Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 90 (2016) 93 – 98

Procedia Computer Science

International Conference On Medical Imaging Understanding and Analysis 2016, MIUA 2016, 6-8 July 2016, Loughborough, UK

Clinical adoption of CAD: exploration of the barriers to translation through an example application

Jonathan Taylor[a,*], John Fenner[a]

[a] *University of Sheffield, Royal Hallamshire Hospital, Glossop Road, Sheffield S10 2JF, UK*

Abstract

Computer aided diagnosis (CAD) software is not yet widely used in clinic. This paper aims to identify possible reasons why. Firstly, the technical maturity of CAD is explored through analysis of diagnostic accuracy metrics in one example application, the automated classification of Ioflupane I123 (DaTSCAN) images. Software is developed for image classification based on well-established eigenimage techniques. Using a publicly available database of images an area under the Receiver Operator Curve (AUROC) of 0.980 is achieved.

Given these impressive results the main blockage to clinical adoption, both in DaTSCAN classification and potentially in other applications, is likely to relate to wider issues. These are explored with reference to the demands of the National Institute for Health and Care Excellence (NICE) evaluation processes. It is postulated that in order to enable wider adoption a greater focus on proving the safety, efficacy and cost effectiveness of CAD may be required.

Peer-review under responsibility of the Organizing Committee of MIUA 2016

Keywords: Computer Aided Diagnosis; DaTSCAN; NICE

1. Introduction

The interpretation of medical images relies largely on the ability of a human observer to visually identify characteristics of an image that are associated with pathological changes. This is not always a straightforward process and is inherently subjective. In extreme cases clinically significant errors may be made (the incidence of such errors is estimated to be 1-20% [1,2]).

Recently, the role of the radiologist has become even harder. With the progression of imaging hardware and technology, the number and complexity of images produced by modern scanners is rapidly increasing [3]. The volume of information generated from individual scans continues to grow and as a practising Clinical Scientist it is clear that interpretation by humans using standard techniques is becoming more demanding (a problem often referred to as data overload [4]). It is envisaged that new ways of working will be required in order to cope [5].

* Corresponding author. Tel.: +44-114-271-2628
E-mail address: jonathan.taylor@sth.nhs.uk

Peer-review under responsibility of the Organizing Committee of MIUA 2016
doi:10.1016/j.procs.2016.07.029

Given these trends it is argued that Computer Aided Diagnosis (CADx) and Computer Aided Detection (CADe) software, henceforth referred to together as CAD, will eventually become a necessity in radiology departments. However, despite the obvious and growing need for computer assistance CAD software is not yet widely used in the clinic outside of mammography in the USA[6], and the number of commercial CAD systems is very limited[5].

This paper attempts to clarify why clinical uptake has been so modest. The first step in this analysis is an exploration of the technical maturity of CAD algorithms. If it can be shown that CAD technology is equal to the task of automated diagnosis then it must be assumed that the blockage to clinical translation lies elsewhere. By way of example, the current suitability of CAD for the automated classification of Ioflupane I-123 (DaTSCAN) images is examined.

A well-established CAD technology, eigenimage analysis, is demonstrated and applied to a DaTSCAN image database. Performance figures usefully quantify diagnostic accuracy and are compared with methods reported in the literature. Beyond this, other aspects of clinical uptake of CAD are considered, with specific reference to the National Institute for Health and Care Excellence (NICE) evaluation processes. Positive recommendations from NICE can be highly influential in persuading hospitals to adopt certain technologies. It is prudent to be aware of such requirements, and their impact in the context of low clinical adoption of CAD.

