SIDDHESH DOSI

Researcher & former Postgraduate Student,
Department of Computer Science and Engineering, IIT Gandhinagar

Portfolio
github.com/git-siddhesh
dosisiddhesh@iitgn.ac.in

Education

Degree	Institution	CPI/%	Year
MTech	IIT Gandhinagar	9.33	2022-2024
BTech	SKIT MG, Jaipur, RTU	79.6%	2017-2021
Class XII	Govt. Nutan School Banswara, RBSE	81.20%	2016-2017
Class X	Springdales School, RBSE	82.17%	2014-2015

Projects

- **Point Cloud completion problem** (Prof Shanmuganathan Raman, IITGN) **[July 2024-Present]**
  o Experimental Training of a combination of X-LSTM (encoder) and PCN (decoder) to estimate the complete geometry of an object from partial point cloud observation.

- **MINI-HINDI LLM for Domain Specific Knowledge Extraction: GANGA-1B** (Prof Mayank Singh, IITGN) **[Dec 2023-Present]**
  o **Customized LLM Pretraining**: Designed a mini **LLM** on optimized mistral architecture on *domain-specific* data sources to build **a compact**, yet *powerful*, *domain-focused* **language model.** [Mtech Thesis]
  o Trained multiple mini-llms (in size, and architecture) on our 70 GB of cleaned and processed hindi dataset on 8 A100 GPUS.
  o **Enhanced Domain Efficiency**: Tailoring and optimizing LLMs for precision and efficiency in the domain contexts, ensuring high performance akin to larger models in accuracy with reduced compute requirements.
  o Links: Indiaai.gov.in, pib.gov.in HF-MODEL: *LingoIITGN/ganga-1b*
  o **FERTILITY-SCORE:** 1.12, **PPL:** 17.92 [Hugging Face, PyTorch, Mistral architecture]

- **Machine Learning Algorithm and their Visualizers from Scratch**, (Prof. Nipun Batra, IITGN) **[Jan-Apr 2023]**
  o Implemented **core ML algorithms**, including **Decision Trees** (using entropy and Gini Index), **Bagging, Boosting,** and **Linear Regression** using Gradient Descent without ML libraries like scikit-learn, to deepen algorithmic understanding
  o *Designed* **Movie recommendation** system with *K-Nearest Neighbor* algorithm on crowd sourced data.
    github.com/git-siddhesh/ML_algo_visualizer [Python, Matplotlib, Numpy]

- **Automated Code Compilation Evaluation System (ACCES)** **[June 2020]**
  o Build a system to *RUN, VERIFY* and *GRADE* programming lab assignments for the institute.
  o Reduced manual grading time by 50% through automated code checks and verification processes.
    [**Python** scripting, **RegEx**, and **Linux** commands].

Internships/Experiences

- **Machine Learning/AI Engineer, Developed two core products based on HYDE and RAG, Neuroreef** **[Oct 2023-Nov 2024]**
  o **Developed Advanced Chat-Engine**: Engineered and optimized the evidence-based model using HyDE on 2k PubMed dataset.
  o **AI-Copilot for Resident Doctors**: Co-developed the tool and apis for the medical professionals outperforming all other products.
  o Implemented core NLU architecture using RAG, HyDE, prompt tuning and fine tuned LLMs on huge pubmed data, with real time *Insights, Audio transcript, Analyze, Medical final notes,* and *Medical coding*.
  o Led a team of 5+ to design, develop, and deploy client-specific feature APIs ensuring seamless integration and high performance across varied healthcare environments.
  Tech stack: [Langchain, Tools and Agent, RAG, LLM fine tuning, Asyncio, FastAPI, AWS botocore]

- **Machine Learning Intern, Train MACH models using NeuralDB for extreme classification, ThirdAI Corp** **[June-July 2023]**
  o Led **MACH** model training using NLP based Neural DB on Amazon-3M data, employing **BOLT** CPU engine for optimization.
  o Trained large **MACH extreme classification models** on Stack overflow data and food UDT data.
  o Developed a deeper knowledge of ThirdAI's BOLT and **Neural Network Database** model training on Amazon-3M.
  o Delivered a high performance model on Amazon-3M with Metrics (P@1=0.695, R@5=0.797, R@10=0.814) in its category.

