Dual Dropout Ranking of Linguistic Features for Alzheimer's Disease Recognition

Xiaoquan Ke*, Man-Wai Mak*, Jinchao Li† and Helen M. Meng†
* The Hong Kong Polytechnic University, Hong Kong SAR
† The Chinese University of Hong Kong, Hong Kong SAR
xiaoquan.ke@connect.polyu.hk, enmwmak@polyu.edu.hk, {jcli, hmmeng}@se.cuhk.edu.hk

***Abstract*—We propose a feature ranking method called dual dropout ranking (DDR) to identify the most discriminative linguistic features for Alzheimer's disease (AD) detection. The proposed DDR is based on a dual-net neural architecture that separates feature selection and recognition into two neural networks (operator and selector), which are alternatively and cooperatively trained to optimize the performance of both feature selection and AD recognition. The operator is trained on the features obtained from the selector to reduce classification loss. The selector is optimized to predict the operator's performance using as few selected features as possible. DDR ranks the features according to the probabilities that the corresponding features should be purged (or kept). The DDR and other feature ranking methods were evaluated on the ADReSS dataset. Results show that the default linguistic feature set in ADReSS comprises many redundant features and that using feature ranking methods can improve the accuracy of AD recognition. Using the most discriminative feature subset (9 features) discovered by DDR, we obtain an F_1 score of 88.9% on the test set of ADReSS, which is 9.8% (absolute) higher than what the default feature set can achieve.**

I. Introduction

Alzheimer's disease (AD) is a severe cognitive impairment seriously affecting the health and daily lives of many older adults.[1] According to the World Alzheimer's Report [1], dementia prevalence in people aged 60 years and over ranges between 4.6% to 8.7% across different regions around the world. It is estimated in 2015 that globally 46.8 million people are living with dementia. This number will roughly double every 20 years, projected to reach 74.7 million in 2030 and 131.5 million in 2050. The global costs of dementia were estimated at $818 billion in 2015, which is about 1.09% of global GDP. This has huge quality of life impact not only for individuals with dementia, but also their families and caretakers.

Currently, AD is diagnosed through brain imaging [2], identification of apolipoprotein E genotypes [3], measuring the level of brain-derived neurotrophic factor [4], cerebrospinal fluid exams [5], and other laboratory measures. In addition to these measures, because AD also manifests language impairment [6, 7], automatic recognition of AD through speech and language analyses has gathered attention in the research community. Some studies used acoustic information (e.g., speech/silence segments [8] and voice quality [9, 10]) from speech waveforms to discover potential AD. Some studies utilized a combination of features, such as rhythm-inspired features with acoustic features [11] and paralinguistic features with linguistic fluency features [12]. Another approach is to integrate the decisions from multiple modalities. In [13], the modalities are based on acoustic, cognitive, and linguistic features, and in [14, 15], they are based on acoustic and textual features. Some other studies used deep neural networks (DNNs) to learn high-level representations from speech transcriptions automatically [16, 17].

While various type of features have been used for AD recognition, it is still unclear which features or combination of features are more effective. Our study investigates feature ranking methods to identify the most discriminative linguistic features that distinguish AD from Non-AD. We propose a novel feature ranking method called dual dropout ranking (DDR) to identify the most representative linguistic features. The proposed DDR is based on a dual-net neural architecture that separates feature selection and recognition into two neural networks which are alternatively and cooperatively trained to integrate feature selection and AD recognition into a coherent process.

The remainder of this paper is organized as follows. Section II presents related work. Section III presents the technical details of DDR. Section IV and Section V describe experimental setup and results, and concluding remarks are given in Section VI.

II. Related Work

There were studies investigating the relevance of various features for AD recognition. For example, Weiner *et al.* [18] extracted features from biographic interviews to predict the development of AD after 5 or 12 years. They reduced the dimensionality of the original feature set by nested forward feature selection. It was found that feature selection can significantly improve prediction performance. Weiner *et al.* [19] also used nested forward feature selection to identify the most commonly selected features during cross-validations for the state screening of AD. Alhanai *et al.* [20] extracted demographic, audio, and text features and used an elastic-net based logistic regression model to identify the discriminative features for cognitive impairment recognition. The method

