Personal Statement (Post-Tenure)

Dongxiao Zhu
Webpage: https://dongxiaozhu.github.io/

1 Introduction

My recent research thrust lies in trustworthy Artificial Intelligence (AI) algorithms with community driven innovations for social good, such as health and wellbeing, mobility equity, better cybersocial behavior, security and privacy. My recent foundational AI research has been published in some of the most competitive **AI conferences** tracked by csrankings.org where my students are the first authors and I am the **senior** and **corresponding author**, including fairness AI algorithms (AAAI-20 [2] and IJCAI-22 [22]), explainable AI algorithms (NuerIPS-22 [24], IJCAI-20 [18] and IJCAI-21 [19]) and adversarial robust AI algorithm (AAAI-21 [12]), just to name a few. Particularly, NeurIPS, AAAI and IJCAI are ranked at **2nd**, **4th**, and **9th** top publication venues by Google Scholar in the category of Artificial Intelligence. I have been supported by **four** NSF research grants and **one** NIH research grant totaling **over 4.0 million** where I am the PI (NSF) or MPI (NIH) on over **1.7 million** federal research grants. I summarize my **post-tenure highlights** below (2015-Present).

Awards and Recognitions:

- WSU College of Engineering: Faculty Excellence in Research Award, 2021-22
- WSU College of Engineering: Faculty Excellence in Teaching Award, 2016-17
- Wiley Journal of Biophotonics: Top Cited Article 2020-2021

Post-Tenure NSF/NIH Research Grants:

- NSF/IIS 2211897, "Collaborative Research: HCC: Small: Understanding Online-to-Offline Sexual Violence through Data Donation from Users", 10/01/2022 - 09/30/2026, Total Amount, $600,000, My Role: **PI** (33%).
- NIH/R61HD105610, "Severity Predictors Integrating salivary Transcriptomics and proteomics with Multi neural network Intelligence in SARS-CoV2 infection in Children (SPITS MISC)", 01/01/2021 - 12/31/2023, Total Amount, $1,433,469, My Role: **MPI** (33%).
- NSF/CNS 2043611, "SCC-CIVIC-PG Track A: Leveraging AI-assist Microtransit to Ameliorate Spatiotemporal Mismatch between Housing and Employment." 01/01/2021 - 12/31/2021, Total Amount, $49,898, My Role: **PI** (25%).
- NSF/IIS 1724227, "S&AS: INT: Autonomous Battery Operating System (ABOS): An Adaptive and Comprehensive Approach to Efficient, Safe, and Secure Battery System Management", 01/15/2017-12/31/2023, Total Amount, $1,249,998, My Role: **Senior Personnel** (10%).
- NSF/IIS 1637312, "S&CC:Promoting a Healthier Urban Community: Prioritization of Risk Factors for the Prevention and Treatment of Pediatric Obesity", Total Amount, $199,996, 09/01/2016-08/31/2019, Total amount: $199,996, My Role: **Co-PI** (33%).

Leadership Roles & Internal Service:

- Founding Director: Wayne AI Research Initiative (2021 – present)
- Co-Director: Master Program in AI, College of Engineering, Algorithms and Systems Track (2022 - present)
- Chair: Computer Science Faculty Search Committee (2021 – 2022)
- Site Director: NSF-funded Michigan Trustworthy AI Institute at Wayne State University (2021-present, in-progress)
- Director: Computer Science Graduate Program (2018 – 2020)

My **foundational AI research** focuses on **explanability**, **adversarial robustness**, and **fairness** of deep neural networks (DNNs), collectively referring to trustworthy AI. The lack of trustworthiness of AI has been measured across technical metrics such as accuracy, robustness, fairness, explainability, privacy, accountability, and ethics. Recent literature has focused much on the marked performance improvement of DNN models, nevertheless overlooking trustworthiness when deploying the AI models in safety and security-critical real-world scenarios. Despite some preliminary encouraging progress, current research outcomes illuminate a continuing need in building theoretical and methodological foundations to forge an authentic trustworthy machine-human better partnership. I posit that closing the 'trust gap' requires tackling a **grand challenge**: the interrelationships of various technical trustworthiness metrics and their positive or negative synergies remain as elusive as ever. For example, there exists a common belief that an AI algorithm would have to hamper one metric in order to optimize another. My foundational AI research will lay the foundations of a new trustworthy AI plane from three dimensions: (i) innovated trustworthiness metrics and their interrelationships, (ii) machine-centric algorithmic advances, and (iii) human-inspired algorithmic innovations.

