Differentiation of speech in Parkinson's disease and spinocerebellar degeneration using deep neural networks

Katsuki Eguchi[a,b,*], Hiroaki Yaguchi[b], Ikue Kudo[a], Ibuki Kimura[a], Tomoko Nabekura[a], Ryuto Kumagai[c], Kenichi Fujita[a], Yuichi Nakashiro[a], Yuki Iida[a], Shinsuke Hamada[a], Sanae Honma[a], Asako Takei[a], Fumio Moriwaka[a], Ichiro Yabe[b]

[a]Hokuyukai Neurological Hospital, 4-30, 2jo, 2cho-me, Nijuyonken, Nishiku, Sapporo 063-0802, Japan

[b]Department of Neurology, Faculty of Medicine and Graduate School of Medicine, Hokkaido University, Kita 15, Nishi 7, Kita-ku, Sapporo 060-8638, Japan

[c]Sapporo Parkinson MS Neurological Clinic, Sapporo Kita sky building F12, 7-6, Kita 7-Nishi 5, Kita-ku, Sapporo, Hokkaido, 060-0807, Japan

***Corresponding author**: Katsuki Eguchi

Hokuyukai Neurological Hospital, 4-30, 2jo, 2cho-me, Nijuyonken, Nishiku, Sapporo 063-0802, Japan

Telephone: 81-11-631-1163

E-mail: k198762@gmail.com

Statements and Declarations

Competing Interests and Funding: The authors have no competing interests to declare that are relevant to the content of this article. This work was supported by JSPS KAKENHI Grant Number JP22K20843.

Acknowledgments: We thank the speech-language pathologists at Hokuyukai Neurological Hospital for providing speech-language therapy and recording the patients' speech.

Authors' contribution

Study concept and design: K. Eguchi and H. Yaguchi.

Examination and treatment of PD and SCD: K. Eguchi, Y. Iida, S. Hamada, S. Honma, A. Takei, and F. Moriwaka.

Predicting PD and SCD from recorded speech data: I. Kudo, I. Kimura, T. Nabekura, R. Kumagai, and K. Fujita.

Provided Speech-Language therapy: Y. Nakashiro.

Drafting of the manuscript: K. Eguchi.

Critical revision of the manuscript for important intellectual content: H. Yaguchi and I. Yabe.

All authors have read and approved the final article.

ABSTRACT

Introduction: Assessing dysarthria features in patients with neurodegenerative diseases helps diagnose underlying pathologies. Although deep neural network (DNN) techniques have been widely adopted in various audio processing tasks, few studies have tested whether DNNs can help differentiate neurodegenerative diseases using patients' speech data. This study evaluated whether a DNN model using a transformer architecture could differentiate patients with Parkinson's disease (PD) from patients with spinocerebellar degeneration (SCD) using speech data.

Methods: Speech data were obtained from 251 and 101 patients with PD and SCD, respectively, while they read a passage. We fine-tuned a pre-trained DNN model using log-mel spectrograms generated from speech data. The DNN model was trained to predict whether the input spectrogram was generated from patients with PD or SCD. We used 5-fold cross-validation to evaluate the predictive performance using the area under the receiver operating characteristic curve (AUC) and accuracy, sensitivity, and specificity.

Results: The average ± standard deviation of the AUC, accuracy, sensitivity, and specificity of the trained model for the 5-fold cross-validation were 0.93 ± 0.04, 0.87 ± 0.03, 0.83 ± 0.05, and 0.89 ± 0.05, respectively.

Conclusion: The DNN model can differentiate speech data of patients with PD from that of patients with SCD with relatively high accuracy and AUC. The proposed method can be used as a non-invasive, easy-to-perform screening method to differentiate PD from SCD using patient speech and is expected to be applied to telemedicine.

Keywords: Parkinson's disease, spinocerebellar degeneration, dysarthria, machine learning, deep learning

Introduction

Parkinson's disease (PD) and spinocerebellar degeneration (SCD) cause motor speech abnormalities. Patients with PD usually develop hypokinetic dysarthria, characterized by monotonous and reduced vocal loudness [1], while ataxic dysarthria in SCD is characterized by difficulty sequencing syllables into fast, smooth, and rhythmically organized larger utterances [2]. Previous studies successfully distinguished PD from spinocerebellar ataxia (SCA) and multiple system atrophy (MSA) by applying acoustic analysis to patients' speech [3, 4]. Therefore, assessing dysarthria features in patients with neurodegenerative diseases helps diagnose underlying pathologies.

A number of studies have attempted to distinguish between voice-affecting disorders and healthy controls using acoustic analysis, especially for Parkinson's disease [5]. Many studies have successfully differentiated patients with PD from normally-speaking controls with high accuracy using features extracted by acoustic analysis and traditional machine learning methods, such as support vector machine, k-nearest neighbors algorithm, and random forest; however, the predictive performance of these studies highly varied [6]. One potential reason for the different performances is that acoustic features used for analysis widely varied among these studies. Although many kinds of acoustic features (such as fundamental frequency, spectrogram, and mel-frequency cepstral coefficients) have been adopted, currently no standard protocols exist to select acoustic features. Hence, it is necessary to select the optimal acoustic features according to the target disease and severity of dysarthria.

