Susan Bataju

Fort Worth, TX • susanbatajoo@gmail.com • 817-435-3850

Summary

Ph.D. in Physics and 7+ years designing production-grade ML systems for rare-event detection. Led ensemble model pipelines (XGBoost, DNNs, BDTs) improving AUC from 60% to 86% (+40%) and architected distributed data processing across 1,000+ CPU nodes, cutting processing time 10x (6 weeks → 4 days). Expertise in model deployment, CI/CD automation, uncertainty quantification, and leading cross-functional technical teams.

Technical Skills

- **Programming Languages:** Python (Expert), SQL (Proficient), C++ (Proficient), R (Basic)
- **Machine Learning:** TensorFlow, PyTorch, Scikit-learn, XGBoost, Keras, Model Selection, Hyperparameter Tuning, Cross-Validation, Ensemble Methods, Neural Networks (DNNs, CNNs, GNNs), SHAP Analysis
- **Data Analysis & Statistics:** Pandas, NumPy, Statistical Inference, Hypothesis Testing, A/B Testing, Bayesian Methods, Profile Likelihood, Uncertainty Quantification
- **Data Visualization:** Matplotlib, Seaborn, Plotly
- **Big Data & Distributed Computing:** HTCondor, Slurm, Kubernetes, High-Performance Computing, Parallel Processing, ETL Pipelines
- **MLOps & Engineering:** Git/GitLab, CI/CD, Unit Testing, Docker (Familiar), Model Validation, Feature Engineering, Data Pipeline Development
- **Databases:** SQL (filtering, joins, aggregations), SQLAlchemy ORM, Structured Data Queries

Selected Projects & Open Source

- **Financial News Analyzer (2024)** • GitHub
  Developed scalable NLP pipeline (Hugging Face transformers) batch-processing financial news across sentiment and NER models; designed relational schema (SQLAlchemy ORM) for per-model sentiment scores and entity storage. Optimized inference throughput by 60% via batching and model-comparison framework. *Tech: Python, Hugging Face, SQLAlchemy, PyTorch*
- **Credit Card Fraud Detection (2024)** • GitHub
  Built end-to-end ML pipeline for highly imbalanced financial dataset (285K transactions); engineered features, handled class imbalance (SMOTE/resampling), and tuned XGBoost to achieve 0.98 AUC-ROC. Delivered production-ready code with 90%+ unit test coverage and comprehensive documentation. *Tech: Python, XGBoost, Scikit-learn, Pandas, Pytest*

Experience

Graduate Research Assistant **2021–2025**

Southern Methodist University, Dallas, TX

- Designed and deployed ensemble ML pipeline (XGBoost, DNNs, BDTs) with custom loss functions to detect rare physics signals in datasets with 1:10,000 signal-to-background ratio; achieved 86% AUC (+40% relative improvement over 60% baseline), directly enabling discovery published in top-tier physics journal (JHEP).
- Applied SHAP analysis and domain expertise to engineer 100+ physics-informed features; reduced feature space by 60% while maintaining 86% AUC, cutting training time by 50% and improving model interpretability for scientific validation.
- Architected distributed data-processing pipeline across 1,000+ CPU clusters using HTCondor and Slurm with automated job management; reduced processing time from 6 weeks to 4 days (10x speedup) while maintaining 99.9% data integrity.
- Built comprehensive validation framework implementing stratified cross-validation, AUC/ROC analysis, and KS-tests across 20+ systematic variations; demonstrated robust generalization with $< 5\%$ performance variance out-of-sample, meeting publication standards.

- Developed SQL-like filtering system with 20+ conditional rules incorporating domain knowledge to clean high-dimensional data (1M+ samples); improved downstream model accuracy by 25% and reduced false positive rate by 35%.
- Implemented GPU-accelerated hyperparameter search (CUDA, parallel grid search); reduced per-model training time from 8 hours to 45 minutes (10x improvement), enabling exploration of 500+ configurations and identifying optimal architecture.
- Developed statistical inference pipeline integrating 15+ systematic uncertainties using profile likelihood methods; enabled hypothesis testing with proper error propagation, contributing to collaboration-wide framework adopted by 100+ researchers.
- Created and delivered weekly technical presentations to international collaboration of 50+ physicists; translated complex ML results into scientific insights with clear visualizations, facilitating feedback-driven development that improved model performance by 15%.

