Cognitive Robotics 2 (2022) 21–29

Contents lists available at ScienceDirect

Cognitive Robotics

journal homepage: http://www.keaipublishing.com/en/journals/cognitive-robotics/

Machine learning model for discrimination of mild dementia patients using acoustic features

Kazu Nishikawa*, Kuwahara Akihiro, Rin Hirakawa, Hideaki Kawano, Yoshihisa Nakatoh*

Kyushu Institute of Technology: Kyushu Kogyo Daigaku, 1 1-1 Sensuicho, Tobata Ward, Kitakyushu, Fukuoka, Japan

ARTICLE INFO

Keywords:
Dementia
Alzheimer's disease
Acoustic feature
Machine learning
Transformer

ABSTRACT

In previous research on dementia discrimination by voice, a method using multiple acoustic features by machine learning has been proposed. However, they do not focus on speech analysis in mild dementia patients (MCI). Therefore, we propose a dementia discrimination system based on the analysis of vowel utterance features. The analysis results indicated that some cases of dementia appeared in the voice of mild dementia patients. These results can also be used as an index for future improvement of speech sounds in dementia. Taking advantage of these results, we propose an ensemble discrimination system using a classifier with statistical acoustic features and a Neural Network of transformer models, and the F-score is 0.907, which is better than the state-of-the-art methods.

1. Introduction

According to the WHO report, dementia cases worldwide will increase by to 130 million by 2050 [1]. Furthermore, global health care costs for dementia are estimated to exceed $2 trillion by 2030 [1]. Dementia changes trigger a decline in thinking skills, also known as cognitive abilities, severe enough to impair daily life and independent function [2–5]. They also affect behaviour, feelings, and relationships. Therefore, dementia is a significant problem for modern healthcare [6].

However, no treatment for dementia has been established [7]. Treatment of dementia is more effective when started early and maintaining good health for a long time [8]. Thus, early identification of the cause is essential. The questionnaire method is generally used for present dementia screening. In particular, the Mini-Mental State Examination (MMSE) is the most used test method globally [9]. During the MMSE, a health professional asks a patient a series of questions designed to test a range of everyday mental skills. The maximum MMSE score is 30 points. A score of 27–30 suggests Normal Controls (NC), 22–26 suggests mild dementia (MCI), and less than 21 indicates severe dementia (AD) [10]. However, this method requires 10–15 min with a doctor and a clinical psychologist, which puts a burden on hospitals and test subjects. As a method for discriminating dementia from speech at present, a method for extracting speech features and discriminating using machine learning is reported. Other works, such as Random Forest [11] and 1DCNN-LSTM [12] are proposed.

In this way, discriminative methods using statistically compressed multiple acoustic features such as LLD have been the mainstream of research, but few studies have focused on speech analysis of older people with dementia tendencies.

MCI, Mild dementia patients; NC, Normal controls.

* Corresponding author.

E-mail addresses: nishikawa.kazu615@mail.kyutech.jp (K. Nishikawa), kuwahara.akihiro135@mail.kyutech.jp (K. Akihiro), hirakawa.rin151@mail.kyutech.jp (R. Hirakawa), kawano@ecs.kyutech.ac.jp (H. Kawano), nakatoh@ecs.kyutech.ac.jp (Y. Nakatoh).

https://doi.org/10.1016/j.cogr.2021.12.003
Received 17 December 2021; Accepted 20 December 2021
Available online 27 December 2021

It can also help people with dementia improve voice quality by analysing their speech. Thus, we propose a dementia discrimination system that clarifies the speech characteristics of patients with mild dementia (MCI) that have not yet been analysed and incorporates them.

2. Related works

Recently, there has been much research on the discrimination of dementia using a machine learning model.

2.1. Discrimination of dementia using machine learning

There have been many studies on the discrimination of dementia speech by machine learning in recent years. Higuchi et al. [13] divided 45 subjects into MCI and NC groups based on MMSE scores, analysed 6552 features from openSMILE's library [14], performed principal component analysis, and created a discriminator with reduced features, and achieved a discrimination rate of 78% from logistic regression. Xue et al. conducted 5-segment cross-validation by training 1264 pieces of conversational speech data for an average of 73 min on a CNN model as 1-dimensional raw data. As a result, an F- value of 0.742 ± 0.033 was achieved [15]. Xue et al. also reported that 5 division cross-validation with the LSTM model using similar speech data resulted in a F-value of 0.596 ± 0.047.

2.2. Speech analysis of people with dementia

Meghanani et al. constructed pBLSTM-CNN and ResNet-LSTM models and reported that MFCC and log-Mel spectrogram, which are acoustic features of vocal tract characteristics, and has root means square errors (RMSE) with 5.9 [16]. The results beat the baseline accuracy (62.5%) and RMSE (6.14) reported for acoustic features on the ADReSS challenge dataset. It suggests that log-Mel spectrograms and MFCCs are valuable features for the AD recognition problem. Kurokawa et al. collected speech data by recording the reading speech of older people aged 60 years or older (approximately 9.2 h) and collecting HDS-R answers and analyzed acoustic characteristics by *t*-test. As a result, it was confirmed that there were significant differences between NC and MCI in speaking speed, silence time, and silence insertion number [17].

