Automated Entry of Paper-Based Patient-Reported Outcomes: Applying Deep Learning to the Japanese Orthopaedic Association Back Pain Evaluation Questionnaire

Kosuke Kita[1], Takahito Fujimori[2*], Yuki Suzuki[1], Takashi Kaito[3], Shota Takenaka[4] Yuya Kanie[2], Masayuki Furuya[2], Tomohiro Wataya[1], Daiki Nishigaki[1], Junya Sato[1], Noriyuki Tomiyama[1], Seiji Okada[2], and Shoji Kido[1]

[1] Osaka University School of Medicine Graduate School of Medicine Diagnostic and Interventional Radiology
[2] Osaka University Graduate School of Medicine Department of Orthopedic Surgery
[3] Osaka Rosai Hospital
[4] Japan Community Health care Organization Osaka Hospital

[*] **Corresponding author**
Takahito Fujimori**,**
takahito-f@hotmail.co.jp
Phone: +81-6- 6879-3552
Fax: +81-6- 6879-3559

Abstract

[Background] Health-related patient-reported outcomes (HR-PROs) are crucial for assessing the quality of life among individuals experiencing low back pain. However, manual data entry from paper forms, while convenient for patients, imposes a considerable tallying burden on collectors. In this study, we developed a deep learning (DL) model capable of automatically reading these paper forms. [Methods] We employed the Japanese Orthopaedic Association Back Pain Evaluation Questionnaire, a globally recognized assessment tool for low back pain. The questionnaire comprised 25 low back pain-related multiple-choice questions and three pain-related visual analog scales (VASs). We collected 1305 forms from an academic medical center as the training set, and 483 forms from a community medical center as the test set. The performance of our DL model for multiple-choice questions was evaluated using accuracy as a categorical classification task. The performance for VASs was evaluated using the correlation coefficient and absolute error as regression tasks. [Result] In external validation, the mean accuracy of the categorical questions was 0.997. When outputs for categorical questions with low probability (threshold: 0.9996) were excluded, the accuracy reached 1.000 for the remaining 65 % of questions. Regarding the VASs, the average of the correlation coefficients was 0.989, with the mean absolute error being 0.25. [Conclusion] Our DL model demonstrated remarkable accuracy and correlation coefficients when automatic reading paper-based HR-PROs during external validation.

Key Words

HR-PRO, back pain, JOABPEQ, deep learning, questionnaire, convolutional neural network, artificial intelligence

Introduction

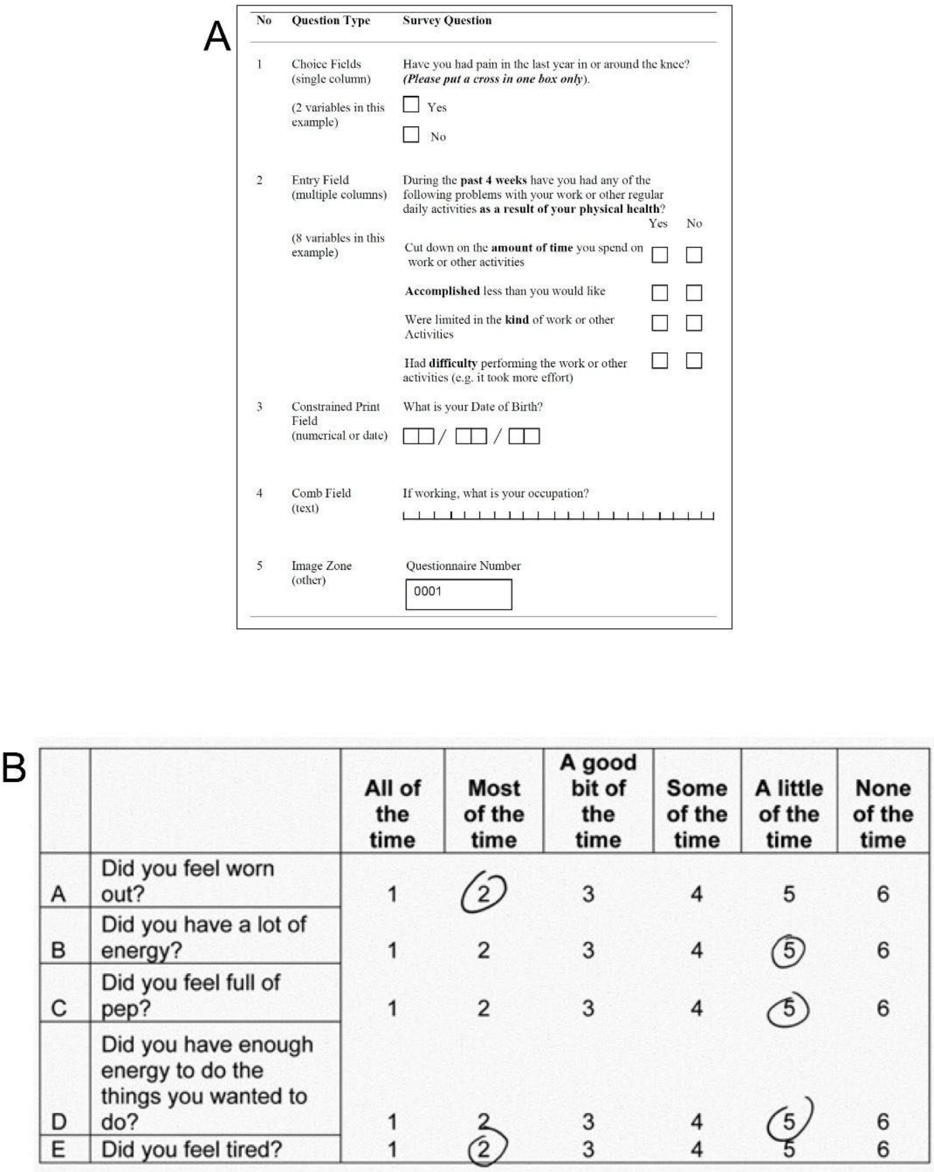
Low back pain is a prevalent musculoskeletal disorder that significantly affects global health. According to the World Health Organization, around 570 million individuals worldwide suffer from this condition, accounting for approximately 7.4 % of years lived with disability [1].

Unlike conditions such as cancer and cardiovascular diseases, which have distinct endpoints related to mortality or death, low back pain does not directly impact life expectancy. Consequently, the use of health-related patient-reported outcomes (HR-PROs) becomes crucial for assessing the quality of life for individuals affected by low back pain. Moreover, the significance of HR-PROs has become increasingly recognized for evaluating outcomes of new drugs and medical devices seeking approval [2].

While some HR-PROs are now accessible on electronic devices like tablet terminals, the elderly population might face challenges with electronic data entry. Consequently, numerous healthcare facilities continue to rely on paper-based data collection methods. Furthermore, paper-based HR-PROs offer advantages such as not requiring tablets or Wi-Fi connectivity and are convenient for both patients and healthcare providers due to their ease of use and handling.

Nevertheless, paper-based HR-PRO can be cumbersome for data collectors. Since HR-PRO data is gathered both before and after interventions, this often leads to processing a substantial volume of HR-PRO forms, sometimes exceeding 1000 cases per year for data collectors. Hence, there is an urgent need for a system capable of automating the information collection process from paper-based HR-PRO.

Existing approaches for collecting data from paper forms typically rely on optical character recognition (OCR) technology. In the healthcare sector, previous studies employing OCR to automatically retrieve data from paper forms utilized software like Teleform [3]–[5]. Notably, Teleform demonstrated high accuracy in collecting data from paper forms originally designed within its framework (**Figure 1A**). However, Wahi et al. (2008) highlighted its inefficiency when applied to other existing forms not designed using the Teleform program [3]–[5] (**Figure 1B**). This poses a significant challenge for OCR technology in retrieving data from HR-PRO, which is also an existing form not designed utilizing the Teleform program.



