IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 17, 2020, accepted September 21, 2020, date of publication October 9, 2020, date of current version November 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3029907

Predicting Prodromal Dementia Using Linguistic Patterns and Deficits

AHMED H. ALKENANI[1,2], YUEFENG LI[1], YUE XU[1], (Member, IEEE), AND QING ZHANG[2], (Member, IEEE)

[1]School of Computer Science, Queensland University of Technology, Brisbane, QLD 4001, Australia
[2]The Australian E-Health Research Centre, CSIRO, Brisbane, QLD 4029, Australia

Corresponding author: Yuefeng Li (y2.li@qut.edu.au)

ABSTRACT Language deficiency is evident in the onset of several neurodegenerative disorders yet has barely been investigated when first occurs on the continuum of cognitive impairment for the purpose of early diagnoses. Alzheimer's disease (AD) is a neurodegenerative pathology that develops years prior to clinical manifestations and typically preceded by prodromal stages such as Mild Cognitive Impairment (MCI). Currently, the manual diagnostic procedures of both types are time consuming, following certain clinical criteria and neuropsychological examinations. Our study aims to establish state-of-the-art performance in the automatic identification of different dementia etiologies, including AD, MCI, and Possible AD (PoAD), and to determine whether patients with initial cognitive declines exhibit language deficits through the analysis of language samples deduced with the cookie theft picture description task. Data was derived from the cookie theft picture corpus of DementiaBank, from which all language samples of the identified etiologies were used, with a random subsampling technique that handles the skewness of the classes. Several original lexical and syntactic (i.e., lexicosyntactic) features were introduced and used alongside previously established lexicosyntactics to train machine learning (ML) classifiers against these etiologies. Further, a statistical analysis was conducted to uncover the deficiency across these etiologies. Our models resulted in benchmarks for differentiating all the identified classes with accuracies ranging between 95 to 98% and corresponding F1 values falling between 94 and 98%. The statistical analysis of our lexicosyntactic biomarkers shows that linguistic deviations are associated with prodromal as well as advanced neurodegenerative pathologies, being greatly impacted as cognitive decline increases and suggesting that language biomarkers may aid the early diagnosis of these pathologies.

INDEX TERMS Alzheimer's disease, prodromal dementia, cognitive decline, clinical diagnosis, neurolinguistics, machine learning, prediction, feature selection.

I. INTRODUCTION AND MOTIVATION

The rising elderly population is a prominent demographic attribute of developed countries [1], [2]. Alzheimer's disease (AD) along with other related neurodegenerative pathologies are considered one of the most common persistent issues facing an aging population due to their nature of being incurable [3]. Without medical breakthrough, early diagnosis is the only hope for people with, or likely to develop, dementia. Therefore, a timely diagnosis of dementia is fundamental for decelerating its progression as well as allowing maximized benefits of pharmaceutical interventions that can mitigate the side effects in certain types of dementia [4]–[6]. On the other hand, it may stabilize or even curtail the decline in some prodromal dementia cases [7]–[10]. However, clinical examinations of dementia typically involve multiple diagnostic procedures, which may be highly stressful and costly thus a major cause of late diagnosis.

The diagnosis of prodromal dementia is currently challenging [11], [12]. Normally, prodromal dementia is diagnosed via traditional pen-and-paper screening tests such as the Montreal Cognitive Assessment (MoCA), which involve a series of questions to assess different cognitive skills (e.g., short memory, attention, repetition, and orientation) [13]. Some of these traditional screening tests are simple to use;

The associate editor coordinating the review of this manuscript and approving it for publication was Hao Ji.

however, they have certain limitations such as relying on the neurologist's level of expertise, being affected by the age and level of education as well as being insensitive to early dementia thus are usually followed by further comprehensive tests [14], [15]. Consequently, there has been a demand for further effective clinical techniques to diagnose prodromal dementia by concerned associations [16].

A well-established literature shows various early-disrupted characteristics of language and speech in patients with prodromal stages of dementia as well as patients with Alzheimer's disease (AD) [17]. Early opening analytical studies of language and speech have underlined deviations in naming and verbal fluency tasks [18]–[21]. More recently, studies using automatic or semiautomatic methods for language and speech analysis have asserted that linguistic analysis can characterize early AD and MCI [22]–[24]. Specifically, as stated by Ball *et al.* [25] and asserted by Rentoumim *et al.* [26], lexical and syntactic (i.e., lexicosyntactic) processing in people with language disorders has revealed promising outcomes, highlighting the necessity of additional investigations for more effective lexicosyntactic biomarkers.

On the basis thereof, this study investigates computational diagnostic models for foreseeing prodromal dementia through linguistic patterns and deficits. It also explores and illustrates the gradual deterioration of lexicosyntactic in patients with different stages of dementia. A potential value of this investigation would be the ability to automatically diagnose preclinical stages of dementia, allowing for early intervention prior to developing irreversible dementia. Additionally, it supports previous related studies on exploring language deficiency caused by cognitive decline.

Using Natural Language Processing (NLP) fused with Machine Learning (ML) algorithms, we propose multiple diagnostic models to predict prodromal dementia. These computational models employ different sets of essential language components extracted from transcripts belonging to Healthy Control (HC), AD, MCI, and PoAD participants from DementiaBank dataset. Firstly, lexicosyntactic representations are investigated to understand the linguistic deviations associated with different stages of dementia, where we introduce original lexicosyntactic features and exploratorily analyze them alongside previously established lexicosyntactics. We then propose a statistical-based feature selection method that handles both normally and non-normally distributed features for selecting the most predictive subset among our lexicosyntactic features. We also introduce *n*-gram language models to seize the sequence of words in the language samples. Our diagnostic models establish new independent benchmarks for pairwise classifications across all the identified classes.

II. RELATED WORK

As reported in various studies, it is possible to identify language changes years prior to developing dementia, which highlights the importance of linguistic analysis for dementia detection [27]–[36]. While most of these studies focused on connected speech of people with developed AD, a relatively small number of computational linguistics studies have attempted to explore early AD and other prodromal dementias. For instance, Orimaye *et al.* [37] carried out a study to distinguish 19 patients diagnosed with MCI from 19 healthy adults from DementiaBank dataset. They introduced a few skip-gram models with different spaces, where word tokens are intermittently skipped when forming *n*-grams. For instance, the one-skip-three-grams of the sentence ''the jar is falling'' are ''the-is-falling'' and ''the-jar-falling''. With different top skip-gram spaces, they conducted several experiments to train a few classifiers, where the top 200 compound skip grams resulted in 98%, 97%, 97% and 99% of precision, recall, F1 score, and AUC, consecutively, when fused with Support Vector Machine (SVM), Logistic Regression (LR), and Naïve Bayes (NB) classifiers.

Later on, the same authors conducted a study using 242 transcripts of patients with probable and possible AD and 242 transcripts of HC, where a SVM classifier was trained with syntactic features and obtained an F1 score of 74% [38]. This work was extended in another study to classify the same classes of AD and HC using *n*-grams combined with lexical and syntactic features. They experimented with different combinations of these features and reported their best model with an Area Under the ROC Curve (AUC) of 93%, using the top 1000 combined features [39]. Similarly, Fraser *et al.* [23] carried out a binary classification study using 240 AD and 233 HC transcripts, where a LR classifier was trained with the top 35 features selected out of 370 acoustic and syntactic features using Pearson's correlation. They reported an accuracy of 81%.

Al-Hameed *et al.* [40] showed that acoustic features could optimally differentiate AD patients from healthy adults in a study involved a total of 264 individuals (i.e., 167 AD and 97 HC). They extracted 263 acoustic features from these transcripts, where the top 20 features led to their highest accuracy of 94.7% and recall of 97% when fused with Bayesian networks. In a subsequent work [41], they focused on differentiating AD and MCI patients from healthy adults, extending the acoustic features to over 811 features in total. An SVM classifier was employed in their experiments that resulted in accuracies failing between 95 and 97%. Similarly, acoustic features were recently investigated by Haider *et al.* [42], where the authors experimented with different sets of acoustic features extracted from transcripts belonging to 82 AD patients and 82 healthy adults. They reported the highest accuracy of 78.7%, which was achieved by applying a ''hard fusion'' technique to these sets of features and fusing them with a Decision Tree classifier.

In a similar study, Ammar and Ayed [43] proposed a model for classifying AD and HC groups, utilizing 242 transcripts per each group. They extracted syntactic, semantic, and pragmatic features then used three feature selection methods (i.e., Information Gain (IG), K-nearest neighbors (KNN), and SVM Recursive Feature Elimination (SVM-RFE)).

The set of features selected using KNN resulted in an accuracy of 79% when used with SVM, as the best model in their study. Contrarily, Yancheva and Rudzicz [44] used topic modeling to classify these two groups (i.e., AD and HC). They formed 10 clusters of verbs and nouns such as C0: window, curtain, kitchen, plate; and D0: baking, cookies, apple. Afterwards, a set of 12 semantic features was extracted and fused with a Random Forest (RF) classifier, resulting in 74% F1 score. However, when combining this set of features with syntactic features, the F1 score increased to 80%. Budhkar and Rudzicz [45] also used topic modeling for the same purpose, where they augmented topic-models with ''pre-trained'' wrod2vec for classifying 167 AD patients (with 240 samples) from 97 healthy adults (with 233 samples). They reported the highest F1 score of 77.5% when fusing 25 topic-induced word2vec with a linear-kernel SVM.

Hernández-Domínguez *et al.* [46] extracted Information Content Units (ICU) from 25 transcripts belonging to healthy adults to form a reference of the main units in the picture. These ICUs were then used to assess how informative the transcripts of AD patients are, where more ICUs coverage means more informativeness. More linguistic and phonetic features were also extracted alongside the ICUs, where they preselected the statistically significant features (i.e., with p-value $<.001$) among these features and used them to the train two classifiers, namely SVM and RF. They reported their best averaged results of a 79% accuracy and AUC as well as 81% precision, recall, and F1 score.

Neural network, on the other side, are seen to perform well in detecting prodromal dementia. An instance is the work of Orimaye *et al.* [24], where they investigated the diagnosis of MCI through newly proposed Deep-Deep Neural Networks Language Model (D2NNLM). They carried out experiments to distinguish between MCI and HC participants, using 43 transcripts per each group, by extracting higher order n-grams and skip-grams and then fusing them with the D2NNLM. Their model scored an accuracy of 87.5%.

Karlekar *et al.* [47] also presented similar models using Convolutional Neural Networks (CNNs), Long Short-Term Memory-Recurrent Neural Networks (LSTM-RNNs), and an integration of them (CNN-LSTM) as an attempt to optimize the automatic classification of patients with AD from HC. They utilized all transcripts from the Cookie Theft Picture description corpus from DementiaBank dataset, belonging to people with different stages of dementia, along with the accompanying Part of Speech (POS) tags that were originally annotated by the dataset custodians. Instead of using these transcripts in their original format, they divided them into single utterances to make utterances-based classification as an attempt to enlarge the samples. As a result, they achieved an accuracy of 91.1%. Later on, their work was reproduced by Di Palo and Parde [48], wherein they attempted to overcome the imbalanced samples resulting from the Karlekar's proposed technique of dividing transcripts into utterances by adding an attention mechanism and class weights. Nonetheless, their approach dropped the performance to 88% accuracy as the best performing model on this metric. They also reported and F1 score of 93.05% and AUC of 95.03%.

