Speech processing for early Alzheimer Disease diagnosis: Machine learning based approach

Randa Ben Ammar
Multimedia Information System and Advanced Computing Laboratory
Sfax university, Tunisia
randabenammar9@gmail.com

Yassine Ben Ayed
Multimedia Information System and Advanced Computing Laboratory
Sfax university, Tunisia
yassinebenayed@gmail.com

***Abstract*— Alzheimer's disease (AD) is a neurodegenerative disease characterized by the insidious onset of cognitive, emotional and language disorders. These attacks are sufficiently intense to affect the daily social and professional lives of patients. Today, in the absence of a reliable diagnosis and effective curative treatments, fighting this disease is becoming a real public health issue, prompting research to consider non-drug techniques. Among these techniques, speech processing is proving to be a relevant and innovative field of investigation. Several Machine Learning algorithms achieved promising results in distinguishing AD from healthy control subjects. Alternatively, many other factors such as feature extraction, the number of attributes for feature selection, used classifiers, may affect the prediction accuracy evaluation. To surmount these weaknesses, a model is suggested which include a feature extraction step followed by imperative attribute selection and classification is achieved using a machine learning classifiers. The current findings show that the proposed model can be strongly recommended for classifying Alzheimer's patient from healthy individuals with an accuracy of 79%.**

Keywords— Alzheimer's disease, Speech analysis, Machine learning, Feature selection.

I. INTRODUCTION

The disease was first described in November 1901 by a German neurologist Alois Alzheimer [1], the patient treated for the first time was Auguste Deter, a 51-year-old woman who suffered from memory problems, language and other psychological disorders (disorientation, hallucinations, etc.) and by 1906, she passed away at the age of 55. The patient's condition met the definition of what was then called dementia, but she was particularly young to present these symptoms, she was then diagnosed with "presenile dementia". The condition was first discussed in medical literature back in 1907 and it was named after Alzheimer in 1910.

Alzheimer disease is most frequent type of dementia, characterized by memory loss and additional cognitive disabilities that are serious enough to intervene with our daily life. Alzheimer's disease accounts for 60 to 80 percent of dementia cases.

The etiology of Alzheimer's disease is still unknown. Many epidemiological data indicate that it is the most common cause of dementia syndromes (70% according to [2]), it evolves over a period of 8 to 12 years or at maximum 2 to 20 years, with the progressive degradation of the brain that leads the patient to the demented state caused by many cognitive disorders. Nowadays we still incapable of preventing its appearance, nor stopping its evolution [3].

Globally, over a 35 million people are affected by the Alzheimer's disease. An estimated 5.5 million Americans are living with Alzheimer's dementia in 2017 and the number is growing in a fast rate. This number includes an estimated 5.3 million people age 65 and older, today, every 66 seconds; someone in the United States develops Alzheimer's dementia [4].

As the population is aging, 60 million people are expected to have the disease by 2030.

While deaths from heart disease (the leading cause of death) is decreasing, death due by Alzheimer is becoming a more common and it is recording a significant increase of 89%. It is considered as the sixth-leading cause of death that cannot be inhibited, cured or even slowed. Alzheimer diseases will only become more prevalent, and the resulting cost of Alzheimer's care will continue rising to unsustainable levels ($259 billion in 2017[4]).

Nowadays the diagnosis of AD is mostly based on clinical and psychometric assessment tests like mini mental state examination (MMSE) and clinical dementia rating scale (CDR) unless the definitive diagnosis of the disease can only be detected by autopsy of the brain. For that, there is a requirement for the processing of novel methods for an early AD diagnosis, and in this respect, an automatic analysis of Spontaneous speech (ASSA) may be considered as one of the most efficient predictors of the stages of Alzheimer's disease[5][38].

The recent literature leans toward the use of (ASSA) with machine learning (ML) classification algorithms to contribute in the interpretation and prediction of AD. It can also help to estimate the seriousness of the disease in the patient. Most of these algorithms have reached promising prediction accuracies.

The process only requires a speech sample of patient, as an input data that will be evaluated and manipulated in order to classify them.

Our proposed approach is based on speech processing for efficient feature extraction, followed by a selection of the significant attributes that lead to increase accuracy and minimize computational costs. The selected attributes are then forwarded to machine learning classifiers, in order to obtain pertinent results in different stages of AD classification.

978-1-5386-9120-5/18/$31.00 ©2018 IEEE

The paper is subdivided as follow. In the second section, we will introduce Alzheimer disease and language impairment, machine learning and detailed related work in automatic AD detection. In the third section, we will formally define our approach. The fourth section will describe the collection of linguistic features, the classification process, including methods for feature selection and classification. Finally, the fifth section will summarize our work and pave the way for future research.

II. THEORETICAL BACKGROUND

A. Alzheimer disease disorder

Alzheimer's disease is a chronic, irreversible disease that attack the brain cells and lead to impairment of mental functioning [7]. The disease typically progresses slowly into three stage; mild cognitive impairment (MCI) stage, moderate stage, and severe stage.

