Learning Predictive Linguistic Features for Alzheimer's Disease and related Dementias using Verbal Utterances

Sylvester Olubolu Orimaye
Intelligent Health Research Group
School of Information Technology
Monash University Malaysia
sylvester.orimaye@monash.edu

Jojo Sze-Meng Wong
Intelligent Health Research Group
School of Information Technology
Monash University Malaysia
jojo.wong@monash.edu

Karen Jennifer Golden
Jeffrey Cheah School of
Medicine and Health Sciences
Monash University Malaysia
karen.golden@monash.edu

Abstract

Early diagnosis of neurodegenerative disorders (ND) such as Alzheimer's disease (AD) and related Dementias is currently a challenge. Currently, AD can only be diagnosed by examining the patient's brain after death and Dementia is diagnosed typically through consensus using specific diagnostic criteria and extensive neuropsychological examinations with tools such as the Mini-Mental State Examination (MMSE) or the Montreal Cognitive Assessment (MoCA). In this paper, we use several Machine Learning (ML) algorithms to build diagnostic models using syntactic and lexical features resulting from verbal utterances of AD and related Dementia patients. We emphasize that the best diagnostic model distinguished the AD and related Dementias group from the healthy elderly group with 74% F-Measure using Support Vector Machines (SVM). Additionally, we perform several statistical tests to indicate the significance of the selected linguistic features. Our results show that syntactic and lexical features could be good indicative features for helping to diagnose AD and related Dementias.

1 Introduction

Ageing and neurodegeneration can be a huge challenge for developing countries. As ageing population continues to increase, government and health care providers will need to deal with the associated economic and social effects such as an increased dependency ratio, higher need for social protection, and smaller workforce. The significance of this increase and demographic transition is a high prevalence of neurodegenerative diseases such as AD and related Dementias. According to Kalaria et al. (2008), 71% of 81.1 million dementia related cases have been projected to be in the developing countries with annual costs of US$73 billion.

Alzheimer's disease is the most common form of dementia (Ballard et al., 2011). However, early diagnosis of dementia is currently challenging, especially in the earlier stages. Dementias have been typically diagnosed through extensive neuropsychological examinations using a series of cognitive tests containing set questions and images (Williams et al., 2013). For example, the MMSE screening tool is composed of a series of questions and cognitive tests that assess different cognitive abilities, with a maximum score of 30 points. A MMSE score of 27 and above is suggestive of not having a Dementia related disease. The challenge with these cognitive tests is that the accuracy depends on the clinician's level of experience and their ability to diagnose different subtypes of the disease as Dementia disease can be classified further into Alzheimer's disease, Vascular Dementia, Dementia with Lewy bodies (DLB), Mixed dementia, Parkinson's disease, as well as other forms[1].

As such, this paper investigates effective computational diagnostic models for predicting AD and related Dementias using several linguistic features extracted from the transcribed verbal utterances produced by potential patients. The premise is that, neurodegenrative disorders (ND) are characterized by the deterioration of nerve cells that control cognitive, speech and language processes, which consequentially translates to how patients compose verbal utterances. Thus, we proposed the diagnostic models using Machine Learning (ML) algorithms that learn such linguistic features and classify the AD and related Dementias group from the healthy elderly group.

[1] http://www.alz.org/dementia/types-of-dementia.asp

Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 78–87,
Baltimore, Maryland USA, June 27, 2014.

2 Related Work

Few ML algorithms have been proposed to automate the diagnosis of Dementias using linguistic features. In a recent study, Williams et al. (2013) experimented with different ML algorithms for learning neuropsychological and demographic data which are then used for the prediction of Clinical Dementia Rating (CDR) scores for different sub-types of Dementia and other cognitive impairments. In that study, four ML algorithms were used comprising of Naïve Bayes (NB), C4.5 Decision Trees (DT), Neural Networks with back-propagation (NN), and Support Vector Machines (SVM). The study reports NB with the highest classification accuracy; however, its accuracy could be biased as the same NB was used for the initial feature selection for all the four ML algorithms. As such, the feature sets would have been optimized for NB.

