

# Arindam Paul

---

## CONTACT INFORMATION

*E-mail:* arindampaul.bits@gmail.com

*LinkedIn:* linkedin.com/in/arndmpaul/

## INTERESTS

Machine Learning, Deep Learning, Natural Language Processing

## PROGRAMMING SKILLS

Languages: Python (Familiar with MATLAB, R)

ML: Scikit-Learn, H2O, XGBoost/LightGBM/CatBoost (Familiar with Gensim, NLTK)

Deep Learning: Keras/Tensorflow (Familiar with PyTorch/FastAI, HuggingFace)

Model Explainability: Shapley, LIME, imodels, interpretML

Data Analytics: Pandas, Numpy, Dash, Plotly, Seaborn, Matplotlib

Web Development : HTML/Markdown (Familiar with CSS, JavaScript)

## EDUCATION

**Northwestern University**, Evanston, Illinois

Ph.D., Computer Science, Sep 2019

Advisors: Prof. Alok Choudhary, Prof. Ankit Agrawal

**Northwestern University**, Evanston, Illinois

Master of Science, Computer Science, Sep 2014

**Birla Institute of Technology & Science**, Pilani, Rajasthan India

Master of Engineering (Hons.), Software Systems, May 2012

**Birla Institute of Technology & Science**, Pilani, Rajasthan India

Bachelor of Engineering (Hons.), Chemical Engineering, Dec 2009

## PROFESSIONAL EXPERIENCE

*Lead Data Scientist (ML Research Team)*, **American Family Insurance**, Boston, MA **Oct'24-Present**

- Generative AI-based Conversational Intelligence for QA:
  - ◇ Designed effective prompts for Gemini Multimodal models to assess claim quality, leveraging its multimodal capabilities for comprehensive analysis.
  - ◇ Collaborated with the Claims QA team to optimize auditing processes and questionnaires, streamlining workflows and improving efficiency.
- Leadership:
  - ◇ Host of the AmFam Fairness & Ethics in ML Seminar.
  - ◇ Collaborated closely with UW-Madison professors as part of the Amfam Data Science Institute, leading to two publications.

*Senior Data Scientist (ML Research Team)*, **American Family Insurance**, Boston, MA **Feb'24-Oct'24**

- Generative AI for Audio Call Summarization & Entity Extraction:
  - ◇ Contributed to Generative AI pipelines for audio call summarization and entity extraction, ensuring robust performance across formats.
  - ◇ Investigated Whisper using pyannote for speaker identification and used models like Gemini-1.5-Pro, Mistral-Au-Large, and GPT-4-Turbo for summarization.
- Custom Generative AI Chatbot Development:
  - ◇ Implemented a custom Generative AI chatbot leveraging technologies like streamlit, langchain, and GPT-4-Turbo/Claude-Sonnet, enhancing customer experience.

*Data Scientist (ML Research Team), American Family Insurance, Madison/Boston* **Oct'19-Feb'24**

- User-Based Insurance (in collaboration with major US automaker):
  - ◊ Developed generalized linear and additive models based on telematics features for auto insurance in collaboration with a major US automaker.
- Claims-Channeling System:
  - ◊ Co-designed a claims channeling system utilizing transfer learning and performed an ablation study to enhance model accuracy for routing claims to experts.
- Financial Forecasting:
  - ◊ Developed long- and mid-term financial forecasts using an ensemble ESRNN+SARIMA+GBM model, deploying a financial dashboard for leadership.
- Motor Vehicle Violation:
  - ◊ Built an ML decision system to predict motor vehicle violation risk using internal and external data.
- Leadership/Outreach:
  - ◊ Mentored rotational associate data scientists.
  - ◊ Participated as a panelist for company-wide data privacy discussions on fairness, privacy, and bias.

**Northwestern Mutual Life Insurance, Milwaukee, Wisconsin**

*Data Science Intern*

**Jun '18 - Aug '18**

- Developed distributed image to text conversion algorithms for detecting responses from scanned questionnaires
- Designed a noise reduction algorithm to denoise scanned and photocopied questionnaires

**Boeing Cybersecurity, Sunnyvale, California**

*Summer Research Intern*

**Jun '13 - Sep '13**

- Understanding Collaboration Among Online Advertising and Analytics Services
- Observed multiple 3rd-party services sharing user's information with each other
- Investigated how these services use means to obfuscate parameter sharing

RESEARCH  
EXPERIENCE

**Northwestern University, Evanston, Illinois USA**

*Research Assistant*

**Fall '12 - Summer'19**

Ensemble Nets on Mixed Representations for Chemical Property Prediction (Keras) **Sep '18 - Feb '19**