In the mild stage, people with Alzheimer disease may have problems in remembering words, organization, and perform tasks independently. In the moderate stage, Alzheimer disease patients might have difficulties remembering their own personal history; they experience alterations in self-control, and become perplex about time and place [8]. In the severe stage, the disease lead to the loss of bodily functions such as the ability to walk or sit, in this stage affected people often need a full-time care to help complete the activities of daily living (ADLs) [9].

In addition to cognitive ability, the emotional response in people affected with AD becomes impaired; this also appears to go through the three stages of AD. In the early stages, affected individuals may show alterations in their social and behavioural responses as being angry for not remembering or completing common tasks [10,11]. In moderate stage, the Alzheimer's patient reacts aggressively and they frequently cry easily, and show appreciation for hugs and smiles. In advanced stages of Alzheimer disease, individual may often seem unconfident and introverted, symptoms are usually associated to memory problems or trouble in expressing oneself.

Despite great efforts focused on disease moderating therapies of Alzheimer's disease (AD), halting the degenerative process has not been possible. Therefore, early diagnosis of AD became essential. Although memory impairment is the major symptom of Alzheimer's disease (AD), language impairment can also be a sensitive index of disease severity.

Following the cognitive and emotional impairment, language disorders in AD patients could be divided in three stages [13]

In the early stage, even in the absence of phonological, phonetic or syntactic perturbation, early anomic phenomena might be noticeable in the patient's speech. This disorder affects the names, the dates, the less frequently used words of the language and then the familiar words [14,15]. Patients establish offset strategies such as semantic paraphasia (replacement of words by another one), or periphrases (Replacing a word with a longer expression). At this state, the speech is generally vague and uninformative.

At an intermediate stage, the frequency of forgetting word increase, forcing patient to misuse words. Some dissociation appears as the denomination of objects, also anomie is also pronounced when the subject is looking for a word. Discourse is then characterized by numerous periphrases and the use of pronouns without detectable referents, which result to an important disorder of pragmatic skill.

At the Advanced Stage, the depletion on speech is confirmed both qualitatively and quantitatively. In fact, the syntax presents many perturbations at the combinatorial level. The word's order is distant from the canonical model; the misuse of pronouns is highly frequent, the phenomena of echolalia (i.e. systematic repetition of all or part of sentences) and of palilalia (i.e. spontaneous and involuntary repetition of syllables or words) are repetitive; the speech is interspersed with numerous pauses and logatomes. At this level, patients present an impaired comprehension, and difficulties in expressing their daily social and emotional needs.

B. Machine learning

Machine Learning (ML) comes under the umbrella of Artificial Intelligence that has become well established and useful in the last 10 years. ML is a field that consist of developing algorithms that allow a machine to learn from a set of data in order to classify new event and predict new patterns [16]. These ML algorithms are also used in several other fields, such as computer vision, form recognition, information retrieval, bioinformatics, data mining and many more.

In machine learning, to get promising results, it is fundamental to have a good understanding of the problem and the limitations of the used algorithms.

Furthermore, all the algorithms and methods in machine learning are somehow made different. For instance, few methods are designed on the basis of certain hypothesis or on the basis of certain type of data which make them inapplicable for other type of data. That is why it is crucial to apply more than one machine learning method on a given training data.

Machine learning generally have three types of learning algorithms:

1. *Supervised learning* is a machine learning technique that seeks to automatically produce rules from a learning database containing "examples".
2. *Unsupervised learning* search to create a model that discovers some hidden structure in the dataset. It is a self-learning based on unclassified and unlabeled data.
3. *Reinforcement learning* refers to a class of machine learning problems, the purpose of which is to learn from experiments what to do in different situations in order to optimize a quantitative reward over time.

Machine learning methods are being widely used in biomedical research to predict and provide good understanding of the classification of disease [38]. The use of classifier systems in medical fields is growing on a daily basis. In fact, machine learning classifier were proven to be effective in the diagnosis of AD through different type of data such as clinical and neuropathological research data, MRI brain

image, and even pathological speech of patient affected by AD [16].

C. Related work

Research in the area of AD proves that language impairment could be demonstrated at a preclinical stage, up to two years before the diagnosis was made [18]. Other longitudinal studies involved in the determination of risk factors show a link between language skills and the probability of developing an AD [19]. Sellal et al. [20] consider language as an epiphenomenon of memory; they suggest that deterioration of language may be one of the earliest symptoms of Alzheimer's disease which may appear either very early during the disease, or at later stage of its evolution [21,22].

Furthermore, recent studies on language in AD marked differences in language abilities between AD patients and healthy control subjects using computational techniques.

Alzheimer's disease present an important disorder in spoken language, including aphasia (difficulty in speech and understanding) and anomia (difficulty recognizing and naming things) [23].