In another study, Chen and Herskovits (2010) proposed different diagnostic models that distinguished the very mild dementia (VMD) group from the healthy elderly group by using features from structural magnetic-resonance images (MRI) to train seven ML algorithms. Their study reported that both SVM and Bayesian Networks (Bayes Nets) gave the best diagnostic models with the same accuracy of 80%. Similarly, a study by Klöppel et al. (2008) reported a better accuracy with SVM on the scans provided by radiologists. In contrast, we study several linguistic features from the transcribed verbal utterances of AD and related Dementia patients. We emphasize that the proposed diagnostic models do not depend on the complex MRI scan processes but a simple verbal description of familiar activities in order to diagnose the disease.

A closely related work to ours is Garrard et al. (2013) research. The study used Naïve Bayes Gaussian (NBG) and Naïve Bayes multinomial (NBM) to classify textual descriptions into a Dementia group and a healthy elderly group. The Information Gain (IG) feature selection algorithm was used in both cases and both algorithms achieved a better accuracy of up to 90% with features such as low frequency content words and certain generic word components. In this paper, we study more exclusive syntactic and lexical features that could distinguish the AD and related Dementia patients from the healthy group. In addition, we build several models by experimenting with different ML algorithms rather than NB alone.

Similarly, Roark et al. (2011) demonstrated the efficacy of using complex syntactic features to classify mild cognitive impairment (MCI) but not AD and Dementia. Also, de Lira et al. (2011) investigated the significance of lexical and syntactic features from the verbal narratives of AD patients by performing several statistical tests based on 121 elderly participants comprising of 60 AD subjects and 61 healthy subjects. Their lexical features comprised of word-finding difficulties, immediate word repetition of isolated words, word revisions, semantic substitutions, and phonemic paraphasias. For syntactic features, coordinated sentences, subordinated sentences, and reduced sentences were examined. Upon performing and making comparison between the parametric Student's t-test (t) and the non-parametric Mann-Whitney test (U), only word-finding difficulties, immediate repetitions, word revisions, coordinated sentences, and reduced sentences were found to be statistically significant with p = 0.001 at a 95% confidence interval (CI). Further post-hoc analysis with the Wald test (Wald X^2) showed that immediate word repetitions, word revisions, and coordinated sentences could be used to distinguish AD patients from the healthy elderly group.

While de Lira et al. (2011) did not perform any evaluation using ML algorithms, we focus on the feasibility of effectively diagnosing AD and related Dementias by learning additional syntactic and lexical features with different ML algorithms. According to Ball et al. (2009), syntactic processing in acquired language disorders such as Aphasia in adults, has shown promising findings, encouraging further study on identifying effective syntactic techniques. Similarly, Locke (1997) emphasized the significance of lexical-semantic components of a language, part of which is observable during utterance acquisition at a younger age. Locke highlighted further that as the lexical capacity increases, syntactic processing becomes automated, hence leading to changes in language. As such, it is almost certain that the effects of a specific language disorder could be observed as changes to the lexical and syntactic processes governing language and verbal utterances.

In this paper, we identify several syntactic and lexical features in addition to the significant features studied by de Lira et al. (2011) and then train five different ML models to predict the like-

lihood of a patient having Dementia. First, we extract predictive syntactic and lexical features from the existing DementiaBank[2] corpus containing a set of transcribed texts from verbal utterances produced by AD and related Dementia patients living in the United States. The transcribed texts are stored in the CHAT system format in the DementiaBank corpus made available by the School of Medicine of the University of Pittsburgh as part of the TalkBank project[3]. We further extract several lexical and syntactic features from the CHAT format and conduct different statistical tests and then learn and evaluate with different ML algorithms. We emphasize that the best model accuracy reported in our study is comparable to the accuracy reported in Garrard et al. (2013) and outperforms a model using only the three significant features reported in de Lira et al. (2011).

The rest of this paper is organized as follows. We present the methodology used in this study in Section 3. The DementiaBank dataset and the participants are described in Section 3.1 and Section 3.2 respectively. Section 4 discusses the feature extraction process that extracts both the lexical and syntactic features used in this study. In Section 5, we perform statistical tests to understand the significant features. Section 6 performs additional feature selection and make comparison with the statistical test results. We discuss the ML models used in this study in Section 7. Finally, results, discussion and conclusion are presented in Section 8, 9, and 10.