No	Question Type	Survey Question
1	Choice Fields (single column) (2 variables in this example)	Have you had pain in the last year in or around the knee? ***(Please put a cross in one box only)***. □ Yes □ No
2	Entry Field (multiple columns) (8 variables in this example)	During the **past 4 weeks** have you had any of the following problems with your work or other regular daily activities **as a result of your physical health**? Yes No Cut down on the **amount of time** you spend on work or other activities □ □ **Accomplished** less than you would like □ □ Were limited in the **kind** of work or other Activities □ □ Had **difficulty** performing the work or other activities (e.g. it took more effort) □ □
3	Constrained Print Field (numerical or date)	What is your Date of Birth? □□ / □□ / □□
4	Comb Field (text)	If working, what is your occupation?
5	Image Zone (other)	Questionnaire Number 0001

		All of the time	Most of the time	A good bit of the time	Some of the time	A little of the time	None of the time
A	Did you feel worn out?	1	[✓ 2]	3	4	5	6
B	Did you have a lot of energy?	1	2	3	4	[✓ 5]	6
C	Did you feel full of pep?	1	2	3	4	[✓ 5]	6
D	Did you have enough energy to do the things you wanted to do?	1	2	3	4	[✓ 5]	6
E	Did you feel tired?	1	[✓ 2]	3	4	5	6

Figure 1. Example forms of questions investigated in previous studies related to optical character recognition (OCR). (A) A form designed using the Teleform designer program, which is a software that employs OCR. Teleform was able to collect data with high accuracy for paper forms structures within the Teleform program. (B) An existing form not designed utilizing the Teleform program. Teleform was not able to collect data from the existing form, which was not designed within the Teleform program.

Deep learning (DL) has emerged as a dominant technique in computer vision, notably featuring the widespread application of convolutional neural networks (CNNs) across various domains [6]–[12]. Nevertheless, the extent to which DL models can effectively overcome the difficulties in collecting data from paper-based HR-PRO while ensuring accuracy and reliability remains relatively unexplored. Understanding the impact of employing DL-based models to collect HR-PRO data on healthcare decision-making, patient outcomes, and resource utilization requires rigorous evaluation. By leveraging CNN, there exists potential to design a system capable of automating data collection from paper-based HR-PRO. In this study, we aimed to develop a DL model tailored for extracting data from paper-based HR-PRO.

The remainder of this paper is organized as follows. The Methods section delineates the collected datasets and offers comprehensive information about our model. In the Results section, we present the performance metrics of the proposed model and showcase representative cases. Subsequently, the Discussion section outlines the advantages of our model as compared to humans or previous technologies.

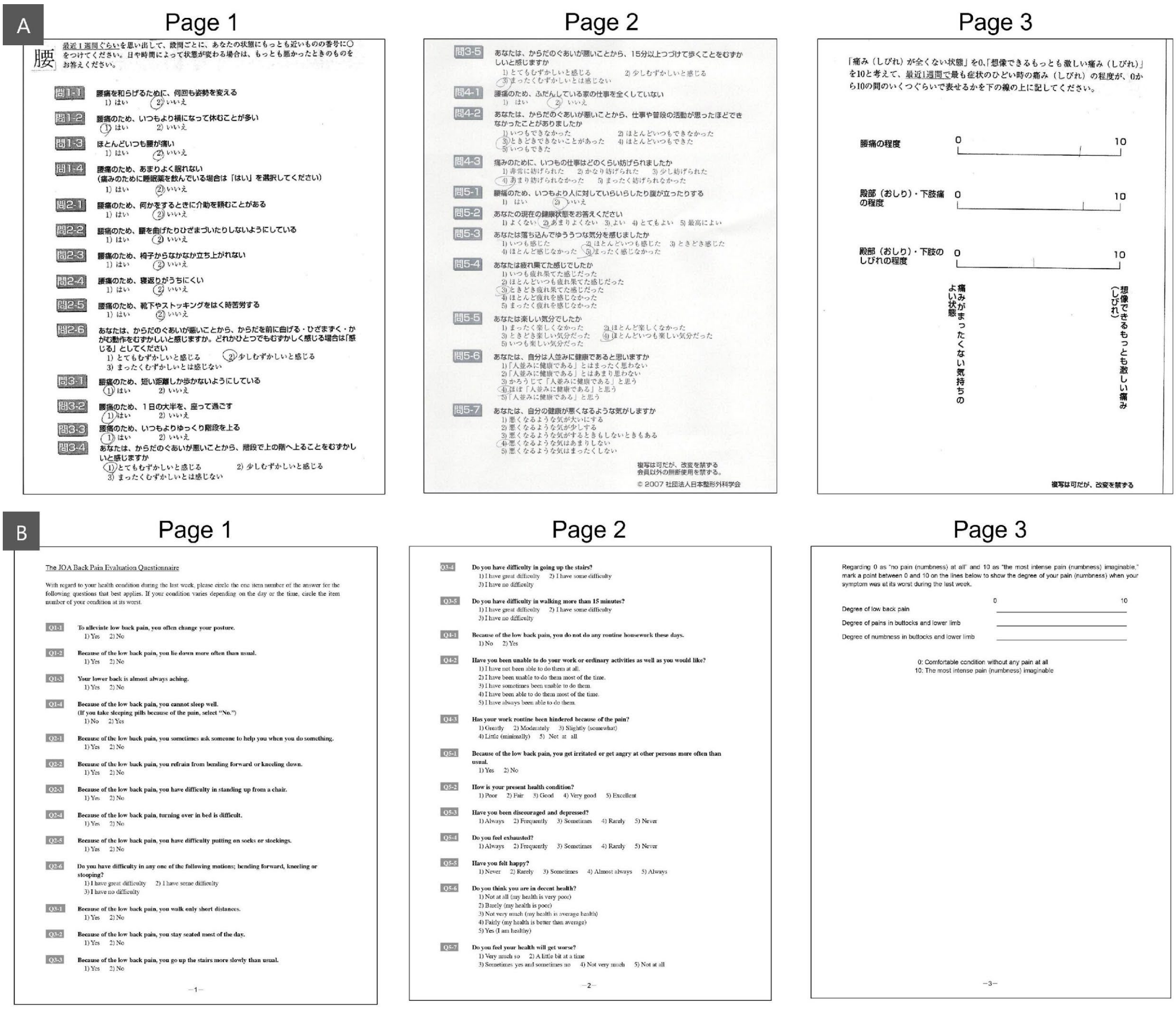
Methods

Questionnaire Dataset

This study received approval from the institutional review board of a blinded institution, and consent requirements were waived due to the retrospective nature of the study. Our study focused on the Japanese Orthopaedic Association Back Pain Evaluation Questionnaire (JOABPEQ). Widely recognized as a common HR-PRO for evaluating back pain [13]–[15] (**Figure 2**), JOABPEQ has been validated in various languages and is used worldwide [14], [16]–[19]. Notably, versions of JOABPEQ are available in English (**Figure 2**), Chinese, Thai, Arabic, and Turkish.

JOABPEQ comprises three pages, and **Figure 2A** presents examples of typical Japanese patient responses. There are 25 categorical questions spread across pages 1 and 2, with three visual analog scales (VASs) on page 3. Patients are required to select the most appropriate response from a series of choices provided for the categorical questions. Though the questionnaire does not specify the method of marking, individuals in Japan and Korea typically use a circled mark to indicate their choice (**Figure 2A**).

For the VASs, the patients were prompted to rate their pain levels by marking a straight line spanning from 0 to 10 cm, with 10 cm representing the most intense pain and 0 cm indicating the least pain (**Figure 2A**). Subsequently, evaluators determine the patient's level of pain by measuring the length of the marked line.