The study by Thomas et al. [41] represented a statistical analysis of the lexical features in the spontaneous speech of patients with Alzheimer Disease. Data were derived from Atlantic Canada Alzheimer's Disease Investigation of Expectations (ACADIE) study of the drug donepezil [24]. They applied the potential of machine learning and natural language processing techniques in assessment Alzheimer Disease. A new classification algorithm called Ordinal CNG was proposed; it showed positive results in detecting AD, with an accuracy of 68.4%.

Lopez-de-Ipena et al. [5] suggested new model for early AD diagnosis by non-invasive methods based on two human issues: Spontaneous Speech and Emotional Response. According to Horley [39], the impairment of spoken language is usually accompanied with alterations in patient's emotional responses. The technique proposed by Lopez-de-Ipina required an automatic and emotional speech analysis, the spontaneous speech analysis is based on three families of features including, Duration, Time and frequency , the emotional speech analysis is based on the studying of a few prosodic and paralinguistic features sets obtained from a temporal segmentation of the speech signal.

The study has been carried out with a subset from the AZTIAHO database. The goal of these experiments was to examine the potential of selected features for automatic diagnosis of AD. Automatic classification by MLP was performed over the speech features; the obtained results were satisfying with an accuracy value of 90%.

Meilán et al.[31] analysed the temporal and acoustic features from specific sentences spoken by 30 patients with mild AD and compared to 36 healthy controls. The measures of intensity, fundamental frequency, and the temporal structure of speech offer a sensitive method in evaluating spontaneous speech of AD. The Discriminant analysis classifier results confirm that these measures are promising tools for early diagnosis of AD, with an accuracy value of 84.8%.

Orimaye et al [24] used several Machine-learning algorithms to construct diagnostic models based on syntactic and lexical features derived from verbal utterances of AD and related Dementia patients. In addition, they performed statistical tests to find out the most significant linguistic features that could differentiate AD from HC. The best result was obtained with Support Vector Machines (SVM) classifier with an F-score of 74%. Their experiment confirms that syntactic and lexical features could be pertinent features for prediction AD and related Dementias.

Along with Orimaye, Fraser et al. [6] in 2015 studied the use of linguistics analysis as a diagnostic of Alzheimer disease. They proposed that semantic impairment, syntactic impairment, information impairment and acoustic abnormality could be a heavily biomarkers of AD.

They presented a machine learning-based approach to classify patients according to patterns in speech and language production, data were derived from DementiaBank[25] (167 patients diagnosed with "probable"AD and 97 healthy controls).The ML process required that speech simple of DementiaBank's patient were transcribed into raw text. Several of linguistic feature extraction tools have been applied to the transcribed speech to produce a Linguistic Feature vector. The extracted features were used into implementations of SVM machine learning classifiers, results were promising, and 78% of speech simples could be correctly classified.

Sirts et al.[26], they refer to the study of Snowdon et al. [27] that a low idea density ID (known as language characteristic) is found to be associated with an increased risk of developing Alzheimer Disease. This information provided the basis for their aim to predict preclinical Alzheimer Disease from language analysis.

Speech sample were derived from two different databases the first collected from Dementia Bank part of the Talk Bank Corpus, The second dataset, obtained at NeuRA3, contains autobiographical memory interviews (AMI) of both AD patients and healthy control subjects [28]. The computational methods DEPID [26], is used to measure propositional idea density (PID) in the text. As hypothesised, idea density values measured by DEPID, were significantly lower in AD patients than healthy control subjects.

Features generated by DEPID of the transcripted speech were trained through a machine learning models with logistic regression classifier, the obtained results were able to predict which individuals went on to develop Alzheimer's Disease, with an F-score of 72%.

Konig et al.[29] used speech signal processing techniques for extracting vocal markers from voice recordings of patients. Data were extracted from dem@Care project [30] and combined between Healthy elderly control (HC) subjects and patients with MCI or AD. Vocals features were trained into machine learning methods, the classification accuracy results of automatic speech analyses were satisfying 87%, demonstrating its assessment utility.

TABLE 1 SURVEY OF STUDIES IN ALZHEIMER DISEASE THROUGH SPEECH ANALYSIS

Author, Year	Feature type	Feature selection technique	Dataset	Classification	(%)	Performance		
						AD vs HC	AD vs MCI	MCI vs HC
Thomas et al. 2005	Lexical	-	95 AD&HC	ZeroR CNG	ACC ACC	63.6 92.7	-	58.8 62.4
De Ipina et al. 2013	Acoustic & Emotional	-	20 AD 20 HC	SVM MLP DT KNN NB	ACC ACC ACC ACC ACC	93,79 93,02 91,47 87,59 87,59	-	-
Meilán et al. 2014	Temporal & acoustic	-	30 AD 36 HC	LDA	SENS	83.3	-	-
Orimaye et al. 2015	Syntaxic & Lexical	Information Gain	242 AD 242 HC	SVM NB DT NN BN	F-score F-score F-score F-score F-score	74 72 73 74 73	- - - - -	- - - - -
Konig et al. 2015	Acoustic	T-test	26 AD 15 HC 23 MCI	SVM	ACC	87	80	79
Fraser et al. 2015	Linguistic	Pearson's correlation	240 AD 233 HC	LR	ACC	81	-	-
Sirts et al. 2017	Idea density	-	169 AD 98 HC	LR	F-score	66	-	-
			20 AD 20 HC	LR	F-score	74	-	-

