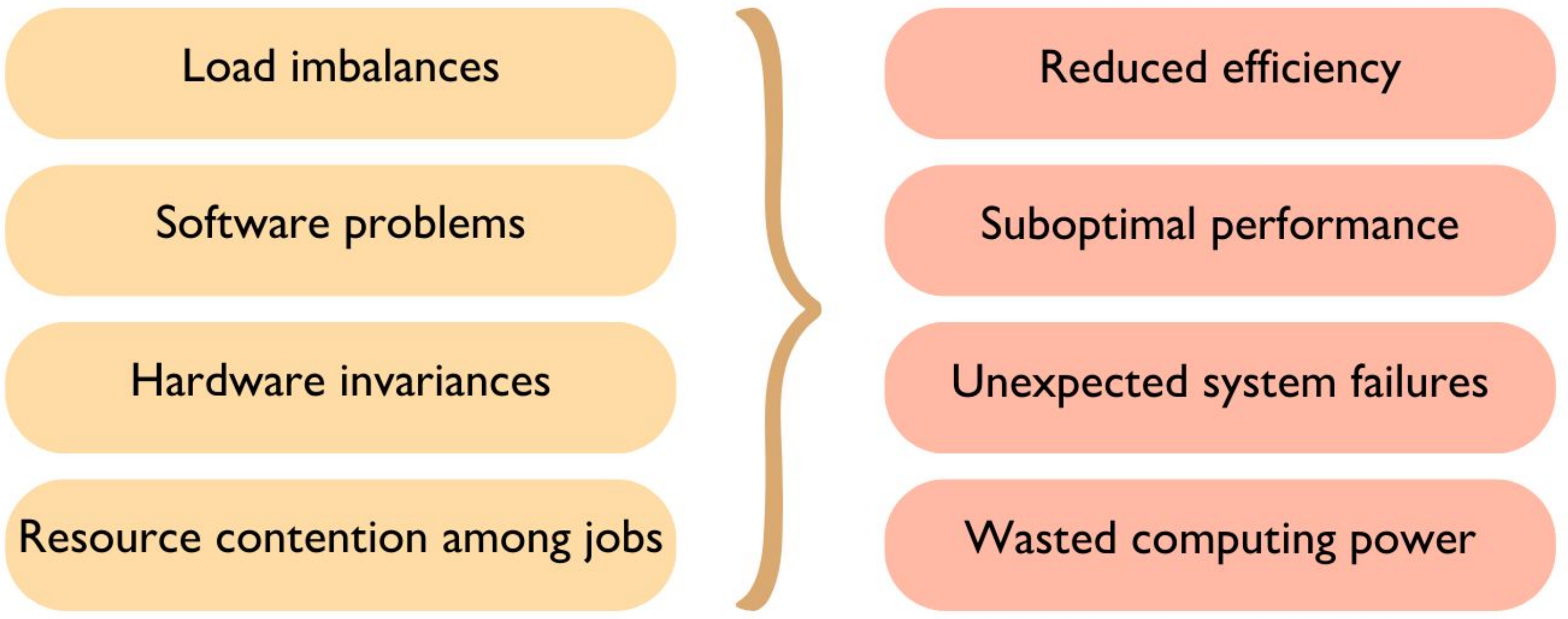
AI-based Scalable Analytics for Improving Performance and Resilience of HPC Systems

Efe Sencan[1], Beste Oztop[1], Benjamin Schwaller[2], Vitus J. Leung[2], Jim Brandt[2], Brian Kulis[1], Manuel Egele[1], Ayse K. Coskun[1]

[1]Electrical and Computer Engineering Department, Boston University, Boston, MA, 02215

