Application of Transfer Learning for Lissencephaly Gene Identification Using a Large-Scale Brain MRI Dataset and Self-Supervised Learning *

Naoki Takahashi and Yoshihiro Sato
Department of Design and Data Science, Tokyo City University, 3-3-1 Ushikubonishi Yokohama, Kanagawa, Japan
Email: yosato@tcu.ac.jp

Mitsuhiro Kato
Department of Pediatrics, SHOWA Medical University, 1-5-8 Hatanodai Shinagawa-ku, Tokyo, Japan

Atsuko Yamaguchi
Department of Design and Data Science, Tokyo City University, 3-3-1 Ushikubonishi Yokohama, Kanagawa, Japan

Lissencephaly is a severe congenital brain malformation characterized by smoothing of the cerebral cortex. Three-dimensional (3D) morphological evaluation based on brain magnetic resonance imaging (MRI) plays a crucial role in identifying causative genes, such as *LIS1* and *DCX*. However, owing to the rarity of this disease, collecting large-scale imaging datasets essential for constructing deep learning models remains difficult. Consequently, training deep neural networks from scratch using limited available datasets presents challenges arising from small sample sizes, such as overfitting in multilayer networks and limited generalization capabilities. To overcome these limitations, we propose a 3D image analysis framework that integrates self-supervised learning (SSL) and transfer learning. Specifically, we applied 3D contrastive learning (3D-SimCLR) to FOMO300K, an unlabeled brain MRI dataset comprising approximately 300,000 scans aggregated worldwide to pretrain a model capable of extracting generalizable and robust anatomical feature representations of human brain structures. Subsequently, fine-tuning was performed using a small lissencephaly cohort (88 T1-weighted images and 98 T2-weighted images) for a three-class causative gene classification task (*DCX*, *LIS1*, and others). The evaluation results demonstrated that the proposed model achieved an average overall accuracy of 84.1% on T1-weighted images and 80.6% on T2-weighted images, significantly outperforming our previous study using training from scratch (approximately 78%).

Keywords: Lissencephaly; MRI; FOMO300K; self-supervised learning.

* This work was supported by JSPS KAKENHI Grant Numbers JP24K11055 (MK and AY) and JP21K12148 (AY); the Ministry of Health, Labour and Welfare Research Program on Rare and Intractable Diseases under Grant Number JPMH23FC0201 (MK); the Showa University Intramural Joint Research Grant based on the Comprehensive Partnership Agreement (MK); and 2025 TOKYO CITY UNIVERSITY Priority Promotion Early-Career Scientists Research (YS).

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

1. Introduction

Lissencephaly, caused by impaired neuronal migration during fetal brain development, is a severe congenital disease characterized by a marked deficiency in the formation of cerebral gyri and sulci [1,2]. Accompanied by prominent morphological abnormalities involving smoothing of the cerebral surface, it causes intractable epilepsy and severe psychomotor developmental delay. Using molecular genetic approaches, researchers have identified specific gene mutations, such as *LIS1* and *DCX*, as the major causes of this condition [3,4]. Furthermore, distinct differences in the spatial distribution of cortical abnormalities, such as gradient structures along the anterior-posterior axis, have been shown to depend on the specific mutated gene [5,6]; thus, three-dimensional (3D) morphological evaluation using brain magnetic resonance imaging (MRI) plays a critical role in understanding the pathology of lissencephaly and estimating causative genes. However, lissencephaly is an extremely rare disease and the number of cases that can be accumulated at a single medical institution remains very small. Additionally, because MRI acquisition in infants involves technical limitations as well as risks associated with sedation, collecting high-quality imaging data suitable for large-scale research is challenging. In recent years, data-driven approaches, particularly deep learning, have become mainstream in medical image analysis, with highly accurate analytical methods that leverage datasets ranging from thousands to tens of thousands of scans [7,8]. However, in the case of rare diseases, such as lissencephaly, the scarcity of training data poses a major technical bottleneck, presenting an inherent challenge that prevents such conditions from fully benefiting from state-of-the-art machine learning techniques. Therefore, an urgent need exists to establish a novel computational framework capable of maximizing the extraction of disease-specific morphological features inherent in MRI scans and elucidating genotype-phenotype relationships, even under sample size constraints resulting from disease rarity.

