Tony Ding

xiayiding.tony@outlook.com · (213)245-5570 · LinkedIn · Personal Website

EDUCATION

Harvard University, Boston, MA *Aug 2022 – May 2024*

Master of Science Degree in Health Data Science

- GPA: 3.96/4.0; Relevant coursework: *Deep Learning*, *Statistical Learning*, *Computing for Big Data*
- Recipient of: Lyman & Grew Memorial Scholarship AND Harvard Central Grant Scholarship
- Cross-registrations at **MIT**: *Machine Learning in Healthcare*, *Advanced NLP*, *Advances in Computer Vision*; GPA: 5.0/5.0

University of Southern California, Los Angeles, CA *Aug 2018 – May 2022*

Bachelor's Degrees in Data Science and in Neuroscience

- Data Science major GPA: 4.0/4.0; Cumulative GPA: 3.9/4.0
- Presidential Scholar (half-tuition awarded throughout 4 years); Renaissance Scholar Distinction

RELEVANT SKILLS

- 3+ years of experience in Data Science and Machine Learning Engineering
- Expert in Python, SQL, and R; Expert in database modeling and ETL
- Extensive experience in deep learning algorithms & libraries (LangChain, PyTorch, Tensorflow, NLTK, spaCy, OpenCV)
- Extensive experience in Natural Language Processing (NLP), including Large Language Model (LLM) deployment and fine-tuning, and Computer Vision
- Proficient in using Google Cloud Platform (BigQuery & Vertex AI), Tableau, Power BI, MySQL, MongoDB, and Excel

RELEVANT PROJECTS

End-to-End Cross-Lingual Summarization (CLS) with Pre-training

Main Affiliation: **MIT Department of EECS** *Sep 2023 – Dec 2023*

- Developed an end-to-end framework for CLS tasks through leveraging mBART, eliminating the traditional pipeline approach that separates CLS tasks into machine translation and monolingual summary sub-tasks.
- Achieved a significantly better ROUGE-1 score of 3.82 on an external dataset by fine-tuning mBART's linear and cross-attention layers on CLS data using techniques such as layer-wise learning rate decay and gradient clipping. Outperformed baseline and alternative models by 5357% and 3720%, respectively.

Pleural effusion diagnosis: multimodal approaches using deep neural networks and transformer-based architectures

Main Affiliation: **MIT Department of EECS** *Dec 2022 – May 2023*

- Combined patients' clinical reports with X-ray images to explore the optimal fusion strategies for a multimodal approach in diagnosing pleural effusion.
- Researched and experimented with early fusion, joint fusion, and late fusion strategies. Fine-tuned VGG16 and DistilBERT using HuggingFace's PEFT and LoRA configurations to enhance parameter efficiency and achieved an AUC of 0.9887. Integrated the late fusion multimodal model with an elastic net regularized logistic regression classifier to further enhance classification performance.

Imaging-based prediction of transcriptional subtypes in Alzheimer's Disease

Main Affiliation: **Harvard Department of Biostatistics & MIT CSAIL** *Jan 2023 – May 2023*

- Applied and fine-tuned Vision Transformers (ViTs) on Alzheimer's MRI imaging dataset using PyTorch.
- Used ViT in 3D brain MRI image analysis by treating each 3D patch as a vector to feed into the transformer. Achieved a 98.57% test accuracy with the best ViT. Extracted feature representations of brain MRI images and combined them with genetic deep learning models to further improve model performance using a self-supervised contrastive learning pipeline.

RELEVANT WORK EXPERIENCES

Data Scientist (Unit Cost & Operations Analytics), **CVS Health**, Wellesley, MA *Apr 2024 – Present*

- Developed and deployed an automated case categorization and resolution generation pipeline using GPT-4, LangChain, and Airflow DAGs. Integrated RAG by leveraging ChromaDB for vector storage and LlamaIndex for structured indexing and retrieval. Further optimized performance through prompt engineering and iterative output refinement techniques. Led the entire project from initial model ideation and development to production pipeline deployment.
- Engineered ETL pipelines using Hadoop and Spark and queried databases using GCP and BigQuery for efficient data retrieval and seamless integration with the automated Generative AI pipelines. Benchmarked various LLMs' performances and conducted token cost and model latency analyses to optimize resource allocation and pipeline efficiency.
- Developed an Streamlit web app to visualize the internal processes of the GenAI prediction pipeline. Designed model demos in Figma tailored for senior business stakeholders. Conducted in-depth model cost and efficiency impact analysis for this GenAI pipeline, revealing an annual savings of $1.7M and a 532% boost in operational efficiency.

