Sharan Sahu

ssahu01@berkeley.edu | sharansahu.com | https://github.com/sharansahu

Research Interests

My research interests lie in statistical machine learning and using statistical/machine learning tools to improve clinical decision-making and bridge the gap between machine learning methodology and clinical practice. In particular, I am interested in theory and methods in the areas of high-dimensional statistics, optimization, reinforcement learning, nonparametric estimation, language and diffusion models, and uncertainty quantification.

Education

Cornell University — Ithaca, NYC
Doctor of Philosophy in Statistics and Machine Learning — *Aug. 2024 – May 2029*

- First Year Support: Cornell University Graduate Fellowship (2024-2025)

University of California, Berkeley — Berkeley, CA
Bachelor of Arts in Computer Science — *Aug. 2020 – May 2024*

- GPA: 3.998 / 4.000
- Advisor: Iain Carmichael, Ryan Tibshirani
- Relevant Graduate Coursework:
  - **Statistics:** Mathematical Statistics (STAT 210A), High Dimensional Statistics (STAT 210B), Statistical Learning Theory (STAT 241B / CS 281B), Mean Field Asymptotics in Statistical Learning (STAT 260)
  - **Computer Science / Electrical Engineering:** Parallel Computing (CS 267), Graduate Algorithms (CS 270), Deep Reinforcement Learning (CS 285), Convex Optimization and Approximation (EE 227C)
  - **Mathematics:** Measure Theory and Topology (MATH 202A), Functional Analysis (MATH 202B), Measure Theoretic Probability Theory (MATH 218A / STAT 205A), Differential Topology (MATH 214)
- Relevant Undergraduate Coursework:
  - **Computer Science / Electrical Engineering:** Machine Learning (CS 189), Deep Learning (CS 182), Algorithms (CS 170), Data Structures (CS 61B), Operating Systems (CS 162), Database Systems (CS 186), Optimization (EECS 127), Probability and Random Processes (EECS 126), Networking (CS 168)
  - **Mathematics:** Real Analysis (MATH 104), Complex Analysis (MATH 185), Introduction to Differential Forms and Integration Theory (MATH 105), Linear Algebra (MATH 110)

Research Experience

Undergraduate Researcher — Jan. 2023 - Present
Berkeley Artificial Intelligence Research Lab — *Berkeley, CA*

- Advised by Iain Carmichael and Ryan Tibshirani on high-dimensional statistical and deep learning models for Computational Pathology applications, specifically nonparametric estimation, conformal inference, and computer vision
- Developing "wsic", a parallelized image segmentation package for whole slide images (WSIs) including H&E, immunohistochemistry, and multiplex immunofluorescence with hyperparameter tuning using distributed computing across GPUs with Ray
- Developing a cell/nuclear segmentation model that can be trained on mixed annotations with masks, bounding boxes, and centroids to significantly reduce the amount of time it takes to annotate data sets for cell segmentation using DETR, ViTs, and U-Net under a representational stitching and multi-task learning framework with various loss balancing techniques such as grid searches and gradient normalization
- Designing tools for the "segt" package to quantitatively and qualitatively evaluate segmentation results, including detailed error metrics and visualization tools for various instance segmentation formulations
- Developing conformal vision language modeling for pathology, integrating multi-task learning and uncertainty quantification techniques to enhance the accuracy and reliability of histopathological image analysis and pathology report generation, ensuring statistically valid confidence measures in predictions

Undergraduate Researcher — May. 2022 - Aug. 2022
Stanford Artificial Intelligence Lab (SAIL) — *Palo Alto, CA*

- Advised by Despoina Paschalidou and Suya You on evaluating generative networks' performance
- Developed reliable and efficient metrics for fidelity, diversity, and authenticity used to evaluate the performance of generative networks (GANs, VAEs, Flow-Based Models, Diffusion Models, etc)

- Refined Precision and Recall metrics using methods such as SimCLR, Barlow Twins, DeepCluster, and SwAV, integrating concepts from nonparametric statistics and manifold theory
- Addressed high-dimensional embedding challenges by implementing PCA and Robust PCA, improving metric reliability and accuracy in evaluating generative models
- Explored theoretical aspects of generative models' precision-recall trade-offs, developing new metrics based on f-divergences and topological significance, and provided formal performance guarantees using concentration inequalities and persistent homology

Undergraduate Researcher Sept. 2021 - May. 2022
Stanford Artificial Intelligence Lab (SAIL) *Palo Alto, CA*