Another recent attempt to address the use of neural networks for detecting dementia is the recent work of Pan *et al.* [49]. They have proposed a hierarchical bidirectional neural network induced with an attention mechanism for extracting different levels of features. A total of 255 recordings of AD patients and 222 of healthy adults were employed in their study, where they firstly used the original manually transcribed transcripts then applied an automatic speech recognition to transcribe these recordings. Their model could reach the best performance of 84.02%, 84.97%, and 84.43% precision, recall, and F1 score, respectively. Similarly, Chen *et al.* [50] have recently introduced a hybrid attention-based neural model for classifying AD and HC groups. They used a total number of 498 transcripts in the development of their model, with 256 transscripts belonging to the AD group and the remaining to the HC group. The hybrid model was formed by integrating a CNN with a bidirectional Gated Recurrent Units (BiGRU) fused with an attention layer, to capture both the semantics and patterns from these transcripts. Interestingly, their model could optimally classify these groups with an accuracy of 97.42%. Table 1 summarizes the existing methods.

III. METHODS AND MATERIALS

In this section, we firstly describe the dataset and groups of participants involved in this study followed by the details of the engineering and selection of diagnostic features, where we introduce our proposed statistical-based feature selection technique. Afterwards, we discuss the employed ML algorithms and present the subsampling technique employed against the skewed classes before concluding the section with the evaluation method.

A. DATASET

DementiaBank[1] dataset is currently considered the largest publicly available dataset for assessing language of AD and related dementia. This dataset was collected through a longitudinal study carried out at the University of Pittsburgh [51]. A subset of this datasets contains recordings, with corresponding transcripts, from English-speaking individuals performing the Cookie Theft Picture description task as a part of the Boston Diagnostic Aphasia Examination ''BDAE'' [27]. These interviews were carried out through multiple visits for a duration of five years from 1983 to 1988 with participants aged from 45 to 90-year old. Those participants were encouraged to thoroughly describe what is happening in the cookie theft picture and were audio-recorded while performing this task. Afterwards, recordings were manually transcribed using the CHAT transcription protocol [52].

The Cookie Theft Picture description task corpus includes 243 language samples from 98 healthy controls, 236 language samples from 189 patients diagnosed with AD, 43 samples

[1] https://dementia.talkbank.org/

TABLE 1. Survey of related studies on DementiaBank dataset.

Author/Year	Features	Feature selection technique	Number of samples* or participants ^	Evaluation	Classification Algorithms	Performance Evaluation (%)			
						Metrics	AD vs HC	HC vs MCI	AD Vs MCI
Orimaye et al. 2014 [38]	Syntactic lexical	Information Gain	242 AD* 242 HC*	10-fold Cross-Validation	SVM	Pr. Re. F1.	75 73 74	-	-
Orimaye et al. 2015 [37]	Skip-grams	N/A	19 MCI^ 19 HC^	10-fold Cross-Validation	SVM Naïve Bayes LR	Pr. Re. F1. AUC	-	98 97 97 99	-
Fraser et al. 2016 [23]	Linguistic phenomena	Pearson's correlation test	240 AD* 233 HC*	10-fold Cross-Validation	LR	Acc	81.92	-	-
Yancheva et al. 2016 [44]	Idea density and efficiency, distance-based, lexiosyntactic and acoustic	N/A	255 AD* 241 HC*	10-fold Cross-Validation	RF	Acc Pr. Re. F1.	80 80 80 80	-	-
Orimaye et al. 2016 [24]	N-grams and skip-grams	N/A	19 MCI^ 19 HC^	Average of 943 iterations	D2NNLM	Acc	-	87.5	
Orimaye et al. 2017 [39]	Syntactic, lexical and n-grams	Information Gain	99 AD^ 99 HC^	LPOCV	SVM	AUC	93	-	-
Al-hameed etal. 2016 [40]	Acoustic	Weka feature selection "SVMAttributeEval"	167 AD^ 97 H^	LOOCV	Bayes Net	Acc.	94.7	-	-
Al-hameed etal. 2017 [41]	Acoustic	Weka feature selection "SVMAttributeEval"	195 AD^ 98 HC^ 19 MCI^	10-fold Cross-Validation	SVM	Acc.	94.3	97.3	95.6
Ammar and Ayed 2018 [43]	Linguistic	K-nearest neighbors	242 AD* 242 HC*	10-fold Cross-Validation	SVM	Acc.	79	-	-
Hernandez-Domınguez et al. 2018 [46]	Linguistic features and information coverage units	Manual preselection of statistically significant features (p-value <.001)	257 AD* 217 HC*	10-fold Cross-Validation	SVM	Acc. F1. AUC Pr. Re.	79 81 79 81 81	-	-
Orimaye et al. 2018 [87]	N-grams	N/A	99 AD^ 99 HC^ 19 MCI^	Average of 500 iterations	D2NNLM 5n-grams	AUC	83	80	-
Karlekar et al. 2018 [47]	POS	N/A	208 PwD^ 104 HC^	N/A	2D-CNN LSTM CNN-LSTM CNN- LSTM	Acc.	82.8 83.7 84.9 91.1	-	-
Di Palo and Natalie 2019 [48]	POS	N/A	208 PwD^ 104 HC^	Three iterations	CNN- LSTM	Acc. Pr. Re. F1. AUC	84.95 85.08 99.65 91.07 92.07	-	-
					CNN-LSTM-ATT	Acc. Pr. Re. F1. AUC	84.66 85.25 98.95 91.58 95.03	-	-
					CNN-LSTM-ATT-W	Acc. Pr. Re. F1. AUC	88.20 93.05 92.98 93.05 94.98	-	-
Fraser et al. 2019 [88]	Cluster features, Summary features, and baseline features	Clustering	166 AD^ 97 HC^ 19 MCI^ 19 HC^	LOOCV	SVM	Acc. F1.	85	63	-
Zhu et al. 2019 [89]	Linguistic features	N/A	234 AD* 240 HC*	N/A	Consensus Networks	F1.	79.9	-	-
Budhkar and Rudzicz 2019 [45]	LDA -induced topics augmented with word2vec	PCA	240 AD* 233 HC*	5-fold Cross-Validation	SVM	Acc.	77.5	-	-
Pan et al. 2019 [49]	Linguistic features	N/A	255 AD* 222 HC*	10-fold Cross-Validation	BHANN	Pr. Re. F1.	84.02 84.97 84.43	-	-
Chen et al. 2019 [50]	Linguistic features	N/A	256 AD* 242 HC*	10-fold Cross-Validation	Att-CNN+Att-BiGRU	Acc.	97.42	-	-
Guo et al. 2019 [90]	Language model derived Perplexities and linguistic features	AUC-based feature selection	256 AD* 242 HC*	LOPO Cross-Validation	LR	Acc.	85.4	-	-
Fritsch et al. 2019 [91]	Language model derived Perplexities	N/A	255 AD* 244 HC*	LOSO Cross-Validation	NNLMs-LSTM	Acc.	85.6	-	-
Haider et al. 2020 [42]	Acoustic features	Pearson's correlation test	82 AD^ 82 HC^	LOSO Cross-Validation	Decision Tree	Acc.	78.7	-	-

HC, Healthy Control; AD, Alzheimer's Disease; MCI, Mild Cognitive Impairment; Acc., Accuracy; Pr., Precision; Re., Recall; AUC, Area-Under-Curve; F1, F1 Score; POS, Part of Speech tags; LPOCV, Leave Pair Out Cross-Validation; LOOCV, Leave One Out Cross-Validation; PwD, People with Dementia; SVM, Support Vector Machine; LR, Logistic Regression; RF, Random Forest; CNN, Convolutional Neural Networks; LSTM, Long Short Term Memory; LDA, Latent Dirichlet allocate; PCA, Principal Component Analysis on ATT, ATTention mechanism; W, class Weight; BHANN, bidirectional hierarchical recurrent neural network combined with an attention mechanism; BiGRU, Bidirectional Gated Recurrent Units; LOPO, Leave-One-Person-Out; NNLMs-LSTM, neural network language models- Long Short Term Memory; LOSO, Leave-One-Subject-Out

TABLE 2. Number of participants and their language samples.

Class	*Subjects*	*Samples*	*Age range*
HC	99	243	46 *to* 81
AD	169	236	49 *to* 90
MCI	19	43	49 *to* 90
PoAD	21	21	60 *to* 85

from 19 patients with MCI, 21 samples from patient with possible AD (PoAD), and 5 samples of patients with Vascular dementia, with all of them received extensive neurological and neuropsychological assessments. As illustrated in Table 2, our study investigated the first four identified etiologies, namely AD, MCI, PoAD, and HC.

We extracted the transcribed word-level sentences from the CHAT files and discarded the annotations as a part of the preprocessing step prior to feature extraction. In addition, we ignored the participants' demographic data included in this dataset since our approach is solely based on linguistic patterns.

B. FEATURE ENGINEERING AND MACHINE LEARNING ALGORITHMS

The designed approach in this study provides a balanced performance and learning speed, resulting in an optimal performance for classifying all the investigated classes. Our approach is described in detail in the next sub-sections.

1) FEATURE ENGINEERING

Our study investigates and combines two types of connected speech measures, namely linguistic features and n-gram vocabulary spaces, as an attempt to optimize the automatic identification of different dementia etiologies. Linguistic features have widely been employed to reveal dementia stages in previous studies [23], [30], [38], [53]–[56]; nevertheless, our study focused on one linguistic perspective that is lexicosyntactic features as being recommended for further investigations by other researchers for their established association with early cognitive decline [25], [26]. N-gram vocabulary spaces, on the other hand, have recently attracted several NLP studies on dementia detection [24], [38], [39], [57]. We explored different spaces of n-gram vocabularies and used them in combination with lexicosyntactic features to train ML classifiers for binary classifications of the involved groups. The description and analysis of these features are provided in the following subsections.

LexicoSyntactic: Our lexicosyntactic processing investigated several features, among which a few features were explored previously and seen to deteriorate in people with dementia (PwD), including the word count from Orimaye *et al.* [39], type-token ratio (TTR) evaluated by Kave and Dassa [58] as well as content density evaluated by Roark *et al.* [30]. Besides, we proposed and investigated new lexicosyntactic features, namely open and closed class ratios, noun to verb index, verb to noun index, active proposition density, and passive proposition density. Another aspect we examined is the ratio of functional words (i.e., stopwords). In contrary to previous studies, we explored these lexicosyntactic features with an additional analysis among different etiologies. Subsequently, our study may reveal the language changes associated with different stages of dementia. Our lexicosyntactic features are described as follows:

- **Word count:** We calculated the total number of terms, including the repeated terms.
- **Type-token ratio:** Type-token ratio (TTR) is the total number of unique POS tagged tokens to the total spoken words. It is a widely used measure for language assessment, to provide insight on the language production and reveal language disabilities.
- **Content density:** Content density is a measurement of language complexity, explored by Roark *et al.* [30], which is the quantity of expressed propositions to the total spoken words.
- **Stopwords ratio:** Stopwords (i.e., functional words) refer to a set of words, typically considered uninformative in NLP tasks, that occur in most documents. However, these words might be useful to our task, given the fact that language production is generally deteriorated as dementia progresses. Therefore, we calculated the count of these words to the total spoken words and named it stopwords ratio.
- **Open and closed classes ratios:** At a higher level of word classes, POS tags form two classes; open and closed classes. Open class refers to an infinite number of new words to be created and added such as nouns, verbs, adjectives, adverbs, and interjection; whereas closed class includes a relatively small fixed sets such as conjunction, determiners, modals, particles, prepositions, ad-positions, auxiliary verbs, and pronoun [59], [60]. We measured the open class ratio by calculating open class words to the total spoken words. Similarly, closed class ratio was measured by calculating the closed class words to the total spoken words.
- **Noun to verb index:** We calculated the noun to verb ratios and name this feature as noun to verb index.
- **Verb to noun index:** Likewise, we calculated the verb to noun ratios and name this feature as verb to noun index.
- **Active proposition density:** Motivated by findings from previous studies that PwD lean towards using less nouns but more verbs and propositions compared to healthy adults [28], [30], [61], [62], we measured verbs, adjective, and adverbs ratios to noun ratio and name this feature as active proposition density.
- **Passive proposition density:** Likewise, we measured the noun ratio to verbs, adjective, and adverbs ratios and term this feature as passive proposition density.