SVM: Support Vector Machine; NB: Naïve Bayes; DT: Decision Tree; NN: Neural Network; BN: Bayes Nets;

LR: Logistic Regression; LDA: Linear Discriminant Analysis; MLP: Multi-layer Perceptron; KNN: k-nearest neighbor

ACC: Accuracy; SENS: Sensitivity;

The presented studies exemplified how speech analysis and machine learning potential could achieve interesting results in the prediction and prognosis of AD. These researches prove that acoustic and linguistic features could be significant biomarkers of the onset of AD.

From the description of the beyond studies in this survey, the most common limitations between them were the input data size, feature selection methods and classification process.

The input data quality is very important for an effective classification results. Although small datasets could easily get a higher accuracies value, larger samples will be required in order to make valid generalizations to the approach. Moreover, feature selection methods are relevant to reduce noise and redundant information; in fact, a vector with the most relevant features from initial dataset is very powerful for effective results generation in machine learning. Furthermore, the choice of classifier may affect the results; it is more suitable to select a classifier following the feature sets and the scalability of training.

Our study differs from previous work in several ways, we examine a much expanded sample size than most previous work, giving a more significant sample for machine learning. Compared to Orimaye et al. [24], we discuss a larger number of features in order to discover the different language impairments presenting AD, and we conduct a feature selection method to improve ML classifier results.

III. PROPOSED APPROACH

Machine learning technology is being adopted in biomedical sciences and research for providing prognosis and deep understanding of the classification of disease [16].

The use of classifier systems in medical diagnosis is daily increasing.

This section describes the computational framework adopted for prediction of AD.

The proposed approach is divided into three steps: Feature extraction, Feature selection, and Classification.

1) **Feature extraction**:

This stage describes the feature extraction process adopted for prognosis of AD. The collected data is being combined for the first time and has never been used for computational study.

As shown in our approach (Figure 1), we performed an extraction of the linguistic feature on transcribed files of participant's speech. Features were computed using the CLAN program [43].

a. Linguistic Features

After speech transcription, we conducted a linguistic analysis of transcribed files, on three levels: Syntactic, Semantic, and Pragmatic. Many researches have demonstrated that there is a statistically important difference in the syntactic, semantic and pragmatic structures of languages between patient with AD and healthy controls subjects [24][33].

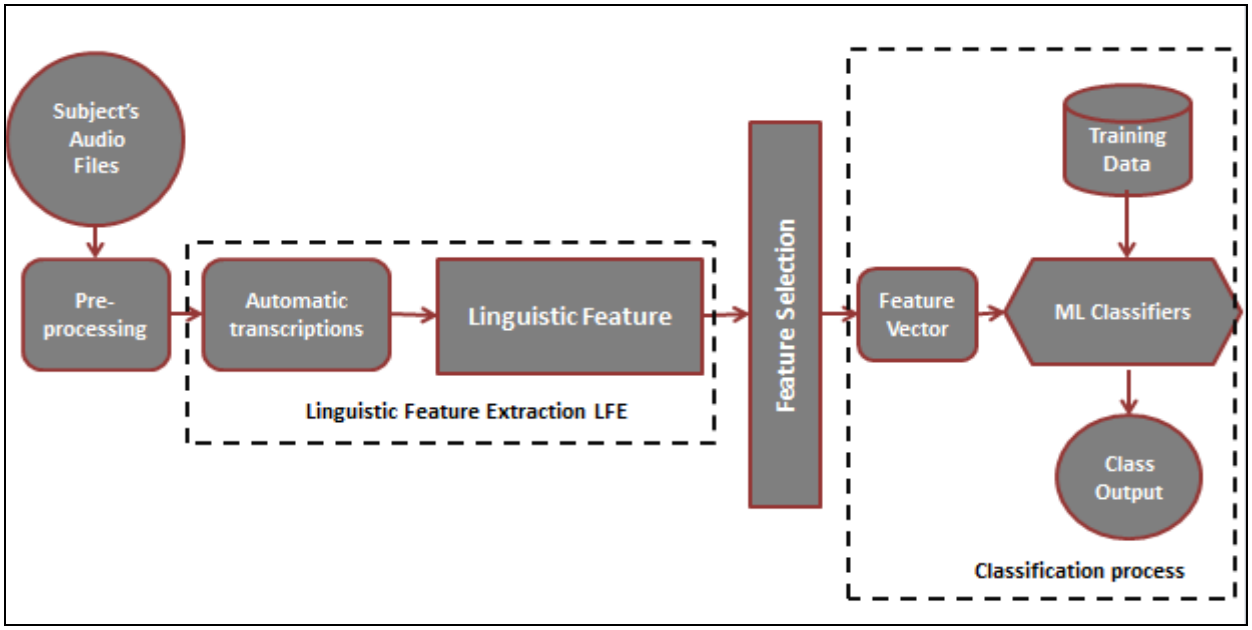

Figure 1 Speech processing architecture for early diagnosis of Alzheimer disease

- Syntactic:

Syntax combines names and actions as a simulation of the sequence of events in the real world it is a study that assess the language evolution (this syntax developed gradually from minimally syntactical utterances). This may include the use of Nouns, Pronouns, Adjectives, and Verbs.

