borderline case. Additional uptake ratios were not tested due to limitations on time and scanning resources After assembly, each phantom was scanned twice on each camera (enabling evaluation of repeatability). Scanning parameters were either the same or as close as possible for each camera (the only difference was in applied zoom, 1.2 for the Infinia and 1.23 for the Symbia).

Non-standard scanning conditions	Tests evaluated three alternative scanning protocols that may be applied for complex patients, including those that are claustrophobic and those with anatomical abnormalities (such as kyphosis), which dictate that detectors cannot be brought close to the patient's head. In the local clinical department, these are the main reasons for protocol deviations or abandoned acquisitions. Two acquisition methods were considered for claustrophobic patients: acquiring from behind the head only using either a single detector (with the other at maximum radial distance), or both detectors set at 90° to each other in 'L-mode' (126). For patients where anatomical abnormalities prevent a small radius of rotation, a standard acquisition at an increased radius was the only alternative considered. The same anatomical template was adopted as in the previous investigation (i.e. striatal uptake was defined from adaptation of the MNI template). The fully assembled head was scanned four times on a GE Infinia camera, as follows: 1) Standard acquisition procedure 2) 'L-mode' acquisition, acquired over 180°, posterior to the patient. 3) Standard acquisition with one active detector acquiring data close to the posterior of the head and the other inactive detector at the maximum radius away from the patient's face 4) Standard acquisition with both detectors at an increased radius of rotation (21.4cm, such that a patient's shoulders could reasonably be expected to be included within the field of view) A schematic of these different acquisition conditions is shown in Figure 7-2 (numbered according to the list above). The anatomical template was printed twice, with different striatum to background ratios, to reflect two different patient appearances.

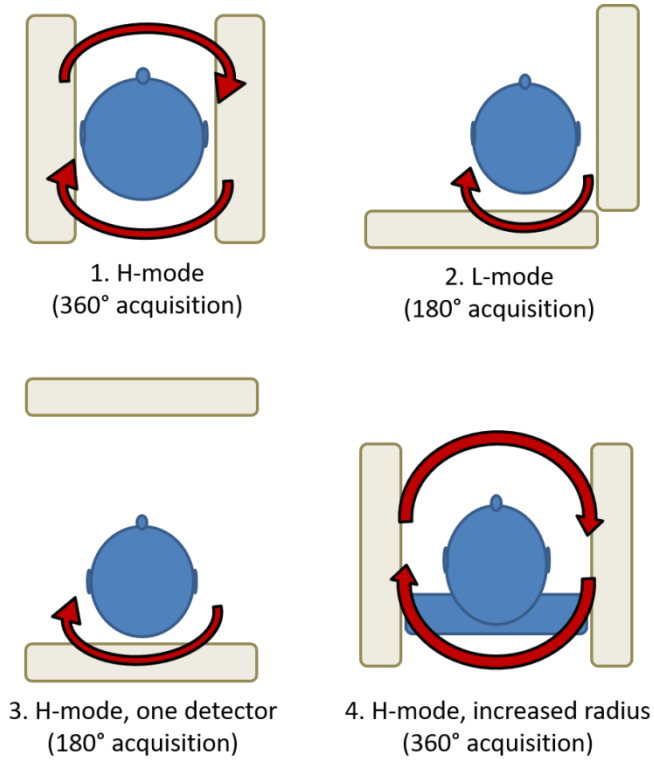
	As before, print 1 had a striatum to background greyscale ratio of 8 to 1 on both sides of the brain (‘normal’ phantom). Print 2 had a striatum to background ratio of 6.5 to 1 on the right and 5 to 1 on the left side (‘borderline’ phantom).  *Figure 7-2 Schematic of the different acquisition protocols.*

Table 7-4 Summary of investigation methods for assessing classification algorithm sensitivity to different acquisition factors

7.3 Results

7.3.1 Camera-collimator design

Figure 7-3 shows the SVM score results for the two phantoms scanned on the different camera systems (for an assumed patient age of 60 years). Note that SVM scores for the borderline phantom are so close to zero that they are barely visible. Table 7-5 summarises repeatability errors. Table 7-6 provides an estimate of between camera differences, taken from differences in the means of the repeated scans on each camera.

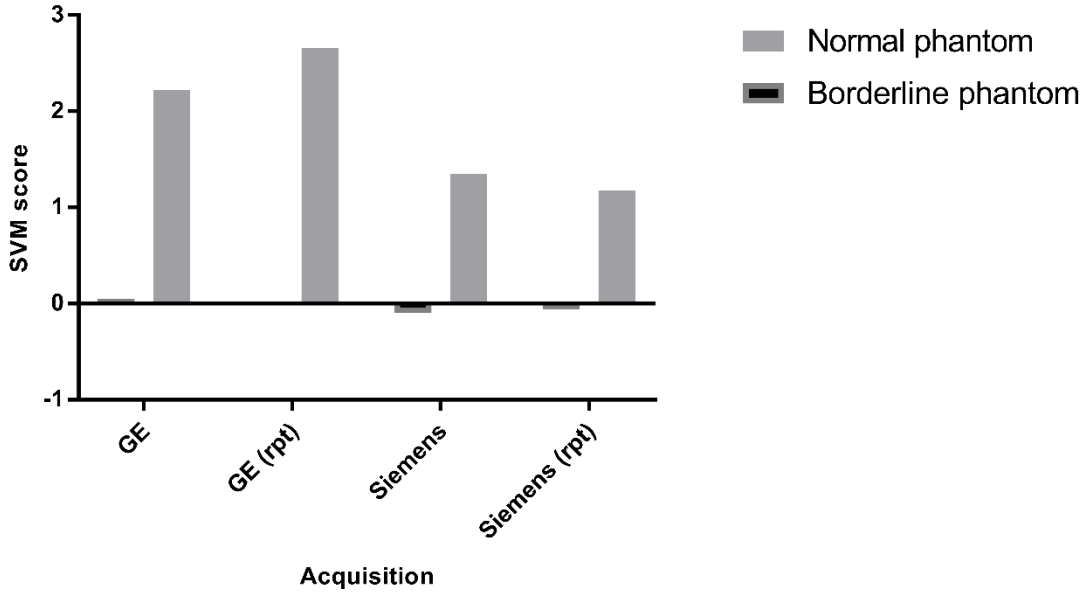


Figure 7-3 Summary of the mean SVM score results for different camera systems (assumed patient age of 60 years)

Camera system	Phantom	SVM score differences between repeat scans
GE	Borderline	0.05
GE	Normal	-0.44
Siemens	Borderline	-0.04
Siemens	Normal	0.17

Table 7-5 Summary of repeatability results

Phantom	Mean SVM score differences between cameras
Borderline	0.10
Normal	1.18

Table 7-6 Summary of camera comparison results (mean GE Infinia result minus mean Siemens Symbia result)

7.3.2 Non-standard scanning conditions

Figure 7-4 shows reconstructed slices following acquisition under standard conditions, and with both detectors set in L-mode. Figure 7-5 summarises the figures output from the SVM algorithm after inputting data acquired under different conditions. Table 7-7 shows differences in SVM scores as compared to standard scanning conditions.

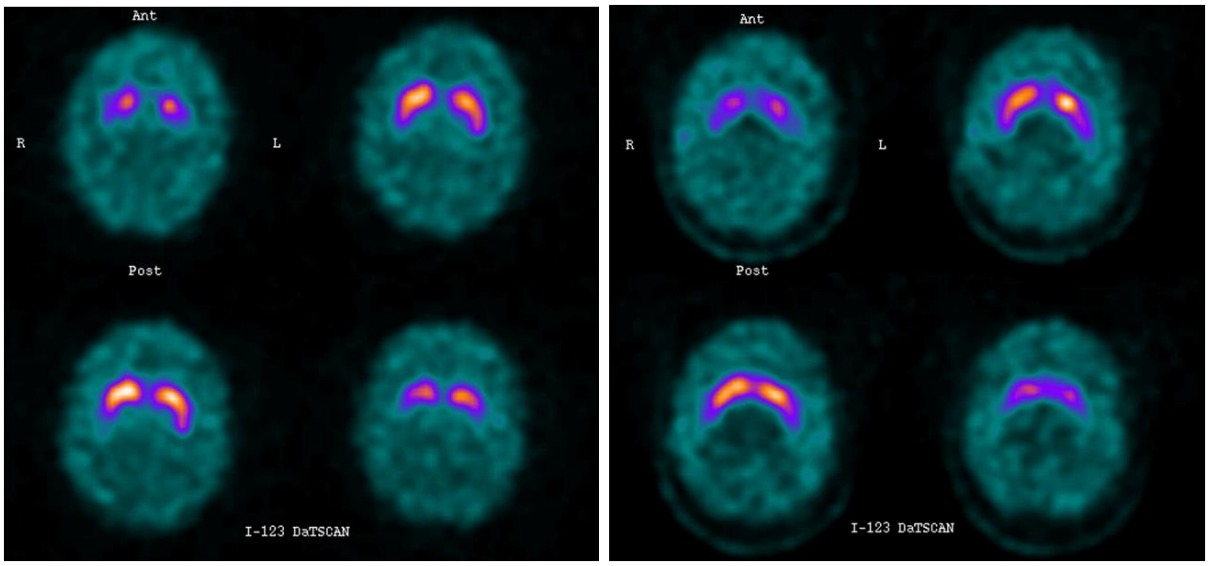


Figure 7-4 Reconstructed slices from the normal phantom acquired in H-mode (left) and L-mode (right)

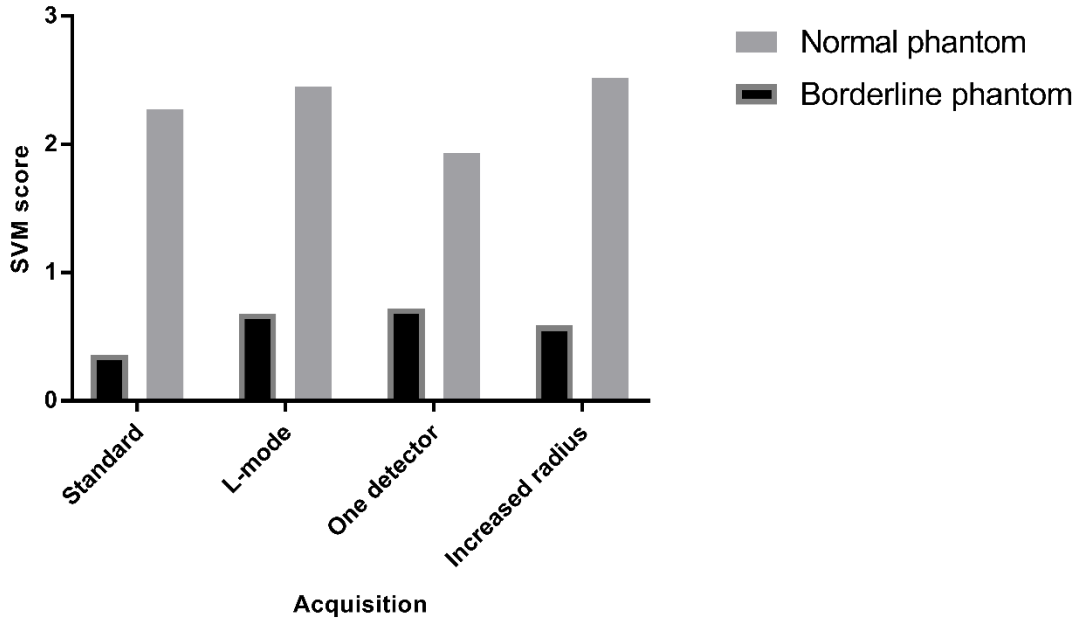


Figure 7-5 Summary of the SVM score results for different acquisition protocols for both phantoms (patient age of 60 years)

	SVM score differences vs standard acquisition		
Phantom	**L-mode**	**One detector**	**increased radius**
Borderline	0.18	-0.35	0.25
Normal	0.33	0.36	0.23

Table 7-7 Summary of acquisition conditions comparison (SVM score from standard acquisition minus SVM score from alternative scenario).

7.4 Discussion

This chapter first sought to develop a strategy for prioritising the different acquisition conditions that could affect classification algorithms performance, and to conduct a sensitivity analysis using developed phantom technology. These endeavours are a response to the research objectives identified in chapter 5.

The sensitivity analysis strategy was largely based on examination of relevant clinical guidelines with respect to equipment factors. Although this is likely to be less reflective of the range of acquisition conditions seen in clinic, findings provide a suitable starting point for evaluation of classification tools. The strategy could perhaps be updated and improved in future through consultation with relevant stakeholders in the (I123)FP-CIT clinical community.

It was intended that sensitivity analysis would be conducted using a number of different phantoms, after further optimising the SSP anatomical template to more closely reflect real patient appearances. Although, due to resource constraints, this was not achievable, results from the application of a simplified phantom to one of the developed classification tools do provide useful insights, as described below. The processes and analysis employed here could be applied to any classification tool, including that used for clinical reporting tests in chapter 4, when more realistic anatomical template designs have been created and optimised.