For syntactic annotations of transcripts, we followed the annotation conventions of Penn Treebank represented in the NLTK suite[2], of which POS tagger was trained on the

[2]http://www.nltk.org/

TABLE 3. Steps of our feature selection and evaluation method.

Stage One

1. Prepare the statistically significant lexicosyntactic features using Algorithm 1

Stage Two

1. Generate *n*-grams with *n* equals to (*2*), (*3*) and (*2, 3*) for generating *bigrams, trigrams*, and a combination of both *bigrams and trigrams*, respectfully and separately
2. Convert the *n*-grams, resulted from the previous step, to a matrix of TF-IDF (term frequency-inverse document frequency) features
3. Using Chi-square, rank the TF-IDF features resulted from the previous step and select the top *t* features, where *t* represents the nominated number of features for each pairwise classes (i.e., 1000 for HC vs AD, 200 for MCI vs HC and AD, and 100 for PoAD vs HC and AD)

Stage Three

1. Combine the statistically significant features, resulted from stage one, with a number of the top *t* features, resulted from stage two, that makes a total equalizing the nominated number of features for each pairwise classes (e.g., 989 *t* features *in addition to* 11 statistically significant features for a total of 1000 to classify HC against AD)

Stage Four

Using GNB, MLP, and SVM classifiers, test the performance of:

1. The entire lexicosyntactic features
2. The statistically significant lexicosyntactic features (i.e., from stage one)
3. The top *t* features (i.e., from stage two)
4. The compound of *t* and statistically significant lexicosyntactic features (i.e., from stage three)

- Repeat all stages for each pairwise classes, with replacing the number of *t* according to the nominated number of features for each pairwise classes (i.e., *1000* for HC vs AD, *200* for MCI vs HC and AD, and *100* for PoAD vs HC and AD)

Penn Treebank corpus with the maximum entropy [63]. NLTK POS tagger revealed 30 unique tags in the Cookie Theft Picture description corpus (highlighted in Supplementary Table 11), with a total of 59480 tagged words and 1808 vocabularies in the corpus, resulted from 543 transcripts in total.

N-gram vocabulary spaces: The use of *n*-gram vocabulary spaces is predominant in many NLP tasks [64]. An *n*-gram is a sequence of *n* tokens that may be either a character or word, with the latter being our concern in this paper. Consequently, and unless specified otherwise, we refer to *n*-gram as a sequence of words where *n* represents the word count in the sequence. For example, it is named ''unigram'' when *n* equals to 1, which consists of only one word. Likewise, when *n* equals to 2 and 3, they are called ''bigram'' and ''trigram'', respectively. Our *n*-gram vocabulary spaces in this study include bigrams and trigrams as $1 < n \leq N$, where N equals to 3, which were seen to perform optimally in other NLP tasks [65]. We left high order spaces of *n*-grams for future investigations.

Due to the enormous nature of *n*-gram features, we would only see how they contributed to the proposed models for classifying the identified groups. It is also worth mentioning that stopwords were removed prior to *n*-grams extraction to capture low-level vocabulary spaces.

2) FEATURE SELECTION

Feature selection is concerned with the extraction of the most relevant features from a set of features. It is considered a critical step to enhancing the efficiency of machine learning models by eliminating redundancy and irrelevant data, especially in the field of text mining where numerical representation of texts leads to high dimensionality that in turn affects the efficacy of learning algorithms. Although, there are filter, wrapper, as well as embedded methods for selecting relevant features [66], filter methods are preferred in text classification tasks due to their independency of the learning algorithms and the associated low computational complexity [67]. Accordingly, we selected filter methods and left the investigation of other methods to future work. Besides, given the nature of the explored features, our feature selection setup bears a close resemblance to the work of Orimaye *et al.* [39]. Our feature selection method is demonstrated in Table 3 and described as follows:

Lexicosyntactic: Intuitively, it is anticipated that lexicosyntactic features deteriorate as cognitive decline increases. As such, we conducted an independent statistical analysis to individually examine these features among pairwise classes of the involved groups (i.e., AD, MCI, PoAD, and HC) and to estimate the deficiency of our lexicosyntactic features across people with different stages of dementia as to reveal the most discriminatory features. It is also our anticipation that linguistic features may differ across participants given the presence and severity of dementia, which may lead to abnormal distributions across the explored transcripts. Consequently, we chose two different two-sample statistical tests: The Student's t-test as a parametric test and Kolmogorov-Smirnov test (KS) as a non-parametric test. These two ''goodness-of-fit'' tests assess whether feature values of transcripts belonging to classes C_1 and C_2, respectively, are drawn from different distributions, with the main difference between them being the assumptions they make. While

TABLE 4. Significant features across pairwise classes through non-parametric "KS" and parametric "t" tests (p<0.05).

Features	*HC vs AD*		*HC vs MCI*		*AD vs MCI*		*HC vs PoAD*		*AD vs PoAD*	
	KS	t	KS	t	KS	t	KS	t	KS	t
Word count	0.106	0.231	0.765	0.323	0.580	0.525	0.531	0.160	0.154	0.110
Stopword R	**0.018**	**0.013**	0.917	0.797	0.765	0.504	**0.029**	**0.030**	**0.018**	**0.008**
Noun R	**0.000**	**0.000**	0.169	0.648	**0.031**	**0.001**	0.304	0.133	0.797	0.687
Verb R	0.201	0.067	**0.002**	**0.000**	0.169	0.086	0.304	0.416	0.973	0.904
Pronoun R	**0.000**	**0.000**	0.580	0.811	**0.031**	**0.001**	**0.029**	**0.047**	0.973	0.462
Adverb R	**0.000**	**0.000**	0.580	0.302	**0.002**	**0.003**	**0.049**	**0.028**	0.071	0.336
Adjective R	0.414	0.321	0.989	0.775	0.765	0.830	0.154	0.111	0.797	0.251
Determiner R	**0.000**	**0.004**	0.408	0.321	**0.008**	**0.005**	0.531	0.202	0.797	0.986
Conjunction R	0.294	0.294	0.169	0.047	0.100	0.060	0.797	0.547	0.531	0.846
Interjection R	0.982	0.856	0.765	0.437	0.989	0.864	0.304	0.116	0.154	0.101
Noun to Verb I	**0.000**	**0.034**	**0.008**	**0.037**	**0.008**	**0.003**	0.797	0.588	0.304	0.653
Verb to Noun I	**0.000**	**0.000**	**0.008**	**0.025**	**0.008**	**0.005**	0.797	0.539	0.304	0.653
Content D	0.072	0.059	0.057	0.080	0.765	0.857	0.973	0.763	0.973	0.749
Open class R	**0.000**	**0.004**	0.057	0.057	0.765	0.720	0.973	0.762	0.973	0.864
Closed class R	**0.000**	**0.006**	0.057	0.058	0.580	0.717	0.973	0.771	0.973	0.844
Active proposition D	**0.000**	**0.000**	0.169	1.131	**0.031**	**0.001**	0.071	0.110	0.531	0.646
Passive proposition D	**0.000**	**0.000**	0.169	0.170	**0.031**	**0.001**	0.071	0.142	0.531	0.667

HC, Healthy Control; AD, Alzheimer's Disease; MCI, Mild Cognitive Impairment; PoAD, Possible Alzheimer's Disease; KS, Kolmogorov-Smirnov test; t, t-Test; R, Ratio; I, Index; D, Density

the t-test presumes an identical variation of both distributions without further assumptions of whether they are discrete or continuous, the KS test is distribution-free but supposes that both distributions are continuous (e.g., ratio level).

The two-sample t-test is widely used to assess the difference between the means of two distributions [68]. It starts with a null hypothesis (H_0) that both distributions have identical means and its p-value within a specific confidence interval can be used to reveal the statistical significance. It is calculated by Eq. (1).

$$t = \frac{f_{C_1} - f_{C_2}}{S_p \sqrt{\frac{1}{n_{C_1}} + \frac{1}{n_{C_2}}}} \tag{1}$$

In Eq. (1), f_{C_1} and f_{C_2} are the two means of the same feature f in the classes C_1 and C_2, with n_{C_1} and n_{C_2} representing the size of samples in C_1 and C_2, respectively. S_p is an estimator of the pooled standard deviation of the two features.

On the other side, the two-sample KS test assesses the variation between two cumulative distribution functions (CDFs) of the distributions (i.e., of a feature f in our case) over the set (X) belonging to the classes C_1 and C_2, respectively [69]. Its initial null hypothesis (H_0) is that both of the C_1 and C_2 share the same cumulative distribution. The two-sided KS statistic applies the maximum absolute variation between the two cumulative distribution functions CDFs of the distributions which can be calculated by Eq. (2).

$$KS = Max \left| F_{C_1}(X) - F_{C_2}(X) \right| \tag{2}$$

where F_{C_1} and F_{C_2} are the distribution functions of a feature f over the set (X) belonging to the classes C_1 and C_2, respectively. The KS value represents the maximum difference between the two distributions and the associated p-value is used to find the statistical significance.

We found that both tests returned similar significance at a degree of freedom equals to N-2, where N represents the total instances of pairwise classes. Given the binary classification task in this study, we select features that independently violet H_0 of both tests at alpha $\alpha < 0.05$ for our experiments as the statistically significant features (Table 4). Besides, we chose the parametric t-test for a further statistical analysis of lexicosyntactic features. All conducted tests are two-tailed using Python version 3.7.3, specifically researchpy and scipy libraries.

