

MOKSHADA NEHETE

+91-9284969160 | mokshadanehete14@gmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

EDUCATION

Indian Institute of Information Technology (IIIT) Pune

B.Tech in Computer Science and Engineering with Honors in AI

Aug 2023 – May 2027

CGPA: 9.79 — Honors CGPA: 10

Indian Institute of Technology (IIT) Madras

B.S. in Data Science and Applications

Jan 2024 – July 2027

CGPA: 9.88

WORK EXPERIENCE

AI Summer Residency Program | IIM Ahmedabad Ventures

May 2026 – Jun 2026

- Selected as one of **43 participants** from **32,000+ applicants** for the AI Summer Residency by IIMA Ventures.
- Built an **MVP of an LLM optimization proxy** for enterprise AI workloads.
- Implemented **prompt canonicalization, context deduplication, semantic caching, and request restructuring**.

Carter AI | Intern

Nov 2025 – Jan 2026

- Built a **mobile interface** for an AI fashion engine leveraging **LLM-driven semantic search**.
- Integrated **embedding-based retrieval (Pinecone)** and **LLM reranking APIs** for contextual recommendations.
- Handled high-scale inference outputs over **1M+ SKUs** with **low-latency caching (Redis)**.
- Supported **multilingual queries (14+ Indic languages)** for personalized fashion discovery.

PROJECTS

Energy-Aware Task Scheduling | | PyTorch, DQN, Cloud Scheduling

Mar 2026 – Apr 2026

- Developed an energy-aware cloud task scheduler using a **Deep Q-Network (DQN)** to optimize task allocation across virtual machines by jointly minimizing energy consumption while maximizing resource utilization and QoS.
- Engineered a **multi-objective reward function** incorporating CPU, RAM, memory, disk utilization, uptime, and response time, enabling adaptive scheduling under dynamic cloud workloads.
- Designed a modular reinforcement learning pipeline comprising environment simulation, agent, scheduler, replay buffer, and evaluation modules, simplifying experimentation with different scheduling policies.
- Implemented an **adaptive ϵ -greedy exploration strategy**, experience replay, and target network updates to improve convergence stability and scheduling performance across varying workload distributions.
- Benchmarked the scheduler on workloads ranging from **50–1000 concurrent tasks**, evaluating energy consumption, throughput, response time, resource utilization, and Pareto-optimal trade-offs against baseline scheduling algorithms.
- Extended the reference IEEE implementation with richer state representations, configurable reward engineering, comprehensive evaluation metrics, achieving improved scheduling efficiency over the baseline implementation.

Flight Operations Data Engineering Pipeline | GitHub | Python, Apache Airflow, Snowflake, Docker

Apr 2026

- Developed an end-to-end ETL pipeline using **Apache Airflow** to orchestrate scheduled ingestion of live flight-state data from the **OpenSky Network API**, with automated retries, task dependencies, and execution every 30 minutes.
- Implemented a **Bronze–Silver–Gold Medallion Architecture**, storing raw API responses in the Bronze layer, performing schema validation, null handling, deduplication, and feature engineering in the Silver layer, and generating aggregated analytical datasets in the Gold layer.
- Built modular Python ETL components to transform high-volume flight telemetry, including aircraft metadata, geospatial coordinates, altitude, velocity, heading, and temporal attributes for downstream analytics.
- Integrated the pipeline with **Snowflake**, automatically creating tables, loading curated datasets, and optimizing analytical queries for reporting and business intelligence dashboards.
- Containerized the complete data engineering stack using **Docker Compose**, enabling reproducible deployment of Airflow services, scheduler, webserver, metadata database, and ETL workers.
- Designed modular DAGs with configuration-driven workflows, environment variable management, comprehensive logging, and fault-tolerant execution for scalable and maintainable data pipelines.

MedAssist: LLM-Based Clinical Intake System | GitHub | Python, FastAPI, React,

Gemini, PubMedBERT, Celery

Feb 2026 – May 2026

- Built an **AI-powered clinical intake platform** for **symptom assessment**, medical report analysis, and medication safety using **LLMs** and **NLP**.
- Designed a **hybrid AI pipeline** combining **PubMedBERT embeddings**, semantic similarity, and **LLM reasoning** for condition ranking and drug interaction detection.
- Automated **SOAP note generation** and **OCR-based** medical report summarization to provide structured clinical insights.
- Implemented **real-time risk scoring**, **QR-based emergency profiles**, and **Celery-powered** medication reminder workflows.