7.4.1 Camera-collimator design

Many clinical departments in the UK do not have the luxury of being able to choose between cameras when scanning patients. Often there may be only one camera available or there may be multiple cameras locally, but all from the same manufacturer. Furthermore, after a gamma camera has been in use for longer than its expected lifetime, a hospital may choose to buy a completely different model as a replacement. Across the country there will be multiple different types of systems in use at any one time, with varying acquisition procedures between centres. Before a CADx or automated classification tool can even be considered for development into widely available software, performance for the most commonly available camera-collimator equipment needs to be understood. This is a key consideration that is often neglected by the machine learning research community.

Camera comparison results demonstrated that differences between GE and Siemens systems can be relatively large for a machine learning algorithm built on SBR features. The maximum recorded difference in SVM score was 1.18, which is approximately one third the size of the mean difference between normal and abnormal patients (see Figure 7-1). This score differential was consistent across all assumed patient ages (due to the linear nature of the adopted classification algorithm). In every case the GE camera produced a higher classification algorithm output than the Siemens camera, thus for every patient there was an

increased probability of belonging to the non-PDD group if they were scanned with GE equipment. These findings are perhaps unsurprising given that the classification algorithm was created with data that were only acquired from GE systems, and so the Siemens data represented a shift in input signal that was not accounted for in the training process. Previous research has shown that GE and Siemens scanners give systematic differences in SBR measurements (50), due to differences in resolution and septal penetration characteristics. In this case it appears that such differences in SBR translated into an overall change in classification algorithm score.

It is interesting to note that the magnitude of the change in SVM score was far larger for the phantom with a count density ratio of 8 to 1 (normal phantom) than for the phantom based on a count density ratio of 5 to 1 (borderline phantom). This can perhaps be attributed to the underlying SBR figures, which changed by a greater magnitude for the normal phantom when scanning on the different systems. There was also a substantial difference in the repeatability error: differences in SVM score were 0.05 and 0.44 for the borderline and normal phantom respectively. Without taking further measurements it is difficult to ascertain whether repeated scanning (and processing) of the more borderline simulated patient is associated with less variability, or whether variability differences between the phantoms are due to chance. However, whatever the underlying pattern in repeatability error, it is clear that differences in acquisition equipment can have a substantial impact on the algorithm under consideration, and would potentially stop it from being developed further into a widely available clinical tool.

The apparent differences in results for the two phantoms emphasises the importance of testing a range of different patient appearances (ideally more than was investigated in this case), in order to fully assess the sensitivity of a classification algorithm. This again emphasises the inadequacy of the fixed Alderson phantom for evaluating machine learning tools. Ideally, further experiments would also be carried out on other commonly used camera systems to evaluate whether patterns seen here are reflected more widely.

In this study a single GE and Siemens camera were tested. However, it is likely that there would be some variability between the performance of individual cameras of the same model type. Tests on multiple identical cameras would be needed to quantify this variability.

The magnitude of the measured change in classification algorithm output needs to be understood in the clinical context. Figure 7-3 shows that although the changes in SVM score

were much smaller for the borderline phantom (and may not be significant beyond repeatability error), scanning on the different systems did cause the SVM output to change between a positive and negative value. Thus, overall binary classification for the phantom changed depending on the camera used (classification was normal when scanned on the GE system, and abnormal group when scanned on the Siemens system). Conversely, the larger changes in SVM score seen for the normal phantom still resulted in a large, overall positive classification algorithm score (and therefore high probability of belonging to the normal class), no matter which camera system was used. Hence, it could be argued that the small changes seen for the more borderline simulated patient are perhaps more critical in the clinical context, particularly if the classification tool was to be used as an automated system for screening out normal scans from the reporting list.

7.4.2 Non-standard scanning conditions

(I123)FP-CIT scanning using a standard protocol requires patients to remain still for approximately 35 minutes, with gamma camera detectors passing close to the face. In addition, the patient's head is firmly strapped into a support. However, some patients cannot tolerate these conditions due to different physical or mental health difficulties. For claustrophobic patients seen at Sheffield Teaching Hospitals this would have previously dictated that the scan was either abandoned or a planar vertex view taken. For those with a physical deformity, preventing scanning close to the head, acquisition would have continued but with detectors at an increased radius (such that the shoulders could be included in the field of view for example). All such patients were excluded from the local databases used in this thesis.

Although not recommended specifically by any clinical guidelines, the local Nuclear Medicine department is currently considering alternative scanning procedures for claustrophobic patients in order to reduce the number of failed acquisitions, namely: scanning from behind the head with either a single detector or with both detectors set at 90° to each other. Other Nuclear Medicine departments are currently using such protocols routinely. It is essential to understand how a classification algorithm performs under these conditions, or when the detector radius is set at a larger value, such that decisions can be made as to whether CADx is contraindicated for these patients or not. Without such evidence or guidance machine learning tools could cause inappropriate diagnosis or patient care when used in the clinic

Figure 7-4 demonstrates one of the problems caused when data is collected asymmetrically, from only one side of the patient, namely that the striata can appear warped. This geometric distortion was as expected given that gamma camera imaging characteristics, including attenuation, scatter and particularly resolution, are depth dependent. In a standard H-mode acquisition, acquired for one full rotation, the effects of improved resolution near the camera face, and degraded resolution far from the detector, are averaged out over the whole scan. However, for a 180° L-mode acquisition where detector heads never pass close to the patient's face, the anterior brain is never sampled at a higher resolution. The distortion effects from an incomplete acquisition rotation are worse the further away the object from the centre of rotation (131,132). Unfortunately, in this case it was not possible to place the detectors close enough to the bed to always keep the striatum within the central field of view. Thus, the geometric distortion appears worse than might otherwise be expected.

However, despite the obvious differences in appearances caused by the alternative scanning conditions for claustrophobic patients, the measured differences in SVM score were not as large as might be expected. For the normal phantom, for example, the SVM score differential as compared to standard scanning conditions was 0.33 and 0.36 for the L-mode and one-detector acquisitions respectively, which are far lower than differences caused by scanning on a different camera system (see 7.3.1). These differences are also lower than the maximum repeatability error seen in the previous section. SVM score differentials for the phantom representing a more borderline patient (5 / 6.5 to 1) were of similar magnitude and are perhaps of more concern in the clinical context given that the machine learning algorithm output is closer to zero (the boundary between normal and abnormal groups).

The fact that there was much less contrast between phantom results in this investigation, as compared to the previous, again shows that reproducing a range of simulated patient appearances is important for fully appreciating the impact of different acquisition settings.

Due to the geometric distortion induced by both forms of 180° acquisition, the image registration step is likely to have played a significant part in either causing or reducing discrepancies in SVM or SBR score. Visual analysis suggested that the affine registration had placed each striatum in approximately the right position as compared to the template. However, given the thin, elongated shape of the striatum in the L-mode 180° acquisition (see Figure 7-4), there is no single 'correct' registration. A number of different scan transformations may have placed the organs in approximately the correct position, each of

which may have led to different scores. Therefore, results presented here are as much a function of image pre-processing as the analysis algorithms themselves, probably more so than for investigations of other factors.

Results for the acquisition carried out at an increased radius of rotation, simulating the effects of including a patient's shoulders within the field of view, showed a consistent but relatively small effect on SVM scores across the two phantoms. This suggests that the loss in resolution caused by this scanning setup may be less important in terms of algorithm performance than acquiring data on different camera systems.

It is interesting to note that in all but one case, the 3 alternative scanning scenarios gave SVM scores that were slightly higher than for the standard acquisition (therefore having a greater probability of belonging to the normal patient group). This was unexpected given that each of the alternative acquisitions was associated with appearances that would normally be considered to be more abnormal. These findings emphasise the fact that a machine learning algorithm built on derived SBR features may not necessarily behave in the same way as a human observer. Classification algorithms which analyse raw pixel values, or principal components of raw pixel values, may give very different results.

This investigation considered only a limited number of patient conditions that would require alternative scanning procedures. There are many more that could be considered, particularly those causing patient movement during acquisition. For example, a patient suffering from DLB may be very forgetful and may try to get up during an acquisition. Furthermore, for the patient group referred for clinical (I123)FP-CIT scanning, Dyskinesia is a common symptom, which may cause constant movement of the striatum during a scan. Any such significant movement is likely to lead to reduced resolution and perhaps reduced contrast, which could result in significantly altered striatal appearances and reduced classification algorithm reliability. Investigations are required to examine the extent of movement that can be tolerated by classification algorithms before their use becomes contraindicated. However, given that there is likely to be wide variability between patients in terms of the magnitude, timing and duration of their movements, investigation of effects on classification algorithm performance would require a large number of acquisitions to be performed to cover these variations. These could be considered in future.

7.5 Summary

This chapter first provided a strategy for prioritising investigations into the sensitivity of classification systems to different acquisition conditions, as part of an overall aim to address heterogeneity of the clinical environment, a vital consideration in the drive towards clinical translation. The two acquisition conditions found to be of highest priority in terms of their potential impact were camera-collimator design and non-standard patient positioning.

Following this, a set of sensitivity analyses were conducted using developed SSP technology. A single machine learning algorithm was considered for all tests, based on binding ratios and patient age input to a linear SVM classifier. Of all the developed classification algorithms, this was most suited to testing with the idealised SSP patient uptake pattern that was developed in previous investigations. However, the testing methodology applied in this chapter is likely to be relevant to any form of classification tool (assuming that sufficiently realistic phantoms can be produced) and so provides a guide to other researchers also grappling with the challenges of clinical translation.

Initial investigations characterised the baseline level of difference in SVM score between PDD and non-PDD patient groups (acquired under a single set of acquisition conditions), to give perspective to the findings. The examinations that followed utilised a limited number of phantom acquisitions, based on an idealised uptake pattern.

Overall, the collected evidence suggested that utilising a different gamma camera system to that which the algorithm was trained on can give a systematic change in SVM score, which can be substantial and may be larger than repeatability error.

Selection of an alternative acquisition protocol, as might be used for claustrophobic patients or those with physical deformities, was generally associated with smaller changes in SVM output, less than the maximum repeatability error. Such protocols are clearly less of an issue for the classification algorithm, even though they were not addressed in the algorithm's training process. The raw SVM scores from each alternative scenario were generally increased as compared to standard scanning conditions. This was despite the more abnormal striatal appearances. These findings emphasised the fact that a classification algorithm built upon derived binding ratios may not always behave in a similar fashion to a human observer.

The SVM score provided a useful metric for analysing the sensitivity of SVM algorithms to different scenarios. However, in the clinical CADx context the impact of the reporter also needs to be considered in order to understand how changing algorithm outputs translates into changing diagnostic decisions. This is a more complex problem, which may vary depending on the experience, skill and confidence of the reporter. In order to make any insights further reporting tests are required, using changing algorithm outputs.

It is clear from the analysis above that different acquisition factors can have a substantial impact on classification algorithm performance. Results imply that, for the algorithm under consideration, use of a Siemens camera may be contraindicated due to the potential substantial shift in SVM score. Thus, theoretically, a large proportion of UK centres would not be able to benefit from the software, if it were released for clinical use in its current form. Even at a local level, results suggest that if a new camera were purchased (with different characteristics), the CADx / classification system may not achieve adequate performance. Such major limitations would be likely to discourage further investment. This again shows the importance of considering heterogeneity of the clinical environment in the journey towards clinical translation.

However, algorithms such as this can be redesigned to some extent in order to ameliorate the impact from different acquisition conditions, in order that the potential for positive clinical impact is maximised. Possible methods for overcoming issues highlighted in sensitivity analyses, for this algorithm and more widely, are discussed in the following section.

7.6 Algorithm adaptations for the clinic

The problem of differing data acquisition conditions is a major concern in other machine learning applications in other imaging modalities (133). Often in the radiological machine learning literature, algorithms are developed using established ‘legacy’ data, acquired using certain equipment and protocols. Frequently this is data from research studies, such as the Alzheimer’s Disease Neuroimaging Initiative (ADNI), http://adni.loni.usc.edu/. However, once trained, the goal is usually to apply machine learning algorithms to images acquired under different (clinical) conditions, using different and often more modern scanning equipment. Under these conditions the performance of even state-of-the-art algorithms, such as Convolutional Neural Networks (CNNs), is often poor (134).

Techniques for adaptation of machine learning algorithms to different acquisition conditions are often referred to as ‘transfer learning’ or ‘domain adaptation’. There are many methods that fall under these headings. The most obvious solution to enable domain adaptation is to collect large databases of images covering all possible data acquisition conditions, where patient classification is known. This data can then be directly implemented within the algorithm learning process such that the relationship between extracted features and different acquisition parameters can be modelled. However, this is usually impractical.

