Pushkar Ambastha

CONTACT INFORMATION	IIT Guwahati Guwahati, India	Portfolio | GitHub | LinkedIn p.ambastha@iitg.ac.in

EDUCATIONAL QUALIFICATIONS

Indian Institute of Technology (IIT) Guwahati, Guwahati, India *(Nov '21 - Present)*
B.Tech - Bio Science and BioEngineering
CGPA: **8.08/10**, Batch Rank: 4th/75

Vidya Sanskar School, Patna, India *(June '19 - Mar '21)*
Higher secondary (12th grade), CBSE Board
Percentage: **92.2%**, School Rank: 2nd/200

Delhi Public School Patna, Patna, India *(June '18 - Mar '19)*
Secondary school (10th grade), CBSE Board
Percentage: **95.2%** , School Rank: 7th/420

PUBLICATIONS

Pushkar Ambastha, *From AlphaFold 2 to AlphaFold 3: A Review on Advancements in Protein Structure Prediction* (Submitted to ACM Computing Surveys (IF: **23.8**). Currently under review)
- Investigated recent advances in protein structure prediction, like AlphaFold 3, which depicted a pattern toward the generalization ability of the models leading toward Large Language Models.
- Reviewed the developments related to current approaches to protein structure prediction and protein design and highlighted a selection of successful applications they have enabled.

INTERNATIONAL RESEARCH INTERNSHIPS

Study of Synthetic Human Memories: AI-Edited Images and Videos which Implant False Memories and Distort Recollection *(Aug. '24 - Present)*
Advisor: Prof. Pattie Maes *(Fluid Interfaces Group, MIT Media Labs, MIT)*
Massachusetts Institute of Technology (MIT) Media Labs
- Developed multi-modal pipeline using original image, AI-enhanced images, image to video, and AI-edited images to video in a successive survey used on old images showing a positive sentiment for potential therapeutic memory reframing. The false recollection is **2.05x** compared to control.
- Created the architecture of the study that examines the impact of AI-altered visuals on false memories, which are recollections of events that didn't occur or deviate from reality.

Automating Detection of APP Abnormalities in Porcine Brain Histology for Post-Traumatic Epilepsy Analysis *(Aug. '24 - Present)*
Advisor: Prof. Ragini Verma *(Center for Biomedical Image Computing & Analytics)*
University of Pennsylvania
- Implemented an automated image annotation pipeline for analyzing Amyloid Precursor Protein (APP) in pig brain histology, significantly reducing manual annotation time and computational resources through efficient tiling and preprocessing techniques.
- Developed a fine-tuned Segformer model for detecting injury patterns in fornix and fimbria regions, incorporating histogram normalization and handling high-resolution histology images (78064 x 65075) with memory optimization (12GB RAM).

Calibrating Agent-Based Models for Tumor-Immune Interactions using Spatial Biopsy Data and Multi-Modal Pipelines with AgentTorch *(June '23 - Present)*
Guide: Ayush Chopra *(Camera Culture Group, MIT Media Labs, MIT)*
Massachusetts Institute of Technology (MIT) Media Labs Report
- Developed methods to calibrate clinical Agent-Based Models (ABM) directly from biopsies to have a mean accuracy of **77%** under the Spatial Agreement Measure *(SAM)* Metric, minimizing the number of biopsy samples taken.
- Designed a novel multi-modal calibrated ABM pipeline to apply gradient-based ABMs to simulate tumor-immune cell interactions. *(for Cytotoxic CD8+ T Cells in multiple carcinoma cases)*
- Extended a novel pip library *AgentTorch* (Pytorch framework to design, simulate, and optimize agent-based models) for the cancer immunotherapy domain. Extracting the parameters from the classification model with an accuracy of **90%**.

Optimizing Medical Segmentation with Integrated Med-SAM and Fast-SAM Models for Enhanced Accuracy in Multi-Modal Imaging *(May '23 - Dec '23)*
Guide: Katie Link *(Hugging Face X Healthcare)*, **Hugging Face, NY** Report
- Developed novel models derived from the cumulative performance and extrapolation of Segment Anything Model *(SAM)*, Medical SAM *(Med-SAM)*, and Fast-SAM.
- Created a streamlined process that reduced the time it takes to analyze images by **68%** and decreased the model size by **82%** compared to the Vanilla SAM model.
- The results observed in Pathology and CT modalities gave an average improvement of **0.48** in mean Intersection of Union *(mIOU)* and **0.42** in Dice Score Coefficient *(DSC)*.

