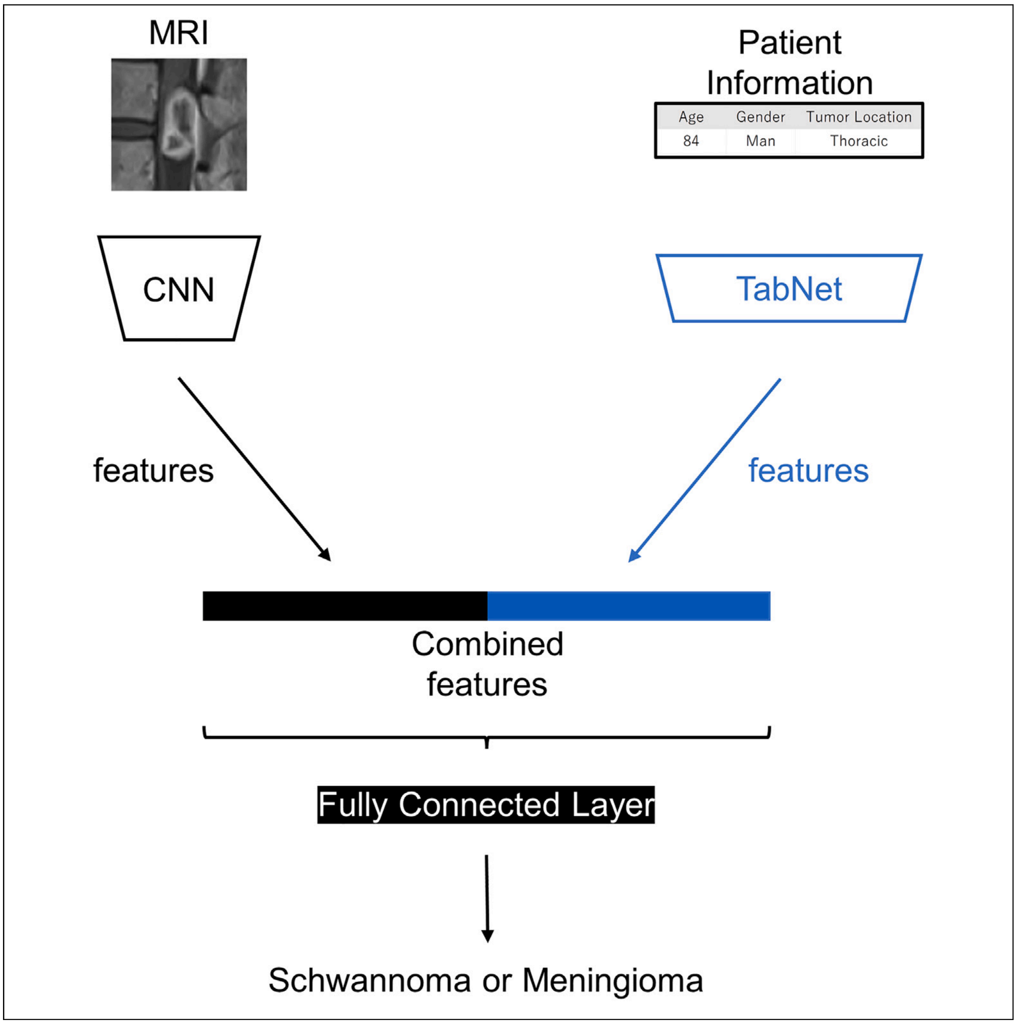
iScience

CellPress
OPEN ACCESS

Article

Bimodal artificial intelligence using TabNet for differentiating spinal cord tumors—Integration of patient background information and images



Kosuke Kita, Takahito Fujimori, Yuki Suzuki, ..., Noriyuki Tomiyama, Seiji Okada, Shoji Kido

kido@radiol.med.osaka-u.ac.jp

Highlights

Spinal cord tumors with 321 MRI and paired patient background information

A bimodal deep learning model for differentiating spinal cord tumors

We propose a bimodal model combining CNN and TabNet

The proposed bimodal model outperformed a conventional model using machine learning

Kita et al., iScience *26*, 107900
October 20, 2023 © 2023 The Authors.
https://doi.org/10.1016/j.isci.2023.107900

iScience

CellPress OPEN ACCESS

Article

Bimodal artificial intelligence using TabNet for differentiating spinal cord tumors—Integration of patient background information and images

Kosuke Kita,[1,16] Takahito Fujimori,[2] Yuki Suzuki,[1] Yuya Kanie,[2] Shota Takenaka,[2] Takashi Kaito,[2] Takuyu Taki,[3] Yuichiro Ukon,[2] Masayuki Furuya,[4] Hirokazu Saiwai,[5] Nozomu Nakajima,[6] Tsuyoshi Sugiura,[7] Hiroyuki Ishiguro,[8] Takashi Kamatani,[9] Hiroyuki Tsukazaki,[10] Yusuke Sakai,[11] Haruna Takami,[12] Daisuke Tateiwa,[13] Kunihiko Hashimoto,[14] Tomohiro Wataya,[1] Daiki Nishigaki,[1] Junya Sato,[1] Masaki Hoshiyama,[15] Noriyuki Tomiyama,[1] Seiji Okada,[2] and Shoji Kido[1,*]

SUMMARY

We proposed a bimodal artificial intelligence that integrates patient information with images to diagnose spinal cord tumors. Our model combines TabNet, a state-of-the-art deep learning model for tabular data for patient information, and a convolutional neural network for images. As training data, we collected 259 spinal tumor patients (158 for schwannoma and 101 for meningioma). We compared the performance of the image-only unimodal model, table-only unimodal model, bimodal model using a gradient-boosting decision tree, and bimodal model using TabNet. Our proposed bimodal model using TabNet performed best (area under the receiver-operating characteristic curve [AUROC]: 0.91) in the training data and significantly outperformed the physicians' performance. In the external validation using 62 cases from the other two facilities, our bimodal model showed an AUROC of 0.92, proving the robustness of the model. The bimodal analysis using TabNet was effective for differentiating spinal tumors.

