Akram Vasighi

Ontario, Canada | +1 (226) 787-1588 | vasighizaker@gmail.com |GitHub: https://www.github.com/vasighiz |
LinkedIn: https://www.linkedin.com/in/Vasighi

Summary

- 6 years of research experience in developing models in machine learning and deep learning, specialized in data analysis.
- 4+ years of professional experience in software developer and data analysis.
- Strong proficiency in Python, R, SQL, and data science domain tools.
- Hands-on experience in designing transformer-based systems, Retrieval-Augmented Generation (RAG), LLM and conversational frameworks.
- Proven track record of publication and presentation in machine learning, leveraging GPU-accelerated environments (HPC).
- Passionate about leveraging cutting-edge LLMs to solve real-world challenges and eager to gain hands-on experience in their practical applications.

Technical Skills

Programming Languages: Python, R, C++, C#, SQL
Machine Learning & AI: PyTorch, Tensorflow , Pandas, NumPy, Scikit-learn, Natural Language Processing (NLP) & Large Language Models (LLMs), Generative AI, Retrieval-Augmented Generation (RAG), Agentic systems, Graph Neural Networks, Fine-tuning LLMs
Frameworks & Tools: DevOps, Hugging Face Transformers, LangChain, Streamlit, FastAPI
Others: AWS, Agile software development, Docker, HPC/GPU Clusters , Git, Linux, REST APIs, Shell Scripting (Bash), Data Pipeline & ETL, SQL Server (SSMS, SSAS, SSIS, SSRS)

Professional Experience

Postdoctoral Fellow, *School of Computer Science & Biomedical Science*
University of Windsor, ON, Canada | **Jul 2024 – Present**
- Developed scalable analysis pipelines for large-scale datasets (CRISPR screen and time-series data) using machine learning and deep learning models.
- Utilized LLMs to assist in scientific finding interpretation, improving research productivity.

- Preprocess and clean large data sets to generate reports and visualizations.

- Executed workflows on GPU clusters ensuring efficient data processing and visualization for collaborators.

AI/ML Developer (*Remote*), InfoSys **Nov 2023 – Present**
- Designed and implemented advanced LLM-powered applications focusing on RAG-based retrieval systems and conversational AI assistants specialized in scientific domains.
- Developed full-stack solutions integrating Hugging Face transformers, vector search (FAISS), and

interactive web UIs (Streamlit).

Research Assistant, *University of Windsor* **Jan 2020 – Apr 2024**

- Led research projects on deep learning models including graph neural networks for large-scale NGS data analysis.
- Addressed challenges such as data sparsity and high dimensionality via embedding methods and model optimization.
- Co-advised graduate and undergraduate students, mentoring in machine learning and bioinformatics.

Graduate Assistant, *University of Windsor* **Jan 2020 – Apr 2024**

- Co-advised grad and undergrad students and helped them to excel in their courses (OO programming using Java, Data Structure and Algorithms, Internship project, Advanced Computing Concepts).

Sessional Instructor, *University of Windsor* **Jan 2023 – Dec 2023**

- Taught courses on Computer Networks and AI for Games.

Software Developer & Data Analyst, *GS1 Iran, Tehran, Iran* **May 2018 – Nov 2019**

- Developed a nationwide web-based data analysis platform automating ETL and reporting pipelines with Python and SQL.
- Integrated NLP for product description auto-suggestion and optimized user experience across multiple platforms.
- Led customer requirement analysis, application maintenance, and database administration using SQL Server.
- Created complex analytical models using SSMS, SSAS, SSIS, SSRS to support supply chain management.

Data Analyst, *Central Insurance Research Centre, Tehran, Iran* **Jan 2019 – Dec 2019**

- Performed data mining and visualization (PowerBI) on customer datasets to improve marketing strategies, resulting in a 70% increase in income.

Data Analyst, *University of Tehran, Tehran, Iran* **Jan 2018 – Dec 2018**

- Conducted thorough data validation and cleaning to support research data quality.