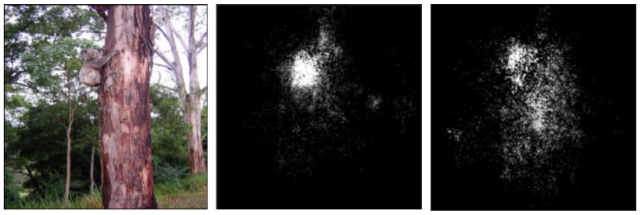

Figure 1: My AGI algorithm [19] mimics how human visual perception detects Koala and discriminates it from the visually similar Polecat.

My foundational research subsequently accelerates my **use-inspired research** to tackle some of more pressing community-driven issues (e.g., disparities in health and mobility), via more efficiently leveraging the limited resources to improve accessibility of the socially vulnerable groups, fostering a thriving community. It has been well-documented that AI models can inherit pre-existing biases and exhibit discrimination against already-disadvantaged or marginalized social groups living in the socially vulnerable regions; be vulnerable to security and privacy attacks that deceive the models and leak the training data's sensitive information; and make hard-to-justify predictions with a lack of transparency. In Detroit and other urban cities, vulnerable communities have been plagued by salient health disparities, food insecurity, housing/job instability, excessive air/water pollution, mobility impairment, and poor-quality K-12 education. Trustworthy AI holds a strong promise to cost-effectively mitigate these pressing problems at scale. I am passionate about leveraging AI for social good research, development and community outreach.

2 Foundational Trustworthy AI Research

DNNs are complex nonlinear functions parameterized by model weights to map inputs to outputs. Explanability and Adversarial Robustness are synergistic concepts in that explanability can indicate adversarial robustness and *vice versa* whereas Fairness is an antagonistic concept since the system has to compromise overall performance for the minority groups in exchange for equity across all groups and individuals.

2.1 Explainability of DNNs

Explainable AI (XAI) research attempts to understand how information flows from input to output. I have made original contributions to both directions of current XAI research: **explainable model prediction** and **interpretable feature mapping**. In explaining AI model prediction, I have developed a general attribution based algorithm, named **Adversarial Gradient Integration (AGI)** [19], to explain the contribution of each pixel, each word and/or each variable to the DNN's prediction, such as image class or sentence sentiment. The existing gradient based XAI approaches such as IG, SHAP and DeepLift suffer from arbitrary choices of the reference examples whereas my AGI approach leverages adversarial examples as references by integrating gradients from adversarial to the benign examples (e.g., images or sentences). It opens a new avenue in XAI research since this work is not only able to explain why the example is predicted as the true class label but also explain why it is so by contrasting with different reference examples. For example, when explaining a Convolutional Neural Network (CNN) based prediction using ImageNet data set with 1000 classes (Figure 1), a prediction of *Koala* is explained as a heatmap with a Koala saliency when contrasting with visually dissimilar classes such as Toilet Tissue whereas the heatmap contains both Koala and Tree Trunk saliency when contrasting with a visually similar class of *Polecat*. Interestingly, our AGI's explanation mimics how human

vision perceives natural images to *detect Koala* and *tell Koala apart from Polecat*. Different from other established feature map based XAI approaches such as Gradient CAM or Gradient CAM ++ that are limited to explain CNN prediction, my AGI algorithm **is sufficiently general** to all DNN prediction regardless of natural language, natural images or structured variables as inputs.

