Interspeech 2022
18-22 September 2022, Incheon, Korea

Automatic Selection of Discriminative Features for Dementia Detection in Cantonese-Speaking People

Xiaoquan Ke[1], *Man-Wai Mak*[1], *Helen M. Meng*[2]

[1]The Hong Kong Polytechnic University, Hong Kong SAR
[2]The Chinese University of Hong Kong, Hong Kong SAR

xiaoquan.ke@connect.polyu.hk, enmwmak@polyu.edu.hk, hmmeng@se.cuhk.edu.hk

Abstract

Dementia is a severe cognitive impairment that affects the health of older adults and creates a burden on their families and caretakers. This paper analyzes diverse features extracted from spoken languages and selects the most discriminative features for dementia detection. The paper presents a deep learning-based feature ranking method called dual-net feature ranking (DFR). The proposed DFR utilizes a dual-net architecture, where two networks (called operator and selector) are alternatively and cooperatively trained to simultaneously perform feature selection and dementia detection. The DFR interprets the contribution of individual features to the predictions of the selector network using all of the selector's parameters. The DFR was evaluated on the Cantonese JCCOCC-MoCA Elderly Speech Dataset. Results show that the DFR can significantly reduce feature dimensionality while identifying small feature subsets with comparable or superior performance than the whole feature set. The selected features have been uploaded to `https://github.com/kexquan/AD-detection-Feature-selection`.

Index Terms: Dementia detection, feature ranking, feature selection, explanatory neural networks

1. Introduction

Dementia is a severe cognitive impairment that seriously affects the health and daily lives of the afflicted individuals.[1] The most common form of dementia is the Alzheimer's Disease (AD). Fortunately, with effective detection of early dementia, disease-modifying medications and interventions are possible [1].

Dementia can be diagnosed through several means, including neuropsychological assessments, brain scans, blood tests, etc. Dementia also manifests itself as spoken language deficits [2]. Studies have found that dementia-induced language impairment could be found in patients years before the disease was diagnosed [3]. Research also showed that individuals with progressive cognitive decline exhibit subtle linguistic impairment even in the pre-symptomatic stages of the disease [4]. These findings suggest that dementia can be detected using spoken language processing (SLP) techniques.

Recently, automatic detection of dementia through speech and language analyses has gathered attention in the research community. Several studies investigated the relevance of various spoken-language features for dementia detection. For example, Weiner *et al.* [5] extracted features from biographic interviews to predict the development of AD after 5 years. They reduced the dimensionality of the original feature set by nested forward feature selection (FS) and found that FS can significantly improve prediction performance. Weiner *et al.* [6] also used nested forward FS to identify the most frequently selected features during cross-validation for the state screening of AD.

In this paper, we investigate various FS methods for dementia detection. We also propose a novel deep-learning based feature-ranking method called dual-network feature ranking (DFR) to rank and select features. The proposed DFR utilizes a dual-net architecture, where two networks (called operator and selector) are alternatively and cooperatively trained to simultaneously perform feature selection and dementia detection. Despite the complex relationship between the network's output and its input variables, the DFR can interpret the contribution of the input variables to the network's prediction. The DFR is evaluated on a Cantonese corpus called JCCOCC-MoCA [7].

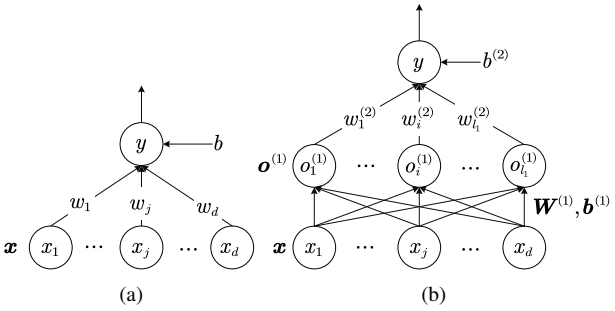

Figure 1: *Network architectures for interpreting the contributions of individual features (input nodes) to the network's prediction. (a) A 1-layer network with linear output is equivalent to the linear regression model. (b) Feature importance can be obtained from the weights of a multi-layer network. See Section 2.1 for details.*