In addition to traditional machine learning methods, deep neural network (DNN) techniques have been widely adopted to process speech, music, and environmental sounds [7]. One advantage of DNN models is that they can automatically extract

features from input data that are useful for solving specific tasks. Because DNN models do not require prior feature selection and can be an end-to-end analysis, they have the potential to overcome the limitations of previous studies in which the performances varied depending on the selected features. One commonly used method for analyzing speech data using DNNs is to apply a convolutional neural network (CNN) to the spectrogram [7]. CNN is a DNN architecture adopted to analyze image data and this kind of architecture has demonstrated high performance in image classification tasks [8]. A spectrogram is a visual representation that displays the strength of a signal over time for a given frequency range. As a spectrogram has two dimensions (one axis represents time, and the other axis represents frequency), DNN models using the CNN architecture can process spectrograms in the same way as images. Previous studies have shown high performance of CNN models in the classification task of audio data, including human voice data [9, 10]. However, recently, a DNN architecture named transformer [11] has rapidly been utilized for various tasks, including image recognition, natural language processing, and speech processing. In a previous study, a DNN model using transformer architectures applied to spectrograms displayed better performance in audio classification tasks than CNN models [12]. We believe that this DNN architecture is potentially useful for analyzing dysarthria due to neurodegenerative disease.

Several studies reported that DNN models using CNN architectures could successfully discriminate patients with PD and healthy controls with spectrograms generated from voice data [13, 14]. However, studies that tested whether DNN models could differentiate neurodegenerative diseases like PD and SCD from speech data are limited. As DNN models do not require prior feature extraction and can perform end-to-end

analysis, we think they are potential easy-to-perform screening tools for differentiating neurodegenerative disease in patients with dysarthria. In addition, studies applying DMM models with transformer architecture to speech data from patients with neurodegenerative diseases are also limited. Therefore, this study aimed to evaluate whether a DNN model using a transformer architecture can differentiate patients with PD from those with SCD from speech data.

Methods

Patients

This retrospective study was based on audio data recorded during speech and language therapy for patients with PD and SCD at Hokuyukai Neurological Hospital between April 2014 and March 2023. The PD diagnosis was based on the United Kingdom Parkinson's Disease Society Bank Criteria [15]. Patients with MSA, hereditary cerebellar ataxia (HCAs), or sporadic adult-onset ataxia (SAOA) were included as those with SCD. The MSA diagnosis was based on consensus diagnostic criteria [16]. Only patients with the cerebellar variant (MSA-C) were included. Patients with HCAs were defined as those who presented with ataxic signs on neurological examination and had a positive family history of similar ataxia syndromes in at least one first- or second-degree relative. The criteria for SAOA were as follows: (1) presenting with progressive ataxia, (2) disease onset after 20 years of age, (3) no similar disorders in first- or second-degree relatives, (4) no established ataxia cause (no ischemia, hemorrhage, or tumor of the posterior fossa; no alcohol abuse; no chronic intake of anticonvulsant drugs; no other toxic causes; no malignancies; no evidence of ataxia onset in association with encephalitis, sepsis, hyperthermia, or heat stroke; normal thyroid function), (5) no

subacute ataxia onset, (6) and no probable MSA based on the consensus diagnostic criteria [16]. Finally, we analyzed speech data from 251 patients with PD and 101 with SCD. The breakdown of SCD diagnoses was: MSA-C, 21; HCAs, 65 (SCA 3, 23; patients without identified pathologic gene, 16; SCA 6, 15; SCA 1, 5; SCA 2, 3; SCA 31, 2; and SCA14, 1); and SAOA, 15. All included patients were native Japanese speakers. The study protocol was approved by the Institutional Review Board of the Hokuyukai Neurological Hospital (approval number: 2023-4), and the requirement for informed consent was waived owing to the study's retrospective nature. Demographic data and modified Rankin Scale (mRS) [17] at the first speech recording were obtained from medical records. In patients with PD, the Hoehn and Yahr (HY) stages [18] at the first speech recording were also obtained.

Speech recording

We used the speech data recorded during speech and language therapy. Speech recordings were performed in a quiet room using a microphone, and speech signals were sampled at 44,100 Hz and resampled at 32,000 Hz. During the recording, the patients were asked to read the Japanese passage "The North Wind and the Sun." The patients were encouraged to maintain a normal pace, cadence, and volume. Speech recordings of patients with PD were performed in the on-drug state. The dysarthria severity was evaluated based on auditory-perceptual judgments using recorded speech data. We used the Speech Intelligibility Rating (SIR) [19] to evaluate dysarthria severity.

The number of patients with PD and SCD was imbalanced, resulting in poor predictive performance of the trained DNN model. Some participants underwent multiple audio recordings to assess rehabilitation effectiveness or changes over time. To reduce the

speech data number differences between PD and SCD, only speech data from the first recording were used for patients with PD. However, multiple recordings from the same patient were used for patients with SCD. In total, 251 speech datasets from 251 patients with PD and 252 datasets from 101 patients with SCD were used.

Deep neural network model

We used the DNN architecture, Patchout faSt Spectrogram Transformer (PaSST) [12]. This model is a transformer-based DNN model applied to the spectrograms obtained from audio data [11]. The transformer is characterized by a series of self-attention layers that calculate the distance between each pair of items in the input sequence. This architecture excels in processing serial data and was originally proposed for natural language tasks. PaSST has exhibited high performance in audio data classification in the "AudioSet" dataset [20], consisting of more than 2 million 10-second audio data clips culled from YouTube videos and labeled with 527 different labels. In this study, PaSST was used to distinguish the input speech data from a patient with PD or SCD. An overview of the model is provided in Fig. 1.