3. Analysis of MCI's acoustic features

In Sections 1 and 2, we discussed the importance of analyzing dementia speech. Therefore, the purpose of this section is to clarify the speech characteristics of MCI that have not yet been analyzed. In this study, We analyzed the acoustic features of MCI and NC in phoneme units, which are the smallest units of speech and, analyzed vowel sounds (/ a /, /i /, /u /, /e /, /o /) among phonemes in this work.

3.1. Acoustic features

Phonology consists of articulatory phonetics, acoustic phonetics, and auditory phonetics. We analyzed the acoustic features of NC and MCI from three academic viewpoints.

***MFCC*:** MFCC is an acoustic analysis method based on auditory filters and is mainly used in speech recognition. It uses the logarithmic power of the fractionated filter bank to represent the power spectrum in a small order efficiently. Similarly, the filter of the trigonometric function arranged at equal intervals along the Mel scale is applied to examine the low-frequency part finely and the high-frequency part roughly by the characteristics of human hearing. The number of trigonometric functions indicates the number of channels in the filter bank [18,19]. The MFCC of 12 dimensions was calculated and analyzed for each dimension in this work.

***F0*:** We use the vocal apparatus from the throat to the lips to make a voice. Especially for the vocal cords, if the tension of the vocal cords is enormous and the air pressure from the lungs is high, the opening and closing cycle of the vocal cords, that is, the oscillation cycle is shortened, and the pitch of the sound source is raised. This oscillation period is called the fundamental period, and the reciprocal is called F0. It corresponds to the pitch of the voice [20]. We performed F0 analysis by autocorrelation method.

***Formant frequency*:** Formant frequencies are the multiple peaks found on the spectral envelope of a speech waveform, which can be used to analyze the vocal tract information of the speech. In this work, the peak points of the spectrum obtained from LPC analysis are defined as formant frequencies, and only the second-order is treated as F1 and F2, starting from the lowest order [21]. Regarding formant frequencies, in speech engineering, it is generally said that F1 is correlated with the mouth opening and F2 with the tongue position.

***Jitter*:** Jitter is a numerical representation of pitch fluctuations in acoustics. The speech waveform is expressed by a combination of sine curves of various frequencies and amplitudes, and it has a roughly periodic waveform. When this repetition is made to be one cycle, the cycle may be disturbed. The numerical result of this turbulence is Jitter. In this study, localJitter and PPQ5 were analyzed. Next, we will describe a specific calculation method. localJitter is the absolute difference between the periods of adjacent cycles [22]. The following formula calculates it.:

$$local\,Jitter(second) = \sum^{i} = 2N\,\frac{|T_i - T_{i-1}|}{(N-1)} \tag{1}$$

Table 1
Dataset details.

		Number ofPeople	Age	MMSE
ALL	**Ave**	80	65.2	28.4
	Std	–	16.1	1.64
NC	**Ave**	30	48.8	28.9
	Std	–	15.8	1.64
NC(70 s)	**Ave**	37	75.2	29.1
	Std	–	3.08	0.69
MCI(70 s)	**Ave**	13	74.3	25.9
	Std	–	2.81	0.99

Table 2
Vowel data numbers.

	/a/	/i/	/u/	/e/	/o/
NC (70 s)	45	48	26	38	42
MCI (70 s)	43	44	30	43	47
Total	88	92	56	81	89

T_i is the duration of the ith cycle, and N is the number of cycles. PPQ5 is the periodic fluctuation for five consecutive cycles. First, absPPQ5 is defined [23]. T_i is the duration of the ith cycle, and N is the number of cycles.

$$absPPQ5(seconds) = \sum_{i=3}^{N-2} \frac{\left| T_i - \frac{(T_{i-2}+T_{i-1}+T_i+T_{i+1}+T_{i+2})}{5} \right|}{(N-4)} \tag{2}$$

The average cycle is then defined as follows:

$$meanPeriod(second) = \sum_{i=1}^{N} \frac{T_i}{N} \tag{3}$$

Finally, the *PPQ5* is calculated by the following equation.

$$PPQ5 = \frac{absPPQ5(second)}{meanPeriod(second)} \tag{4}$$

Shimmer: Shimmer is a numerical representation of the fluctuations in loudness, specifically the differences in amplitude between adjacent periods. A similar principle is obtained by substituting the amplitude for the period of Jitter [24].

