

Hairong Wang

+1 412-482-0131 | hairongcmu@gmail.com | hwcmu.github.io | hairong-wang | hwcmu

Profile

Ph.D.-trained Data Scientist with a Master's in Analytics, bridging statistical rigor with scalable engineering (Spark, AWS, Docker). Expert in developing end-to-end ML solutions (Survival Analysis, NLP) for Healthcare and Environmental domains, translating complex data into explainable AI products that drive measurable business impact.

Education

Ph.D. Environmental Engineering *Carnegie Mellon University*

Pittsburgh, PA Sep 2025

M.S. Data Analytics *Georgia Institute of Technology*

Pittsburgh, PA May 2026

Skills

- Machine Learning & AI:** NLP (BERT/GPT/Hugging Face), Computer Vision (Mask R-CNN), Scikit-learn, PyTorch, XGBoost, Survival Analysis.
- Big Data & Engineering:** Apache Spark, SQL (PostgreSQL), AWS, GCP, Docker, Git/CI-CD.
- Visualization & Deployment:** Tableau (TabPy integration), Power BI, D3.js, FastAPI, Flask, Streamlit.
- Domain & Soft Skills:** Healthcare Analytics (EHR/MIMIC-III), Environmental Data (GIS/ArcGIS), Research Communication, Cross-functional Collaboration.

Professional Experience

Data Scientist Intern *Peachy Day*

Pittsburgh, PA Oct 2025 - Present

- Product Analytics & KPI Framework:** Established the company's core product metrics system by defining and tracking key performance indicators (e.g., Daily Active Users, Retention) using **Supabase**; provided the leadership team with actionable dashboards to guide product roadmap decisions.
- Physician-Guided Forecasting:** Collaborated with **neurologists** to incorporate clinical expertise into feature selection and output smoothing; built a Random Forest pipeline that achieved an **18% accuracy boost** while ensuring medical interpretability for patients.
- Data-Driven Retention Strategy:** Designed and delivered the *"Migraine Wrapped"* data product for **1,000+ users**; visualized longitudinal behavioral trends to personalize user experience, driving a **12% increase** in app engagement.

Data Scientist Intern *University of Pittsburgh*

Pittsburgh, PA Sep 2024 - May 2025

- NLP Pipeline Development:** Built a scalable NLP pipeline to process **13K+ unstructured clinical notes**; implemented **BERT** and **GPT-2** to generate contextual embeddings, enriching structured EHR data for downstream modeling.
- Clinical Predictive Modeling:** Engineered high-dimensional features using **TF-IDF** and interaction terms; trained and tuned ensemble models (**XGBoost**, Random Forest), achieving an **AUC of 0.87**.
- Model Interpretability:** Conducted rigorous validation using **5-fold cross-validation** and applied feature importance analysis (SHAP values) to identify key clinical predictors, ensuring model transparency and clinical validity.

PhD Research

- Automated Inspection System (CV):** Engineered a **Mask R-CNN** pipeline for fiber detection and released a benchmark dataset, reducing manual identification time by **60%+**.
- Geospatial Active Learning:** Developed a **Gaussian Process** framework and deployed a full-stack web app (python-uv) for real-time uncertainty visualization, improving sampling accuracy by **18%**.

Course Projects

Scalable Taxi Trip Analytics on AWS & GCP with PySpark, *CSE 6242: Data and Visual Analytics*

Fall 2025

- Big Data Engineering:** Architected distributed **PySpark** pipelines on **AWS Athena/S3** and **GCP Dataproc** to process **1 Billion+** NYC trip records, optimizing partitioning strategies to reduce execution time by **35%**.
- Deployment & Ops:** Containerized the application using **Docker** for reproducible deployment on Databricks; integrated **CI/CD workflows** (Git) to automate schema validation and testing.

Deep Learning for Healthcare Time-Series, *ISYE 6740 - Computational Data Analytics*

Summer 2025

- Model Architecture:** Designed an end-to-end **PyTorch** pipeline for MIMIC-III sequential data; implemented and compared **GRU**, **LSTM**, and **CNN** architectures with dynamic batching for irregular time-series.
- Performance:** Achieved **AUC = 0.783** on mortality prediction benchmarks; enforced reproducibility via **Pytest** and strictly versioned Conda environments.

Selected Publication

- Wang, H., & Zhang, X. (2025). *Machine learning-based prediction of electrocardiogram (EKG) use in emergency care.* **Journal of Personalized Medicine**.
- Wang, H., Ling, H., & Zhang, X. (2025). *Integrating structured clinical data and GPT-2 embeddings of patient narratives to predict IV fluid utilization.* **PEERJ Computer Science**.
- Wang, H., Piao, W., & Gregory, L. (2025). *AI-assisted screening for asbestos fibers in soil using Mask R-CNN and computer vision on polarized light micrography.* **Journal of Hazardous Materials** (Under Review).