To address the aforementioned data collection challenges inherent to rare diseases, we presented in our previous study an initial approach by aggregating the brain MRI data of patients with lissencephaly through a collaborative effort with multiple pediatric medical institutions in Japan, and constructing a deep-learning-based causative gene classification model [9]. To capture the complex, disease-specific spatial alterations associated with cortical smoothing, we adopted a 3D convolutional neural network (3D-CNN) rather than conventional 2D slice-by-slice analysis. Using this approach, by optimizing the models represented by 3D-ResNet [10] from scratch to classify mutations in LIS1 and DCX (the major causes of classical lissencephaly) as well as other genetic abnormalities, we achieved a promising classification accuracy of approximately 78%. Furthermore, visualization analysis of the decision rationale using Gradient-weighted Class Activation Mapping (Grad-CAM) [11] confirmed that the trained model identified genotypes by comprehensively extracting extensive morphological features across the entire brain, rather than relying solely on specific localized brain regions. Additionally, by introducing thresholding of prediction probabilities, we demonstrated that false positives can be suppressed even in environments with limited data sizes, thereby realizing more reliable predictions. These findings strongly suggest that models that directly process 3D spatial information are valuable for computer-aided diagnosis and pathological elucidation in rare disease fields, where data-driven approaches are typically considered challenging. More broadly, computational frameworks and deep learning models

specifically tailored for lissencephaly remain notably scarce in the literature. Conventional neuroimaging pipelines and general-purpose deep learning models are often inadequate for capturing complex 3D cortical malformations, particularly when constrained by small sample sizes. Consequently, establishing robust, domain-specific computational methods for rare cortical malformations represents a critical unmet clinical need.

However, the achievements of our previous study highlight the fundamental limitations of optimizing deep learning models from scratch using small-scale datasets unique to rare diseases. In general, in order for multilayer deep learning models to exhibit high expressiveness and learn complex fine structures or abstract feature representations, vast training datasets on the scale of tens of thousands to millions of samples are indispensable. Indeed, in our previous validation, although 3D-ResNet18 with a relatively shallow architecture exhibited the best performance, expanding the network to greater depths led to severe overfitting owing to a lack of training data, resulting in the degradation of classification accuracy. Furthermore, a training process relying on a limited cohort of only several dozen to a few hundred cases makes it difficult to acquire sufficient robustness against domain shifts caused by different MRI scanners or imaging protocols across medical institutions, leaving generalization performance based on unseen external datasets an open challenge. In other words, to analyze the pathology of lissencephaly in greater detail and precision, and to elevate the system into a genuine clinical-grade diagnostic support tool, departing from the conventional paradigm of building models solely on limited available data and introducing a novel approach that fundamentally resolves the optimization bottleneck caused by small sample sizes is necessary.

Therefore, to overcome the challenge of data scarcity specific to rare diseases, we incorporated self-supervised learning (SSL) and transfer learning frameworks—quantitative approaches that have advanced rapidly in recent years. Specifically, we utilized FOMO300K [12], a large-scale, heterogeneous dataset containing over 300,000 brain MRI scans aggregated from more than 910 data sources worldwide, as the foundation for pre-training. FOMO300K encompasses diverse imaging conditions and anatomical variations encountered in clinical practice, making it suitable for developing AI models that are robust against real-world noise. By applying 3D-SimCLR [13], a 3D extension of the contrastive learning algorithm, to this large-scale dataset, we enabled the model to acquire generalizable and powerful feature representations of complex 3D brain structures without relying on annotations. Subsequently, the resulting pre-trained weights were transferred to 3D-ResNet50, a deeper architecture that previously suffered from overfitting when trained from scratch, and were fine-tuned for the 3-class causative gene classification task (*LIS1*, *DCX*, and others) using the lissencephaly dataset. We hypothesized that transferring generalizable 3D morphological features acquired through self-supervised learning to a large-scale unlabeled dataset would prevent overfitting in deep networks, even for rare diseases such as lissencephaly, where data collection is extremely challenging, and would enable accurate discrimination of complex, subtle structural differences associated with each causative gene (*LIS1*, *DCX*, and others). By validating this hypothesis, we demonstrate that integrating self-supervised learning and transfer learning provides a promising new approach to overcoming data scarcity in the image analysis of rare genetic diseases.

2. Methods

With the aim of addressing the inherent data scarcity in the image analysis of lissencephaly, we developed a classification method that integrates self-supervised learning using large-scale unlabeled MRI data and transfer learning. The overall two-stage computational framework—encompassing Stage 1 self-supervised pre-training on the FOMO300K dataset, Stage 2 transfer learning with 5-fold cross-validation on the target lissencephaly cohort, architectural modifications (projection and classification heads), and final evaluation procedures—is schematically depicted in Fig. 1. Specifically, we describe the composition of the datasets used and the learning framework based on 3D-SimCLR.

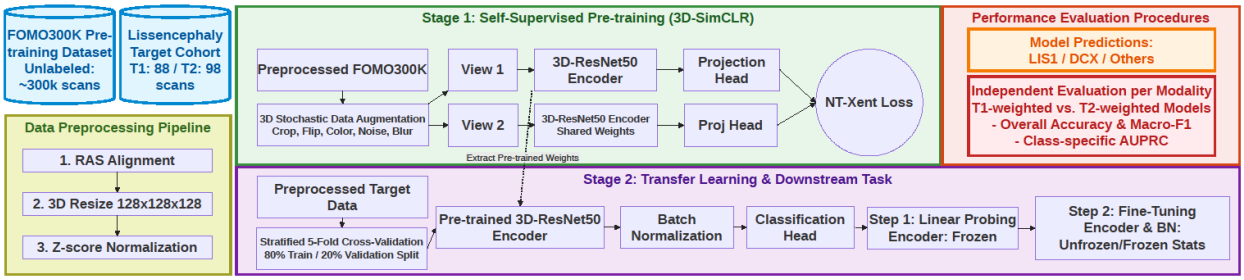


Fig. 1. Schematic diagram of the computational framework.