INTRODUCTION

Spinal cord tumors that originate in the central nervous system can cause severe disruption to the daily lives of patients, and an estimated 70%–80% is intradural extramedullary in location.[1] Because 70% of intradural extramedullary tumors are schwannomas or meningiomas,[2,3] physicians often face the need to discriminate between the two in clinical practice. Differentiation between these tumors is important because the resection methods are different for each. However, differentiation is still challenging because the tumors are relatively rare and do not always present typical imaging findings.[4]

Some researchers have reported that artificial intelligence (AI) models aid in the diagnosis of spinal tumors.[5–9] Automated object detection and segmentation of spinal tumors yielded a high accuracy that was comparable to that of the physicians.[6,7,9] Besides these, classification models of spinal tumors have been established[8] and one of them was a model to discriminate between schwannomas and meningiomas.[5] In this study, magnetic resonance imaging (MRI) of 84 patients was analyzed, and the area under the receiver-operating characteristic curve (AUROC) was 0.88. Although it was reported that the preliminary AI model was comparable to physicians, we believe there is still room for the improvement of AI.

In daily practice, physicians do not always make a diagnosis based on images alone; in addition, they usually gain patient demographic information such as age and gender from a medical interview. Some studies have recently reported an AI model, the so-called *bimodal*

[1] Osaka University School of Medicine Graduate School of Medicine Diagnostic and Interventional Radiology, Suita, Osaka, Japan
[2] Osaka University Graduate School of Medicine Department of Orthopaedic Surgery, Suita, Osaka, Japan
[3] Department of Neurosurgery, Iseikai Hospital, Osaka, Osaka, Japan
[4] Osaka Rosai Hospital, Sakai, Osaka, Japan
[5] Department of Orthopedic Surgery, Graduate School of Medical Sciences, Kyusyu University, Higashi, Fukuoka, Japan
[6] Japanese Red Cross Society Himeji Hospital, Himeji, Hyogo, Japan
[7] General Incorporated Foundation Sumitomo Hospital, Osaka, Osaka, Japan
[8] National Hospital Organization Osaka National Hospital, Osaka, Osaka, Japan
[9] Toyonaka Municipal Hospital, Toyonaka, Osaka, Japan
[10] Kansai Rosai Hospital, Amagasaki, Hyogo, Japan
[11] Suita Municipal Hospital, Suita, Osaka, Japan
[12] Osaka International Cancer Institute, Osaka, Osaka, Japan
[13] Osaka General Medical Center, Osaka, Osaka, Japan
[14] Osaka Police Hospital, Osaka, Osaka, Japan
[15] JCHO Hoshigaoka Medical Center, Hirakata, Osaka, Japan
[16] Lead contact
*Correspondence: kido@radiol.med.osaka-u.ac.jp

https://doi.org/10.1016/j.isci.2023.107900

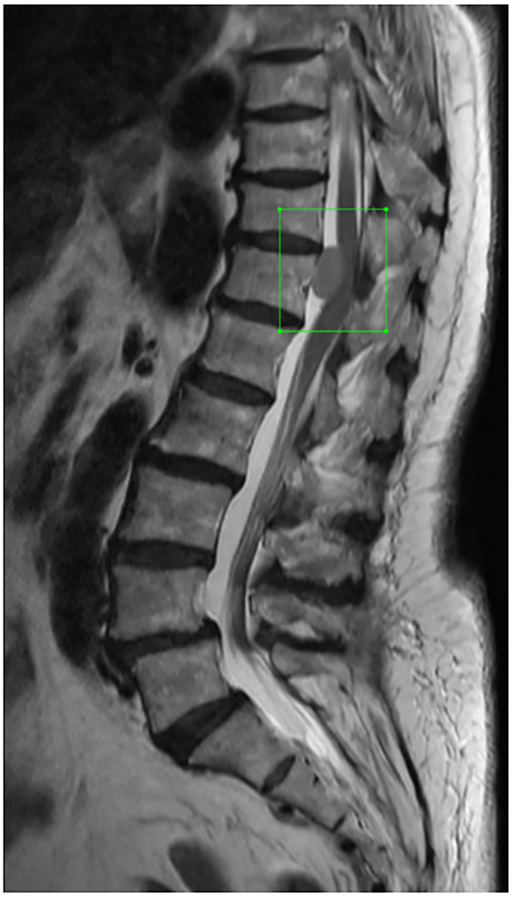


Figure 1. Image preprocessing for convolutional neural network model training
We selected the mid-slices of the tumor and cropped them to a minimal region containing the tumor (green square) on T2-weighted sagittal magnetic resonance images. The cropped images were used for training for and testing of the convolutional neural network models.

model, that can analyze data integrated from medical images and patient information.[10–12] If patient information is combined successfully, the bimodal model has the potential to outperform the previous unimodal model that analyzes images alone.[10–12]

Because patient information contains tabular data, traditional machine learning models such as gradient-boosted decision trees (GBDTs) have been used.[13–15] Furthermore, previous studies discussing bimodal AI in medicine used models combining GBDT and a convolutional neural network (CNN).[10,11] Recently, the Google Cloud AI research team invented TabNet as a state-of-the-art deep learning model for tabular data[16] and TabNet has been shown to be more effective than GBDT in a variety of tasks such as classification of forest cover type and the poker hand.[16,17] Additionally, TabNet has been applied in the medical field.[18,19] Because TabNet can extract latent features from patient information, we can establish an end-to-end bimodal model that combines TabNet with CNN.

In this study, we propose a bimodal model by combining TabNet with CNN and conduct three experiments to investigate the efficacy of the proposed bimodal model. The experiments seek to answer the following questions.

1. Is TabNet, a deep learning model, superior to GBDT, a machine learning model?
2. Is the bimodal model superior to the unimodal model?
3. Is the proposed bimodal model superior to physicians?

