Journal of Neuroscience Methods 256 (2015) 30–40

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Computational Neuroscience

Unsupervised learning based feature extraction for differential diagnosis of neurodegenerative diseases: A case study on early-stage diagnosis of Parkinson disease

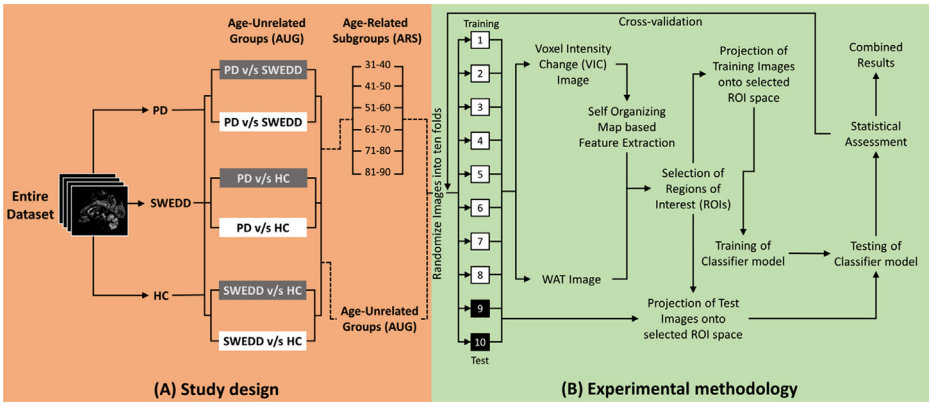
Gurpreet Singh[a], Lakshminarayanan Samavedham[a,b,*]

[a] Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore 117585, Singapore
[b] Residential College 4, 8 College Avenue West, #02-16W, Education Resource Centre, Singapore 138608, Singapore

HIGHLIGHTS

- A new paradigm for early-stage differential diagnosis of PD has been proposed.
- We report accuracy of up to 99.9% with 100% specificity and 99.4% sensitivity.
- We have made the first attempt to distinguish SWEDD subjects from HC and PD.
- This method correctly identifies areas affected in PD as found in literature.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 11 July 2015
Received in revised form 8 August 2015
Accepted 10 August 2015
Available online 21 August 2015

Keywords:
Parkinson disease/diagnosis
Brain mapping
Computer-assisted decision support system
Case-control study
Self-organizing map
Least square support vector machine

ABSTRACT

Background: The development of MRI based methods could prove extremely valuable for identification of reliable biomarkers to aid diagnosis of neurodegenerative diseases (NDs). A great deal of current research has been aimed at identification biomarkers for both diagnosis at early stage and evaluation of the progression of NDs.
New method: We present here a novel synergetic paradigm integrating Kohonen self organizing map (KSOM) and least squares support vector machine (LS-SVM) for individual-level clinical diagnosis of NDs. Feature are extracted in an unsupervised manner using KSOM on preprocessed brain MRIs. Thereafter, these features are fed as input to LSSVM for subject classification.
Results: The applicability of the proposed methodology has been demonstrated using 831 T1-weighted MRIs obtained from Parkinson's Progression Markers Initiative (PPMI) database. We have achieved classification accuracy of up to 99% for differential diagnosis of Parkinson disease with confidence interval of 99.9%.
Comparison with other existing methods: The potential for translation of similar research findings to clinical application is greatly dependent upon two factors (1) accuracy of subject classification achieved and (2) size of the dataset used. Here, we report very high accuracy achieved on one of the largest MRI datasets using multivariate analysis tools.

* Corresponding author at: Residential College 4, 8 College Avenue West, #02-16W, Education Resource Centre, Singapore 138608, Singapore. Tel.: +65 66014927.
E-mail address: laksh@nus.edu.sg (L. Samavedham).

http://dx.doi.org/10.1016/j.jneumeth.2015.08.011

Conclusions: In this paper, we describe a methodology that has the potential to be translated into first-line diagnostic tool for NDs. We also demonstrate the applicability of this methodology for diagnosing PD subjects in early stages of the disease, i.e., subjects in age of 31–60 years.

1. Introduction

Neurodegenerative diseases (NDs) such as Parkinson's disease (PD), Alzheimer's disease (AD) are beginning to become a substantive economic burden. PD is the most common degenerative movement disorder affecting 10 million people worldwide and accounting for an estimated annual expenditure of about €14 billion in Europe only (Tan et al., 2004; Politis, 2014). PD is associated with atrophy of the substantia nigra pars compacta resulting in insufficiency of dopamine hormone (Samii et al., 2004; Obeso et al., 2010). At present there is no definitive test for diagnosing PD (Politis, 2014). The onus of confirmative diagnosis heavily relies only on the expertise of a physician who judges based on subject's history and his/her observable signs and symptoms. In prodromal stages, these signs may range from overpowering to being mild that may go unnoticed. Thus, PD diagnosis still remains a formidable challenge particularly in the early-stages.