2.1. *Datasets*

To validate the proposed two-stage learning approach, we adopted two types of MRI datasets with different purposes and scales. We describe a large-scale pre-training dataset for acquiring generalizable 3D feature representations, and a small-scale clinical target dataset for the downstream task of classifying lissencephaly causative genes.

2.1.1. *Large-scale Pre-training Dataset (FOMO300K)*

To pretrain the encoder through self-supervised learning, we accessed FOMO300K [12], a large-scale and highly heterogeneous brain MRI database. This dataset consists of 306,207 scans from 58,741 subjects aggregated from 920 public data sources worldwide. The database includes standard high-resolution scans for research purposes and low-resolution images acquired in real clinical settings, covering data obtained by diverse scanners with different magnetic field strengths and from major manufacturers. Furthermore, imaging modalities are diverse, including T1-, T2-, and diffusion-weighted images. Healthy individuals accounted for approximately 65% of the subjects; however the database also included rich disease data from patients presenting with severe anatomical abnormalities, such as brain tumors and cerebrovascular diseases.

The reason for adopting FOMO300K as the pre-training foundation in this study was to acquire generalizable and robust anatomical feature representations for the analysis of lissencephaly, a rare disease. If a dataset comprising data from a single institution, limited imaging conditions, or solely normal brains were used, overfitting to specific protocols would easily occur, posing the risk of failing to adapt to domain shifts in real clinical data. In this study, we comprehensively applied in the pre-training stage a cohort of scans encompassing diverse imaging conditions and pathologies without prior filtering based on the modality of the input data. Specifically, we used 300,980 scan

data points that were successfully acquired and preprocessed from the entire collection for pre-training. Thus, by learning from a wide variety of unlabeled data on a large scale without relying on specific pathologies or single institutions, constructing a universal 3D feature space regarding human brain structures is possible, which strongly ensures the academic validity of subsequent transfer learning to a small lissencephaly cohort.

2.1.2. *Lissencephaly Target Dataset*

To evaluate the downstream tasks and classification performance of transfer learning, we used a clinical MRI cohort of patients with lissencephaly; the data was accumulated through a collaborative research network comprising multiple pediatric medical institutions in Japan. The use of the case data was approved by the ethics committee of the leading institution, Showa University, and appropriate consent had been previously obtained from all subjects or their legal representatives.

Building upon the dataset constructed in our previous study [9] (181 3D scans), we conducted a further rigorous clinical review and expanded the cohort to a total of 186 3D MRI scans with genetically confirmed diagnoses. Selecting T1-weighted and T2-weighted modalities, the cohort comprises 88 T1-weighted images (32 *LIS1*, 29 *DCX*, and 27 other mutations) and 98 T2-weighted images (31 *LIS1*, 33 *DCX*, and 34 other mutations). This setup yields a balanced distribution across three target categories: *LIS1* (63 scans in total), *DCX* (62 scans in total), and other genetic mutations (61 scans in total). Because *LIS1* and *DCX* mutations account for the vast majority of classical lissencephaly cases, identifying or ruling out these two primary etiological groups is of paramount clinical importance for prognosis, patient management, and genetic counseling. The "other genetic mutations" class encompasses 61 scans from 23 patients across 14 distinct rare gene mutations: *ARX* (10 scans), *WDR62* (8 scans), *DYNC1H1* (8 scans), *TUBB2A* (7 scans), *FOXG1* (6 scans), *TUBB2B* (4 scans), *KIF11* (4 scans), *CUL1 & WDR37* (3 scans), *ASPM* (2 scans), *DARS2* (2 scans), *PNKP* (2 scans), *RTTN* (2 scans), *WDR81* (2 scans), and *LAMC3* (1 scan). Although this category exhibits genetic heterogeneity, grouping these ultra-rare variants into a single "other" class reflects a pragmatic clinical screening approach, given that sample sizes for individual rare genes are insufficient to train separate neural networks.

To maximally leverage the benefits of the generalizable feature space constructed through pre-training in the target task, exactly matching the input distribution of the downstream task dataset with that of the pre-training data is necessary. Accordingly, a strict image pre-processing pipeline consistent with the pre-training phase, such as standardizing the anatomical orientation and unifying the spatial resolution, was applied to all scans included in this cohort. The detailed procedures for pre-processing and normalization are described later.

