

EDUCATION

PhD in Data Science, School of Data Science, University of Virginia JUL 2020 — MAY 2024

Advisors: Dr. Donald E. Brown, Dr. Sana Syed

Thesis: Combining and evaluating multi-modal natural language processing techniques for use in clinical contexts [Link to dissertation](#)

Master of Science in Data Science, School of Data Science, University of Virginia JUL 2018 — MAY 2019

B.E. (Hons.) in Electrical and Electronics, Birla Institute of Technology and Science, Pilani, Goa Campus, India JUL 2012 — MAY 2016

EXPERIENCE([GOOGLE SCHOLAR LINK](#))

Data Scientist, School of Data Science JUN 2024 — Present
University of Virginia Charlottesville, VA

- Set up an analysis pipeline in HIPAA protected compute environment using Polars, Sci-kit learn, Pandas and VSCode.
- Conducted impact analysis of telemedicine on South West VA patients by analyzing Electronic Health Records using XGBoost. Work included data preprocessing, feature engineering, feature selection, model development with hyper-parameter tuning, evaluation, and interpretation. Collaborated with a team of clinicians and data scientists to help identify issues and validate models.
- Contributed to the development of robust multimodal representation models using Mixture of Experts and LLMs for healthcare data designed to handle incomplete multimodal data **Accepted at ICCV Workshops 2025**.
- Part of the team creating the [Research Data Enclave at the University of Virginia](#).

Graduate Research Assistant, School of Data Science AUG 2020 — MAY 2024
University of Virginia Charlottesville, VA

- Developed novel deep learning architecture using PyTorch for high-resolution histopathology image captioning, creating an encoder-decoder model that uses pre-trained Vision Transformers (ViT) and a BERT-based decoder to generate descriptions for the whole image – achieved 79.98% accuracy in Tissue Type classification and a BLEU-4 score of 0.5818 for caption generation on a dataset of 25,120 Whole Slide Image (WSI)-text pairs.
- Designed the model to incorporate tissue type, gender, and the actual caption, and enhanced interpretability by preserving patch-based localization information to associate generated tokens with specific image regions.
- Pioneered interpretable deep learning methods to identify Long COVID risk factors by analyzing historical diagnosis code data from the National COVID Cohort Collective (N3C) using attention-based Bidirectional LSTM models. This approach achieved an AUROC of 0.93 (0.88 F1 Score) for predicting Long COVID, significantly outperforming models trained on randomly selected controls.
- Implemented Gradient-weighted Class Activation Mapping (GradCAM) and attention scoring to provide interpretable insights into the most important historical diagnoses and their temporal trends contributing to a Long COVID diagnosis, identifying conditions such as Mixed Hyperlipidemia, Essential Hypertension, and Dyspnea. Research funded by NIH.

Data Scientist, Gastro Data Science Lab JUN 2019 — JUL 2020
University of Virginia Charlottesville, VA

- Created Convolutional Neural Network (CNN) based models to detect gastro-intestinal diseases (Crohn's, Celiac) on high resolution biopsy images.
- Lab received the \$100k Litwin IBD Pioneers Program, Crohn's & Colitis Foundation grant based on work.

Associate Consultant, Development Bank of Singapore Project JUL 2016 — OCT 2017
Capgemini Hyderabad, India

- Developed Java-based REST APIs for the Development Bank of Singapore (DBS) using Spring Boot.

PROJECTS

Automated report generation for histopathology using Vision Transformers and BERT ([code](#))([paper](#))

Created novel method for captioning extremely high resolution tissue images by fine-tuning pre-trained Vision Transformers and BioClinical BERT using HuggingFace Transformers and PyTorch Lightning, combining vision and language in histopathology. Achieved 0.12 BLEU-4 score and able to detect tissue type with 90% accuracy. **Accepted at IEEE ISBI 2024**.

Evaluating bias in DeepSeek R1 reasoning model on Pew Research data ([code](#))

Created code to create prompts from [Opinions QA](#) dataset and run DeepSeek R1 models locally to evaluate biases in the model responses.

Stain Transfer and Re-stitching code for Whole slide images([code](#))

Python script to transfer stain between source and target domain using Vahadane stain transfer. Re-stitches the original WSI at native resolution.

