Random forest to differentiate dementia with Lewy bodies from Alzheimer's disease

Dauwan, Meenakshi; van der Zande, Jessica J; van Dellen, Edwin; Sommer, Iris E C; Scheltens, Philip; Lemstra, Afina W; Stam, Cornelis J

Published in:
Alzheimer's & dementia (Amsterdam, Netherlands)

DOI:
10.1016/j.dadm.2016.07.003

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2016

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 09-09-2020
Diagnostic Assessment & Prognosis

Random forest to differentiate dementia with Lewy bodies from Alzheimer’s disease

Meenakshi Dauwan, Jessica J. van der Zande, Edwin van Dellen, Iris E. C. Sommer, Philip Scheltens, Afina W. Lemstra, Cornelis J. Stam

Abstract

Introduction: The aim of this study was to build a random forest classifier to improve the diagnostic accuracy in differentiating dementia with Lewy bodies (DLB) from Alzheimer’s disease (AD) and to quantify the relevance of multimodal diagnostic measures, with a focus on electroencephalography (EEG).

Methods: A total of 66 DLB, 66 AD patients, and 66 controls were selected from the Amsterdam Dementia Cohort. Quantitative EEG (qEEG) measures were combined with clinical, neuropsychological, visual EEG, neuroimaging, and cerebrospinal fluid data. Variable importance scores were calculated per diagnostic variable.

Results: For discrimination between DLB and AD, the diagnostic accuracy of the classifier was 87%. Beta power was identified as the single-most important discriminating variable. qEEG increased the accuracy of the other multimodal diagnostic data with almost 10%.

Discussion: Quantitative EEG has a higher discriminating value than the combination of the other multimodal variables in the differentiation between DLB and AD.

Keywords: Alzheimer’s disease; Dementia with Lewy bodies; EEG; Random forest; Diagnostic accuracy; Beta power; Machine learning

1. Introduction

Alzheimer’s disease (AD) and dementia with Lewy bodies (DLB) are the two most common forms of dementia in the aging population [1,2]. DLB and AD have several overlapping characteristics, making differential diagnosis in clinical practice at times difficult [3]. Compared to AD, consensus criteria [1] in DLB have moderate sensitivity [4,5]. Accurate diagnosis of DLB and AD is essential for patient guidance and appliance of possible early treatment and prevention strategies [6]. Therefore, disease-specific biomarkers from cerebrospinal fluid (CSF) and neuroimaging are increasingly used, but these diagnostic tests can be costly and are not always available [5,7]. Furthermore, the frequent presence of concomitant AD pathology in DLB patients renders amyloid markers and magnetic resonance imaging (MRI) less discriminative [5,8]. In contrast, electroencephalography (EEG) has been proposed as a low-cost and readily available diagnostic tool to distinguish between DLB and AD [9,10]. At present, in a clinical setting, data from patient history and above-mentioned diagnostic tests are weighted differently in each individual patient to make a diagnosis [11]. The exact contribution of the (combinations of) EEG and other diagnostic tests to the differential diagnosis of DLB and AD remains unclear.

Automated classification algorithms can directly provide the most relevant diagnostic variables and estimate their...
relative importance in classifying cognitive impairment, which can improve diagnostic efficiency [12,13]. Ensemble-learning methods construct automated classification algorithms that can learn from and predict data by building a model in the form of input-output relationships of variables (i.e., features in classification algorithms) [14]. Random forest is one such algorithm, developed by L. Breiman, and based on the principle of decision tree learning [15]. In the field of dementia, ensemble-learning methods have mainly been studied to classify patients with AD [13], whereas very little evidence is available on the automated discrimination between DLB and AD [12] or on the combination of different diagnostic modalities in an automated classifier.

This study aimed to build a random forest classifier to discriminate between DLB, AD, and controls and to quantify the importance of (combinations of) different types of diagnostic features (i.e., clinical, neuropsychological, EEG, CSF, and neuroimaging data), with a specific focus on the role of EEG.

