Xuzhe Zhi

alina-zhi.github.io
xzhi2@jh.edu

Education

Johns Hopkins University — Baltimore, MD
M.S.E. in Computer Science (GPA: 4.0/4.0) — Aug. 2024 – Present

University of California, San Diego — La Jolla, CA
B.S. in Data Science (Machine Learning Track, GPA: 3.97/4.0) — Sept. 2020 - Dec. 2023

Research Interests

Computer Vision, Natural Language Processing, AI in Healthcare, Human-Centered AI

Research Experience

Center for Digital Health and AI, Johns Hopkins University — Baltimore, MD
Research Assistant — Dec. 2024 - Present

Project 1: Fine-tuned **RoBERTa** using Hugging Face on **10M+ patient clinical records** of structured blood glucose, behavior, and diet data.

- Explored RoBERTa source code with pretraining strategies (masked token prediction, next-token generation) and tested multiple output layer (single-/multi-class classification, regression) to suit the dataset.
- Applied **post-training calibration** and evaluation metrics to improve the model's generalization across different data collection settings and input noise conditions.
- Built Hugging Face pipelines for multimodal clinical data, enabling robust handling of missingness and integration of structured + unstructured inputs.

Project 2: Engineered temporal features and patient-specific embeddings for **time-series forecasting**, achieving a 14% improvement in blood glucose and behavior prediction accuracy.

- Benchmarked a wide range of models from statistical (ARIMA) to deep learning (Transformer, N-HiTS, PatchTST) and LLM-based approaches (TimeGPT, Gemini, Claude), using the nixtla package.
- Used Optuna for large-scale experiments and hyperparameter tuning; collected zero-shot results from LLM APIs on validation, test, and out-of-distribution datasets.
- Replicated the study *Context is Key: A Benchmark for Forecasting with Essential Textual Information*, reproducing scripts, and adapting group datasets to enable direct comparison with in-house approaches.
- Developed a fairness-aware evaluation framework with subgroup and time-sensitive metrics to assess equity across patient cohorts.
- Conducted multi-perspective data analysis and visualization to identify disparities and support equitable model evaluation.

HEPIUS Innovation Lab, Johns Hopkins University — Baltimore, MD
Research Assistant — Aug. 2024 - Feb. 2025

Project 1: Implemented and optimized a **physics-informed** deep learning framework (DeepONet) for multi-source focused 2D ultrasound pressure field prediction in heterogeneous spinal cord environments, achieving a 2% prediction error rate.

- Implemented DeepONet in PyTorch and TensorFlow with two branch networks (encoding spinal cord images and device locations) and one trunk network (encoding prediction points), ensuring simulation granularity matched desired resolution.
- Incorporated **physics constraints** in network loss function to enforce realistic pressure propagation patterns.
- Designed the matrix multiplication scheme linking branchnet outputs to prediction layers, ensuring correct latent space interactions and stable gradient backpropagation to capture relationships between device location, prediction point, and spinal heterogeneity.
- Developed CNN and Transformer baselines for pressure map simulation, identifying limitations in their ability to integrate device location information.
- Improved prediction accuracy and robustness by replacing the CNN image branch with a customized Transformer.
- Designed input/output layers and hidden dimensions of networks compatible with the imaging data, and minimized training cost by integrating pretrained hidden layers with trained input and output layers.

Project 2: Built a data-centric ETL and annotation pipeline (DALI + SAM2) packaged in Docker to enable stable, scalable, and reproducible pattern extraction across heterogeneous experimental setups.

- Designed and implemented **GPU-accelerated** data preprocessing using NVIDIA DALI.
- Integrated SAM2 for semi-automated region-of-interest annotation and segment-wise image analysis, improving annotation efficiency and consistency.
- Incorporated existing pretrained segmentation algorithms into the transformation pipeline to streamline data ingestion and ensure robust downstream model training.

Project 3: Designed and automated inference scripts for benchmarking and fine-tuning CNN, Transformer, RNN, and FNN models within DeepONet, ensuring optimal compatibility, improved robustness, and validated cross-domain adaptation.

- Implemented NNI for online hyperparameter tuning and monitoring, enabling efficient parameter search and performance tracking.
- Developed Python scripts for batch inference, experiment logging, and checkpoint management, supporting error recovery and extended training.
- Compared porcine (train/val/test) and human (OOD test) performance, analyzing how variations in spinal cord morphology and thickness influence prediction accuracy.
- Enhanced location encoding by applying a Fourier transformation before Branchnet-2 and integrating a Transformer.

