Ketaki Dabade

(651) 384-8787 | kvd2112@columbia.edu | linkedin | github | ketaki.dev | google scholar

Education

Columbia University — New York, NY
Master of Science in Computer Science (Machine Learning Track) — *Aug 2025 – Dec 2026*

- **Coursework:** Neural Networks & Deep Learning, NLP, Analysis of Algorithms, Continual Learning & Memory Models, Financial Engineering, Databases

Dr. Vishwanath Karad MIT World Peace University — Pune, IN
Bachelor of Technology in Computer Science and Engineering, CGPA: 3.74/4.0 — *Jul 2021 – Jul 2025*

- **Coursework:** Data Structures, Operating Systems, Computer Networks, OOP, Discrete Mathematics, AI, Statistics & Probability, Distributed Computing, HPC

Work Experience

Complex Resilient Intelligent Systems Laboratory, Columbia University — New York, NY
Research Assistant under Professor Venkat Venkatasubramanian — *Sept 2025 – Present*

- Building a knowledge-mapping pipeline for STEM textbooks to support Sparse Autoencoder training on structured knowledge representations. Chose BERTopic over LDA after LDA failed to preserve cross-chapter concept relationships in initial experiments.
- Implemented MinerU for PDF-to-Markdown conversion across 3,000+ pages after PyMuPDF and Nougat both broke on multi-column layouts with inline equations. Generated 17,000+ dense embeddings using Qwen3-Embedding for corpus-wide semantic similarity.
- Discovered 493 coherent topics across 102 textbook chapters (0.9791 mean cosine similarity) and built hierarchical clustering that surfaces prerequisite chains, e.g., "Atomic Structure" must precede "Chemical Bonding" in learning paths.

LivingScopeHealth — Remote
Research Collaborator — *Jan 2026 – Present*

- Querying 2M+ patient records in PostgreSQL (hospital visits, prescriptions, lab panels) to find early diabetes indicators that appear before clinical diagnosis. Initial EDA revealed visit frequency spikes 18 months pre-diagnosis as a stronger signal than lab values alone.
- Building XGBoost and Random Forest classifiers on longitudinal patient features. Applied SMOTE after the positive class sat at 8%; SHAP analysis surfaced complaint-pattern features that clinicians had not previously prioritized.

AI4M Technology Private Limited — Pune, IN
Deep Learning Engineer Intern — *Jul 2024 – Dec 2024*

- Trained YOLOv7/v8 for manufacturing defect detection (scratches, dents, misalignments). Deployed on NVIDIA Jetson with TensorRT FP16/INT8 quantization, achieving 3x inference speedup and 25% latency reduction over the baseline Flask pipeline.
- Designed REST APIs serving real-time inference across 3 production lines. Built multi-threaded Docker backend with AWS S3/EC2 and Azure for data ingestion, plus automated daily defect reporting consumed by 4 manufacturing clients.
- Established CI/CD pipeline with 85% code coverage. System runs in production today, catching defects that human inspectors consistently missed during night shifts.

ViLA EmachWirken Private Limited — Pune, IN
Data Analyst Intern — *Jun 2022 – Dec 2022*

- Built K-Means clustering on 100K+ transactions to segment customers into 5 personas that the marketing team used to restructure their email campaigns. Deployed Grafana dashboards tracking 15+ KPIs (revenue, churn, CAC).
- Automated data extraction and reporting pipelines, cutting manual reporting from 10 hours/week to 6. Dashboards are still in active use 3+ years later.

Publications

EEG-Powered Brain-Computer Interface for 3D Hand Gesture Control — *IEEE ICICIS 2025 (First Author)* 2026
SkillSet Sherpa: Career Counseling with Large Language Models — *Springer LNNS* 2025
ViziAssist: Visual Assistance for Visually Impaired Drivers — *Springer CCIS* 2025

Projects

PaperTrail – SEC Filing Contradiction Detector | GitHub 2026

- Real-time contradiction detector for S&P 500 SEC filings. Streams EDGAR ingestion (10-K, 10-Q, 8-K, Form 4) via Kafka-backed microservices with under 5-minute end-to-end latency from filing detection to user alert.
- Fine-tuned FinBERT on 5,000 labeled financial sentences for claim classification. Built hybrid retrieval: pgvector for semantic search over sentence-transformer embeddings, Neo4j temporal knowledge graph for structured queries (CONTRADICTS, TRADED edges).
- Orchestrated LLM agent (Claude/GPT-4 via LangChain) with 5 custom tools: negation detection, temporal reasoning, severity scoring, insider transaction lookup, and filing diff. Next.js dashboard with live WebSocket feed and insider transaction timeline overlay.

