Hudanyun Sheng

hudanyun.sheng@outlook.com | (+86) 13683132915 | https://danniesheng.github.io/

EDUCATION

University of Florida - *M.S. in Electrical and Computer Engineering (GPA: 3.86/4), December 2019*

Master thesis: Switchgrass Genotype Classification using Hyperspectral Imagery

University of Florida - *M.S. in Industrial and Systems Engineering (GPA: 3.87/4), December 2017*

Tongji University - *B.S. in Physics (GPA: 4.45/5), June 2015*

Bachelor thesis: The Correction of the Intensity Unevenness of X-Ray KB Imaging

TECHNICAL SKILLS

- Programming & Development: Python (NumPy, Pandas, SciPy, Seaborn, Plotly, Streamlit), MATLAB, Flask, HTML, CSS, JavaScript, Ajax
- Machine Learning: PyTorch, Scikit-learn, Weights & Bias, Tensorflow, Keras, Deep Learning, Computer Vision, OpenCV, LangChain, Retrieval-Augmented Generation (RAG), NLP, HuggingFace
- Cloud & DevOps: AWS, GCP, Azure, Databricks, BigQuery, Docker, Kubernetes, Git
- Data Analysis & Visualization: PySpark, SQL, BigQuery, Data Pipeline Development, EHR Data Analysis, MS Office

PROFESSIONAL EXPERIENCE

Johnson & Johnson | **Data Scientist (Contractor),** Beijing, China — April 2024-present

Center of Excellence

- Processed and integrated multiple data sources, including basic information of HCO and HCP, drug sales data, and promotional data etc.; managed project data flow using the Kedro framework, ensuring data accuracy and consistency
- Undertook complex feature engineering tasks to provide high-quality input data for model training
- Developed and optimized early adopter classification and prescriber classification models for HCO and HCP
- Collaborated with other data scientists to address data challenges and optimize model performance
- Utilized large language model APIs to perform Retrieval-Augmented Generation (RAG) tasks for a project comparing global FAQs with Japanese drug labels.
- Contributed to the migration of a Kedro-managed project to Databricks and participated in transforming the Databricks Feature Store to improve feature management and scalability.

Zenni Optical | **Data Scientist,** Beijing, China — June 2023-Jan 2024

AI/ML Department

- Led the end-to-end research, development, and deployment of the Rx Extraction API service, overseeing key stages from conceptualization to deployment and continual improvement. Implemented cutting-edge Optical Character Recognition (OCR) techniques to automate eyeglass prescription interpretation and parsing
- Drove the successful launch of multiple iterations of the Rx Extraction API service on the Google Cloud Platform (GCP), showcasing adaptability by delivering a specialized version tailored for the Japanese market. This involved incorporating language nuances and prescription-style variations specific to the Japanese market
- Pioneered the integration of deep learning models to enhance the document quality assessment module within the prescription extraction framework. Specifically, introduced prescription blurriness detection capabilities, contributing to improved accuracy and efficiency in processing prescription data
- Demonstrated analytical prowess by conducting in-depth analysis of the prescription extraction module's usage data. Utilized advanced querying techniques on the BigQuery database to extract valuable insights, aiding in strategic decision-making and further refinement of the service
- Maintained update-to-date comprehensive technical documentation to produce detailed changelogs for software releases
- Developed and implemented Streamlit visualization applications to enhance communication with stakeholders, providing intuitive and interactive data presentations for effective decision-making

University of Texas Southwestern Medical Center | **Data Scientist,** Dallas TX USA — Sep 2021-May 2023

Quantitative Biomedical Research Center

- Spearheaded the implementation of Mask R-CNN in PyTorch, enabling simultaneous nuclei segmentation and classification from H&E-stained histology images. Achieved remarkable results with an 82.5% detection rate and an 82.0% classification accuracy across six classes
- Pioneered the design of custom loss functions tailored for Mask R-CNN, optimizing training procedures particularly for deficiently labeled data, thus ensuring robust model performance
- Successfully established a comprehensive whole slide histology image (WSI) analysis pipeline
- Engineered a Python-based CyTOF image analysis pipeline that drastically improved processing efficiency by 10x through the implementation of parallel processing. Additionally, developed an intuitive and user-friendly graphical user interface (GUI) using Flask, HTML&CSS, and Ajax, enhancing accessibility. Created Docker images to facilitate seamless transitions between development, testing, and production environments
- Innovated the preprocessing protocol for raw doctors' notes, laying the foundation for training an NLP model dedicated to CLASI score prediction—a critical step in advancing cancer research
- Implemented a cutting-edge model for the automated de-identification of Electronic Health Record (EHR) data, contributing to the protection of patient privacy while streamlining data analysis processes

Donald Danforth Plant Science Center | **Data Science Researcher,** St. Louis MO USA Feb 2020-Sep 2021

