Datta Desai Email: dattatreya.mdesai1996@gmail.com , Phone: +491778531338

Software - AI / ML Engineer

LinkedIn | GitHub

SUMMARY

AI/ML Engineer with deep learning experience across NLP and computer vision domains, specializing in building Transformer-based pipelines using PyTorch. Strong mathematical background through projects in signal processing, supervised learning, and biomedical data analysis. Proficient in Python, C++, and JavaScript, with hands-on experience in deploying real-time inference APIs and AI-backed applications in healthcare and research environments. Currently improving German (B1) with the goal of reaching C1. Passionate about using AI in production, sensor systems, and real-world engineering challenges.

STRONG SKILLS

Programming: Python (expert), SQL, Bash, C++
ML & AI Frameworks: PyTorch, Transformers (HuggingFace), LoRA, PEFT, Scikit-learn, TorchVision, (basic TensorFlow/Keras)
Databases: PostgreSQL, SQLAlchemy
DevOps & Deployment: Docker, GitHub Actions, Jenkins, basic AWS (EC2, Lambda, S3), Azure
AI Concepts: Supervised learning, OCR, NLP (prompt engineering, scene captioning), feature extraction, VLMs
Math/Engineering Tools: Signal processing, ECG/EMG/TCR modeling, NumPy, Matplotlib, CUDA, Jupyter
Collaboration Tools: Jira, Confluence, Agile/Scrum
Languages: English (C2), German (B1 – actively progressing to C1)
Soft skills: Proven ability to communicate technical solutions clearly across medical and engineering stakeholders, Self-motivated, creative, and adaptable within fast-paced interdisciplinary environments

Interdisciplinary AI Experience:

- Applied deep learning to clinical datasets (DICOM, HL7, signal waveforms) for diagnostics and workflow understanding
- Built modular Python-based pipelines for inference in real-world constrained environments
- Strong interest in transferring these skills to engineering domains such as mechanical diagnostics, robotics, and automation systems

WORK EXPERIENCE

Master Thesis at AIBE Lab, FAU Erlangen collaboration with ZEISS **December 2024 – CURRENT**

- **ML Engineer – Surgical Workflow Understanding using VLMs & LLMs (LangChain + GPT-4)**
- **Automated Surgical Workflow Understanding** – Finetuned Vision Language Models (VLMs) such as LLaVA, LLaVA-Med, and QWEN for phase recognition and description generation in ophthalmological surgeries, focusing on cataract procedures.
- **Dataset Curation & Annotation** – Preprocessed and annotated cataract surgery videos to define surgical phases, anatomical structures and instruments. Applied data augmentation techniques, and structured datasets for training and evaluation of VLMs.
- **Fine-tuned VLMs for Surgical Scene Segmentation** – Trained models to recognize and describe surgical phases, integrating both visual and textual components to enhance procedural understanding and real-time decision support.
- Developing a **PyTorch-based MultiSourceCaptionDataset** to efficiently load **preprocessed .pt image tensors** and their respective **tokenized captions** from JSON annotations.
- **Python Libraries & Computational Frameworks:** Applied **LoRA** and **PEFT**-based finetuning methods for efficient model adaptation. Using CUDA-enabled pipelines for model optimization. The implementation utilizes **PyTorch** for deep learning, **TorchVision** for image transformations, **Scikit-learn** for K-fold cross-validation and evaluation, **Cython** and **NumPy** for optimized tensor operations, **JSON and CSV handling** for structured annotation processing, and **CUDA** for GPU-accelerated model training.
- Developed real-time AI pipeline using **Transformer-based VLMs** and **LoRA-optimized GPT-4 prompting**, enabling contextual phase transitions in surgical videos.
- Translated model outputs into clinical insights through structured **API interfaces** and integrated them into an experimental medical workflow. Explored performance/efficiency trade-offs between LLaVA and QWEN models, contributing to internal AI benchmarking.