[1] https://www.who.int/news-room/fact-sheets/detail/dementia

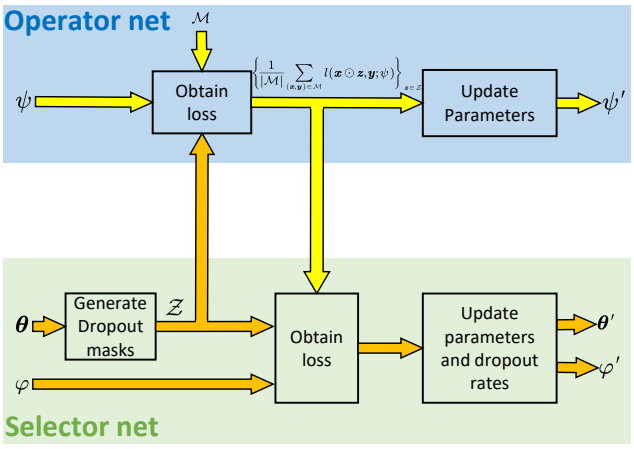

Fig. 1. The dual-net architecture of DDR. ψ and φ represent the network parameters of the operator and selector, respectively. $\boldsymbol{\theta}$ comprises the dropout rates at the input layer of the selector.

ranks features according to the sparsity regularization coefficients of the regression model.

There have been other feature ranking methods based on sparsity regularization, such as LASSO ($L1$ penalty) [21], group LASSO [22, 23], and L1-norm [24]. Sparsity regularization has also been adopted in deep learning-based methods for feature ranking. For example, in deep feature selection (Deep FS) [25], elastic-net regularization is imposed on the weights between the input and the first hidden layer, and in dropout feature ranking (Dropout FR) [26], LASSO regularization is imposed through a penalty term.

There have also been deep learning feature ranking methods that do not have sparsity regularization. For instance, in [27], the authors ranked the features according to their net positive contribution to the classification tasks.

III. Dual Dropout Ranking

A. Dropout for Feature Ranking

Feature ranking aims to rank the importance of individual features according to some criteria, where the criteria typically reflect the features' contributions to the learning performance [28].

In dropout [29], nodes were purged according to their dropout rates. Therefore, the *higher* the dropout rate, the *lower* the rank of the feature, and feature ranking amounts to determining the dropout rates of individual input nodes. To formulate the dropout rate of a feature, we adopt an approach similar to Dropout FR. More specifically, given a dropout rate vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k, \ldots, \theta_d)$ and a dropout mask vector $\boldsymbol{z} = (z_1, z_2, \ldots, z_k, \ldots, z_d)$, we denote the distribution of $\boldsymbol{z}$ as $q(\boldsymbol{z}) = \prod_{k=1}^{d} q\left(z_k \mid \theta_k\right) = \prod_{k=1}^{d} \mathrm{Bern}\left(z_k \mid \theta_k\right)$, where θ_k is the dropout rate for the k^{th} feature, and $z_k \in \{0, 1\}$ is the corresponding dropout mask. This gives us a fully factorized Bernoulli distribution that focuses on feature ranking. Suppose $\boldsymbol{x} = (x_1, x_2, \ldots, x_k, \ldots, x_d)$ is an input feature vector. During the forward pass, we place the dropout mask vector on the input layer, that is $\boldsymbol{x} \odot \boldsymbol{z}$, where $\odot$ is the element-wise product (Hadamard product).

B. Trainable Dropout Rate

In ordinary dropout, the dropout rates are fixed hyper-parameters. Instead of fixing the dropout rates, we treat them as *trainable* parameters. To optimize the dropout rates, we relax the binary dropout masks to *soft* dropout masks as follows:

$$\boldsymbol{z} = \mathrm{sigmoid}\left(\frac{1}{t}\left[\log \boldsymbol{\theta} - \log(\mathbf{1} - \boldsymbol{\theta}) + \log \boldsymbol{u} - \log(\mathbf{1} - \boldsymbol{u})\right]\right), \tag{1}$$

where $\boldsymbol{u} \in \mathbb{R}^d$ follows the $\mathrm{Uniform}(\mathbf{0}, \mathbf{1})$ distribution and t is a normalization constant, which is set to 0.1 in our experiments. Note that this relaxation has also been used in Concrete Dropout [30] and Dropout FR [26]. Eq. (1) suggests that $q(\boldsymbol{z})$ places most of the mass to either $z_k = 0$ or $z_k = 1$ to closely resemble the binary dropout mask. With the continuous relaxation in Eq. (1), the dropout rates can be optimized through back-propagation, and we can gradually select the optimal features $\boldsymbol{x} \odot \boldsymbol{z}$ along with the optimization of the dropout rates.