The novelty of this paper is not in the methods applied to generate CAD output, but the realisation that adoption of CAD in the field requires more than a novel algorithm; acceptance of the technology implies acceptance by the end user, which relies heavily on a structured pathway to adoption

1.1. Eigenimage analysis

Fundamentally, this widely used model involves the application of principal component analysis (PCA) to a set of training images, where the pixel values are the feature of interest. The principal components or eigenimages output from PCA are then used for subsequent processing of test images, in particular for classification purposes. One common method for performing PCA is through eigen-decomposition of the variance-covariance matrix:

$$XX^T = EDE^T \tag{1}$$

Where X is a matrix containing the pixel values of test images, concatenated into vectors, E is an orthonormal matrix containing the eigenvectors of XX^T (i.e. the eigenimages) and D is a diagonal matrix containing the eigenvalues. The eigenimages are usually stated in order of reducing variance such that the first eigenimage describes the largest amount of variance in the data.

Eigenimages can be used as a basis for classification in a number of ways. In face recognition tasks, where eigenimage analysis has been most extensively exploited, the distance between test images and training images in multidimensional eigenimage space is used as the discrimination metric[7]. In medical imaging, similar approaches can be used to establish whether test images are part of normal or pathological groups, usually based on a small subset of the generated eigenimages. In the following section the Mahalanobis distance from a group of normal training data in eigenimage space is used as the discrimination method.

2. Method

DaTSCAN is a radioactive tracer with a high affinity for dopamine active transporter (DaT), a protein which resides in the membrane of presynaptic axon terminals in the striatum. It is administered prior to gamma camera SPECT imaging, for diagnosis of Parkinsonian syndromes such as Parkinsons Disease (PD). The PPMI database includes a collection of reconstructed DaTSCAN images taken from controls and diagnosed PD patients (www.ppmi-info.org/data). This has been used extensively by CAD specialists for validating a range of different algorithms[8,9,10]. Striatal uptake of DaTSCAN reduces in extent and intensity with increasing severity of disease, creating two distinct voxel intensity patterns in the normal and pathological populations: a comma appearance for normal images and a dot appearance for PD patients.

The performance of eigenimage analysis in relation to DaTSCAN imaging was assessed by measuring the area under the receiver operator curve (AUROC), which provides the most comprehensive description of diagnostic accuracy[11]. Although of reduced clinical relevance, for comparison purposes the maximum accuracy of the algorithm on the ROC curve was also reported. Processing was carried out as follows. All steps were performed using Matlab software:

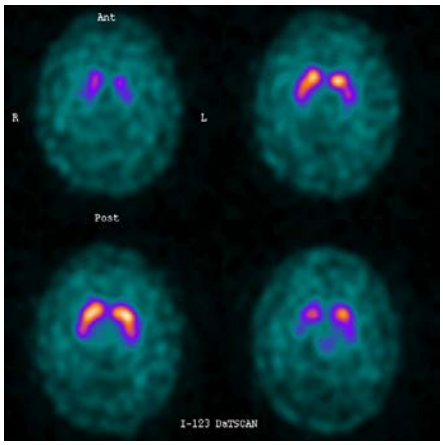

Figure 1. Example DaTSCAN images. Four reconstructed slices are shown from within the 3D brain image.

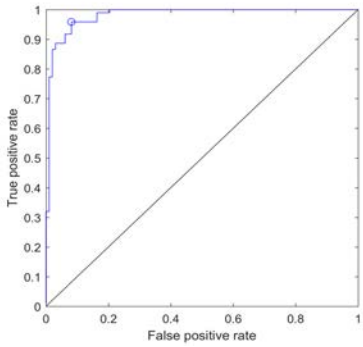

Figure 2. ROC curve derived from Mahalanobis distance measurements. The blue circle is the point of maximum diagnostic accuracy

1. Images in the PPMI database were affine registered to a template (chosen from the controls) so as to remove the effects of differences in patient positioning. Registration was carried out using the Sheffield Image Registration Toolkit (ShIRT[12]). A striatal region of interest (ROI) was manually drawn on the template image.
2. The ROI was used to mask intensity values outside of the striata in all control and PD images, thereby reducing the influence of diagnostically unimportant tissues. Data were intensity normalised to the mean intensity level in the striatal region to account for differences in scaling.
3. 100 eigenimages were generated from the first 100 controls (training images), after mean centring. The coefficients of these datasets in eigenimage space were retained
4. 100 control and 100 PD test images, separate to the training data, were projected on to the first 5 eigenimages. The Mahalanobis distances of each test image from the training images in the eigenimage subspace was measured.
5. An ROC curve was generated from the Mahalanobis distance measurements

3. Results

The ROC curve is displayed in figure 2. AUROC was 0.980 (95% confidence interval 0.961-1.000). The maximum accuracy achieved was 94.0%, highlighted on the ROC curve by the blue circle in figure 2. These results are presented in table 1, alongside results reported by other authors using the same PPMI image database.

