USCAP

ARTICLE

Check for updates

Deep convolutional neural network-based algorithm for muscle biopsy diagnosis

Yoshinori Kabeya[1], Mariko Okubo[2], Sho Yonezawa[1], Hiroki Nakano[1], Michio Inoue[2], Masashi Ogasawara[2], Yoshihiko Saito[2], Jantima Tanboon[2], Luh Ari Indrawati[2], Theerawat Kumutpongpanich[2], Yen-Lin Chen[2], Wakako Yoshioka[2], Shinichiro Hayashi[2], Toshiya Iwamori[1], Yusuke Takeuchi[3], Reitaro Tokumasu[1], Atsushi Takano[1], Fumihiko Matsuda[4] and Ichizo Nishino[2]✉

Histopathologic evaluation of muscle biopsy samples is essential for classifying and diagnosing muscle diseases. However, the numbers of experienced specialists and pathologists are limited. Although new technologies such as artificial intelligence are expected to improve medical reach, their use with rare diseases, such as muscle diseases, is challenging because of the limited availability of training datasets. To address this gap, we developed an algorithm based on deep convolutional neural networks (CNNs) and collected 4041 microscopic images of 1400 hematoxylin-and-eosin-stained pathology slides stored in the National Center of Neurology and Psychiatry for training CNNs. Our trained algorithm differentiated idiopathic inflammatory myopathies (mostly treatable) from hereditary muscle diseases (mostly non-treatable) with an area under the curve (AUC) of 0.996 and achieved better sensitivity and specificity than the diagnoses done by nine physicians under limited diseases and conditions. Furthermore, it successfully and accurately classified four subtypes of the idiopathic inflammatory myopathies with an average AUC of 0.958 and classified seven subtypes of hereditary muscle disease with an average AUC of 0.936. We also established a method to validate the similarity between the predictions made by the algorithm and the seven physicians using visualization technology and clarified the validity of the predictions. These results support the reliability of the algorithm and suggest that our algorithm has the potential to be used straightforwardly in a clinical setting.

Laboratory Investigation (2022) 102:220–226; https://doi.org/10.1038/s41374-021-00647-w

INTRODUCTION

The diagnosis, management, and further study of rare diseases, such as muscle diseases, carry fundamental challenges that are different from those of common diseases, owing to fewer patients and limited expert facilities and clinicians[1]. Novel digital tools, such as artificial intelligence (AI), are expected to circumvent these shortcomings by accelerating the processes of diagnosis, specialist referrals, gathering and sharing of data, and clinical research on rare diseases[2]. Deep learning is a highly reliable AI technology used for specific tasks[3]; analyzing medical and pathological images using deep learning[4–9] is comparable to that by human experts[4–8]. However, thus far, almost all deep learning-based medical image analyses have only dealt with common diseases[4–9] owing to the limited data available on rare diseases. Previous studies using muscle magnetic resonance imaging and AI scored conditions manually and did not diagnose them using the images directly[10,11].

In this study, we aimed to diagnose muscle diseases with histopathology. Evaluating muscle pathology is unique because muscle biopsy specimens require freeze-fixation and a different set of histochemical staining techniques from general pathology. Classifying muscle diseases according to their pathological features remains diagnostically relevant even in today's era of molecular diagnoses[12].

Previous studies required whole-slide pathological images[13,14] and important infrastructural investments. However, real-world pathological diagnoses are performed with analog microscopes. Developing a diagnostic tool with a charge-coupled device (CCD) camera that is much cheaper than a whole-slide scanner and can recruit analog microscope images is the key to establishing a practically applicable system, especially in underserved areas worldwide.

To address this issue, we developed a deep learning, convolutional neural network (CNN)-based algorithm that could differentiate between major muscle diseases using a small amount of training data. To train and evaluate the algorithm, we collected microscopic images of hematoxylin and eosin (H&E)-stained pathological slides that were obtained by CCD cameras. The underlying algorithmic architecture for classifying dominant muscular dystrophies with whole-slide images has been previously established[15].

MATERIALS AND METHODS

Target muscle diseases

We chose 11 muscular diseases: 4 idiopathic inflammatory myopathies (IIM) [dermatomyositis (DM), inclusion body myositis (IBM), immune-mediated necrotizing myopathy (IMNM), and antisynthetase syndrome

[1]IBM Japan Ltd., Tokyo, Japan. [2]Department of Neuromuscular Research, National Institute of Neuroscience, National Center of Neurology and Psychiatry, Tokyo, Japan. [3]Watson Health, IBM Corporation, Cambridge, MA, USA. [4]Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan. ✉email: nishino@ncnp.go.jp

Received: 4 May 2021 Revised: 7 July 2021 Accepted: 9 July 2021
Published online: 2 October 2021

Table 1. Summary of studies, diseases, and number of pathological images and slides.