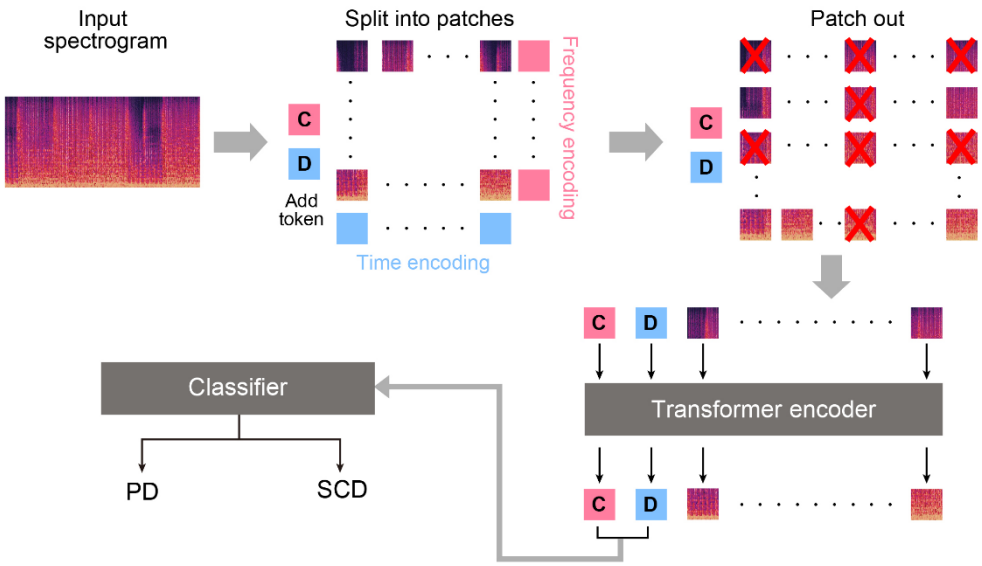

Fig. 1.

Overview of the deep learning model used in this study

C, classification token; D, distillation token; PD, Parkinson's disease; SCD, spinocerebellar degeneration

This study used the PaSST model pre-trained with the "AudioSet" provided in the literature [12]. To match the length of the audio data in the "AudioSet," we split the patients' speech data into 10-second segments. Thus, the input to the PaSST was a log-mel spectrogram obtained from a 10-second clip, a 128 × 1000 matrix (see '**Creating datasets and training the model**' below). The input spectrogram was then split into a sequence of 16 × 16 matrices with six overlaps in both the time and frequency dimensions. Each split 16 × 16 matrix was called a "patch." The patch was flattened to a 1-dimensional patch of size 768 using a linear projection layer. Because the transformer architecture did not capture the location information in the original spectrogram of each

patch, two positional encodings, one representing the frequency and one representing the time, were added to each patch. In addition, we added a “classification token” and a “distillation token” at the beginning of the patch sequence. These tokens were the same size as the 1-dimensional vectors of each patch.

The resulting sequence was then input into a Transformer Encoder block [11]. First, the input was fed into the layer normalization layer. Then, the output of the normalization layer was fed to the multi-head attention layer. We set the number of attention heads to 12 for all self-attention blocks. The output of the multi-head attention layer was added to the input of the block with a skip connection. The normalization layer was then applied again, and the output was fed to two fully connected layers. The first fully connected layer had 3,072 neurons, and the second had 768 neurons. The first fully connected layer followed the activation function of Gelu [21]. The sum of the output of the fully connected layers and the input of the layers with another skip connection was the final output of the self-attention layer. The size of the output of the self-attention block was the same as that of its input. After repeating the process of the self-attention block 12 times, the mean of the “classification token” and the “distillation token” was used for classification with a fully connected layer. Finally, feeding the output to the softmax function, we obtained two values: the probability that the input spectrogram represented PD and the probability that it represented SCD.

For model training, a “patchout” process was applied. In this process, some frequency bins and time frames were randomly selected, and whole values belonging to the column/row of the selected frequency bins and time frames were removed. This process reduced the computational complexity and memory requirements for training the model without degrading the classification performance [12]. We randomly dropped four

frequency bins and 40 frames for model training. For inference, all patches were used for prediction.

Creating datasets and training the model

We created spectrograms from patients' speech data as inputs for the model. As the duration of each audio dataset of the "AudioSet" was 10 s, we clipped each speech dataset into 10-second intervals. The number of clips per speech depended on the length. Each sub-clip was labeled as PD if the pre-divided speech data were from a patient with PD and as SCD if from a patient with SCD. Each divided clip was then converted into a log-mel spectrogram. To calculate the log-mel spectrogram, the fast Fourier transform window length was 25 ms, and the hop length was 10 ms. Thus, a log-mel spectrogram with 1000 frames and 128 mel frequency bins was calculated. A 128 × 1000 matrix was input into the model in our experiments.

All patients were randomly split into 80% training and 20% testing sets, with the patient ratio for PD and SCD maintained. Of the spectrograms obtained from the participants who were split into the training sets, 80% were used as training data and 20% as validation data. The training data were used to train the model, whereas the validation data were used to improve the hyperparameters, such as the number of training epochs and the learning rate. The testing sets were used to evaluate the model's prediction performance using the parameters that showed the best prediction performance in the validation data. Of note, there was no overlap of individuals among training and test sets. Although multiple speech datasets were used from some patients with SCD, we ensured that speech data obtained from the same patient were not assigned to the training and test sets simultaneously.