Technical Stack: Python, TensorFlow, PyTorch, Scikit-learn, XGBoost, Pandas, NumPy, SQL, Git, CUDA, HTCondor, Slurm

Doctoral Researcher — Electron Identification Framework **2022–2025**

CERN

- Architected modular Python/C++ software suite with automated workflows handling 10+ systematic variations across 150+ kinematic bins for 3,000+ international researchers; delivered scalable solution adopted by 100+ researchers, reducing analysis time from weeks to hours and establishing new collaboration standard.
- Implemented comprehensive CI/CD pipeline via GitLab including unit tests (95% code coverage), automated regression testing, and peer code reviews; achieved zero critical bugs in 18 months of production use across 50+ analysis teams.
- Designed intuitive API with extensive documentation and example notebooks; onboarded 30+ new users with 90% reporting independent analysis capability within 1 week, significantly lowering technical barriers to adoption.
- Established GitLab workflow with merge request standards, automated testing gates, and documentation requirements; maintained 2-day average merge time across 20+ developers in 4 time zones while ensuring code quality.

Technical Stack: Python, C++, ROOT, Git/GitLab, CI/CD, Unit Testing, Matplotlib

Doctoral Researcher — Data Systems & Infrastructure **2022–2025**

CERN

- Designed multi-linear bit-encoding strategy for trigger data format optimization, balancing information retention against 10-bit bandwidth constraints; improved energy reconstruction accuracy by 15% while reducing data loss by 20%, enabling real-time processing within latency requirements (< 100 microseconds).
- Provided 24/7 on-call support for critical detector hardware during $1B physics data-taking operations ($50K/hour downtime cost); maintained 99.9% uptime over 6-month rotation by diagnosing and resolving 20+ real-time system anomalies with 15-minute average resolution time.
- Designed custom control interface (WinCCOA, OPCUA protocols) enabling real-time visualization and remote control of 100+ detector components; reduced mean time to diagnosis by 40% through improved system visibility.

Technical Stack: Python, C++, Statistical Modeling, Constrained Optimization, WinCCOA, OPCUA

Teaching Assistant **2020–2022**

Physics Department, Southern Methodist University

- Designed interactive tutorials and hands-on lab sessions teaching experimental methods and statistical analysis to 75+ undergraduate students; achieved 85% average exam scores and 4.5/5 teaching evaluation rating.
- Held weekly office hours providing personalized guidance in Python programming and data analysis; 90% of attendees improved grades by at least one letter grade.

Undergraduate Researcher **2017–2020**

University of Texas at Arlington

- Developed Monte Carlo simulation framework (MadGraph, Pythia8) generating 500K+ simulated events; demonstrated 3-sigma discovery potential in 50% of parameter space, convincing collaboration to approve full analysis.
- Designed systematic testing protocol and SolidWorks test bench infrastructure for 200+ critical hardware components; identified 15% failure rate preventing $100K+ in potential detector damage.

Education

Ph.D. in Physics **2020–2025**

Southern Methodist University, Dallas, TX
Dissertation: Machine Learning Methods for Rare Event Detection in High-Dimensional Data

B.S. in Physics (Minor: Mathematics) **2016–2020**

University of Texas at Arlington
Relevant Coursework: Applications of Deep Learning, Statistical Inference, Intro to Programming

Publications

- Co-author, ATLAS Collaboration. "Search for non-resonant Higgs boson pair production in final states with leptons, taus, and photons in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector." *Journal of High Energy Physics*, vol. 2024, no. 08, article 164, 2024. doi:10.1007/JHEP08(2024)164. arXiv:2405.20040 [hep-ex].

Leadership & Communication

- **Workshop Organizer — BCVSPIN 2024**, Kathmandu, Nepal (Dec 2024)
  Designed hands-on curriculum with Python-based tutorials and coordinated international speaker logistics to introduce particle physics and ML to 30+ undergraduates in underserved region; achieved 95% satisfaction rating and inspired 80% of participants to pursue graduate studies in STEM fields per post-event survey.
- **Physics Outreach Coordinator**, Pokhara, Nepal (Dec 2024)
  Organized educational events featuring international experts to introduce advanced technical topics in accessible ways; inspired students in rural communities to pursue STEM careers.