RESULTS

Demographics of the patient cohort

There were 158 patients with schwannomas (76 men and 82 women) and 101 with meningiomas (22 men and 79 women) who had undergone T2-weighted (T2WI) sagittal magnetic resonance (MR) images. We cropped the region containing the tumor as an input image from each T2WI sagittal slice (Figure 1). We used age, gender, and tumor location as patient information. The demographics of the patients are shown in Table 1. The mean age of patients with schwannoma was 58 years (age range, 42–74 years) and that of patients with meningioma

Table 1. Demographic data of patients with schwannoma and meningioma

	Schwannoma	Meningioma	p value
No. of patients	158	101	
Age (years)	58 ± 16	68 ± 14	<0.0001*
Gender (M/F)	76/82	22/79	<0.0001*
Location of tumor (%)			<0.0001*
Cervical	21.5	24.8	
Thoracic	36.1	68.3	
Lumbar	40.5	7.0	
Sacral	1.9	0	

*p < 0.05.

was 68 years (age range, 54–82 years). The women-to-men ratio was 1.1 in schwannoma and 3.6 in meningioma. We found the following characteristics in the patient information. In patients with meningioma, the mean age, proportion of women, and thoracic occurrence were higher than those in patients with schwannoma. The lumbar occurrence was higher in patients with schwannoma. The incidence ratios of schwannomas and meningiomas, as well as patients' background information, were consistent with previous reports.[5,20,21]

Advantage of TabNet over GBDT

We compared the performance of the unimodal model with TabNet (U_{Tab}) to that of the unimodal model with GBDT (U_{GBDT}). We also compared the performance of the bimodal model with TabNet (B_{Tab}; Figure 2A) to that of the bimodal model with GBDT (B_{GBDT}; Figure 2B). As a result, there was no significant difference in AUROC between U_{Tab} and U_{GBDT} (0.80 vs. 0.79; $p = 0.45$). However, B_{Tab} had significantly larger AUROC than did B_{GBDT} (0.91 vs. 0.88; $p = 0.03$). U_{Tab} outperformed U_{GBDT} in all metrics (U_{Tab} vs. U_{GBDT}; accuracy: 0.76 vs. 0.73, sensitivity: 0.75 vs. 0.75, specificity: 0.87 vs. 0.71, F1 score: 0.79 vs. 0.76), and B_{Tab} outperformed B_{GBDT} in all metrics except for specificity (B_{Tab} vs. B_{GBDT}; accuracy: 0.85 vs. 0.83, sensitivity: 0.84 vs. 0.78, specificity: 0.85 vs. 0.86, F1 score: 0.87 vs. 0.86) (Table 2).

Advantage of the bimodal model over the unimodal models

We compared the performance of the bimodal model to that of the unimodal models, namely the image-only (U_{img}: Table 3) and patient information-only unimodal models. Here we adopted U_{Tab} and B_{Tab} mentioned earlier to represent a patient information-only unimodal

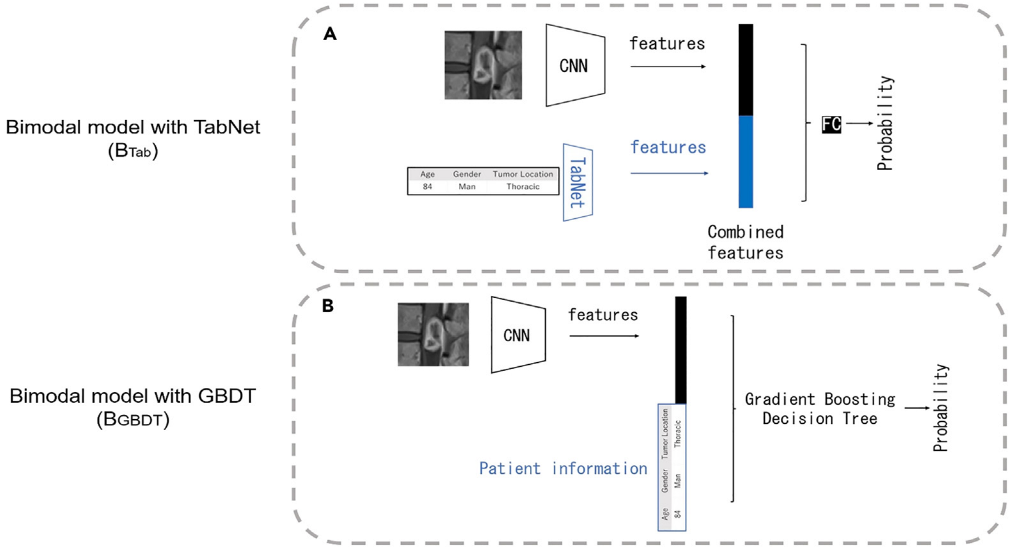


Figure 2. Proposed and baseline models

Proposed and baseline bimodal models.

(A) B_{Tab}: bimodal model with TabNet uses a convolutional neural network (CNN) to encode an image and TabNet to encode patient information. Two extracted features are combined and then passed to a fully connected (FC) layer to output probability.

(B) B_{GBDT}: bimodal model with gradient-boosting decision tree (GBDT) model uses a CNN to encode an image. The extracted feature of the image is concatenated to the raw patient information and then passed to a GBDT to output probability.

Table 2. Comparison of the performance among AI models

Model	Information	Algorithm	Accuracy	Sensitivity	Specificity	F1 score	AUROC	p value of AUROC
Unimodal	Patient demographic	GBDT (U_{GBDT})	0.73	**0.75**	0.71	0.76	0.79	U_{GBDT} VS. U_{Tab} $p = 0.45$
	Patient demographic	TabNet (U_{Tab})	**0.76**	**0.75**	**0.87**	**0.79**	**0.80**	U_{Tab} VS. B_{Tab} $p < 0.0001$*
	Image	CNN (U_{img})	0.81	0.73	0.85	0.84	0.84	U_{img} VS. B_{Tab} $p = 0.003$*
Bimodal	Patient demographics and image	GBDT (B_{GBDT})	0.83	0.78	**0.86**	0.86	0.88	B_{GBDT} VS. B_{Tab} $p = 0.03$*
	Patient demographics and image	TabNet (B_{Tab})	**0.85**	**0.84**	0.85	**0.87**	**0.91**	N.A.

