

KUMAR S V

Seasoned Artificial Intelligence Developer with a robust background in Machine Learning. Aspiring to leverage extensive expertise in a challenging role within the Machine Learning and Artificial Intelligence and Gen AI domain, aiming to develop innovative AI solutions and foster innovation in intricate business landscapes

salashvijay@gmail.com

https://www.linkedin.com/in/vijay-kumar-a2a9b420/



9849333365



https://github.com/salashvijay



Profile Summary

Results-driven AI/ML Engineer with **9+ years in Data Science/AI** delivering production-grade solutions in healthcare and insurance. Specialized across **NLP, Document Intelligence, RAG, multi-agent LLM systems, microservice architecture, and Responsible AI**, consistently achieving **60–90% cycle-time reduction** and generating **>\$1M in business value**. Expert in **Python, FastAPI, Azure AI, and scalable MLOps pipelines** (monitoring, fairness, governance, retraining). Strong foundation with **5+ years in Business Intelligence**, enabling end-to-end data understanding and solution design. Proven ability to design, build, and deploy advanced AI systems and production microservices that solve complex business challenges and scale reliably.

Core Value Proposition

- End-to-end ownership: data → models → APIs → deployment → monitoring & fairness.
- Document Intelligence/NLP: OCR+ML+LLM pipelines, 95–98% extraction/classification accuracy.
- RAG systems: BM25 + vector retrieval with structured, multi-task LLM extraction and precise metadata filtering.
- Agentic AI: architected multi-source LLM-routed assistants (tool-calling + deterministic keyword fallback) with prompt-injection guardrails, self-learning knowledge bases, and correlation-ID tracing.
- Responsible AI: reduced demographic/geographic disparity 8.3% → <5% (audit-ready); calibration (ECE/Brier) & drift (PSI) monitoring.
- Performance: latency 5s → <1s; 60–66% throughput gains; 70% API cost reduction via caching.
- Leadership: architecture, iterative model improvement, mentoring, platform governance (Jira/AIRB).

Selected Impact

- Recon (Claims Reconciliation, incl. ORS & Macess): 87% auto accuracy → 87.5% manual time reduction → ~\$800K/yr savings; ORS pipeline 12K+ docs, 90% review-time reduction, 18% duplicates removed.
- AOR Detector: 98%+ accuracy; 70% manual-review reduction; now a live tool source inside ATS Copilot for deep AOR verification.
- ATS Copilot: expanded from a single-case chatbot into a 6-tool agentic assistant (case data, inventory, document metadata, SOP knowledge, SME knowledge, AOR verification).
- Case Setup RAG: 4-task parallel LLM extraction (member/provider/claims/IC classification) with BM25 page-relevance filtering.
- PDF Q&A: 12.5K docs, 0.6s cached latency, 91% answer accuracy, 150+ DAU.
- Fairness framework: max disparity 8.3% → 4.8% with 85.6% accuracy; 45+ audit reports.



Core Competencies

Technology Roadmap Development



Data Science & Analytics



Machine Learning, Artificial Intelligence & Programming Language



Project Management & Execution



AI Consultation & Implementation



Business Strategy & Execution



Stakeholder Management



Release Management



Delivery Excellence



Soft Skills



Collaborator

Communicator

Planner

Innovator

Intuitive



Education

23-24

Advanced Certification in Generative AI from up Grad

2019

Post Graduate Diploma in Machine Learning and AI from IIT Bangalore

2007

Bachelor of Technology (Electrical and Electronics Engineering) from SVEC, JNTU.

Languages: Adv: Python; SQL Moderate:UNIX shell scripting Basic: JS, HTML, CSS | **ML/NLP:** scikit-learn, PyTorch, TensorFlow, Keras, spaCy, Transformers (BERT/BioBERT), LSTM, ANN, XGBoost | **GenAI/RAG:** Azure OpenAI (GPT-4/3.5), Prompting, ChromaDB/FAISS, LangChain, AutoGen, vector semantics | **Document/OCR:** Azure Document Intelligence, PyMuPDF, Tesseract, EasyOCR, PaddleOCR, RapidFuzz, PDF.js | **APIs/Perf:** FastAPI/Flask, async, Redis caching, concurrency (ThreadPoolExecutor) | **Cloud/DevOps:** Azure (ML, OpenAI, Storage, Container Instances, Functions basics), Docker, IIS, App Insights, Mlflow | **Responsible AI:** Demographic parity, equalized odds, threshold optimization, automated reporting.