HNR: *HNR* is the ratio of harmonic component and noise component of the speech waveform. It is widely used as an indicator of the degree of hoarseness. When the harmonic component of the speech waveform is A%, and the noise component is B%, the *HNR* is obtained by the following equation [25].

$$HNR\ = 10log_{10}\left(\frac{A}{B}\right) \tag{5}$$

3.2. Experiment method

Dataset: We utilized the elderly corpus with the control group [26]. It involves three tasks for participants and consists of speech, the transcribed text, and MMSE scores. Based on the MMSE results, participants with a score of 23 or more and 27 or less are defined as MCI, and the rest of the elderly is defined as Normal Controls (NC). Table 1 show more detail for the dataset.

Preprocess: We analyzed the voice data of 16 people from the dataset. The 16 patients consisted of 6 MCI in their 70 s (3 male and three female) and 10 NC (5 male and five female). As mentioned above, vowels are analyzed in this research. Therefore, phoneme segmentation processing was performed using a speech analysis tool, Audacity [27], while visually confirming the waveform. As a result, the number of data shown in Table 2 was obtained for each vowel. We did not use values outside the quartile range or the normal range.

Evaluation: We conducted a *t*-test to investigate the significance of each feature. To visualize the formant difference between MCI and NC, we plotted the formant area of the five vowels on a two-dimensional coordinate system, with F1 on the horizontal axis and F2 on the vertical axis, for each speaker. In general, the smaller the formant area, the lower the intelligibility of the voice [28].

3.3. Experiment result and consideration

Result of the t-test: Table 3 shows only the feature values with significant differences in the *t*-test of MFCCs, and Fig. 1 shows the power of the MFCC by dimension. MFCCs mainly were significantly different for higher orders (7d-12d). MFCCs of higher orders are generally used for emotion estimation. Since emotional flattening has been reported as a symptom of dementia [29], it is possible that this feature appeared (Table 4). Table 4 shows only the feature values with significant differences in the *t*-test. F0 was the feature quantity of the voice height, and MCI tended to be lower to confirm the significance. The tension of the laryngeal muscles

Table 3
Evaluation results of each feature quantity.

		NC		MCI		
Featurequantity	vowel	Ave	std	ave	std	*t*-testvalue
F0 mean	/i/	166.9	12.68	154.3	12.17	2.783**
	/e/	153.4	18.88	138.2	17.61	2.468*
F0 std	/u/	3.222	1.440	1.832	0.417	3.075**
localJitter	/o/	0.020	0.005	0.027	0.008	−2.299*
PPQ5	/o/	0.007	0.002	0.013	0.005	−3.442**
Jitter						
Local	/a/	0.122	0.031	0.095	0.016	2.604**
Shimmer	/u/	0.140	0.043	0.077	0.013	4.651***
APQ5	/a/	0.059	0.017	0.042	0.003	2.612**
Shimmer	/o/	0.075	0.018	0.048	0.005	3.969**
HNR	/a/	5.282	1.063	7.942	2.094	−4.567***
	/u/	12.57	1.551	14.50	1.551	−2.910**
	/o/	9.705	1.525	11.05	1.916	−2.137*

* $\partial<0.05$,.
** $\partial<0.01$,.
*** $\partial<0.001$.

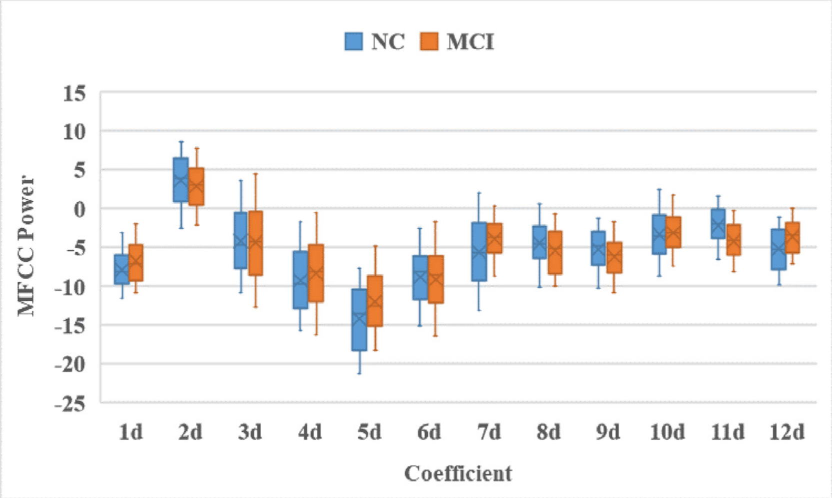

Fig. 1. MFCC power of NC and MCI.

generally produces a high-pitched voice [30]. People with a tendency to dementia may have weakened it. MCI tended to be higher than Jitter when a significant difference was confirmed in Jitter. People who tend to have dementia may have more disturbed voice cycles. It may be due to articulatory dysfunction, a symptom of dementia [31]. Shimmer found that NC tended to have higher values for significant differences. People with a tendency to dementia may have a minor disturbance of voice amplitude. This result needs further investigation, considering the relationship with voice volume. The HNR values tended to be higher in the MCI group than in the MCI group when a significant difference was confirmed. People with a tendency to dementia may have a loud, hoarse voice. In this sample, there was no difference in the effectiveness of the feature quantity by the type of vowel.