AUROC: area under the receiver-operating characteristic curve; GBDT: gradient-boosted decision trees.
*$p < 0.05$. The Holm method was used.

model and a bimodal model, respectively. As a result, B_{Tab} (0.91) had larger AUROC than U_{img} (0.84; $p = 0.003$) and U_{Tab} (0.80; $p < 0.0001$). B_{Tab} outperformed others in metrics except for specificity (B_{Tab} vs. U_{img} vs. U_{Tab}; accuracy: 0.85 vs. 0.81 vs. 0.76, sensitivity: 0.84 vs. 0.73 vs. 0.75, specificity: 0.85 vs. 0.85 vs. 0.87, F1 score: 0.87 vs. 0.84 vs. 0.79) (Table 2).

Advantage of the bimodal model with TabNet over physicians

We ended by comparing the AI and physicians. Three radiologists and three spine surgeons were recruited to represent physicians. For the comparative analysis, we established eight B_{Tab} models using CNN from EfficientNetB0 to EfficientNetB7. We randomly split the whole dataset, namely MR images with patient information, into a training set and a test set at a 4:1 ratio. The training set was used for establishing eight AI models, and the test set was used for evaluating eight AI models and six physicians. As a result, the AUROC, accuracy, and specificity of the AI models were significantly larger than those of the physicians (AUROC: $p = 0.003$, accuracy: $p = 0.002$, specificity: $p = 0.007$) (Table 4). The intraclass correlation coefficient (ICC) for the six clinicians was 0.89, indicating "almost perfect".[22]

External validation

In the above comparison, we investigated the unimodal and bimodal models' performance with 5-fold cross-validation, which means splitting the dataset without separating the facility. This validation is called internal validation, which might include the possibility of overfitting. On the other hand, external validation uses test data from a completely different facility. To prove the robustness of the models, we also performed external validation. We collected additional 62 cases from other two facilities and tested the performance of our proposed models using these other cases (Table 5). Although there were no significant differences between the external validation and the internal validation in terms of age, gender, and the types of tumors (schwannoma/meningioma), there was a significant difference in tumor location.

In the external validation, the AUROCs were 0.77 for U_{GBDT} and 0.74 for U_{Tab} (Table 6), which were slightly lower than those of the internal validation (U_{GBDT}: 0.79, U_{Tab}: 0.80). As for the bimodal model, the AUROC was 0.87 for B_{GBDT} and 0.92 for B_{Tab}, which were comparable with those of the internal validation (B_{GBDT}: 0.88, B_{Tab}: 0.91). These results indicated the robustness of the bimodal models.

Case presentation

For reference, representative MR images of spinal cord tumors are presented in Figure 3.

DISCUSSION

We demonstrated the following points in this study. First, TabNet, which is a state-of-the-art deep learning model for tabular data, was effective in handling clinical data. Second, the bimodal model, which added patient information to images, was superior to the unimodal model, which handles only images. Third, the diagnostic ability of the proposed bimodal model was superior to that of experienced physicians. Another strength of this study is that we included the largest number of cases of AI-based research to distinguish between spinal cord tumors. The robustness of our model was tested in the external validation. Furthermore, the diagnosis (ground truth) was reliable because it was based on pathology in specimens taken at surgery, rather than consensus of image findings between physicians.

Table 3. Comparison of the image-only unimodal models' (U_{img}) performance

	EfficientNetB							
Number	0	1	2	3	4	5	6	7
AUROC	0.911	0.889	0.906	0.910	0.896	0.891	0.901	0.893

AUROC, area under the receiver-operating characteristic curve.

Table 4. Comparison of the performance between AI models (B_{Tab}) and physicians

	AI (B_{Tab})								Physician					
	EfficientNetB								Radiologist			Spine surgeon		
Number	0	1	2	3	4	5	6	7	1	2	3	1	2	3
AUROC	0.93 (0.86–1.0)	0.93 (0.84–1.0)	0.95 (0.87–1.0)	0.92 (0.83–1.0)	0.91 (0.81–1.0)	0.94 (0.86–1.0)	0.93 (0.86–1.0)	0.92 (0.84–1.0)	0.81 (0.69–0.92)	0.84 (0.72–0.96)	0.86 (0.75–0.97)	0.70 (0.55–0.85	0.81 (0.68–0.95)	0.91 (0.83–0.99)
Accuracy	0.88 (0.80–0.97)	0.88 (0.80–0.97)	0.90 (0.82–0.98)	0.90 (0.82–0.98)	0.90 (0.82–0.98)	0.90 (0.82–0.98)	0.92 (0.85–1.0)	0.88 (0.80–0.97)	0.73 (0.61–0.85)	0.79 (0.68–0.90)	0.79 (0.68–0.90)	0.60 (0.46–0.73)	0.67 (0.55–0.80)	0.85 (0.75–0.94)
Sensitivity	0.95 (0.85–1.0)	0.8 (0.62-0,98)	0.9 (0.77–1.0)	0.8 (0.62–0.98)	0.8 (0.62–0.98)	0.85 (0.69–1.0)	0.8 (0.62–0.98)	0.9 (0.77–1.0)	0.8 (0.62–0.98)	1.0 (1.0–1.0)	0.9 (0.77–1.0)	0.9 (0.77–1.0)	0.95 (0.85–1.0)	0.75 (0.56–0.94)
Specificity	0.84 (0.72–0.97)	0.94 (0.85–1.0)	0.91 (0.81–1.0)	0.97 (0.91–1.0)	0.97 (0.91–1.0)	0.94 (0.85–1.0)	1.0 (1.0, 1.0)	0.88 (0.76–0.99)	0.69 (0.53–0.85)	0.66 (0.49–0.82)	0.72 (0.56–0.87)	0.41 (0.24–0.58)	0.5 (0.33–0.67)	0.91 (0.81–1.0)

AI, artificial intelligence; AUROC, area under the receiver-operating characteristic curve.
The numbers in parentheses represent 95% confidence interval.