C. Learning Algorithm

Suppose $\mathcal{M} = \{\mathcal{X}, \mathcal{Y}\}$ is a mini-batch comprising $|\mathcal{M}|$ pairs of $\boldsymbol{x}$ and $\boldsymbol{y}$, where $\boldsymbol{x} \in \mathcal{X}$ is a feature vector of size d, and $\boldsymbol{y} \in \mathcal{Y}$ is the corresponding target. By sampling the uniform distribution in Eq. (1), we obtain *several* soft dropout mask vectors $\boldsymbol{z} = (z_1, z_2, \ldots, z_k, \ldots, z_d)$ and form a dropout mask subset $\mathcal{Z}$ of size $|\mathcal{Z}|$. The learning algorithm of DDR is defined in Eq. (2), where $\mathcal{L}_O\left(\mathcal{M}, \mathcal{Z}; \psi\right)$ is the operator's objective, $l(\boldsymbol{x} \odot \boldsymbol{z}, \boldsymbol{y}; \psi)$ is the cross-entropy loss for binary/multiclass classification or the mean square error (MSE) loss for regression, and ψ is the operator's parameters. $\mathcal{L}_S\left(\mathcal{Z}; \varphi\right)$ is the selector's objective, $f_S(\boldsymbol{z}, \varphi)$ is the selector's output, and φ is the selector's parameters. The training procedure of dual-net is depicted in Figure 1. During training, the operator and selector are trained alternately. The alternate training procedure is depicted in Appendix A.

1) Operator: The *operator* is trained on the features obtained from the selector to reduce classification loss. For each iteration, given the dropout mask subset $\mathcal{Z}$ from the selector, the selected features $\{\boldsymbol{x} \odot \boldsymbol{z}\}_{\boldsymbol{x} \in \mathcal{X}, \boldsymbol{z} \in \mathcal{Z}}$ are fed to the operator, and the operator's learning performance based on the selected features is obtained. Given the selected features $\boldsymbol{x} \odot \boldsymbol{z}$, $\frac{1}{|\mathcal{M}|} \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{M}} l(\boldsymbol{x} \odot \boldsymbol{z}, \boldsymbol{y}; \psi)$ is the learning performance of the operator on the mini-batch $\mathcal{M}$. By enumerating $\boldsymbol{z}$ in $\mathcal{Z}$, we obtain the average learning performance of the operator on the mini-batch. Then, we update the operator's parameters and pass the operator's learning performance to the selector as a feedback indicating how well the operator performs on the selected features. Different from the sparsity regularization methods that also incorporate the regularization into the network, the operator only focuses on reducing classification loss. Given the selected features, the operator's architecture

$$\text{Operator's objective:} \quad \mathcal{L}_O\left(\mathcal{M}, \mathcal{Z}; \psi\right) = \frac{1}{|\mathcal{Z}||\mathcal{M}|} \sum_{\boldsymbol{z}\in\mathcal{Z}} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{M}} l(\boldsymbol{x}\odot\boldsymbol{z}, \boldsymbol{y}; \psi) \tag{2a}$$

$$\text{Selector's objective:} \quad \mathcal{L}_S\left(\mathcal{Z}; \varphi\right) = \frac{1}{|\mathcal{Z}|} \sum_{\boldsymbol{z}\in\mathcal{Z}} \left\{ \left| f_S(\boldsymbol{z}; \varphi) - \frac{1}{|\mathcal{M}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{M}} l(\boldsymbol{x}\odot\boldsymbol{z}, \boldsymbol{y}; \psi) \right| \Big/ \sum_{k=1}^{d} (1 - z_k) \right\} \tag{2b}$$

2) Selector: The *selector* learns to predict the operator's learning performance using as few selected features as possible. The mean absolute error (MAE) between $f_S(\boldsymbol{z}, \varphi)$ and $\frac{1}{|\mathcal{M}|}\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{M}} l(\boldsymbol{x}\odot\boldsymbol{z}, \boldsymbol{y}; \psi)$ requires that the selector closely predicts the operator's learning performance. The constraint $\sum_{k=1}^{d}(1 - z_k)$ on the denominator of Eq. (2b) requires that most dropout masks in $\boldsymbol{z}$ become 0; so the selector only selects a small number of features when predicting the operator's learning performance.

After training and updating the selector's parameters and dropout rates, we have the updated dropout rate vector $\boldsymbol{\theta}'$. Through sampling the uniform distribution in Eq. (1), we obtain several new soft dropout mask vectors $\boldsymbol{z}'$ from the updated dropout rate vector $\boldsymbol{\theta}'$ and form a new dropout mask subset $\mathcal{Z}'$ for the next iteration. In practical implementation, the dropout mask vector fed to the selector is $\boldsymbol{z}\odot\boldsymbol{z}'$, where $\boldsymbol{z}\in\mathcal{Z}$ and $\boldsymbol{z}'\in\mathcal{Z}'$.

IV. Experimental Setup

A. Datasets

We used two datasets in the experiments. One is a synthetic dataset and another was obtained from the AD Recognition Through Spontaneous Speech Challenge (ADReSS) [31].