Key Projects

LLM-Based Question Answering System (RAG-Powered)

- Developed a RAG pipeline using Python that combines: Document processing with LangChain, Semantic search using Hugging Face Transformers and FAISS, Local LLM inference with CTransformers (Mistral 7B), Interactive UI with Streamlit.

Domain-Specific Scientific Chatbot (Single-cell RNA seq Data Analysis)

- Fine-tuned GPT-2 transformers using LoRA and PEFT for genomics datasets to create a specialized conversational assistant.
- Automate NGS data analysis and enabled researchers to query and interpret genomic analysis

results via natural language.
- Implemented efficient few-shot learning and real-time inference optimizations.

Text Data Mining and Topic Modeling of Historical Newspapers Using JupyterHub
Collaborated in an interactive workshop series focused on mining and analyzing digitized historical newspaper data within a JupyterHub environment. Utilized Python tools including NLTK, Whoosh, Pandas, and Non-negative Matrix Factorization (NMF) to preprocess OCR-generated text, perform sentiment analysis, and uncover latent topics. Employed OCR text indexing and search to extract relevant snippets from historical newspapers (Essex County's *Amherstburg Echo*, 1874–2012). Applied sentiment analysis using VADER lexicon to gauge polarity across historical text segments. Implemented NMF-based topic modeling to identify key themes and trends in newspaper content spanning 1920–1930. Interpreted topic clusters by analyzing high-weight terms and corresponding documents, linking findings to historical socio-economic contexts. Combining text mining and machine learning for historical data research in a collaborative JupyterHub setting.

House Price Prediction using ML
Developed a real-world data science project focused on predicting housing prices. Addressed challenges of working with authentic, industry-relevant datasets and implemented various regression models using state-of-the-art tools such as PyCaret to optimize accuracy and speed. Deployed the final model as an interactive web application using Streamlit, showcasing end-to-end data analysis, modeling, and productization skills

Apple Stock Price Prediction Using Deep Learning
Designed and implemented a practical deep learning project to predict Apple Inc.'s stock prices using real historical data sourced from Yahoo Finance. Covered core deep learning architectures and applied advanced feature extraction and data preprocessing techniques. Delivered an end-to-end solution encompassing model building, evaluation, and result interpretation to enhance investment decision-making.

Face Image Generation Using Generative Adversarial Networks (GANs)
Led a hands-on project focused on generating realistic human face images using GAN-based generative AI techniques. Leveraged a celebrity face dataset to train GAN models comprising generator and discriminator networks. Provided deep insights into the architecture, training procedures, and optimization strategies of GANs.

Literary Text Generation Using Recurrent Neural Networks (RNNs)
Conducted a hands-on project to generate Shakespearean-style literary text using advanced generative AI techniques. Utilized a dataset of Shakespeare's works to train RNN-based models, introducing participants to fundamental and advanced recurrent architectures. Explored the core structure and applications of RNNs in text generation. Covered enhanced layers such as LSTM and GRU, and their advantages over traditional RNNs. Investigating practical applications including artistic writing, storytelling, and chatbot content creation. This project fostered deep understanding of sequential generative models and creative text synthesis using RNNs.

Discovering Cell Types Using Manifold Learning and Enhanced Visualization of Single-Cell RNA-Seq Data
Developed a computational method combining nonlinear dimensionality reduction techniques with clustering algorithms to identify representative cell type clusters in high-dimensional, sparse single-cell RNA-seq (scRNA-seq) data. Evaluated the approach on thirteen diverse publicly available scRNA-

seq datasets, demonstrating that combining modified locally linear embedding (LLE) with independent component analysis (ICA) outperforms existing unsupervised methods. Performed gene set enrichment analysis to validate cluster biological relevance, advancing disease module identification through precise cell type characterization.

SEGCECO: Subgraph Embedding of Gene Expression Matrix for Cell-Cell Interation Prediction
Developed SEGCECO, a novel attributed graph convolutional neural network method to predict cell–cell interaction from single-cell RNA-seq data by embedding local subgraphs derived from gene expression profiles. Addressed challenges of high-dimensional, sparse data by applying similarity-based optimization (SoptSC) to construct reliable cell–cell communication networks. Validated on six human and mouse pancreas datasets, SEGCECO achieved superior performance over state-of-the-art link prediction methods, with 0.99 ROC and 99% accuracy.