2. Dual-net Feature Ranking

2.1. Variable Selection in Deep Neural Networks

We consider the usual *linear regression* model. Given d predictor variables $\boldsymbol{x} = (x_1, \ldots, x_j, \ldots, x_d)^{\mathsf{T}}$, the response variable $f(\boldsymbol{x})$ is predicted by

$$f(\boldsymbol{x}) \approx \hat{f}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_j x_j + \cdots + \hat{\beta}_d x_d, \quad (1)$$

where $\hat{f}(\boldsymbol{x})$ is a linear model and $\hat{\beta}_1, \ldots, \hat{\beta}_j, \ldots, \hat{\beta}_d$ are its parameters. Since there is a one-to-one correspondence between the model parameters and the predictor variables, the effect of a given predictor x_j on the model $\hat{f}(\boldsymbol{x})$ can be evaluated through

This work was in part supported by Research Grants Council of Hong Kong, Theme-based Research Scheme (Ref.: T45-407/19-N). We thank Prof. C.Y. Kwok for providing us with the JCCOCC-MoCA dataset.

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

10.21437/Interspeech.2022-10122

the value of $\hat{\beta}_j$ [8]. In particular, when the model parameter $\hat{\beta}_j \gg 0$, the predictor x_j may have a significant positive effect on the model. When $\hat{\beta}_j \approx 0$, we may say that x_j contributes little to the prediction of $\hat{f}(\boldsymbol{x})$, and it can be removed from the model.

A 1-layer fully-connected network with one linear output node (Fig. 1(a)) is equivalent to the linear regression model. Given a d-dimensional input vector $\boldsymbol{x} = (x_1, \ldots, x_j, \ldots, x_d)^\mathsf{T}$, the network's output y is (omitting the bias for simplicity):

$$y = \boldsymbol{w}^\mathsf{T}\boldsymbol{x} = w_1 x_1 + \cdots + w_j x_j + \cdots + w_d x_d, \quad (2)$$

where $\boldsymbol{w} = (w_1, \ldots, w_j, \ldots, w_d)^\mathsf{T}$ is the network's weight vector. As Eq. (2) is equivalent to Eq. (1), we can also interpret the effect of a given input variable x_j on the prediction of the network through the value of w_j. In particular, when $w_j \gg 0$, we may say that x_j has a significant positive effect on the network's output. When $w_j \approx 0$, we may say that x_j is irrelevant to the network's output and can be removed from the network. We formulate a one-to-one correspondence between the input $\boldsymbol{x}$ and the network's weight vector $\boldsymbol{w}$:

$$\text{diag}\{\boldsymbol{x}\}\boldsymbol{w} = (w_1 x_1, \ldots, w_j x_j, \ldots, w_d x_d)^\mathsf{T}, \quad (3)$$

where $\text{diag}\{\boldsymbol{x}\}$ is a diagonal matrix with $\{x_j\}$ in its diagonal. By setting $\boldsymbol{x} = \mathbf{1}$ in Eq. (3), we obtain the *feature importance vector* $\boldsymbol{c}$:

$$\boldsymbol{c} = \text{diag}\{\mathbf{1}\}\boldsymbol{w} = \boldsymbol{w} = (w_1, \ldots, w_j, \ldots, w_d)^\mathsf{T}. \quad (4)$$

Eq. (4) suggests that the bigger the value of w_j, the more important the input variable x_j. Therefore, we can select the important input variables according to $\boldsymbol{c}$.