The synthetic data set was designed for evaluating the capability of classifiers and feature selection algorithms in solving a multi-dimensional XOR problem [32]. By grouping 8 corners of a 3-dimensional hypercube $(v_0, v_1, v_2) \in \{-1, 1\}^3$ into the tuples (v_0v_2, v_1v_2), we have 4 sets of vectors and their negations $\{\boldsymbol{v}^{(c)}, -\boldsymbol{v}^{(c)}\}_{c=1}^{4}$, where c is the class index. For example, the tuple $(v_0v_1, v_1v_2) = (1, -1)$ corresponds to $c = 2$, where $\boldsymbol{v}^{(2)} = [1, 1, -1]^\mathsf{T}$. The points in class c are generated from the distribution $\frac{1}{2}[\mathcal{N}(\boldsymbol{v}^{(c)}, 0.5\boldsymbol{I}_3) + \mathcal{N}(-\boldsymbol{v}^{(c)}, 0.5\boldsymbol{I}_3)]$, where $\boldsymbol{I}_3$ is a 3×3 identity matrix. Each sample is additionally accompanied by 7 Gaussian noise features with zero mean and unit variance, leading to a 10-dimensional feature vector.

The ADReSS dataset comprises the speech recordings and corresponding annotated transcriptions of 78 AD patients and 78 healthy controls (HCs), where 108 age- and gender-matched subjects were grouped into the training set, and the remaining subjects were grouped into the test set. 34 linguistic features were obtained based on the annotated transcriptions using CLAN [33].

B. Training Procedure and Recognition

For Deep FS and Dropout FR, we directly adopted the network architectures "34–170–170–2" and the default hyper-parameters (e.g., batch size, learning rates, etc.) in [26]. The most important hyper-parameters for Deep FS and Dropout FR is the regularization coefficient. A grid search was carried out to optimize the regularization coefficient (from 0.01 to 1) using leave-one-subject-out (LOSO) cross-validations on the training set of ADReSS. For DDR, we adopted the architecture "34–60–30–20–2" for the operator network and "34–100–50–10–1" for the selector network. We adopted the same batch size and learning rates in [26] to train the operator network and selector network. The most important hyper-parameter for DDR is the initial dropout rate. A grid search was carried out to optimize the initial dropout rate (from 0.1 to 0.9) for DDR. All other experimental settings for Deep FS, Dropout FR and DDR are the same except for the optimal hyper-parameters for each method.

The goal is to determine the most discriminative linguistic features that can effectively recognize the AD patients in the test set. We used feature ranking methods to rank linguistic features and identified the most discriminative features. The identified features were then used for training a linear SVM with a box constraint of 0.003 to recognize AD.

V. Results and Analysis

In this section, we evaluate the feature ranking methods on the synthetic and ADReSS datasets.

A. Analysis of Keep Probabilities

In this evaluation, we show that DDR is robust to different random seeds.

1) Synthetic: We trained a DDR network (Figure 1) on the synthetic dataset. After training, the keep probabilities $(\mathbf{1} - \boldsymbol{\theta})$ of the features for 20 random seeds are depicted in Figure 2. It shows that the keep probabilities associated with the valid features (v_0, v_1, v_2) converge to 1, whereas the noise features $(v_3 \sim v_9)$ have keep probabilities close to 0. This result suggests that DDR can effectively identify the valid features.

2) ADReSS: We used the training set in ADReSS to train another DDR network. Figure 3 depicts the keep probabilities of the features for 20 random seeds. The results show that DDR converges to almost the same keep probabilities for different random seeds. The top 5 features have high keep probabilities, while other low rank features have keep probabilities close to 0, which means that DDR can select a small number of discriminative features confidently.

B. LOSO on the Training Set

For the deep learning-based feature ranking methods (Deep FS, Dropout FR, and DDR), we ranked all of the linguistic

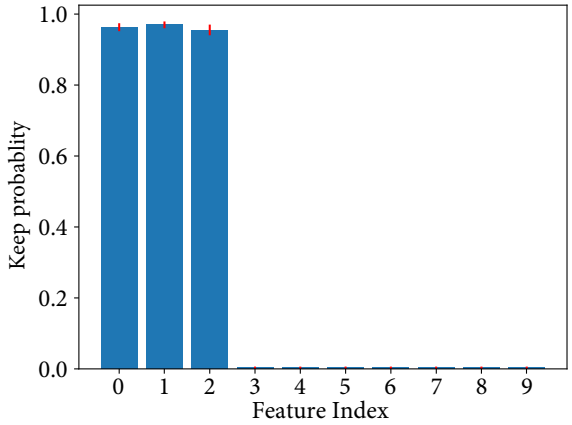

Fig. 2. The keep probabilities ($\mathbf{1}-\boldsymbol{\theta}$) of 10 features in the synthetic dataset for 20 random seeds. Indexes 0–3 and 4–9 correspond to the valid and invalid features, respectively. The blue bars and the red error bars denote the means and two times the standard deviations of 20 random seeds, respectively.