[2]Sandia National Laboratories, Albuquerque, NM, 87123

Introduction



Load imbalances

Software problems

Hardware invariances

Resource contention among jobs

Reduced efficiency

Suboptimal performance

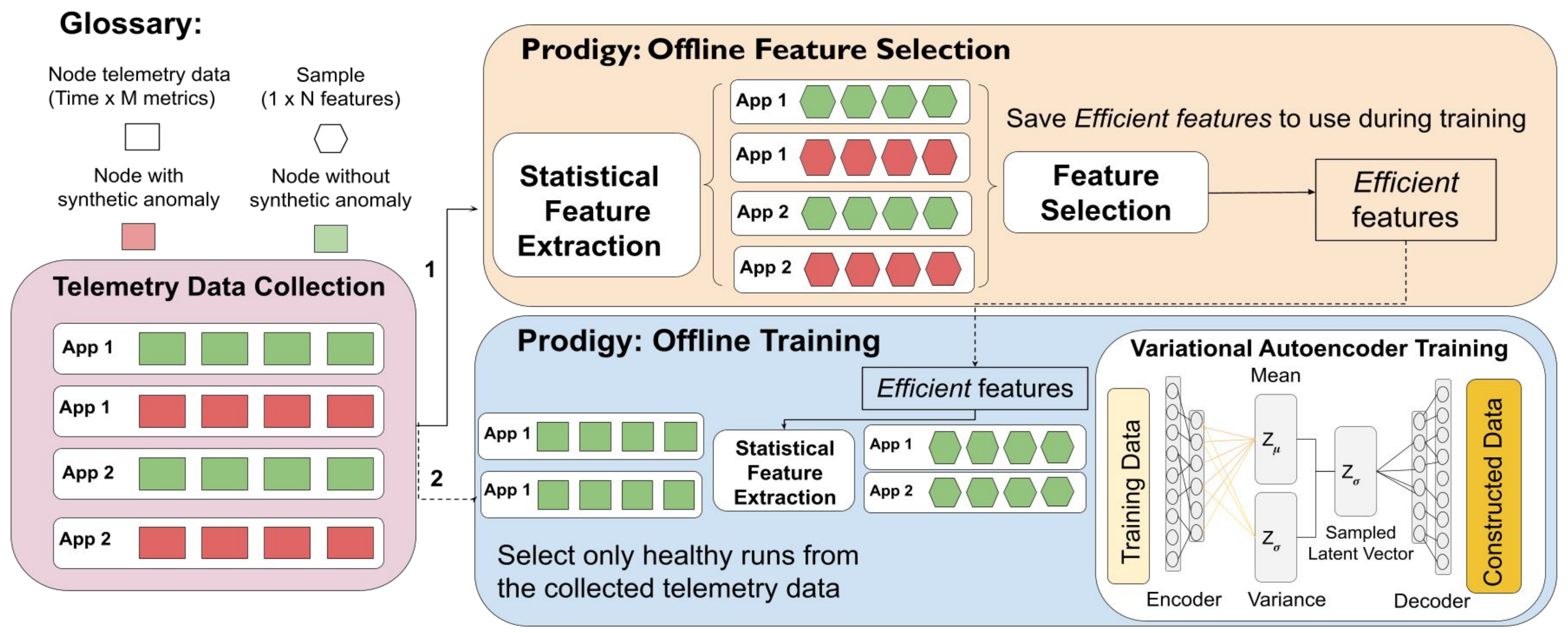
Unexpected system failures

Wasted computing power

Our project aims to create scalable AI frameworks for automatically diagnosing and mitigating performance anomalies in HPC systems.

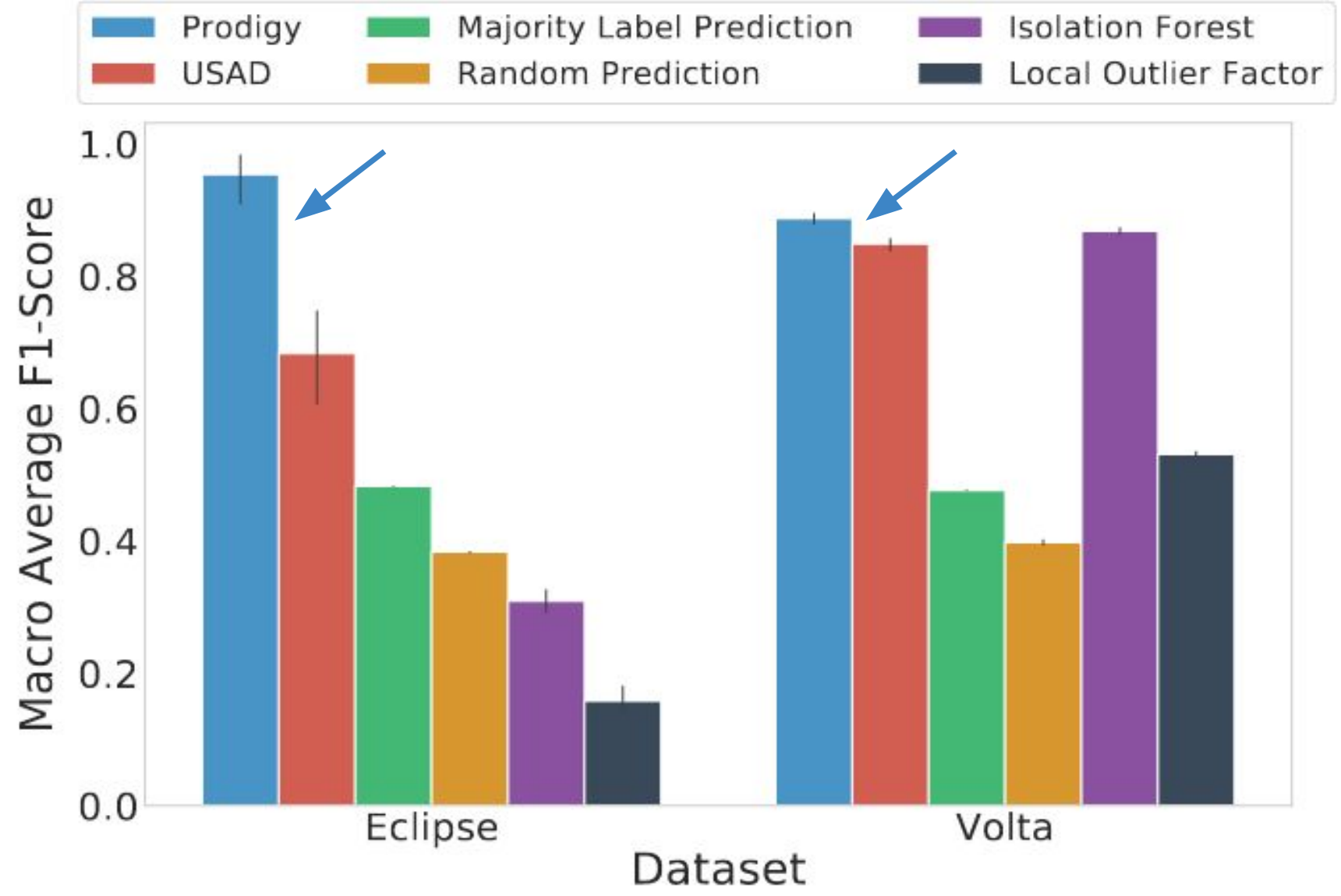
Prodigy

Our recent work involves designing an unsupervised anomaly detection framework for HPC systems, aiming to reduce reliance on extensive labeled data [1].