A

Page 1

腰 最近1週間ぐらいを思い出して、設問ごとに、あなたの状態にもっとも近いものの番号に○をつけてください。日や時間によって状態が変わる場合は、もっとも悪かったときのものをお答えください。

問1-1 腰痛を和らげるために、何回も姿勢を変える
1) はい [✓ 2) いいえ]

問1-2 腰痛のため、いつもより横になって休むことが多い
[✓ 1) はい] 2) いいえ

問1-3 ほとんどいつも腰が痛い
1) はい [✓ 2) いいえ]

問1-4 腰痛のため、あまりよく眠れない
(痛みのために睡眠薬を飲んでいる場合は「はい」を選択してください)
1) はい [✓ 2) いいえ]

問2-1 腰痛のため、何かをするときに介助を頼むことがある
1) はい [✓ 2) いいえ]

問2-2 腰痛のため、腰を曲げたりひざまづいたりしないようにしている
1) はい [✓ 2) いいえ]

問2-3 腰痛のため、椅子からなかなか立ち上がれない
1) はい [✓ 2) いいえ]

問2-4 腰痛のため、寝返りがうちにくい
1) はい [✓ 2) いいえ]

問2-5 腰痛のため、靴下やストッキングをはく時苦労する
1) はい [✓ 2) いいえ]

問2-6 あなたは、からだのぐあいが悪いことから、からだを前に曲げる・ひざまずく・かがむ動作をむずかしいと感じますか。どれかひとつでもむずかしく感じる場合は「感じる」としてください
1) とてもむずかしいと感じる [✓ 2) 少しむずかしいと感じる]
3) まったくむずかしいとは感じない

問3-1 腰痛のため、短い距離しか歩かないようにしている
[✓ 1) はい] 2) いいえ

問3-2 腰痛のため、1日の大半を、座って過ごす
[✓ 1) はい] 2) いいえ

問3-3 腰痛のため、いつもよりゆっくり階段を上る
[✓ 1) はい] 2) いいえ

問3-4 あなたは、からだのぐあいが悪いことから、階段で上の階へ上ることをむずかしいと感じますか
[✓ 1) とてもむずかしいと感じる] 2) 少しむずかしいと感じる
3) まったくむずかしいとは感じない

Page 2

問3-5 あなたは、からだのぐあいが悪いことから、15分以上つづけて歩くことをむずかしいと感じますか
1) とてもむずかしいと感じる 2) 少しむずかしいと感じる
[✓ 3) まったくむずかしいとは感じない]

問4-1 腰痛のため、ふだんしている家の仕事を全くしていない
1) はい [✓ 2) いいえ]

問4-2 あなたは、からだのぐあいが悪いことから、仕事や普段の活動が思ったほどできなかったことがありましたか
1) いつもできなかった 2) ほとんどいつもできなかった
[✓ 3) ときどきできないことがあった] 4) ほとんどいつもできた
5) いつもできた

問4-3 痛みのために、いつもの仕事はどのくらい妨げられましたか
1) 非常に妨げられた 2) かなり妨げられた 3) 少し妨げられた
[✓ 4) あまり妨げられなかった] 5) まったく妨げられなかった

問5-1 腰痛のため、いつもより人に対していらいらしたり腹が立ったりする
1) はい [✓ 2) いいえ]

問5-2 あなたの現在の健康状態をお答えください
1) よくない [✓ 2) あまりよくない] 3) よい 4) とてもよい 5) 最高によい

問5-3 あなたは落ち込んでゆううつな気分を感じましたか
1) いつも感じた 2) ほとんどいつも感じた 3) ときどき感じた
4) ほとんど感じなかった [✓ 5) まったく感じなかった]

問5-4 あなたは疲れ果てた感じでしたか
1) いつも疲れ果てた感じだった
2) ほとんどいつも疲れ果てた感じだった
[✓ 3) ときどき疲れ果てた感じだった]
4) ほとんど疲れを感じなかった
5) まったく疲れを感じなかった

問5-5 あなたは楽しい気分でしたか
1) まったく楽しくなかった 2) ほとんど楽しくなかった
3) ときどき楽しい気分だった [✓ 4) ほとんどいつも楽しい気分だった]
5) いつも楽しい気分だった

問5-6 あなたは、自分は人並みに健康であると思いますか
1)「人並みに健康である」とはまったく思わない
2)「人並みに健康である」とはあまり思わない
3) かろうじて「人並みに健康である」と思う
[✓ 4) ほぼ「人並みに健康である」と思う]
5)「人並みに健康である」と思う

問5-7 あなたは、自分の健康が悪くなるような気がしますか
1) 悪くなるような気が大いにする
2) 悪くなるような気が少しする
3) 悪くなるような気がするときもしないときもある
[✓ 4) 悪くなるような気はあまりしない]
5) 悪くなるような気はまったくしない

複写は可だが、改変を禁ずる
会員以外の無断使用を禁ずる。

© 2007 社団法人日本整形外科学会

Page 3

「痛み（しびれ）が全くない状態」を0､「想像できるもっとも激しい痛み（しびれ）」を10と考えて、最近1週間で最も症状のひどい時の痛み（しびれ）の程度が、0から10の間のいくつぐらいで表せるかを下の線の上に記してください。

腰痛の程度 0 10

殿部（おしり）・下肢痛の程度 0 10

殿部（おしり）・下肢のしびれの程度 0 10

痛みがまったくない気持ちのよい状態

想像できるもっとも激しい痛み（しびれ）

複写は可だが、改変を禁ずる

B

Page 1

The JOA Back Pain Evaluation Questionnaire

With regard to your health condition during the last week, please circle the one item number of the answer for the following questions that best applies. If your condition varies depending on the day or the time, circle the item number of your condition at its worst.

Q1-1 To alleviate low back pain, you often change your posture.
1) Yes 2) No

Q1-2 Because of the low back pain, you lie down more often than usual.
1) Yes 2) No

Q1-3 Your lower back is almost always aching.
1) Yes 2) No

Q1-4 Because of the low back pain, you cannot sleep well.
(If you take sleeping pills because of the pain, select "No.")
1) No 2) Yes

Q2-1 Because of the low back pain, you sometimes ask someone to help you when you do something.
1) Yes 2) No

Q2-2 Because of the low back pain, you refrain from bending forward or kneeling down.
1) Yes 2) No

Q2-3 Because of the low back pain, you have difficulty in standing up from a chair.
1) Yes 2) No

Q2-4 Because of the low back pain, turning over in bed is difficult.
1) Yes 2) No

Q2-5 Because of the low back pain, you have difficulty putting on socks or stockings.
1) Yes 2) No

Q2-6 Do you have difficulty in any one of the following motions; bending forward, kneeling or stooping?
1) I have great difficulty 2) I have some difficulty
3) I have no difficulty

Q3-1 Because of the low back pain, you walk only short distances.
1) Yes 2) No

Q3-2 Because of the low back pain, you stay seated most of the day.
1) Yes 2) No

Q3-3 Because of the low back pain, you go up the stairs more slowly than usual.
1) Yes 2) No

—1—

Page 2

Q3-4 Do you have difficulty in going up the stairs?
1) I have great difficulty 2) I have some difficulty
3) I have no difficulty

Q3-5 Do you have difficulty in walking more than 15 minutes?
1) I have great difficulty 2) I have some difficulty
3) I have no difficulty

Q4-1 Because of the low back pain, you do not do any routine housework these days.
1) No 2) Yes

Q4-2 Have you been unable to do your work or ordinary activities as well as you would like?
1) I have not been able to do them at all.
2) I have been unable to do them most of the time.
3) I have sometimes been unable to do them.
4) I have been able to do them most of the time.
5) I have always been able to do them.

Q4-3 Has your work routine been hindered because of the pain?
1) Greatly 2) Moderately 3) Slightly (somewhat)
4) Little (minimally) 5) Not at all

Q5-1 Because of the low back pain, you get irritated or get angry at other persons more often than usual.
1) Yes 2) No

Q5-2 How is your present health condition?
1) Poor 2) Fair 3) Good 4) Very good 5) Excellent

Q5-3 Have you been discouraged and depressed?
1) Always 2) Frequently 3) Sometimes 4) Rarely 5) Never

Q5-4 Do you feel exhausted?
1) Always 2) Frequently 3) Sometimes 4) Rarely 5) Never

Q5-5 Have you felt happy?
1) Never 2) Rarely 3) Sometimes 4) Almost always 5) Always

Q5-6 Do you think you are in decent health?
1) Not at all (my health is very poor)
2) Barely (my health is poor)
3) Not very much (my health is average health)
4) Fairly (my health is better than average)
5) Yes (I am healthy)

Q5-7 Do you feel your health will get worse?
1) Very much so 2) A little bit at a time
3) Sometimes yes and sometimes no 4) Not very much 5) Not at all

—2—

Page 3

Regarding 0 as "no pain (numbness) at all" and 10 as "the most intense pain (numbness) imaginable," mark a point between 0 and 10 on the lines below to show the degree of your pain (numbness) when your symptom was at its worst during the last week.