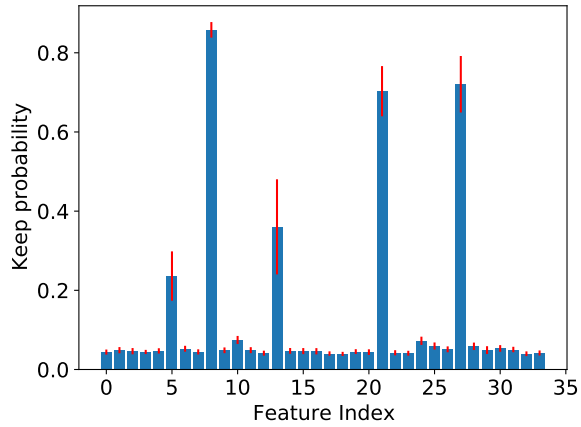

Fig. 3. The keep probabilities ($\mathbf{1}-\boldsymbol{\theta}$) of 34 linguistic features in the ADReSS dataset for 20 random seeds. Refer to the caption of Figure 2 for the meaning of the blue bars and the red error bars.

features and determined their feature relevance based on LOSO cross-validations on the training set of ADReSS. For each fold in the LOSO, we trained the deep neural networks on the training partition (107 subjects) and evaluated them on the test partition (one subject). Because each fold uses different partitions for training, the feature relevance in different folds is not the same. We used the following strategy to select the final set of features. For each fold, we chose n features according to their ranking (keep probabilities). For a K-fold LOSO, this amounts to Kn features after running the K folds. Because some features are more relevant than others, some features may appear in most of the K folds and there are many repeated features in the Kn features. By assuming that frequently selected features are more relevant, we sorted the Kn features

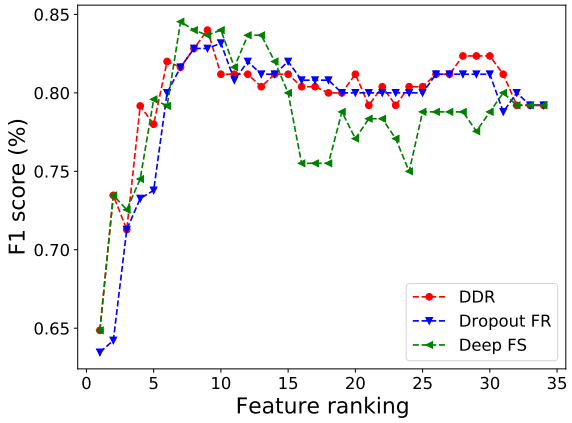

Fig. 4. F_1 scores achieved by different deep learning-based feature ranking methods based on the LOSO cross-validations on the training set of ADReSS.

in descending order of their number of occurrences and picked the top n unique features in the sorted list.

Figure 4 shows the recognition performance of the deep learning-based feature ranking methods. It shows that when using the original feature set, the LOSO F_1 score on the training set was 79.2%. Figure 4 also shows that the original feature set comprises many redundant features and that using feature ranking methods to obtain the most discriminative subsets can improve the accuracy of AD recognition. Deep FS achieves its highest F_1 score (84.5%) when the number of selected feature is 7, but the F_1 score becomes unstable and fluctuating when more features were selected. When the number of selected features is 9, DDR achieves its highest F_1 score (84.0%).

We obtained the *optimal* feature subsets by varying the number of selected features through LOSO cross-validations (Figure 4). The optimal feature subset was obtained when the highest F_1 score was achieved in the LOSO cross-validations. For Deep FS, the size of the optimal feature subset is 7, for Dropout FR it is 10, and for DDR it is 9. In the next subsection, we apply the optimal feature subsets to the test set of ADReSS.

C. Recognition Performance on the Test Set of ADReSS

In this subsection, we evaluate the recognition performance on the test set of ADReSS. Table I shows the optimal feature subsets discovered by the traditional feature ranking methods such as LASSO [21], L1-norm [24], Univ [34], FDR [35], and SVM-RFE [36]. Table I shows that LASSO and L1-norm are not effective for this dataset because their performance is the same as without feature selection. Among all traditional methods, SVM-RFE performs the best. The numbers in the brackets are the sizes of the feature subsets. Finding the optimal number of selected features is challenging because it depends on the nature of the input features. For the traditional feature ranking methods, we directly varied the number of selected features and reported their best recognition performance on the test

TABLE I
THE RECOGNITION PERFORMANCE BASED ON THE FEATURE SUBSETS DISCOVERED BY TRADITIONAL FEATURE RANKING METHODS. THE NUMBERS IN THE BRACKET ARE THE SIZES OF THE FEATURE SUBSETS.