Group	Disease	Images/ Slides
Idiopathic inflammatory myopathy (IIM)	Dermatomyositis (DM)	307/103
	Inclusion body myositis (IBM)	306/102
	Immune-mediated necrotizing myopathy (IMNM)	332/109
	Antisynthetase syndrome (ASS)	317/107
Non-myositis muscle disease	Dystrophinopathy (DYST)	348/126
	Fukuyama-type congenital muscular dystrophy (FCMD)	256/95
	Limb-girdle muscular dystrophy 2A (LG2A)	321/120
	Ullrich congenital muscular dystrophy (UCMD)	213/83
	Limb-girdle muscular dystrophy 2B (LG2B)	512/177
	Congenital myopathy (CM)	629/211
	GNE myopathy (GNEM)	181/60
Others	Neuropathy (NP)	319/107
Total		4041/ 1400

The number of images obtained from one slide was ~3, but it could vary (1–6) depending on the size of the tissue on the slide.

(ASS)] and 7 hereditary muscle diseases [dystrophinopathy (DYST), limb-girdle muscular dystrophy 2A (LGMD2A), limb-girdle muscular dystrophy 2B (LGMD2B), Ullrich congenital muscular dystrophy (UCMD), Fukuyama-type congenital muscular dystrophy (FCMD), congenital myopathy (CM), and GNE myopathy (GNEM)], and neuropathy (Table 1). CM included nemaline myopathy, central core disease, and centronuclear myopathy.

Approach

We employed a two-step approach to distinguish between the diseases: 1) differentiating IIM from other conditions, and 2) subclassifying each category because most IIM conditions are treatable but other hereditary conditions are not (Fig. 1a). In the first step, we combined images of DM, IBM, IMNM, and ASS to create the IIM group, and those of DYST, FCMD, LGMD2A, LGMD2B, UCMD, CM, GNEM, and NP to create the counterpart group. We used a holdout method[16] to train and evaluate the CNNs and compare them to human physicians. In the second step, we subclassified four IIM subtypes and seven hereditary muscle diseases. We evaluated the results using five-fold cross-validation[16].

We also developed a visualization method using Grad-cam[17] to check the prediction accuracy of the CNNs. We used it to create AI-focused images (Fig. 1c and f) that masked the areas that were not evaluated by the CNNs and AI-unfocused images (Fig. 1d and g) that masked the areas to be targeted by the CNNs. We then identified the image group that physicians could use to obtain correct diagnoses and investigated the relationship between the predictions made by the CNNs and the physicians' diagnoses.

Dataset

We utilized H&E-stained frozen muscle sections on glass slides that were mounted for diagnostic purposes in the National Center of Neurology and Psychiatry between 1981 and 2019. All materials used in the present study were obtained for diagnostic purposes with written informed consent. The Ethics Committee of the National Center of Neurology and Psychiatry approved the study. The samples had already been evaluated by immunostaining, western blot, blood biochemistry, and genetic testing.

Data preparation

Images were taken from slides with CCD cameras (DP72/DP74, Olympus, Tokyo) attached to the microscope. The objective lens magnification was 4×, and the image size was 1024 × 1360 or 1600 × 1200. Three pathological images were taken per slide to maximize the coverage of the sample and minimize overlaps between them. If the sample size was small enough to fit into the finder field, only one image was taken.

By contrast, more than three shots were obtained if the sample size was too large. In total, 4041 images were obtained from 1400 slides, with one slide per patient. When the training and validation sets were created, two image patches (1024 × 1024) were cropped from both the left and right edges of the images for data augmentation. Meanwhile, an image patch (1024 × 1024) was cropped from the image center when the test sets were created.

CNN design

The CNN architecture developed for this study is illustrated in Fig. 1h. The CNNs resized the input images from 1024 × 1024 to 640 × 640, divided an imported image into sixteen 160 × 160 image patches, and generated a feature map per patch. The 16 feature maps were concatenated and used for the prediction of each image; we expected that the CNN could be trained effectively even with a small number of images. All feature maps were concatenated, averaged by global average-pooling, and used to generate the final probability. At the same time, each feature map was used to generate the probability per image patch. In the test phase, only the final probability was used for the prediction. In the training phase, the final probability and probabilities of image patches were used to calculate the loss function (L) as:

$$L = -\sum_{i=1}^{M}\left(y_i\left(\log\ \boldsymbol{p}_i^{final} + \sum_{j=1}^{T}\log\ \boldsymbol{p}_j^{image\ patch}\right)\right)$$

where M is the number of classes, T is the number of patches, $\boldsymbol{y}$ is a one-hot vector (true class is 1 and others are 0), $\boldsymbol{p}^{final}$ is a vector of the probability of final prediction, and $\boldsymbol{p}^{image\ patch}$ is a vector of the probability of prediction by image patch.

In this study, a densely connected convolutional network[18], called DenseNet, was used as a feature generator. DenseNet generally consisted of dense blocks and transition layers, with each dense block having two convolutional layers and a concatenation layer, and each transition layer having a convolutional layer and an average-pooling layer. Our DenseNet had 58 dense blocks and three transition layers, and it was pretrained with ImageNet[19], an extensive image database.

Training and testing the algorithm for the IIM differentiation task

For this task, we adopted the holdout method that divides the data into training and test sets. We randomly selected 96 slides as the test set and used the remaining 1304 slides as the training set. In the training phase, the training set was divided into five groups; four were used for training and one to validate the progress of the training set. We conducted the training five times while shifting the evaluation set one by one to train five CNNs (Fig. 1i). In the test phase, we used a model ensemble method[20] that averaged the probabilities of images cropped from the same slide to obtain the image probability by one slide and also averaged the output probabilities of the five CNNs. The probability averaging function is:

$$\boldsymbol{p}_{slide} = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{p}_{image}^{n}\right),$$

where n is the number of images per slide, $\boldsymbol{p}_{slide}$ is a vector of the probability of a slide, and $\boldsymbol{p}_{image}$ is a vector of the probability of an image. The class with the highest probability was adopted as the final prediction.