3 Methods

It is common in clinical research to conduct investigation on the actual patients (or subjects). This process can be achieved over a period of time; however, previous research studies have made available series of clinical datasets that reduce the investigation time considerably. Although, this study does not involve direct interaction with actual patients, we focus on understanding the linguistic patterns from the verbal utterances of existing patients. In Section 2, we have discussed those verbal utterances to be present in the transcription files contained in the DementiaBank dataset and we will describe the dataset further in Section 3.1. In this study, our focus is to use the extended syntactic and lexical features from the transcripts and compare to the features established in de Lira et al. (2011) as our baseline. We identified 21 features including the 3 significant features investigated in de Lira et al. (2011). 9 of those features are syntactic, 11 are lexical features, and 1 is a confounding feature (age). We will describe the features in detail in Section 4. Our feature extraction is followed by statistical tests as performed in de Lira et al. (2011). Both the Student's t-test (t) and the Mann-Whitney test (U) are performed and followed by multiple logistic regression (MLR) that shows the most significant features. In addition, we also perform feature selection using the Information Gain algorithm and compare our results to those achieved by MLR. The final ML models are built using SVM, NB, Bayes Net, DT, and NN.

3.1 Datasets

In this study, an existing DementiaBank clinical dataset was used. The dataset was created during a longitudinal study conducted by the University of Pittsburgh School of Medicine on Alzheimer's and related Dementia and funded by the National Institute of Aging[4]. The dataset contains transcripts of verbal interviews with AD and related Dementia patients, including those with MCI. Interviews were conducted in the English language and were based on the description of the Cookie-Theft picture component which is part of the Boston Diagnostic Aphasia Examination (Kaplan et al., 2001). During the interview, patients were given the picture and were told to discuss everything they could see happening in the picture. The patients' verbal utterances were recorded and then transcribed into the CHAT transcription format (MacWhinney, 2000). Thus, in this study, we extract the transcribed patient sentences from the CHAT files and then pre-process the sentences for feature extraction.

3.2 Participants

The participants in the DementiaBank dataset have been categorized into Dementia, Control, and Unknown patient groups. Our study uses only the Dementia and Control groups as we are interested in the binary diagnosis of the AD and related Dementias. Thus, the Dementia group consists of 314 elderly patients with an approximate age range of 49 to 90 years. The group consists of 239 peo-

[2] http://talkbank.org/DementiaBank/

[3] http://www.talkbank.org/browser/index.php

[4] http://www.nia.nih.gov/

ple diagnosed with probable AD; 21 with possible AD; 5 with Vascula Dementia (VD); 43 with MCI; 3 with Memory problem and 4 other people with an unidentified form of dementia. On the other hand, the Control group consists of 242 healthy elderly without any reported diagnosis and with approximate age range of 46 to 81 years. In order to have a balanced number of participants across groups, we reduced the AD and related Dementias group to the first 242 patients consisting of 189 probable AD, 8 possible AD, 37 MCI, 3 memory problems, 4 Vascular dementia, and 1 other participant with an unidentified form of dementia. In addition, some demographic information was made available in the DementiaBank dataset, however, we have only selected age in order to measure the significance of the disease with respect to age.

4 Features Extraction

Several features were extracted from the transcript files. First, we extracted every CHAT symbol in the transcript files and stored them according to their frequencies and positions in each sentence. We emphasize that some CHAT symbols represent both explicit and implicit features that describe the lexical capability of the patient. For example, having the CHAT symbol [//] at a specific position within a sentence implies that the patient was retracing a verbal error that precedes that position and at the same time attempting to make correction, while the CHAT symbol [/] shows the patient making immediate word repetition (MacWhinney, 2000). On the other hand, it is non-trivial to extract the syntactic features without performing syntactic parsing on the sentences. As such, using the Stanford Parser Klein and Manning (2003), we generated the syntactic tree structure of each sentence and extract features as appropriate.

4.1 Syntactic features

As described below, we investigated a number of features that are seen to demand complex syntactic processing, including the three syntactic features (*coordinated*, *subordinated*, and *reduced* sentences) evaluated by de Lira et al. (2011) and the *Dependency distance* feature evaluated by Roark et al. (2011) and Pakhomov et al. (2011). All syntactic features are extracted from the syntactic tree structures produced by the Stanford Parser.

- Coordinated sentences: Coordinated sentences are those whose clauses are combined using coordinating conjunctions. The number of occurrence for this feature per patient narrative is obtained based on the frequency of the coordinating conjunction PoS tag (`CC`) detected in the parse tree structure.

- Subordinated sentences: Subordinated sentences are those that are subordinate to the independent primary sentence to which they are linked. Similarly, the number of occurrence for this feature per patient narrative is obtained based on the frequency of the subsentences indicated by the PoS tag (`S`) detected in the parse tree structure.