Feature ranking methods	Recognition accuracy
None	81.2% (34)
LASSO [21]	81.2% (33)
L1-norm [24]	81.2% (15)
Univ [34]	83.3% (24)
FDR [35]	83.3% (24)
SVM-RFE [36]	87.5% (24)

TABLE II
RECOGNITION PERFORMANCE OF DEEP LEARNING FEATURE RANKING METHODS ON THE TEST SET OF ADRESS. THE NUMBERS IN THE BRACKET ARE THE SIZES OF THE FEATURE SUBSETS.

Feature ranking methods	F_1 score
None	79.1% (34)
Deep FS [25]	83.7% (7)
Dropout FR [26]	83.7% (10)
DDR	**88.9**% (9)

set. Therefore, the performance of these traditional methods in Table I is slightly over-estimated. Despite this over-optimism, the performance of these traditional methods is still poorer than that of the proposed DDR and is only comparable to deep learning-based methods.

From the last subsection, we have obtained the optimal feature subsets discovered by the deep learning-based feature ranking methods (Deep FS, Dropout FR and DDR) based on LOSO cross-validations on the training set of ADReSS. We now apply the optimal feature subsets to the test set of ADReSS. We show the recognition performance in Table II. Table II shows that using the optimal feature subset discovered by DDR, we obtain an F_1 score of 88.9% on the test set, which is 9.8% (absolute) higher than what the original feature set can achieve. In the LOSO cross-validations, the optimal feature subset discovered by Deep FS achieves the highest F_1 score (84.5%), but the recognition performance of the feature subset is unsatisfactory on the test set, which indicates that the feature subset discovered by Deep FS does not generalize well from the training set to the test set.

We finally depict the optimal feature subset discovered by DDR in Table III. Some of the known specificities of the selected linguistic features are also shown in Table III. Explanations of the selected features can be found in [33].

VI. CONCLUSIONS

In this paper, a novel feature ranking method is presented. The original feature set contains many redundant features. Therefore, using the discriminative feature subsets discovered by feature ranking methods can improve the accuracy of AD detection. The highest F_1 score is achieved by a feature subset (9 features) discovered by the proposed feature ranking method, which is 9.8% (absolute) higher than what the original feature set can achieve. Some of the linguistic features may have intrinsic patterns for distinguishing the AD patients from the healthy controls. In the future, more features are expected to be included in the feature ranking process, which can find more representative features for AD recognition.

ACKNOWLEDGEMENT

This work was in part supported by Research Grands Council of Hong Kong, Theme-based Research Scheme (Ref.: T45-407/19-N).

VII. APPENDIX A

Algorithm 1 Alternate learning algorithm of DDR

Require: Operator network with parameters ψ and selector network with parameters φ
Require: The size of dropout mask subset $|\mathcal{Z}|$, size of mini-batch $|\mathcal{M}|$, and number of training iterations n
Output: Dropout rates $\boldsymbol{\theta}_n$

Initialize dropout rates as $\boldsymbol{\theta}_0$
for $i \leftarrow 1$ to n **do**
  Obtain a dropout mask subset $\mathcal{Z}$ with size $|\mathcal{Z}|$ using Eq. (1)
  for $j \leftarrow 1$ to $|\mathcal{Z}|$ **do**
    Compute the operator loss given $\boldsymbol{z}_i^{(j)}$:

$$\ell_{O,i}^{(j)} = \frac{1}{|\mathcal{M}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{M}} l(\boldsymbol{x} \odot \boldsymbol{z}_i^{(j)}, \boldsymbol{y}; \psi_i)$$

  end for
  Compute the average operator loss on $\mathcal{Z}$:

$$\mathcal{L}_O\left(\mathcal{M}, \mathcal{Z}; \psi_i\right) = \frac{1}{|\mathcal{Z}|} \sum_{j=1}^{|\mathcal{Z}|} \ell_{O,i}^{(j)}$$

  Update operator network's parameters:

$$\psi_i \leftarrow \psi_i - \eta \nabla_\psi \mathcal{L}_O\left(\mathcal{M}, \mathcal{Z}; \psi_i\right)|_{\psi=\psi_i}$$

  for $j \leftarrow 1$ to $|\mathcal{Z}|$ **do**
    Compute the selector loss given $\boldsymbol{z}_i^{(j)}$:

$$\ell_{S,i}^{(j)} = \left| f_S(\boldsymbol{z}_i^{(j)}; \varphi_i) - \ell_{O,i}^{(j)} \right| \Big/ \sum_{k=1}^{d} (1 - z_{i,k}^{(j)})$$

  end for
  Compute the average selector loss on $\mathcal{Z}$:

$$\mathcal{L}_S\left(\mathcal{Z}; \varphi_i\right) = \frac{1}{|\mathcal{Z}|} \sum_{j=1}^{|\mathcal{Z}|} \ell_{S,i}^{(j)};$$