The following is the function of averaging the probabilities of models:

$$\boldsymbol{p}_{image}^{n} = \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{p}_{image}^{m,n}\right),$$

where m is the number of models, and $\boldsymbol{p}_{image}^{m,n}$ is a vector of the probability of an image outputted by a model. Furthermore, $\boldsymbol{p}_{slide}$ is:

$$\boldsymbol{p}_{slide} = \frac{1}{n}\frac{1}{m}\sum_{i=1}^{n}\sum_{i=1}^{m}\left(\boldsymbol{p}_{image}^{m,n}\right).$$

The disease with the highest probability was selected as the final prediction.

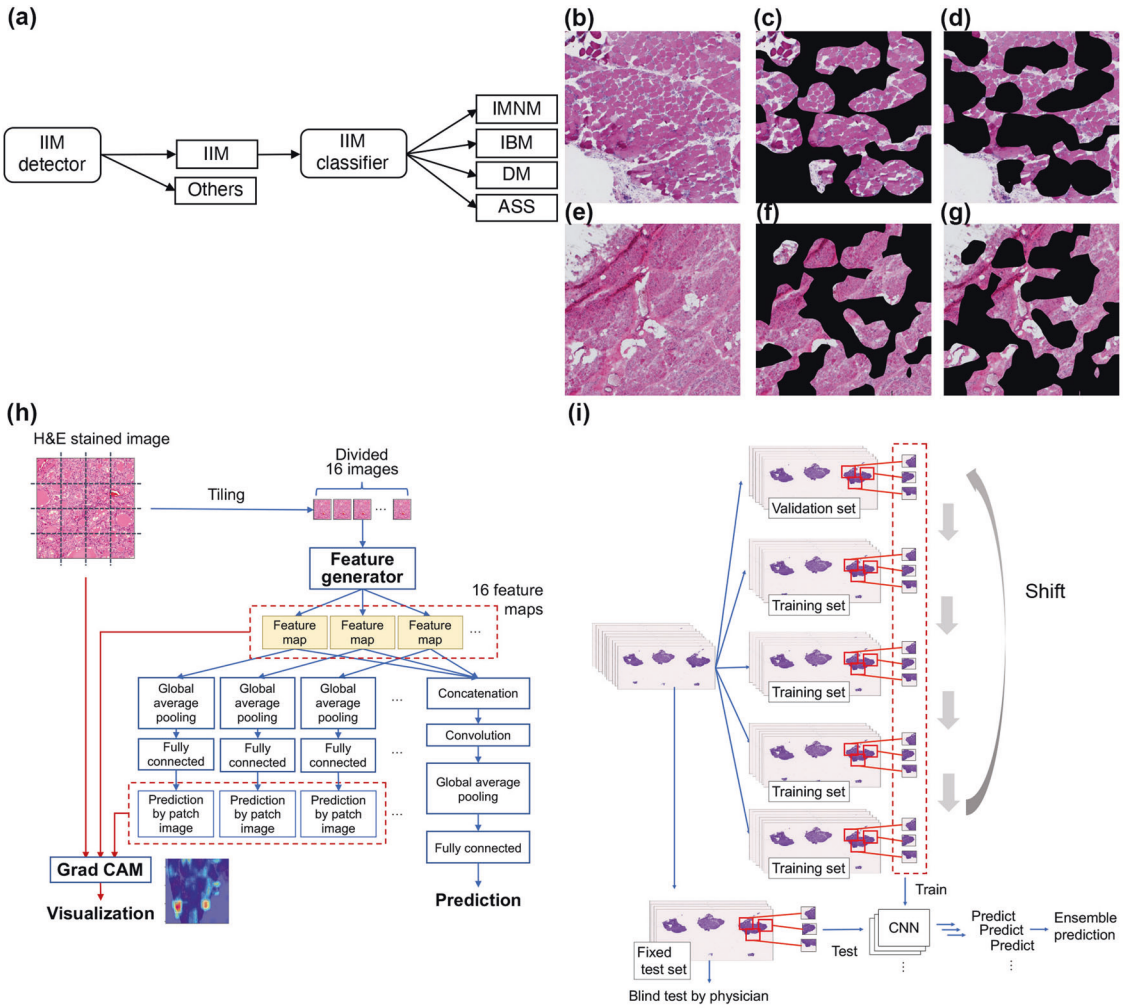

Fig. 1 Strategy, masked sample images, and deep convolutional neural network architecture. a The strategy of IIM identification. **b** An IIM (IMNM) image. **c** AI-focused image with masked areas not focused by CNNs in **b**. **d** AI-unfocused image with masked areas focused by CNNs in **b**. **e** A non-myositis muscle disease (FCMD). **f** AI-focused image with masked areas not focused by CNNs in **e**. **g** AI-unfocused image with masked areas focused by CNNs in **e**. **h** Deep CNN architecture; blue arrows indicate the flow for training and prediction, and red arrows indicate the visualization flow. **i** Approach for training and evaluating CNNs as compared with those of the physicians. The images were divided beforehand into training and test sets.

