

# Samiha Mirza

Samiha.mirza1234@gmail.com | 346-599-6417 | Houston, TX | [in/samiha-mirza](#) | [orcid/samiha](#)

## Data Scientist | Applied ML | PhD Candidate in CS (May 2026)

Data Scientist and PhD candidate with a strong background in applied machine learning, statistical modeling, and data analysis for real-world, high-variability datasets. Experienced in designing experiments, building end-to-end ML pipelines, and evaluating predictive models under distribution shift, noise, and operational constraints. Comfortable working across the ML stack, from data preparation and modeling to analysis and deployment—while collaborating with cross-functional teams to deliver actionable insights.

## WORK EXPERIENCE

<b>Shell</b> <i>AI Research Intern</i>	<b>June 2024 – Aug. 2024</b> <i>Houston, TX</i>
<ul style="list-style-type: none"><li>Developed and evaluated machine learning models for prediction, anomaly detection, and pattern discovery on large-scale operational datasets.</li><li>Built end-to-end ML pipelines including preprocessing, feature engineering, model training, hyperparameter tuning, and evaluation.</li><li>Applied representation learning techniques to extract meaningful embeddings from complex, high-dimensional data.</li><li>Designed controlled experiments and ablation studies to quantify model impact and trade-offs.</li><li>Collaborated with engineers and stakeholders to translate ML outputs into actionable insights and decision-support tools.</li></ul>	

<b>Shell   Quantitative Imaging Lab, University of Houston</b> <i>Research Fellow   Aug. 2022 – Present</i>	<b>Aug 2022 – Present</b> <i>Houston, TX</i>
<ul style="list-style-type: none"><li>Led applied ML research focused on representation learning, generalization, and robustness across large, heterogeneous datasets.</li><li>Built and evaluated deep learning models (CNNs, autoencoders) for learning transferable embeddings under distribution shifts.</li><li>Designed experimental frameworks to study model behavior under distribution shift, noise, and data imbalance—skills directly applicable to ranking and recommendation systems.</li><li>Conducted extensive exploratory data analysis to understand user-like data variability, failure modes, and performance bottlenecks.</li><li>Developed reproducible ML pipelines using Python, PyTorch, and experiment tracking tools.</li><li>Published peer-reviewed research emphasizing evaluation-driven, data-centric approaches to robust perception models.</li></ul>	

<i>Teaching Assistant   Aug. 2022 – Spring, 2024</i>	
<ul style="list-style-type: none"><li>Taught <i>Digital Image Processing</i>, <i>Data Structures</i>, and <i>Computer Architecture</i>; mentored 100+ students on algorithm design and performance-aware coding practices.</li></ul>	

## EDUCATION

<b>University of Houston</b> <i>PhD, Computer Science</i>	<b>Aug 2022 – May 2026</b> <i>Houston, TX</i>
<b>University of Houston</b> <i>MS, Computer Science</i>	<b>June 2025</b> <i>Houston, TX</i>
<b>University of Dammam</b> <i>Bachelor's, Computer Science (Summa Cum Laude)</i>	<b>June 2022</b> <i>Saudi Arabia</i>

## SKILLS & COMPETENCIES

- Technical Skills & AI Modeling:** Python, PyTorch, Linux, deep learning, feature engineering, data visualization & analysis, experimental design, ablation studies, time-series modeling, statistical analysis, model validation, A/B testing, anomaly detection
- Professional:** Cross-disciplinary collaboration, scientific writing, reproducible research, agile development, scientific writing

## PROJECTS

<b>Shell</b> <i>Data-Centric AI for subsalt interpretation Modeling under Domain Shift</i>	<b>Aug 2022 – Present</b>
<ul style="list-style-type: none"><li>Designed experiments to test how instrumentation, data coverage, and feature changes influenced downstream performance metrics. Integrated multi-domain data to improve cross-dataset generalization.</li><li>Used statistical analysis and visualization to identify trends, failure modes, and improvement opportunities.</li></ul>	
<i>Graph-based Application Dependency Mapping</i>	
<ul style="list-style-type: none"><li>Mapped application dependencies across shared databases using machine learning and graph-based analytics to optimize IT resource allocation.</li></ul>	
<i>Risk Categorization and Mitigation Model</i>	
<ul style="list-style-type: none"><li>Designed predictive ML models for operational risk detection using historical, time-dependent data. Focused on model stability, interpretability, and decision relevance.</li></ul>	