# Uma Maheswara Rao Meleti maheshmeleti350@gmail.com

**+1 (864) 765 7139** | https://www.linkedin.com/in/mahesh-meleti | maheshmeleti.github.io/

## EXPERIENCE

**Machine Learning Engineer: In The Loop Inc.** **Feb 2025 – Present**

- Built a browser-based **Agentic** workflow for product price prediction using **Langchain**, **Azure OpenAI** and **Playwright**. Created tools and **prompt-engineered LLMs** for a 20x faster response compared to existing browser-based agents
- Deployed **REST** API on **Google Cloud Kubernetes Engine** (GKE) containerized with **Docker**. Automated **CI/CD** with **Jenkins** while **load balancing** and monitoring, ensuring 99.9% uptime under 50+ concurrent requests
- Finetuned **Vision Language Model** (Qwen2-VL) on Image QA for apparel damage detection using **PEFT** and **LORA**. Deployed model on **Vertex-AI** and optimized inference speed by ~38% using **DeepSpeed-FastGen** and **vllm**

**Lead Engineer: Aira Matrix** **Dec 2020 – Dec 2022**

- Engineered Machine learning-based solutions for life science applications, advancing **Cancer Research** and **Histopathology Workflows**. Worked on architectures like **MobileNet**, **EfficientNet**, **HoverNet**, **YOLO**, **Vision Transformers** etc.
- Created **parallel processing** and **imaging pipelines** for projects. Implemented a **Random Forest** classifier for staging the estrous cycle using **PySpark.** Partnered with pathologists to refine algorithms, improved **Kappa Score** by 30%
- Applied **Semantic Segmentation** and **Object Detection** algorithms for abnormality detection in Whole Slide Images. Devised an **active learning** framework for reducing the annotation time by over 60%. Collaborated with the **HPC** team for deployments
- Customized **Unet** architecture with **ResNet** and **attention blocks** for feature extraction and optimized the processing speed by 20%. Worked on augmentation techniques and **test-time adaptation** using methods like StainMixup, Meckanko, Vahadane etc.
- Implemented outlier detection using **VAE** and integrated monitoring tools using **Prometheus** and **Grafana**. Developed **Generative AI** solutions for addressing **data drift** and reduced the false positive rate by 10%

**Machine Learning Engineer/Data Scientist: Brane Enterprises** **Jul 2019 – Dec 2020**

- Created **computer vision** software for **automated** bank cheque detection using **OCR**, **image processing,** and **Google Cloud** APIs, achieved a 92% accuracy on 500+ cheques from 15 Indian banks
- Designed ML solution to convert workflow images into intuitive entities using **TensorFlow Faster RCNN**, and **zero-shot learning**. Utilized **S3** for storage, **RDS PostgreSQL** for structured transaction data, and **MongoDB** for unstructured metadata
- Attained 80% conversion accuracy on 1500+ documents. Enhanced extraction accuracy by integrating OCR (**Tesseract**, **Google Vision**) and NLP (**spaCy NER**, **rule-based matching**) for identifying and parsing entities and attributes.
- Containerized **Django** REST APIs and ML models using **Docker**, orchestrated with **AWS ECS**, and deployed on **EC2** with **auto-scaling** for high-traffic workloads

**Machine Learning Intern: Musco Sports Lighting** **May 2024 – Aug 2024**

- Constructed adaptive calibration pipeline for **stereo cameras** to work in varying weather conditions (low light, bright sun etc). Accomplished a 93% detection accuracy on the custom-prepared dataset
- Deployed **real-time** ball tracking and **3D pose estimation** model on **Edge Devices** (**Jetson Xavier**) using **YOLOv8** and optimized latency with **Tensor-RT** and Polynomial Interpolation
- Performed **statistical analysis** on **depth estimation** using computer vision algorithms, recommending insights and performance limitations of the product. Proposed **lidar-based** methods for product validation

**Graduate Researcher: Clemson University** **Jan 2023 – Dec 2024**

- **Researched** wildfire management, developing a **drone-based** approach for detecting obscured fires. Improved detection rate by 28% through **experimentation** and utilized **MLflow** for tracking. Published work in a **peer-reviewed** conference
- Proposed a novel **video-transformer** for wildfire localization. Implemented a new **attention** mechanism surpassing SOTA methods by 5% F1 score. Trained models on **AWS SageMaker** using **DDP** and **PyTorch Lightning**
- Developed a pricing prediction model for hotel room reservations, leveraging **exploratory data analysis** (EDA) and **regression** techniques, and achieved an R2 value of 0.33
- Designed data ingestion pipelines with **GCP** and **PostgreSQL** and implemented **ETL** workflows with **Astro Airflow**. Built a high-performing feature store using **Redis** for data retrieval and model training, reducing retrieval time by 90%

## Publications

- Obscured Wildfire Flame Detection by Spatio-temporal Analysis of Smoke Patterns Using Frame-wise Transformers. https://doi.ieeecomputersociety.org/10.1109/DCOSS-IoT61029.2024.00019

## SKILLS & Certifications

- **Areas of Expertise**: Machine Learning, Deep Learning, Computer Vision, NLP, MLOps, Generative-AI, Agentic-AI
- **Programming**: Python, R, Matlab, Java, C++, bash, SQL, NoSQL, PostgreSQL
- **Tools and Systems**: MLFlow, DVC, Docker, Kubernetes, Airflow, Linux, ONNX, PySpark, Redis, SageMaker, VetexAI, AWS, GCP, Azure, Jenkins, Github Actions, CircleCI
- **Libraries and Frameworks**: Pytorch, Tensorflow, LangChain, LangGraph, Numpy, Pandas, Scikit-Learn, Matplotlib
- **Certifications**: AWS Certified Solutions Architect, MLOps Specialization, LLM Engineering

## Education

Master's in Computer Science, **Clemson University (3.93/4.0)** **Jan 2023 – Dec 2024**

Bachelor's in Electronics and Communication Engineering, **RGUKT IIIT (8.61/10.00)** **Jun 2015– Apr 2019**

## Projects

**Agentic RAG for Document Analysis with Self-Evaluation | Github**

- **LangGraph**-based agent responds to queries using **Retrieval-Augmented Generation** (**FAISS** database) and a self-evaluation feedback loop. Grades document retrieval relevance, rewrites queries, and generates answers or quizzes with **on-device Ollama LLM** integration. Containerized with **Docker** for deployment and includes a **Streamlit** frontend for interactive user testing.

**Text to SQL Generation | Github**

- Fine-tuned **Llama-7b** using **Parameter-Efficient Fine-Tuning** (PEFT) to generate **SQL queries** from natural language questions. Trained model using SQL-create-context dataset on **NVIDIA A100 GPUs** using **Quantized Low Rank Adaptation** (QloRA). Evaluated the model over the base model and showed a significant improvement in SQL generation.