- Reduced sentences: Following the definition set out by de Lira et al. (2011), this feature represents those subordinated sentences without a conjunction but with nominal verb forms (which are either participles or gerund). To obtain the count for this feature, the frequencies of PoS tags (`VBG` and `VBN`) are used.

- Number of predicates: The number of predicates found in every patient's narrative can be seen as another estimation of the sentence complexity. The predicates are extracted using a rule-based algorithm that locates transitive verbs which are followed by one or more arguments. We emphasize that the importance of predicate-argument structures has been explored in the literature for text classification tasks (Surdeanu et al., 2003; Orimaye, 2013).

- Average number of predicates: The average number of predicates per patient narrative is investigated as well to study its effect.

- Dependency distance: This feature was used in the study of Pakhomov et al. (2011) as a way to measure grammatical complexity in patients with Alzheimer's disease. The distance value is calculated based on the sum of all the dependency distances, in which each dependency distance is the absolute difference between the serial position of two words that participate in a dependency relation.

- Number of dependencies: For a purpose similar as to the syntactic dependency distance, the number of unique syntactic dependency

relations found in every patient's narrative is examined.

- Average dependencies per sentence: We also consider the average number of the unique dependency relations per sentence.

- Production rules: Production rules derived from parse trees has been explored in a number of NLP related classification tasks (Wong and Dras, 2010; Post and Bergsma, 2013). We investigate this feature by counting the number of unique production rules in the context-free grammar form extracted from each patient's narrative.

4.2 Lexical features

The lexical features used in this study include the *revision* and *repetition* features proposed in Croisile et al. (1996) and evaluated in de Lira et al. (2011). The remaining features are additionally investigated lexical features that show better improvement with our models.

- Utterances: The total number of utterances per patient was computed. Each utterance is identified to start from the beginning of a verbal communication to the next verbal pause length, such as punctuation or a CHAT symbol that represents a specific break in communication (Marini et al., 2008). A sentence could have one or more utterances, and an utterance could be one word, a phrase or a clause. It has been identified that utterance acquisitions form a grammatical lexicon for a language (Locke, 1997). Thus, we hypothesize that the absolute number of utterances in a conversation could show the language strength of a potential patient.

- Mean Length of Utterances (MLU): We measure the structural organization of sentences using the MLU. This was computed as the ratio of the total number of words to the number of utterances (Marini et al., 2008). MLU has been specifically used to measure grammar growth in children with Specific Language Impairment (SLI) (Yoder et al., 2011). In this study, we investigate the significance of MLU in determining language disorder in AD and related Dementias.

- Function words: We compute the total number of function words in the patient's narrative. Function words enable sentences to have meaning and they have been studied as an essential attribute to brain and language processing (Friederici, 2011).

- Unique words: We measure the total number of unique words as the absolute word count minus the number of immediate repeated words.

- Word count: This is measured as the absolute word count including repeated words.

- Character length: We measure the absolute character length of the patient's narrative.

- Total sentences: This is the absolute count of sentences in the patient's narrative.

- Repetitions: This is measured as the number of immediate word repetitions in the patient's narrative (de Lira et al., 2011; Croisile et al., 1996).

- Revisions: This feature is measured as the count of pause positions where the patient retraced a preceding error and then made a correction (MacWhinney, 2000; de Lira et al., 2011; Croisile et al., 1996).

- Lexical bigrams: We take into account the number of unique bigrams in a patient's narrative in order to capture repeated bigram patterns.

- Morphemes: To capture the morphology structure of the patient's narrative, we measured the number of morphemes. Each morpheme represents a word or a part of it that cannot be further divided (Creutz and Lagus, 2002).

5 Statistical Evaluation

One of the challenges that we encountered in evaluating the features above is that some features are not normally distributed. An exception to that is the confounding feature "age". For age, it is our assumption that the DementiaBank study was designed to cover normally distributed participants in terms of age range. For the other generated features, it is understandable, since each patient would give specific attributes that show the severity of the disease overtime. As such, we performed one parametric test (Student's t-test (t)) and one

non-parametric test (Mann-Whitney test (U)) and then compared the results of the two tests similar to the baseline paper (de Lira et al., 2011). Both results achieved the same results as shown in Table 1; thus, we chose the parametric results for further statistical evaluation.