Fig. 1(b) depicts a 2-layer network with the hidden layer having l_1 nodes and the output layer having one node. Suppose $\boldsymbol{W}^{(1)}$ is a $d \times l_1$ weight matrix connecting the input $\boldsymbol{x}$ to the hidden layer and $\boldsymbol{b}^{(1)} \in \mathbb{R}^{l_1}$ is the corresponding bias vector. Also, suppose $\boldsymbol{w}^{(2)} = (w_1^{(2)}, \ldots, w_i^{(2)}, \ldots, w_{l_1}^{(2)})^\mathsf{T}$ is the weight vector of the output layer and $b^{(2)}$ is the bias. Given a d-dimensional input vector $\boldsymbol{x}$, the output of the hidden layer is (omitting the bias for simplicity):

$$\boldsymbol{o}^{(1)} = g\left(\left(\boldsymbol{W}^{(1)}\right)^\mathsf{T}\boldsymbol{x}\right) \in \mathbb{R}^{l_1}, \quad (5)$$

where $g(\cdot)$ is a non-linear activation function, e.g., sigmoid. And the output of the network is:

$$y = (\boldsymbol{w}^{(2)})^\mathsf{T}\boldsymbol{o}^{(1)} = (\boldsymbol{w}^{(2)})^\mathsf{T} g\left(\left(\boldsymbol{W}^{(1)}\right)^\mathsf{T}\boldsymbol{x}\right). \quad (6)$$

Comparing Eq. (2) and Eq. (6) and following Eq. (3), we can also formulate a one-to-one correspondence between the input $\boldsymbol{x}$ and the network's parameters:

$$g\left(\text{diag}\{\boldsymbol{x}\}\boldsymbol{W}^{(1)}\right)\boldsymbol{w}^{(2)} = (v_1 x_1, \ldots, v_j x_j, \ldots, v_d x_d)^\mathsf{T}. \quad (7)$$

Again, by setting $\boldsymbol{x} = \mathbf{1}$, we can obtain the feature importance vector:

$$\boldsymbol{c} = g\left(\text{diag}\{\mathbf{1}\}\boldsymbol{W}^{(1)}\right)\boldsymbol{w}^{(2)} = g\left(\boldsymbol{W}^{(1)}\right)\boldsymbol{w}^{(2)} \in \mathbb{R}^d. \quad (8)$$

Note that $\boldsymbol{c}$ is also a d-dimensional vector with c_j corresponding to the input variable x_j. Similar results can be extended to an L-layer neural network with weight matrices $\{\boldsymbol{W}^{(i)}, i = 1, 2, \ldots, L-1\}$ for the hidden layers and weight vector $\boldsymbol{w}^{(L)}$ for the output layer. The feature importance vector $\boldsymbol{c}$ for the L-layer network is:

$$\boldsymbol{c} = g\left(g\left(g\left(\boldsymbol{W}^{(1)}\right)\boldsymbol{W}^{(2)}\right)\cdots\boldsymbol{W}^{(L-1)}\right)\boldsymbol{w}^{(L)} \in \mathbb{R}^d. \quad (9)$$

2.2. Learning Algorithm

In Section 2.1, we formulate a d-dimensional feature importance vector $\boldsymbol{c}$ that reflects the feature importance of the input variables. We use $\boldsymbol{c}$ to determine the contribution of the input variables to the output of a deep neural network. Specifically, the input variable x_j with a larger c_j will have a greater contribution to the output. Based on the feature importance vector $\boldsymbol{c}$, we propose a deep-learning-based FS method called dual-net feature ranking (DFR). It comprises two deep neural networks (called operator and selector), as shown in Fig. 2. During training, the operator and selector are trained alternately.

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. The learning algorithm of DFR is defined in Eq. (10), where $\mathcal{L}_O(\mathcal{M}, \mathcal{Z}; \psi)$ is the operator's objective, $l(\boldsymbol{x} \odot \boldsymbol{z}, \boldsymbol{y}; \psi)$ is either the cross-entropy loss for classification or the mean squared error (MSE) loss for regression, and ψ denotes the operator's parameters. $\mathcal{L}_S(\mathcal{Z}; \varphi)$ is the selector's objective, $f_S(\boldsymbol{z}, \varphi)$ is the selector's output, and $\varphi = \{\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, \ldots, \boldsymbol{W}^{(L-1)}, \boldsymbol{w}^{(L)}\}$ contains the selector's parameters. The training procedure of DFR is depicted in Fig. 2.

