

Ved Piyush

PhD, Statistics · University of Nebraska–Lincoln

vedpiyush93@gmail.com · +91 85954 69548 · India

Google Scholar · ORCID 0000-0002-5107-9279 · Scopus 58259201600 · LinkedIn

RESEARCH INTERESTS

Uncertainty quantification for deep learning — calibration, prediction intervals, and ensemble Kalman-filter approaches to model stacking. Statistical machine learning for genomics, multi-omics and health applications. Generative AI and large language models, including retrieval-augmented generation and evaluation. Time-series forecasting and state-space models. Design and analysis of experiments.

EDUCATION

Ph.D. in Statistics, University of Nebraska–Lincoln

Aug 2018 – Dec 2023

DISSERTATION [*A Generalized Stacking Method Using Matrix Ensemble Kalman Filter-Based Multi-Arm Neural Network*](#)

Advisor: Dr. Souparno Ghosh · GPA 3.99 / 4.00

M.S. in Statistics, University of Nebraska–Lincoln

Aug 2018 – May 2020

En route to the Ph.D. · GPA 4.00 / 4.00

M.S. in Statistics, University of Minnesota–Twin Cities

Sep 2015 – Jul 2018

Advisor: Dr. Kean Ming Tan · GPA 3.76 / 4.00

B.Sc. (Honours) in Statistics, Sri Venkateswara College, University of Delhi

Jul 2011 – May 2014

First Division

PROFESSIONAL EXPERIENCE

Senior Data Scientist, IBM Chief Analytics Office

Sep 2023 – Apr 2026

Manhattan, New York, USA

Applied statistical and machine-learning methodology to enterprise problems, spanning time-series forecasting through to large language models and multi-agent systems. Four systems reached production.

- **Automated document evaluation.** Led a generative-AI validation pipeline that automated compliance review of business-partner submissions against predefined criteria, saving approximately USD 3M in manual effort; included a multi-format extraction layer and a systematic tradeoff study across context-window size, inference latency, cost and accuracy.
- **Retrieval-augmented assistant.** Designed and deployed a retrieval-augmented generation system serving more than 8,000 employees, with benchmarking of open-source and proprietary foundation models on retrieval quality, latency and cost.
- **Agentic content platform.** Built a multi-agent orchestrator coordinating tool calls across extraction, embedding and grounded generation, so that generated copy remains anchored to its source documents.
- **Revenue forecasting.** Built a weighted-ensemble forecasting system (ARIMA, ETS, regression and tree-based models) for approximately USD 62B of global revenue across business units, with ensemble weights optimised per unit against held-out validation.

Senior Data Science Intern, IBM Chief Analytics Office

May 2022 – Aug 2022

Armonk, New York, USA

Built machine-learning models recommending strategic products to IBM clients, paired with an interpretability framework that explained each recommendation in terms of competitive parity — so that a recommendation could be justified, not merely issued.

Applied Research Data Science Intern, LinkedIn

May 2021 – Aug 2021

Premium Data Science Group, Silicon Valley, California, USA

Developed multi-input deep-learning models for subscription-cancellation prediction, combining several distinct signal types about a member into a single churn model, and designed an optimised A/B testing framework that shortened the measurement cycle for product experiments.

Data Science Intern, Seagate Technology

Jan 2018 – Jul 2018

Global Wafer Systems, Bloomington, Minnesota, USA

Deployed a recommendation system surfacing internal analytics reports to employees across two global sites. Automated the full pipeline from extraction and cleaning through model building to writing recommendations into the production database; identified the metrics by which recommendation quality would be judged and built an R Shiny dashboard to monitor accuracy and relevancy against them; and designed tests comparing competing recommendation algorithms and evaluation strategies.

Energy Data Analysis Intern, Hennepin County Facility Services

Oct 2017 – Dec 2017

Minneapolis, Minnesota, USA

Built statistical models quantifying the reduction in expenditure attributable to energy-saving practices adopted in county-owned buildings, providing evidence-based justification for the programme's cost, and presented the findings to the energy managers of Hennepin County and the City of Minneapolis.

Advanced Analytics Intern, Travelers Insurance

Jun 2016 – Aug 2016

Personal Insurance R&D, St. Paul, Minnesota, USA

Developed models predicting the occurrence and severity of fire claims on home insurance policies, and engineered a substitute for an expensive third-party data attribute using inexpensive data already held — giving the company renegotiation leverage with the vendor. Presented the modelling approach and findings to both technical and non-technical stakeholders.

