Editorial
Introduction for the Special Series (Part I) on AI in Signal & Data Science - Toward Explainable, Reliable, and Sustainable Machine Learning

I. Introduction

MACHINE learning methods are the backbone of AI and data science. As we embrace the era of deep learning and foundation models, marked by improving performance due to enhanced computing hardware and data scaling, numerous challenges remain. In areas such as healthcare, government decision making and scientific fields, there is a pressing need to develop transparent models that generate interpretable results. Equally important is the reliability of these models, which ensures their robustness and their ability to generalize to new datasets. This is particularly crucial in sectors like healthcare and autonomous driving, where the stakes are high. The recent trend of increasing model sizes also raises concerns about the environmental and societal impacts of training and deploying large models. Therefore, it is essential to mitigate these impacts as much as possible.

To address these prevalent issues, we initiated a special series focusing on the development towards Explainable, Reliable, and Sustainable Machine Learning. We received an overwhelming response of 63 submissions following our open calls. After an extensive review process, eight papers were selected for the first installment of this special series. These works cover a range of topics, including fairness and bias analysis in generative models and graph learning, reliable and energy-efficient design of federated learning, and theoretical insights into understanding CNNs, Reservoir computing for RNNs, group imbalance problems, and neural network activation function designs. The second installment is scheduled for publication in December 2024.

II. In This Issue

A. Fair Machine Learning

Research indicates that many machine learning models widely used for automating decision-making tasks may exhibit bias towards sensitive variables such as “race” and “gender”. This is mainly due to the skewed or imbalanced distribution of these attributes in the collected data, often sourced from a limited population. The main object of fair machine learning is to mitigate such issues though interventions before, during or after model training. While classic literature mainly focuses on straightforward classification or regression problems, there is growing attention towards improving fairness in advanced learning paradigms, such as graph learning and generative models.

The paper “Fairness-aware Optimal Graph Filter Design” [A1] focuses on the group fairness in graph learning. In graph learning models, bias can be further amplified using graph topology. The authors introduce a bias metric p that measures the linear correlation between the connectivity patterns and sensitive attributes. Based on theoretical analysis on p’s upper boundary and p-minimizing optimal filter design in the spectral domain, they propose a suite of fairness-aware graph filter designs. These designs are shown to effectively mitigate bias in the data and maintain good computational efficiency. The proposed filters can be implemented as special fair layers within graph-based machine learning and signal processing algorithms at different learning stages, such as inserting them into each Graph Neural Network (GNN) layer and updating them through backpropagation.

The paper “FAIRTL: A Transfer Learning Approach for Bias Mitigation in Deep Generative Models” [A2] focuses on mitigating the bias of generative models. The authors first demonstrate that bias still exists in state-of-the-art GAN models, even when trained on large and diverse datasets, and quantify its magnitude. They then propose a simple and effective transfer learning approach, FairTL, to preserve fairness w.r.t. some sensitive variables. In this approach, the source model is trained using all available data, while unsupervised transfer learning is used to adapt features to a target task using a small set of fair data. Experiments showed that as little as 2.5% of the training data size was sufficient to achieve fairness.

B. Explainable Machine Learning

Explainability involves understanding what a complex machine learning model is doing, which is especially important when there is a mismatch between the model’s original prediction goal and its actual usage in deployment. It allows users to anticipate a model’s inherent limitation, understand when and how to use it, and easily correct by understanding its predictions. In this issue, our primary focuses include providing theoretical

Date of current version 3 July 2024.
Digital Object Identifier 10.1109/JSTSP.2024.3417111

interpretation to existing learning algorithms, introducing new transparent models, and understanding model behavior such as transferability and group imbalance.

The paper "Interpretable Deep Image Classification using Rationally Inattentive Utility Maximization" [A3] interprets CNN for image classification as utility maximizers with information cost. The utility maximizer is learned through system identification for Bayesian decision-making. The numerical results compellingly shows that the proposed model can predict the classification performance of deep CNNs with high accuracy.