Table 1. Performance results from automated diagnosis of images in the PPMI database

Authors	Method	Num. of images	Accuracy (max)	AUROC
J. Taylor, J. Fenner	Eigenimage analysis of pixels. Mahalanobis distance as discrimination metric.	200 controls, 100 PD	94.0%	0.980
F.J. Martinez-Murcia, J.M. Gorriz, J. Ramirez, I.A. Illan, A. Ortiz[9]	Independent component analysis of extracted features, input to Support Vector Machine for discrimination	114 controls, 175 PD	91.3%	
F.J. Martinez-Murcia, J.M. Gorriz, J. Ramrez, I.A. Illan, C.G. Puntonet[8]	Extracted features input to Support Vector Machine for discrimination	209 controls, 445 PD	97.9%	
R. Prashanth, S.D. Roy, P.K. Mandal, S. Ghosh[10]	Extracted features input to Support Vector Machine for discrimination	181 controls, 369 PD	96.1%	

4. Discussion

The results obtained from this exercise show that an accurate CAD algorithm for DaTSCAN can be created with a relatively simple algorithm based on eigenimages. Established methods were deliberately chosen for this study since they are better understood and represent a lower risk for clinical adoption. Validation was carried out using a high-quality, multi-centre, prospective patient database, providing strong evidence that eigenimage analysis could be an effective tool for the clinic. The strength of CAD for automated diagnosis in DaTSCAN imaging is further enhanced by results from other authors (see table 1), who have achieved accuracies of up to 98%.

Although conducted with different data, a previous test of the classification performance of human observers reported an accuracy of 91% for DaTSCAN imaging[13]. The evidence from table 1 suggests that all CAD methods, covering different levels of complexity, exceed this performance. Therefore, algorithm sophistication is not a barrier to adoption in this case.

The results strongly suggest that CAD is likely to be sufficiently effective in this application to be suitable for clinical translation. However, outside of software tools which simply quantify and display ratios of counts in different image regions (e.g. Hermes BRASS, GE DaTQUANT), CAD is not yet used routinely by radiologists in DaTSCAN imaging. This is despite recent audit data showing a disagreement between radiologists in test centres and a panel of experts in 8% of cases[14].

The relative mismatch between high accuracy figures in the literature, from a range of algorithms, and low clinical CAD uptake suggests that other factors must have a significant impact on the route to translation. In this paper a structured approach to understanding the wider issues around translation is proposed.

The consumers of CAD are mostly radiologists and clinical radiology departments. In the UK these consumers are likely to be significantly influenced by NICE guidelines. Therefore, analysis and understanding of what NICE evaluation requires may lead to further insights into the barriers to clinical adoption.

4.1. NICE evaluation processes

One of the key roles of NICE is to produce guidance for health and social care professionals in the UK and further afield. New diagnostic technologies, such as CAD, can be alerted to NICE through the Technology Evaluation Programme or the Diagnostic Assessment Programme. Both programmes are closely related, the major difference being that the Technology Evaluation Programme is focused on technologies that achieve a similar clinical benefit at reduced cost to the health system or more benefit at the same cost as current practice. The Diagnostic Assessment

Table 2. Summary of the main requirements for NICE evaluation (and approval) along with an assessment of the implications for CAD software development