- Semantic:

The meaning of word phrases and the words themselves are the base of the Semantic analyses. This will be measured by the use of Type-Token Ratio, and Idea density.

- Pragmatic:

Determining how language can be used and how Semantic and syntax features could be determined through the different uses of the language, are the main objective of Pragmatic analysis. This includes pronoun use, paraphrasing, and repetition, and syllables-per minute.

2) Feature Selection:

Despite the importance of feature selection process in machine learning results, few of the described studies have applied this technique. Feature selection helps reducing the dataset size before proceeding to the classification phase since the majority of existing classifiers become inefficient if there are too many input variables. It can lead to increase accuracy or to reduce computational costs [34].

Using all described features in previous section is a time consuming and an intensive task. Furthermore, a few features may be more significant in distinguishing AD patient's from Healthy Control subjects. During this process, the most significant features will be used and the inappropriate features will be eliminated.

There are three feature selection (FS) categories;

- *Filters*: Features were selected irrespective to the classifier used (As the information gain filter, the analysis of variance ANOVA)
- *Wrappers*: features were chosen according to the classification algorithm. The method uses cross validation to estimate the accuracy of the classifying algorithm for a given set of attributes
- *Hybrid*: Features were first chosen using a filter method then applied using a wrapper method.

3) Classification

The final stage of the proposed approach is based on classification using a set of classifiers. The goal behind using many classifiers is to improve the reliability and accuracy of our results , by handling deficiency and effectiveness of each classifier in order to give the best possible decision; taking into account all the ensemble of data [37].

The classifiers that we have used are SVM, NN, and Decision Tree.

- SupportVectorMachines (SVM) are one of the supervised multivariate classifiers; they are a new type of learning methods for binary classification, driven by the results of statistical learning theory. SVM methods aspire to structural risk minimization which is a compromise between the complexities of the decision functions space and the quality of adjustment to data learning.

  SVM has shown good performance in many areas of applications such as text grading, pattern recognition, medical diagnosis, etc. They are now recognized as one of the standard tools for learning. The idea of SVM algorithms is to find a hyperplane that best separates the two data groups. The closest observations to the separator hyperplane, are called the "support vectors".

- Neural network (NN): A multi-layer neural network consisting of a series of neurons layers or units to which weights are attached. Each neuron performs a relatively simple task: receiving the external information or outputs of the previous layer's neuron and use them to calculate its own output (called activation) that propagates to the connected neurons of the next layer. NN is very popular for pattern recognition and interpolation. NN is used for Alzheimer's disease detection from spontaneous speech [24].
- Decision Tree: The decision tree classifier is described by a tree structure, which has been widely used to represent classification models due to its ability to break down a complex decision-making process into simpler sub-decisions, thus, providing a solution that is often easier to interpret. Decision trees have some advantages over other learning algorithms, such as low computational cost when building the model and the ability to handle redundant attributes. On the other hand, the constructed model generally has good generalization capacity [39]. This algorithm has already been used for Alzheimer's disease identification with different modalities [40].

The feature selection and classification process will be performed using the WEKA program [44].

IV. Linuistic features for early dianosis of Alzheimer Disease

The proposed model was effectively applied with a concrete case with the DementiaBank database. In this paper we will only focus on linguistic features for AD prediction.

A. Dataset

Our data are extracted from the DementiaBank, one of the largest existing datasets of spontaneous speech with and without dementia. All patients were required to have at least an initial Mini-Mental State Exam (MMSE) score of 10, 44 years of age and 7 years of education.

Each speech sample consists of a verbal description of the Boston Cookie Theft picture. In this task, patients describe a complex kitchen scene. Each speech sample was recorded along with manual transcriptions, following the TalkBank CHAT (Codes for the Human Analysis of Transcripts) protocol [42].

From the original database, a subset of 242 samples of healthy control subjects and 242 samples of AD patients were chosen.

The audio files selected from the conversations were then transformed into .WAV files (16 bits and 16 Khz). The pre-processing step is applied in order to enhance the efficiency of feature extraction and classification process and thus, to ameliorate the overall system [32].

It consists in removing background noise, the beginning and ending breaks and deleting non-analysable effect as segments where patients overlapped, coughing or laughing. The resulting speech simples were transcribed into text files. Finally, the pre-processed speech records are forwarded to the feature extraction step.

B. Feature extraction

As we have mentioned above, we will use syntactic, semantic, and pragmatic features for the onset of AD.