Algorithm 1 The proposed statistical-based feature selection method

Input: $D \leftarrow$ training dataset
$M \leftarrow l$exicosyntactic features, $M = \{\boldsymbol{f_1}, \ldots \boldsymbol{f_n}\}$ in D
$C \leftarrow$ the set of Classes, $C = \{\boldsymbol{c_1}, \boldsymbol{c_2}\}$
$L \leftarrow$ Significance level (0.05)
Output: $SLF \leftarrow$ significant lexicosyntactic features *in* M

initialize a vector S and let $S_i = 0$ for each f_i in M, $\text{t\_test}_{p-value} = 0$, $\text{KS}_{p-value} = 0$
for each feature $f_i \in M$ **do**
  get the weight of f_i in D for c_1 and c_2
  calculate $\text{t\_test}_{p-value}$
  calculate $\text{KS}_{p-value}$
  if ($\text{t\_test}_{p-value} < L$) and ($\text{KS}_{p-value} < L$) **then**
    $S_i = S_i + 1$
  endif
endfor
$SLF = \{f \in S | \text{S}_\text{i} > 0\}$
Return SLF

N-gram vocabulary spaces: In terms of selecting the most informative vocabulary spaces, we employed the chi-square (X^2) test owing to being one of the most widely used and effective methods for feature selection [70]–[72]. The X^2 formula is applicable to data that are not on a numerical scale and its formula is associated with information-theoretic

TABLE 5. A contingency table of feature *t and class c.*

	Class c	*Different c*	*Total c*
Term t	*a*	*d*	*a + d*
Different term t	c	b	c + b
Total t	*a + c*	*d + b*	*Grand Total N*

feature selection methods, which anticipates that the top terms (i.e., *n*-grams in our case) t_p for the class c_i are those differently distributed in positive and negative documents of class c_i. It measures the dependency between a term t and a class c where the higher X^2 score, the higher association between these two factors t and c. For feature selection in text classification tasks, X^2 is used to rank terms based on how informative they are. It calculates the score for a term t and a specific class c as illustrated in the following contingency table (Table 5), where each cell represents an observed value and these observed values are used to calculate the expected value "E" as $E = (total\ t\ *\ total\ c)/grand\ total\ N$. Eq. (3) demonstrates the X^2 statistic for this contingency table.

$$X^2\left(t_p, c_i\right) = \frac{N\,(ab - cd)^2}{(a+c)\,(d+b)\,(a+d)\,(c+b)} \tag{3}$$

where N equals to the entire documents in the dataset, a equals to the number of documents in class c_i that covers the term t_p; d equals to the number of documents related to other classes that cover the term t_p; c equals to the number of documents in class c_i that do not cover the term t_p; and b equals to the number of documents related to other classes which do not cover the term t_p. For each feature in each class, a score is assigned as explained in Eq. (3). All given scores are then formulated in a final score $X^2\left(t_p, c_i\right)$.

3) CLASSIFICATION ALGORITHMS

The core task of this study is binary classification to distinguish patients with different stages of prodromal dementia. A few machine learning algorithms were employed in developing our models for classifying the identified groups, namely Gaussian Naive Bayes (GNB), SVM, and Multilayer Perceptron neural networks (MLP). We selected these machine learning algorithms as seen to perform well in related studies [37], [55]. The implementation of these algorithms was performed in Python version 3.7.3, using Scikit-Learn[3] library. We experimented with different settings of hyper-parameters per each of the employed algorithms in the optimization process of our models and found that default settings had led to the best performance. This, however, was not applicable to SVM, where a linear kernel with enabled probability led to the best performance among other hyper-parameters, which was seen to perform well in a similar task [73].

Since our hyper-parameters tuning process included two search strategies (i.e., grid search and manual search), it is worth highlighting that grid search performed poorly compared to manual search despite assembling every possible combination of parameters [74].

[3]https://scikit-learn.org/stable/

4) TECHNIQUES FOR HANDLING IMBALANCED CLASSES

As indicated earlier, the cookie theft picture corpus includes heavily imbalanced classes, especially for MCI and PoAD groups against HC and AD groups, which typically biases the classification model towards the more common class. We tackled this issue with a subsampling technique, whereby we performed a random selection of a subset of the majority with a matching size to that of the minority. This procedure was repeated 10 times with a different subset from the majority each time to avoid bias. Finally, the mean of these iterations is reported. The followed technique is illustrated in Fig. 1. For PoAD against HC and AD, we had to increase the iterations to 20 times for better handling of the high skewness of the transcripts (i.e., 21 against 236 and 243 samples for PoAD, AD, and HC, respectively).

5) EVALUATION

An extensive evaluation of our proposed models was performed using the common metrics typically used for evaluating text mining and NLP systems, namely Accuracy, Precision, Recall, and F1 score. Besides, we use AUC as seen to be effective in summarizing the overall performance of diagnostic models [75].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{6}$$

$$F1 = 2 * \frac{(precision * recall)}{(precision + recall)} \tag{7}$$

$$AUC = \frac{(s_p - n_p\,(n_n + 1)\,/2}{n_p n_n} \tag{8}$$

Above equations represent the metrics used for the performance evaluation in this study, where the number of samples that have been recognized correctly as positive referred to as the True Positives (TP). Similarly, the number of samples that have been incorrectly recognized as positive represents the False Positives (FP) whereas the False Negative (FN) is the number of samples that have been incorrectly recognized as negative. Eq. (8) is used to calculate the AUC, where: s_p is the sum of all positive samples, n_p and n_n correspond to the number of positive and negative samples respectively.

Each subsampling iteration was evaluated using cross validation, where a unique random subset of the transcripts is selected in each iteration, with fixed seed to ensure a direct comparison of results as well as reproducibility. With *n*-folds, the StratifiedKFold technique uses $n-1$ for training the model and then performs the testing on the held set, then scores the average of all folds as illustrated in Fig. 2. We selected the

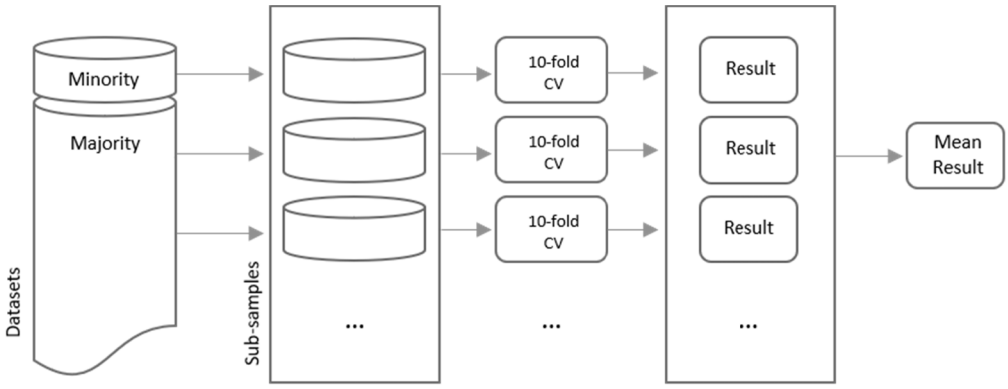

FIGURE 1. High level depiction of the experimental technique.

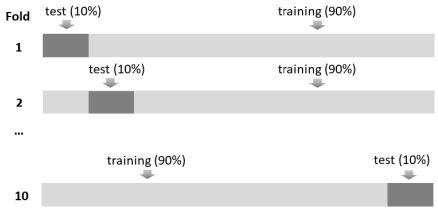

FIGURE 2. The dataset splitting process during cross validation of classifiers evaluation. The dark grey represents the testing set and the lighter grey represents the training set in each fold.

n-fold value as n= 10 for maximum reduction of the difference between true and estimated values of performance, which in turn reduces the bias [76]. Each iteration was assessed using different metrics in-parallel, whereby the final reported results are the average over all iterations. For Precision, Recall and F-Score, we reported the weighted results.

IV. RESULTS AND DISCUSSION

A. INTRODUCTORY ANALYSIS OF LEXICOSYNTACTIC FEATURES

As a part of our lexicosyntactic investigation, we conducted an intuitive exploratory data analysis to understand the correlations between these features with different stages of dementia. As illustrated in Fig. 3, we observed strong correlations between content density and two of our newly proposed features (i.e., open and closed ratios). For example, open class ratio and content density were shown to decrease in tandem as the disease progresses which suggests that while the language samples of PwD may lack an informative content, the usage of nouns, verbs, and other words that forms the open-class domain could be deteriorated. This is closely mirrored in noun ratio with passive proposition density and noun to verb index, where they seemed to decrease jointly as the cognitive decline increases. This strong correlation feasibly shows a concurrent validity between some of the previously validated features (e.g., content density) and that we introduced in our study.

Moreover, it was interesting to see a negative linear association between the number of spoken words and the ratio of functional words. Despite being moderate (i.e., -0.50 to -0.70 [77]), this correlation suggests that PwD lean towards using more functional than content words as the disease advances. This could possibly indicate increased word finding difficulties that leads to less informative or ''empty'' speech which is, the later, associated with semantic impairment [6], [78]. A close noteworthy observation was the moderate positive correlation between pronoun ratio and active proposition density, which implies that PwD may overuse pronouns alongside other propositions as the disease evolves [6].

On the other side, the low correlations of the remaining features suggest that they may complement each other towards enhancing the diagnostic ML models. As such, we extended our data analysis in the next section to reveal the statistically significant lexicosyntactic features for each pairwise groups.

B. STATISTICAL ANALYSIS OF LEXICOSYNTACTIC FEATURES

We performed an independent analysis of our lexicosyntactic features to draw a conclusion of their changes crosswise different etiologies of dementia. The challenge of an uneven distribution of some features was encountered in our evaluation, which is explicable as each participant would give specific characteristics corresponding to the existence and severity of the disease. Consequently, we initiated the analysis with the non-parametric test, which does not assume specific distribution of features across classes, followed by the parametric test that contrarily presumes a normally distributed features, where both tests resulted in similar statistical significance as highlighted in Table 4. It is noteworthy that we performed the statistical analysis of these features based on equal subsamples selected from majority classes with matching individuals' demographic attributes to those of minority classes such as age and education, motivated by the work of Orimaye *et al.* [39].

Our evaluation of the AD and HC groups revealed a total of 11 statistically significant features with alpha $\alpha < 0.05$, among which we observed that the AD group had smaller ratios of nouns, determiners and open-class compared to that of the HC group. This was also extended to the noun-to-verb

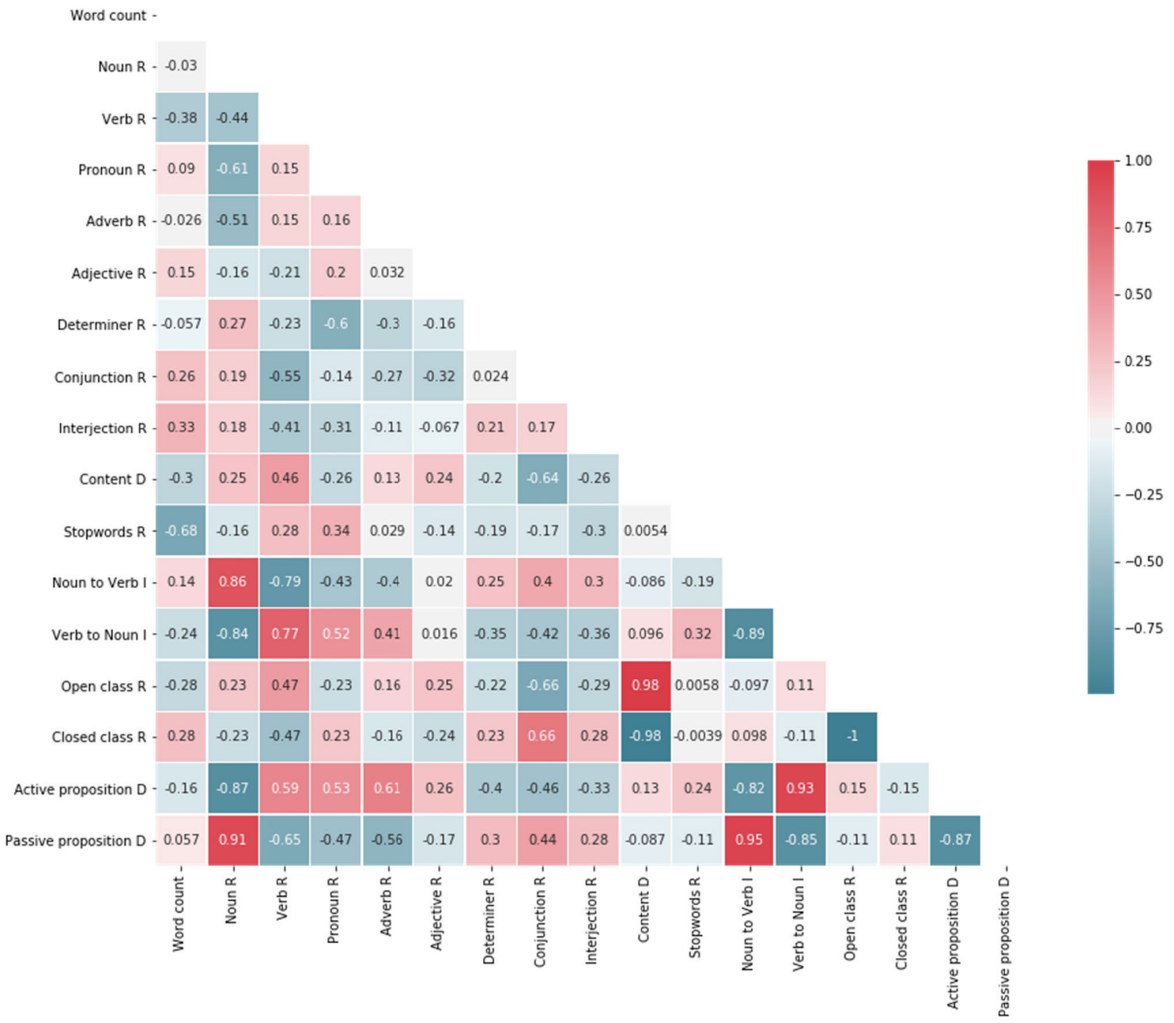

FIGURE 3. Correlations between lexicosyntactic features across all classes (HC, AD, MCI and PoAD) Pearson's correlation coefficient is shown along with direction (blue = negative and red = positive). R, Ratio; D, Density; I, Index. Noun R, Verb R, Pronoun R, Adverb R, Adjective R, Determiner R, Conjunction R, and Interjection R = Type Token Ratios.