- Created CheMixNet - a deep neural network that combines molecular fingerprints to develop a generalized architecture for chemical property prediction
- Designed SINet - a deep network that combines two different textual representations SMILES and InChI for predicting chemical properties
- Expanded SINet for transfer learning tasks from a 2.3 million dataset to a smaller 350 compound dataset
- Developed ChemensembleNet - a deep network that combines different textual, molecular fingerprints and molecular graph representations of molecules to achieve better results than individual representations

Predictive Modeling for Additive Manufacturing (Tensorflow, Keras)

**Nov '16 -**

- Developed iterative bootstrap tree algorithms for temperature prediction in additive manufacturing processes
- Designed Recurrent Neural Network models to predict point-wise temperature information for accelerating additive manufacturing simulations

Solar Cell Efficiency Prediction using Molecular Fingerprints (Tensorflow, Scikit Learn) **Mar '16 - June '19**

- Designed Deep Neural Network and Random Forest models for predicting power conversion efficiency of solar cells using chemical fingerprints, and achieved mean square percentage error between 1.5-2 %
- Designed an online application for material scientists to get an estimation of power efficiency

Very Deep Neural Networks for Predicting Formation Stability (Tensorflow) **Mar '16 - Sept '17**

- Constructed Neural Network Models with 18-25 layers to predict formation energy of a chemical compound
- Attained 20 % higher accuracy than the state-of-the-art models using Random Forests that would allow domain scientists to explore millions of possible compounds

Ensemble Learning-based Guided Optimization for Aircraft Design (MATLAB, Python) **Oct '15 - Dec '17**

- Created intelligent sampling algorithms to explore the constrained search space for candidate microstructures
- Developed Feature Ranking-based Technique for Search Space Reduction of Constrained Non-Convex Optimization
- Achieved 100x candidate microstructures compared to state-of-the-art methods that can accelerate the design-to-experiment life-cycle

Convolutional Neural Nets for Thematic Image Classification in Pinterest (Torch) **Oct '15 - Sep '16**

- Harnessed Association Rule Mining for thematic label curation
- Developed ConvNet Models for hierarchical classification that led to automated image categorization based on themes

Classification of Anonymous Posts using Recurrent Neural Networks (Tensorflow) **Jan '15 - May '16**

- Generated vectorizer models using Word2vec trained on crowd-sourced (Urban Dictionary) & psycholinguistic (LIWC) dictionaries (Gensim)
- Attained prediction accuracy of 79.8 % and 78.1 % using LSTMs and ensemble models respectively

Facebook Confessions & Yik Yak **Jun '14- Dec '14**

- Studied question asking about sensitive topics in anonymous forums
- Designed a system to automatically identify and classify taboo posts in anonymous forums with good accuracy

Learning from Ads: Reverse-engineering demographics and interests **Feb '13-May '14**

- Created synthetic user profiles with different demographic and interest features and collecting ad traffic
- Created a model by using cross-validation which can predict user features from resulting data-set and ground-truth
- Application of the model to cellular web-data to predict user's demographics and interests on-the-fly

### ***Graduate Researcher***

**Spring '10 - Spring '12**

Preventing Sybil attacks in P2P systems using Psychometric Tests

**Fall '09 - Spring '11**

- Suggested a novel approach to use Psychometric Tests (Luscher Color Test and Myers Briggs Type Indicator Test) to evaluate psychometric index of users
- Cluster nodes with similar scores and in case of a particularly high-frequency zone, we treat these nodes as suspicious and further use CAPTCHAs to remove false positives.

Software Quality Evaluation using Fuzzy Multi-Criteria Approach

**Spring '10- Spring '11**

- Employed fuzzy ratings and weights to software attributes and proposed a comprehensive model for calculating overall software quality based on ISO/IEC 9126 model.
- Tested on multiple university softwares and one industrial application.

### TEACHING EXPERIENCE

**Northwestern University**, Evanston, Illinois USA

*Teaching Assistant*

**Winter '14 -**

Assisted the instructor in teaching the following undergraduate level courses. Duties included sharing of responsibilities for lectures, exams, homework assignments, grades, and office hours.

- EECS 510 Social Media Mining, Spring '17, '18 & '19
- EECS 214 Data Structures and Data Management, Fall 2015
- EECS 110 Introduction to Computer Programming (Python) , Winter & Spring 2014, Winter 2015

*Guest Lecturer*

**Spring '16**

EECS 510 Social Media Mining

- Impact of “likes” and “reactions” on social media
- Anonymity in Social Media
- Crawling and scraping the web

*Instructor*

**Summer '16**

**MGLC Transferable Skills Workshop on Machine Learning**

Attended by McCormick Graduate students and faculty

- What and Why of Machine Learning ?
- Algorithms
- Application
- Introduction to Deep Learning

**BITS Pilani**, Rajasthan, India

*Teaching Assistant*

**Jan - May '12**

Assisted the instructor in teaching the following undergraduate level courses. Duties included sharing of responsibilities for exams, homework assignments, grades and leading computer lab exercises.