TABLE 2 EXTRACTED LINGUISTIC FEATURES

Feature's type	Features	Descriptions
Syntactic	Verbs	Total verbs
	Nouns	Total nouns
	adjective	Total adjectives
	adverbs	Total adverbs
	Total Utterances	Includes all utterances used
	MLU Utterancess	Number of utterances used to compute MLU
	MLU Words	MLU in words.
	MLU Morphemes	MLU in morphemes
Semantic	FREQ types	Total word types does not include repetitions and revisions
	FREQ tokens	Total word tokens does not include repetitions and revisions
	FREQ TTR	type/token ratio
	Idea density	Measure of propositional idea density
	Word Errors	Percentage of words that are coded as errors
	Utterance Errors	Number of utterances coded as errors
	Verbs/Utterances	Verbs per utterance
Pragmatic	Words/Min	Number of words per minute
	Auxiliaries	Number of auxiliaries
	3S	Third person singular
	1S3S	Identical forms for first and third person
	PAST	Past
	PASTP	Past participle
	PRESP	Present participle
	Preposition	Number of prepositions
	Conjunctions	Number of conjunctions
	Pronoun	Number of pronouns
	Determiners	Number of determiners
	Retracing	Number of retracing
	Repetition	Number of repetitions

*MLU: mean length of utterance

C. Feature selection

Three feature selection methodologies were studied for the selection of optimal feature set. Information Gain (IG), kNN model-based feature selection and SVM recursive feature elimination (SVM-RFE) were deployed individually in the complete dataset [35][36].

- Information Gain: It is a filter method that looks at each feature separately , computes its information gain and measures how important and relevant it is to the class labeled (Practically, it measures the expected reduction in entropy).
- K-nearest neighbors (KNN): As wrapper methods are time consuming, we will use KNN classifier since it is a simple algorithm that manages to identify the features in each class without a complex parameter to optimize.
- SVM recursive feature elimination (SVM-RFE): stands for support vector machines, which is a hybrid method that computes the ranking weights for all features and sorts them according to weight vectors as the

classification basis. SVM-RFE is an iteration process of the backward removal of features. the feature set selection are as follows, Use the current dataset to train the classifier, Compute the ranking weights for all features, Delete the feature with the smallest weight.

TABLE 3 FEATURE SELECTION RESULT'S

Feature selection methods	Selected attribute
IG (filter)	Words/Min, Nouns, Pronoun, MLU Morphemes, PRESP, Repetition, Words Errors, Preposition
KNN (wrapper)	MLU Morphemes, Word Errors, Noun Auxiliaries, 3S PAST, PASTP, Preposition, Adverb, Conjunctions, Repetition
SVM-RFE (embedded)	Repetition, Preposition, PAST, Determiners, Word Errors, Adverb, Auxiliaries, 3S

The result of this IG analysis is a listing of features ranked by their importance; we have selected the top 8 features. The number of variables was reduced to 11 for the KNN method.

Similar to the IG method, we have limited the results of the SVM-RFE to the top 8 ranked features.

By observing the table we can see that attributes "Word Errors", "Preposition", and "Repetition" also had high rankings in the three feature selection methods.

The classification with each of these optimal sets was done by; the Neural Network, Support Vector Machine and Decision Tree classifier. The results are given in Table 4.

D. Classification Results

TABLE 4 RESULT OF ML CLASSIFIERS

Feature selection	NN	SVM	DT
NONE	**0.58**	**0.60**	**0.56**
IG	**0.64**	**0.68**	**0.61**
KNN	**0.69**	**0.79**	**0.71**
SVM	**0.65**	**0.64**	**0.60**

From the results of the classifiers presented in Table 4, it is clear that the performance for all classifiers was augmented by the feature selection methods. For the NN the global precision increased from 58% to 69%. In similar way, DT increased from 53% to 67%. Amelioration in the results was showed in the SVM too, by the overall precision increased from 60% to 79%. Nevertheless, SVM performed better than NN and DT in both cases (with and without the feature selection methods).

Interestingly the three ML classifiers demonstrated a better precision, using the top 11 features selected by the KNN method.

According to the output from our study, we could prove that using linguistic features extracted from verbal utterances of AD patients could be effective biomarkers of Alzheimer and the related dementia diseases.

However, compared to Orimaye [24], our study identifies more detailed and representative linguistic features. Also contrary to other study, we did not include the MMSE score and age, which considered as the most informative features that could improve the accuracy of ML classifier; our approach is properly based in speech samples only.

Eventhough, the use of the CHAT transcription format has been an available tool for analyzing speech data; it is still not universally used for speech transcription. Therefore, we consider that the use of CHAT symbols for the extraction of the lexical features could be a limitation for this work.

V. CONCLUSION

As the population continues to age, interest in accurate diagnosis of dementia is increasing.

Nowadays, clinical diagnosis may require much time from caregivers, patients, and medical personnel. These demands are expected to increase; and therefore, an early detection for AD would be a valuable tool.