End-to-end bimodal model can fit the parameters efficiently

TabNet is a new deep learning model specialized for tabular data that can extract a subset of semantically meaningful features from tabular data such as patient information.[16] Although some researchers have studied TabNet, there are few reports in the medical field that have combined TabNet with a CNN.[18,19,23] According to previous studies, TabNet outperformed GBDT, the state-of-the-art machine learning model, in several datasets, although it was not superior in all datasets.[17] Their report may support our result that TabNet and GBDT were comparable in the unimodal analysis. In contrast, in the bimodal analysis, TabNet was superior to GBDT. The effectiveness of TabNet in bimodal analysis is supposed to be attributed to its end-to-end approach. When using TabNet, both the CNN and TabNet are simultaneously trained, allowing the CNN to incorporate patient information into the analysis. In contrast, when using the GBDT model in bimodal analysis, the parameters of the CNN are fixed, thereby limiting its ability to learn patient information. This difference in processing method may be the reason for TabNet's superior performance in our dataset compared to GBDT.

Patient demographic is useful for cases with tumors describing atypical intensities

In the present study, the image-only unimodal model (U_{img} accuracy: 0.81, AUROC: 0.84) was comparable to the previously reported image-only unimodal model (accuracy: 0.80, AUROC: 0.88) (Maki et al.). Our proposed bimodal model (B_{Tab}) outperformed both the unimodal models.

Physicians may consider patient information to make a diagnosis of spinal cord tumor; however, they rarely diagnose spinal cord tumors based on patient information alone. As indicated by the superior performance of U_{img} over U_{Tab} (Table 2), the importance of images in the

Table 5. Demographic data of patients in the external validation and internal validation

	External validation (62 cases from two facilities)	Internal validation (259 cases from 10 facilities)	p value
No. of patients	62	259	
Age (years)	61 ± 13	62 ± 16	0.81
Gender (M/F)	25/37	109/150	0.85
Schwannoma/meningioma	45/17	168/91	0.15
Location of tumor (%)			0.002*
Cervical	16.1	26.5	
Thoracic	40.3	48.2	
Lumbar	38.7	24.5	
Sacral	4.8	0.8	

*$p < 0.05$.

Table 6. Comparison of the AUROCs between external validation and internal validation

			AUROC	
Model	Information	Algorithm	External validation (62 cases from two facilities)	Internal validation (259 cases from 10 facilities)
Unimodal	Patient demographic	GBDT (U_{GBDT})	0.77	0.79
	Patient demographic	TabNet (U_{Tab})	0.74	0.80
	Image	CNN (U_{img})	0.86	0.84
Bimodal	Patient demographics and image	GBDT (B_{GBDT})	0.87	0.88
	Patient demographics and image	TabNet (B_{Tab})	0.92	0.91

diagnosis of spinal cord tumors is undisputed. Surprisingly, even in the absence of imaging information, U_{Tab} had a performance of 0.76 for accuracy and 0.80 for AUROC. This fact suggests that epidemiological information such as the common age of onset, the gender, and the favorite location of the tumor cannot be ignored.

Patient demographic information was useful, especially for tumors with atypical images. A typical image of schwannoma has hyper- or mixed-signal intensity on T2WIs[24] (Figure 3A). In this case, both B_{Tab} and U_{img} diagnosed correctly. Figure 3B shows an atypical image of schwannoma with iso-signal intensity, and U_{img} and half of the physicians could not differentiate such an atypical image correctly; however, B_{Tab} diagnosed it correctly. The reason B_{Tab} answered correctly is probably because the tumor occurred at the lumbar spine of a middle-aged male, which is consistent with the epidemiology of schwannomas. But, of course, epidemiological information can sometimes contradict images. Even in such cases, we think that B_{Tab}, a deep learning model, can successfully integrate patient information and images to make a diagnosis. This may be the same reason why an excellent physician had a higher accuracy rate. Although it is difficult to quantify the extent to which a physician relies on specific information when making a diagnosis, an excellent physician probably may be better at integrating imaging and patient information.

Conversely, a less excellent physician may be biased in terms of the cases he or she has experienced (the learning data for the physician). For example, a case in which a physician has recently misdiagnosed a patient may be more memorable to that physician; in contrast, a case that was correctly diagnosed long ago could be less notable. Humans have such recall biases, but AI has the advantage of being able to handle all case data without bias.

Limitations of the study

There were some limitations to the study. Firstly, it was difficult to perfectly align the reading conditions of the physicians and the AI. The physicians did not learn based on data trained by AI. At first glance, this appears to have put the physicians at a disadvantage. However,

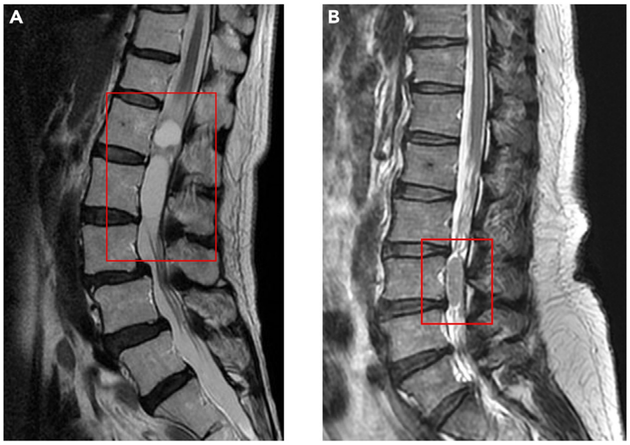


Figure 3. Representative T2-weighted magnetic resonance (MR) images that B_{Tab} (bimodal model with TabNet) diagnosed correctly

The red square describes a region containing the tumor.

(A) Typical T2-weighted MR image (hyper-signal intensity) of schwannoma in a patient with typical demographics (44-year-old male at lumbar spine). In this case, both the U_{img} (image-only unimodal model) and the B_{Tab} (bimodal model with TabNet) diagnosed correctly.