- CS/IS 332 Introduction to Database Systems and Application, Spring 2012.

### PUBLICATIONS: JOURNALS

Fa Li, Qing Zhu, Kunxiaojuan Yuan, Fujiang Ji, **Arindam Paul**, Peng Lee, Volker C. Radeloff, Min Chen. “**Projecting large fires in the western US with a more trustworthy machine learning method**”, *Earth's Future*, 2024

D. Borthakur, **A. Paul**, D. Kapil, M.J. Saikia. “**Yoga Pose Classification Integrating Computer Vision and Discriminative Modelling**”, *Health*, 2023

Y. Mao, M. Hasan, **A. Paul**, V. Gupta, K. Choudhary, F.M. Tavazza, W. Liao, A. Choudhary, P. Acar and A. Agrawal. “**An End-to-End AI-Driven Microstructure Optimization Framework for Elastic Properties of Titanium Beyond Cubic Crystal Systems**”, *Nature Partner Journal on Computational Materials*, 2023

A.Dimri, **A.Paul**, D.Girish, P.Lee, S. Afra and A.Jakubowski. “**A Multi-input Multi-label Claims Channeling System Using Insurance-Based Language Models**”, *Expert Systems With Applications*, 2022

Y.Mao, Z.Yang, D.Jha, **A. Paul**, W. Liao, A. Choudhary and A. Agrawal. “**Generative Adversarial Networks and Mixture Density Networks based Inverse Modeling for Microstructural Materials Design**”, *Integrating Materials and Manufacturing Innovation Journal*, 2022

K.Ness, **A. Paul**, L. Sun and Z. Zhang. “**Towards a generic physics-based machine learning model for geometry invariant thermal history prediction in additive manufacturing**”, *Journal of Materials Processing Technology*, 2022 - *Special Issue on AI in Advanced Manufacturing*

R.Richards, and **A. Paul**. “**An Attention-driven LSTM Network for High Throughput Virtual Screening of Organic Photovoltaic Candidate Molecules**”, *Solar Energy*, 2021

**A. Paul**, W. Liao, A. Choudhary and A. Agrawal. “**Harnessing Psycho-lingual and Crowd-Sourced Dictionaries for Predicting Taboos in Written Emotional Disclosure in Anonymous Confession Boards**”, *Journal of Health Informatics Research*, 2021

**A. Paul**, A. Furmanchuk, W. Liao, A. Choudhary and A. Agrawal. “**Property Prediction of Organic Donor Molecules for Photovoltaic Applications using Extremely Randomized Trees**”, *Journal of Molecular Informatics*, 2019

**A. Paul**, P. Acar, W. Liao, A. Choudhary, V.Sundararaghavan and A. Agrawal. “**Microstructure Optimization with Constrained Design Objectives using Machine Learning-Based Feedback-Aware Data-Generation**”, *Journal of Computational Materials Science*, Apr 2019

D.Jha, L.Ward, **A. Paul**, W. Liao, A. Agrawal, A. Choudhary and C. Wolverton.“**ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition**”, *Nature Scientific Reports*, Nov 2018

M.Mozaffar, **A. Paul**, R. Al-Bahrani, S. Wolff, A. Choudhary, A. Agrawal, K. Ehmann and J.Cao.“**Data-Driven Prediction of the High-Dimensional Thermal History in Directed Energy Deposition Processes via Recurrent Neural Networks**”, *Manufacturing Letters*, Sep 2018

**A. Paul**, P. Acar, R.Liu, W. Liao, A. Choudhary,V.Sundararaghavan and A. Agrawal. “**Data Sampling Schemes for Microstructure Design with Vibrational Tuning Constraints**”, *Journal of American Institute of Aeronautics and Astronautics*, Mar 2018

K Haribabu, C.Hota and **A. Paul** “**GAUR: A Method to Detect Sybil Groups in Peer-to-Peer Overlays**”, *International Journal of Grid and Utility Computing*, 2012

J.S. Challa, **A.Paul**, Y.Dada, V.Nerella, P.R. Srivastava and A.P.Singh “**Integrated Software Quality**

**Evaluation: A Fuzzy Multi-Criteria Approach”, *Journal of Information Processing Systems (JIPS): Korean Information Processing Society, 2011.***

