Interim Report

Project Title:

Data Analysis and Prediction Models for AI Benchmarking Data

Submitted By:

Team 6
Parth Praful Parakhiya | 110179090
Tirth Rameshbhai Patel | 110157043
Sanket Harshadbhai Kothiya | 110156177

Internal Supervisor:

Dr. Olena Syrotkina
School of Computer Science, University of Windsor

Industry Supervisor:

Nishatha Nagarajan, Jaguar Land Rover (JLR)

Submission Date: June 30, 2025

University of Windsor
Summer 2025

Contents

1 Changes in the Scope of the Project

The project has remained faithful to the core goals outlined in the original proposal. However, some refinements and optimizations were made based on real-time insights and feasibility:

- **Refined Modeling Phases:** Prediction efforts were split into three clearer categories: (i) bias/weight-based models, (ii) static KPI predictors, and (iii) neural network-specific models.
- **PostgreSQL Chosen:** Initially considering both SQL and NoSQL, we finalized PostgreSQL for its structured schema and performance with analytical queries.
- **Expanded Metrics:** The number of engineered features expanded from 20+ to 46+ based on exploratory data analysis and correlation studies.
- **Architecture Emphasis:** Due to strong hardware-specific trends, a larger portion of the modeling was allocated to capturing architecture/manufacturer bias.

These refinements have enhanced analytical precision while keeping deliverables aligned with the original scope.

2 Progress Summary

Phase 1: Database & Foundation (100% Complete)

- Constructed a fully normalized PostgreSQL database schema (Figure 1) with over 2,108 unique records spanning 17 GPU/AI architectures from NVIDIA, AMD, and Intel.
- Reduced entries with "Unknown" architecture from 1,252 to 532 using cross-referenced data and pattern-based inference.
- Engineered 26 domain-specific metrics and expanded to 46 normalized features, fully documented in the AI Benchmark Matrix documentation (`ai_benchmark_matrix_column_documentation.pdf`).
- Created workload-specific performance matrices covering AI models like ResNet50, BERT, GPT2, EfficientNet, and MobileNet.
- Documented every column, derived formula, and transformation in the database and matrix documentation for full reproducibility.