The model was trained to predict whether a label on the input spectrogram was from a patient with PD or SCD. For each input of a spectrogram, the model outputs two values: the probability that the spectrogram was obtained from a patient with PD and the probability that it was obtained from a patient with SCD. The prediction error was calculated as the binary cross-entropy loss between the model-predicted values and the label previously assigned to the input spectrogram. Subsequently, the model parameters were updated for binary cross-entropy loss value reduction. Since our dataset was small, we implemented a warm-start strategy using the parameter values of a pre-trained model as the initial values. In the literature, this strategy is known as the transfer learning technique [22]. In this study, we used pre-trained parameters with the "AudioSet" provided in the literature [12] as initial parameters and fine-tuned them using our dataset. The Adam optimizer was used for parameter optimization, with a learning rate of 0.00001 and weight decay of 0.0001. We set a mini-batch size of 16 and trained the model for 100 epochs.

Evaluating the Prediction Performance of the Deep Neural Network Model

After training the model, the prediction performance was evaluated using the testing set with the best predictive parameters for the validation data. This study generated multiple spectrograms from each individual speech dataset depending on its length and then the model output values for the probabilities of PD and SCD for a single spectrogram. Therefore, the average of the outputs for all spectrograms obtained from each set of speech data defined the final model prediction as PD if the average probability for PD exceeded 0.5; otherwise, it was defined as SCD. Additionally, although multiple speech datasets from some patients with SCD were used for model

training, only the first set of speech data was used as test data to evaluate the model performance.

We evaluated predictive performance using the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity. With reference to previous studies [23, 24], we used 5-fold cross-validation for model performance evaluation. To ensure that all the participants were split into a testing set once, we repeated the same experiment five times and reported the average of the above metrics for the five experiments.

Perceptual prediction by experienced speech-language pathologists

To compare the prediction performance of the DNN model, five experienced speech-language pathologists (SLPs) differentiated PD and SCD from speech data. We defined experienced SLPs as practitioners with at least 5 years of clinical experience, per previously established methodology [25]. Each SLP was assigned one of five testing sets and tasked with predicting PD or SCD from speech data within the assigned set. SLPs were blinded to all patient clinical information. We compared the predictive performance of the DNN model and STs.

Statistical analysis

The sex ratio between the patients with PD and SCD was compared using a two-proportional Z-test. The age between the two groups was compared using the Student's t-test. mRS and SIR at the first speech recording between the two groups were compared using the Mann–Whitney U test. Data are presented as mean ± standard deviation. A threshold of statistical significance was set at $p<0.05$.

Results

Characteristics of the participants

Patient characteristics and demographic data are shown in Table 1. There was no significant difference in the sex ratio and mRS between patients with PD and SCD. The age at the first speech recording was significantly higher in patients with PD. SIR values were significantly lower in the SCD group, indicating that the SCD group had more severe dysarthria than the PD group.

Predictive performance of the trained model

The average accuracy, sensitivity, and specificity of the trained model of the testing sets for the five experiments were 0.87 ± 0.03, 0.83 ± 0.05, and 0.89 ± 0.05, respectively (mean ± standard deviation). Overall, in the five experiments, the model correctly predicted the dysarthria label in 307 of the 352 patients. Fig. 2 shows a confusion matrix representing the relationship between the true labels and model predictions. The average AUC for the five experiments was 0.93 ± 0.04. The receiver operating characteristic curves for each of the five experiments are shown in Fig. 3.

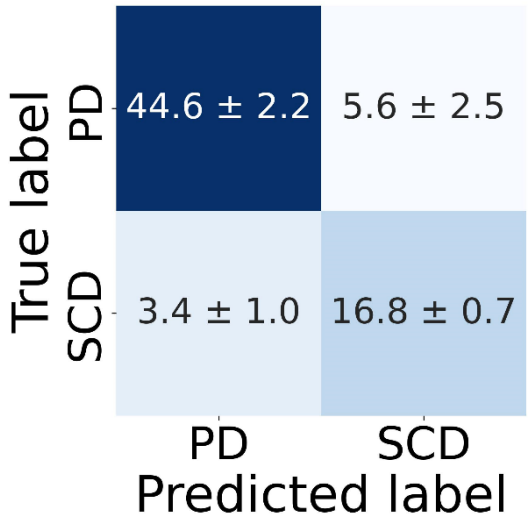

Fig. 2

Confusion matrix of the true label and predicted dysarthria type based on the trained model. Data are presented as the mean ± standard deviation for five experiments. PD, Parkinson's disease; SCD, spinocerebellar degeneration