2. Methods

2.1. Study population

A total of 66 probable DLB patients, 66 probable AD patients, and 66 subjects with subjective cognitive decline (SCD) were selected from the Amsterdam Dementia Cohort [11]. The groups were matched on group level for age and gender. All subjects were referred to the Alzheimer Center of the VU University Medical Center (VUmc) in Amsterdam, The Netherlands, between September 2003 and June 2010. Standardized dementia diagnostic workup included neuropsychological assessment, lumbar puncture, brain MRI, and resting-state EEG. All subjects gave written informed consent for storage and use of their clinical data for research purposes. The Medical Ethics Committee of the VUmc approved this study. A clinical diagnosis and treatment plan was made by consensus in a weekly multidisciplinary meeting [11]. Probable AD was diagnosed according to the revised NINCDS-ADRDA criteria [2], and probable DLB was diagnosed according to consensus guidelines [11]. Subjects were labeled as SCD when they experienced and presented with cognitive complaints, but diagnostic workup was not abnormal and no other neurological or psychiatric disorder known to cause cognitive problems could be diagnosed [11]. These subjects were included as controls.

The EEG data set of the present study population has been previously analyzed focusing on functional and directed connectivity and network topology in DLB and AD [16,17].

2.2. Feature selection

All the non-EEG features (Table 1) for the classification algorithm were manually selected from the diagnostic workup based on availability, and their correspondence with the clinical criteria of DLB and AD [1,2].

2.2.1. Clinical features

Visual hallucinations were assessed with the neuropsychiatric inventory (NPI) [18]. Extrapyramidal signs were assessed by a preformatted checklist and defined as the presence of bradykinesia, rigidity, or tremor. Cognitive functions were assessed using a standardized test battery [11]. From this, the mini-mental state examination (MMSE) was used as a measure of global cognitive function [19], trail making test part A (TMT-A) as a measure of motor speed [20], the visual association test (VAT) as a measure of episodic memory [21], and the forward and backward condition of the Digit Span as a measure of attention [22].

2.2.2. Biomarkers

CSF was collected by lumbar puncture [11]. Amyloid-β1–42 (Aβ42), total tau, phosphorylated tau (p-tau), and a ratio of tau to Aβ42 were included as features [23]. From neuroimaging, medial temporal lobe (MTA) atrophy, global cortical atrophy, and white-matter hyperintensities on MRI were included as features [11].

2.2.3. EEG recordings

As part of the diagnostic workup, all subjects underwent a 20-minute no-task, resting-state EEG recording with OSG digital equipment (Brainlab; OSG B.V. Belgium), according to the international 10–20 system [17].

EEGs of all subjects were rated according to a standard visual rating scheme [24]. The visual rating includes the severity of EEG abnormalities on a 4-point rating scale and the presence of focal, diffuse, and epileptiform abnormalities [11,24]. In addition, all EEGs were assessed for the presence of frontal intermittent rhythmic delta activity (FIRDA) [9,10].

Subsequently, four artifact-free epochs, recorded in an awake state with eyes closed, were visually selected for each subject. Data were converted to American Standard Code for Information Interchange (ASCII) format, and 4 epochs of 4096 samples per subject (i.e., approximately 4 × 8 s EEG data per subject, sufficient to perform qEEG analyses [25]) were loaded into the BrainWave software for further analysis (BrainWave version 0.9.152.2.17, C. J. Stam; available for download at http://home.kpn.nl/stam7883/brainwave.html).

The machine-learning module of BrainWave was used to create a data file containing all the qEEG features shown in Table 1. Phase transfer entropy (PTE) was used as a measure for effective connectivity between EEG channels. PTE measures the strength and direction of phase-based functional connectivity between interacting oscillations [26]. In addition, minimum spanning tree (MST) measures (i.e., highest degree, leaf number, and tree hierarchy) were used as a representation of functional network topology. MST is a unique acyclic subnetwork that connects all nodes in a network such that only the strongest connections in the network are included without forming loops [27].
2.3. Data handling

All the measures from clinical and neuropsychological data, CSF and neuroimaging biomarkers, and visual EEG rating (Table 1) were added to the data file with qEEG data per subject. Missing data were imputed by the average value over the two tested diagnostic groups for a particular feature. Features with ≥33% missing values were excluded from analyses.

2.4. Classification algorithm

The classification algorithm was built in the machine-learning module of BrainWave. The random forest approach was used to build a classifier to differentiate between DLB and AD, DLB and controls, and AD and controls. Each decision tree in the random forest is built using a bootstrap sample (i.e., new training set), with replacement, from the original data (i.e., training set). Each new training set of features is randomly drawn from the original data set of features. This bootstrap aggregating (i.e., bagging) and random feature selection help in reducing the variance of the model, avoid overfitting, and result in uncorrelated trees [15]. Consequently, in random forest, the cross-validation is done internally, and there is no need for a separate test set to estimate the generalization error of the training set [15].