  Update selector network's parameters:

$$\varphi_i \leftarrow \varphi_i - \eta \nabla_\varphi \mathcal{L}_S\left(\mathcal{Z}; \varphi_i\right)|_{\varphi=\varphi_i}$$

  Update dropout rates:[2]

$$\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_i - \eta \sum_{j=1}^{|\mathcal{Z}|} \nabla_{\boldsymbol{z}} \mathcal{L}_S\left(\mathcal{Z}; \varphi_i\right) \nabla_{\boldsymbol{\theta}} \boldsymbol{z}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_i, \boldsymbol{z}=\boldsymbol{z}_i^{(j)}}$$

end for

[2]The gradient is based on the chain rule: $\frac{\partial \mathcal{L}_S}{\partial \boldsymbol{\theta}} = \frac{\partial \mathcal{L}_S}{\partial \boldsymbol{z}} \cdot \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\theta}}$.

TABLE III
THE OPTIMAL FEATURE SUBSET DISCOVERED BY DDR. THE PARENTHESIZED VALUES ARE THE PERCENTAGE OF THE FEATURES BEING SELECTED DURING THE LOSO CROSS-VALIDATIONS. *AD*: ALZHEIMER'S DISEASE, *HCs*: HEALTHY CONTROLS.

Linguistic feature	**Known specificity**
% PresP: Percentage of present participle (100%)	Yuan *et al.* [37] reported that AD patients used relatively fewer *present particles* (*-ing* verbs) compared with the HCs.
Words/min: Words per minute (100%)	AD could be detected through the analysis of voice activity detection and *speech rate* tracking [38].
FREQ types: Total word types (100%)	–
% pro: Percentage of pronouns (100%)	Ahmed *et al.* [39] reported changes in *the number of pronouns*, and Jarrold *et al.* [40] reported an increase in *the proportion of pronouns* in AD patients.
% adv: Percentage of adverbs (100%)	–
% Nouns: Percentage of nouns (100%)	Jarrold *et al.* [40] reported a decrease in *the proportion of nouns* in AD patients.
% Word Errors: Percentage of words that are coded as errors (98.1%)	–
noun/verb ratio: Total no. of nouns / total no. of verbs (81.5%)	AD patients may have more difficulty *naming verbs than nouns* [41], and Robinson *et al.* [42] found that AD patients performed worse on a picture-naming task for *verbs than nouns*.
% conj: Percentage of conjunctions (75.9%)	–

REFERENCES

[1] M. J. Prince, A. Wimo, M. M. Guerchet, G. C. Ali, Y.-T. Wu, and M. Prina, *World Alzheimer Report 2015 - The Global Impact of Dementia: An analysis of prevalence, incidence, cost and trends*. Alzheimer's Disease International, 2015.

[2] K. A. Johnson, N. C. Fox, R. A. Sperling, and W. E. Klunk, "Brain imaging in Alzheimer disease," *Cold Spring Harbor Perspectives in Medicine*, vol. 2, no. 4, pp. a006213, 2012.

[3] C.-C. Liu, T. Kanekiyo, H. Xu, and G. Bu, "Apolipoprotein E and Alzheimer disease: Risk, mechanisms and therapy," *Nature Reviews Neurology*, vol. 9, no. 2, pp. 106–118, 2013.

[4] J.-H. Song, J.-T. Yu, and L. Tan, "Brain-derived neurotrophic factor in Alzheimer's disease: Risk, mechanisms, and therapy," *Molecular Neurobiology*, vol. 52, no. 3, pp. 1477–1493, 2015.

[5] L. M. Shaw, J. Arias, K. Blennow, D. Galasko, J. L. Molinuevo, S. Salloway, S. Schindler, M. C. Carrillo, J. A. Hendrix, A. Ross, J. Illes, C. Ramus, and S. Fifer, "Appropriate use criteria for lumbar puncture and cerebrospinal fluid testing in the diagnosis of Alzheimer's disease," *Alzheimer's & Dementia*, vol. 14, no. 11, pp. 1505–1521, 2018.

[6] D. Kempler, "Language changes in dementia of the Alzheimer type," *Dementia and communication*, vol. 1, pp. 98–114, 1995.

[7] J. Reilly, J. Troche, and M. Grossman, "Language processing in dementia,", *The Handbook of Alzheimer's Disease and Other Dementias*, vol. 7, 2011.