- Advised by Stanford Postdoc Despoina Paschalidou and ARL West research scientist Suya You on generative modeling, computer vision, and reinforcement learning
- Enhanced scene grammar and policy gradient methods in NVIDIA's Meta-Sim framework to improve the quality and relevance of synthetic datasets for downstream tasks
- Implemented differentiable rendering techniques, such as Differentiable Monte Carlo Ray Tracing, to achieve efficient backpropagation through non-differentiable rendering functions, optimizing for visual similarity to real data
- Explored alternative gradient estimation methods, including PPO, TRPO, and Generalized Advantage Estimators (GAE), to reduce variance and improve task performance in synthetic data generation
- Developed probabilistic scene grammar trees and parameter estimation techniques using hierarchical clustering, Bayesian networks, and deep learning algorithms, enhancing the synthesis of plausible scene structures without ground truth annotations

Undergraduate Researcher Jan. 2021 – May 2021
UC Berkeley Electrical Engineering and Computer Science (EECS) *Berkeley, CA*

- Developed and optimized machine learning models to accelerate Density Functional Theory (DFT) computations to identify efficient electrocatalysts in Hydrogen Energy Storage and methaneation
- Redesigned neural network architectures, including SchNet, DimeNet++, and Crystal Graph Convolutional Neural Networks (CGCNN), to model quantum interactions more effectively
- Implemented feature selection using Lasso regularization and stability enhancements with Ridge regularization to improve model interpretability and performance in computational chemistry tasks
- Employed quasi-Newton optimization methods to increase efficiency in handling large datasets and achieve faster convergence, significantly reducing computational resources required for DFT calculations
- Enhanced SchNet framework performance, achieving lower loss and mean-average error through continuous-filter convolutional layers and rotationally invariant energy predictions

Industry Experience

Cofounder & Lead Research Scientist May. 2023 - Present
167 Labs *Fremont, CA*

- Engineered an Employee Self Service framework for HR services, providing policy overviews and detailed information
- Utilized state-of-the-art LLMs (Mistral, OpenAI, Claude) with RAG (FAISS), implementing reranking models and database optimizations for enhanced information retrieval
- Integrated framework with external APIs and models for real-time calculations, enabling context-based function calling and personalized responses based on employee details and organizational hierarchy and bonus and paycheck calculations, incorporating historical performance and fluctuating tax brackets

Data Science Intern May. 2023 - Aug. 2023
Marine Corps Tactical Systems Support Activity (MCTSSA) *Camp Pendleton, CA*

- Engineered Logistic Regression and fine-tuned BERT models to accurately classify various system requirements, slashing testing efforts by 500% and saving $1M in system requirement classification tasks
- Developed a full-stack chat application interfaced with classified Marine Corps documents, employing Node.js, Express.js, Flask, and fine-tuned LLMs like Llama V2 and Falcon 7B
- Created a vector-based retrieval database using the pgvector extension in PostgreSQL for fast similarity search along with experimenting with Annoy, Milvus, and Pinecone

Software Engineering Intern May. 2022 - Aug. 2022
Marine Corps Tactical System Support Activity (MCTSSA) *Camp Pendleton, CA*

- Develped ML solutions utilizing CNN and Sparse Dictionary Learning for robust classification of standard waveforms in the EMF spectrum
- Crafted data preprocessing and image extraction pipelines for a comprehensive application extracting definitions from documentation, employing OpenCV, Tesseract, and Schwartz-Hearst algorithm

- Engineered full-stack cloud applications for streamlined resource management and deconfliction, leveraging the React JS-Framework

Product Management Intern May 2021 – Aug. 2021
Novartis *Libertyville, IL*

- Developed ML and data science innovations including database optimizations and improved statistical modeling, culminating in over $1M savings
- Employed ARIMA and LSTM models to forecast inventory and equipment values, leveraging Pandas, SQL, and PyTorch
- Collaborated with the development team to institute OAuth/SSO protocols for improved security and efficiency in revenue management business applications

Software Engineering and Data Science Intern Sept. 2018 – June 2020
Northrop Grumman Corporation *Rolling Meadows, IL*

- Was part of the data science team in leveraging Kalman Extended and Particle Filtering, AdaBoosting, and PCA with decisive logic for a self-autonomous system
- Crafted a CNN architecture with batch normalization, max pooling, and dropout, enhancing surface deformity detection using Keras and TensorFlow
- Authored a comprehensive 20-page research paper, showcased at the annual Northrop Grumman research conference

Teaching Experience

EECS 16A uGSI Aug. 2021 - Dec. 2021
UC Berkeley Electrical Engineering and Computer Science (EECS) *Berkeley, CA*

- Engaged in EECS 16AB sequence, imparting key knowledge on signal processing, control, circuit design, and machine learning, intertwined with practical linear algebra applications
- Conducted individual and group tutoring sessions, along with orchestrating EECS 16A review workshops
- Guided and supported 50 students in EECS 16A labs, bridging theoretical concepts with practical implementation
- Received an average rating of 4.9 / 5.0 in teaching quality, subject mastery, and proactive assistance from students

