Pengyuan Li

☎(302)-501-0291 ✉ pengyuan@ibm.com 🌐 www.eecis.udel.edu/~pengyuan

RESEARCH INTERESTS

Machine Learning, Biomedical Informatics, Image & Text Mining, Document Analysis

EDUCATION

- **University of Delaware** 2015 – 2021
  Ph.D. in Computer Science Advisor: Prof. Hagit Shatkay
  Dissertation on "Utilizing Image Information for Biomedical Document Classification"
- **Harbin Engineering University** 2011 - 2014
  M.E. in Computer Software and Theory Advisor: Prof. Haiwei Pan
  Dissertation on "Medical Image Retrieval Based on Uncertain Location Graph"
- **Zhengzhou University** 2007 - 2011
  B.E. in Computer Science and Technology

AWARDS & HONORS

- Corporate Special Accomplishment, IBM Research (2023)
- Corporate A-level Accomplishment, IBM Research (2022)
- Frank A. Pehrson Graduate Student Award for Outstanding Computer Science Research, CIS Department, University of Delaware (2021)
- Distinguished Graduate Student Award, CIS Department, University of Delaware (2020)
- Dissertation Fellowship, University of Delaware (2020)
- NSF – ACM CIKM Travel Grant (2018)
- Professional Development Award, University of Delaware (2017, 2018, 2019)
- CLEF Student Travel Grant (2017)
- National Scholarship for Graduate Students, Ministry of Education of China (2013)
- Outstanding Graduates of Zhengzhou University (2011)
- Silver medal, Second ACM-ICPC Henan Province Collegiate Programming Contest (2009)
- First place, Third Programming Contest of Zhengzhou University (2009), etc.

RESEARCH EXPERIENCES

Research Staff Member, **IBM Research - Almaden** 2021-*Present*

- Data Acquisition Lead for developing large language models (~10PT data collected) (***Corporate Special Accomplishment***)
- Large-scale scientific data preprocessing
- Creating data cards for understanding the datasets used for large language model training
- Search engine for matching client requirements with business products (***Corporate A-level Accomplishment***)
- Business document analysis for information extraction and understanding
- Question-Answering system for automatic responding to clients' requirements

Collaborator, **Sternberg Lab, Caltech** 2022-*Present*

- Image manipulation detection for biomedical literature
- Machine learning for accelerating the biocuration process

Collaborator, **Electronic Visualization Laboratory, University of Illinois at Chicago** 2021-*Present*

- ✧ Image search engine for retrieving figures within COVID literature

Research Assistant, **Computational Biomedicine Lab, University of Delaware** Sep 2015-Aug 2021

- ✧ Biomedical document classification utilizing image and text information
- ✧ Figure and caption extraction from biomedical documents (www.eecis.udel.edu/~compbio/FDFigCapX)
- ✧ Compound image separation of published figures (www.eecis.udel.edu/~compbio/FigSplit)
- ✧ Biomedical image classification for supporting the bio-image annotation process
- ✧ Heart disease detection using ECG signals and ultrasound images

Research Intern, **IBM Research – Almaden**, San Jose, USA Jun 2019-Aug 2019

- ✧ Customer review analysis and topic detection

Visiting Student, **Robotics and Control Lab, The University of British Columbia** May 2018-Aug 2018

- ✧ Analysis of ultrasound images for heart disease detection

Research Assistant, **Intelligent Information Processing Center, HEU** Sep 2011-Jun 2015

- ✧ Brain CT image retrieval using an uncertain location graph model
- ✧ Brain CT image classification based on symmetry and content features

Visiting Student, **Fan Lab, David Geffen School of Medicine, UCLA** Sep 2013-Dec 2013

- ✧ Research on chromosome image analysis and gene sequence analysis

Visiting Student, **Stem Cell Lab, School of Medicine, Tongji University** Sep 2012-Feb 2013

- ✧ Research on colored cell image analysis
- ✧ Core algorithm development for colored sperm cell detection and quality evaluation