(B) Atypical T2-weighted magnetic resonance image (iso-signal intensity) of schwannoma in a patient with typical demographics (58-year-old male at lumbar spine). The U_{img} (image-only unimodal model) and three of the six physicians could not differentiate this case correctly. However, the B_{Tab} (bimodal model with TabNet) diagnosed correctly.

because spinal cord tumors are relatively rare, if a physician learned from the more than 200 cases of data that the AI used as training data this time, that physician would not be representative of physicians in regular practice. In addition, the physicians had experience gained from their previous practice, which the AI did not have, so the physicians were not generally at a disadvantage. Secondly, we did not include contrast-enhanced T1 weighted (T1WI) MRI because not all the patients undergoing T2WI sagittal MRI had undergone contrast-enhanced T1WI MRI and, thus, adding contrast-enhanced T1WI would induce the smaller dataset. Physicians typically make decision based on contrast-enhanced T1WI MRI as well as T2WI, therefore it can be challenging for both AIs and physicians to accurately differentiate spinal cord tumors based solely on T2WI MRI. Thirdly, we did not include intramedullary tumor such as ependymoma or astrocytoma. However, 70%–80% of spinal tumor is extramedullary in location,[1] and it is not hard for spine surgeons to tell a difference between intramedullary and extramedullary. Therefore, we focused on intradural extramedullary tumor.

Despite these limitations, AI achieved diagnostic capabilities comparable to or better than those of the experienced physicians, indicating the usefulness of bimodal analysis that combines images with patient information. This AI may be a useful tool to support diagnosis in the future.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Patient cohort
- METHOD DETAILS
  - ○ MR images and patient information preprocessing
  - ○ Overview of the bimodal models
  - ○ Overview of the unimodal models
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Evaluation of the AI models' performance
  - ○ Preparation of the dataset for image assessment by physicians
  - ○ Statistical and data analysis

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP22K09353.

AUTHOR CONTRIBUTIONS

Conceptualization, K.K., S.O., and S.K.; Methodology, K.K., Y.S., Y.K., S.T., T.K., T.T., Y.U., M.F., H.S., N.N., T.S., H.I., T.K., H.T., Y.S., H.T., D.T., K.H., M.H., and S.J.; Formal Analysis, K.K., Y.S., D.N., and T.W.; Writing – Original Draft, K.K.; Writing–Review & Editing, K.K., T.F., Y.S., and S.K.; Funding Acquisition, T.F. and S.K.; Resources, K.K.; Supervision, S.O. and S.K.

DECLARATION OF INTERESTS

Authors declare that they have no competing interests.

Received: December 20, 2022
Revised: February 18, 2023
Accepted: September 8, 2023
Published: September 14, 2023

REFERENCES

1. Özkan, N., Jabbarli, R., Wrede, K.H., Sariaslan, Z., Stein, K.P., Dammann, P., Ringelstein, A., Sure, U., and Sandalcioglu, E.I. (2015). Surgical management of intradural spinal cord tumors in children and young adults: A single-center experience with 50 patients. Surg. Neurol. Int. *6*, S661–S667. https://doi.org/10.4103/2152-7806.171236.

2. Hirano, K., Imagama, S., Sato, K., Kato, F., Yukawa, Y., Yoshihara, H., Kamiya, M., Deguchi, M., Kanemura, T., Matsubara, Y., et al. (2012). Primary spinal cord tumors: review of 678 surgically treated patients in Japan. A multicenter study. Eur. Spine J. *21*, 2019–2026. https://doi.org/10.1007/s00586-012-2345-5.

3. Ozawa, H., Onoda, Y., Aizawa, T., Nakamura, T., Koakutsu, T., and Itoi, E. (2012). Natural history of intradural-extramedullary spinal cord tumors. Acta Neurol. Belg. *112*, 265–270. https://doi.org/10.1007/s13760-012-0048-7.

4. Bhat, A.R., Kirmani, A.R., Wani, M.A., and Bhat, M.H. (2016). Incidence,

histopathology, and surgical outcome of tumors of spinal cord, nerve roots, meninges, and vertebral column - Data based on single institutional (Sher-i-Kashmir Institute of Medical Sciences) experience. J. Neurosci. Rural Pract. *7*, 381–391. https://doi.org/10.4103/0976-3147.181489.