[8] 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 Monitoring*, vol. 1, no. 1, pp. 112–124, 2015.

[9] J. J. G. Meilán, F. Martínez-Sánchez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana, "Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia?" *Dementia and Geriatric Cognitive Disorders*, vol. 37, no. 5–6, pp. 327–334, 2014.

[10] K. L. de Ipiña, J. B. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, C. M. Travieso, M. Ecay-Torres, P. Martínez-Lage, and H. Eguiraun, "On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature," *Cognitive Computation*, vol. 7, no. 1, pp. 44–55, 2015.

[11] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. Blackburn, and H. Christensen, "Improving detection of Alzheimer's disease using automatic speech recognition to identify high-quality segments for more robust feature extraction," in *Proc. Interspeech 2020*, 2020, pp. 4961–4965.

[12] E. L. Campbell, R. Y. Mesía, L. Docío-Fernández, and C. García-Mateo, "Paralinguistic and linguistic fluency features for Alzheimer's disease detection," *Computer Speech & Language*, vol. 68, pp. 101–198, 2021.

[13] U. Sarawgi, W. Zulfikar, N. Soliman, and P. Maes, "Multimodal inductive transfer learning for detection of Alzheimer's dementia and its severity," *arXiv preprint arXiv:2009.00700*, 2020.

[14] A. Pompili, T. Rolland, and A. Abad, "The INESC-ID multi-modal system for the ADReSS 2020 challenge," *arXiv preprint arXiv:2005.14646*, 2020.

[15] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, 'or Alzheimer's dementia through spontaneous speech,"

in *Proc. Interspeech 2020*, 2020, pp. 2222–2226.

[16] 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*, 2019, pp. 4085–4089.

[17] S. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models," *arXiv preprint arXiv:1804.06440*, 2018.

[18] J. Weiner, C. Frankenberg, J. Schroder, and T. Schultz, "Speech reveals future risk of developing dementia: Predictive dementia screening from biographic interviews," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 674–681.

[19] J. Weiner and T. Schultz, "Selecting features for automatic screening for dementia based on speech," in *Proc. International Conference on Speech and Computer*, 2018, pp. 747–756.

[20] T. Alhanai, R. Au, and J. Glass, "Spoken language biomarkers for detecting cognitive impairment," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 409–416.

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[22] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[23] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.

[24] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proc. of the Fifteenth International Conference on Machine Learning*, 1998, pp. 82–90.

[25] Y. Li, C.-Y. Chen, and W. W. Wasserman, "Deep feature selection: Theory and application to identify enhancers and promoters," *Journal of Computational Biology*, vol. 23, no. 5, pp. 322–336, 2016.

[26] C.-H. Chang, L. Rampasek, and A. Goldenberg, "Dropout feature ranking for deep learning models," *arXiv preprint arXiv:1712.08645*, 2017.

[27] D. Roy, K. S. R. Murty, and C. K. Mohan, "Feature selection using deep neural networks," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–6.

[28] M. Wojtas and K. Chen, "Feature importance ranking for deep learning," *arXiv preprint arXiv:2010.08973*, 2020.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[30] Y. Gal, J. Hron, and A. Kendall, "Concrete dropout," *arXiv preprint arXiv:1705.07832*, 2017.

[31] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge," *arXiv preprint arXiv:2004.06833*, 2020.

[32] J. Chen, M. Stern, M. J. Wainwright, and M. I. Jordan, "Kernel feature selection via conditional covariance minimization," *arXiv preprint arXiv:1707.01164*, 2017.

[33] B. MacWhinney, "The CHILDES project: Tools for analyzing talk (third edition): Volume i: Transcription format and programs, volume II: The database," *Computational Linguistics*, vol. 26, no. 4, pp. 657–657, 2000.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[35] S. Wang, D. Li, X. Song, Y. Wei, and H. Li, "A feature selection method based on improved Fisher's discriminant ratio for text sentiment classification," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8696–8702, 2011.

[36] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, 2002.

[37] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease," in *Proc. Interspeech 2020*, 2020, pp. 2162–2166.

[38] S. Luz, "Longitudinal monitoring and detection of Alzheimer's type dementia from spontaneous speech data," in *Proc. IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, 2017, pp. 45–46.

[39] S. Ahmed, A.-M. F. Haigh, C. A. de Jager, and P. Garrard, "Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease," *Brain*, vol. 136, no. 12, pp. 3727–3737, 2013.

[40] 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. of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 27–37.

[41] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, pp. 407–422, 2015.

[42] K. M. Robinson, M. Grossman, T. White-Devine, and M. D'Esposito, "Category-specific difficulty naming with verbs in Alzheimer's disease," *Neurology*, vol. 47, no. 1, pp. 178–182, 1996.