Requirements for NICE evaluation	Implications for CAD software
Regulation Technology must be licensed for the intended purpose. It is likely that software will need to be CE marked (in Europe) in line with the requirements of the Medical Devices Directive (MDD). This is likely to require adherence to appropriate standards and auditing by an external notified body.	Software must be created according to an appropriate quality management system. For risk assessment and validation purposes a detailed understanding of how CAD components perform in different scenarios is likely to be required. Depending on the device classification, clinical investigations are likely to be required to prove that the technology is safe and effective. This will require testing in realistic clinical environments
Clinical impact The evidence base detailing the clinical impact of the technology must be available to enable an informed decision about adoption. In particular a performance comparison of the new technology as compared to standard of care is required. NICE places emphasis on high quality research where bias is minimised. The evidence must be of sufficient quantity and consistency to enable a robust recommendation.	Proof of the diagnostic effectiveness of CAD, as compared to standard care (i.e. visual analysis by radiologists), must be collated. Ideally, test results should reference accurate gold-standard diagnoses and tests should ideally be performed in realistic clinical scenarios with randomly selected data. Small test datasets are unlikely to be suitable.
Health economics Evidence base of economic data in relation to the technology must be available. For the Technology Assessment Programme economic analysis can be presented using a cost-consequence methodology. Overall costs to the NHS must be equal to or less than those of current standard care. For the Diagnostic Assessment Programme more in-depth economic analysis is required, including cost effectiveness analysis. NICE does not use specific thresholds but above £30,000 per Quality Adjusted Life Year (QALY) gained a strong case for support is required.	The impact of CAD in terms of direct costs and indirect costs must be prepared. This may require economic modelling. If costs can be shown to be less or equal to those of standard care then NICEs decision will largely fall on clinical evidence. If overall costs are greater than that of standard care it may be useful to perform cost effectiveness analysis, generating evidence through the care pathway, to estimate the likely cost per QALY gained (and ensure it is less than £30,000). If cost effectiveness figures are uncertain then NICEs decision will again fall on the clinical evidence

Programme, on the other hand, is focused on the introduction of technology that is likely to result in an overall increase in costs to the NHS. Table 2 shows a list of the common requirements of the two NICE evaluation programmes, based on published guidance[15,16], along with an assessment of the implications for the development of approved clinical CAD software. Table 2 illustrates that there are a number of significant hurdles that must be navigated before NICE approval is granted. Although this is not a pre-requirement for adoption within a local centre, without it wider adoption is challenging, whilst the presence of approval is likely to dramatically improve uptake both in the UK and elsewhere.

Only a very small fraction of current CAD algorithms are likely to be able to meet requirements related to regulatory adherence and clinical and economic evidence. NICE approval is only granted if the research base demonstrates that impact in terms of improved care or reduced costs is substantial. Given previous criticisms of CAD research, particularly with regards to the relevance and significance of reported results[5,3,17,18], and the general lack of economic analysis, NICE approval is likely to be unrealistic for most CAD applications, including automated DaTSCAN analysis. Therefore, a greater research effort directed towards proving the safety, efficacy and cost effectiveness of CAD may be required.

Unfortunately, overcoming regulatory hurdles and testing in clinic against the current standard of care (i.e. against unaided radiologists) are likely to be expensive processes. This latter point indicates that one of the biggest blocks on

clinical adoption of CAD may be related to finances, particularly if, as is often the case, CAD development is driven by commercial organisations.

5. Conclusion

This study has shown that a relatively simple, well-established technology, eigenimage analysis, can perform well when applied to the automated classification of DaTSCAN images. Accuracy figures were in line with results achieved through more complex, highly optimised algorithms applied to the same database of images. Due to the simplicity of DaTSCAN image appearances this task is one of the more straightforward problems in CAD research. However, results do suggest that in at least some applications CAD technologies are ripe for clinical adoption.

As with radiological applications in general, the routine use of CAD for DaTSCAN analysis is very limited. It is in this context that the requirements of NICE evaluation processes provide an insight as to which other factors may be a barrier to clinical translation.