2.2. *Pre-training and Transfer Learning Process to Overcome Data Scarcity*

To avoid overfitting in deep learning using small-scale datasets, we designed a pipeline that integrates self-supervised pre-training and transfer learning. In this subsection, we detail the mechanism of 3D contrastive learning using 3D-SimCLR and the transfer learning process for adapting the trained model to downstream tasks.

2.2.1. *Model Architecture*

To acquire generalizable feature representations from medical images without relying on ground-truth labels, we adopted the 3D-SimCLR framework, which is an extension of the SimCLR contrastive learning algorithm to 3D space [13]. Considering that the lack of training data is the greatest technical barrier in the image analysis of rare diseases, we aimed to construct a universal anatomical feature space for the human brain by introducing self-supervised pre-training using a large-scale, unlabeled brain MRI dataset.

Because the encoder forms the foundation for feature extraction, we adopted 3D-ResNet50, which has a multilayer residual connection structure [10]. In a previous study [9], we illustrated the effectiveness of 3D-CNNs in capturing disease-specific morphological abnormalities in lissencephaly. However, optimization from small-scale data imposes a limit on network depth. The goal in developing the proposed method was to acquire higher-order and more abstract feature representations using a deep hierarchical structure by suppressing overfitting through large-scale pre-training.

To maximize the representational capacity for contrastive learning, a non-linear projection head consisting of a multilayer perceptron is placed at the output stage of the encoder. With this structure, the extracted high-dimensional feature vectors are projected onto a lower-dimensional latent space for loss calculation. In the optimization during training, the Normalized Temperature-scaled Cross Entropy Loss (NT-Xent Loss) is used to update the encoder weights such that the similarity between the feature vectors of a pair generated from the same scan (positive examples) is maximized while minimizing the similarity with other scans in the batch (negative examples). Specifically, for a given input 3D scan, two distinct views generated via stochastic data augmentations form a positive pair, whereas all other $2(B-1)$ augmented scans within a mini-batch of size B are treated as negative pairs. Because the FOMO300K pre-training dataset inherently mixes T1-, T2-, and diffusion-weighted scans acquired across diverse magnetic field strengths and patient pathologies, mini-batches naturally present high contrast and image quality variability. Constructing positive pairs exclusively from two augmented views of the same scan forces the encoder to pull together representations of identical anatomical structures despite heavy spatial and intensity perturbations. Concurrently, treating all other distinct subjects and modalities in the mini-batch as negative pairs prevents the network from exploiting superficial shortcuts, such as modality contrast or scanner noise, thereby promoting the learning of a highly robust, modality-invariant 3D anatomical feature space. The specific network architecture and optimization parameters during training are described later.

2.2.2. *Data Augmentation*

To acquire useful feature representations within the 3D-SimCLR framework, generating two highly correlated views (augmented images) for the same input data is necessary, altering their appearance while preserving the anatomical semantic structures. Accordingly, a dynamic and stochastic 3D data augmentation pipeline was applied to each minibatch during training. Specifically, random cropping, which mimics the extraction of local anatomical structures in 3D space, and left-right flipping operations, which consider the structural symmetry of the brain, were introduced. In

addition, to enhance the robustness against domain shifts caused by different MRI scanners or imaging protocols, non-linear variations in brightness and contrast, noise addition, and spatial smoothing processes were combined and applied using random probabilities. By conducting contrastive learning through these diverse appearance variations, the model is encouraged to learn a representation space invariant to the essential morphological features of the disease without overfitting to scanner-dependent superficial noise or image quality fluctuations. The application probabilities and specific variation coefficients for each data augmentation method are detailed in the *Experimental Settings* section.

2.2.3. *Transfer Learning*

To apply the 3D anatomical feature representations acquired through self-supervised pre-training to the analysis of lissencephaly, a framework for transfer learning toward the downstream task was introduced. The downstream task in this study was a three-class classification task that discriminated between the *LIS1* mutation group and the *DCX* mutation group, the major causes of lissencephaly, and the other genetic abnormality group. To rigorously evaluate the contribution of morphological features per modality, we constructed independent classification models for T1-weighted images and T2-weighted images. This single-modality design directly reflects routine clinical workflows, where complete multi-sequence MRI protocols (e.g., paired high-resolution T1 and T2 scans) are not always available for pediatric patients owing to motion artifacts, incomplete acquisitions, or emergency screening conditions. Evaluating each modality independently ensures that our framework remains practical and broadly applicable even when only a single 3D MRI sequence is available, while establishing a diagnostic baseline for each modality in genetic etiology classification.

To adapt the network architecture, a pre-trained 3D-ResNet50 encoder was adopted as the foundation, and the non-linear projection head designed for contrastive learning was discarded. Subsequently, batch normalization is applied to the high-dimensional feature vectors extracted from the encoder to stabilize the distribution of the feature space, and a fully connected layer that outputs the class membership probabilities is integrated as a new classifier.

