PAVAN KUMAR P H V

EMAIL phvpavankumar@gmail.com
MOBILE +91- 9912552285
LINKEDIN https://www.linkedin.com/in/pavan-kumar-phv/
GITHUB https://github.com/phvpavankumar

LOCATION
Bangalore

PROFILE SUMMARY

- **AI/GenAI Engineer with 4 years of experience in architecting high-performance deep learning, computer vision, and generative AI systems**, specializing in edge-AI deployments, LLM workflows, and production-scale model optimization.
- **Expert in neural network architecture engineering, inference acceleration, and embedded-AI optimization**, leveraging TensorRT, ONNX, PEFT, and microcontroller-aware design to deliver real-time, low-latency intelligent systems.
- **Proficient in building end-to-end GenAI solutions**, including Stable Diffusion–based T2I models, LLM-powered analytics pipelines, and vector-database-driven retrieval systems using LangChain and custom prompt engineering frameworks.
- **Developed & optimized RAG pipelines and agentic AI systems** to deliver intelligent retrieval, contextual understanding, and autonomous decision-making workflows.
- **Developed and maintained end-to-end deployment pipelines for AI/ML and Generative AI solutions**, utilizing cloud-native services across **AWS, Azure, and GCP** to ensure scalable, production-ready model delivery.
- **Implemented robust model-serving architectures using Flask or Gunicorn APIs**, enabling seamless interaction with cloud platforms and supporting automated, reliable CI/CD workflows.
- **Delivered enterprise-grade AI solutions across retail, embedded systems, and industrial automation domains**, including shelf-vision intelligence, RFID ML-serving platforms, anomaly detection, and MCU-optimized neural network deployments.

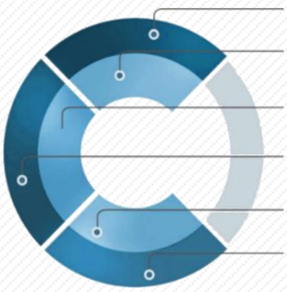
CORE COMPETENCIES

- Deep Learning Engineering
- Computer Vision Algorithms
- Generative AI (GenAI)
- Neural Network Optimization
- Edge AI / Embedded AI
- LLM Fine-Tuning & PEFT
- TensorRT Acceleration
- Model Deployment (GCP/AWS/Azure)
- CI/CD for ML Systems
- Vector Databases & LangChain
- Object Detection & 3D Reconstruction
- Real-Time Inference Pipelines

TECHNICAL SKILLS

- **Programming Languages:** Python, C++, MATLAB
- **Core Competencies:** Deep Learning, Machine Learning, Data Science, Computer Vision, Natural Language Processing (NLP), Generative AI, 3D Object Reconstruction/Registration, Image Registration, AI Model Development, CI/CD, Git
- **Computer Vision:** Image Processing, Object Detection, Semantic Segmentation, 3D Object Reconstruction/Registration
- **Natural Language Processing:** Large Language Models (Falcon, LLaMA2, GPT), BERT, LaBERT, MURIL, Prompt Engineering Techniques, Parameter-Efficient Fine-Tuning (PEFT), LangChain, Vector Databases
- **Model Deployment & Cloud:** GCP, Vertex AI, AWS, Azure, Docker, Kubernetes, Flask
- **Frameworks & Libraries:** TensorFlow, PyTorch, ONNX, MLflow, PyTest, OpenCV, Open3D, NLTK, Scikit-learn, NumPy, Pandas, Matplotlib, Seaborn

SOFT SKILLS



Team Management
Collaborator
Critical Thinking
Innovator
Leadership Qualities
Communicator

EDUCATION DETAILS

2022: Master of Technology (M.Tech.) Data Science Amrita School of Engineering, Coimbatore

2019: Bachelor of Technology (B.Tech.) Mechanical Engineering Amrita School of Engineering, Coimbatore