0 10

Degree of low back pain

Degree of pains in buttocks and lower limb

Degree of numbness in buttocks and lower limb

0: Comfortable condition without any pain at all
10: The most intense pain (numbness) imaginable

—3—

Figure 2. Images illustrating the Japanese Orthopaedic Association Back Pain Evaluation Questionnaire (JOABPEQ) in Japanese and English. The JOABPEQ comprises three pages: (A) Typical responses to the Japanese version of the JOABPEQ (The form was answered by a patient). (B) English version of the JOABPEQ. These images aim to facilitate comprehension of JOABPEQ content for non-Japanese readers.

Eligible Facilities

We retrospectively reviewed the medical lists of patients who completed the JOABPEQs at both the Academic Medical Center and Community Medical Center from April 1, 2016, and April 1, 2021. We recorded 1698 and 537 patients at the academic and community center, respectively. In cases where a patient filled out multiple JOABPEQ forms, such as preoperatively and postoperatively, we randomly selected one form to prevent data duplication or leakage in our DL model. The JOABPEQ forms were scanned and saved in portable document format (PDF) for analysis.

Ground Truth

Data entry (labeling) was outsourced to a specialized company, involving two professional data-entry workers working independently on all JOABPEQ forms. For the categorical questions on pages 1 and 2, a correct value was defined when there was agreement between the professional workers on the independently entered answer. In case of a discrepancy, the answer was reviewed by another supervisor, and a final decision was made by consensus. Regarding the VASs on page 3, two data entry workers independently measured the line lengths. The correct values were determined by averaging the measurements of the two evaluators. We excluded cases with missing answers, which prevented us from determining the final label. In total, we obtained 1305 JOABPEQ forms from the academic center and 483 from the community center.

Accordingly, a correct label was assigned to each question. Questions Q1-1 to Q2-5, Q3-1 to Q3-3, Q4-1, and Q5-1 offered two choices, labeled as 1 or 2; Q2-6, Q3-4, and Q3-5 provided three choices, labeled as 1, 2, or 3. The remaining questions (Q4-2, Q4-3, Q5-2 to Q5-7) offered five choices, with labels 1, 2, 3, 4, or 5.

Preprocessing

The PDFs of the JOABPEQ forms were converted into JPEG format for each page. All images were resized to 640 × 880 pixels (width × height) and normalized, with all pixel values rescaled to a range of 0–1 per image.

Model Construction

A fully connected (FC) layer was established for each categorical question and VASs. We began by extracting a 1000-dimensional feature from each page using a pre-trained CNN with ImageNet (**Figure 3**). The extracted feature was then passed through either each FC in the categorical questions or the FC in the VASs. Each FC associated with the categorical questions generated probabilities for each choice in every question, with the input dimension matching the extracted feature, and the output dimension corresponding to the number of choices in each question. However, the FC linked with the VASs produced prediction values for each VAS question, with the input dimension mirroring the extracted feature. Furthermore, the output dimension was three, corresponding to the number of VAS questions. Notably, each case was processed page by page throughout this procedure.

We trained the CNN and FC using five-fold cross-validation on the training set, thereby obtaining five models (Fold:1–Fold:5). The proposed DL model, which comprised five models, generated prediction probabilities by averaging the outputs of these five models.

We employed cross-entropy as the loss function for categorical questions and mean squared error for the numerical values associated with the VASs. Utilizing the Adam optimizer, we set the learning rate to 3.0×10^{-3} and the maximum number of training epochs to 200. Our DL model was constructed using the Python programming language (version 3.9.7) and Pytorch (version 1.11.0). For the CNN, we utilized EfficientNetB0 [20], which was pretrained with ImageNet. Our workstation was equipped with a Core i7-10710U 1.10-1.61 GHz (Intel) and a GeForce GTX 3090 (NVIDIA).

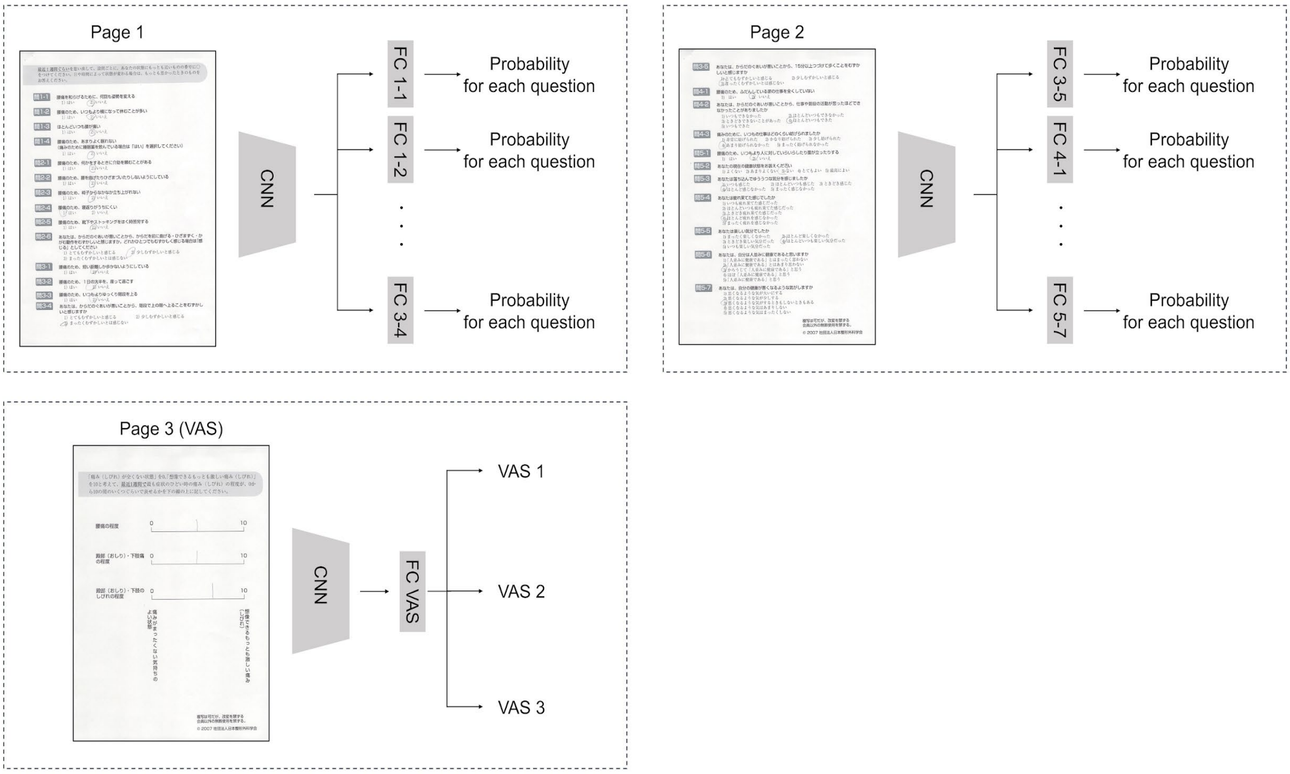

Page 1
CNN
FC 1-1
FC 1-2
FC 3-4
Probability for each question
Probability for each question
Probability for each question
Page 2
CNN
FC 3-5
FC 4-1
FC 5-7
Probability for each question
Probability for each question
Probability for each question
Page 3 (VAS)
CNN
FC VAS
VAS 1
VAS 2
VAS 3

Figure 3. The architecture of our deep learning (DL) model. Our model utilizes EfficientnetB0, which is a convolutional neural network (CNN) pre-trained with ImageNet, to extract features from each page. The extracted feature is passed to the fully connected layers (FC) for each categorical question on pages 1 and 2, or to an FC on page 3 to output each visual analog scale (VAS). FCs for categorical questions output numerical values with the same number of dimensions as the choices for each question, indicating the predicted probabilities for each choice. The FC for VAS generates three numerical values corresponding to the number of questions in the VAS.