Cell Type Annotation Model Selection: General-Purpose vs. Pattern-Aware Feature Gene Selection in Single-Cell RNA-Seq Data
Compared and evaluated state-of-the-art supervised models, including XGBoost and Support Vector Machines (SVM) with information gain feature selection, for automated cell type annotation in single-cell RNA-seq datasets. Demonstrated that XGBoost provides a scalable and effective framework for cell type identification, outperforming traditional methods. Highlighted the potential of combining boosting tree algorithms with deep neural networks to improve marker gene identification and biological insights from scRNA-seq data.

Unsupervised Identification of SARS-CoV-2 Target Cell Groups via Nonlinear Dimensionality Reduction on Single-Cell RNA-Seq Data
Applied advanced unsupervised learning methods combining nonlinear dimensionality reduction (modified Locally Linear Embedding and Independent Component Analysis) with clustering to identify representative lung cell clusters targeted by SARS-CoV-2. Conducted comprehensive preprocessing including normalization and quality filtering on large-scale COVID-19 scRNA-seq datasets. Validated findings by identifying immune-related target cell types and overlapping marker genes linked to COVID-19, Influenza A, and HSV-1 infections, providing insights into disease mechanisms and potential therapeutic targets. Led literature review, data curation, algorithm design, implementation, and result presentation in collaboration with Mitacs and University of Windsor.

Cell Type Identification Using Convolutional Neural Networks and Self-Organizing Maps on Single-Cell RNA-Seq Data
Developed a two-step supervised learning approach combining convolutional neural networks (CNNs) and self-organizing maps (SOMs) for automatic cell type identification in high-dimensional single-cell RNA-seq data. Applied unsupervised feature selection to identify key genes, achieving an average prediction accuracy of 98% across six human pancreas cell types. Biological validation confirmed that the majority of selected genes are relevant markers, demonstrating the method’s effectiveness in accelerating cell type annotation.

Identifying Therapeutically Targetable Tumor-Immune Interactions in Small Cell Lung Cancer
Applied machine learning and single-cell RNA sequencing (scRNA-seq) analysis to investigate tumor heterogeneity and uncover novel therapeutic targets in small cell lung cancer (SCLC). Leveraged published datasets to identify key biomarkers predictive of tumor-immune interactions and gene expression changes across immune and epithelial subtypes. Conducted pathway analysis and literature validation to pinpoint genes such as RBP1 and CD74 with strong protective roles. This

research provides molecular insights to guide pre-clinical validation and supports the development of stage-specific therapeutic strategies with potential clinical translation.

C-PUGP: Cluster-Based Positive Unlabeled Learning for Disease Gene Prediction and Prioritization
Developed a novel semi-supervised machine learning method to identify and prioritize candidate disease genes despite the challenge of lacking reliable negative samples. Introduced a three-step approach combining clustering of positive data, one-class classification to generate a reliable negative set, and SVM-based binary classification for gene ranking. Achieved significant performance improvements with precision, recall, and F-measure exceeding 92%, outperforming existing methods by over 11% in F-measure and enhancing gene prioritization accuracy by approximately 6%.

One-Class Classification Approach for Accurate Prediction of Disease-Gene Associations in Acute Myeloid Leukemia (AML)
Developed a novel one-class SVM method focused on precisely identifying disease-causing genes in AML by treating the problem as a one-class classification task. Unlike traditional binary classifiers, this approach emphasizes high sensitivity and precision in detecting known disease genes while minimizing false negatives. Evaluated on an AML gene expression benchmark dataset, the method demonstrated superior performance in precision, recall, and F-measure compared to existing techniques. The model and benchmark dataset are publicly available, promoting reproducibility and further research.