The NICE evaluation processes most relevant to CAD software require adherence to medical device regulations, the generation of evidence comparing CAD with standard of care and economic evidence showing the impact on healthcare systems costs. In most cases CAD algorithms fall short in all these categories. Therefore, although NICE approval is an ambitious target, results do suggest that more research effort should be expended in generating evidence that is of concern to intended consumers, rather than continually adapting and refining CAD technology.

Acknowledgements

This work was supported by the National Institute for Health Research

References

1. Goddard, P., Leslie, A., Jones, A., Wakeley, C., Kabala, J.. Error in radiology. *The British Journal of Radiology* 2001;**74**(886):949–951.
2. Jolly, B.C., Ayers, B., Macdonald, M.M., Armstrong, P., Chalmers, A.H., Roberts, G., et al. The reproducibility of assessing radiological reporting: studies from the development of the General Medical Councils Performance Procedures. *Medical Education* 2001;**35**:36–44.
3. Wang, S., Summers, R.M.. Machine learning and radiology. *Medical Image Analysis* 2012;**16**(5):933–951.
4. Andriole, K.P.. Addressing the Coming Radiology CrisisThe Society for Computer Applications in Radiology Transforming the Radiological Interpretation Process (TRIP¢) Initiative. *Journal of Digital Imaging* 2004;**17**(4):235–243.
5. van Ginneken, B., Schaefer-Prokop, C.M., Prokop, M.. Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology* 2011;**261**(3):719–732.
6. Martyn, C.. Diagnosis by machine. *BMJ* 2008;**337**:a1703.
7. Navarrete, P., Ruiz-del Solar, J.. Eigenspace-based recognition of faces: comparisons and a new approach. In: *11th International Conference on Image Analysis and Processing, 2001. Proceedings*. 2001, p. 42–47.
8. Martinez-Murcia, F.J., Gorriz, J.M., Ramirez, J., Illan, I.A., Puntonet, C.G.. Texture Features Based Detection of Parkinsons Disease on DaTSCAN Images. In: Vicente, J.M.F., Sanchez, J.R.l., Lopez, F.d.l.P., Moreo, F.J.T., editors. *Natural and Artificial Computation in Engineering and Medical Applications*; no. 7931 in Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2013, p. 266–277.
9. Martinez-Murcia, F.J., Gorriz, J.M., Ramirez, J., Illan, I.A., Ortiz, A.. Automatic Detection of Parkinsonism Using Significance Measures and Component Analysis in DaTSCAN Imaging. *Neurocomput* 2014;**126**:58–70.
10. Prashanth, R., Dutta Roy, S., Mandal, P.K., Ghosh, S.. Automatic Classification and Prediction Models for Early Parkinson's Disease Diagnosis from SPECT Imaging. *Expert Syst Appl* 2014;**41**(7):3333–3342.
11. Metz, C.E.. Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. *Journal of the American College of Radiology: JACR* 2006;**3**(6):413–422.
12. Barber, D.C., Oubel, E., Frangi, A.F., Hose, D.R.. Efficient computational fluid dynamics mesh generation by image registration. *Medical Image Analysis* 2007;**11**(6):648–662.
13. Jennings, D., Seibyl, J., Oakes, D., Eberly, S., Murphy, J., Marek, K.. (123i) -cit and single-photon emission computed tomographic imaging vs clinical evaluation in parkinsonian syndrome: Unmasking an early diagnosis. *Archives of Neurology* 2004;**61**(8):1224–1229.
14. Kemp, P.. Bnms autumn meeting invited talk - BNMS DaTSCAN Audit. 2012.
15. National Institute for Health and Care Excellence, . Diagnostics Assessment Programme manual. 2011.
16. National Institute for Health and Care Excellence, . Medical Technologies Evaluation Programme Process Guide. 2011.
17. Giger, M.L., Karssemeijer, N., Schnabel, J.A.. Breast Image Analysis for Risk Assessment, Detection, Diagnosis, and Treatment of Cancer. *Annual Review of Biomedical Engineering* 2013;**15**(1):327–357.
18. Eadie, L.H., Taylor, P., Gibson, A.P.. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. *European Journal of Radiology* 2012;**81**(1):e70–76.