index as well as passive-proposition densities. In contrast, more usage of pronouns, adverbs, and a higher verb-to-noun index, and active proposition density was noted in the AD group. Besides, we noticed more intense of functional words (i.e., stopwords) in the AD group. We also observed higher standard deviation (SD) values of all statistically significant features in the AD group than that of the HC group, which indicate that AD samples are more dispersed than that of the HC group.

In a different manner, only a few features were statistically significant for distinguishing healthy adults from people with prodromal dementia. For instance, the HC and MCI groups had ratios of verbs and conjunctions as well as noun-to-verb and verb-to-noun indexes as the only significant features, whereas functional words, ratios of pronouns and adverbs were significant for the HC group alongside the PoAD group. This was mirrored in differentiating AD from PoAD patients, with the intense of functional words being the only significant feature. On the other hand, the AD and MCI groups had eight significant features wherein the AD group had higher ratios of pronouns and adverbs, verb index as well as idea and active proposition densities, and lesser ratios of nouns and determiners along with a lower passive proposition density compared to the MCI group. This makes sense since PwD tend to use more propositions than healthy adults, which has been noted in previous studies [28], [61], [62]. For instance, the AD group showed higher ratios of adverbs and pronouns compared to the HC and MCI groups. It also was our expectation to observe higher values of content density in the HC group compared to that of cognitively impaired groups, which was shown to be the case, with higher means in the former. We, thus, suggest that content density decreases in all stages of dementia, which is coherent with the findings of Roark *et al.* [30].

C. DIAGNOSTIC MODELS

Given the multiple classes and combinations of different features in this study, we carried out extensive experiments with

TABLE 6. Top models for classifying HC - AD groups.

	Model	Features	Acc.	Pr.	Re.	F1	AUC
Baselines	**Orimaye et al.** SVM	Top combined 1000-lin.-ngrams	-	-	-	-	93
	Karlekar et al. CNN-LSTM	POS	91.1	-	-	-	-
	Di Palo and Natalie CNN-LSTM	POS	84.95	85.08	99.65	91.07	92.07
	CNN-LSTM-ATT		84.66	85.25	98.95	91.58	95.03
	CNN-LSTM-ATT-W		88.20	93.05	92.98	93.05	94.98
	Chen et al. ATT-CNN+ATT-BiGRU	Linguist patterns and features	97.42	-	-	-	-
Ours	GNB	Top combined 1000-Sig.-Bi.	**97.25**	**97.55**	97.04	**97.21**	98.45
	GNB	Top combined 1000 -Sig.-Bi.-Tri.	96.19	95.33	**97.46**	96.28	96.19
	MLP	Top combined 1000-Sig.-Bi.	95.31	94.42	96.15	95.61	**98.74**

HC, Healthy Control; AD, Alzheimer's Disease; Acc., Accuracy; Pr., Precision; Re., Recall; AUC, Area-Under-Curve; F1, F1 Score; lin, linguistic features; POS, Part of Speech tags; SVM, Support Vector Machine; Bayes Net, Bayesian Networks; CNN, Convolutional Neural Networks; LSTM, Long Short Term Memory; ATT, ATTention mechanism; W, class Weight; BiGRU, Bidirectional Gated Recurrent Units; GNB, Gaussian Naïve Bayes; MLP, Multilayer Perceptron; Sig., Statistically Significant features; Bi., Bigrams; Tri., Trigrams

TABLE 7. Top models for classifying HC - MCI groups.

	Model	Features	Acc.	Pr.	Re.	F1	AUC
Baseline	**Al-hameed et al.** SVM	Acoustic	97. 3	-	-	-	-
	Orimaye et al. SVM NB Logistic	Top compound 200 skip-grams	-	98	97	97	99
Ours	GNB	Top combined 200-Sig.-Bi.	95.25	**100**	90.5	94.60	**99.37**
	GNB	Top combined 200-Sig.-Bi.-Tri.	**97.5**	98	**97.5**	**97.46**	97.5

HC, Healthy Control; MCI, Mild Cognitive Impairment; Acc., Accuracy; Pr., Precision; Re., Recall; AUC, Area-Under-Curve; F1, F1 Score; SVM, Support Vector Machine; GNB, Gaussian Naïve Bayes; MLP, Multilayer Perceptron; Sig., Statistically Significant features; Bi., Bigrams; Tri., Trigrams.

each set of our explored features then with combinations of them across pairwise classes, employing the aforementioned algorithms (i.e., GNB, MLP, and SVM).

While we followed Orimaye *et al.* [37], [39] in selecting the top 1000 features to discriminate AD patients from healthy participants and the top 200 features for distinguishing MCI patients from other groups, we randomly selected the top 100 features for PoAD patients against other groups. Our experiments yielded improved performance with combined features across all pairwise classes. Specifically, with the top combined 1000 statistically significant lexicosyntactic features and bigrams, the GNB model could classify the AD and HC groups with 97.25%, 97.55%, and 97.21% accuracy, precision and F1 score, respectively. However, fusing the same compound of features with MLP resulted in a slightly better AUC of 98.74% than that of the GNB model (i.e., 98.45%). We also noticed that adding trigram to this compound and fusing them with GNB led to the best recall of 97.21%.

In a similar situation, our approach could result in optimum performance for distinguishing MCI patients from healthy adults, achieved with GNB; with 97.5%, 97.5%, and 97.46% scores of accuracy, recall, and F1 score, respectively, when fused with a compound of the top 200 statistically significant lexicosyntactic features, bigrams and trigrams. Nevertheless, removing trigrams from this combination led to the best precision of 100% and AUC of 99.37%. Contrarily, MLP took the lead in classifying the MCI from AD groups with a compound of the top 200 statistically significant lexicosyntactic features and bigrams, where we scored 98.75%, 100%, 97.5%, 98.57% and 99.6% of accuracy, precision, recall, F1 score and AUC, consecutively.

This applies to the task of classifying PoAD patients from the AD and HC groups, where GNB and MLP outperformed SVM. A combination of the top 100 statistically significant lexicosyntactic features, bigrams, and trigrams were sufficient for training GNB to segregate the PoAD and HC groups with the best accuracy of 95%, precision of 96.66%, recall of 95% and F1 score of 94.66%. Yet, the best AUC was achieved with MLP fused with a compound of the top 100 of statistically significant lexicosyntactic features and

TABLE 8. Top models for classifying AD - MCI groups.

	Model	Features	Acc.	Pr.	Re.	F1	AUC
Baseline	**Al-hameed et al.** SVM	Acoustic	95.6	-	-	-	-
Ours	MLP	Top combined 200-Sig.-Bi.	**98.75**	**100**	**97.5**	**98.57**	**99.6**

AD, Alzheimer's Disease; MCI, Mild Cognitive Impairment; Acc., Accuracy; Pr., Precision; Re., Recall; AUC, Area-Under-Curve; F1, F1 Score; SVM, Support Vector Machine; MLP, Multilayer Perceptron; Sig., Statistically Significant features; Bi., Bigrams.

TABLE 9. Top models for classifying HC - PoAD groups.

Model	Features	Acc.	Pr.	Re.	F1	AUC
GNB	Top combined 100-Sig.-Bi. Top combined 100-Sig.-Bi.-Tri.	**95**	**96.66**	**95**	**94.66**	95
MLP	Top combined 100-Sig.-Bi.	90	93.33	90	87.90	**97.5**

AD, Alzheimer's Disease; PoAD, Possible Alzheimer's Disease ; Acc., Accuracy; Pr., Precision; Re., Recall; F1, F1 Score; AUC, Area Under Curve; GNB, Gaussian Naïve Bayes; MLP, Multilayer Perceptron; Sig., Statistically Significant features; Bi., Bigrams; Tri., Trigrams.

TABLE 10. Top models for classifying AD - PoAD groups.

Model	Features	Acc.	Pr.	Re.	F1	AUC
MLP	Top combined 100-Sig.-Bi. Top combined 100-Sig.-Bi.-Tri.	**95**	**93.33**	**100**	**96**	**97.5**

HC, Healthy Control; PoAD, Possible Alzheimer's Disease; Acc., Accuracy; Pr., Precision; Re., Recall; AUC, Area-Under-Curve; F1, F1 Score; MLP, Multilayer Perceptron; Sig., Statistically Significant features; Bi., Bigrams; Tri., Trigrams.

bigrams (i.e., 97.5%). On the other hand, differentiating PoAD patients from the AD group was best performed by MLP trained with a combination of the top 100 statistically significant lexicosyntactic features, bigrams, and trigrams, which resulted in 95%, 93.33%, 100%, 96%, and 97.5%, accuracy, precision, recall, F1 score, and AUC, respectively.

In brief, we found GNB to be superior for distinguishing AD, MCI, and PoAD patients from healthy adults whereas MLP took the lead in the classification of the MCI and PoAD groups against AD patients. We also noticed that bigram showed slightly higher importance than trigram in the training of MLP; yet, both were mostly important to GNB. Interestingly, our approach could reveal the most informative *n*-grams features, leading to optimal performance and showing high discriminative power of bigrams and trigrams when selected with chi-square algorithm [79], emphasizing that *n*-grams are the definite strength of our models. Besides, it showed that statistically significant lexicosyntactic features have positively impacted the performance of our models. Tables 6-10 present a summary of the best models. In addition, detailed experiments are provided in Supplementary Tables 6-10.

On the other hand, aligning with Orimaye *et al.* [39], our experiments show the powerfulness of *n*-grams features, capturing the most informative linguistic deficits between HC and people with different dementia stages. We found both bigram and trigram vocabulary spaces to be effective in the diagnostic task, leading to optimal performance of our models. However, we observed that combining the statistically significant features with *n*-gram vocabulary spaces increases the performance, indicating the usefulness of our approach in the automatic diagnosis of prodromal dementia via the cookie theft picture description task. Orimaye *et al.* [39] highlighted the advantage of *n*-grams features as being easily computed without requiring manual annotation, which suggests that our models could be extended to other clinically recommended pictures for the same purpose.