Lab Member, **ACM-ICPC Lab, Zhengzhou University** Mar 2009-Apr 2010

TEACHING EXPERIENCES

Adjunct Faculty, **Data Science Institute, UD** 2023-*Present*

- ✧ Class design for BINF601: Introduction to Data Sciences
- ✧ Provided lectures and practices about biomedical image analysis

Research Advisor, **UCSC** HCI271: Human-Computer Interaction Capstone Spring & Fall 2023

- ✧ Provided research insights about training Large Language Models (LLMs)
- ✧ Coordinate with students for developing a user-friendly LLM training platform

Intern Mentor, **IBM Research - Almaden** Summer 2022

- ✧ Mentored two PhD students for their summer intern projects
- ✧ Collaborated with interns to conceptualize and submit innovative papers and patents

Teaching Assistant, **UD** CISC436/636: Computational Biology and Bioinformatics Fall 2019

- ✧ Held office hours for graduate and undergraduate students, graded assignments and exams

PUBLICATIONS

[1] **IBM Research**. Granite Foundation Models. https://www.ibm.com/downloads/cas/X9W4O6BM.

[2] Nezamabadi K, Sivalokanathan S, **Li P**, Lee J, Chen M, Lu D, Abraham J, Sardaripour N, Mousavi P, Abraham MR. XplainScar: Explainable artificial intelligence to identify and localize left ventricular scar in hypertrophic cardiomyopathy from 12-lead electrocardiogram. [J] **Nature Biomedical Engineering**. (In submission)

[3] **Li P**, Ren G, Gentile AL, DeLuca C, Tan C. Long-form information retrieval for enterprise matchmaking. **ACM SIGIR 2023**. (Accepted)

[4] Gentile AL, Shbita B, DeLuca C, **Li P**, Ren G. Understanding Customer Requirements - an Enterprise Knowledge Graph Approach. **ESWC 2023**. (Accepted)

[5] Zhang Z, **Li P**, Jin G, Wang J. DAUF: An Attention-Based UNet Framework for Identifying Progressive and Stable Mild Cognitive Impairment Associated with Disease. [J] **Computers in Biology and Medicine**. (Accepted)

[6] Nezamabadi K, Mayfield J, **Li P**, Greenland GV, Rodriguez S, Simsek B, Mousavi P, Shatkay H, Abraham MR. Toward ECG-based analysis of hypertrophic cardiomyopathy: a novel ECG segmentation method for handling abnormalities. [J] **Journal of the American Medical Informatics Association**, 2022, 29(11), 1879–1889.

[7] Bian X, Pan H, Zhang K, **Li P**, Li J, Chen C. Skin lesion image classification method based on extension theory and deep learning. [J] **Multimedia Tools and Applications**, 2022, 81(12), 16389-16409.

[8] **Li P,** Jiang X, Zhang G, Trabucco JT, Raciti D, Smith C, Ringwald M, Marai GE, Arighi C, Shatkay H. Utilizing image and caption information for biomedical document classification. [C] In the Proceedings of the joint conference on Intelligent Systems for Molecular Biology and European Conference on Computational Biology (ISMB/ECCB2021).
Also in [J] ***Bioinformatics***, 2021, 37(S1), i468-i476.

[9] Trabucco JT, **Li P**, Arighi C, Raciti D, Shatkay H, Marai GE. ANIMO: Annotation of biomed image modalities. [C] *In Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine* ***(BIBM2021)***, 1069-1076.

[10] Jiang X, Li P, Kadin JA, Blake JA, Ringwald M, Shatkay H. Integrating image caption information into biomedical document classification in support of biocuration. [J] Database, 2020, 2379-2385.

[11] Trabucco JT, **Li P**, Arighi C, Shatkay H, Marai GE. Modality-classification of microscopy images using shallow variants of deep networks. [C] *In Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine* ***(BIBM2020)***, 2379-2385.

