Shivasankaran Vanaja Pandi

631-949-0169 | sankaran110601@gmail.com | linkedin.com/in/shivasankaran-vp | github.com/Shiva-sankaran

Education

Stony Brook University, SUNY — Stony Brook, NY
Master of Science in Computer Science; GPA: 3.97 — *Aug 2023 – May 2025*
- **Thesis**: Robust Particle Detection for Cryogenic Electron Microscopy.
- **Coursework**: Computer Vision, Machine Learning, Natural Language Processing, Distributed Systems.

Indian Institute of Technology (IIT) — Gandhinagar, India
Bachelor of Technology with honors in Computer Science; GPA: 3.8 — *Aug 2019 – May 2023*

Experience

ML Research Engineer | *LLMs, PLMs, Generative AI* — Sep 2024 – Present
Deep Forest Sciences — *Remote*
- Designed a transformer-based VAE to generate plastic-degrading enzymes; identified **2 novel candidates** with synthesizable sequences and **verified degradation potential**.
- Trained a **distilled large-scale language model**, achieving **95% of original performance** using only **1% of the training data**.
- Explored protein generation using PLMs for **hydrolase design and antibody discovery**, contributing to efficient therapeutic candidate screening.

Open-Source Software Developer | *LLMs, PLMs, Open-Source* — May 2024 – Aug 2024
Google Summer of Code – DeepChem — *Remote*
- Integrated **ProtBERT**, the first PLM in DeepChem, enabling support for **7+ protein sequence tasks** including classification and embedding.
- Refactored and stabilized CI/CD pipelines, reducing build failures by **10%** and streamlining release workflows.
- Contributed over **6000 lines of production-level code**, enhancing PLM usability and expanding DeepChem's bioinformatics capabilities.

Projects

Systematic Alpha Factor Discovery Pipeline | *Python, Pandas, NumPy, Scikit-learn* — Jan 2025 – Apr 2025
- Built an **end-to-end alpha research framework** that systematically generates, tests, and validates predictive factors from market microstructure data using statistical learning techniques and cross-validation methods.
- Implemented **robust backtesting infrastructure** with walk-forward analysis and risk-adjusted performance metrics, identifying 12 statistically significant factors with **Sharpe ratios exceeding 1.8** and low correlation.
- Deployed **automated model selection pipeline** using ensemble methods and feature engineering, achieving **23% improvement in prediction accuracy** over baseline models while maintaining statistical significance.

Multi-Asset Risk Model with Alternative Data | *Python, Pandas, Statsmodels, Matplotlib* Sep 2024 – Dec 2024
- Developed a **systematic risk attribution model** combining traditional financial metrics with alternative datasets, using principal component analysis and factor decomposition to explain **85% of portfolio variance**.
- Implemented **dynamic hedging strategies** through statistical learning approaches, reducing portfolio volatility by **32%** while maintaining target returns using systematic rebalancing and risk parity techniques.
- Created **comprehensive backtesting framework** with performance attribution analysis, stress testing across 5 market regimes, and automated reporting dashboard for systematic strategy evaluation and model validation.

Sentiment-Driven Equity Strategy Research | *Python, NLTK, Pandas, NumPy, Scikit-learn* May 2024 – Jul 2024
- Engineered **systematic sentiment extraction pipeline** from 50,000+ financial articles using NLP techniques, creating quantitative sentiment scores with **0.64 correlation to next-day returns** for large-cap equities.
- Implemented **statistical learning models** combining sentiment factors with technical indicators, achieving **Information Ratio of 1.4** and **15% annual alpha generation** through systematic signal processing

Publications

- Constrastive Loss and Clustering Approach for Particle Detection in Cryo-EM. **ISBI** 2024 [**First Author**]
- Language Models for Function Prediction and Protein Design. AI2ASE@**AAAI** 2024 [**First Author**] *Link*
- LineEX: Data Extraction from Scientific Line Charts. **WACV** 2023 [**First Author**] *Link*
- A Unified Contrastive Learning Approach for Intent Detection and Discovery **EMNLP** 2023. [Third Author]*Link*

Technical Skills

Languages: Python, C++, SQL, R, MATLAB
Scientific Computing: Pandas, NumPy, SciPy, Statsmodels, Scikit-learn, Matplotlib, Seaborn
Quantitative Finance: Backtesting, Risk Management, Portfolio Optimization, Statistical Arbitrage, Machine Learning

Achievements

- Top 100 in Joint Engineering Examination (Math + Physics) among 1.2 million students