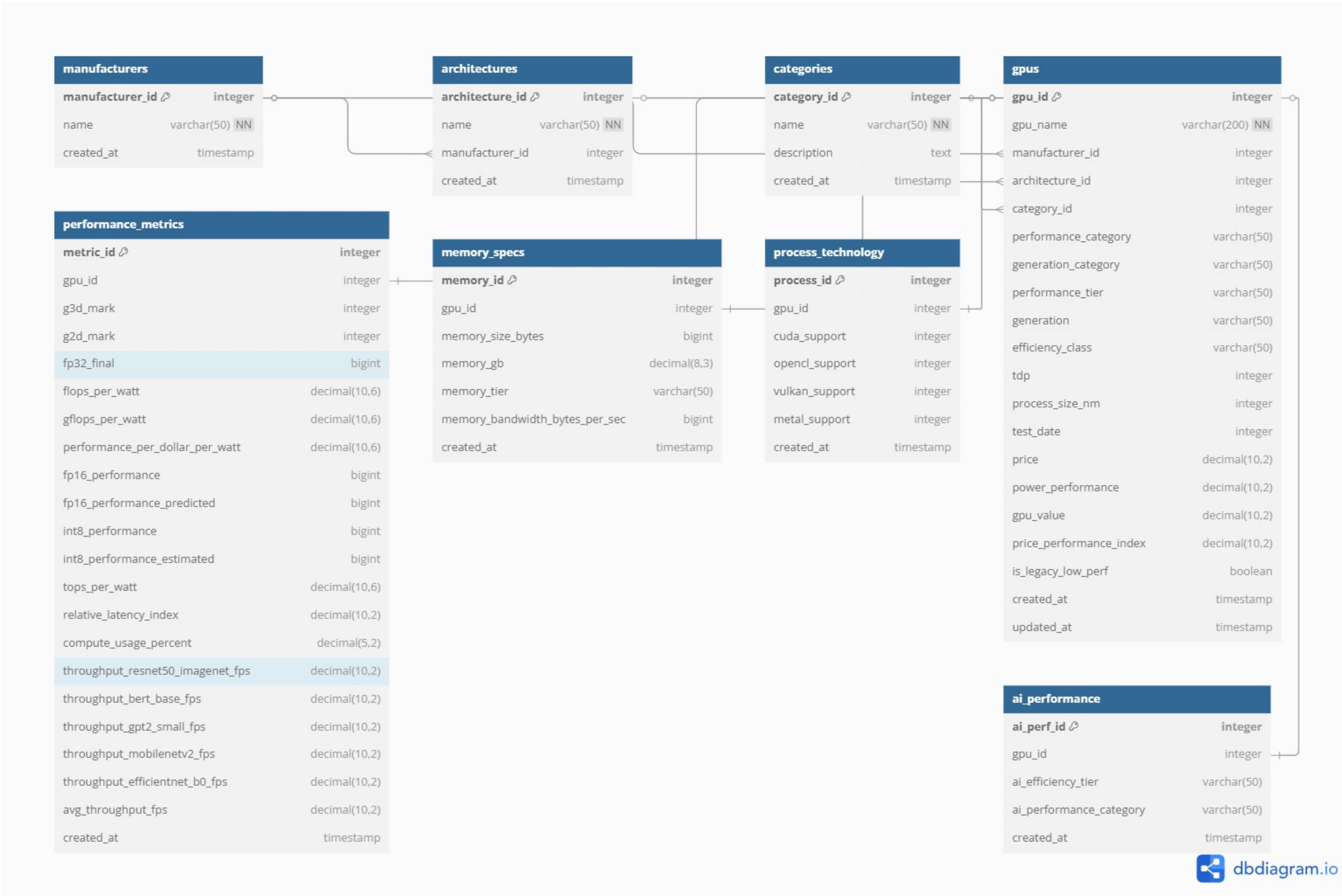

Figure 1: PostgreSQL Schema – AI Benchmarking Dataset

Phase 2: Data Optimization (70% Complete)

- Developed correlation heatmaps (Figure 2) to uncover hidden relationships among metrics like G3Dmark, G2Dmark, FP32_Final, and TDP.
- Conducted exploratory distribution analysis (Figure 3) for key metrics like G3Dmark to design suitable normalization pipelines.
- Engineered 46 features and visualized their importance using model-based feature selection (Figure 4).
- Established performance tiers and architecture categories using unsupervised clustering and categorical encoding (e.g., generation, vendor, AI tier).
- Implemented vendor-specific missing value imputation strategies to resolve data sparsity in attributes like memory, API support, and pricing.