Setting the training algorithm

The number of trainings (the epoch size) was 60. Loss and accuracy were calculated using the validation set in every training. The CNN with the best accuracy was chosen for the test phase. We used rectified Adam[21] as the optimizer because it is more robust against the variance of learning rates than the conventional Adam[22] optimizer. The initial learning rate of the optimizer was 0.0001. The CNNs were trained on Nvidia V100 using TensorFlow 1.13.2 (https://github.com/tensorflow/tensorflow) and Keras 2.2.4 software (https://github.com/keras-team/keras).

Transfer learning

Transfer learning[23] is a technique employed to improve the performance of deep neural networks when the training data are limited. It uses the parameters of CNNs that are trained in one task and applies them to another task. In this study, the CNNs trained in IIM differentiation were used to classify IIMs and non-myositis muscle diseases. First, the CNNs were trained in IIM differentiation. Second, the final output layer of the trained CNNs was changed from two units to four units because there were two classes in the IIM differentiation task and four in the IIM classification task. Finally, the CNNs were trained in IIM classification. Transfer learning can be implemented in one of two ways: one is by having fixed parameters and excluding the output layer, and the other is by having unfixed parameters. We used the unfixed approach while classifying IIMs and non-myositis muscle diseases and set the number of output units to seven.

Metrics

We evaluated the performance of the algorithm using the following metrics: accuracy, receiver operating characteristic (ROC) curves (true- versus false-positive rate), and AUC (area under the ROC). The accuracy was calculated as follows:

Accuracy $= (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})$, where TP is true-positive, TN is true-negative, FP is false-positive, and FN is false-negative. When we used the five-fold cross-validation, we summed all the TP, TN, FP, and FN values in five shots. ROC curves were generated by sweeping the threshold from 0 to 1. In multi-class classifications, one target class was set as a positive class and the other classes as negative.

Visualization

In previous studies, heatmaps were generated to visualize the prediction by deep CNNs[17,24,25]. In this study, we adopted Grad-cam, which can generate visual explanations from any CNN-based network without requiring architectural changes or retraining. This method uses gradients between confidences and feature maps to identify reactive filters. The gradients are calculated as:

$$Gradient = \frac{\partial y^c}{\partial A},$$

where y^c is the confidence of class c, and A is a feature map. We used the following function to calculate gradients because the aggregated classifier

