SIDDHARTH

Delhi, India
siddharth899@gmail.com linkedin.com/in/Siddharth899 github.com/Siddharth7113 siddharth7113.github.io

Education

Cluster Innovation Center **Delhi, India**
B.E. in Information Technology & Mathematics *Nov 2022 – Present*

- **Coursework:** Calculus, Discrete Mathematics, Algorithms, Operating System, Linear Algebra, Java, Statistics, Computer System Architecture, In-Silico Biology, Database Systems, Data Communication & Networks, Software Engineering, Linear Programming & Game Theory, Genomics, Internet of Things, Security & Machine Learning, Numerical Analysis

Indian Institute of Technology Madras **Online Degree Program**
B.S. in Data Science *Sep 2022 – Present*

- **Coursework:** Computational Thinking, Mathematics for Machine Learning, Descriptive Statistics, Machine Learning Foundation, Machine Learning Techniques, Python, Business Data Management, Tools in Data Science, Machine Learning Practice, Business Analytics, DBMS & Data Structures & Algorithms, Modern Application Development, Java

Professional Experience

European Summer of Code (EsoC) & Ecospecs **Remote**
Research Fellow (AI for Drug Discovery) *Jul 2025 – Present*

- Designing *in silico* aptamers for drug development using AI methods in collaboration with the German Center of Open Source and Ecospecs.
- Building data/ML pipelines for sequence design and evaluation; coordinating with mentors across organizations.

Google Summer of Code - OpenClimateFix **Remote**
Machine Learning Intern *Jun 2025 – Sep 2025*

- Training an AI model on public datasets to forecast solar generation across the UK, with planned extension to global coverage.
- Contributing forecasting tooling and data pipelines to OpenClimateFix's open-source ecosystem.

Everstream Analytics **Remote**
Data Science Intern , Solar Forecasting *May 2025 – Jul 2025*

- Built a continental United States solar-generation simulation pipeline producing day-ahead forecasts in <5 minutes.
- Achieved overall error rate $<30\%$ and delivered regional forecasts for grid operators and energy estimators.
- Extended modeling to regional operators; obtained accuracy $>95\%$ in select regions.

FSIL, Georgia Tech **Remote**
Volunteer Research Assistant *May 2024 – Present*

- Working on financial LLM projects, including data collection, preprocessing, and model implementation.
- Initiated a project to evaluate instruction-following ability of LLMs for finance; developing evaluation protocols and exploring mitigation strategies for failure modes.

Student Internship Scheme, Cluster Innovation Center **Delhi, India**
Intern *Feb 2024 – Oct 2024*

- Worked on Micro-Quadcopter drones for autonomous flights.
- Designed and implemented control algorithms for stabilization and navigation.
- Integrated IMUs and Wi-Fi communication for robust real-world performance.

Projects

Genomics Foundation Model

- Designed and implemented transformer-based DNA language models using the ModernBERT architecture, inspired by DNABERT.
- Constructed preprocessing pipelines for large-scale genomic data, including k-mer tokenization and balanced sequence sampling.
- Executed 30K+ optimization steps on high-performance compute resources, generating embeddings for benchmarking.
- Conducted evaluations against baseline models, analyzing embedding quality, phylogenetic signal, and downstream classification potential.
- Identified limitations of generic architectures on biological data, motivating exploration of genomics-specific model designs.

10-K Filing Analysis using LLM

- Engineered a platform to download and analyze 10-K filings of companies, leveraging a fine-tuned Mixtral-7B LLM for natural language processing.
- Implemented deep learning pipelines using PyTorch to extract and summarize critical financial insights for decision-making.
- Developed the user interface using Streamlit for seamless interaction and deployed the platform using OpenRouter API.

Diagnosing Depression using Machine Learning Algorithms

- Developed a robust pipeline for classifying patients into depressed and non-depressed groups based on Burns Depression Checklist (BDC) scores.
- Implemented advanced feature selection techniques like SelectKBest, mRMR, and Boruta to enhance the relevance and quality of input features.
- Applied machine learning algorithms such as KNN, XGBoost, Gradient Boosting, AdaBoost, and Random Forest, achieving high classification accuracy.
- Compared multiple models and optimized hyperparameters using automated tools like Optuna, ensuring state-of-the-art performance in tabular data classification tasks.
- Evaluated model performance with metrics like accuracy, precision, recall, F1-score, and ROC-AUC, while employing SHAP for interpretability.

Publications

Glenn Matlin, **Siddharth**, Anirudh JM, Aditya Shukla, Yahya Hassan, Sudheer Chava. Instruction Following for Finance: Verifying Language Models' Ability to Follow Complex Financial Instructions. *NeurIPS 2025 Workshop on GenAI in Finance*.

Technical Skills

Python, PyTorch, TensorFlow, Hugging Face Transformers, Scikit-Learn, SQL, C, Java, Embedded C, Git, Linux.

Deep Learning: Transformers (BERT, GPT, Mixtral), CNNs, RNNs, Reinforcement Learning.

Applications: NLP, AI Agents, Autonomous Navigation, Time-Series Analysis.