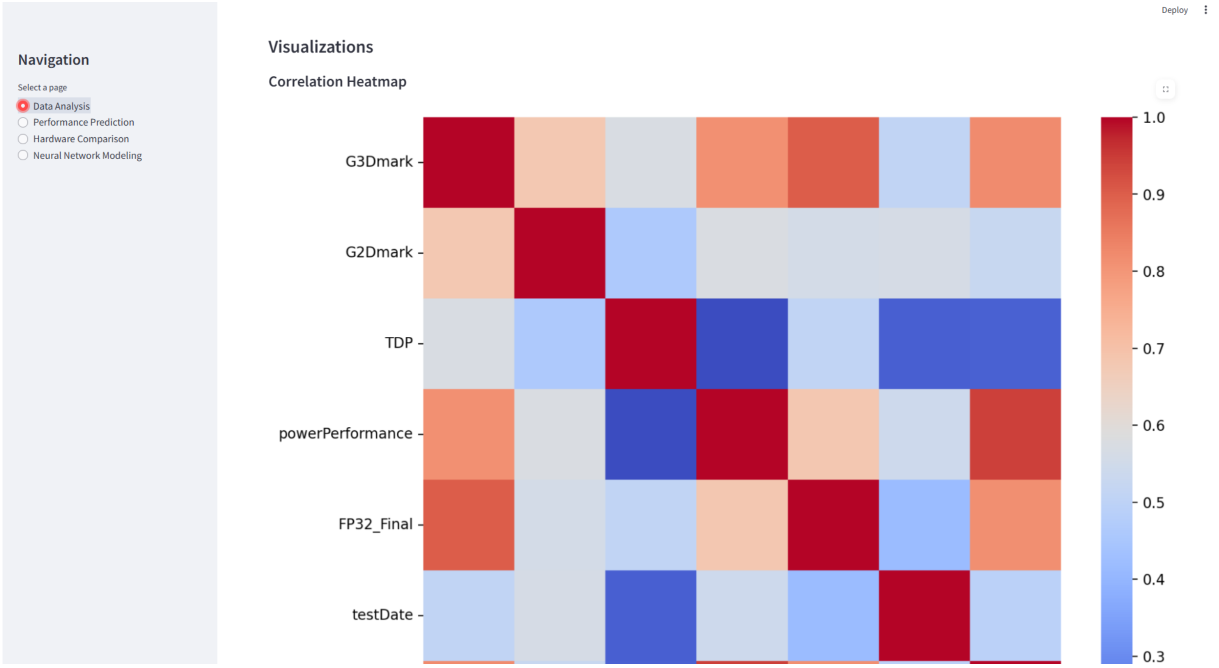

Figure 2: Correlation Heatmap of Core Metrics

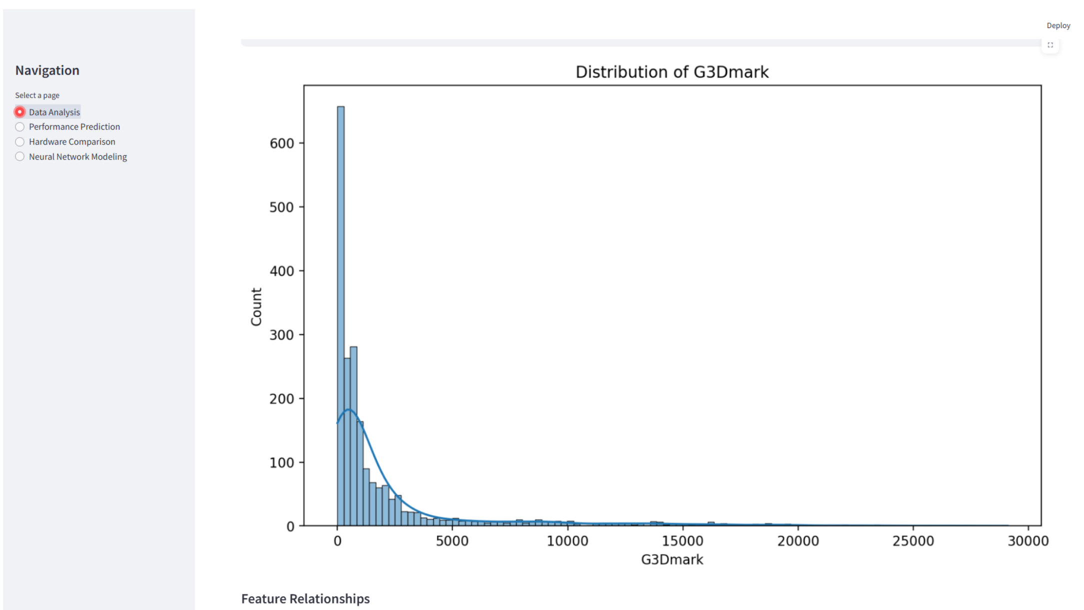

Figure 3: Distribution of G3Dmark Across 2,108 GPUs

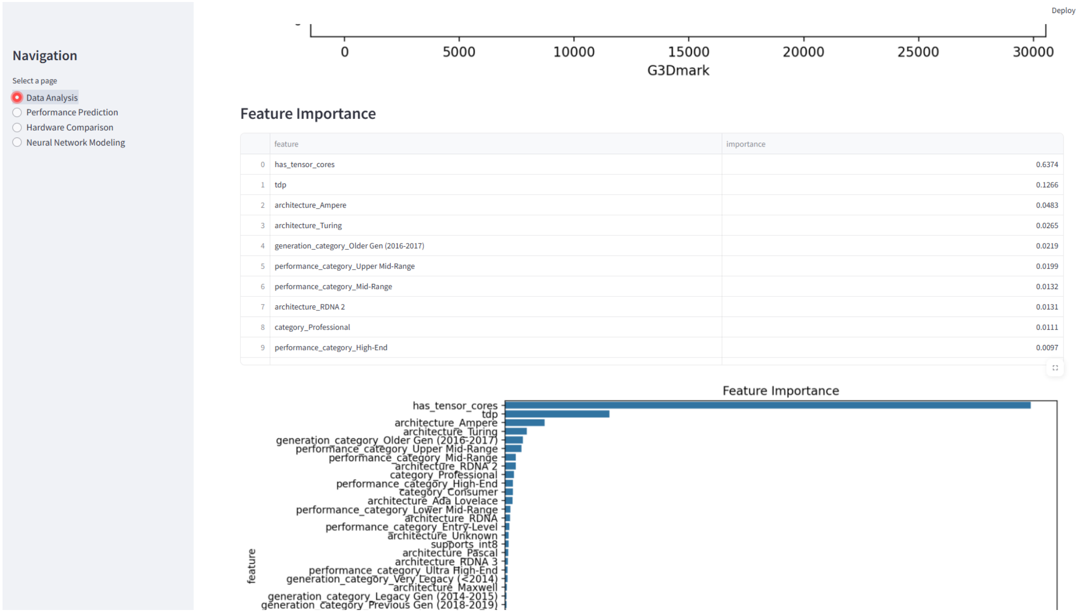

Figure 4: Feature Importance Based on Random Forest Classifier

Web Interface Prototype (Dashboard Preview)

A dashboard web interface has been partially implemented using Streamlit (Figure 5–7), featuring:

- Visualization page with correlation, distribution, and scatterplots.
- Hardware comparison tools by vendor, generation, and category.
- Real-time AI performance prediction for selected architectures and hardware.
- Neural network to hardware mapping and optimization report generation.

