Editorial
Special Issue on Explainable and Generalizable Deep Learning for Medical Imaging

THE rapid advancements in deep learning technologies have profoundly influenced the field of medical image analysis, yet their full integration into clinical radiology practices has not progressed as quickly as expected. A significant hurdle to their widespread adoption among radiologists and clinicians is the prevailing lack of trust and confidence in the outcomes produced by these technologies. This concern primarily stems from concerns regarding the explainability and generalizability of deep learning models within the realm of medical imaging. As part of the responses from the Medical Image Analysis Community to address these critical issues, we organized the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS) Special Issue on explainable and generalizable deep learning for medical imaging. This IEEE TNNLS Special Issue calls for original and innovative methodological contributions that aim to address the key challenges on explainability and generalizability of deep learning for medical imaging. This IEEE TNNLS Special Issue emphasizes the research and advanced development of the technical aspects of new image analysis methodologies, and all the developed new methods should also be evaluated or validated on real and large-scale medical imaging data.

This Special Issue call received 95 submissions, which went through a two-phase review process. In the first phase, each submission was reviewed by the team of Guest Editors. The first round of review resulted in about 80% of papers being desk-rejected without sending out for peer reviews. The remaining submissions were then handled by the Guest Editors in the second phase of peer review, which might include several rounds of revisions and resubmissions. The Editor-in-Chief validated the acceptance/rejection decisions of all submissions in both review phases. Finally, 14 papers were accepted and included in this Special Issue, and their contributions are summarized by technical aspects as follows. Also, these 14 papers are clustered into four groups according to their research topics.

A. Papers That Contribute to Novel Deep Learning Architectures and Feature Extraction Techniques

The study [A1] introduces a cutting-edge method utilizing adversarial learning combined with graph attention networks, specifically designed for the identification of autism spectrum disorder (ASD) from brain imaging data. The method focuses on leveraging adversarial learning strategies to enhance the robustness of the model against variabilities in the dataset, thereby boosting its generalization capabilities across diverse populations. Graph attention networks are utilized to selectively focus on pertinent features within intricate brain network data, which significantly improves identification accuracy. This approach also incorporates both node and edge features from structural and functional MRI data, allowing for a comprehensive analysis that supports better classification results. The study's framework demonstrates strong potential for aiding in more accurate ASD diagnoses through its innovative use of advanced machine learning techniques on neuroimaging data.

The study [A2] introduces a novel decoupled unbiased teacher (DUT) model for medical object detection within a source-free domain adaptation (SFDA) framework. By using a structural causal model, the DUT approach addresses biases at sample, feature, and prediction levels to enhance the performance of medical object detection without requiring access to the original labeled source data. The model's dual invariance assessment (DIA) generates counterfactual synthetics based on unbiased invariant samples, helping to reduce biases from the dataset. Furthermore, a cross-domain feature intervention (CFI) module deconfounds domain-specific biases to yield unbiased features, while a correspondence supervision prioritization (CSP) strategy refines prediction quality using robust box supervision. Extensive experiments across various medical object detection scenarios demonstrate the superior performance of DUT compared to traditional methods, highlighting its potential in real-world clinical applications without the need for source data. This novel framework shows significant improvements over previous state-of-the-art unsupervised domain adaptation (UDA) and SFDA methods.

The paper by Tan et al. [A3] proposes a Fourier domain robust denoising decomposition and adaptive patch MRI reconstruction (DDAPR) method to address the challenges of noise in MRI data and the need for high-quality reconstruction with low undersampled data. The proposed two-step optimization approach includes a low-rank and sparse denoising reconstruction model (LSDRM) and a robust dictionary learning reconstruction model (RDLRM). The LSDRM employs the proximal gradient method to optimize the model using singular value decomposition and soft threshold algorithms, while the RDLRM uses a block coordinate descent method to optimize the variables with valid closed-form solutions. The proposed method demonstrates better performance than previous approaches in image reconstruction based on compressed sensing or deep learning, showing its effectiveness in reconstructing MRI data with noise and at low sampling rates.