The paper "Universal Approximation of Linear Time-Invariant (LTI) Systems through RNNs: Power of Randomness in Reservoir Computing" [A4] tackles the interpretability problem for Reservoir computing. Reservoir Computing is a specialized RNN that addresses issues of vanishing and exploding gradients in standard RNN training. This work aims to provide a theoretical basis that explains its empirical performance enhancement. It offers a clear signal processing interpretation of Reservoir Computing and utilizes this understanding to approximate a generic LTI system. Under this setup, one can analytically characterize the optimal probability density function for configuring the recurrent weights of the underlying RNN in Reservoir Computing, instead of training or randomly generating them. This work marks substantial progress in explainable machine learning by integrating domain knowledge for more efficient learning.

Another important aspect of explainable machine learning in transfer learning is understanding model transferability. i.e. whether the model trained on one dataset is generalizable to another. The paper "Transferability of coVariance Neural Networks" [A5] characterizes the transferability of coVariance Neural Networks (VNNs) across dataset of different dimensions. CoVariance Neural Network is a scale-free graph neural network that operates on the sample covariance matrix of a dataset. The authors demonstrate the quantitative transferability of VNNs over a regression task based on a multiscale dataset of cortical thickness features. To highlight the benefits of VNNs in neuroimaging data analysis, VNNs are also used as regression models for predicting "brain age" from cortical thickness features. The discrepancy between brain age and chronological age, or "brain age gap," can then indicate increased vulnerability or resilience towards neurological disease or cognitive impairments.

The paper "How does promoting the minority fraction affect generalization? A theoretical study of one-hidden-layer neural network on group imbalance" [A6] investigates the group imbalance problem in neural networks from a theoretically interpretable approach. Group imbalance, often seen in machine learning applications like person identification and medical image diagnosis, is a common issue in empirical risk minimization (ERM), where high average accuracy often comes with low accuracy for a minority group. How to improve the minority-group performance, especially when group attributes are unknown, poses a significant challenge. This paper formulates the group imbalance problem with Gaussian Mixture Model and quantify the impact of individual groups on sample complexity for one-hidden-layer neural networks. This simple theoretical model sheds light on previous empirical observations, such as why increasing the fraction of the minority group in the training data does not necessarily improve the generalization performance of the minority group. The paper also offers new insights on improving generalization, such as maintaining group-level co-variance in the medium regime and keeping all means close to zero.

In deep neural network models, non-linear activation functions often render the prediction result uninterpretable. The paper "ENN: A Neural Network with DCT Adaptive Activation Functions" [A7] offers a new way to understand the expressiveness of neural networks from a signal processing perspective. In their novel model Expressive Neural Network (ENN), the authors introduce an activation function using the classic Discrete Cosine Transform (DCT), which can be adapted using back propagation during training. The resulting model offers better interpretation of the network at convergence and demonstrates superior performance in a variety of classification and regression tasks.

C. Reliable Machine Learning

Reliable machine learning aims at developing models that are robust, accurate and able to generalize well to new data. In the context of distributed federated learning, reliability is critically important for the model to generalize across heterogeneous data from different devices.

The paper "Energy-Efficient Connectivity-Aware Learning over Time-Varying D2D Networks" [A8] studies the generalization performance of semi-decentralized federated learning over clustered device-to-device (D2D) networks in an energy constraint scenario. Modeling the network as time-varying digraphs, this work innovatively characterizes the convergence rate of the learning process in terms of the degree distribution of cluster degree. It further uses this finding to design a connectivity-aware FL algorithm and a motion planning algorithm that is highly energy efficient. The proposed framework links its generalization ability of the topological characteristics of the network, contributing to both explainability and reliability in federated learning.

D. Concluding Remarks

In summary, this special series presents the frontier efforts in improving the explainability, reliability and fairness in machine learning. While we cover various learning paradigms using both classic and innovative theoretical tools, many more topics remain to be explored. We will continue to present high-quality work in second part of this special series. Our editorial team envisions a future where AI thrives on the synergy of theory and practice, leading to safer, fairer, and trustworthy applications in signal processing and data science.