*Results of the formant analysis***:** We calculated the mean formants of 5 male and five female NCs and the formants area of 3 subjects. The results are shown in Figs. 2–4. For subject S32 in Fig. 2. We can see that the formant area is smaller than that of NCs. In particular, the F1 value of /a/ is low, indicating that S32 has a low tongue position when he utters /a/. Subject S37 in Fig. 3, it can be confirmed that the formant area is much smaller than NCs, especially the change in the value of F1 is small, and it can be said that the change in the position of the tongue tends to be poor when vocalizing. For subject S44 in Fig. 4, the formant area is not significantly different from NCs, but the formant frequency values for each vowel are small, indicating that the tongue position is generally lower. From these results, it was found that MCI tended to lower the position of tongue movement overall compared with NC and that there were individual differences in the degree of voice articulation.

4. Machine learning models to detect dementia

In the previous sections, proposals have been made using raw voice data and LLD as inputs. Meanwhile it uses dimensionally reduced statistics as inputs, but there is a problem that input data becomes large. In addition, it is questionable whether this method

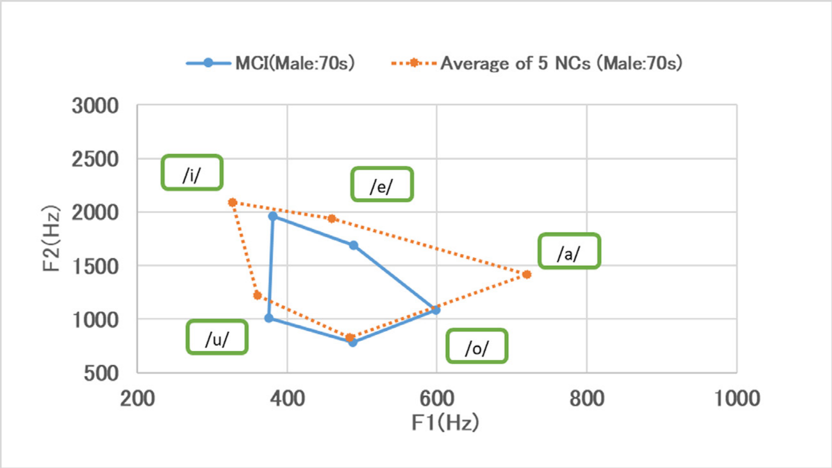

Fig. 2. The formant area of MCI (S32).

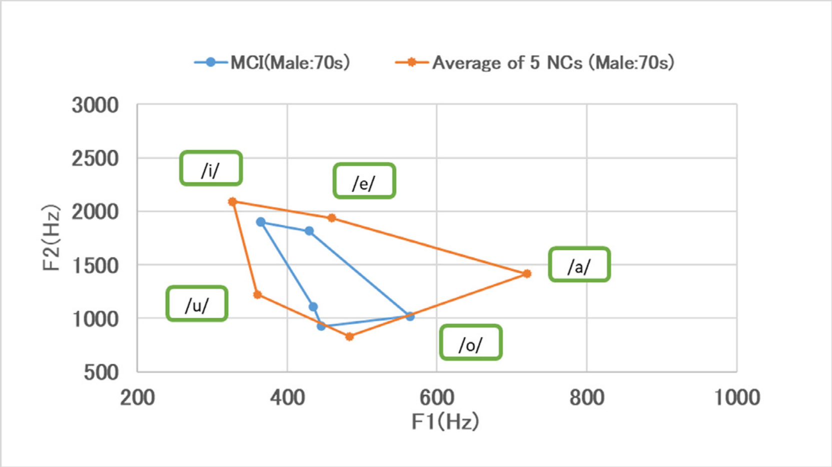

Fig. 3. The formant area of MCI (S37).

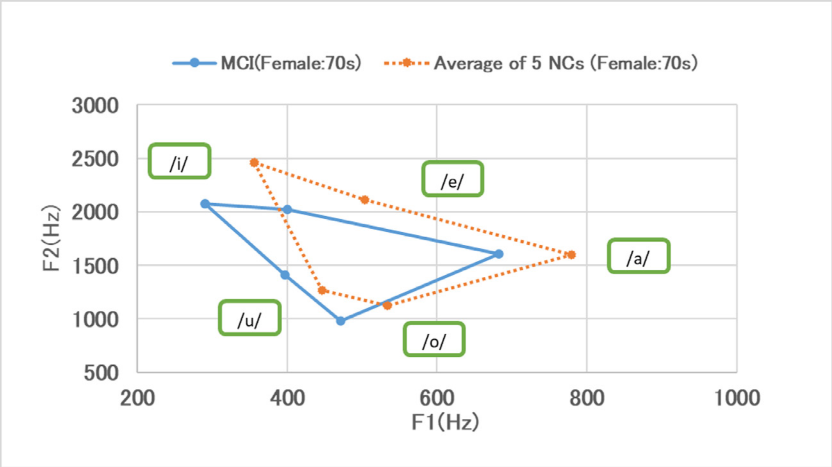

Fig. 4. The formant area of MCI (S44).

