ABHIJITH S

TVM, Kerala ◇ +919497514843
abhijithyess@gmail.com ◇ linkedin.com/in/abhijithyess

PROFESSIONAL SUMMARY

Lead AI/ML Engineer and technical lead building production-grade agentic AI systems across Generative AI, LLMs, predictive modeling, and intelligent decision systems. Currently owns full-stack architecture for a healthcare agentic AI platform. Experienced in research-to-production workflows, including evaluation, fine-tuning, RAG and tool-augmented LLM systems, and low-latency inference, with strong MLOps and Edge AI depth. Google TensorFlow Certified.

EDUCATION

Diploma in Programming and Data Science
Indian Institute of Technology (IIT) Madras

Bachelor of Technology – Engineering
Sree Chitra Thirunal College of Engineering

TECHNICAL SKILLS

Languages	Python, C++, Bash, Node.js, TypeScript
AI / ML	Supervised & Unsupervised Learning, DNN, Model Training and Evaluation, Feature Engineering, PyTorch, TensorFlow, Keras, Scikit-learn, XGBoost, ONNX Runtime, Hugging Face
GenAI & LLMs	LLMs, RAG, Vector Search, LangChain, LangGraph, CrewAI, Ollama, Fine-tuning (LoRA/PEFT), Prompt Engineering, MCP, A2A, Agentic Workflows, Supervisor-Agent Patterns, Agent Skills
Computer Vision & Edge	YOLO, OpenCV, Pose Estimation, Real-Time CV, OCR, CUDA, ROS2, NVIDIA Jetson, DeepStream, TensorRT
Backend & Frontend	FastAPI, REST API Design, React, Vite, Keycloak, SSE / WebSocket Streaming
Data & Cloud	PostgreSQL, MongoDB, Neo4j, Qdrant, Azure AKS, Azure AI Foundry, AWS, Vertex AI, Docker, Kubernetes, Helm, Istio, GitHub Actions, MLflow, Airflow

LEADERSHIP & CERTIFICATIONS

Google TensorFlow Certified (Credential)
Neo4j Certified Professional (Credential)
Claude Certified Architect – Professional (Credly)
Ambassador DeepLearning.AI — Microsoft Student Ambassador
Mentor – RaSoR (Ramanujan Society of Research)

PROFESSIONAL EXPERIENCE

Senior AI Engineer & Technical Lead May 2026 – Present
UST , India

Technical lead and full-stack owner of CarePilot, an agentic AI healthcare platform for a healthcare payor client, spanning database schema, FastAPI backend, React/TypeScript frontend, and Kubernetes infrastructure on Azure AKS. Designed a Supervisor Agent architecture with a shared tool registry and specialized sub-agents. Built a readmission risk framework combining LACE scoring with social determinants of health (SDOH). Leading architecture for a customer-extensible AI platform with reusable base framework and Agent Skills, so new customers can onboard without routing every change through the core dev team.

AI/ML Research Engineer May 2021 – Jul 2026
Neoito, India

Designed and deployed end-to-end AI/ML systems across Computer Vision, LLMs, Reinforcement Learning, Robotics, and Edge AI. Architected LLM-powered multi-agent systems using RAG, vector databases, and NLP pipelines, and built predictive models for trend, anomaly, and bottleneck analysis.

Embedded Systems Developer March 2019 – Jan 2021
Tensorbot Innovations Pvt. Ltd., India

Designed device-to-cloud communication pipelines using ROS, MQTT, and APIs for real-time monitoring and control, integrating sensors and actuators for reliable, low-latency data acquisition.

KEY PROJECTS

CarePilot – Agentic AI Healthcare Platform (UST Global)
Lead full-stack development of a Supervisor Agent system orchestrating specialized sub-agents through a shared tool registry. Designed deterministic autopilot workflows for reliable field extraction and agent routing. Deployed on Azure AKS with PostgreSQL, Keycloak authentication, and Istio service mesh.

Customer-Extensible AI Platform – Base Workable Framework (UST Global)
Leading functional architecture for a reusable UI, API, orchestration, and Helm-based deployment platform that customers can fork and extend, driven by Agent Skills generated from a single customer data schema.

Clinical Document Extraction – CMS-1500 / UB-04 (UST Global)
Built and optimized an LLM-based extraction service for claim forms, covering prompt cost optimization, accuracy validation across large document sets, and production pipeline debugging.

NORTIQ AI – AI-Powered Sales Intelligence & Coaching Platform(Try here)
Led a multi-tenant agentic system for GTM workflows with role-aware coaching agents, shared memory, and RAG over PostgreSQL (pgvector). Built an async FastAPI backend with Redis-backed memory and WebSocket streaming, deployed on AWS with Docker.

BodBox – NVIDIA Edge AI Fitness & Pose Estimation Platform(Try here)
Built a real-time pose estimation system on NVIDIA Jetson Nano. Converted PyTorch models to ONNX and TensorRT, integrated them into DeepStream pipelines, and reduced end-to-end latency through CUDA-level profiling.

CogniAI – Automated PDF Processing & Intelligent Document Platform (Try here)
Developed scalable document intelligence pipelines for automated data extraction, classification, and trend analysis over large PDF collections, implementing asynchronous ingestion and preprocessing workflows for high-volume uploads.

Skoop – AI-Driven Ingredient Monitoring System
Led the development of a real-time, automated order tracking and ingredient verification in high-throughput restaurant environments. Designed and trained YOLO-based object detection models in PyTorch and deployed them through NVIDIA DeepStream for high-throughput, low-latency inference.

Intelligent Robotics Platform – Edge AI & Autonomous Navigation
Developed AI-powered robotic systems using ROS and NVIDIA Jetson for real-time perception, localization, and autonomous navigation. Integrated LiDAR, depth cameras, and vision-based object detection to enable mapping, obstacle avoidance, and human-aware navigation. Optimized deep learning inference with TensorRT and CUDA for edge deployment while implementing cloud-connected monitoring, telemetry, and remote robot management.