By observing previous path integration approaches for interpretation, such as AGI, one can see that they share the same sprite that smoothing the gradient via gradient accumulation. In AGI and others, it is obtained by summing up all gradient integration results from different adversarial points to the input point. I call this type of methods as accumulation based methods, which demonstrate that gradient accumulation is a promising direction for DNN interpretation. However, there are two obstacles that limit the stability of these methods. First, the *ad hoc* choices of baseline (reference) points could yield completely different interpretations. Second, the integration path is not unique, hence causes disagreement between different path choices. In order to overcome these weaknesses of existing accumulation based methods, I propose Negative Flux Aggregation (NeFLAG) [20] that relies on neither baselines nor paths to calculate accumulation by borrowing the concept of divergence and flux in vector analysis. One key contribution of this approach is that it requires neither baseline nor integration path for calculation, eliminating the derived instability issue. Furthermore, it is also the first approach that utilizes the concept of divergence and flux for DNN interpretation. Through extensive experiments, NeFLAG demonstrates superior interpretation performance compared to previous accumulation methods (IG, AGI, etc). Examples of attribution heatmaps obtained by NeFLAG, AGI and IG methods. Compared to AGI and IG, we observe that NeFLAG's output heatmap has clearer shapes and focuses more densely on the target object (Figure 2).

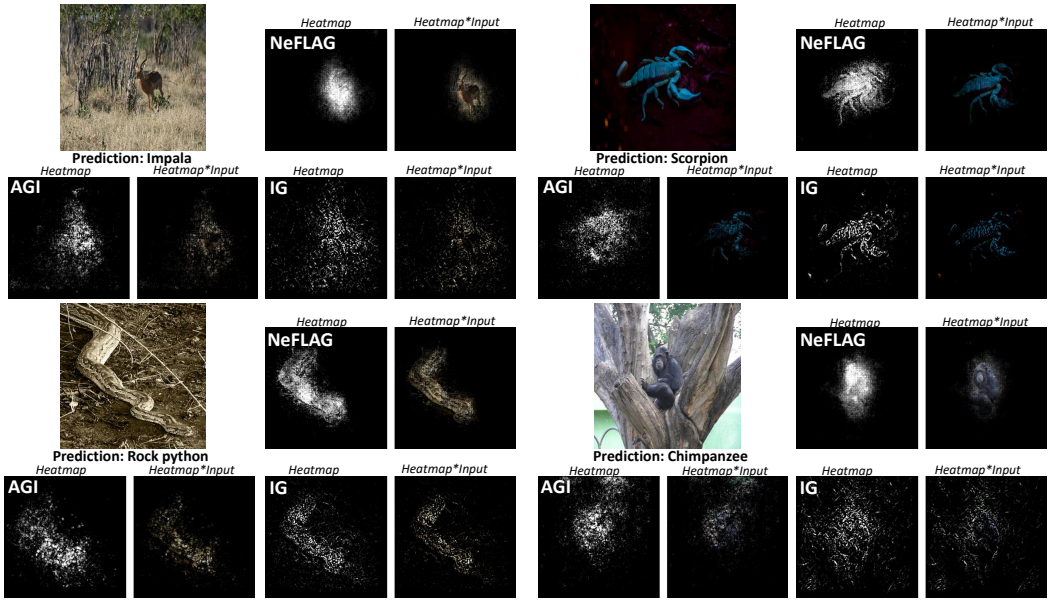

Figure 2: Examples of attribution maps obtained by NeFLAG, AGI and IG methods. The underlying prediction model is InceptionV3. Compared to AGI and IG, we observe that NeFLAG's output heatmap has clearer shapes and focuses more densely on the target object.

In the aforementioned XAI research, explaining DNN prediction follows after the model prediction. In another line of XAI research, DNN architecture itself can become interpretable. I made original contribution, named Interpretable Feature Mapping (IFM) [18], in developing an interpretable AI architecture allowing interpretable feature mapping and demonstrated applications in explainable recommender's system. Notably, this is one of the first papers proposing novel metrics for evaluating of the quality of AI explanation. Specifically, our novel feature mapping approach maps the uninterpretable general features onto the interpretable aspect features, achieving both satisfactory accuracy and explainability in the recommendations by simultaneous minimization of rating prediction loss and interpretation loss. To quantitatively evaluate the explainability, we propose two new evaluation metrics specifically designed to assess our model's explanation quality in terms of general preferences and specific preferences, respectively. In Figure 3, the movie 120 is high-rated by both users 65 and 74, however, with differential explanations: the former user preference is more on the *Action*

genre whereas the latter is more on *Sci-Fi* and *War*. On the other hand, the same movie is low-rated by user 67 mainly due to the dislike of *Action* genre.