Digital Object Identifier 10.1109/TNNLS.2024.3395937

The paper by Hou et al. [A4] presents GCNs-Net, a deep learning framework based on graph convolutional neural networks (GCNs) for decoding EEG motor imagery signals. The model leverages the functional topological relationship of EEG electrodes to enhance decoding performance. It constructs a graph Laplacian from the absolute Pearson's matrix of signals and applies graph convolutional layers followed by pooling and fully connected layers for prediction. GCNs-Net achieves high accuracy in both personalized and groupwise predictions, demonstrating adaptability and robustness to individual variability. The method's performance is stable across cross-validation experiments, indicating its potential as a significant step towards improved brain–computer interface (BCI) approaches.

The paper by Zhao et al [A5] proposes GMILT, a gated multiple instance learning transformer network for predicting epidermal growth factor receptor (EGFR) mutation status from computed tomography (CT) images. GMILT integrates multi-instance learning and discriminative weakly supervised feature learning to utilize pathological invasiveness information as embeddings. The model is trained and validated on a dataset of 512 patients with adenocarcinoma and tested on three datasets, showing superior performance over existing methods and radiomics-based approaches. GMILT also identifies the "core area" related to EGFR mutation status, potentially guiding biopsies and avoiding false negatives.

B. Papers That Contribute to Developments in Explainable and Generalizable Deep Learning Models

The paper by Zhang et al. [A6] presents an innovative multilayer recurrent neural network model, termed MRFS-MRNN, designed to differentiate human–brain states using EEG data. The model leverages a novel strategy combining multiple random fragment search (MRFS) with recurrent neural networks (RNNs), aimed at enhancing both the explainability and generalizability of the results. Specifically, MRFS-MRNN addresses the challenges of high variability and overfitting in EEG data by introducing a method to select random data fragments and process these through RNN layers, thereby capturing temporal dependencies in the EEG signals. The model achieves high classification accuracy, reaching up to 95.18% for binary and 89.19% for four-category classifications at the individual level, and maintains robust performance across different groupings of subjects. This approach not only provides a high level of accuracy but also improves our understanding of the model's decision-making process, making it a valuable tool for both clinical and research applications in neuroimaging and brain–computer interfaces (BCIs).

The paper by Zhou et al. [A7] introduces a novel deep learning architecture named multibranch CNN with an MLP-Mixer-based feature exploration module (ME-Mixer) for high-performance disease diagnosis. The model harnesses both supervised and unsupervised learning techniques to enhance feature extraction and classification accuracy. Key innovations include manifold embedding and dual MLP-Mixer feature projectors, which effectively amplify the network's capability to manage and interpret complex medical imaging datasets. This enables the ME-Mixer to substantially outperform conventional CNNs and various DNN configurations, demonstrating its potential in clinical settings as a sophisticated yet computationally efficient solution for medical imaging analysis. The architecture's effectiveness is validated through comprehensive evaluations on medical datasets, where it achieves notable improvements in classification accuracy.

The paper by Tang et al. [A8] presents a new model that integrates contrastive learning with hierarchical signed graph pooling to analyze brain networks. The approach distinguishes between positive and negative edge weights in the graph, providing more nuanced insights into brain connectivity patterns. By incorporating a hierarchical structure through signed graph pooling, the model effectively captures both local and global network features. It also leverages contrastive learning to enhance model robustness and generalizability, allowing for more accurate detection of connectivity patterns associated with various neurological conditions. The method demonstrates superior performance in identifying brain network abnormalities across different neuroimaging tasks compared to existing models. These improvements offer significant implications for enhancing diagnostic and therapeutic strategies in neuroscience.

The paper by Zhai et al. [A9] introduces MVCNet, a multiview contrastive network designed to enhance the representation learning of 3-D CT lesions using 2-D views from different spatial orientations. The network addresses the challenge of scarce lesion-level annotations in CT data by employing a contrastive loss function that aggregates views of the same lesion and separates those from different lesions. MVCNet also introduces a mechanism to filter out uninformative negative samples, leading to more discriminative features for downstream tasks. The model achieves state-of-the-art accuracies on the LIDC-IDRI, LNDb, and TianChi datasets for unsupervised representation learning. Notably, when fine-tuned with only 10% of labeled data, MVCNet's performance rivals that of supervised learning models, indicating its effectiveness in scenarios with limited annotations. The study's findings suggest that MVCNet's approach to contrasting multiple 2-D views is a promising strategy for capturing the original 3-D information and improving the utilization of valuable annotated CT data.