Visualization of the Judgement of the DL Model

We employed gradient-weighted class activation mapping (Grad-CAM) [21] to visualize the decision-making process of the DL models (**Figure 4**). Grad-CAM visualizes the important pixels by weighting the gradient against the predicted value. We visualized the decision sites of the five models (Fold:1–Fold:5) using Grad-CAM, and integrated them by employing the maximum per-pixel values of the five models.

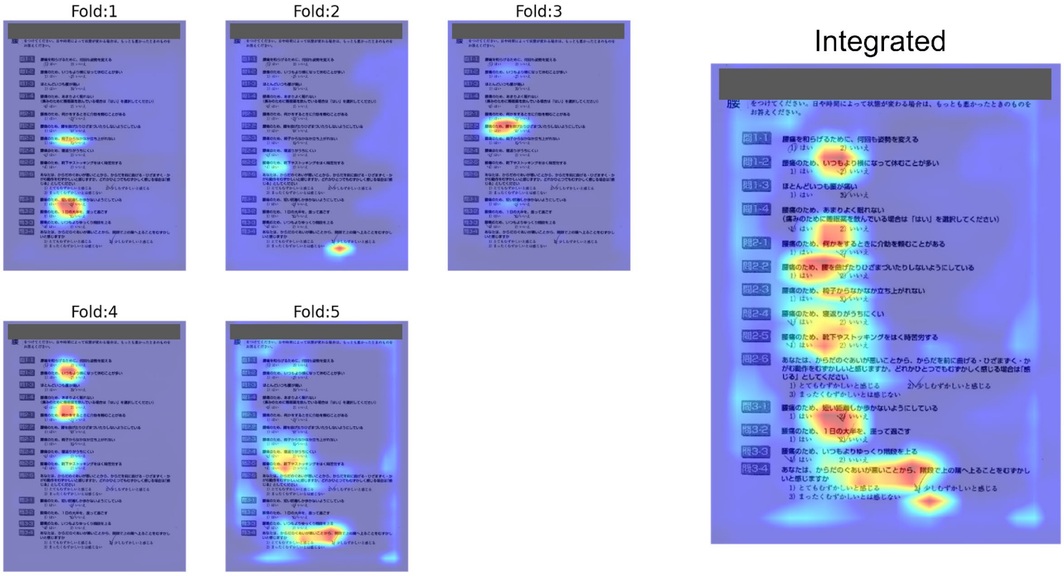


Figure 4. Heatmaps depicting the focus areas of DL models. The images are generated using Grad-CAM based on the trained DL models for one of the test set. Our DL model comprises five models (Fold:1~Fold:5), with the heatmaps on the left representing the focus areas of these models. The heatmap on the right is an integration of the five heatmaps on the left, calculated by employing the maximum per-pixel value of the five heatmaps. Colors signify the degree of activation, ranging from very high (red) and high (yellow) to low (green) and very low (blue).

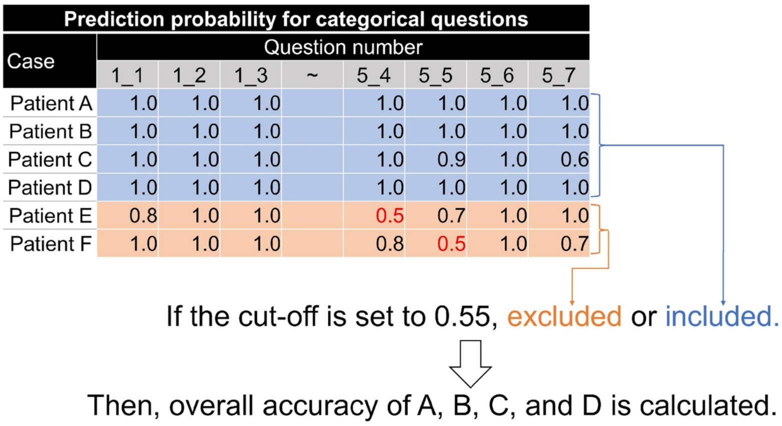
Performance Evaluation

We assessed the performance of our DL model on the test set, which consisted of 483 JOABPEQ forms collected from a community center, distinct from the academic center where the training set was gathered. Our proposed DL model generated the prediction probability by averaging the outputs of the five models derived from the five-fold cross-validation. The prediction label with the maximum probability was adopted as the final answer to the categorical questions. For example, for Q1-1 in **Figure 2A**, our DL model predicted that Label 1 had a probability of 0.002, with 0.998 for Label 2; consequently, Label 2 was adopted. The performance of our DL model was evaluated based on the accuracy of the categorical questions. Accuracy was defined as the number of correct predictions divided by the total number of predictions. Accordingly, we calculated the accuracy for each question, as well as for all the questions (overall accuracy).

For the VASs, the Pearson correlation coefficient (R-value) was calculated between the correct value (ground truth) and the predicted value. Additionally, the absolute error was computed by subtracting the correct value (ground truth) from the predicted value and taking the absolute value. To conduct this analysis, we utilized Scikit-Learn (version 0.24.2) and Scipy (version 1.10.1). Furthermore, we measured the processing speed of our model per page.

Automatic Exclusion System to Increase Accuracy

To improve the accuracy of the categorical questions, we devised an algorithm that automatically excluded cases containing questions having prediction probabilities below a certain cutoff value (**Figure 5**). We examined the relationship between the cut threshold and accuracy.



Prediction probability for categorical questions

Case	Question number							
	1_1	1_2	1_3	~	5_4	5_5	5_6	5_7
Patient A	1.0	1.0	1.0		1.0	1.0	1.0	1.0
Patient B	1.0	1.0	1.0		1.0	1.0	1.0	1.0
Patient C	1.0	1.0	1.0		1.0	0.9	1.0	0.6
Patient D	1.0	1.0	1.0		1.0	1.0	1.0	1.0
Patient E	0.8	1.0	1.0		0.5	0.7	1.0	1.0
Patient F	1.0	1.0	1.0		0.8	0.5	1.0	0.7

If the cut-off is set to 0.55, excluded or included.

Then, overall accuracy of A, B, C, and D is calculated.

Figure 5. The algorithm to improve the overall accuracy. We set a cut-off, removing cases where at least one question's prediction probability is lower than the cut-off value. The overall accuracy from Patients A to D is calculated. However, Patients E and F are excluded because they contain questions with a prediction probability below the cut-off, 0.55.

Results

Construction of Our Proposed DL model

We established our model with a five-fold cross-validation, indicating that we obtained five models (Fold:1–Fold:5). In the test dataset, the overall accuracy of the categorical questions was 0.994 for Fold:1, 0.993 for Fold:2, 0.996 for Fold:3, 0.986 for Fold:4, and 0.992 for Fold:5. By integrating these models (averaging the five-fold model prediction probabilities), our DL model achieved an impressive overall accuracy of 0.997 on the test dataset. This equates to three incorrect predictions out of 1000 questions.

Visualization of the Judgement of the DL Model

The Grad-CAM displayed the heatmaps of these five-fold models and our DL model (integrated) for the JOABPEQ form of a patient in the test set (**Figure 4**). These images illustrate that our proposed model made decisions based on the marks made by the patient on the choices.

Accuracy for Each type of Categorical Questions (Page 1 and 2 on the JOABPEQ form)

The overall accuracy of our DL model for all the categorical questions was 0.997. In contrast, the accuracy of the model for each type of categorical question ranged from 0.981 to 1.000 (**Table 1**). Specifically, questions featuring two choices exhibited a high accuracy of 0.999 (**Table 2**). However, questions offering five choices had a lower accuracy rate (0.992) than those with two choices.

Table 1. Accuracy of our proposed model for each categorical question (pages 1 and 2)

Question item number	Number of choices	Accuracy
Q 1-1	2	1.00
Q 1-2	2	1.00
Q 1-3	2	0.998
Q 1-4	2	0.998
Q 2-1	2	1.00

Q 2-2	2	1.00
Q 2-3	2	1.00
Q 2-4	2	0.998
Q 2-5	2	1.00
Q 2-6	3	0.994
Q 3-1	2	1.00
Q 3-2	2	1.00
Q 3-3	2	0.998
Q 3-4	3	1.00
Q 3-5	3	0.996
Q 4-1	2	1.00
Q 4-2	5	0.994
Q 4-3	5	0.981
Q 5-1	2	0.998
Q 5-2	5	0.990
Q 5-3	5	0.994
Q 5-4	5	0.994
Q 5-5	5	0.992
Q 5-6	5	0.994
Q 5-7	5	0.996
Overall (pages 1 and 2)	-	0.997

Table 2. Accuracy of our proposed model for each number of choices.