5. Maki, S., Furuya, T., Horikoshi, T., Yokota, H., Mori, Y., Ota, J., Kawasaki, Y., Miyamoto, T., Norimoto, M., Okimatsu, S., et al. (2020). A Deep Convolutional Neural Network With Performance Comparable to Radiologists for Differentiating Between Spinal Schwannoma and Meningioma. Spine *45*, 694–700. https://doi.org/10.1097/BRS.0000000000003353.
6. Ito, S., Ando, K., Kobayashi, K., Nakashima, H., Oda, M., Machino, M., Kanbara, S., Inoue, T., Yamaguchi, H., Koshimizu, H., et al. (2021). Automated Detection of Spinal Schwannomas Utilizing Deep Learning Based on Object Detection From Magnetic Resonance Imaging. Spine *46*, 95–100. https://doi.org/10.1097/BRS.0000000000003749.
7. Lemay, A., Gros, C., Zhuo, Z., Zhang, J., Duan, Y., Cohen-Adad, J., and Liu, Y. (2021). Automatic multiclass intramedullary spinal cord tumor segmentation on MRI with deep learning. Neuroimage. Clin. *31*, 102766. https://doi.org/10.1016/j.nicl.2021.102766.
8. Zhuo, Z., Zhang, J., Duan, Y., Qu, L., Feng, C., Huang, X., Cheng, D., Xu, X., Sun, T., Li, Z., et al. (2022). Automated Classification of Intramedullary Spinal Cord Tumors and Inflammatory Demyelinating Lesions Using Deep Learning. Radiol. Artif. Intell. *4*, e210292. https://doi.org/10.1148/ryai.210292.
9. Ouyang, H., Meng, F., Liu, J., Song, X., Li, Y., Yuan, Y., Wang, C., Lang, N., Tian, S., Yao, M., et al. (2022). Evaluation of Deep Learning-Based Automated Detection of Primary Spine Tumors on MRI Using the Turing Test. Front. Oncol. *12*, 814667. https://doi.org/10.3389/fonc.2022.814667.
10. Tiulpin, A., Klein, S., Bierma-Zeinstra, S.M.A., Thevenot, J., Rahtu, E., Meurs, J.v., Oei, E.H.G., and Saarakkala, S. (2019). Multimodal Machine Learning-based Knee Osteoarthritis Progression Prediction from Plain Radiographs and Clinical Data. Sci. Rep. *9*, 20038. https://doi.org/10.1038/s41598-019-56527-3.
11. Han, X., Yu, Z., Zhuo, Y., Zhao, B., Ren, Y., Lamm, L., Xue, X., Feng, J., Marr, C., Shan, F., et al. (2022). The value of longitudinal clinical data and paired CT scans in predicting the deterioration of COVID-19 revealed by an artificial intelligence system. iScience *25*, 104227. https://doi.org/10.1016/j.isci.2022.104227.
12. Joo, S., Ko, E.S., Kwon, S., Jeon, E., Jung, H., Kim, J.-Y., Chung, M.J., and Im, Y.-H. (2021). Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer. Sci. Rep. *11*, 18800. https://doi.org/10.1038/s41598-021-98408-8.
13. Li, K., Yao, S., Zhang, Z., Cao, B., Wilson, C.M., Kalos, D., Kuan, P.F., Zhu, R., and Wang, X. (2022). Efficient gradient boosting for prognostic biomarker discovery. Bioinformatics *38*, 1631–1638. https://doi.org/10.1093/bioinformatics/btab869.
14. Arai, J., Aoki, T., Sato, M., Niikura, R., Suzuki, N., Ishibashi, R., Tsuji, Y., Yamada, A., Hirata, Y., Ushiku, T., et al. (2022). Machine learning-based personalized prediction of gastric cancer incidence using the endoscopic and histologic findings at the initial endoscopy. Gastrointest. Endosc. *95*, 864–872. https://doi.org/10.1016/j.gie.2021.12.033.
15. Seto, H., Oyama, A., Kitora, S., Toki, H., Yamamoto, R., Kotoku, J., Haga, A., Shinzawa, M., Yamakawa, M., Fukui, S., and Moriyama, T. (2022). Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. Sci. Rep. *12*, 15889. https://doi.org/10.1038/s41598-022-20149-z.
16. Arik, S.O., and Pfister, T. (2020). TabNet (Attentive Interpretable Tabular Learning).
17. Shwartz-Ziv, R., and Armon, A. (2021). Tabular Data: Deep Learning Is Not All You Need.
18. Khalili, E., Ramazi, S., Ghanati, F., and Kouchaki, S. (2022). Predicting protein phosphorylation sites in soybean using interpretable deep tabular learning network. Briefings Bioinf. *23*, bbac015. https://doi.org/10.1093/bib/bbac015.
19. Yu, Z., Ye, X., Liu, H., Li, H., Hao, X., Zhang, J., Kou, F., Wang, Z., Wei, H., Gao, F., and Zhai, Q. (2022). Predicting Lapatinib Dose Regimen Using Machine Learning and Deep Learning Techniques Based on a Real-World Study. Front. Oncol. *12*, 893966. https://doi.org/10.3389/fonc.2022.893966.
20. Tish, S., Habboub, G., Lang, M., Ostrom, Q.T., Kruchko, C., Barnholtz-Sloan, J.S., Recinos, P.F., and Kshettry, V.R. (2020). The epidemiology of spinal schwannoma in the United States between 2006 and 2014. J. Neurosurg. Spine *32*, 661–666. https://doi.org/10.3171/2019.10.SPINE191025.
21. Cao, Y., Jiang, Y., Liu, C., Jin, R., Jin, Z., Hong, X., Zhao, L., Zhao, G., and Wang, Y. (2021). Epidemiology and survival of patients with spinal meningiomas: A SEER analysis. Eur. J. Surg. Oncol. *47*, 2340–2345. https://doi.org/10.1016/j.ejso.2021.01.012.
22. Landis, J.R., and Koch, G.G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics *33*, 159–174. https://doi.org/10.2307/2529310.
23. Cahan, N., Marom, E.M., Soffer, S., Barash, Y., Konen, E., Klang, E., and Greenspan, H. (2023). Multimodal fusion models for pulmonary embolism mortality prediction. Review. https://doi.org/10.21203/rs.3.rs-2405211/v1.
24. Iwata, E., Shigematsu, H., Yamamoto, Y., Kawasaki, S., Tanaka, M., Okuda, A., Morimoto, Y., Masuda, K., Koizumi, M., Akahane, M., and Tanaka, Y. (2018). Preliminary algorithm for differential diagnosis between spinal meningioma and schwannoma using plain magnetic resonance imaging. J. Orthop. Sci. *23*, 408–413. https://doi.org/10.1016/j.jos.2017.11.012.
25. Tan, M., and Le, Q.V. (2020). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.
26. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Advances in Neural Information Processing Systems, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. (Curran Associates, Inc.).
27. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinf. *12*, 77. https://doi.org/10.1186/1471-2105-12-77.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Pyhton	https://www.python.org/	Version 3.9.7
Pytorch	https://pytorch.org/	Version 1.11.0
Scikit-learn	https://scikit-learn.org/	Version 1.0.2
Scipy	https://scipy.org/	Version 1.7.3
Statsmodels	https://www.statsmodels.org/	Version 0.12.2
R	https://www.r-project.org/	Version 4.2.2
pROC	https://github.com/cran/pROC	Version 1.18.0
lightgbm	https://lightgbm.readthedocs.io/en/v3.3.2/#	Version 3.3.2
Proposed Bimodal model	https://github.com/kosukekita/bimodal_AI_spinal_tumor	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for model implementation should be directed to and will be fulfilled by the lead contact, Kosuke Kita (k-kita@radiol.med.osaka-u.ac.jp).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Due to privacy concerns and ethical considerations, the data collected and analyzed in this study cannot be publicly shared.
- Source code is publicly available online. The URL is listed in key resources table.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Patient cohort

The institutional review board of the blinded institution approved this study, and the requirement for consent was waived because of its retrospective nature. We retrospectively reviewed the medical lists of patients with spinal cord tumors who had undergone tumor resection at two academic medical centers and eight community medical centers from October 1, 2010 to April 30, 2022. Diagnoses were made based upon the histology of the surgical specimens. There were 259 patients (Man/Female: 109/150), containing 158 patients with schwannoma and 101 patients with meningioma (Table 5).