C. Papers That Contribute to Robustness and Generalization Through Adversarial Learning and Domain Adaptation

This study [A10] introduces a fuzzy attention neural network (FANN) to improve the segmentation of airways from CT images, critical for diagnosing lung diseases. The FANN uses a fuzzy attention layer with a learnable Gaussian membership function to enhance feature focus and reduce uncertainty, leading to more accurate and continuous segmentation. It also presents a new loss function and evaluation metric, the Airway F-score (AF-score), to assess segmentation quality. Tested on various lung disease datasets, the FANN demonstrates strong generalization and robustness, with the potential to enhance objective disease quantification and reduce reliance on manual methods.

The paper by Zeng et al. [A11] introduces a gradient-matching federated domain adaptation (GM-FedDA) method for brain image classification within a federated learning

framework. The method includes a pretraining stage using one-common-source adversarial domain adaptation (OCS-ADA) to pretrain encoders for reducing domain shift at each target site with the help of a common source domain. The fine-tuning stage employs a gradient-matching federated (GM-Fed) method to update local federated models by minimizing the gradient-matching loss between sites. The proposed GM-FedDA method outperforms other methods in diagnostic classification tasks of schizophrenia and major depressive disorder based on multisite resting-state functional MRI (fMRI), indicating its potential in brain imaging analysis and other fields requiring multisite data while preserving data privacy.

The paper by Bi et al. [A12] proposes a novel deep learning method called hypergraph structural information aggregation generative adversarial networks (HSIA-GANs) for the automatic classification and feature extraction of Alzheimer's disease (AD) using imaging genetic data. The method constructs an ROI-gene hypergraph for each subject to represent the complex associations between regions of interest (ROIs) and genes. HSIA-GAN comprises a generator that creates synthetic hypergraphs and a discriminator that extracts high-order structural information from the hypergraph. The discriminator also fuses high- and low-order structural information to classify AD patients and normal controls accurately. The method demonstrates significant advantages in three classification tasks using data from the AD neuroimaging initiative (ADNI) and successfully extracts discriminant features conducive to better disease classification. The proposed HSIA-GAN provides a comprehensive understanding of AD from a multimodal perspective and has potential applications in the diagnosis and treatment of AD.

D. Papers That Contribute to the Integration of Domain Knowledge for Better Explainability and Generalizability

The research [A13] introduces the anatomy-guided spatiotemporal graph convolutional networks (AG-STGCNs), a new framework for analyzing the brain's functional connectivity using advanced graph convolutional networks informed by anatomical data. AG-STGCNs enhance the interpretability of connectivity patterns between gyri and sulci during various cognitive tasks by incorporating anatomical guidance into the network design. The results demonstrate that this method can effectively identify connectivity patterns associated with specific cognitive functions, providing deeper insights into neural mechanisms underlying different brain states across multiple task domains. The study establishes the potential of AG-STGCNs in clinical settings, offering an advanced yet computationally efficient solution for sophisticated medical imaging applications. This approach significantly advances our understanding of brain function, emphasizing the importance of integrating anatomical and functional data.

The paper by Ma et al. [A14] introduces BAI-Net, an algorithm that employs graph neural networks (GNNs) to subdivide individual cerebral cortices into distinct areas using both brain morphology and connectomes. BAI-Net integrates group priors from a population atlas and adjusts areal probabilities using connectivity fingerprints derived from fiber-tract embedding. The method provides reliable and explainable individualized brain areas, overcoming challenges in clinical applications due to individual brain variability. The study demonstrates that BAI-Net outperforms conventional clustering approaches by capturing heritable topographic variations, associating strongly with individual cognitive behaviors and genetics. This research offers a framework for precise brain area localization, potentially benefiting neuromodulation and personalized treatments.