One possible approach to transfer learning, utilising only limited data, involves different weighting of samples in the training process according to their origin. For example, in the adaptation of classification algorithms to patient data acquired from a Siemens camera, images acquired from a GE camera could be given a lower weighting in algorithm training than samples from a Siemens camera. Here the GE data would help to regularise the classification model but not at the expense of reduced performance on Siemens data. This form of sample weighting can help to create reliable classification algorithms even with just a few training examples acquired under the target scanning conditions (133). However, improved performance using such techniques is not necessarily guaranteed. Furthermore, research into the specific use of transfer learning for SPECT data is so far very limited and thus there is little indication of the likely success of this form of algorithm training.

Alternatively, different acquisition factors could be normalised for in image pre-processing. This could be achieved, for example, by incorporating physics models into the processing pipeline or deriving empirical correction factors between camera systems based on limited phantom data, such that semi-quantification figures are transferrable between different equipment. This is the approach used for the PPMI data and in the ENCDAT project (http://earl.eanm.org/cms/website.php?id=/en/projects/enc-dat.htm). Although such corrections ignore differences due to striatal anatomy variations, this technique has been shown to reduce differences in SBR figures between camera systems (49) and would be likely to dramatically reduce the inter-camera variance seen in the sensitivity analysis of section 7.3.1.

Overall, it is acknowledged that classification algorithms could be further adapted, if required, to enable better generalisability to different clinical environments. However, even if machine learning tools are trained to cope with varying acquisition conditions from the outset, methods for validating the impact of clinical heterogeneity are still required. The previous two chapters provide guidance on methods that could be used to assess (I123)FP-

CIT classification algorithms in relation to this important translational issue, and represent an important step forward in clinical machine learning research.

8 Concluding remarks

8.1 Summary

This thesis was composed of two separate but complementary parts. In the first part the main research question was: How effective is a CADx tool, based on established machine learning algorithms, for assisted (I123)FP-CIT image reporting? Defining effectiveness in terms of independent algorithm classification accuracy, and in terms of the impact on reporter performance, the question was addressed through setting and pursuing 5 main objectives:

1) Select and implement machine learning classification tools.
2) Collect a database of (I123)FP-CIT images.
3) Compare the performance of machine learning algorithms with semi-quantification.
4) Develop software for testing of human reporters.
5) Assess the impact of an automated classification tool, implemented as a CADx system, on reporting.

These objectives were met in the following ways:

A selection of binary automated classification algorithms were designed and implemented based on promising techniques identified in the literature. Data for training and testing was collected from Sheffield Teaching Hospitals NHS Foundation Trust, with a minority of the patient cases having diagnosis established through clinical follow-up. This data was supplemented by images from the PPMI database. In the cross validation exercise the developed classification algorithms demonstrated similar or superior standalone performance in classifying both the local and PPMI data, as compared to a wide range of different semi-quantification methods. This is the first comprehensive comparison exercise conducted between these assistive technologies, and provides justification for the pursuit of clinical machine learning tools.

Software for capturing reporting decisions of human observers was manufactured through adaptation of an existing clinical program. A CADx tool based on 5 principal components and a support vector machine classifier was selected for testing. A pilot study and main clinical study quantified the effects on reporting decisions from using this tool, again with both local and PPMI data. CADx output was in the form of a single probability value. It was shown that

reporters gave more consistent image scores (with increased inter-reporter reliability) after being exposed to these figures and that reporting accuracy was higher as a result, particularly when reporters were shown images with unfamiliar appearances. It was also found that intra-reporter variability, using visual analysis alone, was high. This was the case for both junior radiologists and more experienced reporters, which exposes the disadvantage of relying on human visual perception alone in image reporting. Indeed, reduced reporting variance may be one of the most important benefits of CADx, particularly given that (I123)FP-CIT imaging is a relatively rare test which, in smaller nuclear medicine departments, may only be seen by a reporter a few times a year.

These novel, encouraging results for CADx were complemented by analysis of questionnaire results, which provided rich insights into the reporter-CADx relationship (which has not previously been studied in the field of (I-123)FP-CIT imaging). In particular, it was found that reporters were generally highly trusting of the classification system and that reporters from different clinical backgrounds appeared to have differing opinions on the usefulness of a CADx system.

The standalone binary classification performance of the CADx tool was consistently shown to be at a similar level to that of experienced reporters. Results therefore suggested that the machine learning tool could perhaps instead be used independently, possibly as an initial screening tool to remove normal cases from the list of images viewed by reporters. The main advantage of this approach over the CADx paradigm is that the potential efficiency savings are higher.

Despite the large number of published journal articles related to machine learning and (I-123)FP-CIT classification, this work represents the first attempt to consider and evaluate the impact of classification tools in a clinical scenario, with reporters. Such investigations are vital for moving machine learning technology towards clinical use and the positive results add weight to the arguments in favour of clinical translation. In answer to the original research question, the results summarised above demonstrate that this form of CADx is highly effective for assisting (I-123)FP-CIT reporting.

However, although machine learning had been considered here in the clinical context, generating new and important information related to likely algorithm performance, the initial approach had arguably been naïve. There was still a long way to go before the developed classification tools could be used routinely in hospitals. Notwithstanding the need to

generate more robust statistics through collection of more data, testing of more reporters and perhaps even conducting a prospective clinical trial, it was demonstrated that there are many other psychological, economic, legal, management as well as technical issues that needed to be addressed on the path to widespread clinical translation. Although other authors have suggested that clinical translation is often not adequately addressed by the research community, these issues were perhaps more wide ranging than has yet been described in the literature. The scale of the translation challenge is substantial, affecting all machine learning applications in radiology, not just for (I123)FP-CIT imaging.

Frustratingly, many researchers are either unaware of these issues or choose to ignore them. Consequently, clinical use of machine learning tools remains disappointingly low in medical imaging. Undoubtedly a new research approach is needed, focusing more on the translation gap than continual development of the technologies themselves. As shown by this thesis, even well-established machine learning algorithms are already sufficiently mature to offer real benefits to clinical care. Making sure such technology can and does thrive in the clinic should be a greater priority. It is hoped that the research community will in future take heed of this suggestion.

The second half of the thesis considered aspects of the translation burden in relation to (123)FP-CIT classification software, to improve the prospects for future clinical uptake. Given the limited remaining time and resources it was decided that the final chapters would focus on addressing one of the most pressing and most significant technical barriers, namely heterogeneity of the clinical environment. Specifically, it is known that gamma camera imaging characteristics vary between systems in different hospitals, which could lead to variability in (123)FP-CIT classifier output. Investigations are needed to measure this variability to establish whether classification software is likely to be successful outside of the hospital in which it was developed. Without performing such tests it is unlikely that classification / CADx tools would be turned into commercial, clinical products.

With so little previous consideration given to the challenges of clinical translation in the literature, there is little available guidance on how clinical heterogeneity might be addressed. Clearly, repeated scans of patient uptake patterns, under different acquisition conditions are likely to be required. Performing these tests using phantoms rather than real patients is likely to be more viable (due to financial and ethical considerations). However, as shown in chapter 6 there isn’t yet a suitable phantom technology for this task.

Thus, the main focus of the second half of this thesis was on creating novel technologies and strategies that would facilitate investigations into the effects of different acquisition conditions on classifier / CADx performance.

The new objectives for chapters 6 and 7 were to:

- A) Examine and develop phantom technology to provide a toolset that can be adapted to simulate a range of realistic (I123)FP-CIT image appearances.
- B): Use the toolset to demonstrate the influence of heterogeneity by:
    - Analysing and prioritising the individual imaging parameters that may affect classification software performance.
    - Performing sensitivity tests to measure the impact of different imaging parameters on developed classification tools

Therefore, the targets for the remaining research effort were both ambitious, going beyond the scope of the original research question and far beyond the vast majority of machine learning investigations, but they were also necessary given the huge translation challenges that remained.

Chapter 6 provided results from development on a new (I123)FP-CIT phantom based on sub-resolution sandwich phantom technology. A series of investigations demonstrated that this flexible method of phantom manufacture was practical, controllable and repeatable. A fully assembled phantom based on an idealised anatomical template produced image features that were reflective of a cohort of patients.

A strategy for prioritising and selecting image acquisition parameters for sensitivity tests was developed by examining relevant guidelines, and considering the potential for control in the clinical environment. The two highest priority factors were found to be: camera-collimator design and non-standard positioning.

Sensitivity analysis was conducted according to these priority areas, using the idealised SSP anatomical template. For the single classification algorithm that was tested it was shown that use of different image acquisition equipment, from a different manufacturer, can have a substantial impact on algorithm output. This would be likely to preclude the algorithm from being widely used. The impact from using non-standard positioning was found to be smaller.

Although only the simplest of the developed algorithms was tested (that based on SBR features), the investigations demonstrated the suitability of developed processes and technology for assessing classification software. The same methodology could in future be applied to any (I123)FP-CIT classification algorithm once more patient specific anatomical templates had been optimised.

Thus, objectives A, B (part 1 and 2) were largely completed, providing a basis for further work to address the remaining translation gap.

In summary, this thesis has contributed several novel and important findings to the literature:

- Direct, comprehensive comparison between semi-quantification and machine learning tools for classification of (I123)FP-CIT images, demonstrating the superiority of machine learning algorithms
- Evaluation of machine learning software in a clinical (I123)FP-CIT reporting scenario, showing the positive impact on inter-reporter reliability and accuracy, and the ability of machine learning software to match human performance
- Distillation and analysis of the wider barriers to adoption for all machine learning classification tools, that extend far beyond the scope of most machine learning studies
- Development and demonstration of a new, flexible, controllable and repeatable phantom technology, facilitating phantom-based sensitivity analysis of (I123)FP-CIT classification software (which was not previously possible using the Alderson system)
- Creation of a new prioritisation strategy for investigating the impact of different acquisition conditions
- Sensitivity analysis results, demonstrating that differences in camera equipment and acquisition protocols can have a substantial impact on machine learning classification software performance

8.2 Future work

Although a lot has been achieved in this work, there remains a number of avenues that deserve further attention, particularly in relation to the remaining translation gap, which I believe is the most significant problem for machine learning in radiology. The next steps set out below seek to address some of the more immediate questions and issues that have arisen following investigations of automated (I123)FP-CIT classification tools. These ideas

for future work consider both the technologies specifically developed in this thesis but also the wider picture related to translation of machine learning technology.

8.2.1 Mapping out a pathway from initial research to clinical adoption

This work has highlighted the barriers that would prevent immediate widespread adoption of the developed algorithms in clinic. However, only one of these barriers (heterogeneity of the clinical environment) has been partially addressed so far. The necessary steps required to navigate all the other identified hurdles are not clear. What is needed, ideally, is a blueprint for how researchers and other stakeholders could create a clinically successful automated classification or CADx tool. To maximise impact, such a model should ideally be relevant to any application. As a first step towards meeting these goals a workflow diagram was developed, which maps a pathway from initial research idea to clinical adoption for CADx systems. This draft document is introduced below.

In order to create a model for algorithm development, a specific endpoint needs to be defined. One possible choice would be to target inclusion within current clinical guidelines for the specific disease area such that clinical departments wishing to achieve accreditation, or to demonstrate high quality patient care, are incentivised to adopt the technology. However, with numerous national bodies producing clinical guidelines for different areas in medicine, most of which do not have clear criteria for accepting protocols or technologies into guidelines, this would be impractical for creating a generally applicable development model.

A more well-defined endpoint would be achieving NICE approval through either the Medical Technologies Evaluation Programme (MTEP) or the Diagnostic Assessment Programme (DAP). These programmes aim to accelerate the adoption of technologies which have the potential to improve patient outcomes, reduce costs or provide benefits to the healthcare system. Although approval through these routes does not guarantee widespread uptake, it does provide a seal of approval that is likely to dramatically increase pressure on hospitals to invest in the technology.

NICE approval has a list of requirements and assessment processes as set out in relevant guidance documents (135), which provides a basis for creating a pathway to adoption. NICE evaluation covers device regulations, clinical evidence and health economics. Evaluation panels also receive input from multiple different stakeholders. Therefore, many of the barriers to clinical adoption already identified in chapter 5 are also considered as part of

NICE approval procedures. Much of the work undertaken in this thesis (particularly in chapters 2-4) could contribute towards the evidence based requirements of these processes.

Based on the requirements of NICE approval in the context of CAD I produced a peer reviewed publication (5), setting out the main considerations for any new algorithm aiming for clinical adoption. Using the NICE requirements as a target, I also produced a workflow diagram, mapping the whole development pathway, from initial research idea to NICE submission (see Figure 8-1). Although this has not yet been subject to peer review, it could potentially be used by other researchers in future to help ensure that future CAD projects are conducted with clinical translation in mind.