PUBLICATIONS

Google Scholar: 145 citations, h-index 3, i10-index 3.

Peer-Reviewed Journal Articles

Washburn, J. D., Varela, J. I., Xavier, A., Chen, Q., Ertl, D., Gage, J. L., Holland, J. B., et al., including **Piyush, V.** (2025). [Global genotype by environment prediction competition reveals that diverse modeling strategies can deliver satisfactory maize yield estimates](#). *Genetics*, 229(2), iyae195.

Rodene, E., Fernando, G. D., **Piyush, V.**, Ge, Y., Schnable, J., Ghosh, S., & Yang, J. (2024). [Image filtering to improve maize tassel detection accuracy using machine learning algorithms](#). *Sensors*, 24(7), 2172.

Zheng, J., Hu, B., Zhang, X., Ge, Q., Yan, Y., Akresi, J., **Piyush, V.**, Huang, L., & Yin, Y. (2023). [dbCAN-seq update: CAZyme gene clusters and substrates in microbiomes](#). *Nucleic Acids Research*, 51(D1), D557–D563.

Book Chapters

Piyush, V., Ellingson, L., VanderPlas, S., & Ghosh, S. Introduction to Image Data Analysis. In *Festschrift in Honour of Professor Tathagata Bandyopadhyay*. Springer. **Accepted, in press.**

Preprints and Manuscripts Under Preparation

Piyush, V., Yan, Y., Zhou, Y., Yin, Y., & Ghosh, S. A Matrix Ensemble Kalman Filter-based Multi-arm Neural Network to Adequately Approximate Deep Neural Networks. [arXiv:2307.10436](https://arxiv.org/abs/2307.10436). Journal version under preparation.

Piyush, V., Fernando, G. D., Zhang, R., Weil, M. R., Pal, R., & Ghosh, S. A Multi-arm ANN to combine information from multiple deep learners and attach uncertainty. Under preparation.

Refereed Conference Presentations and Abstracts

Piyush, V., Fernando, G., & Ghosh, S. (2023). Approximating multi-arm deep learners with in-situ trained ensembles of shallow artificial neural networks. *Western North American Region of the International Biometric Society (WNAR)*.

Piyush, V., Yan, Y., Zhou, Y., Yin, Y., & Ghosh, S. (2023). A Matrix Ensemble Kalman Filter-based multi-arm neural network. *International Indian Statistical Association Conference*.

Piyush, V., Zhang, R., Pal, R., & Ghosh, S. (2022). Sequence predictability score to boost classification of protein sequences. *Conference on Intelligent Systems for Molecular Biology (ISMB)*.

Fernando, G., **Piyush, V.**, & Ghosh, S. (2021). [Deep learning methods for tassel count time-series](#). *ESS Open Archive*; North American Plant Phenotyping Annual Conference.

RESEARCH EXPERIENCE

AI for Precision Medicine: Novel Cancer Drug Discovery

Jan 2022 – Aug 2023

University of Nebraska–Lincoln & Argonne National Laboratory · Advisor: Dr. Souparno Ghosh

Developed an in-situ model-averaging procedure that matches state-of-the-art transfer-learning deep models in predictive accuracy while producing prediction intervals with higher coverage and tighter width than Monte Carlo Dropout. Applied to cancer drug-response prediction in collaboration with Argonne National Laboratory researchers.

Interpretable Deep Learning for Carbohydrate-Substrate Prediction

Sep 2021 – Aug 2023

University of Nebraska–Lincoln · Advisors: Dr. Souparno Ghosh, Dr. Yanbin Yin

Built sequence embeddings (Word2Vec, Doc2Vec, FastText) over a large unsupervised microbiome corpus, then trained attention-based recurrent architectures for interpretable multiclass classification under limited supervision.

Session-Based Music Recommendation with Sequence-to-Sequence Networks

May 2019 – Jun 2022

University of Nebraska–Lincoln · Advisor: Dr. Souparno Ghosh

Developed seq2seq models combining user behaviour, acoustic track features and playback-device metadata for session-based recommendation.

Predicting Maize Yield from Field Images using Local Counts Regression

Mar 2019 – Dec 2020

University of Nebraska–Lincoln · Advisor: Dr. Souparno Ghosh

Built probabilistic CNN and CNN-LSTM architectures for yield prediction, with the CNN-LSTM variant explicitly modelling temporal structure across image sequences.

