Clinical Neurophysiology 127 (2016) 2695–2703

Contents lists available at ScienceDirect

Clinical Neurophysiology

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

Automatic classification of 6-month-old infants at familial risk for language-based learning disorder using a support vector machine

Marzieh Zare [a,b], Zahra Rezvani [b], April A. Benasich [c,*]

[a] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran 19538, Iran
[b] Institute for Cognitive and Brain Sciences, Shahid Beheshti University, G.C. Evin, Tehran 19393, Iran
[c] Center for Molecular and Behavioral Neuroscience, Rutgers University-Newark, 197 University Avenue, NJ 07102 Newark, USA

See Editorial, pages 2692–2694

ARTICLE INFO

Article history:
Accepted 25 March 2016
Available online 12 April 2016

Keywords:
Infant
Machine learning
Support vector machine (SVM)
EEG
Developmental language disorder
Network analysis

HIGHLIGHTS

- Novel machine learning approaches were used to study selected features within infant resting EEG.
- Two infant groups who differed on familial risk for language learning disorder (LLD) were assessed.
- Identification of infants at higher risk for LLD may facilitate earlier diagnosis and remediation.

ABSTRACT

Objectives: This study assesses the ability of a novel, "automatic classification" approach to facilitate identification of infants at highest familial risk for language-learning disorders (LLD) and to provide converging assessments to enable earlier detection of developmental disorders that disrupt language acquisition.
Methods: Network connectivity measures derived from 62-channel electroencephalogram (EEG) recording were used to identify selected features within two infant groups who differed on LLD risk: infants with a family history of LLD (FH+) and typically-developing infants without such a history (FH−). A support vector machine was deployed; global efficiency and global and local clustering coefficients were computed. A novel minimum spanning tree (MST) approach was also applied. Cross-validation was employed to assess the resultant classification.
Results: Infants were classified with about 80% accuracy into FH+ and FH− groups with 89% specificity and precision of 92%. Clustering patterns differed by risk group and MST network analysis suggests that FH+ infants' EEG complexity patterns were significantly different from FH− infants.
Conclusions: The automatic classification techniques used here were shown to be both robust and reliable and should provide valuable information when applied to early identification of risk or clinical groups.
Significance: The ability to identify infants at highest risk for LLD using "automatic classification" strategies is a novel convergent approach that may facilitate earlier diagnosis and remediation.

1. Introduction

The brain is subject to large structural and functional changes over early development. Rapid auditory processing and auditory change detection abilities in the tens-of milliseconds range are critical to decoding the speech stream and are crucial aspects of speech and language development starting at birth (Aslin, 1989; Eilers et al., 1981; Werker and Tees, 2005; Kuhl et al., 2008). This complex ability to detect subtle sound changes early in infancy, specifically

* Corresponding author at: Center for Molecular and Behavioral Neuroscience, Rutgers University, 197 University Avenue, Newark, NJ 07102, USA. Tel.: +1 973 353 3598; fax: +1 973 353 1760.
E-mail addresses: marziehzare@ipm.ir (M. Zare), za.rezvani@mail.sbu.ac.ir (Z. Rezvani), benasich@andromeda.rutgers.edu (A.A. Benasich).

http://dx.doi.org/10.1016/j.clinph.2016.03.025
1388-2457/

fast sequential changes in the amplitude and frequency composition related to speech, is believed to go awry in a subset of children (Benasich and Tallal, 2002; Choudhury and Benasich, 2011; Tallal, 2004). Such early deviation from normative acoustic processing trajectories is thought to result in language-based learning disorders (LLD) such as specific language impairment and dyslexia (Bush, 2010; Lewis and Elman, 2008; Tallal, 2004), and is suggested to be comorbid with some types of autism (Whitehouse et al., 2008). Familial genetic studies indicate that approximately 30–60% of infants born into families with LLDs are at risk of developing similar problems (Flax et al., 2003; Tomblin, 1989).

Previous studies also indicate that LLDs are associated with detectable differences in brain structure (Chu et al., 2015; Leonard et al., 2011; Westwood, 2004) that may begin before birth. These anatomical differences are thought to be a contributing factor to LLD given the genetic predisposition identified in many of these disorders (Casey et al., 2000; Choudhury and Benasich, 2011; Wong et al., 2013). Therefore, prospective longitudinal studies starting in early infancy and continuing through 3–5 years of age have been designed to detect early precursors and biomarkers of developmental disorders in infants at higher genetic risk of LLD (for review see Benasich and Choudhury, 2012). These early risk markers are difficult to detect using only behavioral testing, however neuroimaging approaches, including EEG, have been effectively used in infant populations (e.g. Benasich et al., 2006; Choudhury and Benasich, 2011; Maitre et al., 2013).