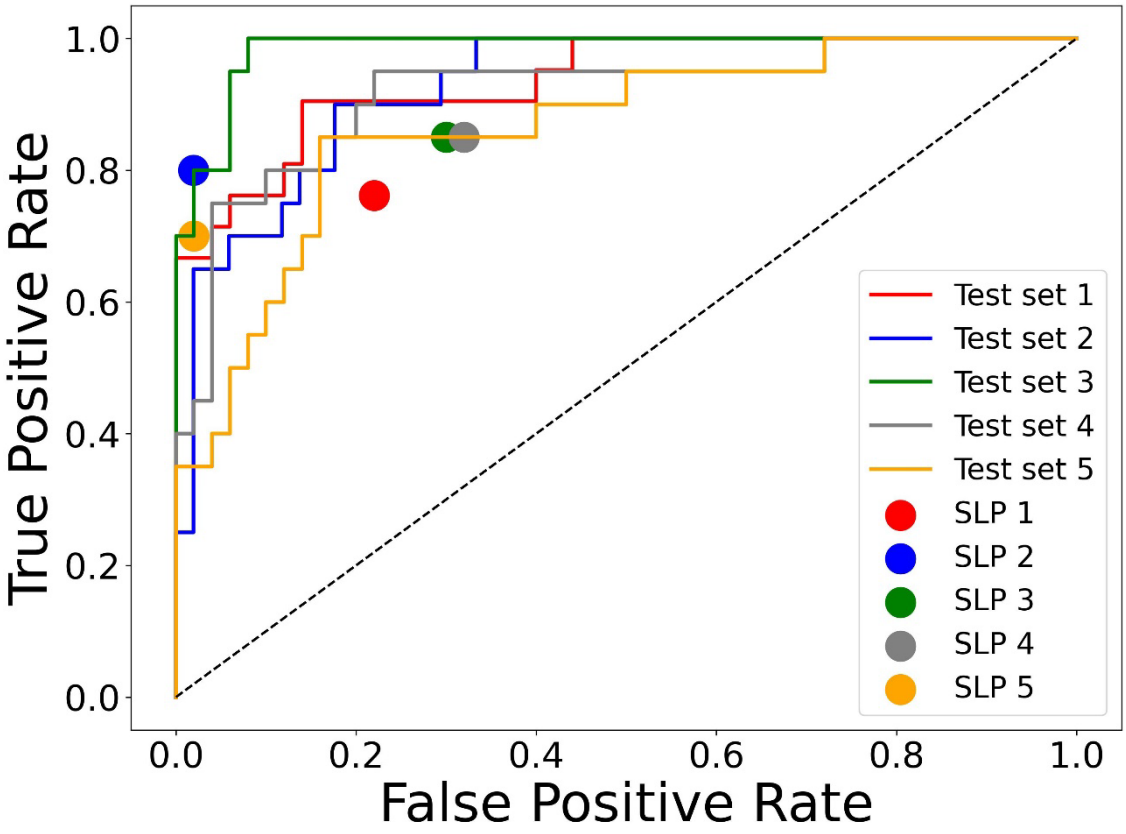

Fig. 3

Solid lines show the receiver operating characteristic curves for five experiments. Round markers show true positive and false positive rates of each SLP. SLP, speech-language pathologist

Comparison of the predictive performance between the trained model and speech-language-hearing therapists

The average accuracy, sensitivity, and specificity of the five SLPs were 0.82 ± 0.08, 0.79 ± 0.06, and 0.82 ± 0.13, respectively. A comparison of the accuracy of the SLPs and the model for each test set is provided in Table 2. The values of true positive rate

and false positive rate for each SLP were also plotted in Fig. 2. These results indicate that the model performed relatively better in differentiating patients with PD from those with SCD, based on speech data.

Additional experiment with propensity score matching

In this study, there were significant differences in age and SIR between the patients with PD and those with SCD. This bias may have favored the model's disease differentiation, and the model's predictive performance could have been overestimated. To balance the clinical characteristics, we conducted propensity score matching [26]. The propensity scores were calculated using logistic regression including age, SIR, and sex. One patient with SCD was matched with one patient with PD, according to nearest neighbor matching with a caliper of 0.2 standard deviations. After propensity score matching, 69 patients with SCD and 69 patients with PD were extracted. Characteristics and demographic data of the extracted patients are provided in Table 3. Because the number of patients with PD and SCD was the same, speech data from the first recording were used for both patients with PD and those with SCD. We performed the same experiment described in the Methods section (see '**Creating datasets and training the model'** and '**Evaluating the Prediction Performance of the Deep Neural Network Model**') with speech data from the extracted patients.

The average accuracy, sensitivity, and specificity of the model of the testing sets for the five experiments in the extracted patients were 0.83 ± 0.06, 0.90 ± 0.07, and 0.77 ± 0.10, respectively. Overall, in the five experiments, the model correctly predicted the PD and SCD in 115 of the 138 patients. Fig. 4 shows a confusion matrix representing the relationship between the true labels and model predictions. The average AUC for the

five experiments was 0.91 ± 0.02. The receiver operating characteristic curves for each of the five experiments are shown in Fig. 5.

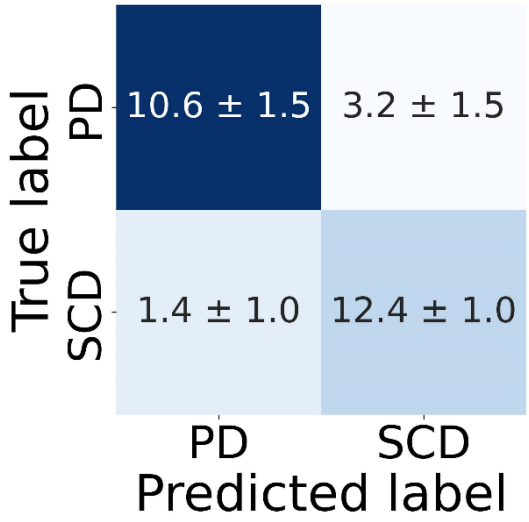

Fig. 4

Confusion matrix of the true label and predicted disease based on the trained model with extracted patients with propensity score matching. Data are presented as the mean ± standard deviation for five experiments. PD, Parkinson's disease; SCD, spinocerebellar degeneration