We would like to express our gratitude to all the authors and coauthors who contributed their high-quality scientific work to this Special Issue on Explainable and Generalizable Deep Learning for Medical Imaging. Our sincere thanks also go to all the reviewers for their insightful feedback and to the current Editor-in-Chief Dr. Yongduan Song and the prior Editor-in-Chief Dr. Haibo He for their vision, guidance, and unwavering support throughout the lifecycle of this Special Issue. We hope that the research presented here will significantly enrich the field, fostering further innovation and engagement among the IEEE TNNLS readers, and advancing the adoption of deep learning technologies in clinical radiology and beyond.

TIANMING LIU, *Guest Editor*
School of Computing
University of Georgia
Athens, GA 30602 USA
e-mail: tianming.liu@gmail.com

DAJIANG ZHU, *Guest Editor*
Department of Computer Science and Engineering
The University of Texas at Arlington
Arlington, TX 76019 USA
e-mail: dajiang.zhu@uta.edu

FEI WANG, *Guest Editor*
Department of Population Health Sciences
Cornell University
Ithaca, NY 14850 USA
e-mail: few2001@med.cornell.edu

ISLEM REKIK, *Guest Editor*
Imperial College London
SW7 2AZ London, U.K.
e-mail: basiralab@gmail.com

XIA HU, *Guest Editor*
Department of Computer Science
Rice University
Houston, TX 77005 USA
e-mail: xia.hu@rice.edu

DINGGANG SHEN, *Guest Editor*
School of Biomedical Engineering
ShanghaiTech University
Shanghai 201210, China
Shanghai United Imaging Intelligence Company Ltd.
Shanghai 201807, China
e-mail: Dinggang.Shen@gmail.com

APPENDIX: RELATED ARTICLES

[A1] Y. Chen et al., "Adversarial learning based node-edge graph attention networks for autism spectrum disorder identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7275–7286, Jun. 2024.

[A2] X. Liu, W. Li, and Y. Yuan, "Decoupled unbiased teacher for source-free domain adaptive medical object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7287–7298, Jun. 2024.

[A3] J. Tan, X. Zhang, C. Qing, and X. Xu, "Fourier domain robust denoising decomposition and adaptive patch MRI reconstruction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7299–7311, Jun. 2024.

[A4] Y. Hou et al., "GCNs-Net: A graph convolutional neural network approach for decoding time-resolved EEG motor imagery signals," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7312–7323, Jun. 2024.

[A5] W. Zhao et al., "GMILT: A novel transformer network that can noninvasively predict EGFR mutation status," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7324–7338, Jun. 2024.

[A6] S. Zhang et al., "An explainable and generalizable recurrent neural network approach for differentiating human brain states on EEG dataset," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7339–7350, Jun. 2024.

[A7] Z. Zhou, M. T. Islam, and L. Xing, "Multibranch CNN with MLP-mixer-based feature exploration for high-performance disease diagnosis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7351–7362, Jun. 2024.

[A8] H. Tang, G. Ma, L. Guo, X. Fu, H. Huang, and L. Zhan, "Contrastive brain network learning via hierarchical signed graph pooling model," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7363–7375, Jun. 2024.

[A9] P. Zhai, H. Cong, E. Zhu, G. Zhao, Y. Yu, and J. Li, "MVCNet: Multi-view contrastive network for unsupervised representation learning for 3-D CT lesions," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7376–7390, Jun. 2024.

[A10] Y. Nan et al., "Fuzzy attention neural network to tackle discontinuity in airway segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7391–7404, Jun. 2024.

[A11] L.-L. Zeng et al., "Gradient matching federated domain adaptation for brain image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7405–7419, Jun. 2024.

[A12] X. Bi, Y. Wang, S. Luo, K. Chen, Z. Xing, and L. Xu, "Hypergraph structural information aggregation generative adversarial networks for diagnosis and pathogenetic factors," *EEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7420–7434, Jun. 2024.

[A13] M. Jiang et al., "Anatomy-guided spatio-temporal graph convolutional networks (AG-STGCNs) for modeling functional connectivity between gyri and sulci across multiple task domains," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7435–7445, Jun. 2024.

[A14] L. Ma et al., "BAI-Net: Individualized anatomical cerebral cartography using graph neural network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7446–7459, Jun. 2024.