We also collected addtional 62 patients (25 men and 37 women, 45 schwannomas and 17 meningiomas) who had undergone tumor resection at other two facilities from October 1, 2010 to April 30, 2022 for external validation (Table 5).

All participants in this study are Japanese.

METHOD DETAILS

MR images and patient information preprocessing

We targeted T2WI sagittal MR images and tabular data of the patients' demographics. An orthopedic surgeon (K.K., 8 years of experience) selected a few mid-slices of a tumor manually (ranging from one to three slices per case) from sagittal MR images in order to capture the tumor in a larger area: there was one slice for small tumors, and there were multiple slices for large tumors near the maximum area. A total of 265 slices were obtained from 158 patients with schwannoma, and 164 slices were obtained from 101 patients with meningioma. We cropped a minimal region containing a tumor with a margin of approximately 1 cm around the tumor on each slice as a region of interest for CNN training (Figure 1). All regions of interest were resized to 224 × 224 pixels, with all pixel values rescaled to a range of zero to one per image.

Patient information comprised three features: age, gender, and tumor location. The tumor location was a categorical variable with four possible values: cervical, thoracic, lumbar, and sacrum. We one-hot encoded the tumor location, and patient information was represented as a six-dimensional feature vector.

Overview of the bimodal models

We proposed a bimodal model with TabNet (B_{Tab}) to differentiate spinal tumors based on integrated MR images and patient information (Figure 2A). The features of the MR images were obtained through a CNN, and those of the patient information was obtained through TabNet. Finally, both features were combined and passed to a fully connected layer to output probability. In addition to B_{Tab}, we established the conventional bimodal model integrating a CNN and a GBDT (B_{GBDT}). We concatenated the feature for the MR images obtained through the CNN and the raw patient information. We then passed it to the GBDT to output probability.

To compare each AI model, we used EfficientNetB0 as a representative of the CNN that had been pre-trained with ImageNet, because EfficientNetB0 performed best among EfficientNetB0-B7 in the cross-validation evaluation comparing AI models (Table 3). For comparison between the AI models and physicians, we adopted eight models of EfficientNet from B0 to B7.[25]

Overview of the unimodal models

We built the following three unimodal baseline models: one analyzing only images (U_{img}) and two models analyzing only patient information (U_{Tab} and U_{GBDT}). As for $U_{img,}$ U_{Tab}, and U_{GBDT}, we built and trained the CNN, TabNet, and GBDT, respectively. We used EfficientNetB0 as the CNN in U_{img} and LightGBM[26] as the GBDT.

QUANTIFICATION AND STATISTICAL ANALYSIS

Evaluation of the AI models' performance

The performance of the two bimodal (B_{Tab} and B_{GBDT}) and the three unimodal models (U_{img}, U_{Tab}, and U_{GBDT}) was evaluated with 5-fold cross-validation. The dataset was split into five non-overlapping folds. The split was performed at patient level to avoid data leakage. In cases where multiple slices were obtained from MRI, the final decision was made based on the mean probability over all slices in the models analyzing MR images (B_{Tab}, B_{GBDT}, and U_{img}). We compared U_{Tab} with U_{GBDT} and B_{Tab} with B_{GBDT} to investigate the performance of TabNet. We also compared the B_{Tab}, U_{img}, and U_{Tab} to investigate the performance of the bimodal model.

Preparation of the dataset for image assessment by physicians

To compare the performance of B_{Tab} with that of physicians, all cases were randomly split into a training set and a test set at a ratio of 4:1. Thus, in the training set, there were 126 cases of schwannomas and 81 cases of meningiomas, and in the test set, there were 32 were cases of schwannomas and 20 cases of meningiomas. All cases included MR images and patient information (age, sex, tumor location). B_{Tab} learned with the training set and was evaluated with hold-out validation.

Three board-certified spine surgeons (K.T., F.T., and K.Y., with 23, 18, and 11 years of experience, respectively) and three board-certified radiologists (W.T., S.J., and T.M., with seven, three, and three years of experience, respectively) participated in this study. All six physicians and B_{Tab} reviewed the test set. For each case, each physician made a diagnosis of whether it was a schwannoma or a meningioma and graded the cases using a stepwise index of 25% increments for suspicion. For example, "meningioma at 75%" meant that the physician believed there was a 75% chance that the case was a meningioma (25% chance of schwannoma).

Because reading only a cropped image is not a typical clinical setting for physicians to form a diagnosis, the physician could refer to the entire image to provide a condition similar to the usual reading environment.

Statistical and data analysis

All statistical analyses were conducted using Scipy 1.7.3, Statsmodels 0.12.2, R 4.2.2, and pROC[27] 1.18.0. As for patient demographic data, parametric variables were evaluated using a Student's *t* test, and categorical variables were evaluated using a χ^2 test. We defined AUROC as the primary outcome to indicate the performance of each model. For comparisons between AI models, the DeLong test was performed. The accuracy, sensitivity, specificity, and F1 score were also calculated to evaluate the AI models' performance. These parameters were calculated at the optimal cut-off point of the ROC curve. For comparisons of AUROC, accuracy, sensitivity, and specificity between the AI models and the physicians, we used the Wilcoxon rank-sum test. We calculated single measures ICC for the six physicians using pingouin 0.5.3. All statistical tests were two-sided, and p values <0.05 were considered statistically significant. The Holm method was used for multiple comparisons.

The learning environment was as follows: Python 3.9.7, Pytorch 1.11.0, PyTorch-Lightning 1.7.0, scikit-learn 1.0.2.