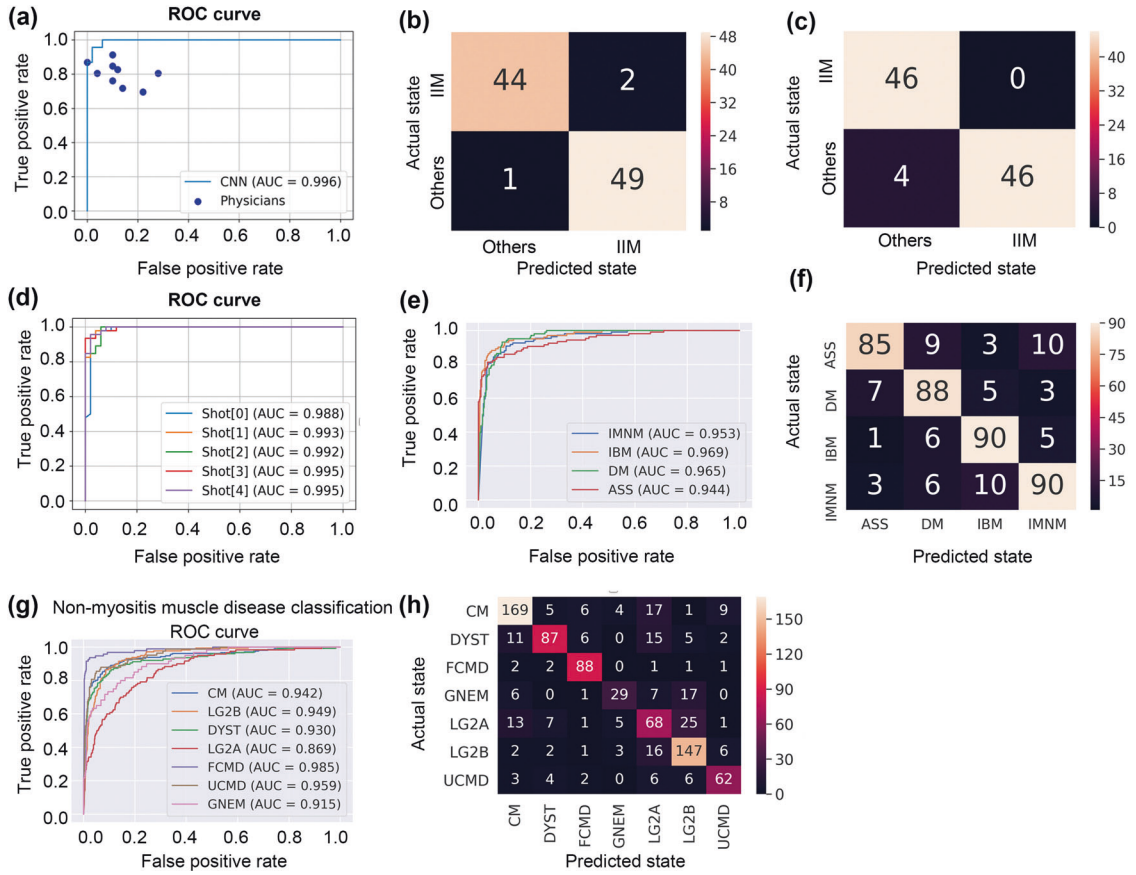

Fig. 2 Differentiation of IIM and classification of IIM and non-myositis muscle disease. a ROC curve of the CNN (blue line) and physicians' (blue dots) performances. **b** The confusion matrix that separates the IIM from the other diseases is 0.5. The vertical is the actual state, and the horizontal is the prediction. **c** The confusion matrix that separates the IIM from the other diseases is 0.3, and there were no misclassifications of IIM. **d** ROC curves without ensemble methods. **e** ROC curves of IIM classification. **f** Confusion matrix of IIM classification. The vertical is the actual state, and the horizontal is the prediction. **g** ROC curves showing the classification of non-myositis muscle disease. **h** Confusion matrix of non-myositis muscle disease classification.

internally generated feature maps per image patch.

$$Gradient = \frac{y_k^c}{\sum_{i=1}^{m} y_i^c} \frac{\partial y_k^c}{\partial A},$$

where m is the number of image patches, and y_k^c is y^c of image patch k.

The representation ability of Grad-cam depends on the size of the feature map: the larger the feature map, the better the representation. The size of the feature map in this study was 5×5, but the aggregated classifier generated 16 (4×4) feature maps, which means that, practically, the size of the feature maps was 20×20.

Participating physicians

Nine physicians (three adult neurologists, four pediatric neurologists, and two pathologists) who were specially trained in muscle pathology differentiated between IIM and other conditions by studying 96 pathology slides. Their training period for muscle pathology was in the range of 1–23 years (average: 5.5 years). For the visualization test, seven physicians (excluding two adult neurologists from the nine physicians) diagnosed 98 slides by studying AI-focused or AI-unfocused images.

RESULTS

Differentiation of IIM

The deep CNN was able to precisely differentiate IIM from other muscle diseases with an AUC of 0.996 and outperform the nine human specialists (Fig. 2a). Its accuracy was 96.9% if a probability of 0.5 or higher was determined as myositis, whereas the physicians' highest accuracy was 93.8%. The average and variability of the accuracies of the nine human specialists were 83.4% and 0.004, respectively. The number of misclassifications was similar between IIM and other conditions (Fig. 2b). If a probability of 0.3 or more was determined as IIM, the accuracy decreased to 95.8%, but the number of incorrectly classified IIMs became zero (Fig. 2c). We also confirmed that the model ensemble was efficient in further improving the CNNs' performance (Fig. 2d). Detailed test conditions and test results are presented in Table 2.

Classification of IIM and classification of hereditary muscle diseases

The CNN successfully classified four subtypes of IIM with AUCs of 0.953 (IMNM), 0.969 (IBM), 0.965 (DM), and 0.944 (ASS) (average = 0.958) (Fig. 2e). We found that the accuracy was 83.9% when the disease with the highest probability was the predicted class; ASS tended to be incorrectly classified as IMNM, and IMNM tended to be erroneously judged as IBM (Fig. 2f). It also classified seven subtypes of hereditary muscle disease with AUCs of 0.942 (CM), 0.949 (LGMD2B), 0.930 (DYST), 0.869 (LGMD2A), 0.985 (FCMD), 0.959 (UCMD), and 0.915 (GNEM) (average = 0.936) (Fig. 2g). We found that the accuracy was 74.5% when the disease with the highest probability was the predicted class; LGMD2A tended to be classified as LGMD2B; and GNEM tended to be misjudged as all other classes (Fig. 2h). Detailed test conditions and test results are presented in Table 2.

Table 2. Number of data samples required for training and validation and the results of every test case.

Test case	Test type	CV no.	Data (image)			Per slide		Per image	
			Train	Val	Test	AUC	Accuracy	AUC	Accuracy
IIM differentiation	Ensemble	N/A	N/A	N/A	288	0.996	0.969	0.992	0.958
	5-CV - Shot1	1	8364	2152	288	0.988	0.948	0.987	0.941
	5-CV - Shot2	2	8412	2104	288	0.993	0.958	0.984	0.934
	5-CV - Shot3	3	8420	2096	288	0.992	0.958	0.986	0.941
	5-CV - Shot4	4	8408	2108	288	0.995	0.948	0.990	0.958
	5-CV - Shot5	5	8460	2056	288	0.995	0.958	0.987	0.938
	5-CV - Average	N/A	N/A	N/A	288	0.993	0.954	0.987	0.942
IIM classification	5-CV - Shot1	1	1520	500	252	0.960	0.812	0.954	0.802
	5-CV - Shot2	2	1510	504	255	0.953	0.800	0.944	0.784
	5-CV - Shot3	3	1510	510	252	0.941	0.857	0.940	0.829
	5-CV - Shot4	4	1514	504	253	0.965	0.881	0.956	0.842
	5-CV - Shot5	5	1518	506	250	0.983	0.843	0.970	0.848
	5-CV - Average	N/A	N/A	N/A	N/A	0.958	0.839	0.970	0.821
Classification of non-myositis muscle disease	5-CV - Shot1	1	2956	944	510	0.896	0.695	0.890	0.688
	5-CV - Shot2	2	2956	1020	472	0.948	0.781	0.942	0.750
	5-CV - Shot3	3	2992	944	492	0.950	0.749	0.942	0.746
	5-CV - Shot4	4	2908	984	514	0.961	0.823	0.952	0.811
	5-CV - Shot5	5	2948	1028	472	0.920	0.676	0.913	0.669
	5-CV - Average	N/A	N/A	N/A	N/A	0.936	0.745	0.928	0.733

