Yiming Shen

yiming.shen001@umb.edu | (774) 578-4554 | Boston, MA (**Open to Relocate**)
GitHub: github.com/yiming-shen | **LinkedIn**: linkedin.com/in/yiming-shen | **Web**: yimings.com.cn

Ph.D. Candidate in Computational Science integrating **Computer Science** engineering standards with **Statistical Learning**. Specialized in building robust, reproducible analytics pipelines for complex time-series data. Experienced in developing public R/Python software packages that address real-world data quality issues like distribution shifts. Combining formal CS training with statistical rigor to deliver production-ready code and reliable model deployment strategies.

TECHNICAL SKILLS

- **Languages: R** (Advanced Package Dev, Rcpp), **Python** (Scikit-Learn, NumPy, PyTorch), **SQL** (MySQL Optimization).
- **Statistical Methods:** Multivariate Statistics, Dimensionality Reduction (PCA, Tensor Decomposition), Hypothesis Testing, Regularized Regression (GLM/ElasticNet), Mixed-Effects Models.
- **Machine Learning:** Domain Adaptation, Transfer Learning, Time-Series Forecasting, Ensemble Methods, Model Evaluation (Nested CV).
- **Computing & Tools:** AWS (Cloud Computing), HPC (Slurm Cluster), Docker, Git, CI/CD (GitHub Actions), Linux.

EDUCATION

University of Massachusetts Boston — **Boston, MA**
Ph.D. in Computational Science – Data Analytics Track — *Expected May 2026*

Rochester Institute of Technology (RIT) — **Rochester, NY**
M.S. Coursework in Computer Science (Transferred to PhD program) — *Aug 2018 – May 2019*

Valparaiso University — **Valparaiso, IN**
M.S. in Analytics and Modelling — *May 2018*

Chongqing University — **Chongqing, China**
B.Eng. in Electrical and Electronics Engineering — *Jun 2014*

OPEN SOURCE SOFTWARE & TOOLS

Lead Developer (Multivariate Time-Series Library) — **2021 – Present**
GitHub (Public) / R

- **Reproducible Pipeline Design:** Architected an object-oriented transformation engine that strictly separates training parameters from testing application. This design prevents **data leakage** and ensures consistent feature generation between historical data and future inference streams.
- **Heterogeneous Data Standardization:** Developed a unified API to harmonize diverse public benchmarks (e.g., Physionet, BNCI). This abstraction allows datasets with varying structures to be processed via a single, standardized interface.
- **Numerical Stability:** Incorporated regularization techniques (epsilon stabilization) to automatically handle singular matrices and noisy input data, preventing runtime failures in automated workflows.

Core Developer (DA4BCI - Domain Adaptation Framework) — **2022 – Present**
GitHub (Public) / R

- **Handling Data Shifts:** Developed a library of Transfer Learning algorithms designed to align statistical distributions across different datasets. This allows models to maintain performance even when the underlying data distribution changes (non-stationarity).
- **Software Quality Assurance:** Established a complete development workflow including Unit Testing (testthat), Continuous Integration (CI/CD), and comprehensive documentation (vignettes), ensuring the package meets production quality standards.

PROFESSIONAL EXPERIENCE

Doctoral Researcher (Data Science & Algorithm Development) — **Sep 2019 – Present**
University of Massachusetts Boston — *Boston, MA*

Scalable Algorithm Design (TMCCA)

- **Complex Data Integration:** Designed a tensor-based algorithm (TMCCA) to integrate heterogeneous datasets. Implemented optimization routines (Coordinate Ascent) to efficiently process high-dimensional data structures that standard statistical methods could not handle.
- **Algorithm Implementation:** Transformed theoretical statistical models into a functional R package (`tensorMCCA`), enabling researchers to perform joint analysis on complex multi-view data without needing deep mathematical expertise.

Model Deployment Strategy & Risk Control

- **Automated Parameter Tuning:** Developed a "Proxy Tuning" mechanism to solve the cold-start problem in model deployment, enabling automatic hyperparameter selection in scenarios where real-time labels are unavailable.
- **Preventing Negative Transfer:** Implemented a statistical gating mechanism (Bootstrap Confidence Intervals) to evaluate source data quality. In simulated deployments, this strategy reduced the failure rate from 20.2% (Random Selection Baseline) to **0%**, ensuring system stability.

Root Cause Analysis & Diagnostics

- **Data Drift Diagnostics:** Developed a framework to decompose model performance drops into specific causes: input data drift versus feature extraction failure.
- **Impact Quantification:** Applied Mixed-Effects Models to quantify how environmental changes impact prediction accuracy, providing data-driven guidance for model retraining schedules.

Benchmarking & Cost Optimization

- **Performance Benchmarking:** Established a rigorous protocol to evaluate models based on a balance of accuracy, latency, and computational cost.
- **"Simple Models First" Strategy:** Conducted extensive analysis demonstrating that well-tuned linear models often match the accuracy of heavy Deep Learning models (e.g., DeepConvNet) while significantly reducing inference latency, advocating for cost-effective edge deployment.

Simulation & Validation

- **Algorithm Validation:** Designed and executed large-scale numerical simulations to stress-test feature matching algorithms under various noise levels and sample sizes, verifying their robustness and scalability before application.

Data Engineer (IoT Analytics) **Jul 2014 – Aug 2016**
China Mobile IoT Company Limited *Chongqing, China*

- **Relational Data Optimization:** Managed **GB-scale** time-series sensor logs in **MySQL**. Optimized heavy aggregation queries via indexing strategies and partitioning, ensuring report generation latency remained under SLA limits.
- **Monitoring Dashboards:** Built automated visualization tools to monitor network health and identify anomalies, supporting daily operational decision-making.

AWARDS, PUBLICATIONS & PRESENTATIONS

Honors & Awards

- **Silver Medal (Rank 98/2767, Top 4%)**, Kaggle (HMS - Harmful Brain Activity Classification) 2024
- **Doctoral Fellowship**, University of Massachusetts Boston (Full Tuition & Stipend) 2019 – Present

Selected Manuscripts

- **Shen, Y.** et al. *Drift-Feature-Performance Decomposition via Structured Geometric Modeling*. (Under Review). 2025
- **Shen, Y.** et al. *Decision-Oriented BCI: Confidence-Gated Adaptation*. (Under Review). 2025
- **Degras, D.**, **Shen, Y.** et al. *Tensor Multiple Canonical Correlation Analysis* (R Package). 2024
- **Degras, D.** & **Shen, Y.** *Scalable Feature Matching Across Large Data Collections*. 2021

Talks & Presentations

- **Invited Talk:** *Benchmarking Classification Pipelines*, MIND Seminar, Inria, France. Jun 2025
- **Poster:** *Statistical Computing Topic*, ENAR, Indiana, USA. Mar 2026

LANGUAGES

- **Chinese:** Native Speaker | **English:** Professional Proficiency