

# Devavrath Sandeep

(412) 430-2092 | dsandeep@andrew.cmu.edu | linkedin.com/in/devavrath-sandeep | devsovague.github.io

## Education

### Carnegie Mellon University

*M.S. in Artificial Intelligence Engineering – Biomedical Engineering | GPA: 3.91/4.00*

Pittsburgh, PA

*Dec 2026*

### Indian Institute of Technology Jodhpur

*B.Tech. in Bioengineering | GPA: 7.97/10 | Top 1% of Department*

Jodhpur, India

*May 2025*

## Technical Skills

**Programming:** Python, C/C++, SQL, JavaScript, R, MATLAB, Bash

**Software Engineering and APIs:** FastAPI, Pydantic, Flask, PostgreSQL, Docker, Kubernetes / GKE, Terraform, CI/CD, PySpark, Kafka, GCP, Git

**Machine Learning:** PyTorch, scikit-learn, XGBoost, LightGBM, HuggingFace Transformers, CNN-Transformer, LSTM/TCN, domain-adversarial (GRL), FiLM, isotonic calibration, SHAP

**LLMs and Agentic AI:** LangGraph, LangChain, RAG, ReAct, Reflexion, LLM-as-judge, tool-calling, MCP, Claude, Gemini, Gemma, Ollama, Whisper, Milvus, BGE embeddings

**Healthcare AI and Clinical Systems:** Clinical decision support, discrete-event simulation of care pathways, medical imaging (CNN), biomedical signal processing (EEG/EMG), external validation on multi-site clinical corpora, calibrated risk scoring, cost modeling over BLS OEWS wage data

## Publications and Patents

- **Patent (pending):** “Upstream AI Governance of an Action Selection Process Through Coherent Propagation,” U.S. Patent Application No. 19/783,601, filed August 21, 2026 (Track One prioritized examination).
- “A Digital Process Twin Simulation of Breast Pathology Workflows: Evaluating the Parametric and Structural Impacts of AI Integration,” **IEEE International Conference on Healthcare Informatics**, 2026.
- “From Desert to Design: Adaptive Nesting of Dusky Crag Martin (*Ptyonoprogne concolor*) and Conservation Implications,” **Ornithology Research**, vol. 34, no. 1, art. 34, 2026.

## Experience

### AI Systems Research Engineer

Chicago, IL

*Stealth AI Startup | Agentic Pipelines, Governance, Interpretability, Cloud and Edge Deployment May 2026 – Jul 2026*

- Engineered self-evaluating agentic pipelines with governance guardrails, multi-step orchestration, tool use, and human-in-the-loop review for production LLM applications.
- Designed governance and evaluation workbenches over mechanistic decomposition methods to make internal model state inspectable through probing classifiers, calibration, tension and coherence signals, and human review.
- Built a mechanistic-interpretability instrument (J-lens): a Hutchinson-estimated Jacobian over a prompt corpus, reading layer-wise token dispositions from Gemma activations; validated against Gemma Scope SAEs under a pre-registered AUROC falsification protocol.
- Owned deployment across edge and cloud, running on-device inference on NVIDIA Jetson and Raspberry Pi and autoscaling GPU instances on Google Cloud (cloud-burst provisioning) for activation capture and model training.

## Healthcare AI and Clinical Decision Support Projects

### Twinbench: Digital Process-Twin Workbench for Clinical Pathology

Carnegie Mellon University

*Agentic LLM to Discrete-Event Simulation | Second Author, IEEE ICHI 2026*

*Jan 2026 – Present*

- Built an LLM pipeline that compiles a clinical process map into a verified, executable SimPy discrete-event simulation; the reflexion check is folded into the spec-generation prompt (the model reconciles the swim-lane diagram against its exported edge list before any code exists), and an LLM-as-judge signs off pass or fail.
- Ran a six-scenario counterfactual sweep over a simulated week across four staffed resource pools: front-loading AI-flagged cases cut 95th-percentile turnaround by roughly 35% (48 to 31 hours), a larger histology batch worsened mean turnaround by roughly 30% and roughly tripled routine-pathologist wait, and the radiologist was the binding constraint across every scenario.
- Wrote the graph-to-SimPy transpiler over a typed graph IR with a validator that closes branch probabilities and checks reachability before code generation; nodes compile to standalone, readable SimPy that re-imports losslessly through an embedded graph parsed back via Python’s AST.
- Added Monte-Carlo replication with 95% confidence intervals across all KPIs and an activity-based cost layer over bundled BLS OEWS wages (445 U.S. areas, 130 occupation profiles), so scenario trade-offs come out in dollars, not just minutes.