In the model optimization process, training is designed in two stages to safely adapt the model to the target task, while preventing the catastrophic forgetting of generalizable and useful feature representations cultivated through pre-training. In the first stage, linear probing, the weights of the pre-trained encoder are completely frozen, and only the parameters of the newly added classifier are updated. This prevents excessive destruction of the encoder weights due to immature backpropagation from the randomly initialized classifier, thereby establishing an initial mapping between the feature space and the new task. In the second stage of fine-tuning, the freezing of the encoder is released, and the weights of the entire network are updated, thereby optimizing the model to minute spatial morphological changes specific to lissencephaly. During this fine-tuning phase, a design that freezes the statistical updates of all batch normalization layers within the network was adopted. This is a measure to forcefully suppress overfitting by preventing the phenomenon in which within-batch statistics become unstable and the model's generalization performance collapses under small batch-sized environments, resulting from the limited sample size in rare disease data and the

memory constraints related to 3D image processing. The loss functions, specific optimization algorithms, and hyperparameter settings for each training stage are detailed in the *Experimental Settings* section.

3. Experiments and Results

In this section, we describe the details of the experiments conducted to verify the effectiveness of the proposed method along with their results. First, we explain the experimental settings, such as image pre-processing and hyperparameters, used to match the input data distribution between the pre-training and downstream tasks. Subsequently, we present the evaluation results of the three-class classification task of the causative genes using T1-weighted and T2-weighted images.

3.1. *Experimental Settings*

Following are the specific procedures for aligning the distribution of the input data between the pre-training and downstream tasks to realize stable model training. In addition to the image pre-processing pipeline applied to all scans, we detail the hyperparameters and computational environment for each training phase.

3.1.1. *Image Pre-processing*

To strictly match the input space distribution between the pre-training data and the target data of the downstream task, a unified image pre-processing pipeline was applied to all scans. First, after removing missing and infinite values contained in the scan data, geometric alignment was performed to standardize the anatomical orientation to the Right-Anterior-Superior (RAS) standard space. Specifically, swapping spatial dimensions and flipping along the anterior-posterior axis were used as fundamental operations. For a subset of scans in which the slice order was recorded from the superior side, an additional superior-inferior flip was applied, thereby achieving complete standardization of the spatial orientation across all data. Subsequently, 3D linear interpolation was applied to resize each scan to a uniform spatial resolution of 128 × 128 × 128 voxels. Lastly, to stabilize the learning process, Z-score normalization, using the mean and standard deviation of the voxel values of the entire volume, was independently applied to each scan to construct an input tensor for the model.

3.1.2. *Model Settings*

During the self-supervised pre-training phase, stochastic data augmentation was applied to the resized input volumes. Specifically, we implemented random cropping to a 96 × 96 × 96 voxel region (100% application probability), left-right flipping (50% application probability), continuous non-linear color transformation (80% application probability), addition of Gaussian noise with a mean of 0 and a standard deviation of 0.1 (50% application probability), and spatial smoothing using a 3D Gaussian filter with a kernel size of 3 × 3 × 3 (50% application probability). With respect to the optimization algorithm, the Adam method [14] (learning rate: 1.0×10^{-4}, $\epsilon = 1.0 \times 10^{-4}$) was adopted with cosine decay scheduling. We utilized the Normalized Temperature-scaled Cross

Entropy (NT-Xent) loss with the temperature parameter set to 0.1. To address memory constraints, mixed-precision computation in the bfloat16 format and gradient accumulation over four steps were introduced, yielding an effective batch size of 256 from a physical batch size of 64. Because Stage 1 involved an unsupervised representation learning process across the entire unlabeled database without target validation labels, early stopping was not applied; instead, pre-training was executed for a fixed duration of 100 epochs, a threshold at which empirical loss convergence was consistently confirmed in preliminary experiments.

In the subsequent transfer learning phase (Stage 2), the target dataset was evaluated using stratified 5-fold cross-validation (80% training, 20% validation per fold) with fixed random seeds for reproducibility. Categorical cross-entropy was adopted as the loss function for downstream three-class genetic classification. Optimization proceeded in two sequential stages: linear probing for 50 epochs using the Adam optimizer [14] (learning rate: 1.0×10^{-3}) while freezing the encoder weights, followed by fine-tuning for an additional 50 epochs with a decayed learning rate of 1.0×10^{-5} after unfreezing the encoder (with batch normalization statistics frozen). Rather than relying on arbitrary final-epoch weights, model selection and termination within each phase were governed by a custom validation checkpointing mechanism at the end of every epoch. The optimal model weights were saved and extracted based on maximizing validation accuracy as the primary criterion and minimizing validation loss as the tie-breaking secondary criterion, thereby ensuring optimal generalization on unseen validation data.