The Cottrell Lab, University of California, San Diego La Jolla, CA
Research Assistant Dec. 2022 - May 2024

Project 1: Investigated biologically inspired divisive normalization as a mechanism to enhance information extraction in early visual processing; developed a trainable HC normalization layer that improved model interpretability and increased AlexNet's ImageNet classification accuracy by 4%.

- Implemented a normalization layer operating across feature channels, based on the cognitively inspired divisive normalization formula.
- Designed flexible neighborhood structures to study the effects of neighborhood size and parameter sharing on training outcomes; ensured normalization within each neighborhood shared hyperparameters and trainable parameters while preventing interference across neighborhoods.
- Modularized the HC normalization layer, extending it from a fixed neighborhood in convolutional layers to a post-convolutional processing layer, enabling seamless integration with 2D, 3D, or higher-dimensional outputs while applying neighborhood separation on the chosen channel dimension.

Project 2: Deployed **multi-GPU** deep learning on Nautilus (via Kubernetes) and built collaborative monitoring tools to enhance interpretability of training results and ensure reproducibility.

- Implemented GPU dataloaders and parallelized training with PyTorch DDP, reducing train time by 50%.
- Utilized Kubernetes (YAML configuration and Kubeflow pipelines) to request GPU resources, manage pods, and schedule jobs.
- Combined with Docker containers and TensorBoard to deploy multiple experiments simultaneously, automating training monitoring and result collection.
- Developed interpretability tools for CNNs, including gradient attribution, activation analysis, and visualization of deep-layer responses.

Work Experience

AiCare Corporation San Jose, California
Data Scientist Intern May 2023 - Aug. 2023
Health Condition assessment for patients recovering from severe diseases using Python, SQL

- Led the development and deployment of a real-time data-visualization and risk-assessment web application, enabling doctors and nurses to monitor patient compliance.
- Embedded time series analysis over months of patient's health data from apple watch and wearable sensors; Implemented Logistic Regression and Random Forest to predict patients' long-term recovery trends with 87% accuracy and 90% specificity
- Explored the characteristics of biometric data of patient symptom cohorts using k-means clustering analysis.
- Developed robust data cleaning and interpolation techniques to handle 30% average missingness

Publications

- Kumar, A., **Zhi, X.**, Ahmad, Z., Yin, M., & Manbachi, A. *Convolutional Deep Operator Networks for Learning Nonlinear Focused Ultrasound Wave Propagation in Heterogeneous Spinal Cord Anatomy.* Accepted for oral presentation at **AAAI 2025** (AI for Accelerating Science and Engineering Workshop).
- Kumar, A., **Zhi, X.**, Davidar, D., Kerensky, M., Theodore, N.,& Manbachi, A. *Accelerating Focused Ultrasound Modeling in Heterogeneous Spinal Cord Anatomy with Vision Transformer Operator Network.* Submitted to IEEE Transactions on Biomedical Engineering
- Luo, J., Han, R., Welivita, A., Di, Z., Wu, J., **Zhi, X.**, Agarwal, R., & Gao, G. *Mapping Patient-Perceived Physician Traits from Nationwide Online Reviews with LLMs.* Submitted to npj Digital Medicine.
- Luo, J., **Zhi, X.**, Han, R., Iyer, A., Agarwal, R., & Gao, G. *FairGlucose: A Benchmark for Fair and Robust Forecasting of Continuous Glucose Monitoring Time-Series Data.* Submitted to **AAAI 2026**.
- Fang, F., **Zhi, X.**, Xu, R.F., & Fee, J. *Method and Apparatus Providing an Ongoing and Real-Time Indicator for Survival and Major Medical Events.* **Patent** (2024).
- **Zhi, X.**, Long, K., Shah, Y., & Cottrell, G. *Modeling Divisive Normalization as Learned Local Competition in Visual Cortex.* Manuscript in preparation.

Teaching Experience

Teaching Assistant at JHU and UCSD 2021 - 2024

- Courses: Machine Learning (Deep Learning), Johns Hopkins University (2024); Data Science, Halıcıoğlu Data Science Institute (HDSI), UC San Diego (2021–2022).
- Assisted with debugging, problem-solving, and student communication for coursework and assignments.
- Managed website maintenance, utilized GitHub for version control on assignments, and designed and graded assignments.

Technical Skills

- **Programming:** Python (PySpark, PyTorch, pandas), Hugging Face, MATLAB, SQL, R, LaTeX, scikit-learn, Tableau, HTML
- **Machine Learning:** Convolutional Neural Networks, Fairness in AI, Model Regularization (LASSO, Ridge, Elastic Net), Time Series Analysis, Classification (Logistic Regression, Random Forest, KNN, AdaBoost)