In a clinical setting, magnetic resonance imaging (MRI) is routinely performed and analyzed to diagnose PD by evaluating structural and functional abnormalities. To achieve high accuracy of diagnosis, there have been attempts to design computer-assisted decision support systems to automate the analysis of medical images. From these images, extraction of features that are clinically relevant and useful to differentiate between different disease classes is the key towards achieving high classification accuracy. Machine learning tools such as principal component analysis (PCA) (Dyrba et al., 2013; Salvatore et al., 2014), Linear discriminant analysis (LDA) (Wolz et al., 2011), non-negative matrix factorization (NMF) (Padilla et al., 2012) have been employed for feature extraction. Typically, these features are then fed into supervised learning based algorithms such as support vector machines (SVM) (Wolz et al., 2011; Padilla et al., 2012; Dyrba et al., 2013; Salvatore et al., 2014; Tong et al., 2014) for subject classification. The potential of these algorithms for clinical applicability is greatly dependent upon two factors (1) size of the dataset used, and (2) accuracy of subject classification achieved. In the previous studies, either the size of chosen dataset has been relatively small (max. 56 subjects) or the classification accuracy obtained after testing was relatively low, nearly 96–97% (Focke et al., 2011a; Haller et al., 2012, 2013; Salvatore et al., 2014). In addition, a group of subjects with scans without evidence of dopaminergic deficit (SWEDD) has puzzled movement disorder specialists. Even though underlying pathophysiology of SWEDD differs from PD, it has similar clinical symptoms. In routine clinical practise SWEDDs are often misdiagnosed for PD (Schneider et al., 2007; Schwingenschuh et al., 2010; Massano and Bhatia, 2012) causing additional burden for subjects. Further, it is important to diagnose PD in the early-stages as medical and psychological treatments are more effective at that time. Thus, there is a pertinent need for a better-suited computer-assisted decision support system for early-stage diagnosis of PD.

In this regard, we propose an innovative and effective technique to analyze medical images for clinical diagnosis of PD. In essence, it is based on combination of two machine learning algorithms viz., (1) an unsupervised, Kohonen's self-organising map (KSOM) and (2) a supervised, least squares support vector machine (LSSVM). KSOM is used as a vector quantization technique to reduce complexity and provide meaningful information when applied to high-dimensional datasets (Kohonen and Kohonen, 1998). On the other hand, LSSVM (Suykens and Vandewalle, 1999) is employed for classification of subjects; it is a reformulation of standard SVM (Cortes and Vapnik, 1995), but it performs faster computationally as it requires solution of a set of linear equations unlike the need to solve quadratic programming problems. We demonstrate the applicability of the proposed technique on 831 subjects obtained from Parkinson progression marker initiative (PPMI) database. We separately focus on grey matter (GM) and white matter (WM) tissues to identify regions of interest (ROI) that are useful to discriminate between PD, SWEDD and HC subjects. This intuitive and easy-to-implement method is clinically relevant and is intended to contribute towards ND diagnosis.

2. Materials and methods

2.1. Data acquisition

We have obtained morphological T1-weighted magnetic resonance images (MRIs) from PPMI database for *de novo* PD, SWEDD and HC subjects acquired at field strength of 3 Tesla. PPMI is a five-year observational, international, multi-centre study aimed at understanding disease aetiology by identifying PD progression biomarkers (Parkinson Progression Marker, 2011). Clinical information on severity of the PD symptoms was assessed by Unified Parkinson's Disease Rating Scale (MDS-UPDRS) and the Hoehn and Yahr Scale. The Montreal Cognitive Assessment (MoCA) test was used for cognitive assessment of the subjects. Table 1 shows the demographic and clinical details of the subjects that comprise the dataset used in this work. For latest information, see www.ppmi-info.org. Details regarding the PPMI IDs of the subjects can be found in the supplementary information (Subject_Information.xlsx).

2.2. Study design

In this study, the entire dataset was divided into binary classification groups. Three classes were considered in this study i.e. PD, SWEDD and HC subjects. A binary classification group compares two subject classes at a time (PD vs. HC or SWEDD vs. HC or PD vs. SWEDD). For each type of classification group, we have made a tissue-by-tissue comparison considering one type of brain matter, i.e. WM or GM, at a time. The two ways in which dataset was divided considering patient class, type of brain matter and age of the subject are:

1. **Age-Unrelated Groups (AUG):** Dataset was divided based on clinically identified subject class and types of brain matter only. For three types of subject classes (HC, PD and SWEDD) considering only two types of brain matter (GM and WM) regardless of age of the subject 6 AUGs were be formed (3 binary classification groups * 2 types of brain tissue). See Table 1 for demographic and clinical details of these subjects.
2. **Age-Related Subgroups (ARS):** As progression of ND occurs in an age dependent manner, we intended to divide entire ensemble of images for each class of subjects into age-related subgroups (ARS) such that each subgroup represents subjects of similar age and at a comparable stage of the disease. Thus, age was included as a factor in addition to subject class and type of brain matter. A total of 36 ARS (instead of the considered 32 ARS) should