Glossary:
Node telemetry data (Time x M metrics)
Sample (1 x N features)
Node with synthetic anomaly
Node without synthetic anomaly
Telemetry Data Collection
App 1
App 1
App 2
App 2
1
2
Prodigy: Offline Feature Selection
Statistical Feature Extraction
Save *Efficient features* to use during training
Feature Selection
Efficient features
Prodigy: Offline Training
Efficient features
Statistical Feature Extraction
Select only healthy runs from the collected telemetry data
Variational Autoencoder Training
Training Data
Mean
Z_μ
Z_σ
Sampled Latent Vector
Constructed Data
Encoder
Variance
Decoder

Aim: Reduce the extensive data labeling in the training process for performance anomaly detection in HPC systems.

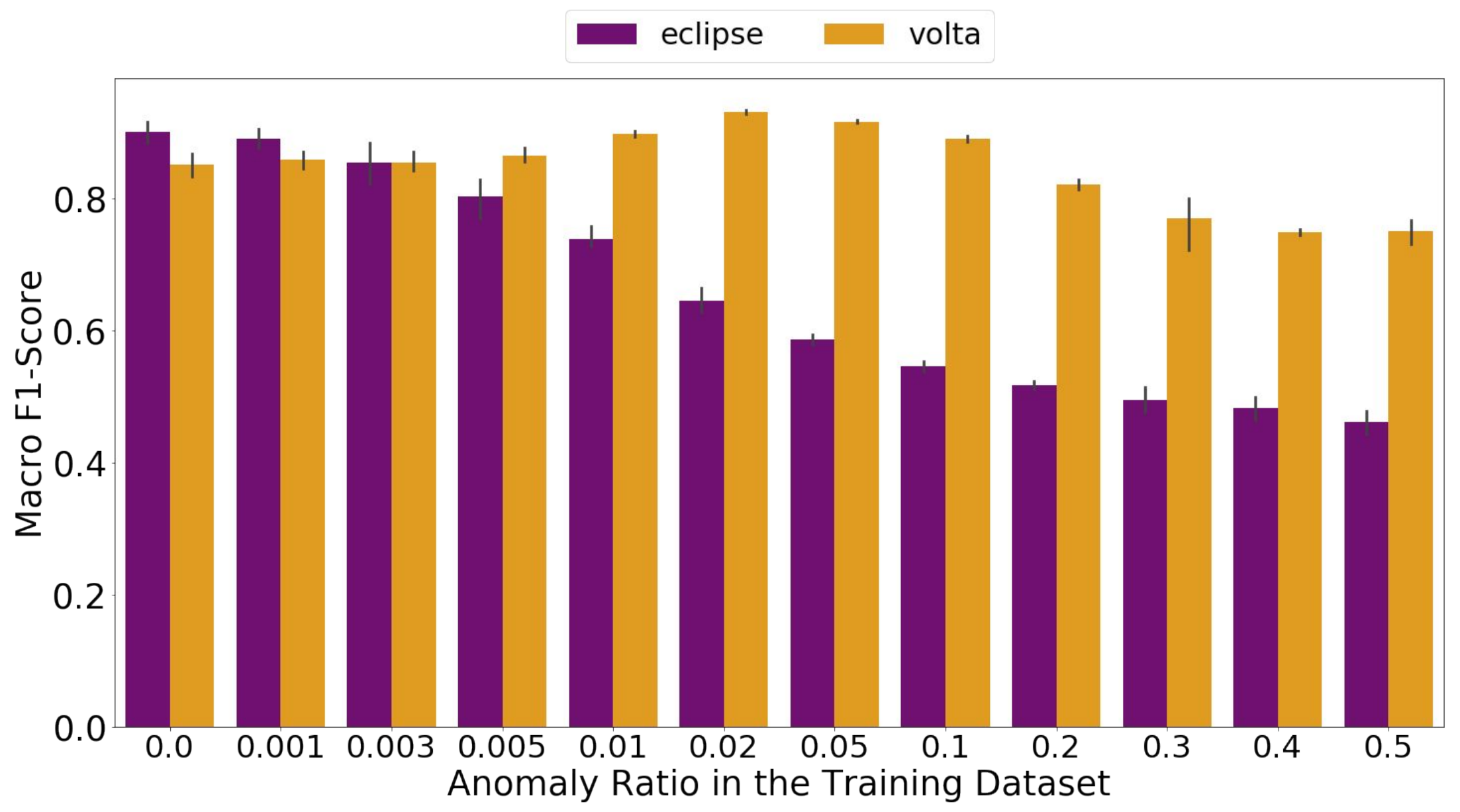


Prodigy achieves **0.95** and **0.88** F1 score for Eclipse and Volta.