The two random forest parameters, namely mTry (i.e., the number of input variables randomly chosen at each split...
calculated by the square root of number of features) and nTree (i.e., the number of trees to grow for each forest) were set to 8 (square root of 61 features) and 500, respectively. Interestingly, the classification outcome is not highly sensitive to the choice of these parameters [28].

In every classification, each feature receives a variable importance (VIMP) score between 0 and 1. For each analysis, it was possible to manually include and exclude features from being used in the tree. By doing so, it was possible to determine the performance of the classifier for various combinations of clinical and EEG features.

Three performance metrics, accuracy, sensitivity, and specificity, were used to assess the performance of the random forest in discriminating DLB, AD, and controls.

Additional methodologic details are provided in the Supplementary Material.

3. Results

3.1. Baseline characteristics

The baseline characteristics of the three groups are shown in Table 2. DLB patients more frequently used medication affecting the central nervous system (CNS), compared to AD and controls (P < .05). EEG power in theta and alpha1 band and peak frequency differed between the three groups (P < .05), whereas power in the alpha2 and beta band differed between DLB and controls, and between DLB and AD, but not between AD and controls.

3.2. Classifier results

For all three data sets (DLB vs. AD, DLB vs. controls, AD vs. controls), performance of the classifier using different combinations of clinical and/or EEG features is shown in Table 3. An example of the machine-learning output is shown in Fig. 1.

It was possible to discriminate between DLB and AD with an accuracy of 87%, a sensitivity of 88%, and a specificity of 86%, when all features were included. qEEG increased the accuracy of the combination of clinical, biomarker, and visual EEG features with 7% (Table 3). Differentiation between DLB and controls was possible with an accuracy of 94% and a sensitivity and specificity of 95% and 93%, respectively. For differentiation between AD and controls, an accuracy of 91% with a sensitivity of 92% and a specificity of 91% was achieved.

3.3. Feature importance

Fig. 2 shows the VIMP scores of the features in the group analyses when all features were included. For discrimination between DLB and AD, EEG highest beta power was the most important feature, followed by mean beta power. Using highest beta power as the only feature resulted in an accuracy, sensitivity, and specificity of 71%. Clinical features, MRI, and CSF biomarkers were of limited value in the discrimination between DLB and AD (Table 3). qEEG solely had a higher diagnostic value in the differentiation between DLB and AD than the combination of the other multimodal variables without qEEG.

For DLB and controls, the qEEG measure theta/alpha ratio was the most important discriminating feature, followed by the visual EEG feature “diffuse abnormalities”. Using theta/alpha ratio as the only feature resulted in an accuracy and specificity of 83% with a sensitivity of 82%.

For discrimination between AD and controls, the clinical feature MMSE showed the highest VIMP score, followed by VAT, TMT-A, and CSF Tau/Aβ42 ratio. MMSE, solely, provided an accuracy of 87% with a sensitivity of 93% and specificity of 81%.

4. Discussion

This study is the first to combine visual and qEEG measures with multimodal diagnostic tests in a machine learning–based classifier. For all three groups (DLB vs. AD, DLB vs. controls, and AD vs. controls), reasonable diagnostic accuracies (>85%) could be achieved when all preselected diagnostic variables were used. However, when studying variable importance, different “profiles” were found. (q)EEG features were identified as the most important for discrimination between DLB and AD and DLB and controls. Interestingly, the accuracy of the classifier for discrimination between DLB and AD was higher when only qEEG features were used, than with a combination of clinical features (including MRI and CSF analysis) and visual EEG. When discriminating AD from controls, cognitive tests (e.g., MMSE) were more valuable. (q)EEG did not have additional value for this discrimination.

4.1. DLB vs. AD and controls

To date, little evidence is available on machine-learning classifiers for the diagnosis of DLB [12]. Nonautomated (q)EEG as a diagnostic modality has been more extensively studied and seems valuable for this diagnosis [9,10,29,30]. Visual EEG abnormalities are a supportive feature in the clinical criteria for DLB [1]. The present findings strongly support the potential of (q)EEG for the differentiation between DLB and AD.

