Sree Dhyuti Nimmagadda

AI Researcher | Machine Learning Engineer | Applied Scientist
sreedhyutin@gmail.com | +1 7738580329 | linkedin.com/in/dhyutin | github.com/dhyutin

EDUCATION

Northwestern University | Evanston, IL Sept 2024 - Dec 2025
Master's, Artificial Intelligence. CGPA: 4.0/4.0.
Focus: NLP, Deep Learning, GenAI, Machine Learning, Graph Neural Networks, Data Science, High Performance Computing.

Indian Institute of Information Technology, Kancheepuram | Chennai, India Jul 2019 - May 2024
M.Tech + B.Tech, Computer Science Engineering. CGPA: 8.86/10.00.
Focus: Big Data Analytics, Digital Image Processing, Computer Graphics, Pattern Recognition, Computer Vision.
Achievements:

- Presented my final year thesis on *Improved Text-Summarization with PEGASUS and Siamese Network Evaluation*" as a paper at IEEE TENCON 2024 (Singapore).
- Ranked 37th globally out of 1,500 teams in the OpenCV AI Competition 2022.

SKILLS

- **Languages/Platforms:** Python, C, C++, R, Matlab, SQL, TypeScript, Verilog, NASM, Microsoft Azure, Google Cloud.
- **Frameworks**: TensorFlow, PyTorch, Keras, Scikit-learn, XGBoost, LightGBM, nltk, LangChain, OpenCV, PySpark.
- **Statistical Analysis:** Hypothesis Testing, Regression, Time Series Analysis, Predictive Modeling, A/B Testing.
- **APIs:** OpenAI GPT-4, Azure OpenAI, HuggingFace Transformers, AWS SDKs, FastAPI.
- **Tools/Optimization:** CUDA, OpenMP, MPI, NVIDIA DeepStream, Docker, Git, CI/CD, Tableau, Figma.

EXPERIENCE

AI Research Intern (Writer | San Francisco, California, USA) Jun 2025 - Sept 2025

- Investigated novel attention mechanism variants to enhance long-context reasoning and model interpretability in LLMs.
- Designed & implemented a vLLM based evaluation framework with different datasets over 7 unique tasks for long-context modeling for the AI Research team's internal Palmyra models evaluations
- Identified and addressed logical and coding discrepancies in Princeton's HELMET evaluation framework (ICLR 2025).

AI Researcher (The Abazeed Lab | Chicago, Illinois, USA) Mar 2025 - Jun 2025

- Built and deployed a DynUNet-based 3D segmentation model for 117 organs-at-risk (OARs) with 94% cross-validation Dice accuracy, now integrated into the radiation oncology treatment planning workflow at Abazeed Lab.
- Engineered a GPU-accelerated, distributed data loading pipeline using PyTorch Lightning that cut down the data ingestion time and model training time from ~3 hours to 8 minutes (x20 times faster) in low-resource environments.
- Designed a scalable CT scan preprocessing pipeline using cube-based volume chunking and conducted experiments with transformer-based models (Swin UNETR, ViT) for performance benchmarking.

Machine Learning Research Intern (BioSystems & Controls Lab | Chennai, India) May 2023 - Oct 2023

- Developed a regression model for predicting apple-sugar levels using Near InfraRed Spectrum, attaining a 0.4 R^2-score.
- Built a semi-supervised autoencoder regression model for real-time *Lactococcus lactis* bacteria fermentation monitoring, improving validation R^2-score from 0.61 to 0.89 (+46%), and prediction R^2-score from 0.49 to 0.82 (+67%).
- Performed multivariate T^2 analysis, uncovering data inconsistencies and refining experimental data curation setup.
- Collaborated with cross-functional biotechnology teammates to integrate regressor to a hardware architecture.

Machine Learning Intern (Tiny Banyan Technologies Pvt Ltd | Chennai, India) Aug 2022 - Dec 2022

- Deployed YOLOv5 model for real-time detecting potholes and cracks, revamping anomaly detection accuracy to 99%.
- Trained a team of interns to manage GCP for model training, hyperparameter tuning, testing & validation of ML models.

PROJECTS

REINFORCE Algorithm based Contradiction Correction in LLMs Generations
Built a BERT-based contradiction detection model fine-tuned with LoRA and improved by a REINFORCE RL framework, achieving ~90% detection accuracy and reduction in contradiction rate from 65% to 44%.

Dialogue Summarizer
Fine-tuned PEGASUS LLM on SAMSUM corpus and deployed a FastAPI- and Docker-based summarization app with AWS and CI/CD integration, streamlining workflows and automated testing.

AI Powered Hotel Recommendation System Using Agents and RAGs
Enhanced natural language understanding in GPT-4o by fine-tuning retrieval mechanisms to reduce hallucinations and improve query disambiguation, and optimized hotel search performance using agents and RAGs, delivering real-time, preference-aligned hotel recommendations with higher precision and relevance.