Messy Mashups: Audio Genre Classification |  GitHub | *PyTorch, torchaudio, AST, librosa* Jan 2026 – Feb 2026

- Built a **noise-robust audio classification pipeline** using **Mel Spectrograms** and **GPU-based preprocessing** for efficient training on music genre data.
- Applied **Mixup**, **SNR-based augmentation**, and **tempo alignment** to improve robustness against noisy and time-shifted audio inputs.
- Scaled the model from **CNN/ResNet** baselines to an **Audio Spectrogram Transformer (AST)** with **self-attention** for stronger feature learning.
- Optimized training with **AdamW**, **Focal Loss**, and **test-time augmentation (TTA)**, achieving a **0.964 Macro F1 score**.

AI-Powered Offline Disk Analyzer |  GitHub | *Python, Flask, React, FAISS, Docker* Sep 2025 – Dec 2025

- Built a **multimodal offline retrieval system** over **1000+ files**, supporting **5+ query modes** including text, image, face, and object search using **BGE-M3 (1024-d)** and **SigLIP (768-d)** embeddings.
- Implemented **FAISS-based vector search** with **hybrid reranking**, achieving **85–90% accuracy** with under **5 s latency** for offline semantic retrieval.
- Enabled **text-to-text**, **text-to-image**, **image-to-image**, and **face/object search** using **YOLO**, **ArcFace (512-d)**, and **LLM-based QA** for structured offline exploration.

Data Analyst Agent |  GitHub | *Python, LLMs, Pandas, SQL, Matplotlib, Docker* Aug 2025

- Built an **LLM-powered autonomous data analysis agent** capable of ingesting and processing large-scale datasets from **CSV files**, **APIs**, **S3**, and **SQL sources**.
- Designed a **multi-stage agentic pipeline** with planner, retrieval, analysis, and visualization modules to automate end-to-end data exploration and reporting.
- Implemented automated **EDA**, **feature extraction**, and **statistical analysis** using **Pandas** and **NumPy** to generate structured insights from heterogeneous data.
- Enabled **natural language querying** with concurrent request handling, delivering analysis outputs and visualizations in under **3 minutes**.
- Containerized the system with **Docker** for reproducible deployment and scalable execution across environments.

Virtual TA Assistant |  GitHub | *FastAPI, FAISS, RAG, LLMs* Jun 2025

- Built an **LLM-powered API** to answer **course-related queries** using **retrieval-augmented generation (RAG)** over course content and **1000+ Discourse posts**.
- Engineered a **FAISS-based retrieval pipeline** with **all-MiniLM-L6-v2** embeddings for semantic search, context grounding, and source-aware answer generation.
- Designed an end-to-end system supporting **JSON-based queries** and **base64 file inputs**, returning responses with **source links** in under **30 seconds**.
- Automated data scraping, chunking, and indexing with date-range filtering to keep the knowledge base continuously up to date.

Cinema Audience Forecasting | *Python, XGBoost, LightGBM, Scikit-learn, Pandas*

- Developed a machine learning pipeline to forecast **movie audience demand** using historical box office, release metadata, seasonal trends, and genre-specific features.
- Performed extensive **feature engineering**, including temporal features, categorical encoding, missing value imputation, and correlation analysis to improve prediction accuracy.
- Trained and evaluated multiple regression models including **XGBoost**, **LightGBM**, and Random Forest, selecting the best-performing model through cross-validation and hyperparameter tuning.
- Built an end-to-end forecasting workflow covering data preprocessing, model training, evaluation, and visualization of predicted versus actual audience trends.

Protein Secondary Structure Prediction | *PyTorch, PyTorch Lightning, BiLSTM, GRU* Dec 2025

- Developed an end-to-end **sequence-to-sequence deep learning pipeline** for predicting protein secondary structures from amino acid sequences, jointly learning **Q8 (8-state)** and **Q3 (3-state)** structural annotations through dual-head sequence labeling.