 Work Experience

Data scientist (AI ML & GenAI) | Optum Global Services, Hyderabad**Feb 2017 – Present | Team: M&R (2017–Till Date)**

ATS Copilot – Agentic Multi-Source Assistant: Evolved ATS Copilot from a single-case lookup chatbot into a multi-source agentic assistant, adding 5 tool sources (inventory lists, document metadata, SOP procedure knowledge, a self-learning SME knowledge base, and AOR document verification) alongside case data. Built an LLM tool-calling orchestrator (Azure OpenAI GPT-5) with deterministic keyword-fallback routing, prompt-injection/input-validation guardrails, BM25 retrieval, and correlation-ID tracing; deployed on IIS Server.

ATS Claims Reconciliation (Recon) : Led the end-to-end Recon system spanning issue-type classification (MNB/ANN), ORS scanned-document classification (fine-tuned BERT/LLMs), and Macess attachment prediction; packaged the platform as an installable Python library (reva-ml) integrating the ORS Stargate API with credentials pulled dynamically from the shared DB for automatic rotation.

AOR Detector: Built a standalone AOR (Appointment of Representative) document classification/verification system (ML + Azure Document Intelligence + LLM extraction) achieving 98%+ accuracy and a 70% cut in manual review; integrated it as a live tool source inside ATS Copilot for deep AOR verification (name/signature-date extraction, manual-review flagging).

Case Setup & Case Setup RAG: Designed the original Case Setup orchestration pipeline (AGIP/Non-AGIP/Binary MR classification & extraction with worker-locking and retry logic), then evolved it into Case Setup RAG – a 4-task parallel LLM extraction pipeline (member, provider, claims, IC classification) using BM25 page-relevance scoring and a mixed gpt-4.1/gpt-5 model strategy to cut cost and latency.

PDF Question Answering: Built a FastAPI RAG service (PyMuPDF text extraction, ChromaDB vector store + query cache, FlashRank reranking, Azure/OpenAI) for PDF document Q&A with page-cited answers.

SMART Business Partner Prediction (legacy): Predictive ANN model for healthcare partner selection, achieving 94% accuracy.

Claims Outcome Prediction (legacy): Deep learning models for claims process automation, achieving 95% accuracy.

AIRB & Governance: Secured AIRB approval across projects (SVI, age, gender fairness reports); led Jira platform upgrade and AI DOJO capstone (Healthcare Insurance Claim Approval Agent).

Software Engineer (Statistics, Business Intelligence) | Optum Global Services, Hyderabad**Nov 2013 – Jan 2017 | Team: 3RDD & CFR**

- Converted Informatica & DataStage ETL pipelines to Python-based ETL, improving processing speed and automation.
- Built and maintained Essbase cubes for financial reporting, budgeting, and forecasting; optimized cube performance & data pipelines.
- Conducted statistical analysis (R, SAS, Python) and data visualization (Excel, Matplotlib, Hyperion) for internal & external financial reporting and trend forecasting.

Associate & Software Engineer | Magna Infotech Pvt. Ltd., Hyderabad**Sep 2010 – Nov 2013 | Team: CFR :** Ensuring seamless financial reporting operations for Essbase cubes, Maintained Unix jobs & TWS workload scheduler,**Selected Achievements & Leadership**

- Automated multi-hour triage into minutes/seconds via ML + parallelization; 70% cost reduction via embedding/caching optimizations; trained 8 team members in deep learning model building.
- Personal initiative (design phase): prototyping a 4-service Healthcare AI microservices portfolio (claims triage, FWA detection, document intelligence, responsible AI governance) sharing a common installable Python package – exploratory future-scope R&D, not yet in production.

Target Roles

Data Scientist (AI/ML Engineer) · AI/ML Engineer · AI Developer · Applied NLP / Document Intelligence Engineer · Generative AI

Personal Published projects: <https://homeo-app.178.105.217.96.nip.io/>, <https://ayurvedic-app.178.105.217.96.nip.io/>.