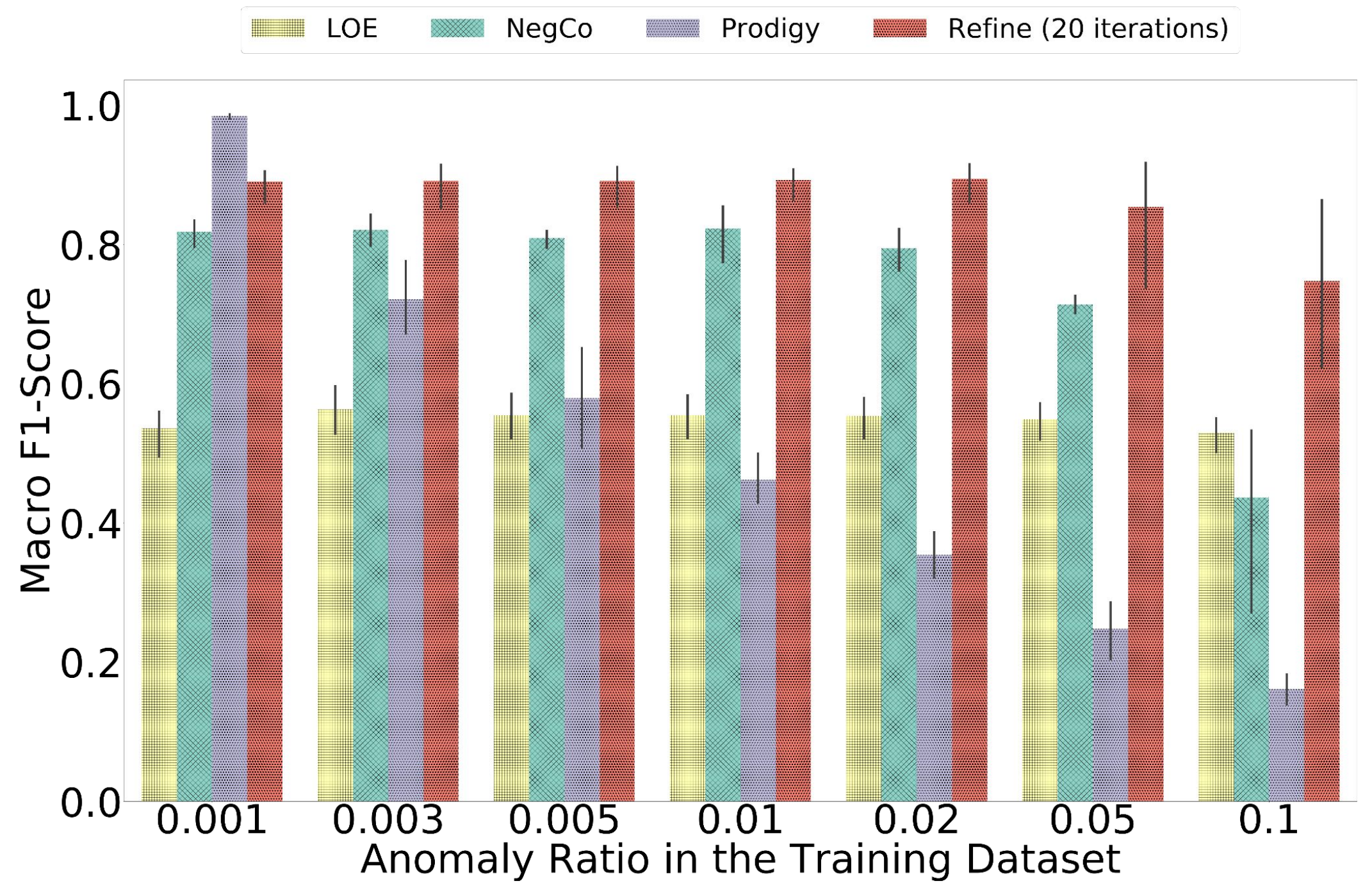
- Variational Autoencoder (VAE) model is trained with the healthy nodes telemetry data.
- Contamination in HPC telemetry data leads to anomalous samples being mislabeled as healthy.
- VAE model accuracy decreases due to the contamination problem.

