Ketaki Dabade

(651) 384-8787 | kvd2112@columbia.edu | linkedin | github.com/ketakiii3 | portfolio-website

Education

Columbia University New York, NY
Master of Science in Computer Science *Aug 2025 – Dec 2026*

- Research Focus: Machine Learning Track
- Laboratory: Complex Resilient Intelligent Systems (CRIS) Lab under Professor Venkat Venkatasubramanian

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*

- Published 2 research papers in Springer conferences (LNNS and CCIS)
- Developed expertise in computer vision, edge computing, and deep learning through hands-on projects

Research Experience

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

- **Scientific Content Analysis Pipeline:** Building infrastructure for understanding how knowledge is structured in educational content and how machines can learn these structures for downstream applications.
- **PDF Extraction Layer:** Implemented MinerU for crash-resistant PDF-to-Markdown conversion, handling complex textbook formatting including equations, figures, tables, and multi-column layouts across 3,000+ pages.
- **Embedding Generation:** Used Qwen3-Embedding to generate 17,000+ dense vector representations of textbook passages, enabling semantic similarity computations across the entire STEM corpus.
- **Topic Discovery:** Applied BERTopic with HDBSCAN clustering to discover 493 semantically coherent topics across 102 textbook chapters in Biology, Physics, and Chemistry. Validated coherence with 0.9791 mean cosine similarity.
- **Knowledge Mapping:** Leveraged Gemma for human-readable topic labeling and built hierarchical clustering to map prerequisite relationships — demonstrating how concepts like 'Atomic Structure' must precede 'Chemical Bonding' in learning paths.
- **Impact:** This work lays the foundation for Sparse Autoencoder (SAE) training on structured knowledge, contributing to our understanding of how LLMs represent and organize information internally.

AI and ML Lab of NITTTR - Siemens Centre of Excellence Bhopal, IN
Research Assistant *Feb 2024 – Mar 2024*

- **Music-Mental Health Correlation Analysis:** Investigated the quantifiable relationship between music listening habits and mental health outcomes using survey data from 1,000+ participants.
- **Feature Engineering:** Extracted meaningful features from music consumption patterns — genre preferences, listening duration, tempo preferences, and contextual factors (when and why people listen).
- **Ensemble Modeling:** Built and compared Random Forest and Gradient Boosting classifiers, achieving 93.19% accuracy in predicting mental health indicators from music behavior patterns.
- **Key Insight:** Identified statistically significant correlations between specific genre preferences and anxiety/depression scores, contributing to the emerging field of music therapy research.

Work Experience

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

- **Challenge:** Build a defect detection system capable of processing 1000+ frames per second on edge hardware with limited compute for real-time manufacturing quality control.
- **Model Development:** Trained and optimized YOLOv7 and YOLOv8 object detection models for identifying manufacturing defects — scratches, dents, misalignments, and color inconsistencies.
- **Edge Deployment:** Deployed models on NVIDIA Jetson GPU using DeepStream SDK for video stream processing. Implemented TensorRT optimization (FP16/INT8 quantization) achieving 3x inference speedup.
- **API Architecture:** Designed REST APIs using Flask for model inference, allowing the production monitoring system to query defect predictions in real-time across multiple production lines.
- **Infrastructure:** Built multi-threaded Python backend with Docker containerization. Established CI/CD pipeline for seamless model updates. Achieved 85% code coverage with comprehensive unit tests.
- **Results:** Reduced detection latency by 25% across 3 production lines. System now runs in production, catching defects that human inspectors miss.

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

- **Customer Segmentation:** Built K-Means clustering models to segment customers based on purchasing behavior. Identified 5 distinct customer personas that informed targeted marketing strategies.
- **Dashboard Development:** Designed and deployed interactive Grafana dashboards tracking 15+ KPIs — revenue trends, customer acquisition costs, churn rates, and operational efficiency metrics.
- **Data Analysis:** Conducted exploratory analysis on 100K+ transaction records using Python (Pandas, NumPy) and SQL. Translated findings into weekly reports for management with actionable insights.
- **Process Automation:** Automated data extraction and reporting pipelines, reducing manual reporting time by 40%.
- **Impact:** Enhanced operational visibility by 30%. Dashboards are still in active use today.