Operator. The operator is trained on the features selected by the selector to reduce the loss $\mathcal{L}_O(\mathcal{M}, \mathcal{Z}; \psi)$. The feature mask vector $\boldsymbol{z}$ in the feature mask subset $\mathcal{Z}$ indicates which features have been selected. For each iteration, given the feature 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 $\mathcal{M}$. Then, we pass the operator's learning performance to the selector as a feedback indicating how well the operator performs on the selected features.

Selector. The selector learns to predict the operator's learning performance using the selected features. The mean absolute error 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 accurately predicts the operator's learning performance. At the beginning of each iteration, the selector produces the feature mask subset $\mathcal{Z}$ using the following steps: (1) *Retain the best feature mask vector.* We retain the best feature mask vector $\boldsymbol{z}_1$ that achieves the best learning performance (e.g., the smallest cross-entropy loss) in the last iteration. (2) *Determine an optimal feature mask vector.* We compute the feature importance vector $\boldsymbol{c}$ using Eq. (9) based on the selector's parameters φ. According to the feature importance vector $\boldsymbol{c}$, we generate an optimal feature mask vector $\boldsymbol{z}_2$ by assigning the top s features with mask 1 and the rest of $d - s$ features with mask 0. (3) *Generate candidate feature mask vectors.* To increase the diversity of the feature mask vectors, we generate several candidate feature mask vectors $\{\boldsymbol{z}_3, \ldots, \boldsymbol{z}_{|\mathcal{Z}|}\}$ by randomly flipping p masks in $\boldsymbol{z}_2$. (4) *Produce the feature mask subset.* Finally, we produce the feature mask subset $\mathcal{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \ldots, \boldsymbol{z}_{|\mathcal{Z}|}\}$.

$$\text{Operator's objective:} \quad \mathcal{L}_O(\mathcal{M}, \mathcal{Z}; \psi) = \frac{1}{|\mathcal{Z}||\mathcal{M}|}\sum_{\boldsymbol{z}\in\mathcal{Z}}\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{M}} l(\boldsymbol{x} \odot \boldsymbol{z}, \boldsymbol{y}; \psi) \quad (10a)$$

$$\text{Selector's objective:} \quad \mathcal{L}_S(\mathcal{Z}; \varphi) = \frac{1}{|\mathcal{Z}|}\sum_{\boldsymbol{z}\in\mathcal{Z}}\left\{\left| f_S(\boldsymbol{z}; \varphi) - \frac{1}{|\mathcal{M}|}\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{M}} l(\boldsymbol{x} \odot \boldsymbol{z}, \boldsymbol{y}; \psi)\right|\right\} \quad (10b)$$

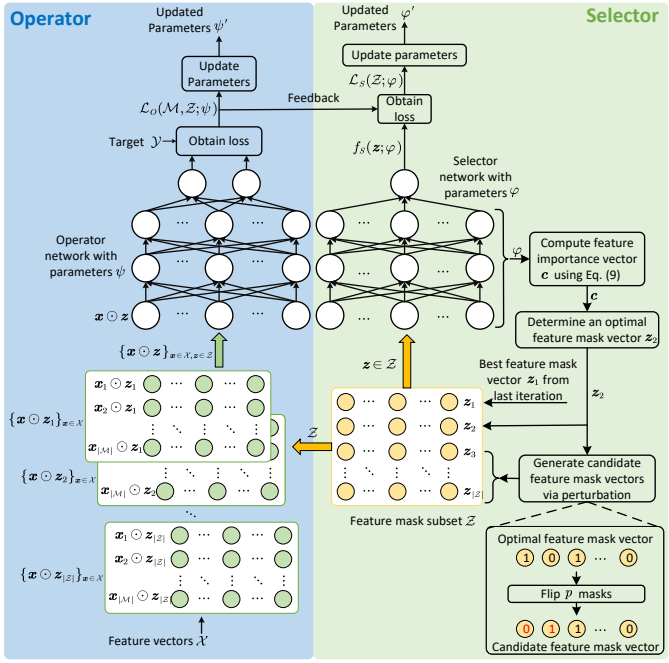

Figure 2: *The dual-net architecture and training procedure of DFR. At the beginning of each iteration, the selector's parameters φ are used to compute the feature importance vector $\boldsymbol{c}$.*