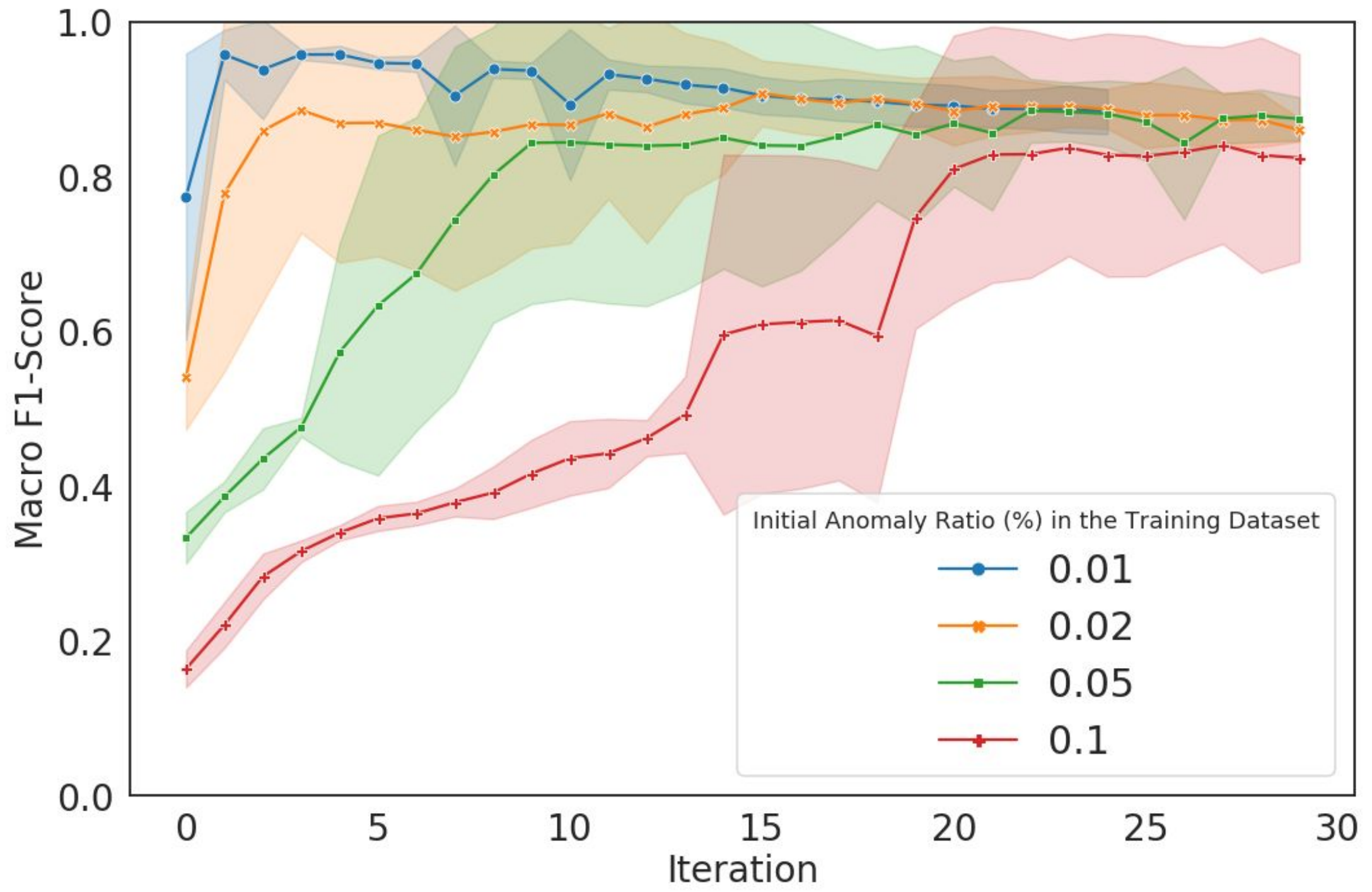
Robust Unsupervised Anomaly Detection for Production HPC Systems

- Unsupervised anomaly detection frameworks often include unhealthy samples and fails the "only healthy data" assumption.
- Our iterative robust VAE method, *Refine*, uses VAE reconstruction error to identify and remove unhealthy samples from the training data.
- Beyond HPC anomaly detection, *Refine* can help any field dealing with contaminated datasets in unsupervised settings.



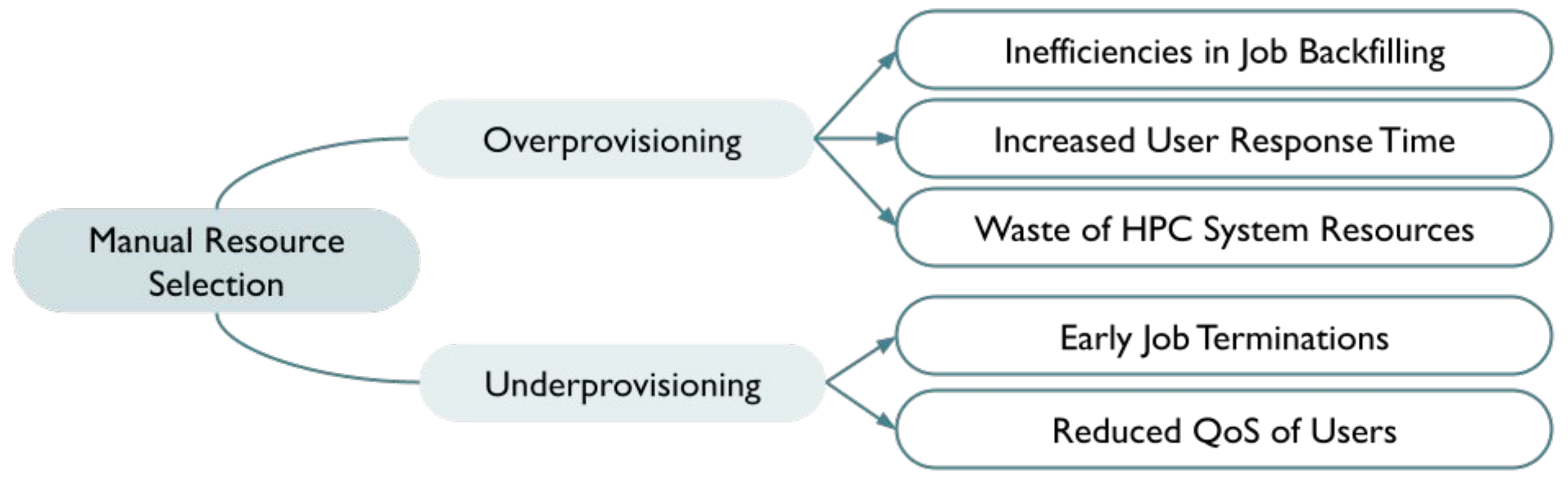
Anomaly ratio in the training dataset	Dataset	Macro Average F1 Score
0 to 10%	Eclipse	from **0.95** to below **0.50**
0 to 30%	Volta	rom **0.88** to below **0.75**





