

# Vyom Pathak

1212 Chianti Ct, Pleasanton, CA 94566, USA

+1(925)6605503 | [angerstick3@gmail.com](mailto:angerstick3@gmail.com) | [linkedin.com/in/01-vyom](https://www.linkedin.com/in/01-vyom) | [01-vyom.github.io](https://01-vyom.github.io) | [github.com/01-vyom](https://github.com/01-vyom)

## EXPERIENCE

### Data Science Research Lab at UF, *Research Assistant*

*Learned Distance Sensitive Bloom Filter for High-Dimensional Similarity Join*

December 2022 - May 2023

- Orchestrated several experimental designs for training SelNet, Mixture of experts (MoE), XGB, LightGBM, and Support Vector Regressor on 3 text based and 3 image based embedding datasets using distributed training framework
- Composed the introduction, related work, and experimental details sections of the research paper for baseline experiments using Latex

### Amazon, *Applied Science Intern*

August 2022 - December 2022

*Alexa Smart Home Team*

- Removed redundant features by performing feature importance analysis using **visualization techniques, and model ablation study**
- Adopted a novel **self-attention** based architecture for **multi-variate time-series classification task** to model user behaviors
- Interpreted attention scores** to find insights on important segments from the history

### Apple, *Machine Learning Research Intern*

May 2022 - August 2022

*Siri Text to Speech Team*

- Implemented an **end-to-end acoustic model for text-to-speech synthesis**
- Built a data synthesis and training pipeline to train deep learning models on **large scale speech corpus (500 hours)**

### Data Science Research Lab at UF, *Research Assistant*

*Multi-Answer Question Answering Benchmarking Dataset*

March 2022 - May 2022

- Annotated 1,000 QA pairs for the **multi-answer QA benchmark dataset**, to track **misinformation, and disinformation**
- Performed **Tweet stance annotation** for 1,000 samples QA pairs from Twitter API Querier using **Dense Distinct Tweet Retriever (DDTR)** model

### AIDA a DARPA funded Question Answering System

September 2021 - March 2022

- Extracted **sentence embeddings** from **100,000 sentences** using **XLNet Roberta for similarity clustering**
- Built pipeline for **cross-lingual Natural Language Inferencing using mT5 model** on LLM **prompt-annotated 5,000 cross-claim pairs**, improving the score by 38%

### ISRO - Indian Space Research Organisation, *Machine Learning Research Intern*

December 2020 - April 2021

*Retrieval of water quality parameters of coastal waters of oceans using satellite data*

- Researched and developed a **modified Neural Network algorithm** to solve the inversion problem of acquiring 6 Inherent Optical Properties from remote surface reflectance data
- Achieved a good settlement with respect to **R-Square values** for each water quality parameter of about **97%**

## PUBLICATIONS

- Neural network based retrieval of inherent optical properties (IOPs) of coastal waters of oceans**, *IEEE India Geoscience and Remote Sensing Symposium (InGARSS 2021)*, [doi.org/10.1109/InGARSS51564.2021.9792013](https://doi.org/10.1109/InGARSS51564.2021.9792013)
- Improving Deep Learning based Automatic Speech Recognition for Gujarati**, *ACM Transactions on Asian and Low Resource Language Information Processing (ACM TALLIP 2021)*, [dl.acm.org/doi/full/10.1145/3483446](https://dl.acm.org/doi/full/10.1145/3483446)
- End-to-End Automatic Speech Recognition for Gujarati**, *17th International Conference on Natural Language Processing (ICON 2020: ACL Anthology)*, [aclanthology.org/2020.icon-main.56](https://aclanthology.org/2020.icon-main.56)

## PROJECTS

### Preliminary Survey on Foundation Language Models, *Research Project*

January 2023 - May 2023

- Performed a thorough analysis on foundation language models, and wrote a 9 page research report
- Trained 10 large language models on NVIDIA A100 using Pytorch in a distributed manner

### mRNA COVID-19 vaccine degradation prediction, *Kaggle Genomics Project*

January 2022 - May 2022

- Established a **hybrid Bi-LSTM Bi-GRU model** to achieve good MCRMSE score of 0.3577 over 5 column values
- Demonstrated that **graph-based architectures better capture sequential patterns in mRNA**, with a **10% improvement in MCRMSE score**

**Image based melanoma detection**, *Medical Research Project* January 2022 - May 2022

- Attained a **sensitivity score of 92.4%** by **ensembling ResNet and EfficientNet based models** by **finding threshold margin maximized over G-Means value**
- Evaluated the application of **self-supervised pre-training based on Bootstrap Your Own Latent (BYOL) model**, achieving an **increase of 3%** for both ResNet and EfficientNet based models

**Schema based dialogue system**, *Spoken Dialogue Systems Research Project* September 2021 - December 2021

- Invented a **zero-shot dialogue system** using a **schema-guided attention model** for **wire-framing dialogue systems**
- Designed **robust entity extraction module** based on **GPT-2 with decent priming** for final round robin evaluation achieving an **average User Satisfaction of 7.3**
- Systematized the study protocol for evaluation for 20 users

**CommonLit Readability Prize**, *Kaggle NLP competition* May 2021 - August 2021  
*Silver medal · 106th/3566 (Top 3%)*

- Formulated **2D attention algorithm for Roberta large** increasing the performance by 15%
- Experimented with fine-tuning techniques by implementing **differential learning rate, gradient accumulation, and custom attention heads** to attain a competitive RMSE of 0.460
- Utilized **Forward Selection OOF to ensemble various models** and boost the overall RMSE to 0.4588 by generalizing the target value

**End-to-End Speech Recognition for low-resource language**, *NLP Research Project* December 2019 - March 2021

- Tailored **beam search decoding** by introducing **multi-level language modeling**, reducing the word error rate by 2.1%
- Innovated **spell correction post-processing using BERT language model**, outpacing the previous performance by 3%
- Analyzed 20,000 words to determine why the system makes erroneous predictions, and to understand how it works
- Scrapped and forged **316M word corpus** for developing **language model using KenLM package with ablation study**

## EDUCATION

**University of Florida, Gainesville, FL, United States** August 2021 - May 2023

Master of Science in Computer Science (*Machine Learning specialization*) (GPA: 3.85/4)

*Relevant coursework: Machine Learning, Machine Learning in Genomics, Spoken Dialogue Systems, Math for Intelligent Systems, Pattern Recognition & Intelligent Systems*

**Dharmsinh Desai University, Nadiad, Gujarat, India** July 2017 - May 2021

Bachelor of Technology in Computer Engineering (GPA: 9.01/10)

*Relevant coursework: Machine Learning, Artificial Intelligence, Big Data Analytics, Digital Image Processing*

## HONORS & CLUBS

- Paper reviewer for 29th International Conference on Neural Information Processing (ACM ICONIP 2022)
- Mentored a group of 5 juniors as the ML Team Head at Developers Student Clubs (DSC) by Google Developers
- Arranged and taught 6+ seminars and workshops on various machine learning concepts at DSC
- Achieved 1<sup>st</sup> place in university for ACM-ICPC, Gwalior-Pune regionals online qualifier Spark

## SKILLS

**Programming Languages & Databases:** Python, Java, C, C++, Latex, Matlab, MongoDB, MySQL

**Framework & Libraries:** Pytorch, Tensorflow, Keras, Pandas, Matplotlib, Numpy, Librosa, Hugging Face, Optuna

**Tools & Services:** Git, Github, GCP, AWS, Microsoft Azure, VS Code, Agile Development Process, CUDA