COMPSCI 61A Mentor (CSM) Jan. 2021 - May. 2021
UC Berkeley Electrical Engineering and Computer Science (EECS) *Berkeley, CA*

- Taught topics in the CS 61A series focusing on software construction, machine operation, and programming abstraction through Python 3, Scheme, and SQL languages
- Conducted weekly tutorials and bi-weekly labs, emphasizing key data structures like trees, binary search trees, and core concepts like recursion and induction
- Prepared and delivered review sessions pre-exams, collaborated in preparing problem sets, quizzes, and exams to ensure a coherent learning curve for students
- Provided individualized guidance during office hours, reflecting on improved student performance over the semester

Academic Honors

Cornell University Graduate Fellowship

Science, Mathematics, and Research For Transformation DoD Scholarship (SMART)

Citadel Datathon Top 5 Placement

UC Berkeley Outstanding GSI Award

UC Berkeley Department of Data, Society, and Computing Data Science Insights Award Winner

USA Physics Olympiad (USAPhO) Qualifier

Math Olympiad Program (MOP) Invitee

USA Mathematics Olympiad (USAMO) Qualifier

USA Junior Mathematics Olympiad (USAJMO) Qualifier

American Invitational Mathematics Examination (AIME) Qualifier

Northrop Grumman Engineering Scholarship

Lockheed Martin Engineering Scholarship

Publications and Preprints, Conference Presentations, and Guest Lectures

Hovenga, V., **Sahu, S.**, Carmichael, I. *Mixed Annotation Cell and Nuclear Segmentation Model With Multi-Task Learning.* Conference on Neural Information Processing Systems (NeurIPS). May 2024. *Paper (In Review)*

Sahu, S., Johnson, T., Marsh, B., Kennedy, R., Kays, J., Das, A. *Enhancing Warfighter Capabilities: Leveraging Machine Learning and NLP in System Requirements Analysis and Clustering for C4ISR.* Naval Applications of Machine Learning (NAML). San Diego, CA. March 2024. *Poster Presentation*

Sahu, S., Li,, J., Hovenga, V., Smith, K., Yin, N., Chen, A., Carmichael, I. *WSIC: A Python Package To Facilitate Running Nuclear/Cellular Segmentation On Whole Slide Images.* Journal on Open Source Software (JOSS). January 2024. *Paper (In Review)*

Sahu, S. *How Do Neural Networks Learn.* Naval Postgraduate School. Monterey, CA. December 2023. *Guest Lecture*

Sahu, S., Wadhwa, R., Yallapragada, L., Das, A. *Systems Requirement Clustering With Machine Learning.* Neptune Office of Naval Research (ONR) Conference. Davis, CA. November 2023. *Oral Presentation*

Sahu, S. *Anchored Intelligence: Navigating the Waters of Machine Learning And Charting the Course to Augmented Decision-Making.* Naval Postgraduate School. Monterey, CA. October 2023. *Guest Lecture*

Sahu, S., Flaherty, D., Vinchure, A., Pei, J., You, S. *Developing Multi-Dimensional Metrics for Precision, Recall, Fidelity, Diversity, and Authenticity in Evaluating Generative Networks Performance using Deep Perceptual Embeddings.* SPIE DCS. Orlando, FL. May 2023. *Poster Presentation*

Das, A., **Sahu, S.**, Johnson, T., Kennedy, R. *Systems Requirements Clustering With Machine Learning and Architecture Design.* Naval Applications of Machine Learning (NAML). San Diego, CA. March 2023. *Oral Presentation*

Sahu, S., Flaherty, D., Vinchure, A., Pei, J., You, S. *Developing Multi-Dimensional Metrics for Precision, Recall, Fidelity, Diversity, and Authenticity in Evaluating Generative Networks Performance using Deep Perceptual Embeddings.* DoD 6.1 Research Conference. Arlington, VA. September 2022. *Oral Presentation*

Li, P., **Sahu, S.**, Shen, T., Rizvi, S., Terrell-Perica, P., You, S. *Advancing Procedural Scene Synthesis through Enhanced Grammars and Gradient Policies in MetaSim.* UC Berkeley Data Science Conference. Berkeley, CA. May 2021. *Oral Presentation*

Tong, S., **Sahu, S.**, Huynh, T., Zhang, E., Majmudar, J. *Using Machine Learning to Model and Discover New Catalysts To Address The Energy Challenges Posed by Climate Change.* UC Berkeley Data Science Conference, Berkeley, CA, May 2021. *Oral Presentation*