Consensus Transcription for the Zooniverse SCOTUS Notes Project

Jun 2017 – Dec 2017

University of Minnesota Libraries and Zooniverse · Research collaborator and consultant

Applied natural language processing to crowd-sourced transcriptions of U.S. Supreme Court Justices' handwritten conference notes, advising the project's programmers and library staff on aggregation methodology.

Interpretable Recommendation for Implicit-Feedback Data

May 2017 – Jul 2018

University of Minnesota–Twin Cities · Advisor: Dr. Kean Ming Tan

Developed collaborative-filtering models for implicit feedback that explain each recommendation in terms of the user's prior purchases.

TEACHING EXPERIENCE

Instructor of Record — University of Nebraska–Lincoln

STAT 218: Introduction to Statistics

S 2020, Su 2020, F 2021

Sole instructor across three terms, with full responsibility for syllabus, delivery, assessment and grading, including a complete transition to online delivery during the pandemic. Reached approximately 400 graduate and undergraduate students across both universities.

Graduate Teaching Assistant and Laboratory Instructor

University of Nebraska–Lincoln

- STAT 802: Design and Analysis of Research Studies — Fall 2018, Spring 2019
- STAT 801A: Statistical Methods in Research — Fall 2019, Fall 2020
- STAT 463: Introduction to Statistical Inference — Spring 2021

University of Minnesota–Twin Cities

- STAT 5201: Sampling Methodology in Finite Populations — Fall 2017
- STAT 4051: Applied Statistics I — Fall 2016
- STAT 5021: Statistical Analysis — Spring 2016

Guest Lectures

STAT 302: Mathematical Statistics and Modeling II

March 2026

University of Nebraska–Lincoln (Prof. Souparno Ghosh)

Three sessions covering deep-learning foundations, convolutional networks on benchmark datasets, and pretrained transformer models.

Courses Proposed

- **Uncertainty Quantification for Deep Learning.** Calibration, conformal prediction, deep ensembles, and Bayesian and Kalman-filter based approaches, with healthcare and enterprise decision-support case studies.
- **Applied Generative AI: Retrieval, Evaluation and Governance.** Foundation-model selection, retrieval-augmented generation, and the statistical problem of evaluating systems whose output is text.

Outreach

Volunteer Teacher, Mathematics and Statistics — Sweccha

2014 – 2015

Non-profit organisation, India

Taught mathematics and statistics to children from underprivileged backgrounds, preparing and delivering lessons pitched to widely varying levels of prior schooling.

INVITED TALKS AND WORKSHOPS

Workshops Delivered

Machine Learning and Deep Learning with Gene Sequences

March 2023

Department of Food Science, University of Nebraska–Lincoln

Three-day workshop for graduate students and postdoctoral researchers on feature extraction (Word2Vec, Doc2Vec, FastText) and recurrent architectures trained via transfer learning.

Invited Talks and Seminars

- A Multi-arm ANN Model for Stacking Deep Learners and Attaching Uncertainty. *Argonne National Laboratory, July 2023.*
- Matrix Ensemble Kalman Filter Multi-arm Neural Network. *Master's Seminar, Department of Statistics, California Polytechnic State University, December 2024.*
- Developing Interpretable Recurrent Neural Networks with Attention using Transfer Learning for Gene Sequence Classification. *Department of Food Science, University of Nebraska–Lincoln, October 2022.*
- Automatic Image Captioning using Convolutional and Recurrent Neural Networks. *Graphics Group, Iowa State University, March 2022; Department of Statistics Seminar, University of Nebraska–Lincoln, October 2021.*
- Few-Shot Learning to Deal with Rare Classes in Multiclass Classification. *Department of Statistics Seminar, University of Nebraska–Lincoln, October 2021.*
- Collaborative Filtering for Implicit Feedback Datasets. *Department of Statistics Seminar, University of Nebraska–Lincoln, October 2018; Syngenta Statistics and Data Science Student Seminar, March 2018.*

Poster Presentations

- Sequential Music Tracks Skip Prediction using Recurrent Neural Networks. *Department of Statistics, University of Nebraska–Lincoln, April 2021.*
- Segmentation and Classification of Overlapping Jellybeans. *Department of Statistics, University of Nebraska–Lincoln, April 2020.*

HONOURS, AWARDS AND FELLOWSHIPS

Fellowships

- **The Othmer Fellowship**, University of Nebraska–Lincoln, Fall 2018 to Fall 2021. USD 8,000 per year for the first three years of the doctoral programme, awarded to recruit exceptional scholars pursuing a terminal degree.
- **Seagate Technology Fellowship**, University of Minnesota–Twin Cities, Fall 2017. USD 5,500, sole recipient from the statistics department that year.