3.1.3. *Computational Environment and Implementation*

The entire process of network construction, pre-training, and transfer learning in this study was implemented in a Python 3.10 environment using the TensorFlow 2.21 and Keras 3.12 libraries. In addition, Scikit-learn 1.7 was utilized for cross-validation data splitting and calculating diverse performance evaluation metrics. All the computational processes were executed on a workstation equipped with an NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition GPU, CUDA 12.9, and cuDNN 9.11.

3.2. *Evaluation Metrics*

Multiple evaluation metrics were adopted to quantitatively evaluate the performance of the three-class classification model for causative genes, which is a downstream task. First, the overall accuracy was calculated as an indicator of the basic classification performance of the entire model. However, in medical image datasets, particularly in studies targeting rare diseases, the number of cases of specific disease subtypes is limited, which tends to cause an imbalance in the number of samples among the classes. If an evaluation relies solely on a single overall accuracy under such conditions, a risk that the overall evaluation will be biased toward the identification performance of the majority class exists, leading to overestimation of the model's performance.

To address the issue of data imbalance, we adopted a macro-average approach, wherein the evaluation metrics for each class were calculated independently, and their arithmetic mean was determined. Specifically, the evaluation was conducted using the macro-precision, macro-recall, harmonic mean, and macro-F1 score. With this approach, the identification performance of minority

classes is incorporated into the evaluation system with the same weight as that of the majority classes, making it possible to measure the objective classification capability of the model without bias toward specific classes.

To verify the detailed identification accuracy for each gene mutation subtype, the area under the precision-recall curve (AUPRC) was calculated for each class. As the generally used Area Under the Receiver Operating Characteristic curve (AUROC) is significantly affected by the number of true negatives, concerns have been raised that it overestimates the performance of minority classes in imbalanced data environments. In contrast, AUPRC focuses on the accuracy of positive predictions, giving it the characteristic of more strictly and accurately reflecting the detection accuracy for minority classes. The reliability of the proposed model in classifying rare diseases with significant clinical importance was comprehensively evaluated by calculating the AUPRC independently for each class.

3.3. *Validation Results*

We describe the evaluation results of the three-class classification model to identify the causative genes of lissencephaly (*DCX*, *LIS1*, and others). A performance evaluation was conducted using a stratified 5-fold cross-validation on independent datasets of T1-weighted images (88 scans) and T2-weighted images (98 scans).

3.3.1. *Classification Performance on T1-weighted Images*

The evaluation results for each fold of the cross-validation on the T1-weighted image dataset as well as the average values across all five folds are shown in Table 1.

Table 1. Cross-validation results for 3-class classification using T1-weighted images.

Fold	Overall Accuracy	Macro-Precision	Macro-Recall	Macro-F1 Score	AUPRC (DCX)	AUPRC (LIS1)	AUPRC (Others)
Fold 1	0.7778	0.7746	0.7746	0.7706	0.7833	0.7919	0.7958
Fold 2	0.8889	0.8968	0.8889	0.8885	0.9151	0.8965	0.9444
Fold 3	0.8333	0.8889	0.8333	0.8364	1.0000	0.6968	0.7949
Fold 4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Fold 5	0.7059	0.7111	0.7238	0.7142	0.9267	0.7704	0.6644
Average	0.8412 (±0.0999)	0.8543 (±0.1011)	0.8441 (±0.0956)	0.8419 (±0.0986)	0.9250 (±0.0793)	0.8311 (±0.1059)	0.8399 (±0.1194)

As shown in the table, the model using T1-weighted images achieved an average overall accuracy of 84.12% (±9.99%) and a macro-F1 score of 84.19% (±9.86%) across five folds. Fold 4 recorded a perfect classification accuracy of 1.0000 across all evaluation metrics. Focusing on the area under the precision-recall curve (AUPRC) by class, the average AUPRC for the *DCX* mutation showed the highest value at 0.9250, followed by the other groups at 0.8399, and the *LIS1* mutation at 0.8311.

3.3.2. *Classification Performance on T2-weighted Images*

In Table 2 we present the results of cross-validation of the T2-weighted image dataset.

Table 2. Cross-validation results for 3-class classification using T2-weighted images.