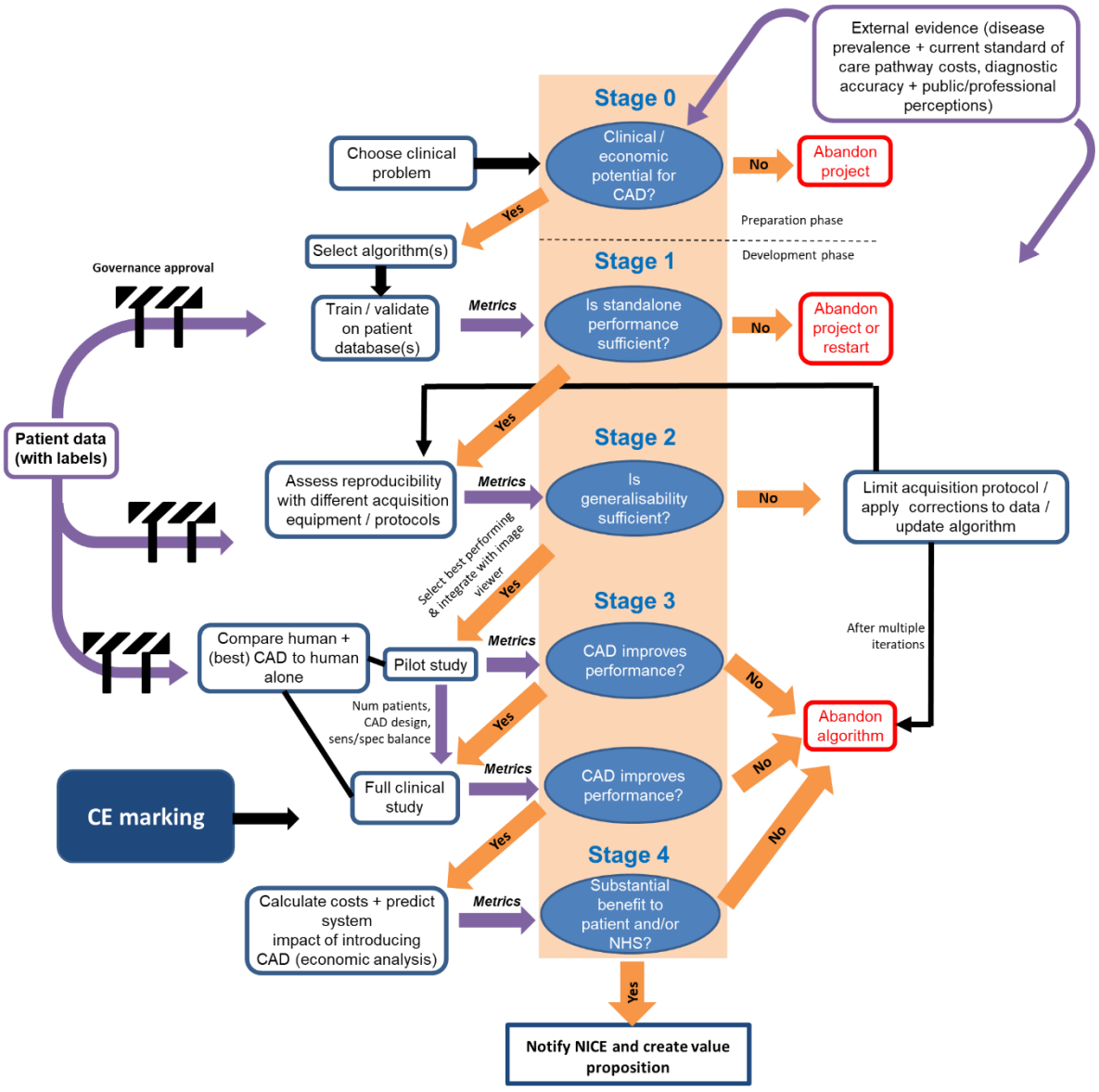


Figure 8-1 CADx development workflow

The workflow is designed for both CADe and CADx applications and is linear, suggesting that steps be taken one after another. Five distinct stages are identified, with each generating additional evidence towards the final goal. Given that it is difficult to know a-priori which technology will work best, it is assumed that multiple possible algorithms will be selected to solve the clinical problem in the early stages. However, only one algorithm will be selected to progress through clinical and economic evaluations (due to the high cost of performing these tasks). The workflow was designed such that early stages are less resource intensive, enabling initial investigations of potential software solutions with minimal outlay. Due to the significant demands of the NICE approval it is anticipated that very few CAD systems would progress through every stage without being rejected.

At each stage a decision must be made as to the sufficiency of performance given the metrics generated from different tests. Should performance be deemed unacceptable the algorithm is either abandoned or updated. Such are the requirements from NICE with regards to compelling empirical evidence, the first stage of the workflow (the preparation phase) involves an assessment of the potential for CAD to have a clinical or economic impact, the implication being that projects should only be undertaken if there is the potential for significant improvements to the current status quo. If standard reporting is quick, inexpensive and effective then there is little point investing time in creating an automated classifier. This strict, impact-focused approach contrasts strongly with much of the current machine literature, where studies are often driven by the availability of data, rather than the potential for improvements to care or efficiency.

Stages 1 and 2 of the workflow involve assessments of standalone performance and analysis of the ability of the algorithms to cope with data derived from different (but realistic) acquisition scenarios. Although these procedures are not specifically referenced in NICE processes they are, as shown by previous discussions related to heterogeneity of the clinical environment, necessary for understanding baseline performance. If such results are not at a sufficiently high level (for example they are lower than human performance unaided) then it is unlikely that clinical investigations will achieve the outcomes necessary to encourage clinicians to adopt the technology. Stage 2 permits adjustments to the algorithm to be carried out (assuming they do not adversely impact on standalone performance) or other compensatory mechanisms to be instigated, including additional pre-processing of data, in order that performance figures are maintained at a high level with the widest possible scope of application.

Stage 3 of the workflow involves the collection of clinical study data through comparison of the performance of standard care procedures (usually radiologists performing visual analysis) and radiologists working with CAD support. This process has been subdivided into a pilot study and a main study, in a similar way to chapter 4. Ideally the main study should be conducted as part of a multicentre trial, whereby a wide range of radiologists can be included in the results, such that results are representative of general clinical usage.

Stage 4 involves analysis of the direct costs of the CAD intervention (including costs of software licences, infrastructure, maintenance, staffing and training) as well as the indirect costs that would result in relation to changes to the patient pathway (for example the additional number of secondary care consultations that may result per patient). A health economic analysis is then conducted to predict the overall system impact from introduction of CAD. It is assumed here that results from clinical evaluations can be extrapolated to estimate the effects on the patient pathway. The uncertainties of this approach are likely to be higher than for a full clinical trial which examined patient outcomes over the longer term following diagnostic assessment with and without CAD. However, the costs to the developer are likely to be much lower using this simplified technique.

Through each stage of the workflow it is assumed that evaluation methodologies are chosen such that uncertainties are minimised. For example, that tests conducted with radiologists use randomly selected patient images, thus minimising recall bias, and that investigations of standalone performance use large databases of images, covering most expected image appearances. Methodological errors that are likely to bias estimates of algorithm performance, such as those highlighted in recent literature (93,94), should be avoided. Maintaining high standards in data collection is crucial, as highlighted by a recent review of the MTEP programme, which showed that the 3 main reasons for technologies failing to progress to guidance development were a lack of evidence, insufficient/uncertain benefit to the NHS and insufficient/uncertain benefit to the patient (136). In addition to maintaining high standards in methodology, it is assumed that patient data used is always anonymised such that data protection and ethical issues are reduced.

The workflow does not make assumptions about who should be responsible for each stage of the process. However, it is assumed that the individuals or groups working on the project are aiming towards the same goal. It is possible that some aspects of the workflow could be carried out by external parties, or could be derived from previous literature. For example, the selection and testing of different algorithms in stage 1 could be largely derived from the

results of ‘grand challenges’ such as those run each year at the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference. Here, databases of images are made freely available to researchers from around the world. Participants submit their created algorithms for evaluation on a (usually held back) test set. Algorithm details are then published and each is ranked according to chosen metrics. Thus, consideration of these results (if available) could be an efficient and cost effective way of comparing the performance of multiple candidate algorithms.

Regulatory authorisation (particularly CE marking) needs to occur once the system design is finalised, which is likely to be after the clinical evaluations are completed. However, compliance processes could instead be completed after the system impact of CAD has been predicted. Thus, the CE marking stage is not strictly tied to either stage 3 or 4 in the workflow. It is also assumed that research publications would be produced at each stage of the workflow, cumulatively adding to the evidence base for the technology.

This model for algorithm development represents my first attempt at trying to create a plausible pathway to help guide researchers on the tortuous path towards clinical translation. Bourne out of a frustration with the current disconnect between the rapidly expanding, exciting machine learning research arena, and the stubbornly low clinical use of machine learning technology, this model undoubtedly has limitations. For instance, the workflow provides no guidance on what type of expertise is required at each stage in order to meet the multidisciplinary challenges presented. Furthermore, there is no guidance on specific processes required in order to meet regulatory requirements (i.e. CE marking), and no consideration of what might be done if prevailing professional and public opinions were largely negative. The model needs to be presented to and critiqued by other experts, and possibly altered, before it can be put forward as guidance for researchers and developers. This would be a priority for future work. However, it undoubtedly offers a useful starting point for driving conversation in the machine learning community towards a more clinic-focused approach.

8.2.2 Development of realistic anatomical templates

The ability to conduct sensitivity tests for more complex (I123)FP-CIT machine learning algorithms is currently hampered by a lack of realistic anatomical template patterns for use with the SSP production method. Given that the most successful machine learning

algorithms developed in this thesis cannot yet be tested with SSP technology, this problem needs to be addressed.

One possible route to creating more realistic templates for (I123)FP-CIT imaging would be to select images from patients who had already undergone both MRI acquisitions and (I123)FP-CIT SPECT within a short time window. The shape of the anatomical template could be defined by segmenting the MRI scan. After setting an assumed count density pattern, printing and scanning, the resultant reconstructed projections could be compared to the patient's real gamma camera images. By taking the difference between the simulated uptake pattern and the real image, on a voxel-by-voxel basis, an update can be made to the design template to iteratively bring it closer to the patient's underlying uptake pattern. The process could be repeated as many times as required. This methodology was adopted by Holmes et al. for the creation of realistic brain perfusion scans (119). Once such templates have been created, sensitivity tests can be conducted for the PCA-based machine learning algorithm used in chapter 4, to add further evidence of its suitability for clinical use. Algorithm adaptations could then be investigated if required.

8.2.3 Evaluation of a (I123)FP-CIT screening tool

As suggested following the clinical study in chapter 4, the performance of developed classification technology was so high (at or above the level of experienced reporters) that the algorithm could perhaps be better exploited as an independent diagnostic device than a CADx tool. It is likely that using such technology as a screening tool (removing the most obviously normal cases from the reporting list) would be the lowest risk and most acceptable way to perform automated, independent image analysis, at least initially. Indeed the US Food and Drug Administration (FDA) recently granted regulatory approval to the first machine-learning based medical image analysis tool which works independently from humans (named IDx-DR), designed to screen out retinal images showing mild or no disease (137). However, no clinical studies have yet been carried out according to this reporting scenario for (I123)FP-CIT imaging and so the potential benefits are not yet clear.

A number of fundamental questions need to be addressed, in particular, what probability value should be used as a cut-off for deciding on whether an image should be shown to a radiologist of not? This could be investigated, in the first instance, through retrospective studies. The available clinical data already extracted from the archives at Sheffield could be split into two halves, the first used for retraining a classification algorithm and the second half

for determining a cut-off in algorithm output that is able to achieve 100% specificity (with a suitable, additional error margin). This could be calculated, for example, from Receiver Operator Curve analysis. The algorithm could then be used in a prospective study, whereby it is applied to all new (I123)FP-CIT scans acquired in the department, working alongside conventional reporting practices. The clinical reports could be compared to the algorithm output to confirm whether the selected cut-off was sufficient for ensuring that all abnormal cases continued to be displayed to reporters, and for quantifying the proportion of studies that could potentially be removed from reporting lists. Such information would be vital for evidencing potential improvements in efficiency and for justifying further algorithm development.

8.2.4 Understanding perceptions of machine learning classification technology

If automated classification software is to achieve widespread clinical adoption, particularly if used as an independent screening system, health professionals and patients need to be accepting of this new approach to radiological reporting. In particular, NICE committees which evaluate new medical technologies (such as the Medical Technologies Advisory Committee), take evidence from patient groups as well as relevant clinicians (136), whose opinions of the technology are likely to have a strong bearing on the final decision. Indeed, the developed pathway to translation (see section 8.2.1) explicitly includes evidence of public and professional opinions.

Although the opinions of such individuals cannot be predicted it is useful to understand whether there are commonly held beliefs which may hamper clinical uptake. The use of machine learning and artificial intelligence in all aspects of life is a topic that is frequently visited in media reports. Such coverage is often negative. For example, the recent data sharing agreement between the Royal Free Hospital in London and Google Deepmind for development of machine learning algorithms was severely criticised in multiple publications. The Information Commissioner's Office deemed that the clinical trial had failed to comply with data protection law (138). It could be that such reports may naturally cause people to be more sceptical about machine learning in general, no matter what the empirical evidence that is presented.