- Designed and benchmarked **six neural architectures**, including Bidirectional RNN, multi-layer BiLSTM, GRU, Inception-BiLSTM, ResNet-BiLSTM, and SE-ResNet-BiLSTM, to improve long-range dependency modeling and contextual sequence representation.
- Engineered a modular **PyTorch Lightning** training framework with custom tokenization, dynamic sequence padding, masking, gradient clipping, checkpointing, mixed-precision training, and configurable learning-rate scheduling for efficient experimentation.
- Implemented a complete experimentation pipeline using **TrackIO** for experiment tracking, model versioning, hyperparameter optimization, and reproducible benchmarking across multiple sequence-modeling architectures.
- Optimized token-level prediction using **Macro-F1** evaluation, cross-entropy based multi-task learning, and extensive architecture comparison to maximize the harmonic mean of Q8 and Q3 prediction performance on variable-length protein sequences.

Multi-Task Facial Attribute Prediction | *PyTorch, PyTorch Lightning, Vision Transformers, CNNs* Nov 2025

- Developed a **multi-task computer vision pipeline** to jointly predict **age (regression)** and **gender (classification)** from facial images using shared feature representations and task-specific prediction heads.
- Implemented and benchmarked both **custom CNNs** and **fine-tuned Vision Transformer (ViT)** architectures, leveraging transfer learning to improve feature extraction and generalization across facial attributes.
- Built an end-to-end **PyTorch Lightning** training framework with advanced image augmentations, mixed-precision training, checkpointing, learning-rate scheduling, and reproducible inference for large-scale experimentation.
- Optimized the joint objective using classification and regression losses, evaluating performance with **Macro-F1** for gender prediction and normalized RMSE for age estimation, maximizing the competition's harmonic mean score.
- Integrated **TrackIO** for experiment tracking, model versioning, and hyperparameter optimization, enabling systematic comparison of CNN and transformer-based models for multi-task learning.

TECHNICAL SKILLS

Languages: Python, Java, C/C++, SQL (PostgreSQL), JavaScript

Frameworks & Backend: FastAPI, Flask, React, React Native, Vue.js

ML/GenAI: PyTorch, HuggingFace Transformers, LLMs, RAG, FAISS, ChromaDB, LangChain

Data & Analysis: Pandas, NumPy, Matplotlib, EDA, Feature Engineering

Systems & Infra: Docker, REST APIs, Vector Databases, Redis, API Design

Tools: Git, Shell Scripting

AWARDS & RECOGNITION

Exceptional Initiative & Learning Award | IIMA Ventures AI Summer Residency 2026

- Awarded for exceptional initiative, continuous learning, and impactful contributions during the AI Summer Residency.

Advanced Certificate in Machine Learning and Data Science (IIT Madras) 2026

- Completed advanced coursework in **Machine Learning, Deep Learning, Transformers, and Applied Data Science**.

Recognized as Female Topper - Diploma in Programming (IIT Madras) 2025

- Recognized for achieving the **highest CGPA** among **1400+** students.

Dover Foundation Scholarship (Global Recipient) 2024

- Selected as **1 of 30** scholars out of applicants worldwide.

Student of the Year Award (2x) 2019, 2021

- Recognized for consistent academic performance and overall excellence.

Post-Graduation in Abacus and Vedic Math 2015–2017

- Completed advanced training in speed calculation and mental arithmetic.

CO-CURRICULAR ACTIVITIES

Club Head | R.O.F.I.E.S (Robotics Club of IIIT Pune) Aug 2025 – Present

- Leading robotics activities, workshops, and technical training sessions.
- Mentoring teams for competitions and managing cross-functional projects.

Design Head | QuantNum (Mathematics Club of IIIT Pune) Aug 2025 – Present

- Created research posters, event designs, and interactive math content.

EXTRA-CURRICULAR ACTIVITIES

Volunteer | Robin Hood Army

June 2024 – Present

- Contributing to education, health, and food donation drives for underprivileged children.

Sports | Throwball & Basketball

2017–18

- Runner-up in Zonal Throwball; active basketball player.