Number of choices	Accuracy
2	0.999
3	0.997
5	0.992
Overall (2, 3, and 5)	0.997

Automatic Exclusion System to Increase Accuracy

When the probability threshold was set to 0.695, 1 % of the JOABPEQ form was excluded, and the overall accuracy reached 0.9995 (**Figure 6**). In other words, the model got five out of 10,000 questions incorrectly. Additionally, when the probability threshold was set to 0.9996, 35 % of the JOABPEQ forms were excluded, and the overall accuracy was 1.000.

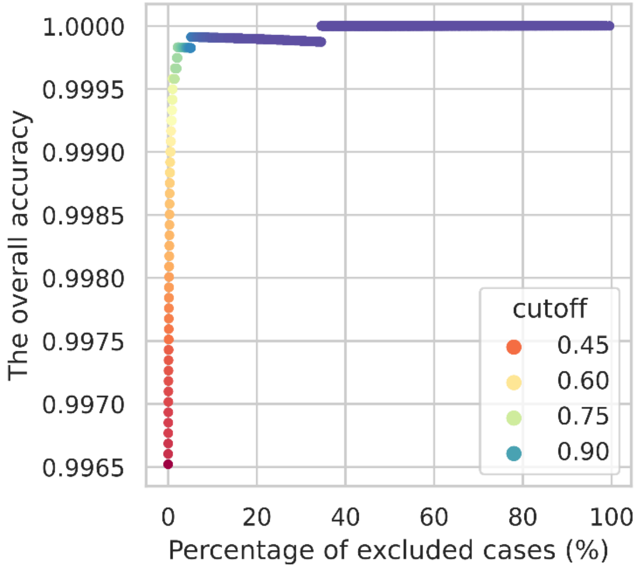


Figure 6. Relationship between the percentage of excluded questions and overall accuracy when the cut-off values of the prediction probability are changed. Increasing the value improves the accuracy, resulting in a higher the number of excluded questions.

VAS (Page 3 on the JOABPEQ form)

The Pearson correlation coefficients between the ground truth (corrected VAS) and DL prediction (Predicted VAS) for the three VASs were 0.985 (VAS_1), 0.991 (VAS_2), and 0.991 (VAS_3) (**Figure 7A**). The average of the three correlation coefficients was 0.989. The absolute errors for the VASs were 0.34, 0.25, 4.18, and 0 for the mean, median, maximum, and minimum values, respectively (**Figure 7B**).

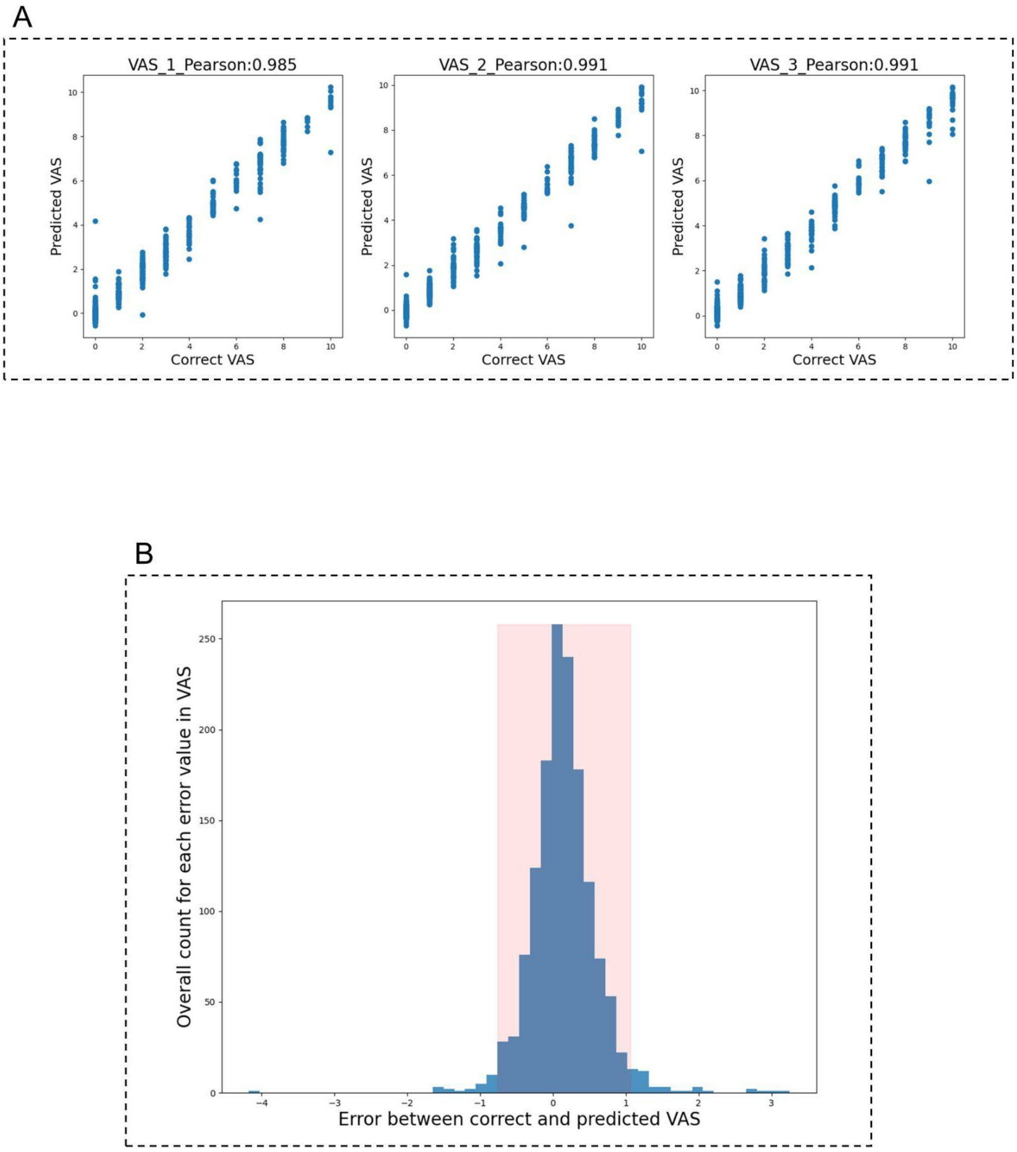


Figure 7. Predictive performance of our DL model in the VASs. The model predictions (Predicted VAS) were compared with the ground truth (Correct VAS). (A) Pearson correlation coefficient measured for VAS questions (Page 3-1, 3-2, and 3-3). (B) Histogram showing the overall count per error value including VAS_1, VAS_2, and VAS_3 count. The red zone represents a 95 % confidence interval, which ranged from – 0.76 to 1.06.

Processing Speed

The processing time for our DL model to handle 483 JOABPEQ forms was 787 s. In terms of time per page, the processing time was 0.88 s to read page 1, 0.69 s for page 2, and 0.06 s for page 3.

Case Presentation

Some patients marked their answers in unconventional ways. This posed challenges for our DL model, which struggled with these varying responses. Specifically, instances where a patient marked two choices (**Figure 8A**), marked multiple choices (**Figure 8B**), or used check marks (**Figure 8C**) presented difficulties for the model. Although the prediction probabilities of the model for these challenging responses were relatively low, it was able to output correct predictions even for certain challenging responses. Additionally, we improved the accuracy by setting a threshold (**Figure 6**).

For the VASs, our model was able to predict with an absolute error of less than 1, even if the style of marking was not according to the instructions, such as using a circle (**Figure 8D**) or a cross (**Figure 8E**). In contrast, the prediction of the model deviated from the correct VAS by an absolute error of two or more in cases where a patient used a number (**Figure 8F**) or made multiple marks (**Figure 8G**).