**CxRisk: Cervical-Cancer Risk Model, Serving API, and Sensitivity Analysis** Carnegie Mellon University  
*Calibrated Four-Model Ensemble and Explainable Screening Triage / Led a 3-Person Team* Jan 2026 – May 2026

- Caught that a first model's near-perfect AUC ( $\approx 0.997$ ) was read off its own training data and rebuilt around a locked 20% stratified holdout, setting honest performance at ROC-AUC  $\approx 0.77$  / PR-AUC  $\approx 0.18$  on the 858-patient, 6.4%-positive UCI cohort.
- Retuned the flag threshold from 81% of patients to at most 20% by maximizing F2 (a worklist a clinic can actually act on), and set soft-vote ensemble weights by Nelder-Mead search over out-of-fold PR-AUC with per-model isotonic calibration; every weight and threshold decision stayed on out-of-fold data so the holdout scored the model exactly once.
- Built a six-term per-patient confidence score with a plain-language explanation naming each model's vote, and turned a fragility finding into a clinic rule: the system warns when STD burden reads zero while other risk factors are elevated.
- Served the Python ensemble over a framework-agnostic FastAPI layer with Pydantic validation, server-rendered explanation charts, and a self-contained joblib bundle, so a client in any language (the R frontend used by the project, or a Streamlit UI) could score against the same contract.

**Leakage-Free Obesity Risk Classifier with RAG Clinical Guidance** Carnegie Mellon University  
*Two-Stage ML plus Retrieval-Augmented Generation / scikit-learn, Milvus, LangGraph* Mar 2026 – May 2026

- Designed a two-stage classifier on a 2,111-record dataset over seven WHO obesity categories: Stage 1 is a Random Forest over 11 lifestyle features using a coarse four-level BMI bucket instead of exact BMI, eliminating algebraic leakage while recovering F1-macro  $0.9535 \pm 0.009$  under five-fold cross-validation.
- Built a four-component confidence score (centroid similarity, model probability, confusion-weighted runner-up, demographic alignment) to route borderline overweight and obese predictions to a Stage 2 clinical classifier trained on NHANES 2021–2023 body-composition and laboratory biomarkers; every Stage 1 miss stayed within one severity category.
- Serialized results as natural-language patient context and prepended them to a Milvus and LangGraph RAG pipeline with a reflection loop, producing class-specific, citation-grounded nutrition and lifestyle guidance for the clinician.

**TC-DANN: Confounder-Audited Voice-Disease Screening** Carnegie Mellon University  
*Domain-Separated, Site-Adversarial Model on NIH Bridge2AI-Voice v3 / PyTorch, DANN + FiLM* Jan 2026 – May 2026

- Built a domain-separated, site-adversarial voice-disease model (20 disease heads across four clinical domains) reaching 0.76 weighted AUROC on Bridge2AI-Voice v3 and 0.83 across the 15 diseases that clear a confounder-separation gate (acoustic AUROC above a demographics-plus-task baseline).
- Externally validated on the German Saarbrücken corpus (283 patients, 1,145 recordings): six of six diseases passed the confounder audit on held-out data, ADHD 0.72 and Parkinson's 0.69 AUROC.
- Ran the audit that motivated the redesign: demographics alone predict disease at 0.96 AUROC for cognitive impairment and the recording task alone at 0.94 for Parkinson's. Then designed a Brier-calibrated five-term confidence score with SLSQP-optimized weights and demographics zeroed for psychiatric heads.

## Software Systems and Data Engineering Projects

**RAG Research Assistant: Multilingual PDF Q&A on a Self-Checking LangGraph Agent** Carnegie Mellon University

*LangGraph, Gemini, Milvus, GKE / Led Team of 2* Jan 2026 – Mar 2026

- Built a retrieval-augmented assistant whose every answer is reviewed by a second Gemini pass against the retrieved chunks; grounding, an exact output format, and a two-paper citation cap are enforced in a bounded reflect-and-revise loop before the answer is shown.
- Structured the pipeline as a seven-node LangGraph state machine with a conditional router and an iteration cap; translation across four languages via a FastAPI microservice with a Gemini fallback so a translator outage degrades to a quality dip, not a failure.
- Shipped containerized services on GKE provisioned by Terraform, with horizontal pod autoscaling (1–5 pods at 70% CPU) configured on both the app and the Milvus vector store.