A separate Ipsos MORI survey commissioned on behalf of the Wellcome Trust (139) was conducted to assess public views on commercial access to health data. This is of relevance here as machine learning tools rely on patient data for adequate training and testing. It was

found that the general public and health professionals are often concerned and sceptical about the use of health data by private companies, in part due to lack of transparency. A significant minority think that health data should never be shared with commercial institutions under any circumstances. This could have implications for the work presented in this thesis if it was ever decided that software should be developed by a private company.

The perils of not fully understanding the opinions of stakeholders in relation to digital technology are shown by the recent, high-profile failure of the care.data initiative in the NHS, which was designed to create a central database of primary care records. One of the key reasons identified for the failure was the lack of adequate information provided to the public on how their data would be used (140).

As part of the promotion strategy for (I123)FP-CIT classification software, in the push towards clinical translation, it would therefore be advantageous to seek out clinical and patient opinions of automatic diagnosis in relation to Parkinson's Disease. These could be gathered through structured focus groups.

A pilot study was conducted to provide initial data on public perceptions, and to assess the suitability of developed questions for a larger study. Four individuals from the Sheffield Parkinson's Disease society group volunteered to attend the focus group. Initial questions presented to the volunteers assessed current understanding of technical terms such as artificial intelligence and machine learning. Different scenarios were described, highlighting the use of computer software to augment or replace human work. These were discussed as a group, with questions posed to elicit views on acceptability.

The main themes identified from discussions with the volunteers were:

- Humanity in medicine (and elsewhere) should always been maintained i.e. there always needs to be some human interaction between patients and healthcare experts
- Machines and software should augment what a human does, not replace or downgrade human work
- Trust is important. If a patient trusts a doctor then the fact that CADx or classification software is used to inform their PD diagnosis is a minor consideration.

- Acceptance of automated diagnostic software is likely to increase as it is used more routinely.
- Some patients may expect the latest technology to be used on them. These individuals may be disappointed if machine learning wasn't used to inform their PD diagnosis.

The findings were somewhat contradictory in that there was a universal desire to maintain and protect human functions and skills, but that use of automated diagnostic algorithms could be tolerated or even desired so long as the patient's main clinical contact was with a human that they trusted. This suggests that marketing and promotion of any classification tool needs to be conducted carefully, emphasising the benefits of algorithms to clinicians and the patient. It may also be easier to induce positive opinions of classification tools if used as part of a CADx system, rather than as a screening tool where human input is removed.

A larger scale qualitative study is needed to confirm these findings and to further explore where the boundary of acceptability lies between human and software based diagnosis. Ideally this would include interviews with radiologists and neurologists too, whose work would be directly affected by new (I123)FP-CIT classification software, and whose opinions are likely to have a substantial impact on whether such technology will flourish or not. Findings from such work would also be valuable to researchers and developers working in other areas of machine learning in medicine.

The four suggested areas for further work described above need to be prioritised. Of these projects, perhaps the highest priority should be given to creation of an accepted development workflow, mapping the pathway from initial CAD research to clinical translation, as such a document could have a big impact. Also of high priority is the need to develop realistic anatomical templates for the newly developed phantom. Without such work the momentum behind the work completed in early chapters of this thesis would be lost. Evaluating a screening tool for (I123)FP-CIT and performing more in-depth qualitative analysis of public / professional perceptions are topics that are arguably less urgent, particularly as the latter relies upon a viable clinical tool being in place first.

8.3 Conclusion

At the beginning this thesis focused on assistive reporting technology in (I123)FP-CIT imaging. The main research question asked how effective was a CADx tool, based on

established machine learning algorithms, for assisted (I123)FP-CIT image reporting. Validation test results showed for the first time that machine learning tools outperform a wide range of semi-quantification approaches in terms of binary classification performance, and that a CADx tool built on such algorithms offered increased consistency between reporters and increased accuracy. Thus, in answer to the research question, machine learning for CADx in (I123)FP-CIT imaging proved to be highly effective.

However, following a realisation that the path to clinical translation would be highly challenging, both for this application and others in medical imaging, there was a subsequent shift in focus towards addressing translation barriers. Driven by a desire to prevent developed technologies from being forever confined to the literature, wider questions related to heterogeneity of the clinical environment were considered. As a result, new phantom technologies and new strategies were created, which facilitated sensitivity testing.

The suggested future work also focuses on wider considerations in relation to clinical translation, for both (i123)FP-CIT imaging and other machine learning applications in radiology. I hope that other researchers will also come to the realisation that in order for machine learning to make an impact in clinic, these areas need to be more of a priority.

9 References

1. Taylor JC, Fenner JW. Comparison of machine learning and semi-quantification algorithms for (I123)FP-CIT classification: the beginning of the end for semi-quantification? EJNMMI Phys. 2017 Dec;4(1):29.

2. Taylor JC, Romanowski C, Lorenz E, Lo C, Bandmann O, Fenner J. Computer-aided diagnosis for (123I)FP-CIT imaging: impact on clinical reporting. EJNMMI Res [Internet]. 2018 May 8;8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5940985/

3. Taylor JC, Vennart N, Negus I, Holmes R, Bandmann O, Lo C, et al. The subresolution DaTSCAN phantom: a cost-effective, flexible alternative to traditional phantom technology. Nucl Med Commun. 2018 Mar;39(3):268–75.

4. Taylor J, Fenner J. The challenge of clinical adoption—the insurmountable obstacle that will stop machine learning? BJR|Open. 2019 Jan;1(1):20180017.

5. Taylor J, Fenner J. Clinical adoption of CAD: exploration of the barriers to translation through an example application. 20th Conf Med Image Underst Anal Miua 2016. 2016;90:93–8.

6. Darcourt J, Booij J, Tatsch K, Varrone A, Borght TV, Kapucu ÖL, et al. EANM procedure guidelines for brain neurotransmission SPECT using 123I-labelled dopamine transporter ligands, version 2. Eur J Nucl Med Mol Imaging. 2010 Feb 1;37(2):443–50.

7. Gibb WRG, Lees AJ. A comparison of clinical and pathological features of young- and old-onset Parkinson's disease. Neurology. 1988 Sep 1;38(9):1402–1402.

8. Williams DR, Litvan I. Parkinsonian Syndromes. Contin Lifelong Learn Neurol. 2013 Oct;19(5 Movement Disorders):1189–212.

9. Wickremaratchi MM, Perera D, O'Loghlen C, Sastry D, Morgan E, Jones A, et al. Prevalence and age of onset of Parkinson's disease in Cardiff: a community based cross sectional study and meta-analysis. J Neurol Neurosurg Psychiatry. 2009 Jul;80(7):805–7.

10. Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. J Neurol Neurosurg Psychiatry. 1992 Mar;55(3):181–4.

11. Rizzo G, Copetti M, Arcuti S, Martino D, Fontana A, Logroscino G. Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis. Neurology. 2016 Feb 9;86(6):566–76.

12. Joutsa J, Gardberg M, Röyttä M, Kaasinen V. Diagnostic accuracy of parkinsonism syndromes by general neurologists. Parkinsonism Relat Disord. 2014 Aug;20(8):840–4.

13. Hughes AJ, Daniel SE, Ben-Shlomo Y, Lees AJ. The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service. Brain J Neurol. 2002 Apr;125(Pt 4):861–70.

14. Marshall VL, Reininger CB, Marquardt M, Patterson J, Hadley DM, Oertel WH, et al. Parkinson's disease is overdiagnosed clinically at baseline in diagnostically uncertain cases: A 3- year European multicenter study with repeat [123 I]FP-CIT SPECT. Mov Disord. 2009;24(4):500–8.

15. Nelson PT, Jicha GA, Kryscio RJ, Abner EL, Schmitt FA, Cooper G, et al. Low sensitivity in clinical diagnoses of dementia with Lewy bodies. J Neurol. 2010 Mar;257(3):359–66.

16. Adler CH, Beach TG, Hentz JG, Shill HA, Caviness JN, Driver-Dunckley E, et al. Low clinical diagnostic accuracy of early vs advanced Parkinson disease. Neurology. 2014 Jul 29;83(5):406–12.

17. Hauser RA, Grosset DG. [123I]FP-CIT (DaTscan) SPECT brain imaging in patients with suspected parkinsonian syndromes. J Neuroimaging Off J Am Soc Neuroimaging. 2012 Jul;22(3):225–30.

18. Vann Jones SA, O'Brien JT. The prevalence and incidence of dementia with Lewy bodies: a systematic review of population and clinical studies. Psychol Med. 2014 Mar;44(4):673–83.

19. Lundkvist C, Halldin C, Swahn C-G, Hall H, Karlsson P, Nakashima Y, et al. [O-methyl-11C]β-CIT-FP, a potential radioligand for quantitation of the dopamine transporter:

Preparation, autoradiography, metabolite studies, and positron emission tomography examinations. Nucl Med Biol. 1995 Oct 1;22(7):905–13.

20. Günther I, Hall H, Halldin C, Swahn C-G, Farde L, Sedvall G. [125I]β-CIT-FE and [125I]β-CIT-FP are superior to [125I]β-CIT for dopamine transporter visualization: Autoradiographic evaluation in the human brain. Nucl Med Biol. 1997 Oct 1;24(7):629–34.

21. O'Brien JT, Oertel WH, McKeith IG, Grosset DG, Walker Z, Tatsch K, et al. Is ioflupane I123 injection diagnostically effective in patients with movement disorders and dementia? Pooled analysis of four clinical trials. BMJ Open. 2014 Jul 3;4(7):e005122.

22. Sharifi S, Nederveen AJ, Booij J, van Rootselaar A-F. Neuroimaging essentials in essential tremor: A systematic review. NeuroImage Clin. 2014 May 9;5:217–31.

23. O'Brien JT, Colloby S, Fenwick J, Williams ED, Firbank M, Burn D, et al. Dopamine transporter loss visualized with FP-CIT SPECT in the differential diagnosis of dementia with Lewy bodies. Arch Neurol. 2004 Jun;61(6):919–25.

24. Shin H-W, Chung SJ. Drug-Induced Parkinsonism. J Clin Neurol Seoul Korea. 2012 Mar;8(1):15–21.

25. Bouwmans AEP, Vlaar AMM, Mess WH, Kessels A, Weber WEJ. Specificity and sensitivity of transcranial sonography of the substantia nigra in the diagnosis of Parkinson's disease: prospective cohort study in 196 patients. BMJ Open. 2013;3(4).

26. Brigo F, Matinella A, Erro R, Tinazzi M. [^{123}I]FP-CIT SPECT (DaTSCAN) may be a useful tool to differentiate between Parkinson's disease and vascular or drug-induced parkinsonisms: a meta-analysis. Eur J Neurol. 2014 Nov;21(11):1369-e90.

27. Cilia R, Rossi C, Frosini D, Volterrani D, Siri C, Pagni C, et al. Dopamine Transporter SPECT Imaging in Corticobasal Syndrome. PLOS ONE. 2011 May 2;6(5):e18301.

28. Latoo J, Jan F. Dementia with Lewy bodies: clinical review. Br J Med Pract. 2008;1(1):10–4.

29. Cummings JL, Henchcliffe C, Schaier S, Simuni T, Waxman A, Kemp P. The role of dopaminergic imaging in patients with symptoms of dopaminergic system neurodegeneration. Brain J Neurol. 2011 Nov;134(Pt 11):3146–66.

30. Siepel FJ, Rongve A, Buter TC, Beyer MK, Ballard CG, Booij J, et al. (123I)FP-CIT SPECT in suspected dementia with Lewy bodies: a longitudinal case study. BMJ Open. 2013 Jan 1;3(4):e002642.

31. Thomas AJ, Attems J, Colloby SJ, O'Brien JT, McKeith I, Walker R, et al. Autopsy validation of 123I-FP-CIT dopaminergic neuroimaging for the diagnosis of DLB. Neurology. 2017 Jan 17;88(3):276–83.

32. Djang DSW, Janssen MJR, Bohnen N, Booij J, Henderson TA, Herholz K, et al. SNM practice guideline for dopamine transporter imaging with 123I-ioflupane SPECT 1.0. J Nucl Med. 2012 Jan;53(1):154–63.

33. de la Fuente-Fernández R. Role of DaTSCAN and clinical diagnosis in Parkinson disease. Neurology. 2012 Mar 6;78(10):696–701.

34. Benamer TS, Patterson J, Grosset DG, Booij J, de Bruin K, van Royen E, et al. Accurate differentiation of parkinsonism and essential tremor using visual assessment of [123I]-FP-CIT SPECT imaging: the [123I]-FP-CIT study group. Mov Disord Off J Mov Disord Soc. 2000 May;15(3):503–10.

