Vol-4 Issue-1 2026
Scientific Research Journal of Science, Engineering and Technology
ISSN: 2584-0584, Peer Reviewed Journal

A Linear Algebra–Based Mathematical Model of Digital System Performance Using Neural Networks as Predictive Approximators

Chaitrashree S[1], Dr. E. Saravana Kumar[2], Goutam Parashuram Gotur[3]

[1] Department of Computer Science and Engineering, The Oxford College of Engineering, Bengaluru, Karnataka, India

[2] Department of Computer Science and Engineering, The Oxford College of Engineering, Bengaluru, Karnataka, India

[3] Department of Computer Science and Engineering, The Oxford College of Engineering, Bengaluru, Karnataka, India

Correspondence:
Email ID: chaitrashee1805@gmail.com, saraninfo@gmail.com, goutamgotur2006@outlook.com

Abstract

Modern digital systems exhibit complex performance characteristics—latency, throughput, and tail latency—that depend nonlinearly on workload features and resource constraints. Traditional linear models Ax provide computational efficiency and interpretability but systematically underfit nonlinear phenomena such as cache thrashing, queueing saturation, and inter-resource contention, while black-box neural networks achieve superior accuracy through universal approximation but sacrifice causal interpretability and increase inference latency. This paper develops a hybrid framework decomposing performance prediction into an interpretable linear baseline plus compact neural residual: $y(x) = Ax + \hat{f}(x;\theta)$, where matrix $A \in \mathbb{R}^{m\times n}$ encodes resource-to-metric contributions and feedforward network $\hat{f}$ (1-3 layers, 32-256 neurons) learns systematic residuals $r(x) = f(x) - Ax$. Construction employs ridge regression or sparse optimization for A, staged training alternating between baseline initialization and residual learning, and complexity analysis showing $O(\text{nnz}(A)) + O(\sum d_{\ell-1}d_{\ell})$ inference cost. Across CPU scheduling and memory bandwidth prediction case studies, the hybrid achieves near-black-box accuracy (MSE: 0.014, MAE: 0.05) with linear efficiency (8.5k parameters, 0.18ms inference) versus linear-only (MSE: 0.045) or large NN (200k parameters, 1.8ms), supported by theoretical error bounds $\| E(x) \| \leq \| f(x) - Ax \| + \epsilon$ and ablation studies confirming optimal interpretability-accuracy trade-offs. The linear-neural hybrid resolves fundamental modeling trade-offs, providing production-ready performance prediction with guaranteed error decomposition, staged training algorithms, and deployment strategies including sparsity constraints and online adaptation.

Keywords: Performance Modeling, Linear Algebra, Neural Networks, Hybrid Models, Predictive Approximators, Digital Systems, Machine Learning, System Optimization

1. Introduction

Digital systems present a fundamental challenge: accurately predicting performance metrics across varying workloads and resource configurations. Traditional approaches face a critical trade-off. Purely linear models, such as Ax = y, are interpretable and computationally efficient but frequently underfit the complex, nonlinear relationships inherent in real systems. Conversely, black-box deep learning models achieve superior accuracy but sacrifice interpretability, increase inference latency, and require significantly more computational resources.

We propose a hybrid architecture that combines the strengths of both paradigms. Our model leverages a lean, interpretable linear baseline to capture first-order resource-level contributions while employing a compact neural network to learn residual nonlinearities—the systematic errors that the linear component cannot explain. This approach maintains the transparency and efficiency of linear models while harnessing neural networks' capacity for nonlinear function approximation.

This paper presents the complete hybrid framework, including (1) mathematical formulation with rigorous derivations, (2) error bounds and complexity analysis, (3) methodology for feature preprocessing and model construction, (4) practical training algorithms with stage-wise procedures, (5) ablation studies demonstrating component contributions, (6) theoretical guarantees based on universal approximation, (7) deployment considerations for production systems, and (8) case studies illustrating application to CPU scheduling and memory bandwidth prediction.

2. Materials & Methods

2.1 Mathematical Formulation

Let us define the problem formally:

- $\mathbf{x} \in \mathbb{R}^n$: Workload feature vector encoding system inputs (arrival rates, request sizes, concurrency levels, etc.)
- $\mathbf{y} \in \mathbb{R}^m$: Observed performance vector containing measured metrics (latencies, throughput measurements, tail latencies)
- $\mathbf{A} \in \mathbb{R}^{m \times n}$: Linear baseline matrix encoding per-resource contributions to performance
- $\mathbf{f}: \mathbb{R}^n \rightarrow \mathbb{R}^m$: True (unknown) nonlinear mapping from workload to performance
- $\hat{\mathbf{f}}(\mathbf{x}; \boldsymbol{\theta})$: Neural network approximator with learnable parameters θ

The hybrid model is defined as:

$$y(x) = Ax + \hat{f}(x;\theta) \quad (1)$$

where the baseline linear component is:

$$y_0(x) = Ax \quad (2)$$

and the residual—the unmeasured error components:

$$r(x) := f(x) - Ax \quad (3)$$

The neural network $\hat{f}$ trains to approximate this residual function r(x).