FREELANCE EXPERIENCE

Teaching Assistant Freelancer, Coimbatore, India **Aug 2020 – Dec 2021**

CERTIFICATIONS & PROFESSIONAL COURSES

- **Fundamentals of Digital Image and Video Processing** — Northwestern University, *2024*
- **Generative AI Mentorship Program** — Growth School, *2024*
- **ChatGPT 101-B26 Course** — Growth School, *2023*
- **Getting Started with CSS** — Codekaro, *2023*
- **Tableau for Data Science** — Scaler, *2022*
- **Machine Learning** — Stanford University, *2021*
- **Building Transformer-Based Natural Language Processing Applications** — NVIDIA, *2021*
- **Fundamentals of Deep Learning** — NVIDIA, *2021*
- **Introduction to Programming using JavaScript – Certified** — Microsoft, *2020*
- **Programming with Python** — Internshala, *2020*

WORK EXPERIENCES

Solum, Bangalore, India | Aug 2024 – Present
AI/CV & ML Systems Engineer

Key Result Areas:

- Designing and operationalizing advanced shelf-vision pipelines leveraging YOLOv7/YOLOv9, TensorRT, and LightGlue, ensuring high-precision calibration and robust end-to-end CI/CD for real-time product recognition.
- Delivering production-grade computer-vision systems optimized for low-latency inference, scalable deployment, and consistent accuracy across diverse retail shelf environments.
- Developing high-performance RFID ML-serving APIs, seamlessly integrating Event Hubs and MongoDB to enable hot-reload capabilities, enhanced observability, and fully containerized deployments.
- Spearheading resilient ML service architectures with monitoring, logging, and automated scalability, improving system reliability and reducing operational overhead.
- Building LLM-powered log-intelligence frameworks that transform unstructured logs into structured insights, significantly accelerating issue detection and root-cause analysis.
- Reducing manual investigation effort by implementing intelligent log-analytics pipelines, enabling proactive anomaly detection and data-driven operational decision-making.

Significant Highlights:

- Delivered **<200 ms response times** and **15–20 FPS inference** by refining sensor-driven session flows and GPU-optimized execution.
- Brought log-processing down to seconds by leveraging fuzzy clustering, ReAct-based LangChain agents, and on-device LLaMA models.
- Developed **performance-focused ML pipelines** that boosted real-time throughput, diagnostics, and system intelligence.

Ignitarium, Bangalore, India | Aug 2022 – July-2024
AI Engineer

Key Result Areas:

- Led architectural refinement of the FastestDet neural network to boost accuracy, efficiency, & support ultra-lightweight microcontroller deployment.
- Designed next-gen AI model development pipelines leveraging advanced optimization, training, and evaluation methodologies.
- Optimized end-to-end innovation in model performance tuning, latency reduction, and deployment readiness.
- Collaborated closely with multidisciplinary teams—including hardware, software, and research—to ensure smooth integration and execution of AI solutions.
- Oversaw project delivery with a focus on quality, scalability, and alignment with technical and business goals.

Sony, Bangalore, India | Feb 2023 - Dec 2023
AI Consultant

Key Result Areas:

- Built and optimized custom object detection and segmentation models for Sony's embedded devices, enhancing speed, accuracy, and on-device performance.
- Accelerated AI deployment using transfer-learning workflows, significantly reducing training time and integration effort.

INTERNSHIP

Ignitarium, Bangalore, India | AI Intern **Jun 2021 – Jul 2022**

KEY PROJECTS UNDERTAKEN

Solum

- Deployed multi-camera shelf-vision system (YOLOv7/YOLOv9 + TensorRT + LightGlue) with sub-200 ms, 15–20 FPS product detection and OOS analytics.
- Built hot-reload RFID ML APIs (Flask + Gunicorn) integrated with Event Hubs + MongoDB for real-time shelf mapping.
- Implemented LLM-driven log analytics & SME automation using LangChain ReAct and local LLaMA, streamlining 1000+ review cycles.

Ignitarium & Sony

Key Result Areas:

- Improved FastestDet accuracy by 35% and deployed high-performance object-detection models on Renesas MCUs and Sony IMX500 cameras for real-time edge AI.
- Designed RGB-D reconstruction and binocular vision pipelines enabling precise alignment and robust defect detection.
- Developed domain-specific LLM Q&A systems and custom neural networks, accelerating early defect detection across Railway, Telecom, and Solar industries.