Further, we conducted a post-hoc test using multiple logistic regression analysis in order to identify specific features that distinguish the AD and related Dementias group from the healthy elderly group. We present the results of the analysis using the Wald test (Wald X^2) and the Odds Ratio or $Exp(B)$ as shown in Table 2. A 95% confidence interval (CI) was computed for both lower and upper bound of $Exp(B)$ and $p < 0.05$ shows statistical significance. All tests performed are two-tailed using the IBM Statistical Package for the Social Sciences (SPSS) version 20.0.0[5].

The result of our analysis is in agreement with the study conducted by de Lira et al. (2011); however, we examined more features in our study. Our analysis shows that the statistically significant syntactic features of the ADAG have *lower* mean compared to the HAG. This indicates that the disease group have difficulties in constructing complex sentences unlike the healthy group. We suggest that effective use of predicates and reduced structures could be of vital importance to appropriately measure healthy language in Alzheimer's disease and related Dementia patients. On the other hand, statistically significant lexical features of the ADAG have *higher* mean compared to the HAG, except for MLU with just 0.91 difference. This makes sense, for example, the disease group performed more immediate word repetitions and made more revisions on grammatical errors in their narrative. More utterances were also noticed with the disease group as they tend to make several pauses resulting from syntactic errors and attempts to correct or avoid those errors in the first place.

The multiple logistic regression analysis indicates that number of utterances, reduced sentences, MLU, revisions, and number of predicates significantly distinguish the disease group from the healthy elderly group leaving out repetitions and average predicates per sentence. Interestingly, repetitions was found to be significant in de Lira et al. (2011), albeit with just 121 patients. In our case, we suspect that repeated words could be less common with both groups given the combined 484 patients, while the absolute count of predicates in a discourse (not at the sentence level) could be more representative of the groups. The confounding feature age was used because of the age difference between ADAG and HAG. The resulting odd ratios OR emphasize the likelihood of having Alzheimer's and the related Dementia diseases when the distinguishing features are used. Lower β values for MLU, predicates, and reduced sentences decreases the likelihood of having Alzheimer's disease and related Dementias.

6 Feature Selection

To further support that the features selected through statistical testing from the previous section (Section 5) are indeed significant, one of the widely adopted metrics for feature selection in the ML-based text classification paradigm — Information Gain (IG) — is explored. We could adopt the feature selection approach taken by Williams et al. (2013), in which the subset of indicative features were selected based on a specific classifier, NB in their case; we chose to use IG instead given that the IG value for each feature is calculated independent of the classifiers and thus reduces the chance of bias in terms of the model performance. By ranking the IG values for each of the extracted features (both lexical and syntactic), the top eight features with the highest IG values are the same as the subset of the eight significant features identified through the statistical tests.

7 Machine Learning Models

In order to conduct an informed comparison with the findings from the previous related work, we evaluate the same four ML models investigated by Williams et al. (2013) which include Support Vector Machines (SVM) with radial basis kernel, Naïve Bayes (NB), J48 Decision Trees (DT), and Neural Networks (NN) with back propagation. In addition, Bayesian Networks (Bayes Nets), which has also been found useful in the work of Chen and Herskovits (2010), is also evaluated. Using the ML models, we performed three sets of experiments[6] to confirm the hypothesis that the identified significant syntactic and lexical features could give effective diagnostic models. First, we experimented with the three significant features reported in de Lira et al. (2011). Second, we performed