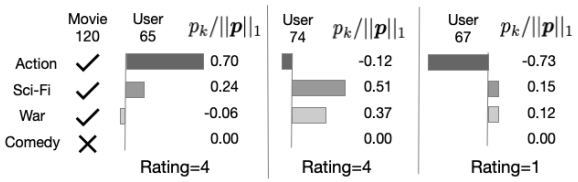

Figure 3: My IFM algorithm [18] qualitatively and quantitatively explains recommendation.

Both AGI and IFM work have been recently published at the top AI conferences IJCAI-21 and IJCAI-20 with stuningly low acceptance rates of 13.9% and 12.6% selected from over 4,000 submissions. In addition, I also developed a novel multiple attention network (MAN) algorithm to explain the model's sentiment prediction from a number of different aspects and published it at IJCNN-20 [23]. Different from existing Aspect Based Sentiment Analysis (ABSA) algorithms, my algorithm does not need the words to be tagged instead it uses the aspect specific sentiment label as the weak supervision to learn the aspect level ratings and explain them by designing a Multiple Attention Network (MAN). MAN explains customer's review of restaurant from four *aspects*: Atmosphere, Food, Service and Value, by visualizing the multiple attention map for each word in the review text. The words "food" and "good" are the most important ones for positive polarity from the aspect of Food whereas "service", "terrible", "wait", "long" and "time" are the most important words for negative polarity from the aspect of Service with large negative scores, dominating over the positive polarity from the aspect of Food, and leading to the overall negative polarity.

We currently continue on XAI research in developing a novel method to explain Transformer's prediction. Transformer has becoming the new standard AI architecture for both NLP and CV tasks due to its robustness and manageable parallel training. However, it is still an open problem to understand a Transformer prediction due to the complexity of the stacked multi head self-attention architectures. We had overcome the major technique bottleneck in the formulation by finding the partial derivatives of each multiple attention heads for each word's embedding. Current experiment results using large-scale NLP benchmark data sets SST-2 are very promising.

2.2 Adversarial robustness of DNNs

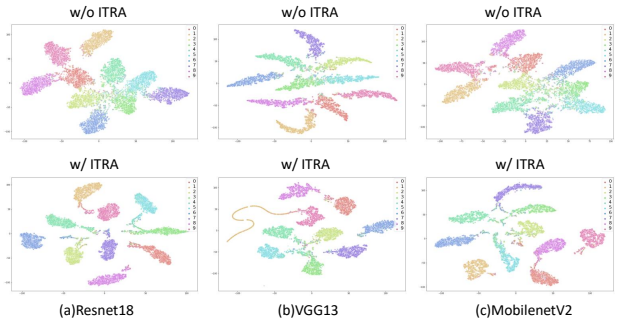

Figure 4: Visualization of my ITRA algorithm for [3] learning compact feature representation.

Despite their success, CNNs are highly vulnerable to adversarial examples. With imperceptibly small perturbation added to a clean image, adversarial samples can drastically change models' prediction, resulting in a significant drop in DNN's predictive performance. I develop novel feature representation learning approaches to improve the adversarial robustness of the DNN via **learning compact feature representation**. I achieve this goal via designing novel **natural training** and **adversarial training** schemes. An important factor impacting feature representation learning is sampling mini-batches. Although a mini-batch of any size is an unbiased estimator of the true loss gradient, large variance between different mini-batches exists. Motivated by this, I developed a novel algorithm, named In-Training Representation Alignment (ITRA), to regularize the training loss with the intra-class mini-batch variations by aligning a pair of feature distributions from two mini-batches to each other, resulting in compact feature representations. Compared with the existing method such as Center Loss, ITRA does not rely on a center of the class; the latter may not be available during each epoch of training before it converges. Figure 4 visualizes the quality of feature representation with or without ITRA. This work received positive reviews from ICML-21, e.g., "Overall, I see values and interesting points and I believe the paper is worth publication at ICML or other similar venues.", with a request for more experiments. I am currently adding more experiments to resubmit it soon. Importantly, ITRA is **sufficiently general** to enhance any mini-batch based training objectives (both natural and adversarial training) for learning compact feature representation.