Adaptive Biomedical Segmentation: Enhancing model Explainability through Domain Shift Analysis *(Nov. '22 - Feb. '23)*

Guide: Tushar Kataria *(Computer Vision Group)*, **University of Utah** Report

- Analysed Domain Shift in biomedical image segmentation models as a critical insight into Model Explainability. (for both binary and multiclass semantic segmentation instances)
- Fine-tuned U-Net, DeepLabV3 model on Dataset like GlaS from MICCAI *(2015)*, CRAG, CPM15 to observe domain dependency of models on the dataset, created a pipeline to improve Image masks mIOU and Dice Score.
- The results achieved for metrics like mIOU Loss, Pixel Accuracy, Jaccard Loss, and Dice Score met the standards of prior field experiments, like **0.96** for CRAG and **0.89** for GLAS *(mIOU)*.

SCHOLASTIC ACHIEVEMENTS	- **Inter IIT Tech Meet 11.0 :** Secured **Gold** Medal in building QA Chatbot by *DevRev.ai*, 2023 - **Kaggle 4X Expert:** Reached the Expert tier in all the categories of Kaggle, 2023 - **Bronze Medal (85th/1100):** Open Problems – Single-Cell Perturbations, Kaggle, 2023 - **Bronze Medal (99th/1025):** HuBMAP – Hacking the Human Vasculature, Kaggle, 2023 - **IITG.ai Hackathon:** Secured 6th/25 rank in hackathon conducted by IITG.ai club, 2022 - **Convolve Hackathon:** Secured 28th/231 rank in hackathon among 6 IITs by *Cisco*, 2022 - All India Rank **140** out of 100,000 candidates in National Entrance Screening Test (NEST), 2021 - Qualified for the Pre-Regional Mathematics Olympiad in **PRMO**, 2019

COMPUTATIONAL SKILLS

Programming languages	: Python, C, C++, R; MATLAB
Frameworks	: PyTorch, Tensorflow, Keras
Machine Learning/Data Analysis	: Deep Learning, including CNNs, Transformers

KEY PROJECTS

Classifying SARS-CoV-2 genomes using advanced NLP *(July '23 - Nov '23)*

Course: Bio-informatics Poster

- Developed DNABERT, a pre-trained bidirectional encoder that achieved state-of-the-art performance in predicting promoters, splice sites, and transcription factor binding sites with a **5%** increase in F1-score over traditional models, using minimal fine-tuning on small datasets.
- Implemented a transformer-based pipeline with **99%** accuracy for SARS-CoV-2 genome classification, outperforming traditional random forest and SVC models by **40%** in precision.
- Presented a poster at Conference Hall, IITG and achieved 1st/23 presenting groups

ProteoSynth & GrooveSynth: Advanced Protein Structure Analysis *(Jan '24 - May '24)*

Course: Computational Biology GitHub

- Developed integrated Flask-based platform combining protein sequence generation and active site analysis: generated 10,000+ custom proteins with configurable amino acid sequences.
- Analyzed protein active binding sites, improving ligand-binding prediction accuracy by **20%**.
- Enhanced functionality via custom sequence length, amino acid exclusion, bulk generation (up to 100 sequences/request), and an automated .pdb file generation with downloadable .zip option.
- Secured 1st/75 position in the course with **AS** (Outstanding) grade for the project presentation.

PRODUCT DEVELOPMENT PROJECTS

Domain-specific Question Answering chatbot *(Dec. '22 - Feb. '23)*

11th Inter IIT TechMeet, IIT Kanpur GitHub

- Developed a question-answering system by retrieving the top candidate sentences from the corpus, used up to free collab resources, and *quantized* models to keep *latency* less than *1000 ms*.
- Developed question-answering pipeline using techniques like model distillation, sparsification, pruning, and fine-tuning the *DebertaV3-Base* model to decrease net inference time and have a minimum loss in accuracy.

Cover Generation using all the OpenAI tools *(Jan. '23 - March '23)*

IITG.ai Club, IIT Guwahati GitHub

- Developed a multi-modal pipeline that converts audio/text input into images. Generated optimal transcripts for the podcasts and songs with *OpenAI Whisper* to create prompts.
- Designed pipeline with Latent Diffusion Models (*DALL-E*) to generate aesthetic cover images from created prompts using *ChatGPT/GPT-2* models, and deployed this on Hugging Face Spaces.

RELEVANT COURSE WORK

Bioengineering (Computational):- Bioinformatics ● Biological Data Analysis ● Biophysics ● Computational Biology ● Systems Biology[1] ● Research Methodology[1] ● Analytical Biotechnology

Computer Science:- Intro to Computing ● Computing Lab ● Pattern Recognition and ML

Mathematics:- Linear Algebra ● Basic Calculus ● Complex Analysis

Bioengineering (Wet lab):- Genetic Engineering ● Biochemistry ● Microbiology ● Immunology

MOOCs:- Machine Learning[2] ● Deep Learning[2] ● Computer Vision[2] ● Natural Language Processing (NLP)[2] ● Reinforcement Learning[3]

---

[1] Ongoing course   [2] Course undertaken from Coursera DL Spec.   [3] Course undertaken from Kaggle