Hudanyun Sheng

Email: hudanyun.sheng@outlook.com
Phone: (+86) 13683132915
Website: https://danniesheng.github.io/
Location: Beijing, China

Summary

Applied AI Engineer and Data Scientist with experience designing and deploying product-oriented AI systems across healthcare, life sciences, and enterprise applications. Specialized in retrieval-augmented generation (RAG), document understanding, information retrieval, and machine learning pipelines, with a strong focus on building reliable, explainable, and maintainable AI solutions. Experienced in translating business and research problems into end-to-end AI products by integrating LLMs, classical machine learning, software engineering, and cloud-native data platforms.

Technical Skills

- **Machine Learning & GenAI:**
  classical ML (classification, ranking, feature engineering), Retrieval-Augmented Generation (RAG), prompt engineering, LLM evaluation, embedding-based retrieval, cross-encoder reranking, uncertainty handling
- **Data & System Engineering:**
  end-to-end ML / GenAI pipeline design with quality gating and fallback strategies, noisy document understanding (PDF parsing, OCR, table extraction), human-in-the-loop systems for review, validation, and error correction
- **Platform & Cloud:**
  Databricks-based ML workflows and feature pipelines, containerized deployment with Docker, Azure-based ML / LLM platforms in regulated environments, cloud-native data workflows
- **Tools & Stack:**
  Python, Pandas, NumPy, scikit-learn, FAISS, Hugging Face (E5 embeddings, rerankers), OpenAI / Azure OpenAI APIs. Kedro, Streamlit

Professional Experience

Johnson & Johnson

Data Scientist (Contractor) | Center of Excellence Beijing, China | April 2024 – Present

Commercial ML System for Drug Launch & HCP Targeting

- Developed machine learning models to identify high-potential healthcare professionals likely to adopt newly launched products.
- Designed feature engineering and scoring pipelines to support large-scale prioritization under business constraints.

- Implemented modular Kedro-based data pipelines with validation and reproducibility controls.
- Collaborated with commercial stakeholders to translate predictive outputs into targeting strategies.

GenAI / RAG Systems for Medical Content Governance

- Developed an LLM-assisted workflow to identify inconsistencies between localized Japanese medical FAQs and approved regulatory labels.
- Built retrieval and comparison pipelines combining document chunking, semantic search, and GPT-based reasoning to support scalable compliance review.
- Generated evidence-backed outputs to support human validation and reduce manual review burden.

Doctor Lookup & Q&A System (Japanese Medical RAG)

- Designed and deployed a Japanese medical RAG system integrating FAISS, E5 embeddings, history-aware retrieval chains, cross-encoder reranking, and uncertainty gating.
- Built multilingual name normalization and fuzzy-matching pipelines.
- Developed an interactive UI for retrieval tuning and evidence tracing.
- Modularized retriever and LLM components with Dockerized deployment.

Document Understanding & PDF Table Extraction Pipeline

- Designed resilient pipeline for extracting hospital schedule tables from heterogeneous PDFs (digital, scanned, malformed).
- Integrated Camelot, OCR-based detection, and LLM-based image reasoning with quality gating and fallback logic.
- Implemented failure-aware handling for encoding errors, broken PDFs, OCR degradation, and oversized images to prevent pipeline interruption.
- Developed reviewer-facing UI for batch validation and annotation.

Feature Store Migration and Optimization

- Participated in migration from legacy ML pipelines to Databricks-based feature management workflows.
- Supported feature standardization and reusable feature architecture to improve maintainability and model development efficiency.

Zenni Optical

Data Scientist, AI/ML Department
Beijing, China
Jun 2023 – Jan 2024

Prescription (Rx) Recognition System

- Built FastAPI-based inference service to extract structured optical parameters (sphere, cylinder, axis) from user-uploaded images.
- Integrated OCR outputs with rule-based and heuristic post-processing.

- Adapted parsing logic for Japanese prescription formats to support multilingual layout variation.
- Integrated CloudSQL and BigQuery for monitoring and multi-region usage analytics.

Human-in-the-Loop Evaluation & Quality Diagnostics

- Built internal Streamlit tool to visualize OCR and extraction outputs for rapid debugging.
- Analyzed API logs and user behavior to diagnose extraction failures related to image quality and capture issues.

University of Texas Southwestern Medical Center

Data Scientist, Quantitative Biomedical Research Center
Dallas, TX, USA
Sep 2021 – May 2023

CyTOF Image Analysis Toolkit

- Designed end-to-end pipeline integrating multiplexed imaging with single-cell expression data.
- Built Flask-based service with Dockerized deployment and parallelized computation.
- Achieved approximately 10× runtime speed-up for large-scale experiments.

Digital Pathology: Nuclei Segmentation & Classification

- Developed Mask R-CNN–based pipeline for nuclei segmentation and 6-class classification on H&E slides.
- Customized loss functions for partially labeled data, recovering approximately 20% of training samples.
- Achieved 82.5% detection and 82.0% classification performance under weak supervision.

Donald Danforth Plant Science Center

Data Science Researcher, Data Science Facility
St. Louis, MO, USA
Feb 2020 – Sep 2021

Multimodal Imaging & Analysis Systems for Plant Phenotyping (open-source software)

- Built automated pipelines for large-scale multimodal plant imaging (RGB, thermal, hyperspectral).
- Developed instance-level segmentation and temporal tracking workflows for growth analysis.
- Contributed to the open-source PlantCV project (modules, tests, documentation, community support).
- Supported interdisciplinary teams with reproducible analysis and visualization.

University of Florida Academic Health Center

Data Science Intern
Gainesville, FL, USA
May 2019 – Aug 2019

- Processed and cleaned patient vital-sign time-series data.
- Extracted predictive features and evaluated clustering methods.
- Reproduced interpolation-based models for irregular clinical data.

Selected AI Projects

AI News Agent

LLM-based News Summarization and QA Assistant

Open-source project for daily AI news aggregation, summarization, and question answering.

- Designed and implemented an end-to-end LLM-powered news agent for daily AI news aggregation, summarization, and question answering.
- Built a lightweight RAG pipeline integrating web scraping, local embedding models, hybrid retrieval, and OpenAI APIs.
- Implemented scheduled crawling, automatic topic tagging, and corpus-based QA over historical news data.
- Developed interactive Streamlit front-end and deployed on Hugging Face Spaces.
- Explored Docker-based deployment and embedding model switching.

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

Publications

- 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.
- Rong, R., Sheng, H., Jin, K.W., Wu, F., Luo, D., Wen, Z., Tang, C., Yang, D.M., Jia, L., Amgad, M., & Cooper, L.A. (2023). A deep learning approach for histology-based nucleus segmentation and tumor microenvironment characterization. Modern Pathology.

Awards & Leadership

- 1st Place, Siemens Tech for Sustainability Campaign (2023)
- Google Data Analytics Certificate
- Co-Chair, Committee for Scientific Training and Mentoring, Donald Danforth Plant Science Center