V. COMPARISION WITH RELATED WORK

The comparison of our results alongside previous studies on automated assessment of picture descriptions may be challenging due to multiple reasons. Firstly, NLP tasks do not typically necessitate the usage of specific metrics, which resulted in different metrics across related studies. Some metrics are illustrative; yet, may mislead the evaluation on skewed datasets. Another challenge would be the different distributions of datasets across studies, which is the case in related studies despite using the same corpus of DementiaBank, meaning that results may vary when classes are differently distributed in the training and testing sets. In addition, as reported in the related work and summarized in Table 1, the number of samples varies across previous studies, which is a major constraint to direct comparisons. Furthermore, we noted that some authors had reported their best performing models instead of reporting the average of their results over multiple iterations. Most of these challenges were note clearly justified; for instance, regarding why they used a certain number of samples or only reported their best models. Consequently, we evaluated our work against the

most related studies with the highest performing models as follows:

A. AD VS HC

In the task of differentiating AD patients from healthy adults, we selected four baselines against our models; the work of Orimaye *et al.* [39], the work of Karlekar *et al.* [47] along with the work of Di Palo and Parde [48], and the recent study of Chen *et al.* [50].

Orimaye *et al.* [39] proposed a similar work in which an SVM classifier was trained with a compound of the top 1000 lexical, syntactic, and *n*-grams features to classify 242 samples of patients with possible and probable AD alongside 242 samples of healthy adults. They achieved 93% AUC, showing that such features could aid the diagnosis of AD. Karlekar *et al.* [47], on the other hand, applied multiple neural networks models (i.e., CNNs, LSTM-RNNs, and a combination of them), where they used transcripts belonging to all dementia etiologies with the accompanying POS tags, which were originally annotated by the custodian of the dataset. They dealt with these transcripts differently by dividing them into utterances and considering each utterance as a sample. Their best model was a combination of CNNs and LSTM with an accuracy of 91.1%. On the other side, Di Palo and Parde [48] reproduced the work of Karlekar *et al.* [47] and added a class weight and an attention mechanism to handle the skewness of the classes; however, the highest accuracy was 88%, achieved through an CNN-LSTM model fused with both the attention mechanism and class weight. They also reported a corresponding F1 score of 93.05% and AUC of 95.03%, setting benchmarks scores on this dataset with these metrics. Lately, Chen *et al.* [50] integrated a CNN model with a bidirectional GRU and added attention layer to distinguish 256 transcripts belonging to AD patients from 242 of healthy participants, where they reported an accuracy of 97.42%. We believe their work represents the state-of-the-art accuracy on this dataset for classifying people with AD from heathy adults.

When comparing our approach with these baselines, several perspectives must be pointed out: at first, with regard to Orimaye *et al.* [39], our approach differs in a few ways. We firstly utilized the entire dataset whereas they selected an equal subset of the majority to that of the minority to overcome the skewness of the classes. Additionally, more detailed experiments were provided in our study to show the behavior of our models with different sets of features. To the best of our knowledge, Orimaye *et al.* [39] used the original POS-annotated transcripts and also involved functional words when computing *n*-grams whereas both of them were discarded in our approach so, therefore, their work is not directly comparable to our work. Moreover, they used the age of participants as a feature, which could possibly contribute to the performance of their models. In addition, our models were evaluated with other metrics alongside AUC, showing better performance in comparison to their models.

In comparison with the studies of Karlekar *et al.* [47] and Di Palo and Parde [48], we used only transcripts belonging to the AD group rather than using combined transcripts of all dementia etiologies. More importantly, we did the classification at the transcript level rather than the utterance level, stressing on the fact that dividing a single transcript into utterances and considering each utterance as a sample might result in irrelevant samples thus unfair distributions across classes. For example, some utterances might be produced by a healthy participant yet have a degree of linguistic deficiency. This approach has not escaped criticism by other researchers [50]. Moreover, unlike all these three baselines, our study did not involve the originally POS-annotated transcripts, which could aid the diagnostic process yet might decrease the generalizability of the model to an unannotated dataset. To the best of our knowledge, they also considered filler words (e.g., uh, um, and hmm) which seen to be heavily used by patients with dementia thus contribute to the overall model performance [56]. On top of that, two of our models (i.e., GNB and MLP) outperformed these baselines, putting a new benchmark for classifying AD and HC groups from DementiaBank dataset.

As to the recent study of Chen *et al.* [50], we both present relatively comparable results for classifying the AD and HC groups. Nevertheless, our study differs in a few perspectives; at first, we analyzed the language ability across multiple stages, showing how it decreases as the cognitive impairment increases. Besides, while their study involved imbalanced classes, we tackled this issue with a subsampling technique thus made a use of the entire dataset. More importantly, they used the accuracy measure to evaluate their model despite using imbalanced classes, which perhaps could be misleading in such a case [80]. We contrarily assessed our models using multiple metrics.

In fact, to the best of our knowledge, an accuracy score achieved through an approach using imbalanced classes is not obviously representative as it would probably drop when using the same approach with balanced classes. In addition, the performance of such model may be limited in real applications as most of the population is dementia-free. As such, the performance of our models differs from that of Chen *et al.*

Furthermore, despite being promising, neural networks are generally considered computationally expensive than traditional algorithms. This is increased when it comes to RNNs, including both of LSTM and GRU recurrent units, due to its incapability to be parallelly computed, which leads to much more training time of RNNs [81]. Accordingly, we believe that traditional learners confront neural networks when both perform similarly. Table 6 summarizes these baselines alongside our models.

B. MCI VS HC AND AD

We evaluated our experiments for distinguishing the MCI group from the HC and AD groups against the work of Orimaye *et al.* [37] and the work of Al-hameed *et al.* [41]. Orimaye *et al.* [37] they introduced a few skip-gram models,

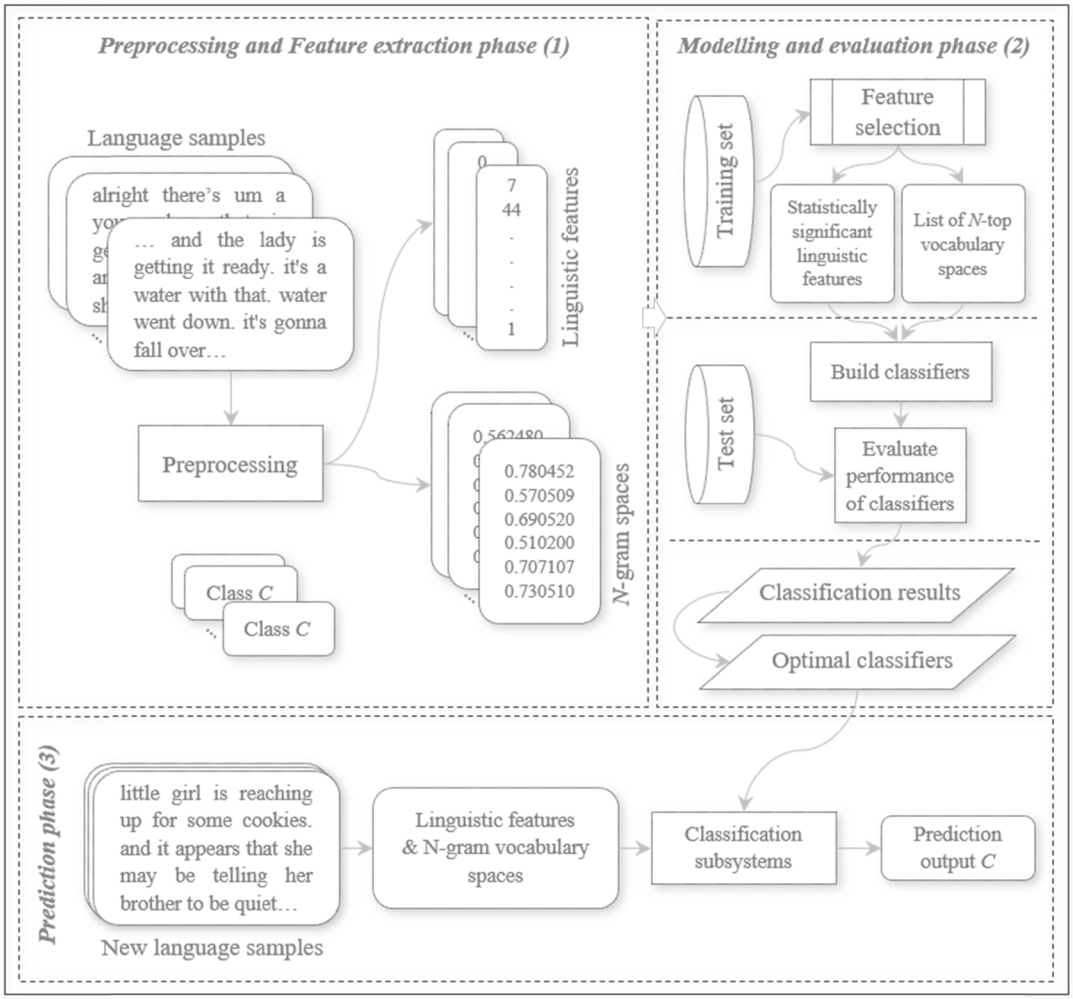

FIGURE 4. Illustration of the proposed framework.

where words are intermittently skipped while forming *n*-grams, to differentiate 19 MCI patients from 19 healthy adults. With the top 200 compound skip grams, they scored 98%, 97%, 97% and 99% of precision, recall, F1 score, and AUC, consecutively, employing SVM, NB, and LR classifiers. On the other side, Al-hameed *et al.* [41] used acoustic features to identify these groups from their speech recordings, where they extracted a total of 811 features then performed feature selection using ''SVMAttributeEval'' method under Weka software package. Besides, their study is the only identified study that utilized DementiaBank dataset to classify the MCI and AD groups. They reported an averaged accuracy of 97.3 and 95.6% across multiple visits with SVM classifier to distinguish MCI patients from healthy adults and form AD patients, respectively.

Our study has a few different aspects in contrast to these baselines. With regards to Orimaye *et al.* [37], while their study is concerned with classifying the MCI and HC groups, our study investigated the language deficiency of MCI patients against both HC and AD participants thus may shed light on the correlation between the language deficiency and progression of dementia. Besides, despite showing promising results, the skip-gram technique at word-level may fail to capture the grammatical complexity in patients with cognitive impairment. Another aspect that differentiates our approach would be the subsampling technique for handling the skewness of the classes. Notwithstanding that we both achieved comparable performance, our approach surpassed this baseline, slightly.

In respect to Al-hameed *et al.* [41], on the other end, the set of features is dissimilar as we used linguistic features, which may be more informative compared to acoustic markers. We also dealt with the skewed dataset in its original status whereas they used the Synthetic Minority Oversampling Technique (SMOTE) which artificially creates samples out of the minority class. Despite being an intelligent oversampling method for imbalanced data, SMOTE is considered time-consuming and complex [82]. Furthermore, in contrast to Al-hameed *et al.*, we evaluated our models using different metrics along with accuracy, since accuracy may mislead the performance when dealing with such skewed dataset as mentioned earlier. At last, two of our models outperform

these baselines (i.e., GNB for the classification of the MCI and HC groups and MLP for the MCI and AD groups) putting a benchmark for classifying these classes, demonstrated in Tables 7 and 8.