Professional Awards

- **IBM Outstanding Innovation Award**, October 2025, for the GenAI Document Validation System, DocEval.
- **Brandon Hall Group Excellence Awards**, December 2025. Gold, Best Advance in Generative AI for Business Impact, for DocEval; Bronze, Best Advance in AI for Business Impact, for the Enterprise RAG assistant.
- **Stevie Awards**, 2025. American Business Awards, August 2025, and International Business Awards, November 2025, for DocEval.
- **Mutual of Omaha Data Science Competition**, Third Place and Best in Show, March 2019.
- **Travelers Analytics Insurance Modeling Competition**, University of Minnesota, First Place, December 2015.

SOFTWARE

Research code for the dissertation and collaborative projects is maintained as a set of documented, reproducible repositories covering ensemble Kalman-filter stacking, uncertainty quantification for multi-arm deep learners, interpretable sequence models for microbiome data, and genotype-by-environment yield prediction. Available at github.com/vedpiyush93-stack.

STATISTICAL CONSULTING

Statistical and NLP Consultant — SCOTUS Notes Transcription Project

2017

University of Minnesota Libraries and Zooniverse

Advised on and designed consensus-aggregation methodology for crowd-sourced transcriptions of U.S. Supreme Court Justices' handwritten conference notes, and explained the algorithm and its behaviour on edge cases to political scientists and library staff.

Design of Experiments and Applied Statistical Consulting

Ongoing

University of Nebraska–Lincoln

Developed through formal statistical-consulting training in STAT 8801 and STAT 825, and through instructing STAT 802, Design and Analysis of Research Studies.

Enterprise AI Systems

2023 – 2026

IBM Chief Analytics Office

Translating research into production tools for internal and business-partner use — scoping stakeholder questions, selecting methods to match them, and characterising what the resulting answers can and cannot support.

PROFESSIONAL SERVICE

- Student Panelist, STAT 810 Internship Panel, Department of Statistics, University of Nebraska–Lincoln, October 2021 (selected by Prof. Erin Blankenship) and Fall 2022 (selected by Prof. Kent Eskridge).
- Search Committee Member, Business Associate position, Department of Statistics, University of Nebraska–Lincoln, Spring 2019.
- Officer, Statistics in the Community (STATCOM), University of Minnesota–Twin Cities, 2017–2018. Provided pro-bono statistical consulting to community organisations.

- Mentored junior graduate students informally on machine-learning and deep-learning projects during doctoral studies. Prepared to supervise across artificial intelligence, machine learning, uncertainty quantification and applied statistics.

TECHNICAL EXPERTISE

Statistics	Generalized linear mixed models; ARIMA, ETS and state-space models; design and analysis of experiments; Bayesian inference; dimensionality reduction; uncertainty quantification and calibration
Machine learning	Deep learning, ensemble methods, recommendation systems, natural language processing, computer vision, time-series forecasting, A/B testing
Generative AI	Retrieval-augmented generation, foundation-model evaluation and selection, prompt engineering, multi-agent orchestration
Programming	Python, R, SAS, SQL, MATLAB, LaTeX
Libraries	PyTorch, TensorFlow, Keras, Hugging Face Transformers, scikit-learn, statsmodels, pmdarima, statsforecast, LangChain, LlamaIndex, spaCy, NLTK; Milvus, FAISS, ChromaDB; ggplot2, R Shiny, Plotly, Dash
Certifications	SAS Certified Base Programmer for SAS 9; SAS Certified Statistical Business Analyst (Regression and Modeling)
Languages	English (full professional), Hindi (native), Odia (native)

REFERENCES

Dr. Souparno Ghosh	Doctoral advisor
Professor, Department of Statistics, University of Nebraska–Lincoln sghosh5@unl.edu · +1 402 472 2084	
Dr. Yanbin Yin	Research supervisor
Professor, Department of Food Science and Technology, University of Nebraska–Lincoln yyin@unl.edu · +1 402 472 4303	
Dr. Mohammad Rashedul Hasan	Collaborator
Assistant Professor, Electrical and Computer Engineering, University of Nebraska–Lincoln hasan@unl.edu · +1 402 472 5026	