Fold	Overall Accuracy	Macro-Precision	Macro-Recall	Macro-F1 Score	AUPRC (DCX)	AUPRC (LIS1)	AUPRC (Others)
Fold 1	0.9000	0.9167	0.8889	0.8889	0.9094	0.7788	0.9617
Fold 2	0.7000	0.7460	0.6984	0.6976	0.8214	0.5418	0.7810
Fold 3	0.8500	0.8571	0.8492	0.8498	0.9201	0.8857	0.9240
Fold 4	0.7368	0.7639	0.7460	0.7463	0.9242	0.5208	0.7177
Fold 5	0.8421	0.8413	0.8492	0.8419	0.9484	0.8179	0.8516
Average	0.8058 (±0.0750)	0.8250 (±0.0627)	0.8063 (±0.0718)	0.8049 (±0.0713)	0.9047 (±0.0436)	0.7090 (±0.1492)	0.8472 (±0.0896)

As for the model using T2-weighted images, an identification accuracy with an average overall accuracy of 80.58% (±7.50%) and an average macro-F1 score of 80.49% (±7.13%) were obtained. Compared with the results of the T1-weighted image model, although the overall average values were slightly lower, the standard deviation of the overall accuracy was 0.075, and that of the macro-F1 score was 0.071, confirming that the performance variation among folds was suppressed. Analyzing the performance by class, similar to the T1-weighted images, the classification performance for the *DCX* mutation was the highest, with an average AUPRC of 0.9047. In contrast, the average AUPRC for the *LIS1* mutation was 0.7090, and that for the "Others" group was 0.8472.

3.3.3. *Comparison with Training from Scratch*

To isolate the diagnostic gain provided by self-supervised pre-training from the effect of dataset expansion, we evaluated baseline 3D-ResNet50 models trained from scratch (random initialization without pre-training) using the same stratified 5-fold cross-validation setup on both single-modality target datasets.

When trained from scratch, the T1-weighted model achieved a mean accuracy of 59.15% (±7.84%) and a macro-F1 score of 56.63% (±9.11%), whereas the T2-weighted model yielded a mean accuracy of 68.47% (±9.61%) and a macro-F1 score of 68.35% (±9.55%). In our prior work [9], combining T1 and T2 scans into a unified dataset without pre-training achieved an accuracy of 78.0%. When splitting the expanded cohort into single modalities to address real-world imaging constraints, training from scratch on reduced per-modality sample sizes (88 T1 scans and 98 T2 scans) led to a severe degradation in diagnostic performance due to overfitting on small 3D volumes.

Crucially, introducing self-supervised pre-training on FOMO300K yielded massive absolute accuracy improvements of +24.97% on T1 images (59.15% to 84.12%) and +12.11% on T2 images (68.47% to 80.58%). These direct comparisons demonstrate that self-supervised pre-training—rather than sample size expansion alone—is the primary driver enabling the network to overcome small-sample limits and to acquire generalizable 3D cortical feature representations.

4. Discussion

The results of our analysis demonstrate that the proposed self-supervised pre-training framework leveraging a large-scale unlabeled brain MRI cohort effectively addresses the severe data-scarcity bottleneck in rare disease classification. In our previous study [9], small sample sizes constrained the model to a shallow network architecture, thereby limiting classification accuracy to 78.0%. By transferring generalizable anatomical representations acquired from hundreds of thousands of unannotated scans, the present approach successfully suppressed overfitting within a deep 3D-ResNet50 architecture, achieving a mean overall accuracy of 84.12% on T1-weighted images. These findings substantiate that self-supervised pre-training on large-scale datasets provides a robust strategy for training deep neuroimaging models under severe sample constraints.

Class-specific evaluations revealed that the *DCX* mutation group achieved consistently high accuracy (AUPRC > 0.90 across modalities), indicating that the pre-trained feature space effectively captures characteristic 3D cortical malformations associated with specific genetic etiologies. The fold-to-fold performance variability observed during cross-validation—including the perfect score in Fold 4 of the primary T1 evaluation—is primarily attributable to the small validation sample size (~17–18 scans per fold), where specific splits occasionally exhibit favorable geometric separability. Re-evaluating the pipeline across alternative random seeds yielded consistent overall accuracies, confirming that the framework maintains stable generalizability across different data partitions without data leakage.

Performance discrepancies were, however, observed between modalities, with the T1-weighted model outperforming the T2-weighted model (84.12% vs. 80.58% overall accuracy). Beyond physical MRI contrast properties, this gap is largely driven by modality imbalance within the FOMO300K pre-training cohort, which contained 91,917 T1-weighted scans compared to 33,781 T2-weighted scans. This imbalance yielded richer representation quality for T1-weighted sequences during pre-training, directly influencing downstream task performance.

4.1 *Limitations and Future Directions*

Some limitations of this study should be noted. First, although evaluating single modalities independently addresses clinical scenarios where paired sequences are unavailable, developing unified multimodal models (e.g., combining T1, T2, and FLAIR) represents an important next step to improve diagnostic power. Second, incorporating 3D visual explainability techniques to map mutation-specific morphological variations was not implemented in the present framework. Third, benchmark comparisons against other emerging brain MRI foundation models were not performed in this study. Finally, external validation on larger, multi-center cohorts is needed to confirm domain robustness across clinical protocols. Future work should be focused on balancing pre-training modality distributions, incorporating multimodal architectures, and validating model generalizability across multi-institutional datasets.

