Keshav Krishna

New York City, US | +1 (479) 685 5463 | kk6081@nyu.edu | Linkedin | Github | Website

Summary

Graduate software engineer specializing in machine learning and computer vision, with experience prototyping image processing pipelines and deploying models at scale. Delivered deep learning solutions on real-world datasets, including models achieving 85%+ accuracy. Seeks to advance applied research in next-generation semiconductor imaging at KLA.

EDUCATION

New York University Courant **Aug 2025 - May 2027**
MS, Computer Science (GPA: 4.0) *New York City*
- **Coursework:** Machine Learning, Honors Analysis of Algorithms, Programming Languages

Indian Institute of Technology, Ropar **Aug 2019 - Jun 2023**
Bachelors, Computer Science (GPA: 8.74/10) *Ropar, India*
- **Achievements:** Merit Scholarship for top performance

SKILLS SUMMARY

- **Frameworks and Libraries:** PyTorch, TensorFlow, Hugging Face, pandas
- **Data & Tools:** Docker, GIT, JIRA, Postgres, redis, FastAPI, MySQL, Airflow, pyspark, hive, Qdrant, Annoy, ETL
- **Machine Learning & Computer Vision:** Deep Learning, YOLO, BERT, NudeNet, Transformers, Multimodal Learning, Image Processing, SetFit
- **Languages:** C, C++, Python, SQL, Unix scripting, Java

EXPERIENCE

JioSaavn | *Software Engineer(Data Science)* **Jul 2023 - Jul 2025**
- User Content Mixes: Built early-morning and late-night content mixes for 12M daily users using Hive, Spark, Redis and PHP; drives 1M impressions, 800K clicks and 500K streams per day.
- Topic Mixes (E2E): Generated mood-based topic mixes for 7.6M daily users via Hive, Spark and TF-IDF; added deduplication filters and Flask endpoints for PHP backend; served to 100M app users with 12.81% CTR and 8.5 streams/user.
- Deployed transformer and few-shot learning models, including SetFit, hing-roberta and prompt-engineered Llama3-8B-instruct, to improve playlist moderation, boosting accuracy to 85% on manually labeled data.
- Created a custom BERT-based model for user-song embeddings, increasing top-5 and top-10 accuracy to 61.82% and 67.21%, and enhanced large-scale recommendations by integrating deep neural networks.
- Developed a retrieval-augmented generation pipeline using LLaMA3-7B for prompt intent classification and multimodal cover art generation with the segmid/SSD-1B diffusion model, demonstrating applied research and prototyping skills.
- Reduced infrastructure costs by migrating data from Annoy indices to Qdrant, purging unused data, and cleaning tables for improved lifecycle management

GE Healthcare | *Software Intern* **May 2022 - Jul 2022**
- Face De-identification: Applied YOLO algorithm to train a model for identifying and blurring faces in medical data, achieving over 90% accuracy on 20,000+ images.
- Person Counting and Nudity Detection: Developed person and nudity detection models using YOLO and NudeNet, reaching over 85% mAP and accuracy on medical image data.
- Packaged research prototypes as Docker images with FastAPI endpoints, streamlining global access and integration into engineering workflows.

RESEARCH

- Single-author research paper, "Advancements in Cache Management: A Review of Machine Learning Innovations for Enhanced Performance and Security," surveying 40+ works on ML-based cache management (replacement, dynamic policies, edge caching, and security). Reviewed techniques including RL, LSTMs, imitation learning, and deep neural networks for cache optimisation and attack detection; published in Frontiers in AI (2025)..*https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1441250/abstract*

ACADEMIC PROJECTS

Pico-LLM: Unified SLM Training & Interpretability | *NYU Courant* **Sep 2025 - Dec 2025**
- **Unified Training Framework:** Developed a codebase for Small Language Models (SLMs) supporting **SFT** and **DPO/GRPO** alignment on reasoning datasets like GSM8K.
- **Mechanistic Interpretability:** Implemented visualization tools (**Logit Lens**, Attention Maps) to diagnose feature activations and internal model states.
- **Inference Optimization:** Engineered a robust inference engine with **KV-Caching** and automatic architecture detection for efficient batched generation.

Patient Audit and FeedBack App | *IIT Ropar* **Jan 2022 - May 2022**
- Focused on making record keeping process 100% automatic by residents,professors
- Used Flutter to make secure persistent login, option to star patients, sort, search, get all info and add new info for a patient, make graph of data
- Used NodeJS and PostgreSQL for backend. 100% secure OTP generation by email auth
- Deployed database and backend on Heroku, Able to handle more than 100 patients data and multiple users simultaneously
- Presented to PGI officials, got good response and used by doctors

ACHIEVEMENTS

- **Merit Scholarship:**IIT Ropar Merit Scholarship Cash scholarship provided to top 7% students by CGPA all over the batch(2019-20) link
- **Mathematics Olympiad:**Gold Medal at RMO Top Scorer at RMO (Regional Mathematics Olympiad) conducted by HBCSE (2017)
- **Hackathon Win:**Jumpstart'22 Finalist Hackathon conducted by Publicis Sapient. Passed all rounds from among 29864 people(2022)
- **Competitive Coding Win:**Competetive Coding Codechef 5 star, rating 2021, Codeforces Specialist (2022)
- **Google competition Top performance:**Google Kickstart Round D Got Global rank 1547 out of 7500+ registrations (2022)
- **IIT Entrance exam top rank:**JEE Mains & Advanced Mains - Got AIR 2505 out of 1.14 million & Advanced - Got AIR 3293 out 161,000 (2019)