Table 1: *The characteristics of the JCCOCC-MoCA dataset. HCs: healthy controls.*

Spoken languages	Cantonese
Tasks	Fluency tests (animals, fruits, and vegetables)
Number of participants	43 HCs and 43 possible dementia
Number of samples	129 HCs and 129 possible dementia
Manual transcriptions provided	Yes

3. Feature Engineering

3.1. Cantonese JCCOCC-MoCA Elderly Speech Dataset

The JCCOCC Montreal Cognitive Assessment (MoCA) Cantonese Speech corpus was collected by the CUHK Jockey Club Centre for Osteoporosis Care and Control. A MoCA test was given to each participant for assessing their cognitive capability. According to the assessment results and MoCA scores, the participants were divided into four groups: (1) 205 healthy older adults (HCs); (2) 16 older adults having mild neurocognitive disorders (mild NCD); (3) 17 older adults suffering from mild cognitive impairment (MCI); and (4) 10 older adults suffering from major NCD.

For detecting dementia, we combined the mild NCD, MCI, and major NCD into the possible dementia category and randomly picked 43 healthy participants from the HCs. From the speech recording of each participant, after excluding the assessor, we extracted three 1-minute fluency tests (animals, fruits, and vegetables), resulting in 3 samples for each participant. The transcriptions corresponding to the fluency tests were also extracted. The data used for dementia detection are shown in Table 1.

3.2. Feature Extraction

We differentiate two categories of features: transcription-based and speech-based. The transcription-based features are extracted from the manual transcriptions, which capture the semantic, syntactic, and lexical aspects of the speaker's spoken language. The speech-based features are extracted from the corresponding speech recordings, which contain a variety of acoustic characteristics of the speakers.

老虎，<PAU>獅子，<PAU>駱駝，<PAU>犀牛，<PAU>海馬，<PAU>諗車，豺狼，<PAU>諗，<PAU>豹，<PAU>大笨象，<PAU>諗，水牛，諗，冇喇，唔記得，天空有嘢啲諗，燕諗，燕子，諗，諗，<PAU>諗，海鷗，諗，<PAU>係咁多喇，冇啦，諗唔到，喎，唔記得喇。

Figure 3: *An example of 1-minute transcription tagged with pauses '<PAU>'.*

Table 2: *The 5 statistical characteristics of pauses that are extracted from 6 duration groups.*

Statistical characteristic	Description
#p	Number of pauses per minute
%p/word ratio	Pause-to-word ratio
p duration	Total duration of pauses per minute
p mean duration	Mean duration of pauses
%p duration/word duration	Pause-duration-to-word-duration ratio

3.2.1. Transcription-based Features

(1) *Lexical features*. Based on the transcriptions, the following lexical features were extracted:[2] the number of sentences per minute and the average number of words per sentence. Then, the pycantonese library was utilized to parse the transcriptions.[3] After that the following features were appended to the feature set: part-of-speech (POS) counts per minute, POS ratio, the ratio of pronoun to noun, and the ratio of noun to verb.

(2) *ELECTRA features*. We consider the ELECTRA model [9] pre-trained on a large Cantonese corpus as a feature extractor.[4] More specifically, we fed the transcriptions to the pre-trained ELECTRA model and extracted the representations from the last layer of the model. Similar pre-trained language models for other languages, e.g., pre-trained BERT models, have also been used for dementia detection [10].