Adversarial training with ITRA is expected to greatly enhance robustness of DNNs, however, the computing resource demand for sampling the extra mini-batch and/or using augmented adversarial training set

can often limit its scalability. Inspired by the unique insight that the predictive behavior of DNN on adversarial samples that the former tends to misclassify the latter into the first several most probable classes, I formulate the problem of improving adversarial robustness by proposing a new loss function, i.e., Probabilistically Compact (PC) loss with logit constraints [12] to improve DNN's adversarial robustness. The former enforces the gap between the true class and the most probable false classes at the penultimate layer where the latter complements PC loss, which suppresses logit to ensure that the gaps are not only large, but also difficult to be crossed (Figure 5). These two components are systematically integrated and simultaneously optimized during training process. Our PC loss can be used as a drop-in replacement of the essentially any loss to supervise DNN training without extra procedure nor additional computational burden for improving adversarial robustness. The PC loss paper was published in AAAI-21 with a 21.4% acceptance rate from a **record high** 8,000 submissions. One of my major ongoing work is to establish the theoretic linkage between DNN's explanability and adversarial robustness. Adversarial examples are not only able to attack the DNN's prediction but also DNN's explanation. How to design the novel feature representation learning approach to mitigate the attacks on both prediction and explanation? I address the problem in both In-Distribution (ID) and Out-Of-Distribution (OOD) settings by learning both semantic and discriminating features.

2.3 Fairness of DNNs

Fairness of DNN prediction is another active area of trustworthy AI research including both group fairness and individual fairness. A fairness-indifferent AI model often overlooks **class-imbalance** and/or **group-imbalance** issues, and hence is biased against minority group or minority class.

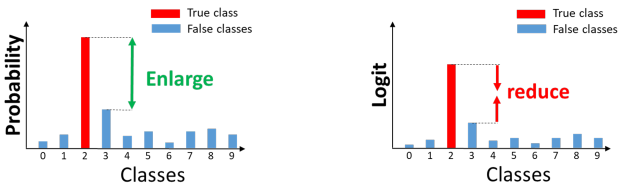

Figure 5: My algorithm of probablistically compact (PC) loss [12] with logistic constraint.

To mitigate **group-imbalance** issue, I developed Multi-Task Learning (MTL) algorithms to predict time-to-event [26] and ordinal outcomes [27]. In the former, I proposed a novel multi-task survival analysis approach that takes advantage of both censored instances and task relatedness towards mitigating biases derived from group-imbalance. Specifically, based on two common used task relatedness assumptions, i.e., low-rank assumption and cluster structure assumption, I formulated two concrete models, the global COX-TRACE and the local COX-cCMTL models, which can be further regularized to ensure group and individual fairness, respectively ([26], ICDM-17). In [27], I developed another MTL algorithm for heterogeneous data with ordinal outcomes instead of more common categorical or continuous outcomes ([27], DMKD). These work demonstrated MTL as an effective framework to overcome group-imbalance challenges to DNN debiasing. To overcome **class-imbalance** issue, I developed a cost-sensitive approach to design optimal weight schemes ([2], AAAI-20). I explicated the learning property of logistic and softmax loss functions by analyzing the necessary condition (e.g., gradient equals to zero) after training converges. Our analysis provides explanations for understanding (1) quantitative effects of the class-wise reweighting mechanism: deterministic effectiveness for binary classification using logistic loss yet indeterministic for multi-class classification using softmax loss; and (2) disadvantage of logistic loss for single-label multi-class classification via one-vs.-all approach, which is due to the averaging effect on predicted probabilities for the negative class (e.g., non-target classes) in the learning process. With the disadvantage and advantage of logistic loss disentangled, I developed a novel debiasing algorithm to reweight logistic loss for multi-class classification to tackle the class-imbalance issue, noted as **"an elegant proof"** by reviewers. This foundational AI research was published in AAAI-20 with a low acceptance rate 20.6% from **nearly 8,000 submissions**.