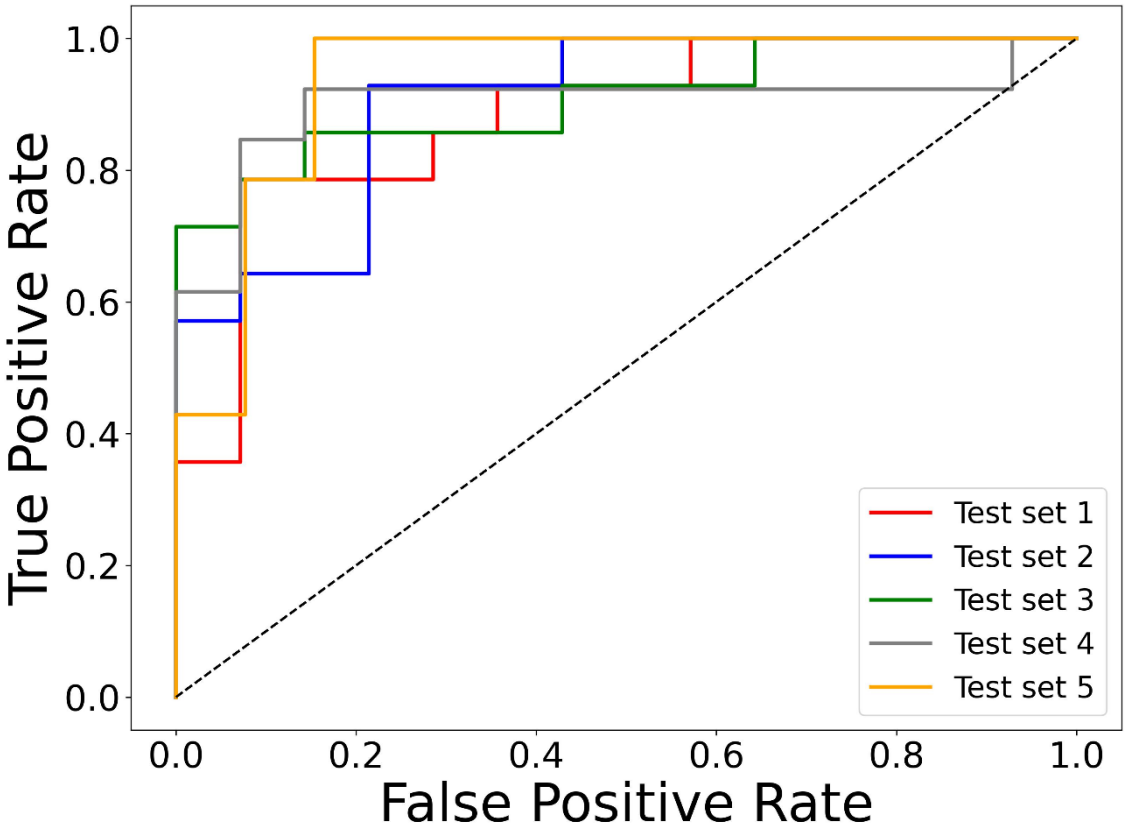

Fig. 5

Receiver operating characteristic curves for five experiments with extracted patients with propensity score matching

Discussion

We investigated whether a DNN model with a transformer architecture could differentiate between PD and SCD based on recorded speech data. The DNN model had relatively higher performance in classifying speech data from patients with PD and SCD than experienced SLPs. Our results indicate that the model has potential use in differentiating PD from SCD from speech data. The prevalence of PD is increasing worldwide [27], and early diagnosis is crucial for appropriate disease management. As

recording patients' voices is non-invasive and inexpensive, our method could be a potential easy-to-perform screening tool for differentiating PD from other neurodegenerative diseases that cause cerebellar ataxia. However, the accuracy of our model was 0.87 and hence, requires improvement prior to use in actual clinical practice.

In this study, SLPs predicted PD and SCD using only speech data in reading the Japanese passage "The North Wind and the Sun", which did not reflect a normal, clinical situation. Usually, physicians and SLPs do not evaluate dysarthria using only the text-reading task, but combine several tasks, such as continuous vowel utterance, diadochokinetic tasks, and free conversation. Additionally, each SLP assessed the speech data of a large number of patients, approximately 70 each, to differentiate diseases, which is also not a realistic clinical situation. Although the test conditions could have reduced the performance of the SLPs, the model performed relatively better than the SLPs in differentiating patients with PD from those with SCD from speech data. We believe this result suggests that the proposed model could be useful for differentiating neurological disorders based on patient speech data, and may be a useful aid in disease diagnosis in situations where voices are the main source of data, such as telemedicine.

In this study, patients with SCD were significantly younger and had a higher dysarthria severity compared to patients with PD. This raised concerns that the model predicted PD and SCD based on age estimated from voice and dysarthria severity, rather than on the dysarthria features themselves. Therefore, we performed propensity score matching to assess whether the model could differentiate PD and SCD in a population that did not differ in age and dysarthria severity. Although this reduced the model's performance compared to the original dataset, the model still showed higher accuracy than the STs.

The reduction of the model's performance in extracted patients could be because of the reduced number of data as well as the alignment of age and dysarthria severity. Generally, the performance of DNN models is expected to improve with an increase in training data. By increasing the amount of data, the model's performance for an age- and dysarthria severity-matched group of patients could approach that of the original data. Additionally, we expect that increasing the number of participants would improve the model performance to the point where it could be used in actual, clinical practice. Further studies with larger numbers of patients are needed to validate this.