(3) *Pause features*. In [11], the authors demonstrated that pauses can function as word-finding, as planning at the word, phrase, and narrative levels, and as pragmatic compensation when other interactional and narrative skills deteriorate. Thus, we included the pause features for dementia detection. In the JCCOCC-MoCA dataset, pauses and their durations have been tagged. An example of 1-minute transcription tagged with pauses is shown in Fig. 3. We divided the pauses into six groups according to their durations: G_1 (pauses between 0.05s–0.5s), G_2 (pauses between 0.5s–1s), G_3 (pauses between 1s–2s), G_4 (pauses between 2s–3s), G_5 (pauses between 3s–4s), and G_6 (pauses longer than 4s). We used the statistical characteristics of the pauses as the pause features, as illustrated in Table 2. For each duration group, we extracted the 5 statistical characteristics.

3.2.2. Speech-based Features

(1) *Acoustic features*. We follow the standard pipelines in the COVFEFE toolbox [12] to extract the acoustic features, which include formants, loudness, pitch, zero-crossing rate, etc.

(2) *COVAREP features*. COVAREP features [13] are comprehensive acoustic features, which include prosodic features

[2]The lexical features were extracted using the toolbox: https://github.com/SPOClab-ca/COVFEFE.

[3]https://pycantonese.org/

[4]https://huggingface.co/toastynews/electra-hongkongese-base-discriminator

Table 3: *Classification performance of different feature types on the JCCOCC-MoCA dataset. The numbers in the brackets are the sizes of the feature sets.*

Feature set	5 repetitions of leave-n-subject-out CV			
	ACC	PRE	REC	F1
Lexical (113)	0.541	0.555	0.556	0.532
ELECTRA (768)	0.572	0.576	0.577	0.557
Pause (30)	0.550	0.557	0.560	0.539
Acoustic (30)	0.481	0.486	0.488	0.463
COVAREP (518)	0.389	0.409	0.419	0.376
IS10 (1582)	0.519	0.535	0.539	0.508
Emobase (988)	0.531	0.544	0.551	0.518
eGeMAPS (88)	0.533	0.545	0.548	0.522
All features (4117)	**0.584**	**0.590**	**0.591**	**0.566**

(fundamental frequency and voicing), voice quality features, and spectral features. Rohanian *et al.* [14] used the COVAREP features from the audio modality for multi-modal cognitive impairment detection.

(3) *INTERSPEECH 2010 Paralinguistic Challenge Features (IS10)*. IS10 is a feature set useful for emotion recognition [15] and bipolar disorder recognition [16], which include PCM loudness, eight log Mel-frequency bands, eight line spectral frequency pairs, F0 envelope, voicing probability, jitter, and shimmer [17].

(4) *Emobase*. The Emobase feature set [18] comprises mel-frequency cepstral coefficients (MFCC), fundamental frequency (F0), F0 envelope, line spectral pairs (LSP), etc. Wang *et al.* [19] used the Emobase features in a multi-modal attention network for AD Detection.

(5) *eGeMAPS*. The Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [20] contains 88 features that are selected based on their potential for characterizing physiological changes in voice production.

4. Experiments and Results

4.1. Performance of Different Feature Types

We first evaluate the recognition performance of all the features *before* FS on the JCCOCC-MoCA dataset. We adopted a leave-n-subject-out cross-validation (CV) in which the samples of the same speakers are grouped into either the training partition (TR) or the test partition (TS) for each fold. We adopted a Gaussian SVM with a box constraint of 1 to identify the possible dementia and the HCs. We repeated the leave-n-subject-out CV 5 times and averaged the results, as shown in Table 3. Table 3 shows that the transcription-based features generally outperform the speech-based features and that combining all the feature sets achieves the highest detection accuracy.

4.2. Applying DFR for Dementia Detection

This subsection reports the performance of DFR and some strong supervised feature ranking methods on the JCCOCC-MoCA dataset. These supervised feature ranking methods include deep feature selection (DFS) [21], dropout feature ranking (DropoutFR) [22], dual dropout ranking (DDR) [23], deep feature importance ranking (DeepFIR) [24], and random forest (RF) [25]. We combined all the feature sets listed in Section 3.2 to form 4117-dimensional feature vectors and applied a leave-n-subject-out CV on the feature vectors. On the training partitions (TR) of individual folds, we applied the feature ranking methods described above to rank and select features. The selected features were then used to train a Gaussian SVM with a box constraint of 1 to identify the possible dementia and the HCs.