C. PoAD VS HC AND AD

In the tasks of identifying patients with initial AD (i.e., PoAD) from the HC and AD groups, we could not identify any previous work against our study. Consequently, we believe that this is the first investigation of its kind to explore language samples of patient with possible AD for the purpose of early detection. Despite their given small volume, we could achieve optimal performance for differentiating these classes with our proposed approach as shown in Tables 9 and 10.

VI. CONCLUSION

Our study shows that early stages of dementia can be efficiently diagnosed through linguistic patterns and deficits. Given the optimal performance achieved with our approach, we suggest that traditional screening tests for the initial diagnosis of prodromal dementia (e.g., MoCA) could be replaced with ML models, as depicted in Fig 4. One advantage of this study is the fact that it introduces original lexicosyntactic features and investigates their representations, in conjunction with other well-known lexicosyntactics, across different dementia etiologies. Besides, previous automated work barely discussed different stages of dementia in parallel, making of our study the first to investigate all transcripts belonging to different etiologies in the corpus of cookie theft picture description from DementiaBank. Another advantage of the current work would be the proposed statistical-based feature selection that handles normally and oddly distributed features. We also recommend random search over grid search for configuring the employed classifiers (i.e., GNB, MLP, and SVM) on this dataset.

On the other hand, a limitation we might think of would be the selection of samples for the lexicosyntactic statistical analysis, where we followed the technique performed by Orimaye *et al.* [39] in selecting individuals from the majority classes with matching demographic attributes to that of the minority classes. We also limited our linguistic features to lexicosyntactic, which is obviously one aspect of linguistic biomarkers, leaving others for future investigations. In addition, a current out-of-control limitation is the sparse nature of related datasets, which happens to be a limitation across various computational diagnostic studies [83], [84]. Besides, the most informative vocabulary spaces in our study are bounded to the cookie theft picture, meaning that a picture with different information units may result in different sequences of *n*-gram features. Finally, this study utilized a subset of DementiaBank dataset inherited from English speakers which warrants further investigations of whether findings in this paper are applicable to other languages.

Our scope of future research will address some of abovementioned limitations. First, we plan to extend and examine our lexicosyntactic features alongside other features including the language ability in social situations ''or so called pragmatic deficits'' which has been rarely investigated [85]. Another planned avenue for future work would be the usage of synthetic augmentation for prodromal dementia samples. Finally, yet importantly, we also plan to explore the effect of linked data for the same purpose, involving other related datasets such as dem@care [86].

ACKNOWLEDGMENT

This study utilizes a subset of DementiaBank dataset, which was collected at the University of Pittsburgh with support from NIH. DementiaBank is being maintained at Carnegie Mellon University under support by NICHD, NIDCD, NSF, SBE, and RIDIR.

REFERENCES

[1] *Older People Projected to Outnumber Children for First Time in US History*, US Census Bureau, Suitland-Silver Hill, MD, USA, 2018.
[2] Australian Bureau of Statistics. (Sep. 1, 2016). *Feature Article: Population by Age and Sex, Australia, States and Territories*.[Online]. Available: https://www.abs.gov.au/AUSSTATS/abs@.nsf/Previousproducts/3101.0Feature%20Article1Jun%202016
[3] *Future Directions for the Demography of Aging: Proceedings of a Workshop*, E. National Academies of Sciences and Medicine, Washington, DC, USA, 2018.
[4] N. Herrmann, K. L. Lanctôt, and D. B. Hogan, ''Pharmacological recommendations for the symptomatic treatment of dementia: The canadian consensus conference on the diagnosis and treatment of dementia 2012,'' *Alzheimer's Res. Therapy*, vol. 5, no. 1, p. S5, 2013.
[5] *What are the Treatments for Dementia?*, NHS, London, U.K., 2018.
[6] D. Kempler, ''Language changes in dementia of the Alzheimer type,'' in *Dementia and Communication*, R. Lubinski, Ed. San Diego, CA, USA: Singular, 1995.
[7] A. J. Mitchell and M. Shiri-Feshki, ''Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies,'' *Acta Psychiatrica Scandinavica*, vol. 119, no. 4, pp. 252–265, 2009.
[8] M. Canevelli, G. Grande, E. Lacorte, E. Quarchioni, M. Cesari, C. Mariani, G. Bruno, and N. Vanacore, ''Spontaneous reversion of mild cognitive impairment to normal cognition: A systematic review of literature and meta-analysis,'' *J. Amer. Med. Directors Assoc.*, vol. 17, no. 10, pp. 943–948, Oct. 2016.
[9] M. Malek-Ahmadi, ''Reversion from mild cognitive impairment to normal cognition,'' *Alzheimer Disease Associated Disorders*, vol. 30, no. 4, pp. 324–330, 2016.
[10] M. Ganguli, Y. Jia, T. F. Hughes, B. E. Snitz, C. C. H. Chang, and S. B. Berman, ''Mild cognitive impairment that does not progress to dementia: A population-based study,'' *J. Amer. Geriatrics Soc.*, vol. 67, no. 2, pp. 232–238, Feb. 2019.
[11] A. Abbott, ''Dementia: A problem for our age,'' *Nature*, vol. 475, pp. S2–S4, Jul. 2011.
[12] R. L. H. Handels, C. A. G. Wolfs, P. Aalten, M. A. Joore, F. R. J. Verhey, and J. L. Severens, ''Diagnosing Alzheimer's disease: A systematic review of economic evaluations,'' *Alzheimer's Dementia*, vol. 10, no. 2, pp. 225–237, Mar. 2014.
[13] Z. S. Nasreddine, N. A. Phillips, V. BÃdirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, ''The montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment,'' *J. Amer. Geriatrics Soc.*, vol. 53, no. 4, pp. 695–699, Apr. 2005.
[14] A. J. Mitchell, ''The mini-mental state examination (MMSE): Update on its diagnostic accuracy and clinical utility for cognitive disorders,'' in *Cognitive Screening Instruments: A Practical Approach*, A. J. Larner, Ed., 2nd ed. London, U.K.: Springer, 2017, pp. 37–48.
[15] R. F. Coen, D. A. Robertson, R. A. Kenny, and B. L. King-Kallimanis, ''Strengths and limitations of the MoCA for assessing cognitive functioning: Findings from a large representative sample of Irish older adults,'' *J. Geriatric Psychiatry Neurol.*, vol. 29, no. 1, pp. 18–24, 2016.

[16] M. S. Albert, S. T. DeKosky, and D. Dickson, "The diagnosis of mild cognitive impairment 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 Dementia*, vol. 7, no. 3, pp. 270–279, 2011.

[17] V. Taler and N. A. Phillips, "Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review," *J. Clin. Exp. Neuropsychol.*, vol. 30, no. 5, pp. 501–556, Jun. 2008.

[18] D. Kempler, E. L. Teng, M. Dick, I. M. Taussig, and D. S. Davis, "The effects of age, education, and ethnicity on verbal fluency," *J. Int. Neuropsychol. Soc.*, vol. 4, no. 6, pp. 531–538, Nov. 1998.

[19] T. Tombaugh, "Trail making test a and B: Normative data stratified by age and education," *Arch. Clin. Neuropsychol.*, vol. 19, no. 2, pp. 203–214, Mar. 2004.

[20] J. D. Henry, J. R. Crawford, and L. H. Phillips, "Verbal fluency performance in dementia of the Alzheimer's type: A meta-analysis," *Neuropsychologia*, vol. 42, no. 9, pp. 1212–1222, 2004.

[21] K. J. Murphy, J. B. Rich, and A. K. Troyer, "Verbal fluency patterns in amnestic mild cognitive impairment are characteristic of Alzheimer's type dementia," *J. Int. Neuropsychol. Soc.*, vol. 12, no. 04, pp. 570–574, Jul. 2006.

[22] C. Elsey, P. Drew, D. Jones, D. Blackburn, S. Wakefield, K. Harkness, A. Venneri, and M. Reuber, "Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics," *Patient Edu. Counseling*, vol. 98, no. 9, pp. 1071–1077, Sep. 2015.

[23] K. Fraser, J. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *J. Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[24] S. O. Orimaye, J. S.-M. Wong, and J. S. G. Fernandez, "Deep-deep neural network language models for predicting mild cognitive impairment," in *Proc. BAI@ IJCAI*, 2016, pp. 14–20.

[25] M. J. Ball, M. R. Perkins, N. Müller, and S. Howard, *The Handbook of Clinical Linguistics*. Hoboken, NJ, USA: Wiley, 2008.

[26] V. Rentoumi, G. Paliouras, E. Danasi, D. Arfani, K. Fragkopoulou, S. Varlokosta, and S. Papadatos, "Automatic detection of linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: A computational linguistics analysis," in *Proc. 8th IEEE Int. Conf. Cognit. Infocommun. (CogInfoCom)*, Sep. 2017, pp. 000033–000038.

[27] E. Giles, K. Patterson, and J. R. Hodges, "Performance on the boston cookie theft picture description task in patients with early dementia of the Alzheimer's type: Missing information," *Aphasiology*, vol. 10, no. 4, pp. 395–408, May 1996.

[28] X. Le, I. Lancashire, G. Hirst, and R. Jokel, "Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three british novelists," *Literary Linguistic Comput.*, vol. 26, no. 4, pp. 435–461, Dec. 2011.

[29] S. Ahmed, C. A. de Jager, A.-M. Haigh, and P. Garrard, "Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease," *Neuropsychology*, vol. 27, no. 1, p. 79, 2013.

[30] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2081–2090, Sep. 2011.

[31] M. Lehr, I. Shafran, E. Prud'hommeaux, and B. Roark, "Discriminative joint modeling of lexical variation and acoustic confusion for automated narrative retelling assessment," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2013, pp. 211–220.

[32] A. Satt *et al.*, "Evaluation of speech-based protocol for detection of early-stage dementia," presented at the 14th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), F. Bimbot *et al.*, Eds. Lyon, France, Aug. 2013, pp. 1692–1696.

[33] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert, and R. David, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's Dementia: Diagnosis, Assessment Disease Monit.*, vol. 1, no. 1, pp. 112–124, Mar. 2015.

[34] M. Asgari, J. Kaye, and H. Dodge, "Predicting mild cognitive impairment from spontaneous spoken utterances," *Alzheimer's Dementia: Transl. Res. Clin. Interventions*, vol. 3, no. 2, pp. 219–228, Jun. 2017.

[35] L. Toth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatloczki, Z. Banreti, M. Pakaski, and J. Kalman, "A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech," *Current Alzheimer Res.*, vol. 15, no. 2, pp. 130–138, Jan. 2018.

[36] M. Lehr, E. Prud'hommeaux, I. Shafran, and B. Roark, "Fully automated neuropsychological assessment for detecting mild cognitive impairment," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 1–4.

[37] S. O. Orimaye, K. Y. Tai, J. S.-M. Wong, and C. P. Wong, "Learning linguistic biomarkers for predicting mild cognitive impairment using compound skip-grams," presented at the Neural Inf. Process. Syst. (NIPS) Workshop Mach. Learn. Healthcare, 2015.

[38] S. O. Orimaye, J. S.-M. Wong, and K. J. Golden, "Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances," in *Proc. Workshop Comput. Linguistics Clin. Psychol., Linguistic Signal Clin. Reality*, 2014, pp. 78–87.

[39] S. O. Orimaye, J. S. Wong, K. J. Golden, C. P. Wong, and I. N. Soyiri, "Predicting probable Alzheimer's disease using linguistic deficits and biomarkers," *BMC Bioinf.*, vol. 18, p. 34, Jan. 2017.