The recent focus has been on improving event-related potential (ERP) recording from EEG and magnetoencephalography (MEG) as well as introducing more fine-grained analyses of continuous EEG, particularly within the context of studying atypical or at-risk groups (Barttfeld et al., 2011; Bosl et al., 2011; Stahl et al., 2012). However, the standard procedure of EEG and MEG analysis continues to be averaging of a large number of artifact-free trials and then using group grand averages to compute statistics. Unfortunately, this can result in a number of problems given the underlying assumptions of this technique. Violation of these basic assumptions arises due to inconsistency of the brain response, variability across trials due to cognitive processes and loss of statistical power as well as statistical bias due to major alterations in the ERP components (Stets et al., 2012). Moreover, in developmental studies, a large proportion of subjects may need to be excluded because the infants have not provided a sufficient number of noise-free trials per condition, thus precluding computation of stable averaged ERPs.

Another approach has focused on using source localization and time/frequency analyses in resting or spontaneous EEG to identify predictors in at-risk populations (e.g. Benasich et al., 2008; Gou et al., 2011) as well as those already diagnosed with a LLD (e.g. Heim and Benasich, 2011; Schiavone et al., 2014). These studies examine particular frequency profiles using Fast Fourier Transform (FFT) or wavelet analyses, thus using more complex oscillatory characteristics. This technique overcomes some of the difficulties of averaged ERPs, however, the information obtained is restricted to the oscillatory domain and must be first computed on a case-by-case basis and then averaged to obtain group data.

In order to improve statistical power and avoid the issues that emerge in MEG, EEG and ERP studies it would be advantageous to have novel statistical methods that would permit *detailed discrimination of individual characteristics* (using raw or minimally pre-processed EEG data) as well as supporting robust classification of groups that may differ on the level of risk for a particular disorder (Stahl et al., 2012).

Automatic classification strategies using machine-learning classifiers have been suggested as just this type of diagnostic tool (Riaz et al., 2013). Machine learning techniques enable leverage of widely distributed, but potentially less robust information to improve the ability to separate individuals into risk groups. Analytic approaches such as FTT do not support group classification and are restricted to features within the oscillatory domain. However, the automatic classification strategies described here use many different features, including oscillatory characteristics, which are driven by the networks detected in the EEG or MEG signal. For example, in one study a multivariate pattern classification approach was applied to network based fcMRI data from a large and unique infant study sample processed with current motion correction procedures (Pruett et al., 2015); significant changes were demonstrated in the structure of large-scale functional brain networks over 6- to 12- months, a period of dramatic cognitive, motor, and social transformation. Hence, this early time period was shown to hold great importance for understanding typical and atypical social-developmental trajectories (Elison et al., 2013). In the present dense-array EEG study, we focus on network connectivity of two infant groups with differing LLD risk levels and deploy a machine learning classifier called a support vector machine (SVM) to separate and classify these infants at low and high risk of LLD. Subjects were two groups of infants: *Family History Positive* (FH+) *group*: Infants born into families with a history of LLD and *Family History Negative* (FH−) *group*: Typically developing control infants without such a family history. We set out to address (i) whether the brain network topology is affected by degree of LLD risk, (ii) which features derived from network analysis contribute to classifications (iii) whether a classifier for prediction of impairment can be implemented to aid early risk assessment of LLD.

2. Participants

Participants in the current study were a subset of the children who participated in a larger prospective study that assessed the effects of early auditory processing skills on later language and cognitive development (Benasich et al., 2006; Choudhury and Benasich, 2011; Choudhury et al., 2007). The studies described here have been reviewed and approved by the Institutional Review Board of our University, were in accordance with the ethical standards of our University and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all parents following a full explanation of the experiment and prior to their child's inclusion in the reported studies. All children were tested at 6, 9, 12, 16, 24, 36 and 48 months of age using both behavioral information processing tasks and EEG assessments. Results from the behavioral information processing assessments are presented elsewhere (Choudhury et al., 2007) as are the analysis of the EEG event-related responses (ERPs) to rapidly modulated auditory stimuli (Choudhury and Benasich, 2011). In the present study we focus only on the 6-month-old infants in the two risk groups: FH+ and FH−).

Families were recruited from urban and suburban communities in New Jersey and assigned to one of the two groups based on parental report of family history of LLD. The FH+ group consisted of 12 full-term normal birth weight healthy infants (10 males, 2 females). Infants from FH− families were recruited from local newspaper birth announcements and pediatric clinics (7 males, 5 females). In order to be classified as FH+, families were asked to provide clinical reports of expressive and receptive language scores and a general cognitive score for at least one affected and diagnosed immediate family member (the "proband"); 75% of the probands were siblings of the infant participant and the remainder were parents who had been clinically diagnosed with a LLD. All probands for this sample had diagnoses of either specific language impairment or developmental dyslexia. (Further information