Table 4
Evaluation results of MFCC.

	NC		MCI		t-
Featurequantity	ave	std	ave	std	testvalue
MFCC_1d	−7.961	2.264	−6.841	2.692	−3.282**
MFCC_5d	−14.23	4.281	−12.01	3.832	−3.987***
MFCC_7d	−5.655	4.233	−3.971	2.435	−3.562***
MFCC_8d	−4.521	2.862	−5.442	2.851	−2.345*
MFCC_9d	−5.295	2.599	−6.216	2.441	−2.663**
MFCC_11d	−2.226	2.346	−4.222	2.270	−6.306***
MFCC_12d	−5.227	2.593	−3.685	2.083	−4.785***

* p<0.05,.
** p<0.01,.
*** p<0.001.

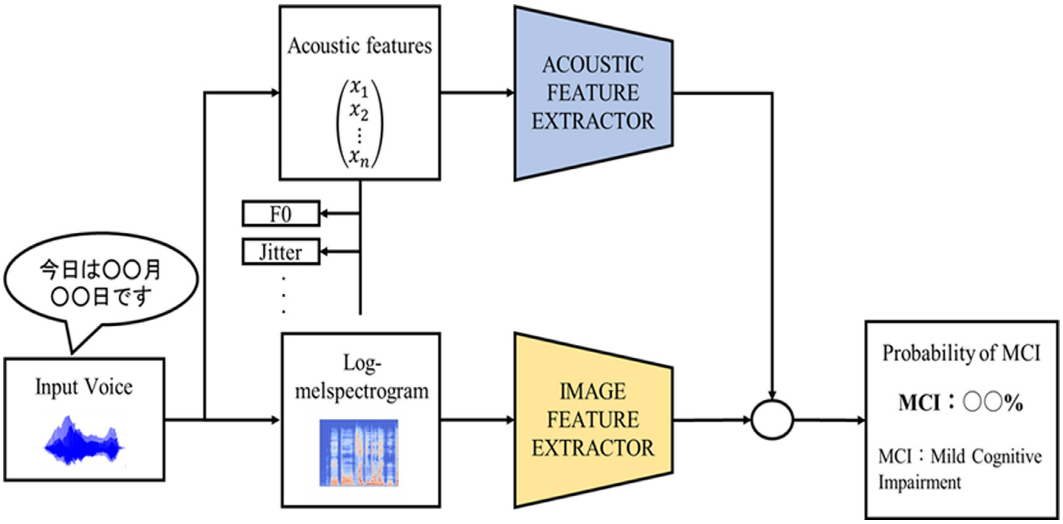

Fig. 5. An overview of the proposed method.

is effective for the dementia discrimination model because it only applies the machine learning technique, which has been used widely. We propose a dementia discrimination model based on the results of Section 3.

4.1. Proposed machine learning models

We propose an ensemble dementia discrimination model that combines acoustic features, for which significant differences were confirmed in the previous chapter, and image processing, which has seen remarkable development in recent years. An overview of the proposed discrimination models is shown in Fig. 5. Initially, the input voice is a verbal response to a commonly used dementia test. We extract the acoustic features of 192 vectors based on the input speech. Furthermore, since significant differences were found in the higher dimensions of MFCCs in the previous chapter, the waveform images of log-Mel spectrograms were output before being converted to MFCCs for the further abstraction of the features. The Acoustic feature extractor uses various machine learning models to discriminate. The output is the weight of the input prediction. The image feature extractor uses a deep learning model specialized for image processing to perform discrimination. SoftMax is used as the output layer. Finally, these outputs are weighted to produce the dementia probability. In this study, we obtained the probability of dementia by using the weight of the Acoustic feature extractor as 30% and the weight of the image feature extractor as 70%. Finally, the threshold for evaluation by machine learning is set at 50%.

Acoustic features: We built an extractor and calculated the statistics of the acoustic features for which a significant difference by t-test was confirmed in the previous experiment. The extractor takes an audio frame every 20 ms and shifts it every 10 ms, and 192 feature vectors per audio data. The list of features and statistics is shown in Tables 5 and 6.

Acoustic feature extractor: In this extraction part, a machine learning model using one-dimensional statistics as an input is used. In this study, models of SVM (linear), Random Forest, and lightGBM were constructed for validation. These were made with skit-learn of python module [32]. SVM is a method of constructing a two-class pattern discriminator using linear input elements. The parameters of the linear input elements are learned from the training samples with the criterion of finding the margin-maximizing hyperplane that maximizes the distance to each data point [33]. Random Forest is an ensemble learning algorithm using a decision tree as a weak learner and is used for classification, regression, and clustering [34].