Data Science Facility

- Enhanced and innovated the tool set for image analysis, object segmentation, classification, and feature detection within the PlantCV framework (Plant Phenotyping using Computer Vision, open source). Conducted unit testing and ensured version control through GitHub collaboration
- Established a robust processing protocol for the automated analysis of RGB, thermal, and hyperspectral imagery, encompassing preprocessing, analysis, post-processing, statistical analysis, and visualization, streamlining the workflow for efficient and accurate data interpretation.
- Pioneered instance-wise leaf segmentation, developing cutting-edge algorithms to track leaf growth over time, contributing to advancing our understanding of the intricate life cycles of plants
- Collaborated with the research team to present data effectively, employing advanced visualization tools to communicate complex statistical outcomes

University of Florida Academic Health Center | **Data Science Intern** May 2019-Aug 2019

Precision and Intelligent Systems in Medicine Partnership Lab, Gainesville FL USA

- Led the formulation of cohorts and conducted meticulous data preprocessing for the analysis of time-series data derived from hospital records
- Extracted key time-series features, focusing on patients' vital signs during their initial 24-hour hospital admission
- Employed advanced algorithms to effectively address irregularities within the time-series data, ensuring data integrity and reliability
- Conducted automated statistical analyses, leveraging cutting-edge methodologies to derive meaningful insights
- Generated comprehensive comparison tables to facilitate a detailed evaluation of time-series clustering results, aiding in the identification of patterns and trends.

PUBLICATION

- Sheng, H., Wang S., et al. "MTIA: An open-source python package for systematic multiplexed tissue image analysis" (in preparation)
- Sheng, H., Gutierrez, J., Schuhl, H., Murphy, K. M., Acosta-Gamboa, L., Gehan, M., & Fahlgren, N. (2023). Increasing the Throughput of Annotation Tasks Across Scales of Plant Phenotyping Experiments. Authorea Preprints.
- Rong, R., Sheng, H., Jin, K.W., Wu, F., Luo, D., Wen, Z., Tang, C., Yang, D.M., Jia, L., Amgad, M. and Cooper, L.A., 2023. A deep learning approach for histology-based nucleus segmentation and tumor microenvironment characterization. Modern Pathology, 36(8), p.100196.
- Panda, K., Mohanasundaram, B., Gutierrez, J., McLain, L., Castillo, S. E., Sheng, H., ... & Slotkin, R. K. (2023). The plant response to high CO2 levels is heritable and orchestrated by DNA methylation. New Phytologist, 238(6), 2427-2439.
- Yu, G., Zare, A., Sheng, H., Matamala, R., Reyes-Cabrera, J., Fritschi, F.B. and Juenger, T.E., 2020. Root identification in minirhizotron imagery with multiple instance learning. Machine Vision and Applications, 31, pp.1-13.

ACADEMIC RESEARCH EXPERIENCE

Machine Learning and Sensing Lab | **Graduate Research Assistant**, Gainesville FL USA Mar 2017-Dec 2019

- Enhanced the efficiency of root detection in mini-rhizotron images by spearheading the development of machine learning algorithms, effectively reducing labor and time requirements. Established a streamlined processing protocol for the automated analysis of hyperspectral and thermal imagery of plants, contributing to the advancement of data processing techniques
- Pioneered the creation of algorithms for automated plant detection from hyperspectral images, utilizing hyperspectral endmember detection and un-mixing methods. Innovatively proposed and implemented a classification-friendly dimensionality reduction algorithm, facilitating the accurate classification of genotypes within identical plant species
- Demonstrated versatility by contributing to the academic environment through the arrangement and preparation of lecture materials, assignments, exams, and solutions for the course on Data Analysis and Data Mining.

Institute of Precision Optical Engineering | **Undergraduate Research Assistant**, Shanghai, China June 2014-June 2015

- Conducted simulations of X-Ray KB (Kirkpatrick-Baez Microscope) imaging through programming, addressing and rectifying irregularities in the imaging process
- Developed expertise in correcting the unevenness of X-Ray KB imaging for improved accuracy and precision.
- Designed and implemented a user-friendly Graphical User Interface using MATLAB, facilitating a seamless and intuitive experience for navigating the simulated imaging process.

SOFT SKILLS

- **Analytical Thinker and Fast Learner**: Possess a robust ability for independent thinking and working, coupled with a keen aptitude for rapid assimilation of new knowledge and mastery of emerging technologies
- **Effective Communication and Team Collaboration**: Demonstrate exceptional communication and teamwork skills, with the ability to articulate technical concepts to both technical and non-technical audiences. Proven success in fostering collaborative environments that enhance project outcomes.
- **Multilingual Proficiency**: Capable of effectively communicating in both written and spoken English and Chinese, facilitating seamless interactions in diverse and international settings

CERTIFICATES, HONORS, REWARDS AND MISCELLANEOUS

- 1st Place of the “Swarm Behavior on the Grid” track in the Siemens "Tech for Sustainability Campaign 2023” (2023)
- Google Data Analytics Certificate– a rigorous, hands-on program that covers the entire scope of the data analysis process
- Co-Chair of the Committee for Scientific Training and Mentoring at Donald Danforth Plant Science Center