Although the relative power in the delta, alpha2, and beta band was significantly different between DLB and AD, power in the beta band turned out to be the most important differentiating feature. This finding is remarkable, as beta power has not been reported previously to have specific discriminative value between these two forms of dementia. Medication use does not seem to be a likely underlying cause, because most medication that was used in the DLB group (e.g., cholinesterase inhibitors) would be expected to cause an increase in beta power [31], while instead a decrease was found. Likewise,
Power was presented as time needed to complete the task; higher scores mean worse performance. Diagnoses, including “subjective cognitive decline” for the control group, were assessed using the NPI. Cognitive fluctuations, extrapyramidal signs, and RBD were qualitatively assessed on their presence or absence at the first clinical presentation. Fazekas score is a measure of white-matter hyperintensities on T2-weighted fluid-attenuated inversion recovery (FLAIR) imaging.

Muscle artifacts cannot explain the difference since motor symptoms in DLB would have resulted in increased beta power. A more plausible explanation is that defective dopaminergic networks in DLB, that are intact in AD, could be related to the lower beta power found in this group and therefore be a cause for the high discriminative value of this EEG measure. Previous studies have linked changes in beta power, and beta peak frequency to linked changes in beta power, and beta peak frequency to linked changes in beta power, and beta peak frequency.
PD, PD dementia, and dopaminergic medication [32–34]. Furthermore, beta power can also be influenced by the cholinergic system [35,36]. Both the cholinergic system and the beta band have been related to the processes of attention [16,37]. The cholinergic system is more severely impaired in DLB brains than in AD brains [38,39], and this cholinergic deficit and associated attentional deficits might also be a discriminating aspect between the two types of dementia. Finally, the lower beta power in DLB could be caused by the overall shift in EEG activity from higher to lower frequency bands in DLB that has been shown by previous work [9,10,29].

For DLB and controls, theta/alpha ratio was the most important discriminating feature. In DLB, theta power is higher, and alpha power is lower than in AD and controls (Table 2). Therefore, EEG slowing seems to be more remarkable in DLB, which is in line with previous results [9,30]. The greater EEG slowing in DLB makes theta/alpha ratio a potentially important discriminating factor between DBL and controls.

### 4.2. AD vs. controls

For classification between AD and controls, an accuracy of 91% is in line with accuracies reported earlier in machine learning–based classification techniques for AD [13,40,41]. Although most of these studies were imaging based, some studies used EEG data. Established neuropsychological tests and CSF biomarkers [23] have high VIMP scores in the present classifier, which is in accordance with their importance in current clinical decision making. However, the observed accuracy of 63% when including only EEG features is low compared to previous studies on qEEG-based classifiers reporting accuracies of 80%–95% [40,41]. A possible explanation for this could be the use of healthy

### Table 3

<table>
<thead>
<tr>
<th>Group and feature selection</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLB vs. AD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All features</td>
<td>87</td>
<td>88</td>
<td>86</td>
</tr>
<tr>
<td>Only clinical features</td>
<td>66</td>
<td>65</td>
<td>67</td>
</tr>
<tr>
<td>Clinical features + biomarkers</td>
<td>71</td>
<td>71</td>
<td>70</td>
</tr>
<tr>
<td>Clinical features + biomarkers + visual EEG</td>
<td>78</td>
<td>76</td>
<td>80</td>
</tr>
<tr>
<td>Quantitative and visual EEG</td>
<td>85</td>
<td>86</td>
<td>84</td>
</tr>
<tr>
<td>Only quantitative EEG</td>
<td>85</td>
<td>86</td>
<td>85</td>
</tr>
<tr>
<td>DLB vs. controls</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All features</td>
<td>94</td>
<td>95</td>
<td>92</td>
</tr>
<tr>
<td>Only clinical features</td>
<td>89</td>
<td>92</td>
<td>86</td>
</tr>
<tr>
<td>Clinical features + biomarkers</td>
<td>86</td>
<td>87</td>
<td>85</td>
</tr>
<tr>
<td>Clinical features + biomarkers + visual EEG</td>
<td>90</td>
<td>87</td>
<td>94</td>
</tr>
<tr>
<td>Quantitative and visual EEG</td>
<td>91</td>
<td>93</td>
<td>89</td>
</tr>
<tr>
<td>Only quantitative EEG</td>
<td>92</td>
<td>95</td>
<td>89</td>
</tr>
<tr>
<td>AD vs. controls</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All features</td>
<td>91</td>
<td>92</td>
<td>91</td>
</tr>
<tr>
<td>Only clinical features</td>
<td>90</td>
<td>93</td>
<td>88</td>
</tr>
<tr>
<td>Clinical features + biomarkers</td>
<td>93</td>
<td>94</td>
<td>92</td>
</tr>
<tr>
<td>Clinical features + biomarkers + visual EEG</td>
<td>93</td>
<td>93</td>
<td>92</td>
</tr>
<tr>
<td>Quantitative and visual EEG</td>
<td>63</td>
<td>62</td>
<td>64</td>
</tr>
<tr>
<td>Only quantitative EEG</td>
<td>62</td>
<td>63</td>
<td>62</td>
</tr>
</tbody>
</table>