I thank all the authors and reviewers whose contributions have made this special series possible. My special thanks go to Mikaela Langdon at the IEEE publications office for her tireless and relentless efforts in keeping the special series on track. I am very grateful to our special series editorial team for this issue: Zheng-Hua Tan (Aalborg University), Bhuvana Ramabhadran (Google), Shuvra Bhattacharyy (University of Maryland), Wenbo Ding (Tsinghua University), Yonina C. Eldar (Weizmann Institute of Science), Zhu Han (University of Houston), Yi Ma (Hong Kong University), Helen Meng (Chinese University of Hong Kong), Maria Sabrina Greco (University of Pisa), Dacheng Tao (University of Sydney), Zhou Wang (University of Waterloo), Aylin Yener (Ohio State University). Without their tireless and meticulous work, this issue would not have been possible.

XIAO-PING (STEVEN) ZHANG, *Editor-in-Chief*
Institute of Data and Information,
Tsinghua Shenzhen International Graduate School
Tsinghua University
Shenzhen, China

APPENDIX
RELATED ARTICLES

[A1] O. D. Kose, G. Mateos, and Y. Shen, "Fairness-aware optimal graph filter design," *IEEE J. Sel. Top. Signal Process.*, vol. 18, no. 2, pp. 142–154, Mar. 2024, doi: 10.1109/JSTSP.2024.3350508.

[A2] C. T. H. Teo, M. Abdollahzadeh, and N.-M. Cheung, "FAIRTL: A transfer learning approach for bias mitigation in deep generative models," *IEEE J. Sel. Top. Signal Process.*, vol. 18, no. 2, pp. 155–167, Mar. 2024, doi: 10.1109/JSTSP.2024.3363419.

[A3] K. Pattanayak, V. Krishnamurthy, and A. Jain, "Interpretable deep image classification using rationally inattentive utility maximization," *IEEE J. Sel. Top. Signal Process.*, vol. 18, no. 2, pp. 168–183, Mar. 2024, doi: 10.1109/JSTSP.2024.3381335.

[A4] S. Jere, L. Zheng, K. Said, and L. Liu, "Universal approximation of linear time-invariant (LTI) systems through RNNs: Power of randomness in reservoir computing," *IEEE J. Sel. Top. Signal Process.*, vol. 18, no. 2, pp. 184–198, Mar. 2024, doi: 10.1109/JSTSP.2024.3387274.

[A5] S. Sihag, G. Mateos, C. McMillan, and A. Ribeiro, "Transferability of covariance neural networks," *IEEE J. Sel. Top. Signal Process.*, vol. 18, no. 2, pp. 199–215, Mar. 2024, doi: 10.1109/JSTSP.2024.3378887.

[A6] H. Li, S. Zhang, Y. Zhang, M. Wang, S. Liu, and P.-Y. Chen, "How does promoting the minority fraction affect generalization? A theoretical study of one-hidden-layer neural network on group imbalance," *IEEE J. Sel. Top. Signal Process.*, vol. 18, no. 2, pp. 216–231, Mar. 2024, doi: 10.1109/JSTSP.2024.3374593.

[A7] M. Martinez-Gost, A. Pérez-Neira, and M. Á. Lagunas, "ENN: A neural network with DCT adaptive activation functions," *IEEE J. Sel. Top. Signal Process.*, vol. 18, no. 2, pp. 232–241, Mar. 2024, doi: 10.1109/JSTSP.2024.3361154.

[A8] R. Parasnis, S. Hosseinalipour, Y.-W. Chu, M. Chiang, and C. G. Brinton, "Energy-efficient connectivity-aware learning over time-varying D2D networks," *IEEE J. Sel. Top. Signal Process.*, vol. 18, no. 2, pp. 242–258, Mar. 2024, doi: 10.1109/JSTSP.2024.3374591.