[40] S. Al-Hameed, M. Benaissa, and H. Christensen, "Simple and robust audio-based detection of biomarkers for Alzheimer's disease," in *Proc. SLPAT Workshop Speech Lang. Process. Assistive Technol.*, Sep. 2016, pp. 32–36.

[41] S. Al-Hameed, M. Benaissa, and H. Christensen, "Detecting and predicting Alzheimer's disease severity in longitudinal acoustic data," in *Proc. Int. Conf. Bioinf. Res. Appl. (ICBRA)*, 2017, pp. 57–61.

[42] F. Haider, S. de la Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 272–281, Feb. 2020.

[43] R. Ben Ammar and Y. Ben Ayed, "Speech processing for early Alzheimer disease diagnosis: Machine learning based approach," in *Proc. IEEE/ACS 15th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Oct. 2018, pp. 1–8.

[44] M. Yancheva and F. Rudzicz, "Vector-space topic models for detecting Alzheimer's disease," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2016, pp. 2337–2346.

[45] A. Budhkar and F. Rudzicz, "Augmenting word2vec with latent Dirichlet allocation within a clinical application," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4095–4099.

[46] L. Hernández-Domínguez, S. Ratté, G. Sierra-Martínez, and A. Roche-Bergua, "Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task," *Alzheimer's Dementia: Diagnosis, Assessment Disease Monit.*, vol. 10, pp. 260–268, Mar. 2018.

[47] S. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, vol. 2, 2018, pp. 701–707.

[48] F. Di Palo and N. Parde, "Enriching neural models with targeted features for dementia detection," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics: Student Res. Workshop*, 2019, pp. 302–308.

[49] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Automatic hierarchical attention neural network for detecting AD," in *Proc. Interspeech*, Sep. 2019, pp. 4105–4109.

[50] J. Chen, J. Zhu, and J. Ye, "An attention-based hybrid network for automatic detection of Alzheimer's disease from narrative speech," in *Proc. Interspeech*, 2019, pp. 4085–4089.

[51] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis," *Arch. Neurol.*, vol. 51, no. 6, pp. 585–594, 1994.

[52] B. MacWhinney, *The Childes Project*. New York: Psychology Press, 2000, doi: 10.4324/9781315805641.

[53] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features," *Comput. Speech Lang.*, vol. 53, pp. 181–197, Jan. 2019.

[54] K. C. Fraser, J. A. Meltzer, N. L. Graham, C. Leonard, G. Hirst, S. E. Black, and E. Rochon, "Automated classification of primary progressive aphasia subtypes from narrative speech transcripts," *Cortex*, vol. 55, pp. 43–60, Jun. 2014.

[55] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proc. Workshop Comput. Linguistics Clin. Psychol., Linguistic Signal Clin. Reality*, 2014, pp. 27–37.

[56] S. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models," 2018, *arXiv:1804.06440*. [Online]. Available: http://arxiv.org/abs/1804.06440

[57] S. Wankerl, E. Nöth, and S. Evert, "An N-gram based approach to the automatic diagnosis of Alzheimer's disease from spoken language," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2017, pp. 1–5.
[58] G. Kavé and A. Dassa, "Severity of Alzheimer's disease and language features in picture descriptions," *Aphasiology*, vol. 32, no. 1, pp. 27–40, 2018.
[59] C. Rao and V. N. Gudivada, *Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications*. Amsterdam, The Netherlands: Elsevier, 2018.
[60] J. Eisenstein, *Introduction to Natural Language Processing*. Cambridge, MA, USA: MIT Press, 2019.
[61] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance," *Aphasiology*, vol. 14, no. 1, pp. 71–91, Jan. 2000.
[62] S. V. Pakhomov, G. E. Smith, D. Chacon, Y. Feliciano, N. Graff-Radford, and R. Caselli, "Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration," *Cogn. Behav. Neurol.*, vol. 23, no. 3, p. 165, 2010.
[63] N. Hardeniya, *NLTK Essentials*. Birmingham, U.K.: Packt Publishing, 2015.
[64] M. Albathan, Y. Li, and Y. Xu, "Using extended random set to find specific patterns," in *Proc. IEEE/WIC/ACM Int. Joint Conferences Web Intell. (WI) Intell. Agent Technol. (IAT)*, Aug. 2014, pp. 30–37.
[65] F. Dipaola, M. Gatti, V. Pacetti, A. G. Bottaccioli, D. Shiffer, M. Minonzio, R. Menè, A. Giaj Levra, M. Solbiati, G. Costantino, M. Anastasio, E. Sini, F. Barbic, E. Brunetta, and R. Furlan, "Artificial intelligence algorithms and natural language processing for the recognition of syncope patients on emergency department medical records," *J. Clin. Med.*, vol. 8, no. 10, p. 1677, Oct. 2019.
[66] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–45, Jan. 2018.
[67] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, Mar. 2003.
[68] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[69] A. Ivanov and G. Riccardi, "Kolmogorov-smirnov test for feature selection in emotion recognition from speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 5125–5128.
[70] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," *Knowl.-Based Syst.*, vol. 36, pp. 226–235, Dec. 2012.
[71] M. L. McHugh, "The chi-square test of independence," *Biochemia Medica*, vol. 23, no. 2, pp. 143–149, 2013.
[72] Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao, "A chi-square statistics based feature selection method in text classification," in *Proc. IEEE 9th Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Nov. 2018, pp. 160–163.
[73] Y. Li, L. Zhang, Y. Xu, Y. Yao, R. Y. K. Lau, and Y. Wu, "Enhancing binary classification by modeling uncertain boundary in three-way decisions," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 7, pp. 1438–1451, Jul. 2017.
[74] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.
[75] J. N. Mandrekar, "Receiver operating characteristic curve in diagnostic test assessment," *J. Thoracic Oncol.*, vol. 5, no. 9, pp. 1315–1316, Sep. 2010.
[76] M. Kuhn and K. Johnson, "Over-fitting and model tuning," in *Applied Predictive Modeling*. New York, NY, USA: Springer, 2013, pp. 61–92.
[77] M. Mukaka, "Statistics corner: A guide to appropriate use of correlation in medical research," *Malawi Med. J.*, vol. 24, no. 3, pp. 69–71, 2012.
[78] D. Kempler, S. Curtiss, and C. Jackson, "Syntactic preservation in Alzheimer's disease," *J. Speech, Lang., Hearing Res.*, vol. 30, no. 3, pp. 343–350, 1987.
[79] B. Liu, X. Wang, L. Lin, Q. Dong, and X. Wang, "A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis," *BMC Bioinf.*, vol. 9, no. 1, p. 510, Dec. 2008.
[80] J. Akosa, "Predictive accuracy: A misleading performance measure for highly imbalanced data," in *Proc. SAS Global Forum*, 2017, pp. 2–5.
[81] Z. Yu and G. Liu, "Sliced recurrent neural networks," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 2953–2964.
[82] A. Kumar, R. Bharti, D. Gupta, and A. K. Saha, "Improvement in boosting method by using RUSTBoost technique for class imbalanced data," in *Recent Developments in Machine Learning and Data Analytics* (Advances in Intelligent Systems and Computing), vol. 740, J. Kalita, V. Balas, S. Borah, and R. Pradhan, Eds. Singapore: Springer, 2019. [Online]. Available: http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-981-13-1280-9_5
[83] H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. M. Summers, "Improving computer-aided detection using convolutional neural networks and random view aggregation," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1170–1181, May 2016.
[84] J. Shi, S. Zhou, X. Liu, Q. Zhang, M. Lu, and T. Wang, "Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset," *Neurocomputing*, vol. 194, pp. 87–94, Jun. 2016.
[85] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski, "Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease," *Frontiers Aging Neurosci.*, vol. 7, p. 195, Oct. 2015.
[86] A. Karakostas, A. Briassouli, K. Avgerinakis, I. Kompatsiaris, and M. Tsolaki, "The Dem@Care experiments and datasets: A technical report," 2016, *arXiv:1701.01142*. [Online]. Available: http://arxiv.org/abs/1701.01142
[87] S. O. Orimaye, J. S.-M. Wong, and C. P. Wong, "Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia," *PLoS ONE*, vol. 13, no. 11, Nov. 2018, Art. no. e0205636.
[88] K. C. Fraser, K. Lundholm Fors, and D. Kokkinakis, "Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment," *Comput. Speech Lang.*, vol. 53, pp. 121–139, Jan. 2019.
[89] Z. Zhu, J. Novikova, and F. Rudzicz, "Detecting cognitive impairments by agreeing on interpretations of linguistic features," in *Proc. Conf. North*, 2019, pp. 1431–1441.
[90] Z. Guo, Z. Ling, and Y. Li, "Detecting Alzheimer's disease from continuous speech using language models," *J. Alzheimer's Disease*, vol. 70, no. 4, pp. 1163–1174, 2019.
[91] J. Fritsch, S. Wankerl, and E. Nöth, "Automatic diagnosis of Alzheimer's disease using neural network language models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5841–5845.

AHMED H. ALKENANI is currently pursuing the Ph.D. degree in data science with Queensland University of Technology, Brisbane, QLD, Australia. He is also a joint Ph.D. at the Australian e-Health Research Centre, CSIRO, Brisbane. His research interests include machine learning, conversational AI, computational linguistics with applications to biomedical and clinical text processing, smart environments, and healthcare technologies.

YUEFENG LI received the Ph.D. degree in computer science from Deakin University, Melbourne, VIC, Australia, in 2001. He is currently a Professor and the HDR Director with the School of Computer Science, Queensland University of Technology, Australia. He has published more than 190 refereed journals and conference papers. His work received 5,190 career citations (2,418, since 2015) with an overall h-index of 33 and an i10 index of 95. He has published ten articles with more than 100 citations and three articles with more than 200 citations (Google Scholar 10/2020). He has demonstrable experience in leading large-scale research projects and achieved many established research outcomes that has been published and highly-cited in top data mining journals and conferences, such as the IEEE Transactions on Knowledge and Data Engineering, IEEE Access, ACM SIGKDD Conference on Knowledge Discovery and Data Mining, CIKM, and ICDM. His research interests include text mining, machine learning, ontology learning, and AI-based data analysis. He serves as an Editor-in-Chief for *Web Intelligence*, an International Journal.

YUE XU (Member, IEEE) received the Ph.D. degree in computing from the University of New England, Armidale, NSW, Australia, in 2000. She was an Active Researcher in web intelligence and data mining. She is currently an Associate Professor with the School of Computer Science, Queensland University of Technology, Brisbane, QLD, Australia. She has published over 180 refereed articles. Her current research interests include recommender systems, text mining, pattern and association mining, and user interest and behavior modeling.

QING ZHANG (Member, IEEE) received the B.S. degree from Tsinghua University, Beijing, China, and the Ph.D. degree in computer science from the University of New South Wales, Sydney, NSW, Australia. He is currently a Principle Research Scientist and a Research Team Leader with the Australian E-Health Research Centre, CSIRO. He also leads the Health Internet of Things Research Team and the Research and Development of CSIRO's Smarter Safer Homes Platform, which aims at supporting seniors' ageing-in-place through wireless sensor and advanced AI techniques. He is evaluated in the world largest random control clinical trial in Australia. He holds four patents in Australia, China, U.S., and Europe. His research interests include AI, data mining, the IoT design, and big data analytic. His platform received the QLD State iAward and the National iAward. He served as the Chair for the EMBC Chapter of the IEEE QLD Section and a Regular Reviewer for *Journal of Medical Internet Research*, the IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, and so on.