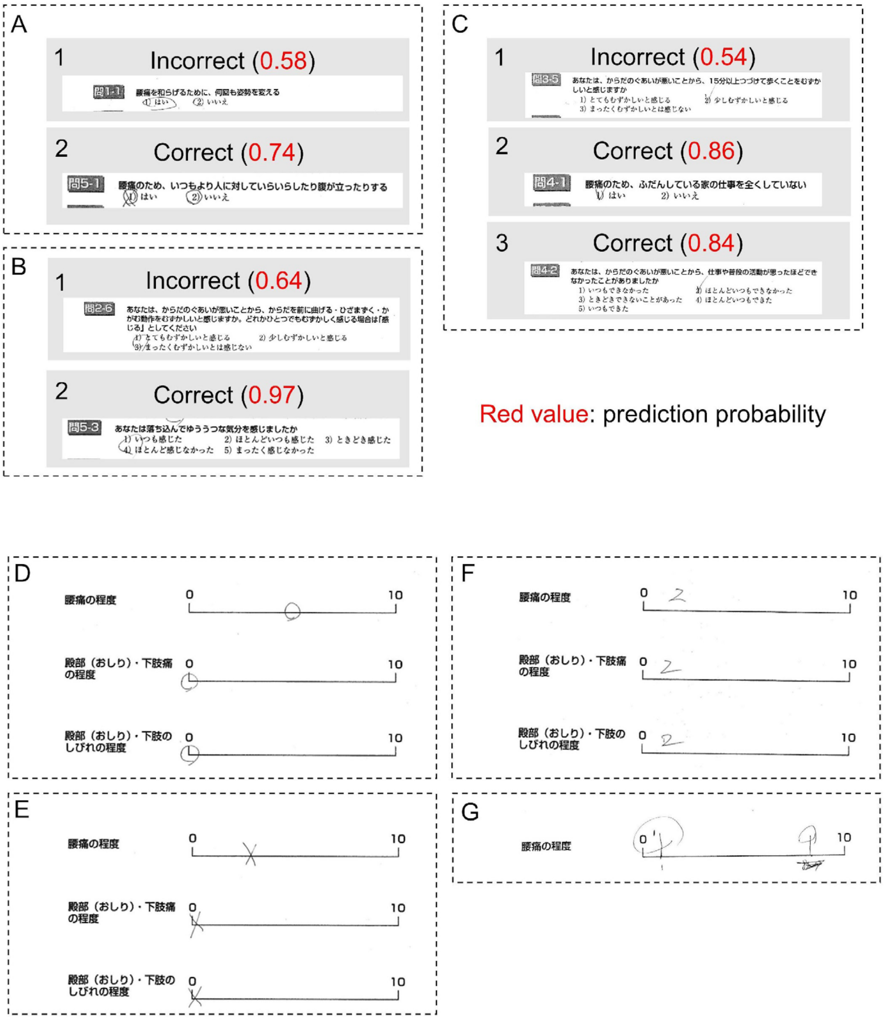

A
1 Incorrect (0.58)
問1-1
2 Correct (0.74)
問5-1
B
1 Incorrect (0.64)
問2-5
2 Correct (0.97)
問5-3
C
1 Incorrect (0.54)
問3-5
2 Correct (0.86)
問4-1
3 Correct (0.84)
問4-2
Red value: prediction probability
D
E
F
G
腰痛の程度
殿部（おしり）・下肢痛の程度
殿部（おしり）・下肢のしびれの程度
0
10

Figure 8. Illustration of challenging responses to the categorical questions and the VASs. Our DL model struggled with responses where a patient marked with (A) two choices, (B) a circle overlapping on two choices, and (C) check marks, in the categorical questions. Some of the model's predictions were correct, while others were incorrect. The model also struggled with responses in VASs marked with (D) circles, (E) crosses, (F) numbers, and (G) multiple marks. Notably, some of these responses were difficult to answer even for humans.

Discussion

This study aimed to develop a DL model for extracting data from paper-based HR-PRO forms. To the best of our knowledge, this is the first study to employ DL for automated data collection using an HR-PRO. Notably, our model achieved impressive results, with a high accuracy rate of 0.997 for the categorical questions. Moreover, a strong Pearson correlation coefficient of 0.989 was observed for the VASs. The performance of the model was evaluated through external validation, demonstrating its suitability for real-world clinical practice.

We manually collected data from the paper-based HR-PRO; however, this work was labor-intensive. The OCR is one traditional method for automatically retrieving data from paper-based HR-PRO. Teleform, a software that leverages OCR technology, has been widely used to collect data from paper forms in the healthcare sector [3]–[5]. It has demonstrated remarkable accuracy, ranging from 99.9 – 100 % when detecting circle marks in paper forms originally designed employing the Teleform designer program [3], [4]. However, when Teleform collects data from existing paper forms not generated by the Teleform program, the accuracy is low, ranging from 46.1 to 84.5 % [5]. For OCR to collect data from a form, it is necessary to recognize the positions of the responses. However, for forms not created by the Teleform program, the OCR struggles in identifying the positions. This leads to reduced accuracy, where the software often misinterprets populated fields as blank and incorrectly reads values from the forms [5]. In our study, we encountered difficulties with forms featuring multiple circle checks (**Figure 8A**) or circle checks spanning across several choices (**Figure 8B**). These complexities within the JOABPEQ forms, designed without specific considerations for OCR recognition, made it difficult for OCR to accurately identify response locations. In contrast, our model recognized not only a variety of marks but also the location of responses, which improved its accuracy. Initially, we attempted to establish an object detection model, such as YOLO [22], capable of detecting the responses to each question. However, this procedure requires an additional dataset to train the detection model, which is labor intensive. As a result, we developed an alternate strategy where the model reads an entire page, rather than focusing on each question, and extracts a feature from the page. By leveraging CNN to extract features from entire pages, we enabled the FC corresponding to each question to output a prediction probability (**Figure 3**), allowing the model to process an entire page in one step. Furthermore, it is difficult for OCR to accurately read the VAS, which is widely used to measure patient pain [23]. Multiple patients indicated their pain levels by drawing vertical lines on the form (**Figure 2A, Page 3**), and a rater measured the length of the line using the VAS. While the OCR could detect the lines, it could not measure the length accurately, a task that our DL model could effectively process. Thus, our model is superior to OCR in that it can accommodate atypical markings and distance measurements. Although not attempted in this study, the OCR output from an entire page could potentially be processed by a language model, such as bidirectional encoder representations from transformers [24], which is a topic for future work.

Our model demonstrates an impressive accuracy of 0.997 for overall categorical questions, surpassing the reported accuracies of previous studies involving manual data collection from paper forms, which ranged from 0.990 to 0.972 [25]–[27]. Instances where our model's prediction was wrong mostly corresponded to atypical responses, characterized by lower prediction probabilities compared to typical responses. This suggests that the model was unsure about its decision (**Figure 8**). These atypical cases can be difficult to identify even for humans. In such cases, a human can withhold input and check it directly with colleagues, supervisors, or patients. In contrast, our model cannot withhold the output, so it outputs a prediction regardless of its confidence level. To address this, it is necessary for the model to minimize wrong predictions stemming from low confidence levels as much as possible. In this study, we propose a method to prevent such unfounded confidence in the model by setting a threshold for the prediction probability (**Figure 5**). By adopting only outputs with a high prediction probability, it was possible to exclude incorrect predictions when the model was not confident, resulting in 100 % accuracy (**Figure 6**). For responses with a low prediction probability, human intervention is necessary to make the final decision.

Notably, while humans excelled in their ability to handle atypical cases, our model outperformed in terms of processing speed and reproducibility. The human processing speed averaged approximately 10 s per page for pages 1 and 2, and approximately 3 s per page for page 3. In contrast, our model processed pages 1 and 2 within 1 s each, and page 3 within 0.1 s. In addition, the model is superior in terms of reproducibility since it does not fatigue like humans.

We have published the program on GitHub (https://github.com/kosukekita/JOABPEQ-AI/tree/main) so that it can be widely used.