Patrona – AI Voice Safety Companion | ***1st Place, Columbia AI for Good Hackathon ($5K)*** | GitHub 2026

- Voice-first AI companion for walking safety using ElevenLabs Conversational AI, WebRTC, Supabase, and real-time GPS. Built in 36 hours by a team of 2; awarded $5,000 in ElevenLabs credits to continue development.
- Designed multi-tier silence detection (90s check-in, then escalation) and NLP-based safe word system that silently triggers emergency SMS with reverse-geocoded street addresses. Chose OpenStreetMap over Google Maps API to avoid rate-limit costs during a hackathon with no budget.

Quant Portfolio Returns Dashboard | GitHub 2025

- Portfolio analytics dashboard computing 15+ risk metrics (Sharpe, Sortino, VaR, CVaR, Beta, Alpha, Max Drawdown). Built the mean-variance optimizer from scratch using SciPy SLSQP rather than a library like PyPortfolioOpt to understand the constraint formulation directly.
- Monte Carlo engine runs 1,000+ scenario simulations with configurable horizons. Modular architecture: separate calculation engine, yfinance data layer, SQLite caching, Streamlit frontend, Docker containerization.

Cross-Lingual Indic Hate Speech Detection | GitHub 2025

- Investigated whether a model trained on Hindi hate speech can detect Marathi hate speech with zero target-language examples. Compared IndicBERT-v2 (corpus scale) vs MuRIL (translation-aware pretraining) for cross-lingual transfer.
- Key result: LoRA achieved F1 $\approx$ 0.80 updating only 0.95% of parameters, while full fine-tuning collapsed to F1 $\approx$ 0.39. LoRA acts as a structural regularizer, not just a compute shortcut. MuRIL outperformed by 2.1% F1 with just 50 target-language examples.

Pinterest Duplicate Detector | GitHub 2024

- Image retrieval system using CLIP embeddings + FAISS indexing for sub-second similarity search across 10K+ images. Combined perceptual hashing, SSIM, and neural embeddings because no single metric handled both near-duplicates and semantically similar images reliably.
- FastAPI backend with Streamlit frontend for interactive duplicate detection and analysis.

Technical Skills

Languages: Python, C/C++, Java, JavaScript/TypeScript, SQL, R, Bash
ML & NLP: PyTorch, TensorFlow, Scikit-learn, XGBoost, JAX, HuggingFace Transformers, LangChain, LlamaIndex, LoRA/PEFT, RAG, fine-tuning, BERTopic, CLIP, YOLOv7/v8, TensorRT, ONNX
Quantitative & Data: Monte Carlo simulation, mean-variance optimization, VaR/CVaR, ARIMA/GARCH, Black-Scholes, backtesting, QuantLib; Pandas, NumPy, SciPy, Polars, Plotly, Tableau
Infrastructure: FastAPI, Flask, React, Node.js; PostgreSQL, MongoDB, Neo4j, Redis, pgvector, FAISS, Kafka, Spark; Docker, Kubernetes, CI/CD
Cloud: AWS (S3, EC2, Lambda, SageMaker), Azure, GCP; Git, Vercel

Awards & Recognition

1st Place, Columbia AI for Good Hackathon 2026 — Patrona AI Voice Safety Companion ($5,000 ElevenLabs credits)
2nd Place, HACKMITWPU 2024 — CanMan Canteen Management System
Top 100 Nationally, KPIT Hackathon 2022 — ViziAssist ADAS Project
3 Peer-Reviewed Publications — 1 IEEE, 2 Springer conference proceedings

Organizations

Columbia Lioness Quantitative — Member
Society of Women Engineers (SWE), Columbia — Member