[5] http://www-01.ibm.com/software/analytics/spss/

[6] https://github.com/soori1/ADresearch

	ADAG MEAN(SD)	HAG MEAN(SD)	t	df	p	95% CI(Difference)
Syntactic features						
Coordinated sentences	5.21(3.51)	4.73(3.11)	1.59	482	0.11	-0.11 to 1.07
Subordinated sentences	5.37(3.41)	5.12(2.84)	0.85	482	0.40	-0.32 to 0.81
Reduced Sentences	3.24(2.47)	4.12(2.67)	-3.77	482	<0.000*	-1.34 to -0.42
Number of Predicates	5.77 (3.33)	7.03(3.63)	-3.99	482	<0.000*	-1.89 to -0.64
Avr.Predicates per sentence	0.46(0.19)	0.57(0.23)	-5.48	482	<0.000*	-0.15 to -0.07
Number of Dependencies	104.67(53.76)	104.12(50.20)	0.11	482	0.91	-8.75 to 9.83
Avr.dependency per sentence	8.84(2.71)	8.82(2.47)	0.09	482	0.932	-0.44 to 0.48
Dependency distance	18.57(8.71)	18.12(8.04)	0.59	482	0.56	-1.05 to 1.95
Production rules	128.36(50.68)	126.83(44.68)	0.35	482	0.73	-7.01 to 10.05
Lexical features						
Utterances	43.56(28.22)	32.31(15.42)	5.44	482	<0.000*	7.19 to 15.31
MLU	2.66(1.22)	3.57(1.31)	-7.87	482	<0.000*	-1.13 to -0.68
Function words	59.18(34.82)	58.98(32.46)	0.07	482	0.948	-5.81 to 6.21
Unique words	115.54(60.93)	116.17(55.61)	-0.12	482	0.905	-11.05 to 9.79
Word count	127.28(68.42)	127.25(63.24)	0.005	482	0.996	-11.74 to 11.79
Character length	567.01(303.59)	580.87(292.07)	-0.512	482	0.61	-67.07 to 39.35
Total sentences	13.24(7.03)	12.86(5.29)	0.67	482	0.502	-0.73 to 1.49
Repetitions	1.64(2.44)	0.64(0.99)	5.92	482	<0.000*	0.67 to 1.34
Revision	3.77(4.36)	1.93(2.22)	5.87	482	<0.000*	1.23 to 2.47
Lexical bigrams	104.84 (52.55)	106.79 (50.61)	-0.42	482	0.677	-11.17 to 7.26
Number of Morphemes	104.23(60.73)	107.90(55.74)	-0.694	482	0.488	-14.09 to 6.74

ADAG = Alzheimer's disease and related Dementia group (n=242); HAG = Healthy elderly group (n=242); SD = standard deviation; df = degree of freedom; CI = confidence Interval.

Table 1: Statistical analysis of linguistic features based on Student's t-test.

Features	β	S.E	Wald X^2	p	OR	95% CI of OR
Age	-0.11	0.02	39.53	<0.000*	0.90	0.87 to 0.93
Utterances	-0.03	0.01	5.55	0.018*	0.97	0.95 to 0.99
MLU	0.374	0.137	7.39	0.007*	1.45	1.11 to 1.90
No of Predicates	0.25	0.059	17.64	<0.000*	1.28	1.14 to 1.44
Revisions	-0.143	0.069	4.33	0.037*	0.87	0.76 to 0.99
Reduced Sentences	0.121	0.055	4.89	0.027*	1.129	1.01 to 1.26
Constant	5.23	1.18	19.67	<0.000*	187.25	-

ADAG, n=242; HAG, n = 242; S.E = standard error; OR = Odds ratio or Exp(β); CI = confidence Interval.

Table 2: Multiple logistic regression analysis on significant and confounding features.

an experiment with the eight significant features identified by the parametric test reported in Table 1. Finally, we used the six distinguishing features identified by MLR in Table 2.

Given the relatively small size of the dataset used in this study, we conduct a 10-fold cross validation on each of the ML models by using a balanced data set with 242 instances for each group: the AD and related Dementias group and the healthy (Control) group. Performance of the ML models were measured in terms of *precision*, *recall*, and *F-measure*. All the ML experiments including the IG ranking are conducted using the Weka toolkit[7] with the default settings.

[7] http://www.cs.waikato.ac.nz/ml/weka/

8 Results

The results of the three experiments are shown in Table 3, 4, and 5 respectively. In addition, Table 6 shows a summary of the performance of the best ML model (SVM) for predicting Alzheimer's disease and the related Dementia diseases.

Our results show that SVM gave better F-Measure and recall in most cases compared to other ML algorithms. Interestingly, DT, Bayes Nets, and NB showed better precision on the disease group using the 6 and 8 significant features. Specifically, using the 6 significant features, DT showed 78% precision but 69% recall on the disease group. Similarly, Bayes Nets showed 77% precision but 66% recall on the disease group. Overall, SVM takes the lead as it showed the highest F-Measure of 74% on the disease group with 75% precision and 73% recall.

Model	Precision (ADAG/HAG)	Recall (ADAG/HAG)	F−Measure (ADAG/HAG)
SVM	0.70/0.65	0.59/0.75	0.64/0.70
NB	0.72/0.57	0.34/0.87	0.47/0.69
DT	0.67/0.65	0.62/0.69	0.65/0.67
NN	0.70/065	0.60/0.74	0.64/0.69
Bayes Nets	0.66/0.68	0.71/0.64	0.68/0.66

Table 3: Results of different ML models using the three significant features reported in (de Lira et al., 2011) on both disease and healthy elderly groups.