Python Engineer (Working student) at Siemens Healthineers GmbH, Erlangen **November 2022 – CURRENT**

- **Automated report generation pipelines using Python and ETL workflows for patient analytics, improving data latency by 40%.**
- **Created a REST API in Flask** to enable seamless integration between Sensis Vibe and external healthcare platforms, improving interoperability by 35% and reducing manual data entry errors by 20%.
- **Developed and deployed data-intensive back-end systems in** Python**, handling** structured patient data **through DICOM/HL7 pipelines**
- Built and containerized RESTful APIs using Flask; improved system interoperability and data exchange by 35%
- Managed PostgreSQL databases and optimized queries for real-time analytics dashboards (Dash/Plotly)
- Maintained and deployed services via Docker with CI pipelines using GitHub Actions
- Supported internal test automation and code quality validation, enabling faster deployment and QA alignment

- Collaborated across data science, QA, and product teams to align features with medical compliance guidelines
- Built and packaged Transformer-based models into modular inference engines with FastAPI and PyTorch, enabling seamless integration into data processing pipelines.

SOFTWARE DEVELOPER III at Cognizant Technology Solutions, Bangalore, India **May 2021 – May 2022**

- Collaborated as **Full-stack Developer** for **AMGEN Healthcare Corp. USA** CoE engagement.
- Developed a **User responsive application** for entering the results of Microbiological Experiments into database, thereby improving the efficiency of experiments by 15%.
- Configured and maintained **CI/CD pipelines** using **Jenkins** and **GitHub** Actions to automate the **build**, **test**, and **deployment** processes, which **increased deployment** frequency by 50% and **reduced integration issues**.
- Actively used **Jira for sprint planning**, **tracking bugs**, and managing user stories, which led to a **20% improvement** in meeting **sprint goals** and deadlines due to better task prioritization and **resource allocation**.

EDUCATION

Master of Science in Medical Engineering **April 2022 – CURRENT**
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany **Grade: 2.0**

Bachelor of Engineering in Medical Electronics **July 2017 – June 2020**
Visvesvaraya Technological University, Karnataka, India **Grade: 1.8**

Relevant Courses: Pattern Recognition and Analysis, Machine Learning, CNN, RNN, Signal Analysis, Natural Language Processing, Machine Learning in Time Series, Advanced C++, Interfacing the Neuromuscular System, AI in Medical Robotics, Human Computer Interaction, Project Management (Agile, Scrum, Waterfall), CI/CD

Projects:

1. Comparison of Simpson's Diversity Indices of TCR samples with and without down-sampling (Python)

- Simulated **T cell clonal dynamics** using a **stochastic birth-death model with logistic growth dynamics**, generating datasets comprising over **1 million clonal events** for **diversity analysis**. Implemented the **Gillespie stochastic algorithm** to model **temporal evolution**, achieving biologically realistic simulations of **contracting, persistent, and late-emerging clones**.
- Developed and **validated down sampling and normalization strategies** to mitigate bias in TCR sequencing data, reducing sample variability effects by 30% and ensuring reliable diversity indices. Applied the **delete-one jack-knife method to assess diversity stability**, revealing a 10% variance reduction in diversity indices after normalization.
- Analysed **Simpson's Diversity Index and Morisita-Horn similarity** metrics to evaluate clonal behaviour across multiple time points, achieving a 25% improvement in identifying key clonal trends.
- Created **predictive visualizations** for **clonal population trends**, enabling the identification of stable and transient clonal dynamics across 200+ simulation scenarios. Contributed insights into immunological diversity and **clonal persistence**, laying the groundwork for advanced statistical frameworks and machine learning applications in TCR repertoire studies.

2. Analysis of muscle activation based on the complex EMG signal analysis

- **Preprocessed an 8x8 channel surface EMG dataset** using advanced Python libraries, achieving a 30% reduction in noise artifacts and ensuring data integrity for subsequent analysis.
- **Analysed muscle size and activation patterns** based on EMG signals, quantifying activation levels with a 95% accuracy rate compared to clinical benchmarks.
- **Developed Python-based algorithms** to assess muscle condition through activation metrics, identifying early signs of muscular fatigue with a detection precision of 90%.
- **Enhanced signal processing efficiency** by 25% through optimized filtering techniques, reducing computation time for large datasets and enabling real-time analysis capabilities.
- **Generated insightful visualizations** of activation trends and muscle size variations using Matplotlib and Seaborn, improving the interpretability of findings for medical professionals by 40%.

CERTIFICATIONS

Certified Angular Developer (Udemy), Certified C# Developer (Udemy), ASP .NET Core Web API Designer, Python Certified Entry – Level Programmer, Certified Relational Database Designer (Udemy), Certified API Developer (Udemy), Microsoft Excel from Beginner to Advanced, Machine Learning A-Z,