AI & Data Scientist, **Mayo Clinic**, Boston, MA *Sep 2023 – Dec 2023*

- Deployed an LLM (PaLM) on Google Cloud Platform (Vertex AI) using Python for automating the data abstraction pipeline for the patient management and pathology data of the oropharyngeal cancer clinical registry.
- Utilized Google BigQuery for data integration and processing. Leveraged the Python Regex library and a semi-supervised topic modeling algorithm, Guided Latent Dirichlet Allocation (Guided LDA), to help LLM locate the correct clinical text to attend to.
- Conducted extensive clinical prompt engineering and multiple LLM output refinement strategies to reduce LLM hallucination and to improve overall LLM performance and consistency. Ultimately increased the accuracy of LLM's output by 47%.

R&D Data Science Intern, **Johnson & Johnson**, Cambridge, MA *June 2023 – August 2023*

- Conducted comprehensive network analysis of multiple myeloma patient referral networks using claims data. Implemented the Leiden algorithm to detect patient referral communities and fine-tuned weighted PageRank algorithms to quantitatively rank the Healthcare organizations (HCOs).
- Identified the top influential HCOs for multiple myeloma patient referrals, supporting company's clinical trial site selection process and multiple myeloma patient outreach programs.
- Designed and built 5 large-scale interactive network visualizations and dashboards using the Dash framework in Python.

Data Scientist Intern, **AstraZeneca**, Shanghai, CN *May 2021 – August 2021*

- Built a supervised machine learning program using Python to predict users' behavior, such as users' follow or unfollow activity, for AZ_MedInfo (a medical information exchange platform with over 1M users). Increased the AUC score by 26.3% and decreased the error rate by 28.5% by applying and tuning a Random Forest classifier. Successfully identified 12 out of 161 most significant and meaningful variables that impact users' decisions.
- Designed and implemented a weighted association rule mining program, for Prof. Binghe Xu, MD, by extracting keywords like "ctDNA" and "Breast Cancer" from publications' titles and abstract sections. Identified and ranked his associations with fellow colleagues and determined his individual academic rankings among peers in the research field.

Data Science & Visualization Intern, **Takeda**, Los Angeles, CA *May 2020 – August 2020*

- Used Python to pre-process the clinical patient data and implemented unsupervised machine learning models on raw baseline data for PANDA, a Takeda's oncology program for Ponatinib. Identified 4 significant patient subpopulations for Ponatinib and the key risk drivers ($p<0.001$) of MACE (Major Adverse Cardiac Event) occurrences among Chronic Myelogenous Leukemia patients in the clinical trials.
- Employed MS SQL Server Management Studio and Excel to quantitatively analyze and gain insights toward Takeda R&D partnerships around the globe. Created a novel R&D partnerships visualization paradigm for Takeda by designing 8 complex time-based interactive network visualizations on Kumu (a visualization platform) to better inform strategic decision-making in partnership management.

RESEARCH EXPERIENCE

Statistics Research Assistant, **USC Health, Emotion, & Addiction Lab** *Dec 2020 – Jul 2022*

- Wrote thousands of rows of SPSS syntax and conducted various statistical tests and analysis in R for ADVANCE (*Assessing Developmental Patterns of Vaping, Alcohol, Nicotine, and Cannabis Use and Emotional Well-being*) School Reports project to determine the significance of associations among variables and among different demographics. Calculated RCADS (Revised Children's Anxiety and Depression Scales) scores from hundreds of variables for each participant. Applied syntax to and created analytical reports for all 6 high schools in the Greater Los Angeles area.
- Completed thorough literature review on RCADS (Revised Children's Anxiety and Depression Scales), CAST (6-item Cannabis Abuse Screening Test), DAST (10-item Drug Abuse Screening Test), RAPI (23-item Rutgers Alcohol Problems Index), and the 15-item Mood Disorder Questionnaire.

LEADERSHIP EXPERIENCE

Data Team Lead, **LinkedIn Campus Editing Team, USC,** Los Angeles, CA *Sep 2020 – May 2021*

- Led a team of 7 undergraduate and graduate students for collecting and processing career outcome data.
- Pre-processed our data using Python and built a fine-tuned XGBoost model in RapidMiner to analyze post-graduation career plans and job offers received by seniors and graduate students at USC across all academic fields.
- Built visualizations and interactive dashboards on Tableau and showcased our findings at the annual Post-graduation Outcome event hosted by USC Career Center.

ADDITIONAL AWARDS

- Bright Futures Award in the 2023 NNLM Data Visualization Challenge - Complex Visualization Category
- USC Academic Achievement Award for highly motivated students with excellent academic records
- Alpha Prize (2nd Place in the World) in AoCMM Mathematical Modeling Contest