Canadian Medicinal Plant Detection via Transfer Learning
Utilized convolutional neural networks with transfer learning to classify plant images. Leveraged transfer learning on the InceptionV3 convolutional neural network architecture to efficiently train the model on a limited dataset of plant images. Collected and curated a high-quality image dataset representing various medicinal plant species native to Canada. Applied transfer learning to fine-tune a pre-trained InceptionV3 model, reducing the need for large-scale data and computational resources. Implemented data augmentation techniques to enhance model robustness and generalization. Evaluated model performance using standard metrics such as accuracy, precision, recall, and F1-score.

Online Data Visualization and Analysis Tool
Designed and implemented a comprehensive data visualization and analysis platform tailored for the insurance industry, leveraging ETL (Extract, Transform, Load) processes and OLAP (Online Analytical Processing) systems. Designed OLAP cubes to facilitate multi-dimensional data analysis, enabling users to perform dynamic slicing, dicing, and drill-down operations. Created interactive PowerBI dashboards with rich visualizations (charts, maps, KPIs) to highlight key insurance metrics such as customer segmentation, claims trends, and risk assessment. Enabled non-technical stakeholders to explore and interpret data easily, improving operational efficiency and strategic planning. Collaborated with domain experts to tailor analytics solutions addressing industry-specific challenges and regulatory requirements.

Education

Ph.D. in Computer Science, *University of Windsor, ON, Canada* **Jan 2020 – Apr 2024**
Thesis: Advanced Machine Learning Methods to Analyse Single-cell RNA-seq Data

M.Sc. in Computer Engineering, *University of Tehran, Iran* **Sep 2015 – Jan 2017**

Thesis: C-PUGP: A Cluster-Based Positive Unlabeled Learning Method for Disease Gene Prediction and Prioritization.

Publications

[publications accessible via Google scholar profile *(https://scholar.google.com/citations?user=mJSJoqIAAAAJ&hl=en)*.]

1. **A. Vasighizaker,** S. Hora, R. Zeng, L. Rueda, "SEGCECO: Subgraph Embedding of Gene expression matrix for CEll cell COmmunication prediction", Briefings in Bioinformatics, (**2024)**
2. **Vasighizaker, A**.; Trivedi, Y. Rueda, L., "Cell Type Annotation Model Selection: General-purpose vs Pattern-aware Feature Gene Selection in Single-cell RNA-seq Data". MDPI GENES (**2023**)
3. **Vasighizaker, A**., Danda, S., & Rueda, L. "Discovering cell types using manifold learning and enhanced visualization of single-cell RNA-Seq data". Scientific Reports, Nature Portfolio, (**2022**).
4. M. Naik, L. Rueda, **A. Vasighizaker**, "Identification of Enriched Regions in ChIP-seq Data via a Linear-time Multi-level Thresholding Algorithm", IEEE/ACM Transactions on Computational Biology and Bioinformatics, **(2021)**.
5. **A.Vasighizaker**, A. Sharma, A. Dehzangi. "A novel one-class classification approach to accurately predict disease-gene association in acute myeloid leukemia cancer", PlosOne Journal, 14 (12), **(2019).**
6. **A.Vasighizaker**, S. Jalili. "C-PUGP: A Cluster-Based Positive Unlabeled Learning Method for Disease Gene Prediction and Prioritization**",** Journal of Computational Biology and Chemistry, (**2018).**