This study applies a Machine learning models for the early diagnosis of Alzheimer's disease, based on linguistic features indicative of language impairment extracted from the verbal utterances of individuals affected and non-affected with AD.

We believe that the proposed model will help improving the prediction performance in detecting AD and cover the limitations discussed in the previous researches.

Other features will be discussed in our future research, such as the acoustical features (prosodic, temporal, emotion) that can be used to AD diagnosis and to emotion responses analysis. Other classifiers may be also used, to better evaluate our features and distinguish between AD and HC classes.

REFERENCES

[1] A. Alzheimer, " eine eigenartige Erkrankung der Hirnrinde. Allgemeine Zeitschrift fűr Psychiatrie", 64, 146-148, 1907.

[2] J. Ankri, "Prévalence incidence et facteurs de risque de la maladie d'Alzheimer. Gérontologie et société", 128/129, 129-141, 2009.

[3] https://www.alzheimers.net/resources/alzheimers-statistics/

[4] Alzheimer's-Association. "2016 Alzheimer's Disease Facts and Figures", In: Alzheimer's and Dementia 11.3 (2016).

[5] K. López-de-Ipiña, Alonso, Travieso J-B, Solé-Casals C-M, Egiraun J., Faundez-Zanuy H. , Ezeiza M. , Barroso A. , Ecay-Torres N. , Martinez-Lage M. , Lizardui P. , "On the Selection of Non-Invasive Methods Based on Speech Analysis Oriented to Automatic Alzheimer Disease Diagnosis", Sensors 2013, 13, 6730-6745.

[6] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features differentiate Alzheimer's from controls in narrative speech," Journal of Alzheimer's Disease, vol. 49, no. 2, pp. 407–422, 2015.

[7] J. L. Cummings, "Alzheimer's disease", New England Journal of Medicine, 351(1):56–67, 2004.

[8] C. Ballard, S. Gauthier, A. Corbett, C. Brayne, D. Aarsland, E. Jones, "Alzheimer's disease", The Lancet, 377(9770):1019– 1031, 2011.

[9] R. N. Kalaria, G. E. Maestre, R. Arizaga, R. P. Friedland, D. Galasko, K.. Hall, J. A. Luchsinger, A. Ogunniyi, E. K Perry, F. Potocnik, "Alzheimer's disease and vascular dementia in developing countries: prevalence, management, and risk factors", The Lancet Neurology,7(9):812–826, 2008.

[10] K. Horley, A. Reid, D. Burnham, "Emotional Prosody Perception and Production in Dementia of the Alzheimer's Type", Journal of Speech Language and Hearing Research. 2010; 53(5):1132-1146. doi:10.1044/1092-4388(2010/09-0030).

[11] MS. Goodkind, A. Gyurak, M. McCarthy, BL. Miller, RW. Levenson, "Emotion regulation deficits in frontotemporal lobar degeneration and Alzheimer's disease" Psychol Aging. 2010; 25(1):30-37. doi: 10.1037/a0018519.

[12] MY. Savundranayagam, ML. Hummert, RJ.Montgomery, “Investigating the effects of communication problems on caregiver burden”, J Gerontol B Psychol Sci Soc Sci. 2005;60(1):S48–S55.

[13] M. Barkat-Defradas, S. Martin, L. Rico-Duarte, D. Brouillet, “Les troubles de la parole dans la maladie d’Alzheimer”, 27èmes journées d’études sur la Parole. Avignon, France, 2008.

[14] D., Cardebat, B.,Aithamon, & M., Puel, “Les troubles du langage dans les démences de type Alzheimer”, In F. Eustache & A. Agniel (Eds.), Neuropsychologie cliniques des démences : Evaluation et prises en charge (pp. 213-223). Marseille: Solal, 1995.

[15] A. Rosser , J. R. Hodges, “Initial letters and semantic category fluency in Alzheimer's disease, Huntington's disease, and progressive supranuclear palsy”, Journal of Neurology, Neurosurgery and Psychiatry, 57, 1389-1394,1994

[16] R. Chen, E. H. Herskovits, “Machine learning techniques for building a diagnostic model for very mild dementia”, Neuroimage, 52(1):234–244, 2010.

[17] L. Khedher, J. Ramírez, J. M. Górriz, A. Brahim, and F. Segovia, “Early diagnosis of Alzheimer disease based on partial least squares , principal component analysis and support vector machine using segmented MRI images,” Neurocomputing, vol. 151, pp. 139–150, 2015.

[18] L. Mickes, J. T. Wixted, C. Fennema-Notestine, Galasko, D. Bondi, M. W., L. J., Thal, “Progressive impairment on neuropsychological tasks in a longitudinal study of preclinical Alzheimer's disease”, Neuropsychology, 21(6), 696-705, 2007.

[19] J S., Kemper, L., Greiner, J., Marquis, K., Prenevost, & T., Mitzner, “Language decline across life span: findings from the nun study”, Psychology and Aging, Vol. 16(2), 227-239, 2001.

