Aryan Raj

+91-8287276911 | aryanraj2713@gmail.com | linkedin.com/in/aryanraj13 | github.com/aryanraj2713

Education

SRM Institute of Science and Technology Chennai, IN

B.Tech in Computer Science and Engineering with specialization in Software Engineering *May 2021 – May 2025*

CGPA: 8.4/10

Experience

Machine Learning Engineer Sep. 2025 – Present

SEOstack (WT Softech) *Remote*

- Built production-grade LLM orchestration agent with industry-standard architecture, reducing latency and failure rates by 20% for business workflow automation
- Designed and deployed serverless customer chatbots on AWS infrastructure, cutting operational costs by 15% while improving response accuracy and user engagement
- Delivered multiple POCs spanning LLM fine-tuning(RLHF, LoRA, QLoRA), agentic systems, model deployment, and containerized AI applications using Docker and Kubernetes

Machine Learning Engineering Intern Aug. 2024 – July 2025

HyperVerge *Bengaluru, IN*

- Developed and fine-tuned domain-specific LLM solutions for automated KYC and fraud detection workflows, ensuring compliance with financial regulatory requirements across enterprise clients
- Led benchmarking and evaluation of LLMs on 1M+ real-world data points, building scalable monitoring pipelines and achieving industry-standard FAR and FRR thresholds for identity verification
- Optimized and deployed computer vision and NLP models for production identity verification systems, implementing MLOps practices including CI/CD pipelines, model versioning, and continuous monitoring

Machine Learning Intern Jan. 2024 – Aug. 2024

EmendoAI *California, USA (Remote)*

- Engineered and deployed 10+ Generative AI microservices using AWS Lambda, API Gateway, and OpenTelemetry, delivering scalable client-facing applications with comprehensive observability
- Architected serverless backends for retrieval-augmented generation (RAG) workflows, optimizing for scalability and efficiency across diverse AI application domains
- Developed automated AI testing framework that reduced production bugs by 20% through systematic evaluation pipelines and quality assurance protocols

Research Intern Feb. 2023 – July 2023

Indian Institute of Technology, Madras (IIT M) *Chennai, IN*

- Designed and implemented ML-based anti-collision system for unmanned surface vehicles (USVs) in collaboration with Ocean Engineering department, improving detection accuracy by 28% over existing solutions
- Developed marine object detection, tracking, and localization pipeline using stereo vision cameras and state-of-the-art computer vision models for real-time maritime navigation
- Deployed edge-based solution on ESP32 IoT hardware and validated system performance through physical wave basin experiments and Unity-based virtual marine simulations

Projects

Open-KYC | *Next.js, OpenCV, TensorFlow, Tesseract, ShadCN, WebRTC* 2024

- Built end-to-end AI-powered KYC verification portal with facial authentication, Aadhaar/PAN card OCR using Tesseract, and real-time liveness detection via WebRTC video streams
- Implemented identity verification workflows using OpenCV and TensorFlow, achieving 98% accuracy in document validation and facial matching with industry-standard security protocols
- Secured 1st place at Standard Chartered Hackathon 2024 among 200+ teams for production-ready automation and scalable architecture supporting 10,000+ daily verifications

Educative.AI | *Python, FastAPI, React.js, TailwindCSS, LLMs, TensorFlow, OCR* 2024

- Developed AI-powered student assistance platform integrating OCR and 10+ fine-tuned open-source LLMs to process handwritten and blackboard notes with 95% text recognition accuracy
- Implemented FastAPI backend and React.js frontend with TailwindCSS, delivering features including automated MCQ generation, speech-to-text doubt resolution, and structured note organization
- Enabled automatic resource retrieval and summarization from 1000+ educational sources, enhancing accessibility and productivity for 500+ student users

OneMed | *Python, Next.js, JavaScript, LLMs, Pinecone, MongoDB, AWS, React Native* 2024

- Built full-stack AI-powered Electronic Health Record (EHR) platform with LLM-based medical summarization reducing documentation time by 60% and voice-to-text consultations with 92% transcription accuracy
- Developed React Native mobile application for real-time emergency data synchronization across 15+ hospitals, enabling instant access to critical patient conditions and medical history
- Implemented scalable vector search using Pinecone and MongoDB for intelligent retrieval across patient records with sub-second query response; ranked Top 10 at MozoHack 2024

PaperPilot | *React.js, scikit-learn, AWS, Vercel, Python, TensorFlow, JavaScript* 2023

- Designed personalized academic paper recommendation engine using TF-IDF vectorization and KNN-based algorithms, achieving 85% recommendation relevance score across diverse research domains
- Curated and preprocessed dataset of 5,000+ IEEE papers spanning 30+ research areas, implementing data cleaning pipelines and feature extraction for optimal model performance
- Integrated AWS SES for automated personalized email delivery system, distributing 1,000+ tailored research paper recommendations weekly based on user preferences and citation patterns

AI-RoadGuard | *React.js, CNN, Flask, Python, TensorFlow, OpenCV* 2022

- Engineered CNN-based real-time accident detection system achieving 90%+ accuracy with 50% reduction in false alerts compared to traditional rule-based solutions
- Developed Flask REST API backend and React.js dashboard for real-time emergency service alerting, reducing average response time by 40% through automated geolocation-based dispatch
- Recognized as Best Project in Open Innovation at MLH MesoHack 2022 among 1500+ submissions for practical impact and technical innovation in road safety automation

Technical Skills

Languages: Python, JavaScript, Java, C/C++, SQL (Postgres), HTML/CSS, R
Frameworks: React.js, Next.js, Node.js, Flask, FastAPI, TailwindCSS, Material-UI
Machine Learning: TensorFlow, PyTorch, Scikit-learn, OpenCV, LangChain, LlamaIndex, Transformers, HuggingFace
Developer Tools: Git, Docker, AWS, Azure, GCP, Terraform, CDK, Kubernetes, MongoDB, Pinecone, Redis
Libraries: NumPy, Pandas, Matplotlib, Tesseract OCR, WebRTC

Publications

Structured Relevance Assessment for Robust Retrieval-Augmented Language Models
ITC4SD 2025, Springer Nature — arXiv: 2507.21287

Achievements & Leadership

Hackathons & Awards

- 1st place at Standard Chartered Hackathon 2024 for OpenKYC, an AI-powered KYC verification solution
- 2nd place at Hack Nova 2024 with Educative.AI, selected to represent at Innverve 2024, AIT Pune
- Best Project Award in Open Innovation at MLH Meso Hack 2022 for AI-Roadguard accident detection system
- Top 10 finalist at MozoHack 2024 for OneMed, an AI-powered EHR platform for hospitals

Open Source & Technical Writing

- Contributed to Dify open-source project with 100,000+ GitHub stars
- Authored technical articles on AI and machine learning for DataX Journal on Medium platform

Research & Leadership

- Researcher at Next Tech Lab, Norman and McCarthy Labs (2021-2025), specializing in deep learning and computer vision applications
- Technical Director at Data Science Community SRM (2022-2023), organized DS Hack 2.0 and led technical workshops for 500+ students