Intelligent Resource Allocation

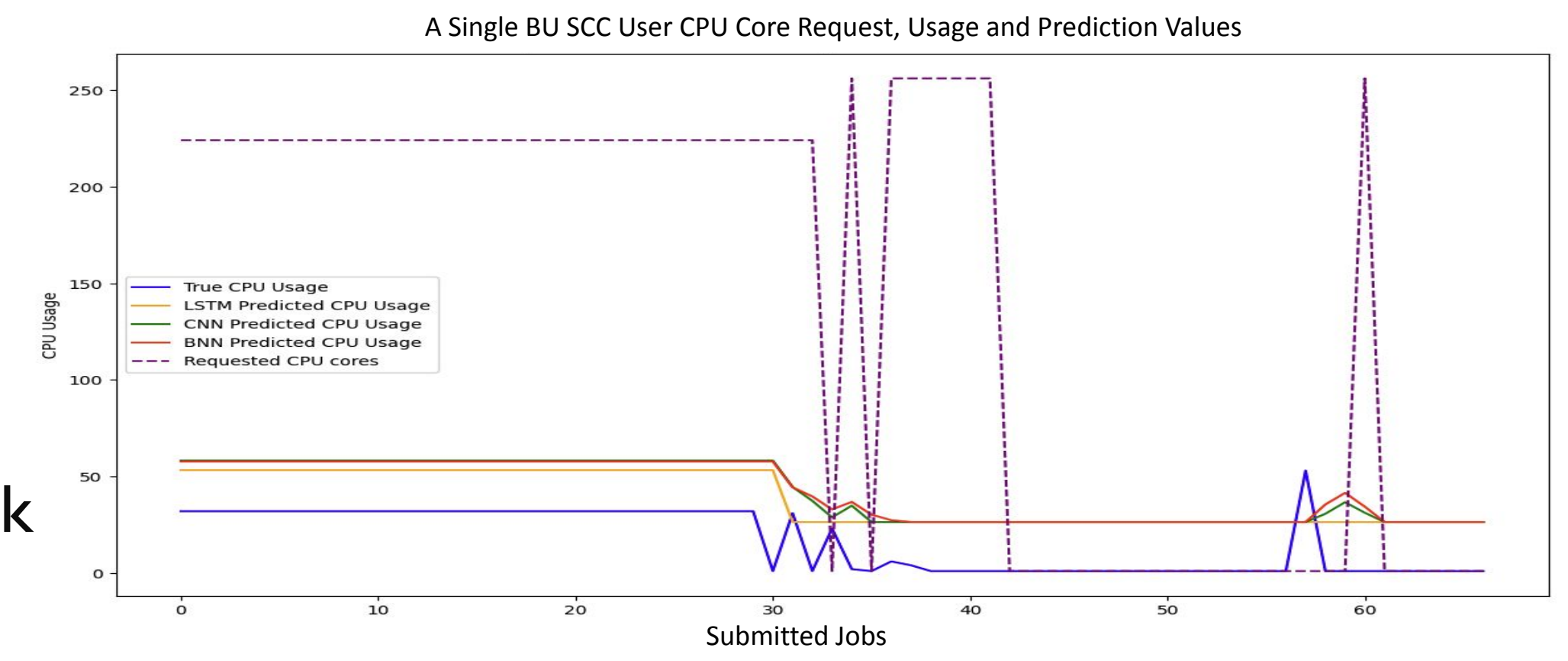
We aim to reduce resource waste in HPC systems [2] by developing an online tool that predicts resource needs for future jobs based on historical usage data.