Publications

SkillSet Sherpa: AI-Powered Career Guidance Platform with LLM Integration
Springer Lecture Notes in Networks and Systems (LNNS), 2024
DOI: 10.1007/978-981-97-9523-9_23

ViziAssist: Assistive Driving System for Visually Impaired Individuals
Springer Communications in Computer and Information Science (CCIS), 2022
DOI: 10.1007/978-3-031-93688-3_6

Projects

Quant Portfolio Returns Dashboard | GitHub 2025

- **Technologies:** Python, Streamlit, Plotly, SciPy, NumPy, Pandas, yfinance, SQLite, SQLAlchemy, Docker
- Built comprehensive real-time portfolio analytics dashboard for quantitative finance enthusiasts seeking institutional-grade analytics without Bloomberg terminal costs.
- **Risk Metrics:** Computes 15+ metrics including Sharpe Ratio, Sortino Ratio, VaR (95%), CVaR/Expected Shortfall, Beta, Jensen's Alpha, Maximum Drawdown, Information Ratio, Treynor Ratio, and R-Squared.
- **Portfolio Optimization:** Implements mean-variance optimization using SciPy's SLSQP solver to visualize the efficient frontier and identify maximum Sharpe/minimum volatility portfolios.
- **Monte Carlo Simulation:** Runs 1,000+ scenario simulations for probabilistic future projections with configurable time horizons and confidence intervals.
- **Return Calculations:** Calculates both time-weighted returns (TWR) and money-weighted returns (IRR) with benchmark comparisons against S&P 500 and Nasdaq.
- **Architecture:** Modular design with separate calculation engine, data fetching layer (yfinance API), SQLite-backed caching and Streamlit frontend. Containerized with Docker.

Cross-Lingual Indic Hate Speech Detection | GitHub 2025

- **Technologies:** PyTorch, HuggingFace Transformers, LoRA, PEFT, IndicBERT-v2, MuRIL
- Research project investigating cross-lingual transfer learning — specifically whether a model trained on Hindi hate speech can detect Marathi hate speech with zero or minimal examples.
- **Research Question:** Compared pretraining strategies — massive monolingual corpora (IndicBERT-v2) vs translation-aware pretraining (MuRIL) for cross-lingual transfer efficiency.
- **Key Finding 1:** LoRA (Low-Rank Adaptation) achieves F1 $\approx$ 0.80 while updating only 0.95% of model parameters.
- **Key Finding 2:** Full fine-tuning catastrophically fails (F1 $\approx$ 0.39), collapsing to majority-class predictions.
- **Key Finding 3:** LoRA acts as a crucial structural regularizer in low-resource settings — parameter-efficient fine-tuning isn't just computationally cheaper, it's necessary for stable training.
- **Transfer Analysis:** IndicBERT-v2 excels at zero-shot transfer due to corpus scale; MuRIL shows superior few-shot adaptability, outperforming by 2.1% F1 with just 50 target-language examples.

ViziAssist ADAS - Assistive Driving System | GitHub 2022

- **Technologies:** NVIDIA Jetson Nano, YOLOv7, TensorRT, OpenCV, Raspberry Pi Camera
- Assistive driving system designed to help visually impaired individuals navigate safely with real-time obstacle detection on edge hardware.
- **Model:** Custom YOLOv7 model trained for road obstacles (pedestrians, vehicles, potholes, barriers) achieving 0.681 mAP on test dataset.
- **Optimization:** Ooptimization for real-time inference on NVIDIA Jetson Nano with limited power and compute.
- **Integration:** Raspberry Pi camera for live video feed with audio feedback system to alert users of detected obstacles.
- **Recognition:** Published in Springer CCIS; Top 20 nationally at KPIT Hackathon 2022; Led team of 4 engineers.

SkillSet Sherpa - Career Guidance Platform | *Published in Springer LNNS* 2024

- **Technologies:** Flask, GPT-3, LangChain, EasyOCR, OpenCV, NLTK, HTML/CSS, JavaScript
- AI-powered career counselor that analyzes resumes and personality assessments to suggest personalized career paths.
- **Resume Parsing:** Upload PDF/image resumes processed by EasyOCR for text extraction, OpenCV for preprocessing, and NLTK for entity extraction (skills, education, experience).
- **Psychometric Assessment:** RIASEC personality assessment integration to understand work style preferences.
- **AI Recommendation:** GPT-3 with custom prompt engineering synthesizes skills, experience, and personality to suggest matching career paths with detailed reasoning.

CanMan - Canteen Management System | GitHub 2024

- **Technologies:** Flask, MongoDB, React, D3.js, NLTK, HTML/CSS, JavaScript
- Full-stack canteen management system with intelligent NLP chatbot for natural language food ordering.
- **NLP Chatbot:** Natural language ordering — ”I want a coffee and samosa” gets parsed into structured order.
- **Analytics Dashboard:** D3.js visualizations for sales trends, inventory levels, and demand forecasting.
- **Architecture:** React frontend, Flask REST API backend, MongoDB database with real-time updates.
- **Recognition:** Won 2nd place at HACKMITWPU 2024; Led team of 5 developers.