In addition to the model regularization approaches, I further developed data augmentation approaches, Counterfactual Interpolation Augmentation (CIA) [22], to enhance both fairness and explainability of DNN. Bias in the training data can jeopardize fairness and explainability of DNN prediction on test data. My CIA approach attempts to debias the training data by d-separating the spurious correlation between the target variable and the sensitive attribute. CIA generates counterfactual interpolations along a path simulating the distribution transitions between the input and its counterfactual example. CIA as a pre-processing approach enjoys two advantages: First, it couples with either plain training or debiasing training to markedly increase fairness over the sensitive attribute. Second, it enhances the explainability of deep neural networks by generating attribution maps via integrating counterfactual gradients. I demonstrate the superior performance of

the CIA-trained deep neural network models using qualitative and quantitative experimental results.

Figure 6 shows a qualitative example demonstrating CGI is capable of generating higher quality attribution map. Note that there is substantial noise in IG's attribution map due to the arbitrary choice of the baseline. Both CGI and BlurIG have captured the meaningful facial features (e.g., eyes and lips) related to the target attribute $HeavyMakeup$. While CGI's attribution map has higher density masks demonstrating a focus more densely on these facial features.

2.4 Classical machine learning and data mining research

My classical machine learning research focus on **feature selection from high dimension data** for both supervised classification ([5], PRL) and regression tasks ([4], PRL), respectively. The former solved unique problem of allowing the feature sets for each class to be overlapping and the latter is one of the first algorithms to tackle multi-modal continuous outcomes. My classical data mining research focuses on **model-based clustering** with mixed feature types ([28], SADM) as well as over-dispersion in clustering network nodes ([29], ICMLA-17). The latter paper wins the **Best Paper Top 3 Award** at ICMLA-2017. I also extended the solution to allow the clusters to be overlapping, and the algorithm, named supervised biclustering ([17], ICMLA-17), wins the **Best Poster Top 3 Award** at ICMLA-2017.

3 Use-inspired Trustworthy AI Research

Healthcare data is featured with high dimension, heterogeneity and scarce labels. My AI applications in healthcare research lies in patient subgroup identification and risk factor prioritization. To overcome label scarcity issue, I used primary labels together with auxiliary labels as regularization to learn features and improve prediction performance [7, 8]. I also tackle the label scarcity issue using semi-supervised [16] and active learning [15] approaches using EHR data. To address the data heterogeneity issue, I developed multi-task deep feature learning approaches to learn general features for predicting population-wide and task-specific features for predicting group-specific health outcomes [8]. When patient groups are undefined, I generalized it with a deep mixture neural network model to predict health outcomes for latent groups [9]. My student received the **Best Student Paper Award** from American Medical Informatics Association (AMIA-20) Summit on Clinical Research Informatics for this work [9]. To prioritize risk factors from high dimension of features for different yet related tasks such as patient groups, I employed MTL approaches to achieve it through regularization across the groups. [25, 30] developed linear model based MTL approaches using $l_{2,1}$ norm as the sparsity regularization across tasks, and applied the methodology to prioritize risk factors affecting obesity [25, 30] and heart failure [6] in different patient groups.

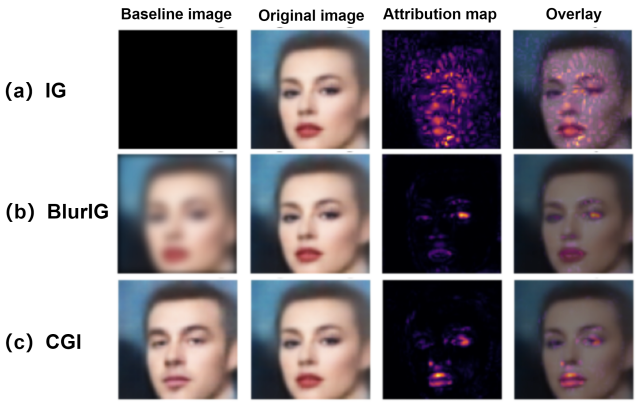

Figure 6: Examples of attribution maps obtained by IG, BlurIG, and CGI. The target attribute is $HeavyMakeup$.