[12] **Li P**, Jiang X, Shatkay H. Extracting figures and captions from biomedical documents. [J] ***Bioinformatics***, 2019, 35(21), 4381-4388.

[13] **Li P**, Jiang X, Kambhamettu C, Shatkay H. Compound image segmentation of published biomedical figures. [J] ***Bioinformatics***, 2018, 34(7), 1192-1199.

[14] **Li P**, Jiang X, Shatkay H. Figure and caption extraction from scientific documents. [C] *In Proceedings of the 27th ACM International Conference on Information and Knowledge Management* ***(CIKM2018)***, 1595-1598.

[15] Zhang G, Roychowdhury D, **Li P**, Wu HY, Zhang S, Li L, Shatkay H. Identifying experimental evidence in biomedical abstracts relevant to Drug-Drug Interactions. [C] *In Proceedings of the 9th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* ***(BCB2018)***, 414-418.

[16] **Li P**, Jiang X, Kambhamettu C, Shatkay H. Segmenting compound biomedical figures into their constituent panels. [C] *In Proceedings of the 8th Cross-Language Evaluation Forum for European Languages* ***(CLEF2017)***, 199-210. ***(Best of Lab paper track)***

[17] Li W, Pan H, **Li P**, Xie X, Zhang Z. A medical image retrieval method based on texture block coding

tree. [J] **Signal Processing: Image Communication**, 2017, 59, 131-139.

[18] Zhang G, Bhattacharya M, Wu HY, **Li P**, Li L, Shatkay H. Identifying articles relevant to Drug-Drug Interaction: Addressing Class Imbalance. [C] *In Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine* ***(BIBM2017)***, 1141-1147.

[19] Gao L, Pan H, Han Q, Xie X, Zhang Z, Zhai X, **Li P**. Finding frequent approximate subgraphs in medical image database. [C] *In Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine* ***(BIBM2015)***, 1004-1007**.**

[20] Pan H, **Li P**, Li Q, Han Q, Feng X, Gao L. Brain CT image similarity retrieval method based on Uncertain Location Graph**.** [J] ***IEEE Journal of Biomedical and Health Informatics***, 2014, 18(2):574-584.

[21] **Li P**, Pan H, Li J, Han Q, Xie X, Zhang Z. A novel model for medical image similarity retrieval. [C] *In Proceedings of the 14th Conference on Web-Age Information Management* ***(WAIM2013)****,* 595-606.

[22] Wang R, Pan H, Han Q, Gu J, **Li P**. Medical Image Retrieval Method Based on Relevance Feedback. [C] *In Proceedings of the 8th International Conference on Advanced Data Mining and Applications (****ADMA2012****),* 650-662.

PATENTS

[1] **Li P**, Ren G, Huang L, Gentile AL. Generation of graphical icons for taxonomy nodes. (Filed)

[2] **Li P**, Ren G, Cai L, Moore R, Tan D. Generating diagrams for visualizing structured documents. (Filed)

[3] Moore R, Ren G, Tan C, Lee A, **Li P**. Navigation guide using different vehicle components. (Filed)

[4] Pan H, **Li P**, Feng X, et al. Patent: Medical Image Similarity Retrieval Method Based on Uncertain Location Graph. Publication Number: CN103226582A.

SERVICE & ACTIVITIES

Journal Reviewer:

Bioinformatics | Bioinformatics Advances | PeerJ Computer Science | Multimedia Tools and Application | MicroPublication Biology | Applied Sciences | Big Data and Cognitive Computing |

PC member / Conference Reviewer:

SIGKDD 2023 | AMIA 2023 | ISMB/ECCB 2023 | WWW 2022, 2023, 2024 | BIBM 2020, 2021 (Session Chair), 2022, 2023 | RECOMB 2020 |SIGIR 2024 |

Organizing Committee:

IBM Almaden Spirit Team (Academic talks, social events, and return-to-work activities organization)