### Abbreviations

AD, Alzheimer’s disease; DLB, dementia with Lewy bodies; EEG, electroencephalography.

NOTE. Clinical features include hallucinations, extrapyramidal signs, and neuropsychological test results (mini mental state [MMSE] score; visual association test [VAT] score; Trail-making-test [TMT]-A score; and digit span forward and backward); Biomarkers include MRI (medial temporal lobe atrophy [MTA] score; global cortical atrophy [GCA] score; Fazekas score), and cerebrospinal fluid (CSF) (amyloid-β 1–42 (Aβ42); total Tau, phosphorylated Tau (p-Tau), and tau to Aβ1-42 ratio) data.
subjects as controls in these studies [40,41]. In the present study, subjects with SCD were used as controls, and possibly, they have more EEG abnormalities than healthy elderly without SCD. Furthermore, earlier studies have shown that EEG results in AD can be heterogeneous, abnormalities are less profound than in DLB, and normal EEGs frequently occur [29,39]. These factors could all be contributing to the lower accuracy found.

4.3. Strengths and limitations

A strength of this study is that the random forest algorithm produces a highly accurate classifier (with cross-validation build into the method and only two required parameters, none of which is critical for the results). The method is easy to use, has a high interpretability, runs efficiently on large data sets, and gives estimates of which variables are important in the classification [14]. Moreover, an essential part of the feature selection is done internally in the random forest algorithm and thus helps in reducing the variance of the model and avoids overfitting [15]. Second, this study compared a relatively large group of DLB patients with a carefully matched group of controls and AD patients. Third, subjects with SCD were used as a control group. These subjects visited the memory clinic with subjective memory complaints, and therefore represent a heterogeneous group. In clinical practice, this group is the exact population from which patients with a diagnosis of dementia need to be distinguished.

This study has some limitations. First, not all possibly relevant variables (e.g., REM sleep behavior disorder, cognitive fluctuations, results of DAT-SPECT scans and EEG variability) were available or scored quantitatively to be included as features in the classification algorithm. Second, some features had a relatively high number of missing values, which were excluded from the analysis or imputed. For instance, cognitive fluctuations and REM sleep behavior disorder were excluded from the analyses for this reason. In the case of CSF biomarkers, hallucinations, and extrapyramidal signs, missing values were imputed. Notably, CSF biomarkers turned out to be important discriminating features between AD and controls. This not only implies that CSF biomarkers could have had a higher VIMP score if no data had been missing, but also that when a continuous feature is important in distinguishing two groups, its performance is not fully influenced by the number of missing values. In contrast, in the case of categorical features, it is more difficult to impute the missing values with a meaningful average. Therefore, missing data in these types of features can result in an underestimation of the importance of categorical features.

Finally, initial selection of available features by the authors could have resulted in circular reasoning by including features that are already known to be important discriminating variables between two groups, whereas there is no autopsy confirmed diagnosis as a gold standard. Nonetheless, during a follow-up period of 0–7 years, none of the DLB diagnoses and only two AD diagnoses (mean follow-up of 21.2 months) were changed (the diagnosis was converted in 15 SCD subjects).

In summary, machine learning–based diagnostic classifiers show that qEEG is a valuable contribution in differentiating between DLB and AD. EEG data are easily obtained, as EEG is a low-cost and noninvasive procedure. Further research should elucidate the diagnostic value of relative beta power and theta/alpha ratio for the diagnosis of DLB. VIMP scores quantify the importance of a variable but do not provide information about the actual value, such as a cutoff, and this needs to be studied separately. Still, the current findings suggest a bigger role for (q)EEG in the diagnostic criteria for DLB.

Acknowledgments

The authors thank Dr. H.d.Waal for her contribution in the EEG data collection.

This study was partly supported by ZONMW TOP grant 40-00812-98-13009.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.dadm.2016.07.003.
References