The amount of training and validation data was increased by data augmentation. Performance per slide was better than performance per image in most of the test cases.

Test with visualization

We created masked images to investigate the relationship between the physicians' predictions and diagnoses. First, we generated heatmap images with Grad-cam from H&E-stained images. Second, we calculated the median for each heatmap image. Finally, we transformed some heatmap images into mask images by masking areas below the median and applying them to H&E-stained images to create AI-focused images (Fig. 1c and f) and created the remaining mask images by using the same method to obtain AI-unfocused images. The number of AI-focused IIM images, AI-unfocused IIM images (Fig. 1d and g), AI-focused non-myositis muscle disease images, and AI-unfocused non-myositis muscle disease images was 25, 25, 24, and 24, respectively.

Seven physicians differentiated the IIM images into four subtypes (ASS, IBM, DM, and IMNM) and the non-myositis images into seven subtypes (DYST, LGMD2A, LGMD2B, CM, FCMD, UCMD, and GNEM). The pathologists were not informed about which images were AI-focused during testing. Each pathologist's accuracy was calculated and averaged in each group to calculate the significant difference between AI-focused and AI-unfocused images using the paired *t*-test. The function of the test was:

$$t = \frac{\sqrt{\mathrm{n}} \cdot \mu_d}{s},$$

where $\mu_d$and s are the mean and standard deviation of differences between all pairs, and n is the number of samples. We assumed that the results of the pathologists followed a normal distribution.

Result of test with visualization

Red and yellow areas were the most important to visualize the CNN predictions (Fig. 3b and d) because they were interspersed within the specimen (Fig. 3a–d). The average accuracy of the physicians' diagnosis with AI-focused and AI-unfocused images in myositis was 0.674 and 0.526, respectively (Fig. 3e). The accuracy of DM decreased significantly (Fig. 3f and g). The average accuracies of non-myositis diseases were 0.458 and 0.440, respectively (Fig. 3h). Comparatively, the AI-focused and AI-unfocused images of myositis conditions were significantly different ($p = 0.003$) but those of non-myositis conditions were not ($p = 0.629$) (Fig. 3f–g and i–j).

DISCUSSION

It is difficult to diagnose rare diseases such as muscle diseases because very few experienced specialists are available; therefore, there is an urgent need for accurate and cost-effective technological diagnostic systems that can be used in remote regions[2]. Herein, we reported a novel AI-based, CNN-assisted system to diagnose muscle diseases with pathology using an algorithm trained with limited image data (H&E-stained, CCD-shot slides). Our major finding is that the CNNs outperformed human physicians under limited diseases and conditions, indicating their potential for cost-effective clinical use, especially in underserved areas.

Most IIMs are treatable, but hereditary muscle diseases are untreatable; therefore, accurately differentiating between them is very important but often challenging even for experts (e.g., IMNM clinically and pathologically mimics muscular dystrophy, especially in children[26,27]). Moreover, each IIM subtype requires specific treatment, further highlighting the need for differentiation. In this study, CNNs successfully differentiated IIMs from other muscle diseases and classified them into its four subtypes (IBM, IMNM, DM, and ASS), suggesting that the system was effective for IIMs. We did not include polymyositis in this analysis because muscle pathologists are increasingly skeptical about its histopathological definitions—a recent conceptual change in IIMs[27,28].

The CNNs also classified seven major hereditary muscle diseases, demonstrating that the system is compatible with conventional muscle pathology diagnoses. This also suggests that