Model	Precision (ADAG/HAG)	Recall (ADAG/HAG)	F−Measure (ADAG/HAG)
SVM	0.74/0.73	0.73/0.74	0.73/0.74
NB	0.77/0.62	0.46/0.86	0.58/0.72
DT	0.74/0.69	0.66/0.77	0.70/0.73
NN	0.75/0.72	0.69/0.77	0.72/0.74
Bayes Nets	0.75/0.69	0.65/0.78	0.70/0.73

Table 4: Results of different ML models using the eight statistically significant features in Table 1 on both disease and healthy elderly groups.

9 Discussion

The results of our ML experiments and statistical evaluations suggest that using ML algorithms by learning syntactic and lexical features from the verbal utterances of elderly people can help the diagnosis of Alzheimers and the related Dementia diseases. The outcome of our evaluations is similar to the study conducted in de Lira et al. (2011). However, our study identifies more indicative and representative linguistic features compared to de Lira et al. (2011). Furthermore, the results of our statistical evaluation agree with the feature selection results (using IG). That is, all the statistically significant features discussed in Section 5 are also the top ranked features using the IG feature selection algorithm in Section 6. Following the identification of additional linguistic features, we emphasize that the best ML model with six significant linguistic features (age, utterances, MLU, reduced sentences, revisions, and predicates) outperforms a three-feature model (repetitions, revisions, and coordinated sentence). More importantly, unlike de Lira et al. (2011), repetitions and coordinated sentences did not contribute to the accuracy of our diagnostic models. Finally, in comparison to Williams et al. (2013), SVM obtained the highest prediction accuracy, albeit on linguistic features. Moreover, unlike Williams et al. (2013), our feature selection process is independent of the best ML algorithm (SVM) in our case. Again, this avoids unnecessary bias especially in clinical diagnosis. A limitation of this study could be the use of a binary classification between a combined Dementia related diseases group with different subtypes (such as AD, MCI and memory problems) and a control group of healthy participants. Although MCI could sometimes (but not always) be a precursor to AD and Dementia, we suggest that it could be important to exclude patients with MCI and other minor memory problems from the AD and related Dementia patients in future study.

10 Conclusion and Future Work

We have investigated promising diagnostic models for Alzheimer's and the related Dementia diseases using syntactic and lexical features from verbal utterances. We performed statistical and ML evaluations and show that the disease group used less complex sentences than the healthy elderly group. Additionally, following our regression analysis, we show that the disease group makes more grammatical errors and at the same time makes reasonable attempts to correct or avoid those errors in the first place. We also emphasized that utterances, reduced sentences, MLU, revisions, and number of predicates, significantly distinguish the disease group from the healthy elderly group. In the future, we plan to investigate indexical cues, prosodic cues, and semantic cues in order to capture the perspectives in a patient's narrative. Furthermore, we intend to evaluate our models against the MMSE and MoCA diagnostic thresholds on actual AD and Dementia patients in a developing country. More importantly, there is a need to train the diagnostic models on a larger dataset, which

Model	Precision (ADAG/HAG)	Recall (ADAG/HAG)	F−Measure (ADAG/HAG)
SVM	0.75/0.74	0.73/0.76	0.74/0.75
NB	0.79/0.65	0.53/0.86	0.63/0.74
DT	0.78/0.71	0.69/0.76	0.71/0.73
NN	0.74/0.70	0.67/0.76	0.71/0.73
Bayes Nets	0.77/0.70	0.66/0.80	0.71/0.75

Table 5: Results of different ML models using the six statistically significant features in Table 2 on both disease and healthy elderly groups.

Model	Precision	Recall	F−Measure
6-feature	**0.75***	**0.73***	**0.74***
8-feature	0.74	0.73	0.73
3-feature(Baseline)	0.70	0.59	0.64

Table 6: Summary of SVM performance with the best predictive features for diagnosing AD and related Dementias.

could lead to better accuracy. Furthermore, longitudinal studies are recommended in order to improve sample sizes and follow the course of the disease overtime.

References

Martin J Ball, Michael R Perkins, Nicole Müller, and Sara Howard. 2009. *The handbook of clinical linguistics*, volume 56. John Wiley & Sons.