This study has several limitations. First, participants' cognitive function was not assessed. Cognitive decline reduces reading fluency and speed, resulting in difficulty distinguishing between PD and SCD. Second, the SCD group included patients with MSA-C, which usually manifests as mixed ataxic-spastic dysarthria rather than purely ataxic dysarthria [4]. The DNN model did not distinguish between pure hypokinetic dysarthria and ataxic dysarthria, and the speech of patients with spastic dysarthria could possibly be misjudged as coming from patients with SCD. Third, because DNNs are generally highly complex mathematical models and process high-dimensional data, it is difficult to understand which features from the speech data were used by the DNN model to differentiate PD and SCD. The black-box nature of DNNs makes it difficult for clinicians to decide if the DNN's predictions are trustworthy. To address this issue, future studies should identify features in the data that the DNN focuses on to differentiate diseases. Fourth, the relatively large amount of speech data we used placed a high burden on assessment, and it was not possible to assess inter-rater reliability among SLPs with respect to differentiating PD and SCD from speech data. Finally, normal control participants were not enrolled in this study. Therefore, we were unable

to verify whether the model could differentiate between the presence and absence of dysarthria. Additionally, it would be more clinically useful if the model could differentiate not only PD and SCD but also other voice-affecting disorders. Further studies that include normal controls and increasing target diseases are necessary to elucidate this.

In conclusion, this study showed that a DNN model with a transformer architecture could differentiate speech data from patients with PD and SCD with relatively high accuracy and AUC. The proposed method is a potential non-invasive, easy-to-perform screening method using patient speech to differentiate PD from other neurodegenerative diseases that cause cerebellar ataxia. This method is expected to be used in situations where voice is important information, for example, in telemedicine. However, it is desirable to improve the predictive performance of the model and to verify whether it can be differentiated from healthy controls in order to use it in actual clinical practice. To achieve this, further studies that include normal controls and more participants are needed.

Statements and Declarations

Ethics approval

The study protocol was approved by the Institutional Review Board of the Hokuyukai Neurological Hospital (approval number: 2023-4)

Consent to participate

Informed consent was waived owing to the study's retrospective nature.

Data availability

The data that support the findings of this study are available on reasonable request from the corresponding author.

References

[1] Darley FL, Aronson AE, Brown JR (1969) Differential diagnostic patterns of dysarthria. J Speech Hear Res 12:246–269. https://doi.org/10.1044/jshr.1202.246

[2] Ackermann H (2008) Cerebellar contributions to speech production and speech perception: psycholinguistic and neurobiological perspectives. Trends Neurosci 31:265–272. https://doi.org/10.1016/j.tins.2008.02.011

[3] Schmitz-Hübsch T, Eckert O, Schlegel U, Klockgether T, Skodda S (2012) Instability of syllable repetition in patients with spinocerebellar ataxia and Parkinson's disease. Mov Disord 27:316–319. https://doi.org/10.1002/mds.24030

[4] Rusz J, Tykalová T, Salerno G, Bancone S, Scarpelli J, Pellecchia MT (2019) Distinctive speech signature in cerebellar and parkinsonian subtypes of multiple system atrophy. J Neurol 266:1394–1404. https://doi.org/10.1007/s00415-019-09271-7

[5] Idrisoglu A, Dallora AL, Anderberg P, Berglund JS (2023) Applied machine learning techniques to diagnose voice-affecting conditions and disorders: systematic literature review. J Med Internet Res 19:e46105. https://doi.org/10.2196/46105.

[6] Ngo QC, Motin MA, Pah ND, Drotár P, Kempster P, Kumar D (2022) Computerized analysis of speech and voice for Parkinson's disease: A systematic review. Comput Methods Programs Biomed 226:107133. https://doi.org/10.1016/j.cmpb.2022.107133

[7] Purwins H, Li B, Virtanen T, Schulter J, Chang S, Sainath T (2019) Deep learning for audio signal processing, IEEE J. Sel Top Sig Process 13:206–219. http://doi.org/10.1109/JSTSP.2019.2908700

[8] Ajit A, Acharya K, and Samanta A (2020) A Review of Convolutional Neural Networks. International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE). pp. 1-5. http://doi.org/10.1109/ic-ETITE47903.2020.049

[9] Piczak KJ (2015). Environmental sound classification with convolutional neural networks. In: Proc. 25th Int. Workshop Mach. Learning Signal Process. pp. 1–6. http://doi.org/10.1109/MLSP.2015.7324337

[10] Jianfeng Z, Xia M, Lijiang C (2018) Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomed Signal Process Control 47: 312-323. https://doi.org/10.1016/j.bspc.2018.08.035

[11] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. Presented at: Proceedings of the Advances in Neural Information Processing Systems pp. 1–11. https://doi.org/10.48550/arXiv.1706.03762

[12] Koutini K, Schlüter J, Eghbal-zadeh H, Widmer G (2022) Efficient training of audio transformers with patchout. Proc Interspeech pp. 2753–2757. https://doi.org/10.48550/arXiv.2110.05069

[13] Hireš M, Gazda M, Drotár P, Pah ND, Motin MA, Kumar DK (2022) Convolutional neural network ensemble for Parkinson's disease detection from voice recordings. Comput Biol Med 141:105021. https://doi.org/10.1016/j.compbiomed.2021.105021

[14] Zhang X, Ma J, Li Y, Wang P, Liu Y (2021) Few-shot learning of Parkinson's disease speech data with optimal convolution sparse kernel transfer learning. Biomed Signal Process Control 69:102850. https://doi.org/10.1016/j.bspc.2021.102850