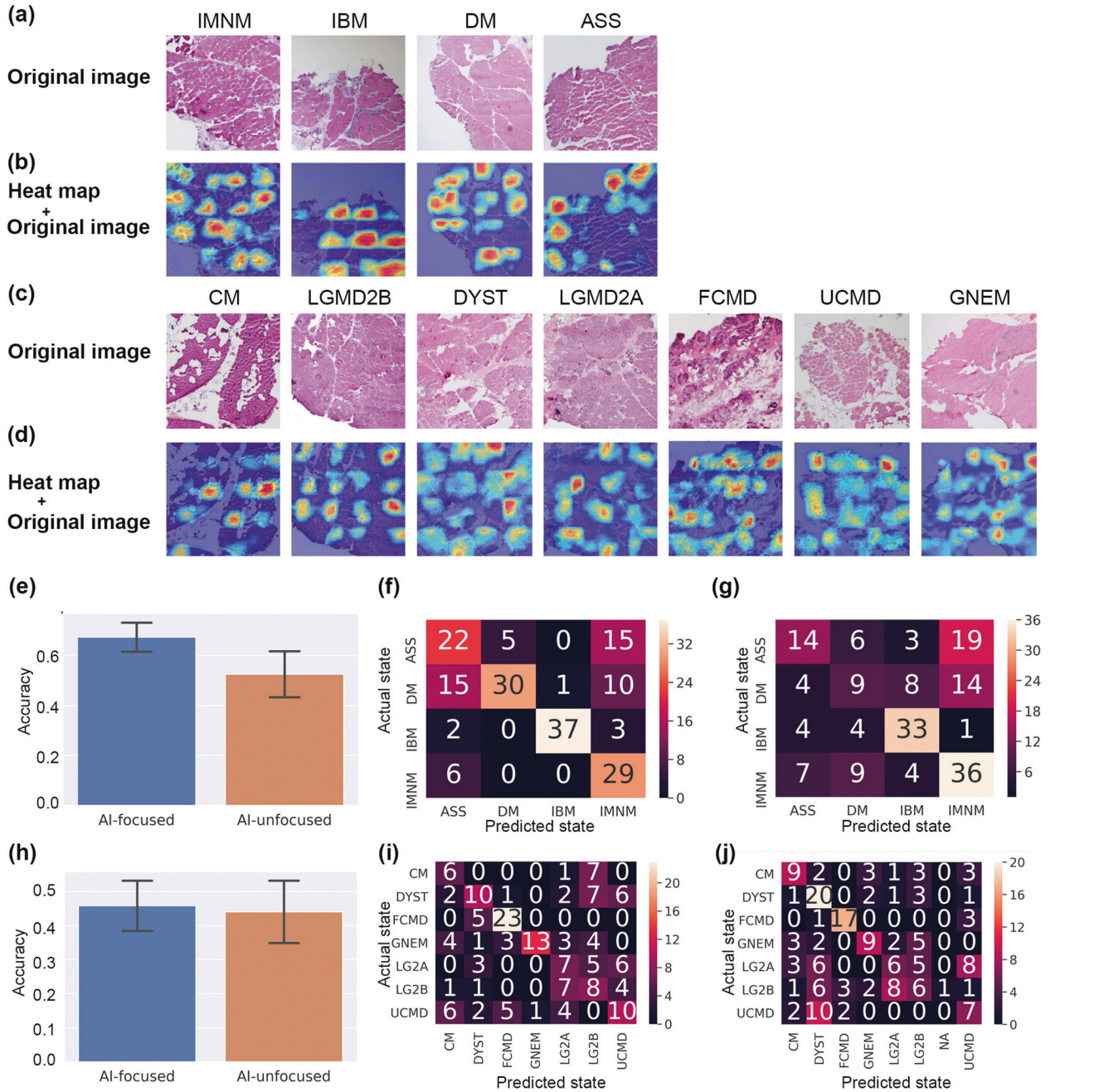

Fig. 3 Visualization of CNNs predictions. a Sample H&E-stained images of IIM. **b** Merged images of H&E-stained images and heatmap images created with Grad-cam. Red color indicates CNN focus areas essential for prediction. **c** Sample H&E-stained images of non-myositis muscle diseases. **d** Merged images of H&E-stained and heatmap images created with Grad-cam. **e** Physicians' test results in myositis. The left (blue) bar shows the average accuracy of the physicians' diagnosis with AI-focused images, while the right (orange) bar shows the AI-unfocused images. The error bar indicates the standard deviation. **f** Confusion matrix of results with AI-focused images in IIM. **g** Confusion matrix of results with AI-unfocused images in IIM. **h** Physicians' test results in non-myositis. The left (blue) bar shows the average accuracy of the pathologists' diagnosis with AI-focused images, and the right (orange) bar shows the AI-unfocused images. The error bar indicates the standard deviation. **i** Confusion matrix of results with AI-focused images in non-myositis. **j** Confusion matrix of results with AI-unfocused images in non-myositis. NA indicates no answer.

genetic differences can be computationally predicted based on histological features.

To visualize the accuracy of CNN predictions, Coudray et al. manually highlighted cropped image patches from a whole-slide photo[5]. In this study, we used Grad-cam[17] to automatically identify the critical regions for CNN predictions and create images to help investigate the relationship between the CNN predictions and physicians' diagnoses.

AI-focused IIM areas were useful for the physicians and CNNs; however, there was no significant difference in the non-myositis images. We speculated that the physicians were not as accurate as CNNs for non-myositis diseases because (1) the CNNs may have considered findings that were unknown to the physicians; (2) the CNNs may have recognized microstructures, such as rimmed vacuoles and nemaline bodies, that are important diagnostic features of hereditary diseases and are usually recognized by

physicians only at high magnifications; and (3) physicians are usually trained to observe muscular findings on stains other than H&E. The pathological findings of IIM, such as perifascicular atrophy in DM, are large enough to observe even in low-magnification H&E images, but hereditary muscle diseases are easier to identify with other stains such as the modified Gomori trichrome stain (vacuoles and nemaline bodies) and nicotinamide adenine dinucleotide dehydrogenase (NADH)-tetrazolium reductase stain (central cores).