35. Tolosa E, Borght TV, Moreno E, DaTSCAN Clinically Uncertain Parkinsonian Syndromes Study Group. Accuracy of DaTSCAN (123I-Ioflupane) SPECT in diagnosis of patients with clinically uncertain parkinsonism: 2-year follow-up of an open-label study. Mov Disord Off J Mov Disord Soc. 2007 Dec;22(16):2346–51.

36. Kemp PM, Clyde K, Holmes C. Impact of 123I-FP-CIT (DaTSCAN) SPECT on the diagnosis and management of patients with dementia with Lewy bodies: a retrospective study. Nucl Med Commun. 2011 Apr;32(4):298–302.

37. Neilly B. 2015 BNMS National DaTSCAN Audit [Internet]. British Nuclear Medicine Society; 2017 [cited 2017 Aug 18]. Available from: https://www.bnms.org.uk/images/2015_BNMS_National_DaTSCAN_Audit_NEW.pdf

38. Bernheimer H, Birkmayer W, Hornykiewicz O, Jellinger K, Seitelberger F. Brain dopamine and the syndromes of Parkinson and Huntington Clinical, morphological and neurochemical correlations. J Neurol Sci. 1973 Dec 1;20(4):415–55.

39. Tossici-Bolt L, Hoffmann SMA, Kemp PM, Mehta RL, Fleming JS. Quantification of [123I]FP-CIT SPECT brain images: an accurate technique for measurement of the specific binding ratio. Eur J Nucl Med Mol Imaging. 2006 Dec;33(12):1491–9.

40. Varrone A, Dickson JC, Tossici-Bolt L, Sera T, Asenbaum S, Booij J, et al. European multicentre database of healthy controls for [123I]FP-CIT SPECT (ENC-DAT): age-related effects, gender differences and evaluation of different methods of analysis. Eur J Nucl Med Mol Imaging. 2013 Jan;40(2):213–27.

41. Albert NL, Unterrainer M, Diemling M, Xiong G, Bartenstein P, Koch W, et al. Implementation of the European multicentre database of healthy controls for [(123)I]FP-CIT SPECT increases diagnostic accuracy in patients with clinically uncertain parkinsonian syndromes. Eur J Nucl Med Mol Imaging. 2016 Jul;43(7):1315–22.

42. Ueda J, Yoshimura H, Shimizu K, Hino M, Kohara N. Combined visual and semi-quantitative assessment of (123)I-FP-CIT SPECT for the diagnosis of dopaminergic neurodegenerative diseases. Neurol Sci Off J Ital Neurol Soc Ital Soc Clin Neurophysiol. 2017 Jul;38(7):1187–91.

43. Suárez-Piñera M, Prat ML, Mestre-Fusco A, Fuertes J, Mojal S, Balaguer E. [Interobserver agreement in the visual and semi-quantitative analysis of the 123I-FP-CIT SPECT images in the diagnosis of Parkinsonian syndrome]. Rev Esp Med Nucl. 2011 Aug;30(4):229–35.

44. Booij J, Dubroff J, Pryma D, Yu JQ, Agarwal R, Lakhani P, et al. Diagnostic performance of the visual reading of (123)I-ioflupane SPECT images when assessed with or without quantification in patients with movement disorders or dementia. J Nucl Med. 2017 May 4;

45. Söderlund TA, Dickson JC, Prvulovich E, Ben-Haim S, Kemp P, Booij J, et al. Value of semiquantitative analysis for clinical reporting of 123I-2-β-carbomethoxy-3β-(4-iodophenyl)-N-(3-fluoropropyl)nortropane SPECT studies. J Nucl Med 2013 May;54(5):714–22.

46. Pencharz DR, Hanlon P, Chakravartty R, Navalkissoor S, Quigley A-M, Wagner T. Automated quantification with BRASS reduces equivocal reporting of DaTSCAN (123I-FP-CIT) SPECT studies. Nucl Med Rev Cent East Eur. 2014;17(2):65–9.

47. Dickson JC, Tossici-Bolt L, Sera T, Erlandsson K, Varrone A, Tatsch K, et al. The impact of reconstruction method on the quantification of DaTSCAN images. Eur J Nucl Med Mol Imaging. 2010 Jan;37(1):23–35.

48. Dickson JC, Tossici-Bolt L, Sera T, Booij J, Ziebell M, Morbelli S, et al. The impact of reconstruction and scanner characterisation on the diagnostic capability of a normal database for [123I]FP-CIT SPECT imaging. EJNMMI Res. 2017 Jan 24;7:10.

49. Tossici-Bolt L, Dickson JC, Sera T, Booij J, Asenbaun-Nan S, Bagnara MC, et al. [123I]FP-CIT ENC-DAT normal database: the impact of the reconstruction and quantification methods. EJNMMI Phys [Internet]. 2017 Jan 28 [cited 2017 Oct 3];4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5272851/

50. Morton RJ, Guy MJ, Marshall CA, Clarke EA, Hinton PJ. Variation of DaTSCAN quantification between different gamma camera types. Nucl Med Commun. 2005 Dec;26(12):1131–7.

51. Shen D, Wu G, Suk H-I. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng. 2017 Jun 21;19:221–48.

52. Augimeri A, Cherubini A, Cascini GL, Galea D, Caligiuri ME, Barbagallo G, et al. CADA-computer-aided DaTSCAN analysis. EJNMMI Phys. 2016 Dec;3(1):4.

53. Bhalchandra NA, Prashanth R, Roy SD, Noronha S. Early detection of Parkinson's disease through shape based features from 123I-Ioflupane SPECT imaging. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI). 2015. p. 963–6.

54. Choi H, Ha S, Im HJ, Paek SH, Lee DS. Refining diagnosis of Parkinson's disease with deep learning-based interpretation of dopamine transporter imaging. NeuroImage Clin. 2017 Sep 10;16:586–94.

55. Oliveira FPM, Faria DB, Costa DC, Castelo-Branco M, Tavares JMRS. Extraction, selection and comparison of features for an effective automated computer-aided diagnosis of Parkinson's disease based on [123I]FP-CIT SPECT images. Eur J Nucl Med Mol Imaging. 2017 Dec 23;

56. Oliveira FPM, Castelo-Branco M. Computer-aided diagnosis of Parkinson's disease based on [123 I]FP-CIT SPECT binding potential images, using the voxels-as-features approach and support vector machines. J Neural Eng. 2015;12(2):026008.

57. Prashanth R, Roy SD, Mandal PK, Ghosh S. High-Accuracy Classification of Parkinson's Disease Through Shape Analysis and Surface Fitting in 123I-Ioflupane SPECT Imaging. IEEE J Biomed Health Inform. 2017 May;21(3):794–802.

58. Tagare HD, DeLorenzo C, Chelikani S, Saperstein L, Fulbright RK. Voxel-based logistic analysis of PPMI control and Parkinson's disease DaTscans. NeuroImage. 2017 Feb;152:299–311.

59. Palumbo B, Fravolini ML, Buresta T, Pompili F, Forini N, Nigro P, et al. Diagnostic Accuracy of Parkinson Disease by Support Vector Machine (SVM) Analysis of 123I-FP-CIT Brain SPECT Data. Medicine (Baltimore) [Internet]. 2014 Dec 12;93(27). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4602813/

60. Prashanth R, Dutta Roy S, Mandal PK, Ghosh S. Automatic Classification and Prediction Models for Early Parkinson's Disease Diagnosis from SPECT Imaging. Expert Syst Appl. 2014 Jun;41(7):3333–3342.

61. Martínez-Murcia FJ, Górriz JM, Ramírez J, Illán IA, Puntonet CG. Texture Features Based Detection of Parkinson's Disease on DaTSCAN Images. In: Natural and Artificial Computation in Engineering and Medical Applications [Internet]. Springer, Berlin, Heidelberg; 2013 [cited 2017 Jul 31]. p. 266–77. (Lecture Notes in Computer Science). Available from: https://link.springer.com/chapter/10.1007/978-3-642-38622-0_28

62. Zhang YC, Kagen AC. Machine Learning Interface for Medical Image Analysis. J Digit Imaging [Internet]. 2016 Oct; Available from: ://MEDLINE:27730415

63. Rojas A, GóRriz JM, RamíRez J, IlláN IA, MartíNez-Murcia FJ, Ortiz A, et al. Application of Empirical Mode Decomposition (EMD) on DaTSCAN SPECT Images to Explore Parkinson Disease. Expert Syst Appl. 2013 Jun;40(7):2756–2766.

64. Martinez-Murcia FJ, Górriz JM, Ramírez J, Ortiz A. Convolutional Neural Networks for Neuroimaging in Parkinson's Disease: Is Preprocessing Needed? Int J Neural Syst. 2018 Jul 26;1850035.

65. Towey DJ, Bain PG, Nijran KS. Automatic classification of I-123-FP-CIT (DaTSCAN) SPECT images. Nucl Med Commun. 2011 Aug;32(8):699–707.

66. Segovia F, Gorriz JM, Ramirez J, Alvarez I, Jimenez-Hoyuela JM, Ortega SJ. Improved Parkinsonism diagnosis using a partial least squares based approach. Med Phys. 2012 Jul;39(7):4395–403.

67. Martinez-Murcia FJ, Gorriz JM, Ramirez J, Illan IA, Ortiz A, Parkinson's Progression Markers I. Automatic detection of Parkinsonism using significance measures and component analysis in DaTSCAN imaging. Neurocomputing. 2014 Feb 27;126:58–70.

68. Kim DH, Wit H, Thurston M. Artificial intelligence in the diagnosis of Parkinson's disease from ioflupane-123 single-photon emission computed tomography dopamine transporter scans using transfer learning. Nucl Med Commun. 2018;39(10):887–93.

69. Illán IA, Górriz JM, Ramírez J, Segovia F, Jiménez-Hoyuela JM, Ortega Lozano SJ. Automatic assistance to Parkinsonˈs disease diagnosis in DaTSCAN SPECT imaging. Med Phys. 2012 Oct 1;39(10):5971–80.

70. Palumbo B, Fravolini ML, Nuvoli S, Spanu A, Paulus KS, Schillaci O, et al. Comparison of two neural network classifiers in the differential diagnosis of essential tremor and Parkinson's disease by (123)I-FP-CIT brain SPECT. Eur J Nucl Med Mol Imaging. 2010 Nov;37(11):2146–53.

71. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006 Feb 23;7:91.

72. Kohavi R. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2 [Internet]. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1995. p. 1137–1143. (IJCAI'95). Available from: http://dl.acm.org/citation.cfm?id=1643031.1643047

73. Kim J-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Comput Stat Data Anal. 2009 Sep 1;53(11):3735–45.

74. The Parkinson Progression Marker Intiative. Protocol Amendment 5 [Internet]. 2012 [cited 2017 Aug 18]. Available from: http://www.ppmi-info.org/wp-content/uploads/2013/02/PPMI-Protocol-AM5-Final-27Nov2012v6-2.pdf

75. The Parkinson Progression Marker Intiative. Imaging Technical Operations Manual [Internet]. The Parkinson Progression Marker Initiative; 2010 [cited 2017 Aug 18].

Available from: http://www.ppmi-info.org/wp-content/uploads/2010/07/Imaging-Manual.pdf

76. Wisniewski G, Seibyl J, Marek K. DatScan SPECT Image Processing Methods for Calculation of Striatal Binding Ratio. Parkinson's Progression Markers initiative; 2013.

77. Jolliffe IT. Principal Component Analysis. Springer Science & Business Media; 2002. 524 p.

78. Stuhler E, Merhof D. Principal Component Analysis Applied to SPECT and PET Data of Dementia Patients - A Review. In: Sanguansat P, editor. Principal Component Analysis - Multidisciplinary Applications. Intech; 2012.

79. Schölkopf B, Smola AJ. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press; 2002. 658 p.

80. Yang ZR. Biological applications of support vector machines. Brief Bioinform. 2004 Dec;5(4):328–38.

81. Mountrakis G, Im J, Ogole C. Support vector machines in remote sensing: A review. ISPRS J Photogramm Remote Sens. 2011 May 1;66(3):247–59.

82. E B, G S. Support vector machine applications in bioinformatics. Appl Bioinformatics. 2003;2(2):67–77.

83. Raghavendra. N S, Deka PC. Support vector machine applications in the field of hydrology: A review. Appl Soft Comput. 2014 Jun 1;19:372–86.

84. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995 Sep 1;20(3):273–97.

85. Chang C-C, Lin C-J. LIBSVM: A Library for Support Vector Machines. ACM Trans Intell Syst Technol. 2011 May;2(3):27:1–27:27.

86. Chang LT. A Method for Attenuation Correction in Radionuclide Computed Tomography. IEEE Trans Nucl Sci. 1978 Feb;25(1):638–43.

87. Oliveira FPM, Tavares JMRS. Medical image registration: a review. Comput Methods Biomech Biomed Engin. 2014 Jan 25;17(2):73–93.

