STEVENS
INSTITUTE OF TECHNOLOGY
1870

Ph.D. DISSERTATION DEFENSE

Candidate:	Bingyang Wen
Degree:	Doctor of Philosophy
School/Department:	Charles V. Schaefer, Jr. School of Engineering and Science / Department of Electrical and Computer Engineering
Date:	Tuesday, April 22nd, 2025
Time/Location:	2:30 pm – 4:00 p.m. / Burchard 219
Title:	Trustworthy Deep Learning via Causal Intervention, Information Theory, and Visual-Linguistic Attention and Applications to Alzheimer's Disease Detection
Chairperson:	Dr. Koduvayur Subbalakshmi, Department of Electrical and Computer Engineering, School of Engineering and Science
Committee Members:	Dr. Shucheng Yu, Department of Electrical and Computer Engineering, School of Engineering and Science Dr. Yada Zhu, IBM T. J. Watson Research Center Dr. Zhuo Feng, Department of Electrical and Computer Engineering, School of Engineering and Science Dr. Ping Wang, Department of Computer Science Engineering, School of Engineering and Science

ABSTRACT

Building trustworthy AI systems is essential in domains where model decisions carry significant consequences. This dissertation explores how principles from causal inference, information theory, and attention mechanisms can be integrated into deep learning frameworks to enhance interpretability, robustness, and transparency—core aspects of trustworthiness. Four contributions spanning both textual and tabular modalities are presented. First, an information-theoretic framework is proposed to evaluate the faithfulness of attention mechanisms in deep neural networks. Using mutual information, we demonstrate that attention distributions in certain encoder-attention configurations (e.g., BiLSTM with additive attention) correlate strongly with the most informative latent representations. This finding offers theoretical and empirical validation for attention as a mechanism for localized interpretability. Second, we introduce the One-Intervention Causal Explanation (OICE) method to generate faithful, fine-grained attributions in Alzheimer's Disease (AD) detection from speech transcripts. OICE models intra-feature causal dependencies and decomposes total effects into direct and indirect components. Applied to part-of-speech features, OICE identifies previously unreported linguistic biomarkers and offers cognitively aligned interpretations for clinical insight. Third, we present Causal-TGAN, a generative adversarial framework for causal-aware data synthesis. Causal-TGAN integrates discrete and continuous mode-specific encoding with a hybrid causal discovery mechanism, enabling the generation of synthetic tabular data that preserves the underlying causal structure. This supports the development of more robust, causally grounded models and enhances the reliability of synthetic datasets used in downstream tasks. Finally, we develop a set of interpretable visual-linguistic attention features to model cognitive behavior in AD detection tasks. By combining eye-tracking-inspired attentional cues with interpretable modeling techniques, we achieve high predictive performance and uncover cognitively grounded indicators of AD that extend beyond traditional linguistic markers. Collectively, these contributions demonstrate a multi-perspective approach to improving the trustworthiness

of deep learning systems through interpretable modeling, causal reasoning, and robust generative techniques, with applications to real-world challenges in health and behavioral AI.