In this study, we collected data from one of the world's largest muscle biopsy collections. The performance of CNNs can be influenced by differences in the image data collection, staining protocols, and cameras at various centers. Therefore, it is necessary to conduct further studies with a larger number of pathological images from several laboratories. We expect that CNNs will be used in prospective studies and will be embedded directly into medical equipment, as has been done in augmented reality microscopes[29]. We believe that this study provides promising outcomes that support the use of an AI-assisted system for diagnosing neuromuscular disorders.

DATA AVAILABILITY

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

REFERENCES

1. Ghebreyesus, T. A. Statement for Rare Disease Day. World Health Organization. https://www.who.int/mediacentre/news/statements/2018/rare-disease-day/en/ (2018).
2. Rare Diseases International. Rare diseases feature for first time at World Health Assembly. Available from https://www.rarediseasesinternational.org/rare-diseases-feature-for-first-time-at-world-health-assembly/ (2019).
3. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
4. Bejnordi, B. E. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
5. Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
6. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
7. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
8. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
9. Iizuka, O. et al. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Sci. Rep.* **10**, 1504 (2020).
10. Morrow, J. M. & Sormani, M. P. Machine learning outperforms human experts in MRI pattern analysis of muscular dystrophies. *Neurology* **94**, 421–422 (2020).
11. Verdu-Dıaz, J. et al. Accuracy of a machine learning muscle MRI-based tool for the diagnosis of muscular dystrophies. *Neurology* **94**, e1094–e1102 (2020).
12. Nishino, I. ABC in muscle pathology. *Rinsho Shinkeigaku* **51**, 669–676 (2011).
13. Xie, Y., Liu, F., Xing, F. & Yang, L. Deep learning for muscle pathology image analysis. In: Deep learning and convolutional neural networks for medical imaging and clinical informatics (eds Le, L., Wang, X., Carneiro, G. & Yang, L.) Ch. 2. 23–41 (Springer, Cham, 2019)
14. Tizhoosh, H. R. & Pantanowitz, L. Artificial intelligence and digital pathology: challenges and opportunities. *J. Pathol. Inform.* **9**, 38 (2018).
15. Kabeya, Y. et al. Physician-level aggregated classifier for genetic muscle disorders. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* 1850–1854. https://doi.org/10.1109/ISBI.2019.8759409 (IEEE, 2019).
16. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence* (IJCAI) 1137–1143 (1995).
17. Selvaraju, R. R. et al. Grad-cam: visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* 618–626. https://doi.org/10.1109/ICCV.2017.74 (IEEE, 2017).
18. Huang, G., Liu, Z., Maaten, L. V. D. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2261–2269. https://doi.org/10.1109/CVPR.2017.243 (IEEE, 2017).
19. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 248–255. https://doi.org/10.1109/CVPR.2009.5206848 (IEEE, 2009).
20. Dietterich, T. G. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, 1–15. https://doi.org/10.1007/3-540-45014-9_1 (2000).
21. Liu, L. et al. On the variance of the adaptive learning rate and beyond. Conference paper at ICLR 2020. Available from https://arxiv.org/abs/1908.03265 (2020).
22. Kingma, D. & Ba, J. Adam: a method for stochastic optimization. Published as conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. https://arxiv.org/abs/1412.6980 (2015).
23. Oquab, M., Bottou, L., Laptev, I. & Sivic, J. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1717–1724. https://doi.org/10.1109/CVPR.2014.222 (IEEE, 2014).
24. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2921–2929. https://doi.org/10.1109/CVPR.2016.319 (2016).
25. Mori, K. et al. Visual explanation by attention branch network for end-to-end learning-based self-driving. In *2019 IEEE Intelligent Vehicles Symposium*, 1577–1582. https://doi.org/10.1109/IVS.2019.8813900 (IEEE, 2019).
26. Liang, W. C. et al. Pediatric necrotizing myopathy associated with anti-3-hydroxy-3-methylglutaryl-coenzyme A reductase antibodies. *Rheumatology* **56**, 287–293 (2017).
27. Tanboon, J. & Nishino, I. Classification of idiopathic inflammatory myopathies: pathology perspectives. *Curr. Opin. Neurol.* **32**, 704–714 (2019).
28. Mariampillai, K. et al. Development of a new classification system for idiopathic inflammatory myopathies based on clinical manifestations and myositis-specific autoantibodies. *JAMA Neurol.* **75**, 1528–1537 (2018).
29. Chen, P. H. C. et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nat. Med.* **25**, 1453–1457 (2019).

AUTHOR CONTRIBUTIONS

Y.K. designed and conducted the experiments and wrote the manuscript; M.O. reviewed the pathological and genetic data and performed the physicians' test; H.N. and S.Y. analyzed the data and contributed to writing the manuscript; M.I., M.O., Y.S., J.T., A.I., T.K., YL.C., W.Y., and S.H. performed the physicians' test; T.I. designed the deep learning algorithm; Y.T. prepared the experimental environment; R.T. and A.T. directed the project; F.M. contributed to obtaining grants and designing the framework of the study; I.N. made the pathological diagnoses for all cases and supervised the study. All authors read and approved the final paper.

FUNDING INFORMATION

This study was supported by AMED under Grant Number 20ek0109348s0503 and Intramural Research Grant (2-5) for Neurological and Psychiatric Disorders of NCNP.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to I.N.