Pinterest Duplicate Detector | GitHub 2025

- **Technologies:** CLIP, FAISS, PyTorch, FastAPI, Streamlit, OpenCV
- Content-based image retrieval system for finding duplicate and similar images at scale.
- **Embeddings:** CLIP (Contrastive Language-Image Pretraining) for semantic visual embeddings that understand image content.
- **Indexing:** FAISS vector indexing for efficient similarity search across 10K+ images with sub-second query times.
- **Multi-Metric Scoring:** Combines perceptual hashing, structural similarity (SSIM), and neural embeddings for robust duplicate detection.
- **Architecture:** FastAPI backend for API endpoints + Streamlit frontend for real-time interactive analysis.

EEG Brain-Computer Interface | GitHub 2025

- **Technologies:** Python, Scikit-learn, Emotiv EPOC X, Blender, Unity
- Control a virtual 3D hand using only brainwaves — end-to-end BCI pipeline from signal acquisition to 3D visualization.
- **Data Acquisition:** Collected EEG signals from participants (ages 20-22) using Emotiv EPOC X headset at 256Hz sampling rate.
- **Feature Extraction:** Applied FFT and wavelet transforms to extract frequency-domain features from raw EEG signals.
- **Classification:** KNN classifier achieving 97.63% accuracy in distinguishing between hand gesture intentions.
- **Visualization:** Real-time 3D hand animation in Blender with Unity integration for interactive demo.

One View - Event Management System 2024

- **Technologies:** Flask, MongoDB, DBSCAN, OpenCV
- Web-based event management application with intelligent photo organization using facial clustering.
- **Smart Photo Clustering:** DBSCAN clustering on facial embeddings to automatically group event photos by person.
- Helps event organizers quickly find and share photos with specific attendees without manual tagging.

Automated Door Lock System 2023

- **Technologies:** Arduino, R307 Fingerprint Sensor, C++
- Secure biometric door lock system built from scratch for home security and lab access control.
- **Hardware:** R307 optical fingerprint sensor for biometric authentication with Arduino microcontroller.
- **Software:** Optimized fingerprint matching algorithm for faster authentication with secure enrollment system.

Technical Skills

Programming
Python (primary), C/C++, JavaScript, SQL, R, HTML/CSS

Machine Learning, Deep Learning & NLP
PyTorch, TensorFlow, Keras, Scikit-learn, HuggingFace Transformers, LangChain, BERTopic, spaCy, NLTK; Transformer fine-tuning (LoRA, PEFT, SFT), text classification, NER, topic modeling, semantic search, cross-lingual transfer, RAG

Computer Vision & Edge AI
YOLOv7/v8, OpenCV, CLIP embeddings, image segmentation, perceptual hashing, SSIM, data augmentation; CUDA, TensorRT, DeepStream SDK, Triton Inference Server, NVIDIA Jetson

Quantitative & Statistical Analysis
Risk metrics (VaR, CVaR, Sharpe, Sortino, Beta, Alpha), Monte Carlo simulation, mean-variance optimization, time-series analysis

Systems, Web & Infrastructure
Flask, FastAPI, REST APIs, React, D3.js, Streamlit; PostgreSQL, MongoDB, SQLite; Docker, Git, Linux, CI/CD, pytest, Grafana, Hadoop, Plotly, yfinance

Hardware & Embedded Systems
Arduino, Raspberry Pi, Emotiv EPOC X (EEG)

Certifications

Google Project Management Professional Certificate (Google/Coursera)	Nov 2024
Machine Learning Specialization (DeepLearning.AI / Stanford)	Jul 2024
Data Analytics & Visualization Job Simulation (Accenture / Forage)	Mar 2024
Introduction to AI in the Data Center (NVIDIA DLI)	Feb 2024
Git & GitHub Bootcamp (Udemy)	Feb 2024
Google Data Analytics Professional Certificate (Google/Coursera)	Dec 2023
Data Structures & Algorithms (C/C++) (Udemy)	Sep 2023

Awards & Recognition

2nd Place, HACKMITWPU 2024 — CanMan Canteen Management System
Top 100 Nationally, KPIT Hackathon 2022 — ViziAssist ADAS Project
2 Springer Publications — Research published in LNNS and CCIS conference proceedings