88. Barber DC, Hose DR. Automatic segmentation of medical images using image registration: diagnostic and simulation applications. J Med Eng Technol. 2005 Apr;29(2):53–63.

89. Redgate S, Barber DC, Al-Mohammad A, Tindale WB. Using a registration-based motion correction algorithm to correct for respiratory motion during myocardial perfusion imaging. Nucl Med Commun. 2013 Aug;34(8):787–95.

90. Ireland RH, Dyker KE, Barber DC, Wood SM, Hanney MB, Tindale WB, et al. Nonrigid image registration for head and neck cancer radiotherapy treatment planning with PET/CT. Int J Radiat Oncol Biol Phys. 2007 Jul 1;68(3):952–7.

91. Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology. 1945 Jul 1;26(3):297–302.

92. Sørensen T. A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons. I kommission hos E. Munksgaard; 1948. 34 p.

93. Aerts HJ. Data Science in Radiology: A Path Forward. Clin Cancer Res Off J Am Assoc Cancer Res. 2017 Nov 2;

94. Chalkidou A, O'Doherty MJ, Marsden PK. False Discovery Rates in PET and CT Studies with Texture Features: A Systematic Review. PLOS ONE. 2015 May 4;10(5):e0124165.

95. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinforma Oxf Engl. 2005 Aug 1;21(15):3301–7.

96. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media; 2001. 560 p.

97. Bergstra J, Bengio Y. Random Search for Hyper-parameter Optimization. J Mach Learn Res. 2012 Feb;13(1):281–305.

98. Platt JC. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: Advances in Large Margin Classifiers. MIT Press; 1999. p. 61–74.

99. Eadie LH, Taylor P, Gibson AP. Recommendations for research design and reporting in computer-assisted diagnosis to facilitate meta-analysis. J Biomed Inform. 2012;45(2):390–7.

100. Eadie LH, Gibson AP, Taylor P. A systematic review of computer- assisted diagnosis in diagnostic cancer imaging. Eur J Radiol. 2012;81(1):e70–6.

101. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. Clin Rehabil. 1998 Jun;12(3):187–99.

102. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016 Jun;15(2):155–63.

103. Bron EE, Smits M, van der Flier WM, Vrenken H, Barkhof F, Scheltens P, et al. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. NeuroImage. 2015 May 1;111:562–79.

104. Giger ML, Karssemeijer N, Armato SG. Guest editorial computer- aided diagnosis in medical imaging. Med Imaging IEEE Trans On. 2001;20(12):1205–8.

105. Wang SJ, Summers RM. Machine learning and radiology. Med Image Anal. 2012;16(5):933–51.

106. Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty L, Ganott MA, et al. The ‘ Laboratory’ effect: Comparing radiologists’ performance and variability during prospective clinical and laboratory mammography interpretations. Radiology. 2008;249(1):47–53.

107. Nishikawa RM, Pesce LL. Fundamental limitations in developing computer-aided detection for mammography. Nucl Inst Methods Phys Res A. 2011;648:S251–4.

108. Castellino RA. Computer aided detection (CAD): an overview. Cancer Imaging. 2005 Aug 23;5(1):17–9.

109. Kostkova P, Brewer H, de Lusignan S, Fottrell E, Goldacre B, Hart G, et al. Who Owns the Data? Open Data for Healthcare. Front Public Health [Internet]. 2016 Feb 17 [cited 2017 Nov 29];4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4756607/

110. Collins B. Adoption and spread of innovation in the NHS [Internet]. 2018 Jan [cited 2018 Jan 22]. Available from: https://www.kingsfund.org.uk/publications/innovation-nhs

111. Kotter JP. Leading Change. Harvard Business School Press; 1996. 206 p.

112. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology. 2018 Jan 8;171920.

113. Van Ginneken B, Schaefer-Prokop CM, Prokop M. Computer- aided diagnosis: How to move from the laboratory to the clinic. Radiology. 2011;261(3):719–32.

114. Ross C, Swetlitz I. IBM pitched Watson as a revolution in cancer care. It's nowhere close. Stat [Internet]. 2017 Sep 5 [cited 2018 Oct 3]; Available from: https://www.statnews.com/2017/09/05/watson-ibm-cancer/

115. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017 Nov 14 [cited 2018 Aug 6]; Available from: https://arxiv.org/abs/1711.05225

116. Sousa MAZ, Matheus BRN, Schiabel H. Development of a structured breast phantom for evaluating CADe/Dx schemes applied on 2D mammography. Biomed Phys Eng Express. 2018;4(4):045018.

117. Van Laere KJ, Versijpt J, Koole M, Vandenberghe S, Lahorte P, Lemahieu I, et al. Experimental performance assessment of SPM for SPECT neuroactivation studies using a subresolution sandwich phantom design. Neuroimage. 2002 May;16(1):200–16.

118. Larsson SA, Jonsson C, Pagani M, Johansson L, Jacobsson H. A novel phantom design for emission tomography enabling scatter- and attenuation-"free" single-photon emission tomography imaging. Eur J Nucl Med. 2000 Feb;27(2):131–9.

119. Holmes RB, Hoffman SMA, Kemp PM. Generation of realistic HMPAO SPECT images using a subresolution sandwich phantom. NeuroImage. 2013;81:8–14.

120. Berthon B, Marshall C, Holmes R, Spezi E. A novel phantom technique for evaluating the performance of PET auto-segmentation methods in delineating heterogeneous and

irregular lesions. Ejnmmi Phys [Internet]. 2015 Dec;2(1). Available from: ://WOS:000379208500013

121. Negus IS, Holmes RB, Jordan KC, Nash DA, Thorne GC, Saunders M. Technical Note: Development of a 3D printed subresolution sandwich phantom for validation of brain SPECT analysis. Med Phys. 2016 Sep;43(9):5020–7.

122. Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K, et al. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). Philos Trans R Soc B-Biol Sci. 2001 Aug 29;356(1412):1293–322.

123. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage. 2002 Jan;15(1):273–89.

124. Ashburner J, Barnes G, Chun-Chuan C, Daunizeau J, Flandin G, Friston K, et al. SPM 12 manual [Internet]. London, UK: University College London; 2016. Available from: http://www.fil.ion.ucl.ac.uk/spm/doc/manual.pdf

125. Koch W, Unterrainer M, Xiong G, Bartenstein P, Diemling M, Varrone A, et al. Extrastriatal binding of I-123 FP-CIT in the thalamus and pons: gender and age dependencies assessed in a European multicentre database of healthy controls. Eur J Nucl Med Mol Imaging. 2014 Oct;41(10):1938–46.

126. Notghi A, O'Brien J, Clarke EA, Thomson WH. Acquiring diagnostic DaTSCAN images in claustrophobic or difficult patients using a 180 degrees configuration. Nucl Med Commun. 2010 Mar;31(3):217–26.

127. Koch W, Bartenstein P, la Fougere C. Radius dependence of FP-CIT quantification: a Monte Carlo-based simulation study. Ann Nucl Med. 2014 Feb;28(2):103–11.

128. Lawson RS, White D, Cade SC, Hall DO, Kenny B, Knight A, et al. An audit of manufacturers' implementation of reconstruction filters in single-photon emission computed tomography. Nucl Med Commun. 2013 Aug;34(8):796–805.

129. Abdulkadir A, Mortamet B, Vemuri P, Jack CR, Krueger G, Klöppel S, et al. Effects of hardware heterogeneity on the performance of SVM Alzheimer's disease classifier. NeuroImage. 2011 Oct 1;58(3):785–92.

130. Soret M, Koulibaly PM, Darcourt J, Buvat I. Partial volume effect correction in SPECT for striatal uptake measurements in patients with neurodegenerative diseases: impact upon patient classification. Eur J Nucl Med Mol Imaging. 2006 Sep 1;33(9):1062–72.

131. Liu Y-H, Lam PT, Sinusas AJ, Wackers FJT. Differential effect of 180 degrees and 360 degrees acquisition orbits on the accuracy of SPECT imaging: quantitative evaluation in phantoms. J Nucl Med. 2002 Aug;43(8):1115–24.

132. Knesaurek K, King MA, Glick SJ, Penney BC. Investigation of causes of geometric distortion in 180 degrees and 360 degrees angular sampling in SPECT. J Nucl Med 1989 Oct;30(10):1666–75.

133. de Bruijne M. Machine learning approaches in medical image analysis: From detection to diagnosis. Med Image Anal. 2016 Oct 1;33(Supplement C):94–7.

134. Ghafoorian M, Mehrtash A, Kapur T, Karssemeijer N, Marchiori E, Pesteie M, et al. Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation. ArXiv170207841 Cs [Internet]. 2017 Feb 25 [cited 2017 Nov 13]; Available from: http://arxiv.org/abs/1702.07841

135. National Institute for Health and Care Excellence. Diagnostics assessment programme manual [Internet]. 2011 Dec [cited 2017 Nov 14]. Available from: https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-diagnostics-guidance/Diagnostics-assessment-programme-manual.pdf

136. Keltie K, Bousfield DR, Cole H, Sims AJ. Medical Technologies Evaluation Programme: A review of NICE progression decisions, 2010–2013. Health Policy Technol. 2016 Sep 1;5(3):243–50.

137. U.S. Food and Drug Administration. Press Announcements - FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems [Internet]. [cited 2018 Oct 28]. Available from: https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm604357.htm

138. Information Commissioner’s Office. Royal Free - Google DeepMind trial failed to comply with data protection law. Information Commissioner’s Office website [Internet]. 2017 Jul 3 [cited 2017 Nov 29]; Available from: https://ico.org.uk/about-the-ico/news-

and-events/news-and-blogs/2017/07/royal-free-google-deepmind-trial-failed-to-comply-with-data-protection-law/

139. Ipsos MORI. Commercial access to health data: key findings [Internet]. Ipsos MORI; 2016 [cited 2017 Nov 29]. Available from: https://www.ipsos.com/ipsos-mori/en-uk/commercial-access-health-data

140. House of Commons Science and Technology Committee. The big data dilemma. Fourth report of session 2015-16 HC468. The stationary Office; 2016.

10 Appendix 1 – handouts provided as part of the pilot reporting study

Computer aided diagnosis (CAD) for DaTSCAN SPECT imaging

John Fenner, Jonathan Taylor

Introduction

Welcome to the SPECT / CAD lab exercise. This brings together many facets of the taught material in the FRCR course:

- Ionising radiation and dose
- Gamma camera
- Diagnostic protocols
- Tomographic reconstruction (SPECT)
- Image interpretation
- Image quantification
- Diagnostic performance
- New techniques

The session will last 2.5 hours and involve diagnosis of brain scan images, after an initial period of training.

Learning objectives:

- Consolidation of SPECT imaging
- Introduction to DaTSCAN and its clinical rationale
- Training in how to interpret DaTSCAN images
- Introduction to quantitative aids in diagnosis
- Practical exposure to CAD diagnosis and its implications for diagnostic practice
- Introduction to metrics of diagnostic performance

Timetable:

Part 1 (14:00-15:00)

- The exercise will start with a reminder of SPECT techniques and an introduction to DaTSCAN imaging.
- You will be trained, as a group, to recognise the appearances of normal and abnormal DaTSCAN images using a series of 15 training datasets.
- After training you will each be asked to interpret a series of 30 further patient images, displayed automatically on a computer. Interpretation will be in the form of a score, from 1-5, representing the degree to which you think the particular image is normal or abnormal.

Break for coffee (15:00-15:15)

Part 2 (15:15-16:15)

- The reporting exercise will be repeated but this time the opinion of the computer aided diagnosis software will also be displayed for each of the 30 datasets (the CAD output will be displayed in terms of a probability value).
- At the end of the exercise summary statistics on your performance will be provided (and compared to that of experienced reporters).
- Discussion about diagnosis and the use of CAD for assisted reporting.

This work is contributing to quantitative image developments within the department of Nuclear Medicine

DaTSCAN background information

The following information is largely taken from the GE website (md.gehealthcare.com).

Indications

DaTSCAN (Ioflupane I 123 Injection) is a radiopharmaceutical indicated for striatal dopamine transporter visualization. Single photon emission computed tomography (SPECT) brain imaging is used to assist in the evaluation of adult patients with suspected Parkinsonian syndromes (PS). In these patients, DaTSCAN may be used to help differentiate essential tremor from tremor due to PS (idiopathic Parkinson's disease, multiple system atrophy and

progressive supranuclear palsy). In addition, DaTSCAN is also used to differentiate between dementia with Lewy Bodies and other forms of dementia.

Patient pathway (Sheffield)

Approximately 2 patients per week are referred for DaTSCAN tests at Sheffield Teaching Hospitals. This represents approximately 1% of the department's total workload. Most referrals are from Neurologists based in secondary care in the local region. Approximately a quarter of referrals come from primary care.