- **SDE Intern Implement Golang Handlers and T-SQL, InTimeTec Visionsoft** **[Mar-Aug 2021]**
  o Worked on Golang web application, leveraging Go routines for concurrency and Go Handlers for request handling, optimizing performance and responsiveness.
  o Acquired understandings in T-SQL through hands-on implementation of key concepts within SQL Server for enhanced data management.

- **Summer Intern, Hand Gesture Recognition tool, AdHoc Networks** **[June-July 2019]**
  o Created a Gesture Recognition Tool with OpenCV and image processing enabling the translation of hand gestures' specific patterns and motion into system commands like scrolling, media search, voice search activation *[python and bash scripting]*.

Hackathon and Competitions

- **Hackathon, TEGD: The Third AI Search Engine for Google Drive, ThirdAI Corp** (*First Prize*) **[Aug 2023]**
  o Developed a **Google Drive** *within-file* **search engine,** providing confidential **retrieval of files** led to a hackathon **winning project.**
  o Leveraging **NLP and NLU** based **ThirdAI's** *NeuralDB* architecture significantly enhances user's *accessibility* and *file retrieval* efficiency with *privacy.* Tech-Stack: [NeuralDB, BOLT, python, pytorch, HPC]

- **Supply-Bot, [E-Yantra]() Robotic Competition-2019, IIT Bombay** *(Runner-up)*   **[Aug-Feb 2020]**
  - o **Designed** and **programmed** *a line-following robot* with an **Arduino** controller having a centralized **satellite camera** for **live navigation**.
  - o Pre-processing live image, (with **ArUco** detection), for real-time analysis of surroundings for navigation and tasks.
    Tech-Stack: [OpenCv, C++, assembly]

Achievements

- 1st Prize **among** 70**+ teams, ThirdAI India Hackathon.Used** *Neural-DB to in-depth google drive* **search for text, audio data**   **[Aug 2023]**
- Secured **AIR 1012 (**out of **100k** participants) with a **score of 640** in **GATE** – 2022.   **[2022]**
- **Runner-up** in **Smart India Hackathon** 2020 (**Intra-college**) (by M**i**nistry o**f** E**du.** and **AICTE**)**.**   **[2020]**
- **Semi-Finalist** out of **8585** teams nationwide in **E-Yantra** Robotics Competition (**eYRC**) by **IIT-Bombay** (supported by **MHRD)**   **[2019]**
- Honored with Highest **Academic Award** for excellence in the **Department** of IT during undergraduate.   **[2019]**

Skill Summary

- *Languages*:   **Python** (Scripting and automation), C, C++, SQL, Golang
- *Tools*:   **Langchain**, **Streamlit**, Notion, Git, **Arduino,** Autodesk(basic) **NLTK, Hugging Face, Transformers, BERT, GPT**
- *General*:   Machine Learning, NLP, **Large Language Models** (LLMs), Deeplearning, Image Processing
- *Skills at a Glance***:** RegExp, FastAPI, Latex, SQL, Bash scripting(basic), Arduino, LFRs

Positions of Responsibility

- **Teaching Assistant, IIT-Gandhinagar**   **[Aug 2022-Present]**
  [Computing In Python, Engineering Mathematics, Engineering Graphics]
  - o Supervised a cohort of 20+ students from a pool of 300, leading respective lab sessions, administered viva-voce assessments for **quizzes** and one to one mentorship**.**
  - o Used the **Autodesk** inventor to model 3D objects in the Eng. Graphics lab of **60+** students
- **Course Instructor and Coordinator, Swami Keshvanand Institute of Technology**   **[May-July 2020]**
  - o **Science and Technology club:** Initiated and led co-curricular activities (weekly Python and DSA classes), as the Course Instructor**.**
  - o **Voluntarily organized** a 30-day programming **virtual bootcamp** during **covid-lockdown** to facilitate fundamental learning for students, significantly improving the class performance
- **Organizer of TechFest-2018 North-Zone, IIT-Bombay**   **[2018]**