Conference & Presentations

1. Nakul Pandya, Raymond Zeng, Biren Dave, **Akram Vasighizaker**, Swati Kulkarni, Ming Pan, Junaid Yousuf, Luis Rueda, "Identifying Therapeutically Targetable Tumor-Immune Interactions in Small Cell Lung Cancer", WE-SPARK's Health Research, Canada, **2025**.
2. **A. Vasighizaker,** S. Hora, L. Rueda, "Exploring Cell-Cell Communication in Pancreas Tissue via Attributed-graph Convolutional Neural Networks on Single-cell RNA-sequencing Data", ISMB/ECCB **2023**, Lyon, France
3. A. **Vasighizaker**, S. Hora, L. Rueda, "Unravelling the Complexity of Cellular Interactions using Underlying Graph Representations of Single-Cell Transcriptomics Data", GLBio **2023**, Montreal, Canada
4. A. **Vasighizaker**, S. Danda, Luis Rueda, "Discovering cell types using manifold learning and enhanced visualization of single-cell RNA-Seq data", Highlight paper presentation in ECCB **2022**, Barcelona.
5. A. **Vasighizaker**, S. Hora, L. Rueda, "A Novel Method to Predict Intercellular Signaling in Single-cell RNA-seq Data via Graph Convolutional Network", ISMB **2022**, Madison, USA
6. A. **Vasighizaker**, S. Danda, Luis Rueda, "Discovering cell types using manifold learning and enhanced visualization of single-cell RNA-Seq data", Highlight paper presentation in ACM/BCB **2022**, Chicago, IL, USA
7. S. Hora, A. **Vasighizaker**, L. Rueda, "SEGCECO: Subgraph Embedding of Gene expression matrix for CEll cell COmmunication prediction", RECOMB **2022**, La Jolla, USA
8. A. **Vasighizaker**, S. Hora, Y. Trivedi, L. Rueda, "Supervised Cell Type Heterogeneity Detection in Single-cell RNA-seq Data", in IWBBIO conference (virtual) **2022**.
9. A. **Vasighizaker**, S. Danda, G. Peralta Milla, Luis Rueda, "Cell Type Identification Single-cell RNA-Seq Data via Modified Locally Linear Embedding" in RECOMB conference (virtual) **2021**.
10. A. **Vasighizaker**, L. Zhou, L. Rueda, "Cell Type Identification via Convolutional Neural Networks and Self-Organizing Maps on Single-Cell RNA Sequencing Data" in ACM/BCB conference (virtual) **2021**.
11. A. **Vasighizaker**, L. Zhou, L. Rueda, "Prediction of Human Pancreas Cell Types via ConvNet on Two-

dimensional Mapping of Single-cell RNA-seq Data" in ISMB/ECCB conference (virtual) **2021**.

12. A. **Vasighizaker**, S. Hora, A. Nagarajan, Y. Trivedi and L. Rueda, "Supervised Cell Type Heterogeneity Detection in Single-cell RNA-seq Data", ICML **2021**, virtual.

S. Danda, A. **Vasighizaker**, L. Rueda, "Unsupervised Identification of SARS-CoV-2 Target Cell Groups via Nonlinear Dimensionality Reduction on Single-cell RNA-Seq Data" in IEEE BIBM conference, Seoul, South Korea, (virtual) **2020**

Delivered Workshop on "Introduction to Generative AI: Concepts, Architectures, and Applications"

Conducted a comprehensive 4-hour online workshop introducing core concepts and architectures of Generative AI, targeting data science and AI practitioners with programming and deep learning experience. Covered fundamental differences of Generative AI from other AI paradigms and detailed key models including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Deep Boltzmann Machines (DBMs). The workshop empowered participants to understand foundational theory and practical relevance of Generative AI, preparing them for advanced exploration in this dynamic AI field.

Delivered Workshop on Applied Machine Learning with Python

Conducted a comprehensive series of workshops introducing core machine learning concepts and hands-on Python programming. Topics included data preprocessing with Pandas, implementing classification models using Scikit-learn (e.g., k-NN and Decision Trees), and performance evaluation through confusion matrices, cross-validation, and precision-recall curves. Participants gained practical experience in model training, hyperparameter tuning, and evaluation using Jupyter Notebooks on Google Colab. The workshop concluded with a comparative overview of machine learning platforms including Microsoft Azure ML Studio and Weka. Workshop materials and code were provided via GitHub to facilitate active participation and post-workshop learning.

Honors & Awards

- Mitacs Research Training Award (RTA), 2020
- Best Paper Award, Conference on Computer and IT, Tabriz, Iran, 2018
- Entrance Scholarship, MSc, University of Tehran, 2017
- High score (94.5%) in SAGE Program on University Teaching and Learning in STEM

Other Activities

- Guest Speaker, University of Detroit Mercy (2021, 2022)
- IEEE Member since 2020
- Reviewer for Scientific Reports (Nature Portfolio), BIBM 2021 Conference and IEEE/ACM Transactions on Computational Biology and Bioinformatics
- Mentorship of graduate and undergraduate students in machine learning and bioinformatics projects