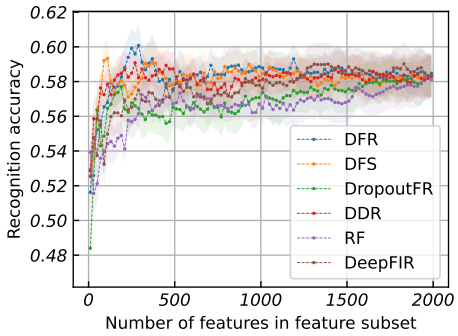

Figure 4: *Classification performance under different number of selected features on the JCCOCC-MoCA dataset.*

Table 4: *Classification performance of the selected feature subsets on the JCCOCC-MoCA dataset. The numbers in the brackets are the sizes of the feature subsets.*

Method	Highest accuracy	Average accuracy
DFS [21]	0.593 (110)	0.581
DropoutFR [22]	0.583 (1970)	0.569
DDR [23]	0.591 (270)	0.581
DeepFIR [24]	0.590 (1830)	0.577
RF [25]	0.583 (1970)	0.566
DFR	**0.601** (290)	**0.583**

Considering that the feature dimensionality is very high, filtering methods were utilized to reduce the feature dimensionality before applying the strong supervised feature ranking methods. On the TR of individual folds, we utilized mutual information (MutInfo) to reduce the feature dimensionality from 4117 to 2000.

We used the same network architecture "2000–512–128–32–2" and their default hyper-parameters for all the deep learning-based feature ranking methods. For the DFR, DDR, and DeepFIR that have the dual-net architecture, we adopted the architectures "2000–512–128–32–2" for the operator network and "2000–512–128–32–1" for the selector network. We repeated the leave-n-subject-out CV 5 times and averaged the results of all folds under different numbers of selected features,[5] as shown in Fig. 4. Fig. 4 shows that the average accuracy exhibits a curve that initially ascends and then becomes stable when the number of selected features increases. The average accuracy of all the feature subsets and the highest accuracy that can be achieved by the methods are reported in Table 4. Table 4 shows that DFR identifies some small feature subsets that achieve the highest accuracy among the methods investigated. These small feature subsets also achieve comparable or superior performance compared with the full, default feature set.

5. Conclusions

We presented a deep-learning based feature ranking method and evaluated the method on the Cantonese JCCOCC-MoCA Elderly Speech Dataset. The method interprets the contribution of the input variables to the prediction of the deep neural network using the network's parameters. The discriminative features selected by the method achieve comparable or superior performance compared with the full feature set. Future work may investigate the biological aspects of the selected features.

[5] We created different sizes of feature subsets by including features in the order of their rankings.

6. References

[1] J. L. Cummings, R. Doody, and C. Clark, "Disease-modifying therapies for Alzheimer disease: Challenges to early intervention," *Neurology*, vol. 69, no. 16, pp. 1622–1634, Oct. 2007.

[2] J. Reilly, J. Troche, and M. Grossman, "Language processing in dementia," in *Handbook Alzheimer's Disease Other Dementias*, Sep. 2011, pp. 336–368.

[3] L. Mickes, J. T. Wixted, C. Fennema-Notestine, D. Galasko, M. W. Bondi, L. J. Thal, and D. P. Salmon, "Progressive impairment on neuropsychological tasks in a longitudinal study of preclinical Alzheimer's disease." *Neuropsychology*, vol. 21, no. 6, pp. 696–705, Nov. 2007.

[4] D. Beltrami, L. Calzà, G. Gagliardi, E. Ghidoni, N. Marcello, R. R. Favretti, and F. Tamburini, "Automatic identification of mild cognitive impairment through the analysis of Italian spontaneous speech productions," in *Proc. Int. Conf. Lang. Resourc. and Eval. (LREC)*, May 2016, pp. 2086–2093.