[20] Sellal, F., & Kruczek, E. (2007). Maladie d’Alzheimer. Paris : Doin.

[21] J. L., Cummings, D. F., Benson, M., Hill, & S. Read, “Aphasia in dementia of the Alzheimer type”, Neurology*, 35*, 394-397, 1985.

[22] F O. A., Selnes, K., Carson, B., Rovner, & M. D., Gordon, “Language dysfunction in early- and late-onset possible Alzheimer's disease”, Neurology*, 38*, 1053-1056, 1988.

[23] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R Jack, C. H. Kawas, W. E Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, “The diagnosis of dementia due to Alzheimer’s disease: Recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer’s disease”, Alzheimer’s and Dementia: The Journal of the Alzheimer’s Association, 7(3):263–269, 2011.

[24] S. Orimaye, J. S.-M. Wang, and K. J. Golden, “Learning predictive linguistic features for Alzheimer’s disease and related dementias using verbal utterances,” in Proc. of the ACL 2014 Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Baltimore, USA, Jun. 2014, pp. 78–87.

[25] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland, “Aphasiabank: Methods for studying discourse,” Aphasiology, vol. 25, no. 11, pp. 1286–1307, 2011.

[26] K. Sirts, O. Piguet, M. Johnson, “Idea density for predicting Alzheimer's disease from transcribed speech”, Proceedings of CoNLL 2017

[27] D. A. Snowdon, S. J. Kemper, J. A. Mortimer, L. H. Greiner, D. R. Wekstein, W. R. Markesbery. “Linguistic ability in early life and cognitive function and Alzheimer’s disease in late life”, Findings from the Nun Study. JAMA 275(7):528–532, 1996.

[28] Neuroscience Research Australia, https://www.neura.edu.au/

[29] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P.H. Robert, R. David, “Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease”, Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 1(1), pp.112-124, 2015.

[30] Dementia Ambient Care: Multi-Sensing Monitoring for Intelligent Remote Management and Decision Support. Available at: http:// www.demcare.eu/. Accessed March 20, 2015.

[31] J. Meilan, F. Martnez-Snchez, J. Carro, D. Lopez, L. Millian- Morell, and J. Arana, “Speech in Alzheimer’s disease: can temporal and acoustic parameters discriminate dementia?” Dementia and Geriatric Cognitive Disorders, vol. 37, no. 5-6, pp. 327–334, 2014.

[32] D. He, Y. Hou, Y. Li, “Key technologies of pre-processing and post-processing methods for embedded automatic speech recognition systems”, Proceedings of 2010 IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications.

[33] J. Onofre de Lira, K. Zazo Ortiz, A. Carvalho Campanha, P. Henrique Ferreira Bertolucci, Th. Soares Cianciarullo Minett, 2011, “Microlinguistic aspects of the oral narrative in patients with alzheimer’s disease, International Psychogeriatrics, 23(03):404–412.

[34] H. Wang, Sh-H. Lo, T. Zheng, I. Hu, “Interaction-based feature selection and classification for high-dimensional biological data”, Bioinformatics. 2012 Nov 1; 28(21): 2834–2842

[35] S. Visalakshi, V. Radha, “A literature review of feature selection techniques and applications: Review of feature selection in data mining”, IEEE International Conference on Computational Intelligence and Computing Research, 2014.

[36] K. Tejeswinee, J. Shomona Gracia, R. Athilakshmi, “Feature Selection Techniques for Prediction of Neuro-Degenerative Disorders: A Case-Study with Alzheimer’s And Parkinson’s Disease”, 7th International Conference on Advances in Computing & Communications, India, ICACC-2017.

[37] M. P. Ponti-Jr. , J. P. Papa, “Improving accuracy and speed of optimum-path forest classifier using combination of disjoint training subsets,” in 10th Int. Work. on Multiple Classifier Systems (MCS 2011) LNCS 6713. Naples, Italy: Springer, 2011, pp. 237–248.

[38] R. Ben Ammar, Y. Ben Ayed, “Machine Learning Based-Approach for Early Diagnosis of Alzheimer Disease”, The 12 th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE’17), 2017.

[39] J. Han, M. Kamber, “Data mining : concepts and techniques”, San Francisco, CA : Morgan Kaufmann, 2001.

[40] M. Dyrba, M. Ewers, M. Wegrzyn et al., “Combining DTI and MRI for the automated detection of Alzheimer’s disease using a large Europeanmulticenter dataset,” in Multimodal Brain Image Analysis, vol. 7509 of Lecture Notes in Computer Science, pp. 18–28, 2012.

[41] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, E. Asp, Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In Mechatronics and Automation, 2005 IEEE International Conference ,Vol. 3, pp. 1569-1574, 2005.

[42] B. MacWhinney (2000) The CHILDES Project: Tools for analyzing talk, 3rd edition. Lawrence Erlbaum Associates, Mahwah, New Jersey.

[43] B. MacWhinney, "Tools for Analyzing Talk Part 2: The CLAN Program." (2017).

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