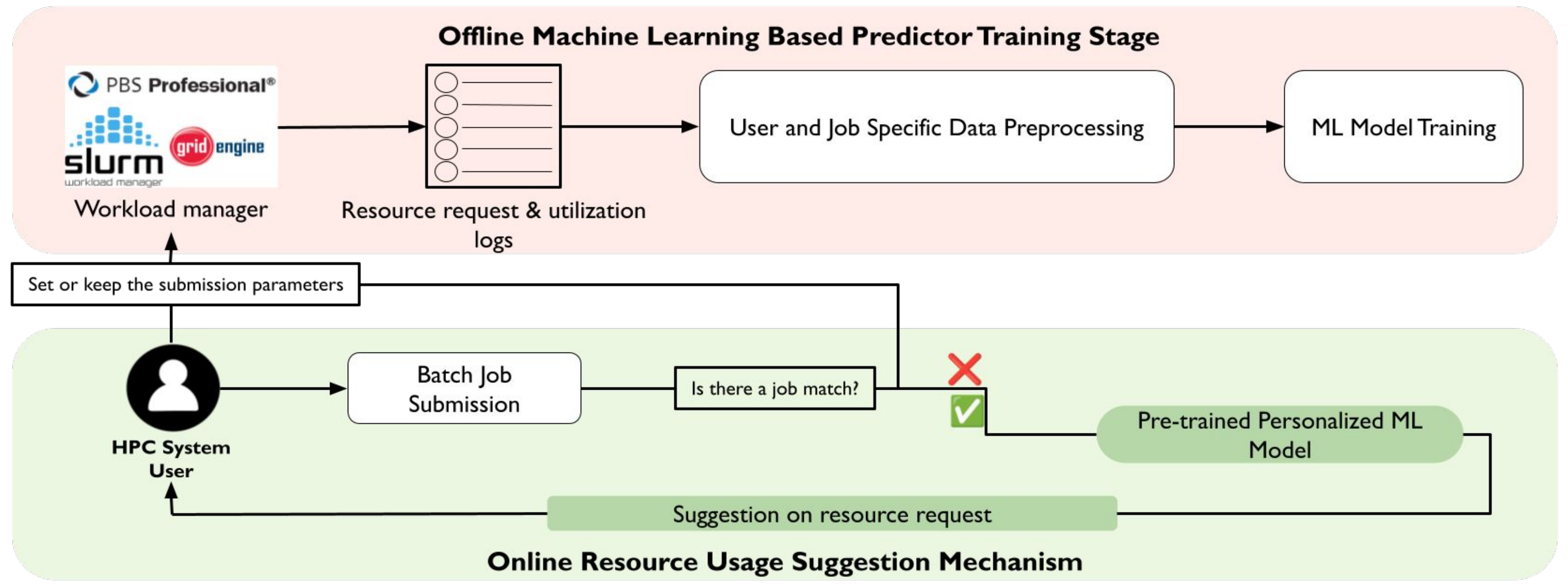


Manual Resource Selection
Overprovisioning
Underprovisioning
Inefficiencies in Job Backfilling
Increased User Response Time
Waste of HPC System Resources
Early Job Terminations
Reduced QoS of Users

Time series prediction of CPU usage using the 2023 Boston University Shared Computing Cluster Dataset.

Reducing the over requests and resource waste using Neural Network models.

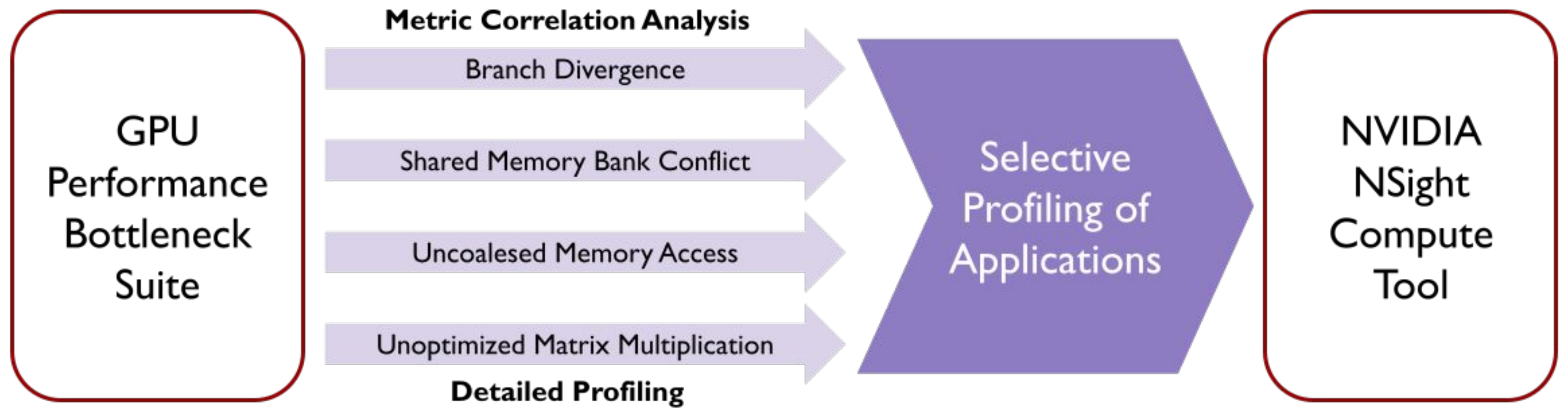




Offline Machine Learning Based Predictor Training Stage
PBS Professional
slurm workload manager
grid engine
Workload manager
Resource request & utilization logs
User and Job Specific Data Preprocessing
ML Model Training
Set or keep the submission parameters
HPC System User
Batch Job Submission
Is there a job match?
❌
✅
Pre-trained Personalized ML Model
Suggestion on resource request
Online Resource Usage Suggestion Mechanism