(434) 284-3794
Charlottesville, VA
sauravsengupta7@gmail.com

Saurav Sengupta

Data Scientist, UVA School of Data Science github.com/ssen7 linkedin.com/in/saurav-sengupta

Portfolio: ssen7.github.io
github.com/ssen7

Multiple instance learning for Whole slide image classification using ResNet-50([code](#))

Resnet-50 model trained using fastai to detect gastrointestinal diseases like Celiac and Environmental Enteropathy.

SELECTED PAPERS AND PRE-PRINTS ([GOOGLE SCHOLAR LINK](#))

1. *Towards Robust Multimodal Representation: A Unified Approach with Adaptive Experts and Alignment*. *arXiv preprint arXiv:2503.09498*. ([paper](#))

Accepted at CVAMD @ ICCV 2025

Moradinasab, N., Saurav Sengupta., Liu, J., Syed, S., & Brown, D. E.

2. *Automatic Report Generation for Histopathology images using pre-trained Vision Transformers and BERT*, *IEEE International Symposium on Biomedical Imaging (ISBI) 2024* ([paper](#))

Saurav Sengupta and Donald E. Brown

3. *Automatic Report Generation for Histopathology images using pre-trained Vision Transformers, Machine Learning for Health (colocated with NeurIPS 2023), Findings Track* ([paper](#))

Saurav Sengupta and Donald E. Brown

4. *Determining risk factors for Long COVID using Positive Unlabeled learning on Electronic Health Records data from NIH N3C*, *IEEE ICMLA 2023* ([paper](#))

Saurav Sengupta, Johanna Loomba, Suchetha Sharma, Donald Brown et al.

5. *Analyzing historical diagnosis code data from NIH N3C and RECOVER Programs using deep learning to determine risk factors for Long Covid*, *IEEE BIBM 2022* ([paper](#))

Saurav Sengupta, Johanna Loomba, Suchetha Sharma, Donald E Brown et al.

6. *MACHINE LEARNING FOR CROHN'S DISEASE PHENOTYPE MODELING USING BIOPSY IMAGES*. *Inflammatory Bowel Diseases*, 27(Supplement_1), S10-S11.(2021)

Sana Syed, Saurav Sengupta, Lubaina Ehsan, Erin Bonkowski, Christopher Moskaluk, Anne Griffiths, Anthony Otley

7. *Artificial intelligence-based analytics for diagnosis of small bowel enteropathies and black box feature detection*. *Journal of pediatric gastroenterology and nutrition* 72, no. 6 (2021): 833-841.

Sana Syed, Lubaina Ehsan, Aman Shrivastava, Saurav Sengupta, Marium Khan, Kamran Kowsari, Shan Guleria et al

8. *Deep Learning for Visual Recognition of Environmental Enteropathy and Celiac Disease (IEEE BHI 2019)*

Aman Shrivastava, Karan Kant, Saurav Sengupta, Sung-Jun Kang, Marium Khan et al.

9. *Deep learning for detecting diseases in gastrointestinal biopsy images. (IEEE SIEDS 2019)*

Aman Srivastava, Saurav Sengupta, Sung-Jun Kang, Karan Kant, Marium Khan, S. Asad Ali, Sean R. Moore et al.

10. *Mo1992-solving the stain dilemma: Computational image analysis to address differential tissue staining color bias in duodenal biopsies*, *Gastroenterology* 156, no. 6 (2019): S-914.

Sana Syed, Aman Shrivastava, Karan Kant, Saurav Sengupta, Luke Kang, Marium N. Khan, Najeeha T. Iqbal et al

SKILLS

Tools and Languages

Python, R, Java 7 (Oracle Certified Associate), C, Verilog

Packages/Tools

PyTorch, PyTorch Lightning, TensorFlow, AWS, PySpark, Palantir Foundry, SQL, Weights and Biases (W&B), TensorboardX, Polars, Pandas, Sci-kit learn, XGBoost, Plotly, Huggingface

ACTIVITIES

Raven Honors Society Member (Oldest Honors Society at UVA)

2024

Reviewer for *Journal of Medical Internet Research* and *ACM Transactions on Computing for Healthcare*

2024

ML4H ([link](#)) 2024 Reviewer

2024

2023 MIDAS Future Leaders Summit Cohort, University of Michigan, Ann Arbor

APR 2023

ICMLA ([link](#)) 2023, 2024 Reviewer

2023

School of Data Science MSDS Capstone Mentor, University of Virginia

JUL 2020 — JUL 2022

Python Instructor, SOAR Scholars Program, University of Virginia

Spring 2021