[5] J. Weiner, C. Frankenberg, J. Schröder, and T. Schultz, "Speech reveals future risk of developing dementia: Predictive dementia screening from biographic interviews," in *Proc. IEEE Autom. Speech Recognit. Underst. Workshop (ASRU)*, Dec. 2019, pp. 674–681.

[6] J. Weiner and T. Schultz, "Selecting features for automatic screening for dementia based on speech," in *Proc. Lect. Notes Comput. Sci.*, Sep. 2018, pp. 747–756.

[7] S. S. Xu, M. W. Mak, K. H. Wong, H. Meng, and C. Y. Kwok, "Speaker turn aware similarity scoring for diarization of speech-based cognitive assessments," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2021, pp. 1299–1304.

[8] Y. Hechtlinger, "Interpretation of prediction models using the input gradient," 2016, *arXiv:1611.07634*. [Online]. Available: http://arxiv.org/abs/1611.07634

[9] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," 2020, *arXiv:2003.10555*. [Online]. Available: https://arxiv.org/abs/2003.10555

[10] J. Li, J. Yu, Z. Ye, S. Wong, M. W. Mak, B. Mak, X. Liu, and H. Meng, "A comparative study of acoustic and linguistic features classification for Alzheimer's disease detection," in *Proc. IEEE Int. Conf. Acoust, Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6423–6427.

[11] B. H. Davis and M. Maclagan, "Examining pauses in Alzheimer's discourse," *Am. J. Alzheimers Dis. Other Dement.*, vol. 24, no. 2, pp. 141–154, Apr. 2009.

[12] M. Komeili, C. Pou-Prom, D. Liaqat, K. C. Fraser, M. Yancheva, and F. Rudzicz, "Talk2me: Automated linguistic data collection for personal assessment," *PLoS One*, vol. 14, no. 3, p. e0212342, Mar. 2019.

[13] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP — a collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust, Speech Signal Process. (ICASSP)*, May 2014, pp. 960–964.

[14] M. Rohanian, J. Hough, and M. Purver, "Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech," 2016, *arXiv:2106.09668*. [Online]. Available: https://arxiv.org/abs/2106.09668

[15] Z. S. Syed, S. A. Memon, M. S. Shah, and A. S. Syed, "Introducing the Urdu-Sindhi speech emotion corpus: A novel dataset of speech recordings for emotion recognition for two low-resource languages," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 4, pp. 1–6, 2020.

[16] Z. S. Syed, K. Sidorov, and D. Marshall, "Automated screening for bipolar disorder from audio/visual modalities," in *Proc. Audio/Vis. Emotion Challenge Workshop*, Oct. 2018.

[17] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. Interspeech*, Sep. 2010, pp. 2794–2797.

[18] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - the Munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. on ACM Multimedia*, 2010, pp. 1459–1462.

[19] N. Wang, Y. Cao, S. Hao, Z. Shao, and K. Subbalakshmi, "Modular multi-modal attention network for Alzheimer' disease detection using patient audio and language data," in *Proc. Interspeech*, Aug. 2021, pp. 3835–3839.

[20] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr. 2016.

[21] Y. Li, C.-Y. Chen, and W. W. Wasserman, "Deep feature selection: Theory and application to identify enhancers and promoters," *J. Comput. Biol.*, vol. 23, no. 5, pp. 322–336, May 2016.

[22] C.-H. Chang, L. Rampasek, and A. Goldenberg, "Dropout feature ranking for deep learning models," 2017, *arXiv:1712.08645*. [Online]. Available: https://arxiv.org/abs/1712.08645

[23] X. Ke, M. W. Mak, J. Li, and H. M. Meng, "Dual dropout ranking of linguistic features for Alzheimer's disease recognition," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2021, pp. 743–749.

[24] M. Wojtas and K. Chen, "Feature importance ranking for deep learning," in *Proc. Advances Neural Inf. Process. Syst. (NIPS)*, Oct. 2020, pp. 5105–5114.

[25] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.