5. Conclusion

In order to computationally analyze the pathology of lissencephaly, a rare disease, from brain MRI scans and identify its causative genes, we developed a novel classification model that combines

self-supervised learning using a large-scale unlabeled brain MRI dataset and a 3D-CNN. The primary findings of this study are as follows:

1. As a result of acquiring generalizable feature representations through pre-training using a heterogeneous dataset of over 300,000 scans (FOMO300K), the proposed model achieved an average overall accuracy of 84.1% on T1-weighted images and 80.6% on T2-weighted images in a 3-class classification task (*LIS1*, *DCX*, and others) using the 3D-ResNet50 architecture.
2. The proposed method clearly surpasses the classification accuracy of our previous study (approximately 78%), which involved training from scratch using a limited dataset. This demonstrates that leveraging large-scale data through self-supervised learning is an extremely effective method of preventing overfitting in deep networks and overcoming the barriers of training data scarcity in rare diseases.
3. In detailed evaluations by class, the AUPRC for the *DCX* mutation group exhibited high values exceeding 0.9 in T1 and T2 modalities, confirming that the proposed model can robustly extract 3D abnormal cortical structures associated with specific causative genes.
4. Performance differences were observed between the modalities, particularly in the accuracy between T1- and T2-weighted images in discriminating *LIS1* mutations. This suggests that the imbalance in the number of samples for each modality in the pre-training dataset (T1: approximately 90,000 scans; T2: approximately 30,000 scans) influenced the quality of the acquired feature representations.

The results of this study demonstrate that a novel paradigm, constructing large-scale foundation models combined with transfer learning, serves as a powerful solution for the image analysis of rare genetic diseases where data collection is fundamentally difficult. Future developments involving pre-training with optimized data balance across modalities and multimodal approaches integrating different MRI contrasts are expected to contribute to the realization of truly generalizable and reliable automated diagnostic support systems for clinical practice.

6. Acknowledgments

This study was conducted using a database of clinical cases of lissencephaly provided by Showa University and its affiliated medical institutions. We would like to express our deepest gratitude to all those involved for their invaluable cooperation in data collection and tremendous support for this study. Additionally, this study was supported by JSPS KAKENHI Grant Numbers JP24K11055 (MK and AY) and JP21K12148 (AY); the Ministry of Health, Labour and Welfare Research Program on Rare and Intractable Diseases under Grant Number JPMH23FC0201 (MK); and the Showa University Intramural Joint Research Grant based on the Comprehensive Partnership Agreement (MK); 2025 TOKYO CITY UNIVERSITY Priority Promotion Early-Career Scientists Research(YS).

References

1. A. J. Barkovich et al., A developmental and genetic classification for malformations of cortical development: update 2012, Brain 135, 1348 (2012).
2. M. Kato and W. B. Dobyns, Lissencephaly and the molecular basis of neuronal migration, Hum. Mol. Genet. 12, R89 (2003).
3. O. Reiner et al., Isolation of a miller-dieker lissencephaly gene containing G protein beta-subunit-like repeats, Nature 364, 717 (1993).
4. V. des Portes et al., A novel CNS gene required for neuronal migration and involved in X-linked subcortical laminar heterotopia and lissencephaly syndrome, Cell 92, 51 (1998).
5. D. T. Pilz et al., LIS1 and XLIS (DCX) mutations cause most classical lissencephaly, but different patterns of malformation, Hum. Mol. Genet. 7, 2029 (1998).
6. M. S. Forman et al., Genotypically defined lissencephalies show distinct pathologies, J. Neuropathol. Exp. Neurol. 64, 847 (2005).
7. G. Litjens et al., A survey on deep learning in medical image analysis, Med. Image Anal. 42, 60 (2017).
8. I. D. Mienye et al., Deep convolutional neural networks in medical image analysis: a review, Information 16, 195 (2025).
9. 9.N. Takahashi et al., Deep learning-based identification of causative genes in lissencephaly using 3D-MRI volumetric datasets, Neuroimage Rep. 6, 100375 (2026).
10. 10.K. He et al., Deep residual learning for image recognition, in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16). Institute of Electrical and Electronics Engineers: New York (2016). p. 770.
11. 11.R. R. Selvaraju et al., Grad-CAM: visual explanations from deep networks via gradient-based localization, in Proc. IEEE International Conference on Computer Vision (ICCV'17). Institute of Electrical and Electronics Engineers: New York (2017). p. 618.
12. 12.S. Cerri et al., A Large-Scale Heterogeneous 3D Magnetic Resonance Brain Imaging Dataset for Self-Supervised Learning, arXiv Preprint ArXiv:2506.14432 (2026).
13. 13.T. Chen et al., A simple framework for contrastive learning of visual representations, in Proc. International Conference on Machine Learning (ICML'20), 1597 (2020).
14. 14.D. P. Kingma and J. Ba, Adam: a method for stochastic optimization, in Proc. International Conference on Learning Representations (ICLR'15) (2015).