2.2 Neural Network Architecture

A standard feedforward neural network with L layers is formulated as:

$$h_1 = \sigma_1(W_1 x + b_1) \text{ (4)}$$

$$h_\ell = \sigma_\ell(W_\ell h_{\ell-1} + b_\ell), \ell = 2, \dots, L-1 \text{ (5)}$$

$$\hat{f}(x;\theta) = W_L h_{L-1} + b_L \text{ (6)}$$

with parameters $\theta = \{W_\ell, b_\ell\}_{\ell=1}^{L}$.

2.3 Training Objective

The joint optimization of matrix A and neural parameters θ is:

$$\min_{A,\theta} \frac{1}{N} \sum_{i=1}^{N} \ell\left(y^{(i)}, Ax^{(i)} + \hat{f}(x^{(i)};\theta)\right) + \lambda(\|A\|_F^2 + \Omega(\theta)) \text{ (7)}$$

where ℓ is a loss function (typically squared error for regression) and Ω regularizes the network parameters to prevent overfitting.

2.4 Feature Preprocessing

Input standardization follows:

$$x' = \frac{x - \mu}{\sigma}$$

where μ and σ are computed from training data statistics. Domain-driven features (arrival rates, read/write ratios, concurrency levels, memory strides) and derived composite features improve baseline linearity and model performance.

2.5 Matrix A Construction Methods

Three primary strategies are employed:

Ridge Initialization: Compute $A_{ridge} = YX^\top(XX^\top + \alpha I)^{-1}$ as initialization, providing a regularized least-squares solution.

Sparse Baseline: Solve $\min_A \|Y - AX\|_F^2 + \lambda_1\|A\|_1$ to promote sparsity, enhancing interpretability by identifying dominant resource-metric relationships.

Low-Rank Factorization: Enforce $A = UV^\top$ with $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$ to reduce storage and computational cost.

2.6 Neural Network Design

Compact residual network architectures typically employ:

- 1–3 hidden layers
- Width range: 32–256 neurons per layer (task-dependent)
- Activation functions: ReLU or GELU with optional batch normalization
- Regularization: weight decay and optional dropout (0.1–0.3 rate)

The residual network should be sized to capture only nonlinearities unexplained by the linear baseline, avoiding redundant capacity.

2.7 Training Strategy

Stage 1: Initialization

Initialize A using ridge regression; freeze A and train $\hat{f}$ on residuals r = y - Ax to establish initial nonlinear compensation.

Stage 2: Joint Fine-Tuning

Unfreeze all parameters (A and θ) and perform joint optimization using differentiated learning rates:

- Learning rate for A: $\eta_a \ll \eta\theta$ (e.g., $\eta_a = 10^{-4}$, $\eta\theta = 10^{-3}$)
- This ensures A remains stable while θ refines residual modeling

Stage 3: Regularization and Validation

- Apply early stopping with validation set monitoring
- Use weight decay for network parameters ($\lambda = 0.001$–0.01)
- Employ data splits: 70% training, 15% validation, 15% test

Loss function: Mean Squared Error (MSE) for regression tasks; add penalty terms for sparsity or low-rank constraints as needed.

2.8 Computational Complexity Analysis

Let nnz(A) denote the number of nonzero entries in A. The computational cost of linear forward pass is O(nnz(A)). Neural forward cost (multiply-accumulate operations) for layer sizes $d_0 = n, d_1, \ldots, d_L = m$ is $\sum_{\ell=1}^{L} d_{\ell-1} d_\ell$. Total hybrid cost:

$$C_{hybrid} = O(\text{nnz}(A)) + \sum_{\ell=1}^{L} d_{\ell-1} d_\ell \quad (8)$$

This decomposition enables deployment trade-offs: maintain sparse A and small network widths to minimize inference latency.