I have developed AI algorithms to improve CNN's robustness against adversarial attacks on medical imaging visualization and interpretation system, both generated from ID examples [14] and OOD examples [13] with a new definition and formulation of adversarial risk. In addition to security-critical applications, I also developed AI workflow to automatically generate radiologist reports from chest radiographs [10] and estimate the age of pre-natal babies from MRI images [1, 22]. Besides healthcare, I developed novel AI approaches for **social good**. For example, I have developed efficient and tiny AI approaches that are robust to adversarial attacks to deploy on resource constrained wearables and sensors including on-device COVID detection using CXR images [11]. I have developed novel progressive distillation techniques to compress the AI models with hundreds of millions of parameters to just thousands so that it can be used for edge devices for real-time inference [21]. I have developed multi-modal geospatial feature learning algorithms for predicting human mobility patterns and leverage it for Point of Interest (POI) recommendation.

4 Leadership, Mentorship, and Service

I have been assuming leadership roles in both education and community service. I have served as the Graduate Program Director for our computer science (CS) graduate programs from 2018 to 2020. During my tenure, I have substantially restructured our MS program with both depth and breadth requirements. For latter, I have led the effort to create two new CS concentrations: AI and Autonomous Driving (AD). Recently, I am leading the effort to develop the Wayne AI Research Initiative and in the progress of creating the Michigan Center for Trustworthy AI - Wayne State University. I serve as the faculty search committee chair to hire a AI researcher to join our current research group and I am a co-director of the Master Program in AI, algorithms and systems track.

I have regularly served on various top AI conferences (e.g., AAAI, IJCAI, ICML, NuerIPS, ICLR, ACL, EMNLP, MICCAI) as (Senior) Program Committee members. I co-chaired the New Frontiers in Adversarial Machine Learning workshop @ICML-22. In addition, I am the Associate Editor for several biomedical informatics journals (e.g., *BMC Genomics*, *Frontiers in Genetics*, *Scientific Reports*).

References

[1] X. Li, J. Hect, M. Thomason, and D. Zhu, "Interpreting age effects of human fetal brain from spontaneous fmri using deep 3d convolutional neural networks," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 1424–1427. 3

[2] X. Li, X. Li, D. Pan, and D. Zhu, "On the learning property of logistic and softmax losses for deep neural networks," in *AAAI*, 2020, pp. 4739–4746. 1, 2.3

[3] X. Li, D. Pan, X. Li, and D. Zhu, "Improve sgd training via aligning min-batches," *arXiv preprint arXiv:2002.09917*, 2020. 4

[4] X. Li and D. Zhu, "Robust feature selection via l2, 1-norm in finite mixture of regression," *Pattern Recognition Letters*, vol. 108, pp. 15–22, 2018. 2.4

[5] X. Li, D. Zhu, and M. Dong, "Multinomial classification with class-conditional overlapping sparse feature groups," *Pattern Recognition Letters*, vol. 101, pp. 37–43, 2018. 2.4

[6] X. Li, D. Zhu, M. Dong, M. Z. Nezhad, A. Janke, and P. D. Levy, "Sdt: A tree method for detecting patient subgroups with personalized risk factors," *AMIA Summits on Translational Science Proceedings*, vol. 2017, p. 193, 2017. 3

[7] X. Li, D. Zhu, and P. Levy, "Predictive deep network with leveraging clinical measure as auxiliary task," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2017, pp. 786–791. 3

[8] ——, "Leveraging auxiliary measures: a deep multi-task neural network for predictive modeling in clinical research," *BMC medical informatics and decision making*, vol. 18, no. 4, p. 126, 2018. 3

[9] ——, "Predicting clinical outcomes with patient stratification via deep mixture neural networks," *AMIA Summits on Translational Science Proceedings*, vol. 2020, p. 367, 2020. 3