Table 5
Evaluation results of each feature quantity.

Acoustic features	Description
F0	Fundamental frequency computed from the cepstrum
Jitter	Temporal fluctuation of speech waveform
Shimmer	Fluctuation in the amplitude of a speech waveform
HNR	Signal to noise ratio of the audio signal
MFCC	The spectral outline that reflects human auditory characteristics
(a first-order delta coefficient of the above features)	

Table 6
Function details.

Functional	Maximum, Minimum, Average, stddev, skewness, kurtosis

Table 7
Parameters for these image feature extractors.

Items	VGG16 and VGG19	ViT_16
Image Size	224	160
Weights	ImageNet	–
Epoch	120	120
Active function	Relu	Sigmoid
Learning Rate	e^{-4}	e^{-4}
Gradient descent	Adam	Adam

Log-mel spectrogram: The MFCCs, which showed significant differences in the previous section, emphasized vocal tract characteristics by performing discrete cosine transform. For image processing, it is necessary to grasp local features and general features. Therefore, we extracted an image of the logarithmic melt spectrum that could be obtained before this processing.

Image feature extractor: In this extraction section, this research was validated with several deep learning models that are widely used in image processing. The machine learning models are VGG16, VGG19 and ViT_b16 [35,36]. VGG16 is a convolutional neural network consisting of 16 layers, 13 convolutional layers, and three fully-connected layers proposed in ILSVRC (ImageNet Large Scale Visual Recognition Challenge) in 2014. VGG19 is a model of VGG16 with three additional convolutional layers in the middle layer. ViT_b16 is a Transformer model that is optimized for image processing. According to this method, the computational cost of the Transformer is about 1/4 to 1/5 of that of conventional CNN models, and the computational time has been improved.

4.2. Experiment method

Dataset: We experimented with the elderly corpus with the control group used in Chapter 3. Details can be found in the dataset in Section 3.2.

Preprocess: We analyzed the voice data of 16 people from the dataset. The 16 patients consisted of 6 MCI in their 70 s (3 male and three female) and 10 NC (5 male and five female). We divided each of these data into syllables of 3 s, and the missing segments were padded with silence. As a result, 2970 NC and 2970 MCI data were obtained. The statistics of acoustic features described in the previous section were calculated for these data, and the heat map of the log-mel spectrogram was plotted. The statistics of acoustic features used the statistics in Table 7 obtained. The log-mel spectrum is calculated using the python module librosa [37].

Evaluation: We evaluated the proposed system using various classifiers. We calculated the accuracy, F-score, and AUC from the Confusion Matrix, a metric commonly used in machine learning evaluation. We trained three classifiers, SVM (linear), Random Forest, and lightGBM, in terms of statistics to compare our results. These were made with skit-learn of the python module [38]. For the log-mel spectrogram, we trained on three networks: ViT_b16 with roughly the same number of parameters, fine-tuning of VGG16 and VGG19. The parameters for these image feature extractors are shown in Table 7. VGG 16 and VGG 19 were used to freeze the intermediate layer, add a dense layer to the output layer, and adjust the parameters. In VGG 16 and VGG 19, the intermediate layer was frozen, the high-density layer was added to the output layer, and the parameter was adjusted.

4.3. Experiment result and consideration

We evaluated each classifier experiment by input. The experimental results are shown in Table 8. In the case of acoustic features, the SVM, which is a linear model, did not provide good accuracy. It indicates that a simple linear model is difficult to discriminate in this sample. The F-score of lightGBM was 0.844, higher than that of other classifiers. Then, we calculated the essential features of lightGBM and found that most of the ΔMFCCs were at the top of the list. It may be due to the large fluctuations in the vocal tract component of dementia speech. When the input was a Log-mel Spectrogram, ViT_b16 had the best F value. In this experimental result,

Table 8
Evaluation results of each feature quantity.

Input	Classifier	Accuracy	F-score	Param
Acoustic features (192 feature vectors)	**SVM(Linear)**	0.657	0.657	–
	Random Forest	0.723	0.720	
	lightGBM	0.845	0.844	–
Log-mel Spectrogram	**VGG16 (fineturning)**	0.812	0.807	91M
	VGG19 (fineturning)	0.877	0.876	97M
	ViT_b16	0.894	0.893	86M

Table 9
Evaluation results of ensemble models.