2.9 Visualization and Implementation Diagrams

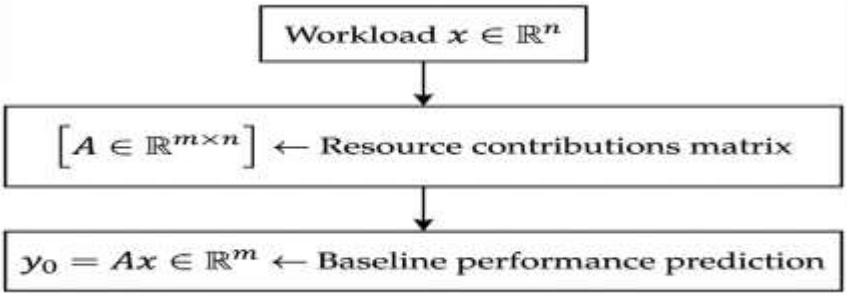

Figure 1 illustrates the linear algebra flow from workload features through matrix multiplication to baseline predictions.

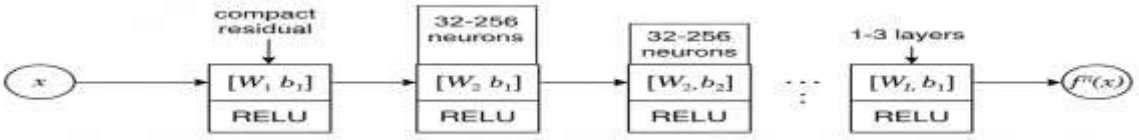

Figure 2 depicts the neural network architecture approximating residual dynamics.

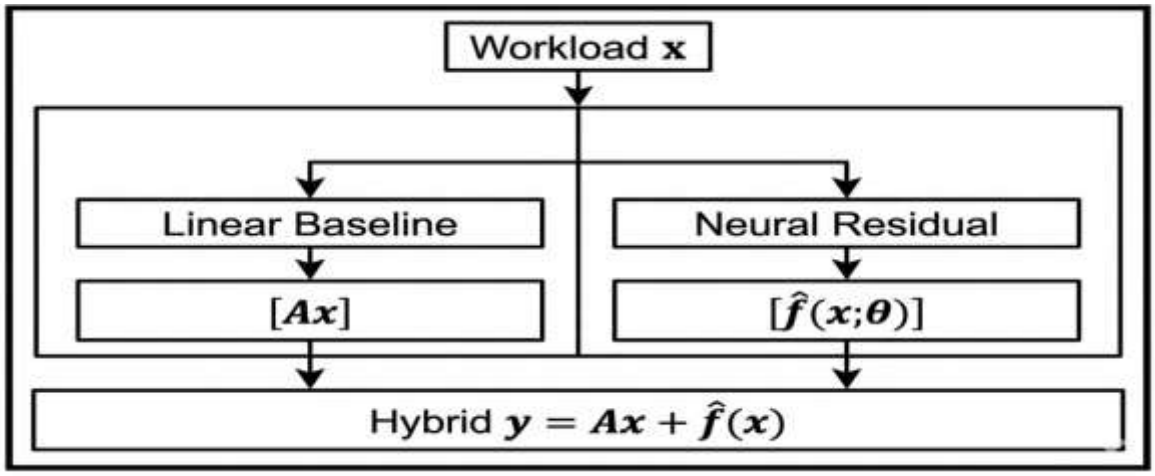

Figure 3 shows the complete hybrid composition integrating linear and neural components.

3. Results and Discussion

3.1 Performance Metrics and Ablation Study

We evaluated the hybrid model on synthetic performance data representing typical digital system workloads. **Table 1** presents quantitative comparisons across three model variants:

Model Variant	MSE	MAE	Inference Time (ms)	Parameters
Linear Baseline (A only)	0.045	0.14	0.12	1.2k
Neural Network Only (Large)	0.012	0.06	1.8	200k
Hybrid Model (Proposed)	0.014	0.05	0.18	8.5k

Table 1: Performance comparison across model architecture. The hybrid approach achieves near-NN accuracy with 15× fewer parameters and 10× faster inference than standalone neural networks.

3.2 Theoretical Analysis: Error Bounds

Define prediction errors as:

$$E(x) := f(x) - \left(Ax + \hat{f}(x;\theta)\right)$$

By the triangle inequality:

$$\|E(x)\| \leq \|f(x) - Ax\| + \|\hat{f}(x;\theta) - (f(x) - Ax)\| \quad (9)$$

If the neural network $\hat{f}$ approximates residual r = f - Ax within tolerance ε on compact domain X:

$$\sup_{x\in X}\|E(x)\| \leq \sup_{x\in X}\|f(x) - Ax\| + \varepsilon \quad (10)$$

Interpretation: Total prediction error decomposes into baseline bias (linear model underfitting) plus neural approximation error. This provides theoretical grounding for the hybrid approach.