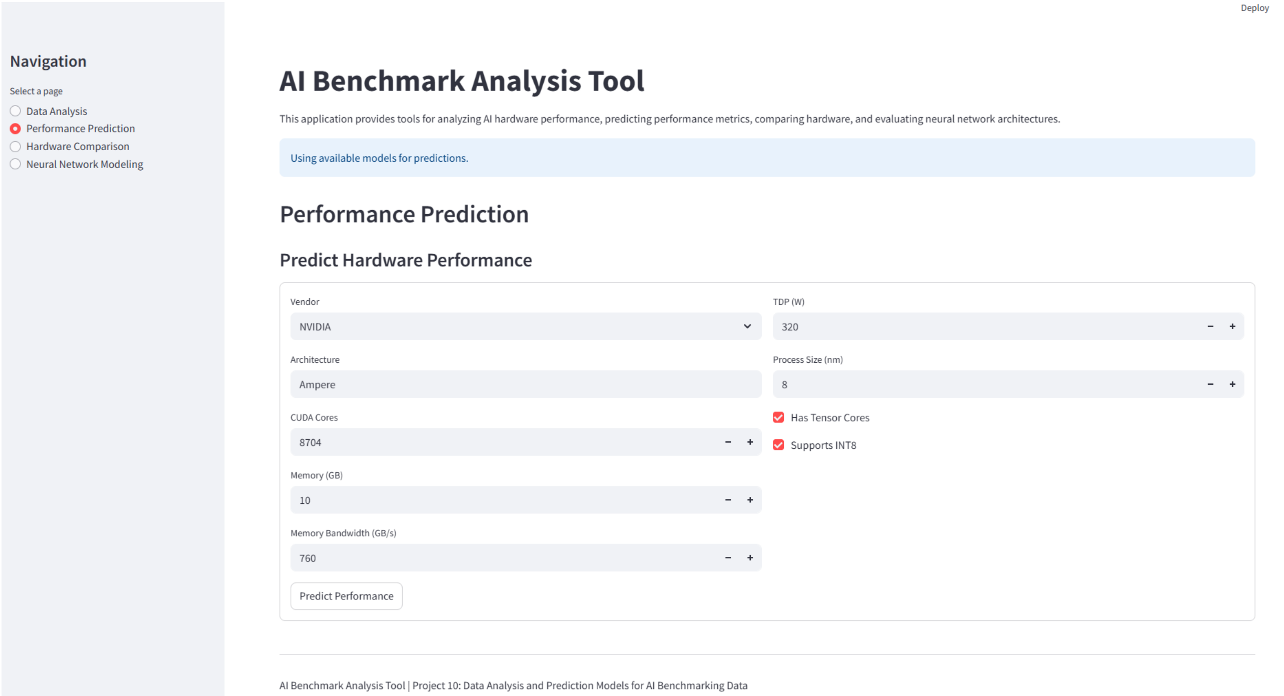

Figure 5: Performance Prediction Panel – User Input for NVIDIA Ampere

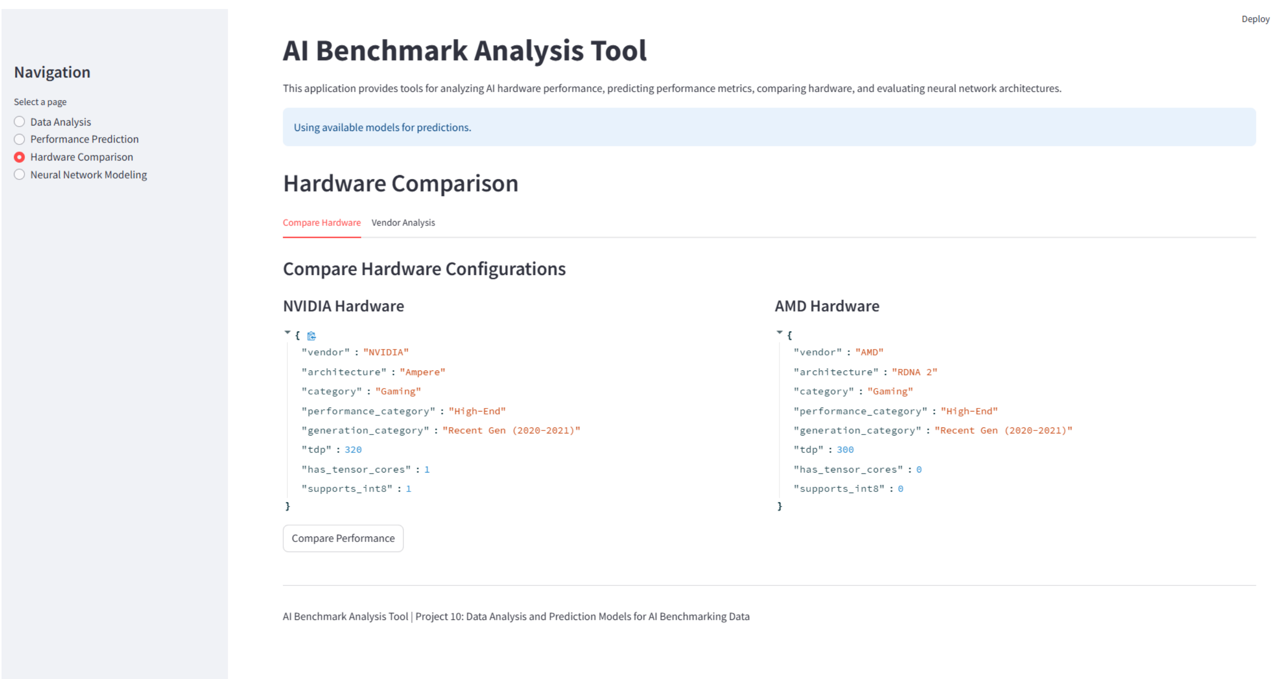

Figure 6: Hardware Comparison Panel – NVIDIA vs AMD Example

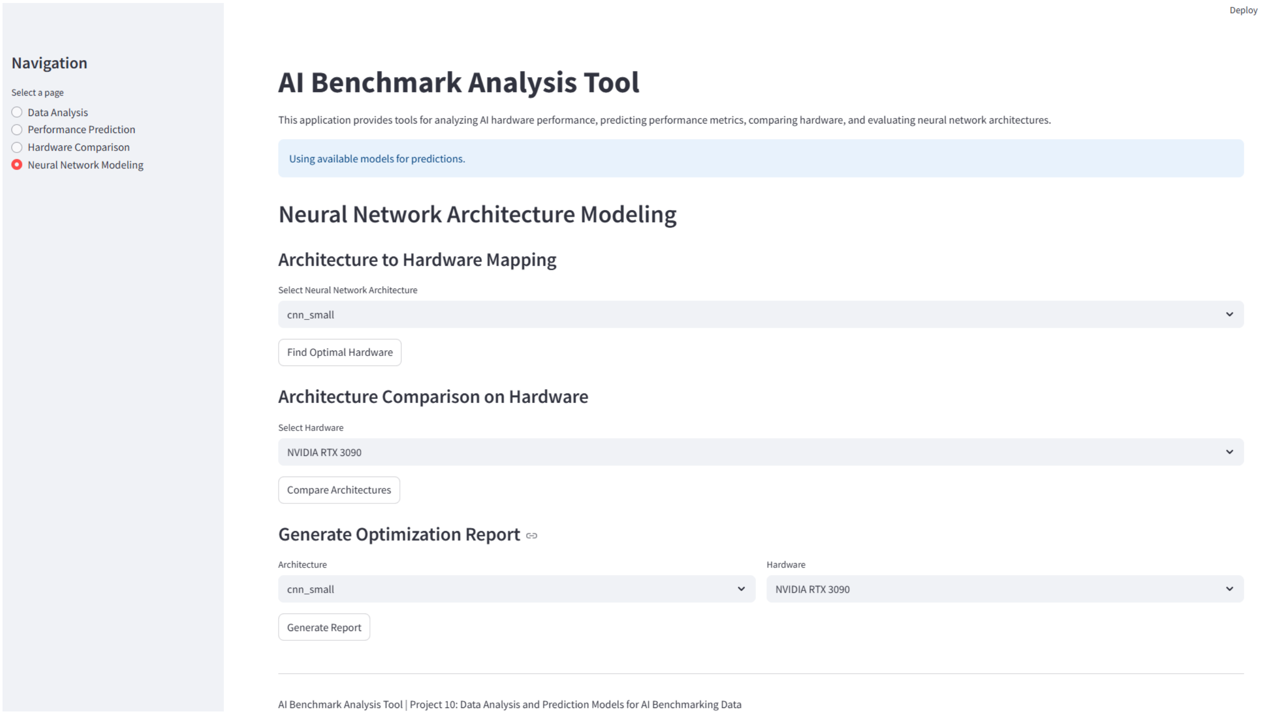

Figure 7: Neural Network Architecture Mapping Panel – CNN to Hardware

3 Challenges and Issues

- **Incomplete Architecture Data:** Many GPUs lacked proper architecture labels. We addressed this via vendor-specific heuristics and external validation.
- **Feature Overlap:** 46 engineered features introduced redundancy. A feature importance ranking system using mutual information is being developed.
- **Cross-Vendor Comparison Difficulty:** Performance varies widely across NVIDIA, AMD, and Intel due to platform-specific optimizations. To solve this, bias correction layers and architecture-specific weighting models are under construction.
- **Model Generalizability:** Ensuring the trained models work across unseen architectures is an ongoing concern. We're mitigating this with cross-validation using hold-out architecture subsets.

4 Next Steps (July 1–30)

Week 1 (July 1–5):

- Finalize implementation of bias/weight-based prediction models.
- Analyze manufacturer-specific trends (e.g., CUDA vs ROCm).
- Script: `bias_weight_models.py`

Week 2 (July 6–12):

- Begin static model development for latency, throughput, and efficiency.
- Benchmark against decision tree and XGBoost baselines.
- Script: `static_prediction_models.py`

Week 3 (July 13–19):

- Extend static models to predict memory usage and cost-performance ratios.
- Integrate historical pricing datasets for economic insights.

Week 4 (July 20–26):

- Build neural network-specific predictors for ResNet50, BERT, and GPT-2.
- Include hybrid architectures and mixed-precision models.
- Script: `neural_network_predictors.py`

Week 5 (July 27–30):

- Validate all models and compile final results.
- Draft deployment script for AI workload prediction pipeline.
- Prepare materials for the final presentation.