Clinical pharmacology – mechanism of action

The active drug substance in DaTSCAN is N-ω-fluoropropyl-2β-carbomethoxy-3β-(4- [123 I]iodophenyl)nortropane or ioflupane I 123. In vitro, ioflupane binds reversibly to the human recombinant dopamine transporter (DaT). Autoradiography of post-mortem human brain slices exposed to radiolabeled ioflupane shows concentration of the radiolabel in striatum (caudate nucleus and putamen). Parkinsonian syndromes reduce DaT availability, enabling DaTSCAN to be used as a tracer to detect these conditions.

DaTSCAN also accumulates in other parts of the body, particularly the liver and lungs. Over time the tracer is washed out of the body, mostly via urinary excretion (60% over 48 hours).

Dosage and administration

The recommended DaTSCAN dose is 111 to 185 MBq (delivered intravenously). Images should be acquired between 3 and 6 hours post-injection (when tracer binding is maximised and stable).

The Effective Dose resulting from a DaTSCAN administration (activity of 185 MBq) is 3.94 mSv in an adult

Physical characteristics

Iodine 123 is a cyclotron-produced radionuclide that decays to 123 Te by electron capture and has a physical half-life of 13.2 hours. The most abundant emission is a gamma ray at 159keV, which is used for imaging.

The first half-value thickness of lead (Pb) for iodine 123 is 0.005 cm. The half-value thickness in soft tissue is approximately 5.0 cm

Imaging parameters (Sheffield)

3D SPECT images provide a map of the concentration of radioactive tracer within the body. DaTSCAN images are reconstructed from multiple 2D projections, taken from different angles around the patient. This is a tomographic reconstruction technique in which images are acquired for 30s per projection with a matrix size of 128 x 128. Each of the 2 detector heads of the gamma camera are positioned 180 degrees apart, on opposite sides of the patient. Each detector acquires 60 images, with 3 degrees rotation between each, in a circular orbit around the patient's head such that the detector gantry rotates 180 degrees over the course of the scan. An energy window of 159keV (+/- 10%) is used. Total imaging time is approximately 35 minutes. Once the scan is finished the projection data are converted into a tracer concentration map through iterative reconstruction. This form of reconstruction has advantages over more traditional filtered back projection algorithms, often leading to reduced noise and improved contrast

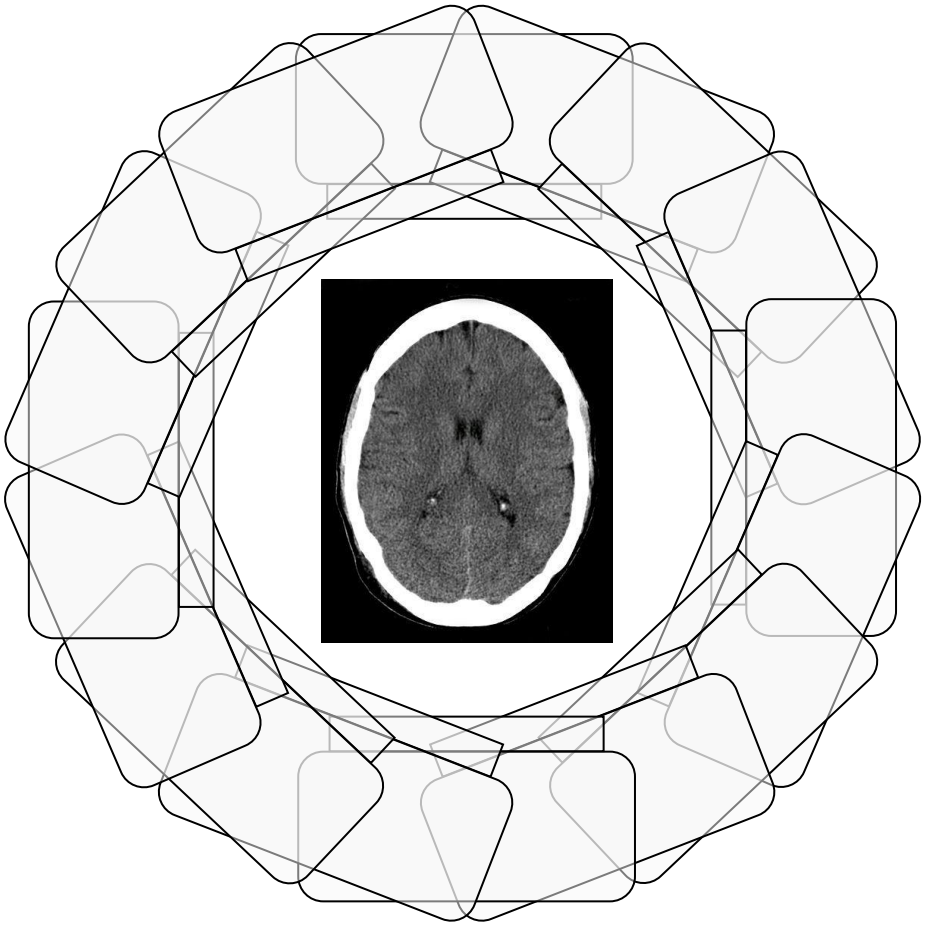


Figure 1. Circular orbit of gamma camera heads

Image interpretation

Determination of whether an image is normal or abnormal is made by assessing the extent (as indicated by shape) and intensity of the striatal signal. Image interpretation does not involve relating the striatal image appearance with clinical signs and/or symptoms.

Normal: In transaxial images, normal images are characterized by two symmetric comma or crescent-shaped focal regions of activity mirrored about the median plane. Striatal activity is distinct, relative to surrounding brain tissue (Figure 2).

Abnormal: Abnormal DaTSCAN images fall into at least one of the following three categories (all are considered abnormal).

- Activity is asymmetric, e.g. activity in the region of the putamen of one hemisphere is absent or greatly reduced with respect to the other. Activity is still visible in the caudate nuclei of both hemispheres resulting in a comma or crescent shape in one

and a circular or oval focus in the other. There may be reduced activity between at least one striatum and surrounding tissues (Figure 3).

- Activity is absent in the putamen of both hemispheres and confined to the caudate nuclei. Activity is relatively symmetric and forms two roughly circular or oval foci. Activity of one or both is generally reduced (Figure 4).
- Activity is absent in the putamen of both hemispheres and greatly reduced in one or both caudate nuclei. Activity of the background with respect to the striata is more prominent (Figure 5).

Figure 2

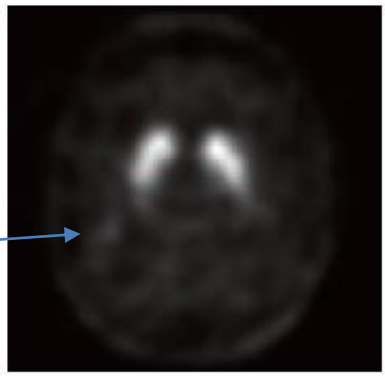


Figure 3

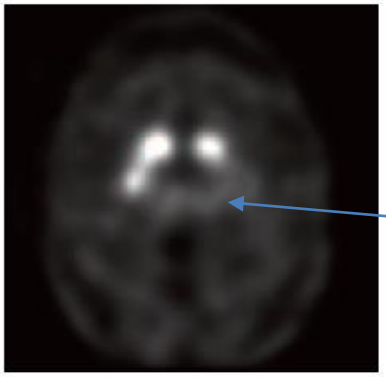


Figure 4

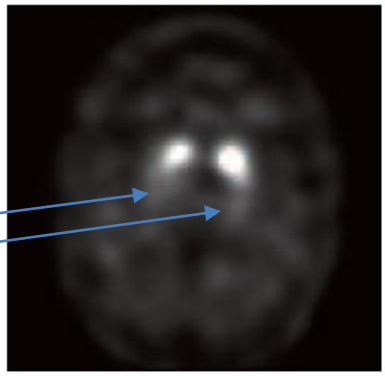
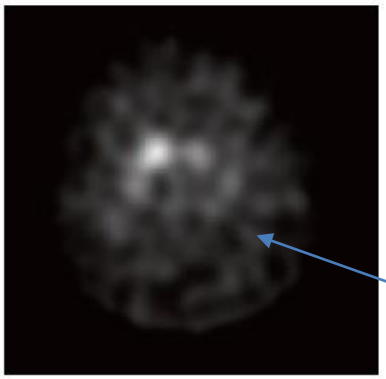


Figure 5



Visual interpretation is inherently qualitative. More objective image analysis is available with the use of quantitative software, based on region of interest (ROI) tools. This is useful in quantifying tracer uptake and can be a helpful aid to diagnosis.

Semi-quantification with an ROI tool

Semi-quantification refers to the measurement of a particular quantity within one region of interest in an image relative to that of a standard. In DaTSCAN imaging, tracer uptake within the striatum (or subregions of the striatum) is measured with respect to a reference. The European Association of Nuclear Medicine (EANM) procedure guidelines recommend that semiquantitative analysis is performed to objectively assess striatal DaT binding, in addition to visual interpretation. It is now used routinely in many UK hospitals.

Commonly, regions of interest are defined on DaTSCAN images, over left and right putamena and caudates, with an additional region drawn over the visual cortex or cerebellum as a reference. ROIs may be determined automatically or may require user intervention. Semi-quantitative figures are derived by dividing detected counts within striatal regions by those measured in the reference area. Comparison with normal ranges enables an objective assessment of the presence or absence of disease.

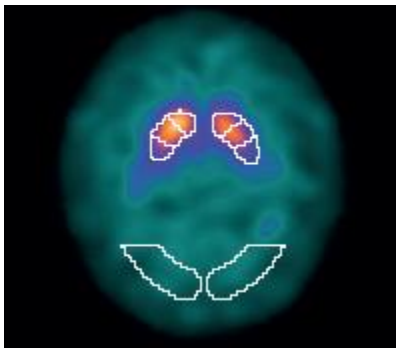


Figure 5 Typical regions of interest used to measure tracer uptake

	Striatum Right SBR	Striatum Left SBR
Measured	+1.62	+1.18
Mean (±1 SD)	+2.44 (±0.42)	+2.43 (±0.43)
Deviation	-33%	-52%
Z-Score	-1.96	-2.93

Figure 6 Example semi-quantification output (relative tracer uptake in the striatum as compared to a normal database)

Quantification and semi-quantification are becoming more common in many other areas of imaging. For instance, cardiac ejection fraction is routinely derived from MRI, CT, Ultrasound and Nuclear Medicine data by drawing regions of interest at systole and diastole and comparing volumes. In addition, tumour volumes are regularly measured on CT or MRI data to enable assessment of the progression of disease. Another example is the calcium score (Agatston score), used to quantify the extent and severity of calcium build up in the coronary arteries.

The DaTSCAN semi-quantification tool is a simple but robust addition to the diagnostic armoury, but more sophisticated methods are available. Computer aided diagnosis (CAD) is one example and its potential is explored in this lab exercise.

Computer aided diagnosis

Computer aided diagnosis (CAD) software generates more than a potentially relevant tracer uptake value. It can be considered to be an objective assessment of an image. It generates a diagnostic score which is shared with the reporting clinician in order to improve reporting accuracy and consistency. It is an extension of simple quantification since the image data is analysed by an independent entity (the software) to come up with an objective output related to the patient's state of disease. It makes a decision as to whether the patient is likely to have a particular disease or not. For instance, the output from a CAD algorithm may be a probability value related to the likelihood of disease being present. CAD algorithms can be more effective than quantification techniques in the detection of disease.

CAD methods have been developed and refined over several decades (Doi, 2007). However, use in the clinic has historically been very limited. CAD for assisted interpretation of mammograms is one of the few areas where commercial software tools have found widespread uptake (prevalent in the USA). Recent technological advancements have made the effectiveness of CAD algorithms much greater. In several areas CAD algorithms have shown evidence of performance that surpasses that of human observers. With significant recent investment many companies are now actively developing CAD software for use in the clinic and these are likely to become mainstream in the future. However, currently there remains a significant gap in the evidence base in terms of CAD's impact on reporting performance.

The aim of this lab session is to introduce you to CAD assisted diagnosis. In this study you will observe how a CAD system, which reports a single probability value, affects reporting decisions. Findings will be augmented through a feedback session, where wider implications associated with the software will be discussed. The output from the exercise will be used to inform a wider clinical study, both in terms of the CAD system design and the study protocol.

Metrics of performance

An important consideration when introducing new clinical tools is performance. Does the new tool improve diagnostic performance? Does it have adverse outcomes? What kind of metrics are appropriate for judging the merit of a new tool / protocol?

In many respects this is a cost-benefit exercise and might include simple measures such as time to diagnose. A rigorous approach will often incorporate the use of ROC curves, requiring an appreciation of true/false positives/negatives, the setting of diagnostic thresholds and an assessment of diagnostic impact (on the patient pathway). This exercise will make use of these approaches to expose the utility of CAD as a diagnostic tool.

References:

Doi, K. 2007. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31: pp198-21