[15] Hughes AJ, Daniel SE, Kilford L, Lees AJ (1992) Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. J Neurol Neurosurg Psychiatry 55:181–184. http://dx.doi.org/10.1136/jnnp.55.3.181

[16] Gilman S, Wenning GK, Low PA, Brooks DJ, Mathias CJ, Trojanowski JQ, Wood NW, Colosimo C, Dürr A, Fowler CJ, Kaufmann H, Klockgether T, Lees A, Poewe W, Quinn N, Revesz T, Robertson D, Sandroni P, Seppi K, Vidailhet M (2008) Second consensus statement on the diagnosis of multiple system atrophy. Neurology 71:670–676. https://doi.org/10.1212/01.wnl.0000324625.00404.15

[17] van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van Gijn J (1988) Interobserver agreement for the assessment of handicap in stroke patients. Stroke 19:604–607. https://doi.org/10.1161/01.STR.19.5.604

[18] Hoehn MM, Yahr MD (1967) Parkinsonism: onset, progression and mortality. Neurology 17:427–442. https://doi.org/10.1212/WNL.17.5.427

[19] Zhou H, Chen Z, Shi H, Wu Y, Yin S (2013) Categories of auditory performance and speech intelligibility ratings of early-implanted children without speech training. PLoS One 8:e53852. https://doi.org/10.1371/journal.pone.0053852

[20] Gemmeke JF, Ellis DPW, Freedman D, Jansen A, Lawrence W, Moore RC, Pakal M, Ritter M (2017) Audio Set: An ontology and human-labeled dataset for audio events. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 776–780. https://doi.org/10.1109/ICASSP.2017.7952261

[21] Hendrycks D, Gimpel K (2016) Gaussian Error Linear Units (Gelus) https://doi.org/10.48550/arXiv.1606.08415

[22] Naseer A, Rani M, Naz S, Razzak MI, Imran M, Xu G (2020) Refining Parkinson's neurological disorder identification through deep transfer learning. Neural Comput Appl 32:39–854. https://doi.org/10.1007/s00521-019-04069-0

[23] Abou Jaoude M, Jing J, Sun H, Jacobs CS, Pellerin KR, Westover MB, Cash SS, Lam AD (2020) Detection of mesial temporal lobe epileptiform discharges on intracranial electrodes using deep learning. Clin Neurophysiol 131:133-141. https://doi.org/10.1016/j.clinph.2019.09.031

[24] Fast L, Temuulen U, Villringer K, Kufner A, Ali HF, Siebert E, Huo S, Piper SK, Sperber PS, Liman T, Endres M, Ritter K (2023) Machine learning-based prediction of clinical outcomes after first-ever ischemic stroke. Front Neurol 14:1114360. https://doi.org/10.3389/fneur.2023.1114360

[25] Nakayama K, Yamamoto T, Oda C, Sato M, Murakami T, Horiguchi S (2020) Effectiveness of Lee Silverman Voice Treatment® LOUD on Japanese-Speaking Patients with Parkinson's Disease. Rehabil Res Pract 24:6585264. https://doi.org/10.1155/2020/6585264

[26] Pattanayak CW, Rubin DB, Zell ER (2011) Propensity score methods for creating covariate balance in observational studies. Rev Esp Cardiol 64:897–903. https://doi.org/10.1016/j.recesp.2011.06.008

[27] Ascherio A, Schwarzschild MA (2016) The epidemiology of Parkinson's disease: risk factors and prevention. Lancet Neurol 15:1257–1272. https://doi.org/10.1016/S1474-4422(16)30230-7

[28] Zhou H, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 12, pp. 11106-11115).

Tables

Table 1. Patients' characteristics and demographic data

	PD (n=251)	**SCD (n=101)**	**p-value**
Male (%)	39.8 (n=99)	42.2 (n=41)	0.39
Age (years)	71.6 ± 7.4	64.3 ± 11.1	< 0.001
Disease duration (year)	8.4 ± 5.5	10.6 ± 9.5	NA
Modified Rankin Scale	2.9 ± 0.87	3.0 ± 0.97	0.86
Hoehn and Yahr stage	3.6 ± 0.90	NA	NA
Speech Intelligibility Rating	4.3 ± 0.59	3.8 ± 0.60	< 0.001

Data are at the first speech recording. Data are presented as mean ± standard deviation.

PD, Parkinson's disease; SCA, spinocerebellar ataxia

Table 2. Comparison of the accuracy of the deep neural network model and speech-language pathologist for each test set

	DNN model	**SLP**
Test set 1	88.7%	77.5%
Test set 2	83.1%	93.0%
Test set 3	92.9%	74.3%
Test set 4	87.1%	72.9%
Test set 5	84.3%	90.0%
average	87.2%	81.5%

DNN, deep neural network; SLP, speech-language pathologist

Table 3. Characteristics and demographic data of patients, extracted by propensity score matching

	PD (n=69)	**SCD (n=69)**	**p-value**
Male (%)	45.0 (n=31)	46.4 (n=32)	0.87
Age (years)	67.9 ± 8.8	68.4 ± 8.6	0.72
Modified Rankin Scale	3.0 ± 0.81	2.9 ± 0.98	0.69
Speech Intelligibility Rating	4.0 ± 0.62	4.0 ± 0.51	0.62

Data are presented as mean ± standard deviation.

PD, Parkinson's disease; SCA, spinocerebellar ataxia