PUBLICATIONS:  
CONFERENCES

**A. Paul, M. Mozaffar, Z. Yang, W. Liao, A. Choudhary, J.Cao and A. Agrawal. “A real-time iterative approach for temperature profile prediction in additive manufacturing processes”, *6th IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2019***

**A. Paul, D.Jha, W. Liao, A. Choudhary and A. Agrawal. “Transfer Learning Using Ensemble Neural Nets for Organic Solar Cell Screening”, *International Joint Conference on Neural Networks, 2019***

**Z.Yang, D. Jha, A. Paul, W. Liao, A. Choudhary and A. Agrawal. “A General Framework Combining Generative Adversarial Networks and Mixture Density Networks for Inverse Modeling in Microstructural Materials Design”, *NeurIPS Workshop on Machine Learning for Engineering Modeling, Simulation and Design, 2020***

**A. Paul, D.Jha, R. Al-Bahrani, W. Liao, A. Choudhary and A. Agrawal. “CheMixNet: Mixed DNN Architectures for Predicting Chemical Properties using Multiple Molecular Representations”, *NeurIPS Workshop on Machine Learning for Molecules and Materials, 2018***

**R. Liu, D. Palsetia, A. Paul, R. Al-Bahrani, D. Jha, W. Liao, A. Agrawal, and A. Choudhary. “Pinterest: A Thematic Label Curation Tool for Large Image Datasets”, *Proceedings of the Workshop on Open Science in Big Data at IEEE Bigdata Conference, 2016.***

**A. Paul, A. Agrawal, W. Liao, and A. Choudhary. “AnonyMine: Mining anonymous social media posts using psycho-lingual and crowd-sourced dictionaries”, *Proceedings of the Workshop on Issues of Sentiment Discovery and Opinion Mining at 22nd Annual ACM Conference on Knowledge Discovery and Data Mining, 2016.***

**J.Birnholtz, N.A.R. Merola, and A. Paul. “Is it Weird to Still Be a Virgin??: Anonymous, Locally Targeted Questions on Facebook Confession Boards”, *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, 2015.***

**A. Paul, Varuni G., J.S. Challa, and Y. Sharma “HADCLEAN: A Hybrid Approach for Data Cleaning Techniques in Data Warehouses”, *Proceedings of the IEEE International Conference on Information Retrieval and Knowledge Management(CAMP),Kuala Lumpur, March,2012***

**J.S. Challa, A.Paul, Y. Dada, V. Nerella, and P.R. Srivastava “Quantification of Software Quality Parameters using Fuzzy Multi-Criteria Approach,” *Proceedings of the IEEE International Conference on Process Automation Control and Computing (PACC) 2011, Coimbatore, July, 2011***

**K Haribabu, A.Paul, and C. Hota “Detecting Sybils in Peer-to-Peer Overlays using Psychometric Analysis Methods,”, *Proceedings of the 25th IEEE International Conference on Advanced Information Networking and Applications(AINA), Singapore, March 2011***

AWARDS & HONORS

- McCormick Dean’s Commendation Fellowship, during 6th year of PhD (2017-2018)

- Predictive Science and Engineering Design Fellowship, during 5th year of PhD (2016-2017)
- Segal Design Cluster Fellowship, during 3rd year of PhD (2014-2015)
- Walter P. Murphy Fellowship, during 1st year of PhD (2012-2013)
- BITS Pilani Merit-cum-Need Scholarship during undergraduate study
- Among 10 doctoral students across Northwestern selected for summer-long Research Communication Workshop, 2016
- All India Rank 1 in BITS HDSAT (admission test for graduate programs at BITS) in Software Systems
- All India Rank 64 & State Rank 9 in National Science Olympiad among more than half million participants during freshmen year of high-school

#### LEADERSHIP

- President and Treasurer, Northwestern Toastmasters Club (2015-16, 2016-17)
- President and Founding Member, Northwestern Creative Writing Club
- President, Northwestern University Cricket Club
- Co-Facilitator, Northwestern Dialogue Group
- Mentor, Brave Initiatives (<http://www.braveinitiatives.com/>)

#### SELECTED COURSE & SIDE PROJECTS

- Developed a Sentiment Analysis Tool to find the most interesting or controversial events at the 2014 Golden Globe Awards from user-Tweets (Python) **Spring '14**
- Developed a web application to track a portfolio of a user's stocks. Used data mining techniques to analyze and predict stock and portfolio performance using historical data. (Perl, SQL) **Fall '12**
- Developed a real-time tool starts an alarm when a designated bus is 'x' (customizable) min away from the closest bus stop by scraping CTA bus tracker webpage (Python). **Fall '14.**