Input(Acoustic feature + Image feature)	Accuracy	F-score
SVM(Linear)+VGG16(fine turning)	0.823	0.818
SVM(Linear)+VGG19(fine turning)	0.877	0.877
SVM(Linear)+ViT_b16	0.904	0.904
Random Forest+ VGG16(fine turning)	0.829	0.826
Random Forest+ VGG19(fine turning)	0.872	0.872
Random Forest+ ViT_b16	0.895	0.895
lightGBM+VGG16(fine turning)	0.860	0.858
lightGBM+VGG19(fine turning)	0.889	0.889
lightGBM+ViT_b16	**0.907**	**0.907**

the Transformer model ViT_b16 performed better than the conventional CNN models VGG16 and VGG19. From the results of VGG16 and VGG19, it was confirmed that the discrimination rate was improved by deepening the middle layer.

At last, we performed ensemble learning between each classifier. The results are shown in Table 9. As a result, all the ensemble models except Random Forest + VGG 19 (fine turning) improved the discrimination rate compared with the case where discrimination was performed individually. It also achieved the higher Accuracy and F-score at lightGBM + ViT_b16. It seems to complement features that could not be captured only by deep learning using log-mel spectrum as an input. Besides, the parameters of the deep learning model are so large and do not change for processing time.

5. Conclusion

In this paper, we proposed the ensemble dementia discrimination system based on the analysis of MCI speech features. The analysis results proved that MCI voices represent several cases of dementia, and the proposed system achieved the F-score of about 90% even with several thousand voices. These results will provide an opportunity to improve the voice quality of dementia patients. In the further research, we will increase the number of voice data and study the construction of a model that can reduce the parameters of the Neural Network.

Statement

We declare that the work "Machine Learning Model for Discrimination of Mild Dementia Patients using Acoustic Features" is entirely our own and no part of it has been quoted from other researchers (Fig. 5).

Declaration of Competing Interest

We declare that we have no professional or other personal interests of any nature or kind with any products, services, or companies that could be construed to influence the position or peer review presented in the manuscript entitled "Machine Learning Model for Discrimination of Mild Dementia Patients using Acoustic Features".

CRediT authorship contribution statement

Kazu Nishikawa: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Visualization. **Kuwahara Akihiro:** Conceptualization, Methodology. **Rin Hirakawa:** Resources, Writing – review & editing. **Hideaki Kawano:** Resources, Writing – review & editing. **Yoshihisa Nakatoh:** Writing – review & editing, Supervision, Project administration.

References

[1] M. Prince, A. Wimo, M. Guerchet, G. Ali, Y. Wu, M. PrinaWorld Alzheimer Report, in: The Global Impact of Dementia, OECD, 2015, pp. 25–70.

[2] E. Rundqvist, E. Severinsson, Caring relationships with patients suffering from dementia — an interview study, J. Adv. Nurs. 29 (4) (2001) 800–807, doi:10.1046/j.1365-2648.1999.00955.x.

[3] C. Qin, B. Winblad, L. Fratiglioni, The age-dependent relation of blood pressure to cognitive function and dementia, Lancet Neurol. 4 (8) (2005) 487–499, doi:10.1016/S1474-4422(05)70141-1.

[4] V. Crooks, J. Lubben, D. Petitti, D. Little, V. Chiu, Social network, cognitive function, and dementia incidence among elderly women, Am. J. Public Health 98 (7) (2008) 1221–1227, doi:10.2105/AJPH.2007.115923.

[5] R. Mahurin, B. DeBettignies, F. Pirozzolo, Structured assessment of independent living skills: preliminary report of a performance measure of functional abilities in dementia, J. Gerontol. 46 (2) (1991) 58–66, doi:10.1093/geronj/46.2.P58.

[6] O. Indicators, in: Health at a Glance, OECD, 2019, pp. 1–249, doi:10.1787/19991312.

[7] D. strooper, E. Karran, The cellular phase of Alzheimer's disease, Cell 164 (2016) 603–615, doi:10.1016/j.cell.2015.12.056.

[8] P. Panegyres, R. Berry, J. Burchell, Early dementia screening, Diagnostics 6 (1) (2016) 1–13, doi:10.3390/diagnostics6010006.

[9] T. Tombaugh, N. Mclntyre, The mini-mental state examination: a comprehensive review, J. Am. Geriatr. Soc. 40 (9) (1992) 922–935, doi:10.1111/j.1532-5415.1992.tb01992.x.

[10] M. Folstein, S. Folstein, P. McHugh, Mini-mental state". A practical method for grading the cognitive state of patients for the clinician, J. Psychiatr. Res. 12 (3) (1975) 189–198, doi:10.1016/0022-3956(75)90026-6.

[11] K. Nishikawa, R. Hirakawa, H. Kawano, K. Nakashi, Y. Nakatoh, Effective speech features for distinguishing mild dementia patients from healthy person, in: Proceedings of the International Conference on Human Interaction and Emerging Technologies, 1253, 2021, pp. 356–361, doi:10.1007/978-3-030-55307-4_54.

[12] K. Nishikawa, R. Hirakawa, H. Kawano, K. Nakashi, Y. Nakatoh, Detecting system Alzheimer's dementia by 1d CNN-LSTM in Japanese speech, in: Proceedings of the IEEE International Conference on Consumer Electronics, 2021, pp. 1–3, doi:10.1109/ICCE50685.2021.9427692.