**Gemma4: Local-First LLM Chat and Agent Workbench** Personal Project  
*Sandboxed Human-in-the-Loop Coding Agent / JavaScript, Flask, Ollama, Whisper* Apr 2026 – Sep 2026

- Built a local coding agent with a human-in-the-loop gate: every write, edit, and shell command runs only after diff approval, enforced under a macOS `sandbox-exec` profile that denies writes outside the project root and denies network entirely.
- Wrote one message-and-streaming layer speaking Ollama, OpenAI-compatible, and Anthropic-compatible backends through a single internal format, including reassembling OpenAI's index-fragmented tool-call arguments across streamed chunks, so tool use behaves identically across providers.
- Managed the context window client-side with pinning, budgeted pruning, and background summarization, after finding the backend truncates over-budget prompts from the front and silently drops the system prompt; kept tool-call and tool-result pairs structurally intact across pruning.

## Big-Data Predictive Analytics on Longitudinal Cohort-Style Data

Carnegie Mellon University

*Distributed ETL, Streaming, and CI/CD at Scale | PySpark, Kafka, PostgreSQL, GCP*

*Aug 2025 – Nov 2025*

- Designed a distributed PySpark ETL pipeline ingesting 130K+ longitudinal records (2015–2022) into a unified PostgreSQL schema, and trained Spark ML and PyTorch regression models reaching 0.95 MSE and 0.98  $R^2$ .
- Built a Kafka Pub/Sub pipeline with an external-API ingestion arm on GCP to cross-validate model predictions against real-world reference signals from a live upstream feed.
- Reconciled schema drift across seven seasons of source data (attribute renames, unit changes, taxonomy shifts) before joining to a unified panel, so a late-arriving season could be re-ingested into an idempotent per-season partition without rewriting the historical cohort.
- Implemented a CI/CD workflow for automated feature engineering, model retraining, and validation across the longitudinal cohort, ensuring reproducibility and version-controlled experiment tracking.

## Wet-Lab and Molecular Biology Project

### B.Tech Capstone: Identifying the Role of CK1 $\epsilon$ in Cisplatin-Resistant SCLC Cells

IIT Jodhpur

*Cell Culture, Lentiviral Gene Silencing, Drug-Resistance Mechanism | MTT, shRNA, GFP Selection* Aug 2024 – May 2025

- Established cisplatin-resistance profiles in small-cell lung cancer lines (H69 parental vs. H69AR resistant) using MTT viability assays, generating the dose-response baseline the rest of the study was built on.
- Performed lentiviral shRNA transfection for stable CK1 $\epsilon$  knockdown in resistant cells, using GFP-based selection to confirm transduction efficiency and isolate the knockdown population.
- Maintained mammalian cell culture under aseptic technique across the full study, including line passaging, resistance maintenance under drug pressure, and viability tracking.
- Designed the downstream validation plan: qPCR and Western blot for knockdown confirmation, Annexin V staining for apoptosis, and a cisplatin+umbralisib combination arm to test whether CK1 $\epsilon$  loss re-sensitizes resistant cells.
- Positioned CK1 $\epsilon$  as a candidate therapeutic target for chemoresistance in SCLC, with the knockdown system built to be reusable for other kinase targets in the same lines.

## Leadership, Teaching, and Selected Extracurriculars

- **Teaching Assistant**, Introduction to Biomedical Engineering, CMU (Jan 2026 – May 2026); previously TA for Introduction to Bioengineering, IIT Jodhpur, managing 236 students across assignments and grading (2022–2023).
- **Student Ambassador**, Biomedical Engineering Program, CMU (May 2026 – Current): represent the BME graduate program to prospective students and at recruitment events.
- **The Latent Map**, CMU x Anthropic Claude Builder Club Hackathon (Apr 2026): built a Claude-powered adaptive-interview system mapping user archetype and latent interests, with a Flask backend and d3.js visualization frontend.
- **Non-Invasive Malnutrition Screening Device**, ACM ICMI 2023, in collaboration with AIIMS Jodhpur clinicians: portable IoT and computer-vision device integrating BMI, weight-for-age, and height-for-age metrics into a real-time digital interface for point-of-care nutritional assessment in resource-limited settings.
- **InterIIT Tech Meet, IIT Jodhpur Contingent** (2023–2025): built a real-time financial agentic system on the Pathway framework with streaming RAG, live signal ingestion, and LLM-grounded insights.