Limitation

This study had some limitations. First, the JOABPEQ forms used in this study were written in Japanese. Although an English version of the JOABPEQ is available online, our model was evaluated using only Japanese data. However, it is possible to create a model that can handle the English version by fine-tuning the model if an English dataset is available. Second, when our model made errors, the exact reasons for them were not always clear, which is a common challenge in DL approaches. However, we managed to visualize the basis for our model's decision using GradCAM (**Figure 4**).

Conclusion

We developed the DL model as a potential alternative to conventional OCR. The DL model demonstrated the capability to retrieve data from HR-PRO without a fixed answer field position, as seen in the JOABPEQ. The DL model is particularly effective for VAS. This system holds promise for future clinical studies.

References

[1] A. Cieza, K. Causey, K. Kamenov, S. W. Hanson, S. Chatterji, and T. Vos, "Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019," *Lancet*, vol. 396, no. 10267, pp. 2006–2017, Dec. 2021, doi: 10.1016/S0140-6736(20)32340-0.

[2] U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, and U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health, "Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance," *Health Qual Life Outcomes*, vol. 4, p. 79, Oct. 2006, doi: 10.1186/1477-7525-4-79.

[3] C. K. Jørgensen and B. Karlsmose, "Validation of automated forms processing," *Computers in Biology and Medicine*, vol. 28, no. 6, pp. 659–667, Nov. 1998, doi: 10.1016/S0010-4825(98)00038-9.

[4] C. Jinks, K. Jordan, and P. Croft, "Evaluation of a computer-assisted data entry procedure (including Teleform) for large-scale mailed surveys," *Computers in Biology and Medicine*, vol. 33, no. 5, pp. 425–437, Sep. 2003, doi: 10.1016/S0010-4825(03)00012-X.

[5] M. M. Wahi, D. V. Parks, R. C. Skeate, and S. B. Goldin, "Reducing Errors from the Electronic Transcription of Data Collected on Paper Forms: A Research Data Case Study," *Journal of the American Medical Informatics Association*, vol. 15, no. 3, pp. 386–389, May 2008, doi: 10.1197/jamia.M2381.

[6] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep Learning for Computer Vision: A Brief Review," *Computational Intelligence and Neuroscience*, vol. 2018, pp. 1–13, 2018, doi: 10.1155/2018/7068349.

[7] K. Kita, T. Fujimori, Y. Suzuki, S. Okada, and S. Kido, "Bi-modal network combining

convolutional neural network and TabNet, differentiating spinal tumors based on images and clinical risk factors," in *Medical Imaging 2023: Computer-Aided Diagnosis*, K. M. Iftekharuddin and W. Chen, Eds., San Diego, United States: SPIE, Apr. 2023, p. 4. doi: 10.1117/12.2646940.

[8] T. Fujimori *et al.*, "Development of artificial intelligence for automated measurement of cervical lordosis on lateral radiographs," *Sci Rep*, vol. 12, no. 1, p. 15732, Sep. 2022, doi: 10.1038/s41598-022-19914-x.

[9] K. Uemura *et al.*, "Development of an open-source measurement system to assess the areal bone mineral density of the proximal femur from clinical CT images," *Arch Osteoporos*, vol. 17, no. 1, p. 17, Dec. 2022, doi: 10.1007/s11657-022-01063-3.

[10] S. F. Abbasi *et al.*, "EEG-Based Neonatal Sleep-Wake Classification Using Multilayer Perceptron Neural Network," *IEEE Access*, vol. 8, pp. 183025–183034, 2020, doi: 10.1109/ACCESS.2020.3028182.

[11] S. F. Abbasi, Q. H. Abbasi, F. Saeed, and N. S. Alghamdi, "A convolutional neural network-based decision support system for neonatal quiet sleep detection," *MBE*, vol. 20, no. 9, pp. 17018–17036, 2023, doi: 10.3934/mbe.2023759.

[12] S. Farooq Abbasi, H. Jamil, and W. Chen, "EEG-Based Neonatal Sleep Stage Classification Using Ensemble Learning," *Computers, Materials & Continua*, vol. 70, no. 3, pp. 4619–4633, 2022, doi: 10.32604/cmc.2022.020318.

[13] Clinical Outcomes Committee of the Japanese Orthopaedic Association, Subcommittee on Evaluation of Back Pain and Cervical Myelopathy *et al.*, "JOA back pain evaluation questionnaire: initial report," *J Orthop Sci*, vol. 12, no. 5, pp. 443–450, Sep. 2007, doi: 10.1007/s00776-007-1162-x.

[14] M. Fukui *et al.*, "Japanese Orthopaedic Association Back Pain Evaluation Questionnaire. Part 2. Verification of its reliability : The Subcommittee on Low Back Pain and Cervical Myelopathy Evaluation of the Clinical Outcome Committee of the Japanese Orthopaedic Association," *J Orthop Sci*, vol. 12, no. 6, pp. 526–532, Nov. 2007, doi: 10.1007/s00776-007-1168-4.

[15] H. Hashizume *et al.*, "Japanese orthopaedic association back pain evaluation questionnaire (JOABPEQ) as an outcome measure for patients with low back pain: reference values in healthy volunteers," *J Orthop Sci*, vol. 20, no. 2, pp. 264–280, Mar. 2015, doi: 10.1007/s00776-014-0693-1.

[16] T. Fujimori, T. Miwa, and T. Oda, "Responsiveness of the Japanese Orthopaedic Association Back Pain Evaluation Questionnaire in lumbar surgery and its threshold for indicating clinically important differences," *Spine J*, vol. 19, no. 1, pp. 95–103, Jan. 2019, doi: 10.1016/j.spinee.2018.05.013.

[17] T. Poosiripinyo *et al.*, "The Japanese Orthopedic Association Back Pain Evaluation Questionnaire (JOABPEQ): A validation of the reliability of the Thai version," *J Orthop Sci*, vol. 22, no. 1, pp. 34–37, Jan. 2017, doi: 10.1016/j.jos.2016.10.001.

[18] P. Azimi, S. Shahzadi, and A. Montazeri, "The Japanese Orthopedic Association Back Pain Evaluation Questionnaire (JOABPEQ) for low back disorders: a validation study from Iran," *J Orthop Sci*, vol. 17, no. 5, pp. 521–525, Sep. 2012, doi: 10.1007/s00776-012-0267-z.

[19] A.-F. Zhou *et al.*, "Cross-cultural adaptation of The Japanese Orthopaedic Association Back Pain Evaluation Questionnaire: A methodological systematic review," *J Orthop Sci*, pp. S0949-2658(22)00234–2, Sep. 2022, doi: 10.1016/j.jos.2022.08.003.

[20] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." arXiv, Sep. 11, 2020. Accessed: Jun. 29, 2022. [Online]. Available: http://arxiv.org/abs/1905.11946

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2015, doi: 10.48550/ARXIV.1506.02640.

[23] M. M. Wertli *et al.*, "Validity of outcome measures used in randomized clinical trials and

observational studies in degenerative lumbar spinal stenosis," *Sci Rep*, vol. 13, no. 1, p. 1068, Jan. 2023, doi: 10.1038/s41598-022-27218-3.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, May 2019, Accessed: Mar. 23, 2022. [Online]. Available: http://arxiv.org/abs/1810.04805

[25] D. D. J. Smith, *Reliability, Maintainability and Risk: Practical Methods for Engineers including Reliability Centred Maintenance and Safety-Related Systems*, 7th ed. Amsterdam Boston Paris: Butterworth-Heinemann, 2005.

[26] X. Wu, C. Wu, K. Zhang, and D. Wei, "Residents' numeric inputting error in computerized physician order entry prescription," *International Journal of Medical Informatics*, vol. 88, pp. 25–33, Apr. 2016, doi: 10.1016/j.ijmedinf.2016.01.002.

[27] M. K. H. Hong *et al.*, "Error rates in a clinical data repository: lessons from the transition to electronic data transfer—a descriptive study," *BMJ Open*, vol. 3, no. 5, p. e002406, 2013, doi: 10.1136/bmjopen-2012-002406.

Acknowledgement

This work was supported by JSPS KAKENHI, Grant Numbers JP22K09353 and JP21K20966.