[10] X. Li, R. Cao, and D. Zhu, "Vispi: Automatic visual perception and interpretation of chest x-rays," in *International Conference on Medical Imaging with Deep Learning: MIDL 2020*. IEEE, 2020, p. XXX. 3

[11] X. Li, C. Li, and D. Zhu, "Covid-mobilexpert: On-device covid-19 patient triage and follow-up using chest x-rays," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 1063–1067. 3

[12] X. Li, X. Li, D. Pan, and D. Zhu, "Improving adversarial robustness via probabilistically compact loss with logit constraints," in *AAAI*, 2021. 1, 2.2, 5

[13] X. Li, D. Pan, and D. Zhu, "Defending against adversarial attacks on medical imaging ai system, classification or detection?" in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, p. in press. 3

[14] X. Li and D. Zhu, "Robust detection of adversarial attacks on medical images," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 1154–1158. 3

[15] M. Z. Nezhad, N. Sadati, K. Yang, and D. Zhu, "A deep active survival analysis approach for precision treatment recommendations: Application of prostate cancer," *Expert Systems with Applications*, vol. 115, pp. 16–26, 2019. 3

[16] M. Z. Nezhad, D. Zhu, X. Li, K. Yang, and P. Levy, "Safs: A deep feature selection approach for precision medicine," in *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*. IEEE, 2016, pp. 501–506. 3

[17] M. Z. Nezhad, D. Zhu, N. Sadati, K. Yang, and P. Levi, "Subic: A supervised bi-clustering approach for precision medicine," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2017, pp. 755–760. 2.4

[18] D. Pan, X. Li, X. Li, and D. Zhu, "Explainable recommendation via interpretable feature mapping and evaluation of explainability," p. IJCAI, 2020. 1, 2.1, 3

[19] D. Pan, X. Li, and D. Zhu, "Explaining deep neural network models with adversarial gradient integration," p. IJCAI, 2021. 1, 1, 2.1

[20] ——, "Interpreting deep neural network models with negative flux aggregation," p. Under review, 2022. 2.1

[21] Y. Qiang, S. T. S. Kumar, M. Brocanelli, and D. Zhu, "Tiny rnn model with certified robustness for text classification," *IJCNN-22*, 2022. 3

[22] Y. Qiang, C. Li, M. Brocanelli, and D. Zhu, "Counterfactual interpolation augmentation (cia): A unified approach to enhance fairness and explainability of dnn," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2022, pp. 732–739, main Track. 1, 2.3, 3

[23] Y. Qiang, X. Li, and D. Zhu, "Toward tag-free aspect based sentiment analysis: A multiple attention network approach," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8. 2.1

[24] Y. Qiang, D. Pan, C. Li, X. Li, R. Jang, and D. Zhu, "Attcat: Explaining transformers via attentive class activation tokens," *NuerIPS-22*, 2022. 1

[25] L. Wang, M. Dong, E. Towner, and D. Zhu, "Prioritization of multi-level risk factors for obesity," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, p. In Press. 3

[26] L. Wang, Y. Li, J. Zhou, D. Zhu, and J. Ye, "Multi-task survival analysis," in *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2017, pp. 485–494. 2.3

[27] L. Wang and D. Zhu, "Tackling ordinal regression problem for heterogeneous data: sparse and deep multi-task learning approaches," *Data Mining and Knowledge Discovery*, pp. 1–28, 2021. 2.3

[28] L. Wang, D. Zhu, and M. Dong, "Clustering over-dispersed data with mixed feature types," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 11, no. 2, pp. 55–65, 2018. 2.4

[29] L. Wang, D. Zhu, M. Dong, and Y. Li, "Modeling over-dispersion for network data clustering," in *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*. IEEE, 2017, pp. 42–49. 2.4

[30] L. Wang, D. Zhu, E. Towner, and M. Dong, "Obesity risk factors ranking using multi-task learning," in *Biomedical & Health Informatics (BHI), 2018 IEEE EMBS International Conference on*. IEEE, 2018, pp. 385–388. 3