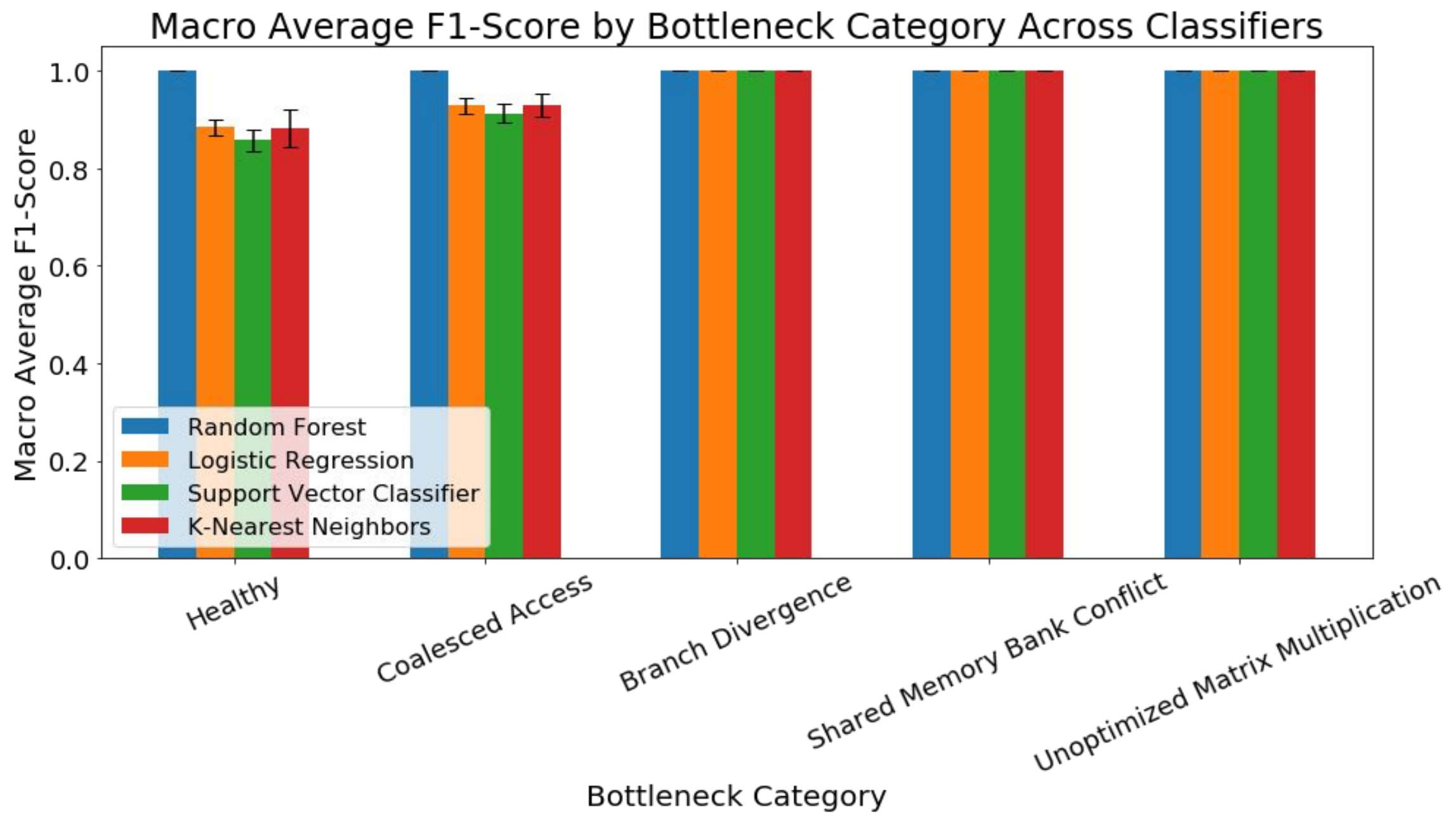
GPU Performance Bottleneck Suite

- GPU-based applications often face performance bottlenecks from branch and memory divergence [3].
- Existing ML methods do not address GPU-specific performance inefficiencies.



GPU Performance Bottleneck Suite
Metric Correlation Analysis
Branch Divergence
Shared Memory Bank Conflict
Uncoalesed Memory Access
Unoptimized Matrix Multiplication
Detailed Profiling
Selective Profiling of Applications
NVIDIA NSight Compute Tool

We aim to replicate and identify common GPU performance bottlenecks while minimizing the profiling overhead of the NVIDIA Nsight Compute tool.



The F1 score for our synthetic dataset of 1380 samples is nearly 100% with the Random Forest model.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525

References:

[1]Burak Aksar, Efe Sencan, Benjamin Schwaller, Omar Aaziz, Vitus J. Leung, Jim Brandt, Brian Kulis, Manuel Egele, and Ayse K. Coskun. 2023. Prodigy: Towards Unsupervised Anomaly Detection in Production HPC Systems. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '23). Association for Computing Machinery, New York, NY, USA, Article 26, 1–14. https://doi.org/10.1145/3581784.3607076

[2] Md Nahid Newaz and Md Atiqul Mollah. 2023. Memory Usage Prediction of HPC Workloads Using Feature Engineering and Machine Learning. In Proceedings of the International Conference on High Performance Computing in Asia-Pacific Region (HPCAsia '23). Association for Computing Machinery, New York, NY, USA, 64–74. https://doi.org/10.1145/3578178.3578241

[3]Rong Zheng, Qi Hu, and Hai Jin. 2018. Gpuperfml: A performance analytical model based on decision tree for GPU architectures. In Proceedings of the IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, Exeter, UK, 602–609. https://doi.org/10.1109/HPCC/SmartCity/DSS.2018.00110

Prodigy Github