[13] M. Higuchi, M. Nakamura, T. Okazaki, J. Takemura, T. Takano, Y. Omiya, … S. Tokuno, Detection of mild cognitive impairment through voice analysis, Trans. Jpn. Soc. Med. Biol. Eng. 59 (2021) 495, doi:10.11239/jsmbe.Annual59.495.

[14] openSMILE: https://www.audeering.com/research/opensmile/, (accessed Nov. 27, 2021)

[15] C. Xue, C. Karjadi, I. Paschalidis, R. Au, V. Kolachalama, Detection of dementia on raw voice recordings using deep learning: a Framingham Heart Study, Alzheimers Res Ther 13 (146) (2021), doi:10.1101/2021.03.04.21252582.

[16] A. Meghanani, C. Anoop, A. Ramakrishnan, An exploration of log-Mel spectrogram and MFCC features for Alzheimer's dementia recognition from spontaneous speech, in: Proceedings of the IEEE Spoken Language Technology Workshop (SLT), 2021, doi:10.1109/SLT48900.2021.9383491.

[17] Y. Kurokawa, Y. Iribe, Analysis of Dementia Trends in the Elderly Using Acoustic Features, Aichi Prefectural University, 2016 Summary of Graduation Thesis.

[18] P. Mermelstein, Distance measures for speech recognition, psychological and instrumental, Pattern Recognit. Artif. Intell. 92 (3) (1976) 374–388.

[19] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust. 28 (4) (1980) 357–366, doi:10.1109/TASSP.1980.1163420.

[20] S. Furui, Acoustics and Speech Engineering, Kindai Kagaku-sha, 1992.

[21] H. Itahashi, M. Akahane, S. Ishikawa, T. Ookouchi, H. Kasutani, N. Kuwahara, … T. Watanabe, in: Voice Engineering, Morikita Publishing Co., 2008, pp. 6–12.

[22] Voice 2. Jitter: https://www.fon.hum.uva.nl/praat/manual/Voice_2__Jitter.html, (accessed Nov. 27, 2021)

[23] PointProcess: Get jitter (local)...:https://www.fon.hum.uva.nl/praat/manual/PointProcess__Get_jitter__local___.html, (accessed Nov. 27, 2021)

[24] Voice3. Shimmer: https://www.fon.hum.uva.nl/praat/manual/Voice_3__Shimmer.html, (accessed Nov. 27, 2021)

[25] Harmonicity: https://www.fon.hum.uva.nl/praat/manual/Harmonicity.html, (accessed Nov. 27, 2021)

[26] D. Shibata, K. Ito, S. Wakamiya, E. Aramaki, Detecting early-stage dementia based on natural language processing, Trans. Jpn. Soc. Artif. Intell. 34 (4) (2019) 1–9, doi:10.1527/tjsai.B-J11.

[27] Audacity: https://www.audacityteam.org/, (accessed Nov. 27, 2021)

[28] Y. Masuko, A study on the formants and the tone quality of Japanese vowels, Tokyo Univ. Foreign Stud. J. 82 (2011) 105–121.

[29] H. Takechi, H. Yamada, Y. Sugihara, T. Kita, Behavioral and psychological symptoms, cognitive impairment and caregiver burden related to Alzheimer's disease patients treated in an outpatient memory clinic, Geriatr. Gerontol. Int. 43 (2) (2006) 207–216, doi:10.3143/geriatrics.43.207.

[30] K. Honda, Biological mechanisms for tuning voice fundamental frequency, Jpn. Laryngol. Assoc. 8 (2) (1996) 109–115, doi:10.5426/larynx1989.8.2_109.

[31] M. Ikeda, Communication difficulties in patients with neurodegenerative dementia, Jpn. Soc. High. Brain Dysfunct. 35 (3) (2015) 292–296, doi:10.2496/hbfr.35.292.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, … É. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830, doi:10.5555/1953048.2078195.

[33] V. Vapnik, A. Lerner, Pattern recognition using generalized portrait method, Autom. Remote Control 24 (1963) 774–780.

[34] B. Leo, Random forests, Mach. Learn. 45 (1) (2001) 5–32, doi:10.1023/A:1010933404324.

[35] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the International Conference on Learning Representations, 2015, pp. 1–14.

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, ….N. Houlsby, An image is worth 16×16 words: transformers for image recognition at scale, in: Proceedings of the ICLR, 2021, pp. 1–24.

[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16×16 words: transformers for image recognition at scale, in: Proceedings of the ICLR Conference Blind Submission, 2021, pp. 1–22.

[38] M. Brian, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: audio and music signal analysis in python, in: Proceedings of the 14th python in science conference, 2015, pp. 18–25, doi:10.5281/zenodo.4792298.