Clive Ballard, Serge Gauthier, Anne Corbett, Carol Brayne, Dag Aarsland, and Emma Jones. 2011. Alzheimer's disease. *The Lancet*, 377(9770):1019 – 1031.

Rong Chen and Edward H Herskovits. 2010. Machine-learning techniques for building a diagnostic model for very mild dementia. *Neuroimage*, 52(1):234–244.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6*, pages 21–30. Association for Computational Linguistics.

Bernard Croisile, Bernadette Ska, Marie-Josee Brabant, Annick Duchene, Yves Lepage, Gilbert Aimard, and Marc Trillet. 1996. Comparative study of oral and written picture description in patients with alzheimer's disease. *Brain and language*, 53(1):1–19.

Juliana Onofre de Lira, Karin Zazo Ortiz, Aline Carvalho Campanha, Paulo Henrique Ferreira Bertolucci, and Thaís Soares Cianciarullo Minett. 2011. Microlinguistic aspects of the oral narrative in patients with alzheimer's disease. *International Psychogeriatrics*, 23(03):404–412.

Angela D Friederici. 2011. The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392.

Peter Garrard, Vassiliki Rentoumi, Benno Gesierich, Bruce Miller, and Maria Luisa Gorno-Tempini. 2013. Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. *Cortex*.

Raj N Kalaria, Gladys E Maestre, Raul Arizaga, Robert P Friedland, Doug Galasko, Kathleen Hall, José A Luchsinger, Adesola Ogunniyi, Elaine K Perry, Felix Potocnik, et al. 2008. Alzheimer's disease and vascular dementia in developing countries: prevalence, management, and risk factors. *The Lancet Neurology*, 7(9):812–826.

Edith Kaplan, Harold Goodglass, Sandra Weintraub, Osa Segal, and Anita van Loon-Vervoorn. 2001. *Boston naming test*. Pro-ed.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430. Association for Computational Linguistics.

Stefan Klöppel, Cynthia M Stonnington, Josephine Barnes, Frederick Chen, Carlton Chu, Catriona D Good, Irina Mader, L Anne Mitchell, Ameet C Patel, Catherine C Roberts, et al. 2008. Accuracy of dementia diagnosisa direct comparison between radiologists and a computerized method. *Brain*, 131(11):2969–2974.

John L Locke. 1997. A theory of neurolinguistic development. *Brain and language*, 58(2):265–326.

Brian MacWhinney. 2000. *The CHILDES Project: The database*, volume 2. Psychology Press.

Andrea Marini, Ilaria Spoletini, Ivo Alex Rubino, Manuela Ciuffa, Pietro Bria, Giovanni Martinotti, Giulia Banfi, Rocco Boccascino, Perla Strom, Alberto Siracusano, et al. 2008. The language of schizophrenia: An analysis of micro and macrolinguistic abilities and their neuropsychological correlates. *Schizophrenia Research*, 105(1):144–155.

Sylvester Olubolu Orimaye. 2013. Learning to classify subjective sentences from multiple domains using extended subjectivity lexicon and subjective predicates. In *Information Retrieval Technology*, pages 191–202. Springer.

Serguei Pakhomov, Dustin Chacon, Mark Wicklund, and Jeanette Gundel. 2011. Computerized assessment of syntactic complexity in alzheimerś disease: A case study of iris murdochś writting. *Behavior Research Methods*, 43(1):136–144.

Matt Post and Shane Bergsma. 2013. Explicit and implicit syntactic features for text classification. In *Proceedings of the 51st Annual Meeting on Association for Computational Linguistics - Volume 2*, ACL '13, pages 866–872, August.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2081–2090.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 8–15. Association for Computational Linguistics.

Jennifer A Williams, Alyssa Weakley, Diane J Cook, and Maureen Schmitter-Edgecombe. 2013. Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia. In *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Sze-Meng Jojo Wong and Mark Dras. 2010. Parser features for sentence grammaticality classification. In *Proceedings of the Australasian Language Technology Association Workshop 2010*, pages 67–75, December.

Paul J Yoder, Dennis Molfese, and Elizabeth Gardner. 2011. Initial mean length of utterance predicts the relative efficacy of two grammatical treatments in preschoolers with specific language impairment. *Journal of Speech, Language, and Hearing Research*, 54(4):1170–1181.