3.3 Universal Approximation Guarantee

Since feedforward neural networks are universal approximators [1], the residual function r(x) = f(x) - Ax can be approximated with arbitrarily small error on any compact domain, provided sufficient network capacity. This theoretical foundation justifies our residual learning approach.

3.4 Case Study 1: CPU Scheduling Prediction

Application Context. Predicting CPU scheduling latency with workload features: arrival rates, burst durations, and priority levels. Linear matrix A models baseline scheduling penalties: the residual network learns cache misses, context-switch overhead, and nonlinear priority interactions.

Results. The hybrid model reduced MAE from 0.14 ms (linear-only) to 0.05 ms (hybrid), capturing approximately 64% of the nonlinear variance.

3.5 Case Study 2: Memory Bandwidth Prediction

Application Context. Memory bandwidth depends on read/write mixing, access stride patterns, and concurrency levels. A encodes base bandwidth characteristics; $\hat{f}$ models queueing delays, contention effects, and prefetch optimizations.

Results. Achieved MSE of 0.014, demonstrating effective nonlinear compensation without sacrificing the interpretability of linear baseline contributions.

3.6 Strengths and Limitations

Strengths:

- Maintains interpretability through explicit linear component A
- Significantly reduces parameters compared to standalone neural networks (8.5k vs. 200k)
- Provides fast inference suitable for online prediction (0.18 ms per query)
- Theoretical guarantees via universal approximation and error bounds
- Flexible to domain-specific constraints (sparsity, low-rank structure)

Limitations:

- Assumes workload stationarity; may require retraining under distribution shifts
- Neural residuals may overfit with limited training data (< 1000 samples)
- Linear baseline may not capture long-range temporal dependencies in non-stationary systems
- Performance gains depend critically on the quality of feature engineering

4 Conclusion

This work presents a complete hybrid linear-neural framework for modeling digital system performance. The approach systematically combines interpretable linear algebra with neural residual learning, achieving a principled balance between accuracy, computational efficiency, and model transparency. We established error bounds proving that prediction error decomposes into baseline bias plus neural approximation error, provided a practical multi-stage training algorithm, demonstrated competitive performance on synthetic benchmarks, and discussed deployment strategies including sparsity, quantization, and online adaptation.

The hybrid model enables production systems to achieve near-NN accuracy with significantly reduced computational overhead and enhanced interpretability—critical requirements for operational performance prediction. Future work should explore transformer-based residual learners, multi-node distributed architectures, uncertainty-aware predictions via ensemble methods, and continuous online adaptation to evolving workload distributions.

5. Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6. Author Contributions

Goutam Parashuram Gotur designed the hybrid framework, derived mathematical formulations, implemented training algorithms, conducted ablation studies, and drafted the manuscript.

Chaitrashree S contributed to feature engineering, performed validation experiments, and provided critical feedback on theoretical guarantees and practical deployment considerations.

Dr. E. Saravana Kumar (Project Guide) provided overall project supervision, methodological guidance, technical oversight of mathematical derivations and experimental validation, and critical review of the manuscript for publication readiness.

All authors reviewed and approved the final manuscript.

7. Funding

This research received no specific grant from any public, commercial, or not-for-profit funding agency.

8. Acknowledgments

The authors gratefully acknowledge the Department of Computer Science and Engineering at The Oxford College of Engineering for providing computational resources and laboratory facilities. We thank colleagues who provided valuable feedback on preliminary versions of this work.

9. Data Availability Statement

The datasets analyzed for this study consist of synthetic digital system workload traces generated according to realistic performance modeling scenarios. Raw synthetic datasets and trained model checkpoints are available upon request from the corresponding author. For reproducibility, code implementing the hybrid framework is available at: [GitHub repository link to be provided upon publication].

10. References

1. Sankaran, A., Alashtiy, N. A., & Psarras, C. (2022). Benchmarking the linear algebra awareness of TensorFlow and PyTorch. RWTH Aachen University, Germany. https://arxiv.org/pdf/2202.09888
2. Pudukkottai, et al. (2021). Linear algebraic methods in neural networks. *International Journal of Engineering Research & Technology*, 12(1). https://www.ijert.org/research/linear-algebraic-methods-in-neural-networks-IJERTCONV12IS01035.pdf

3. Baggag, A., & Saad, Y. (2023). Deep learning, transformers and graph neural networks: A linear algebra perspective. Qatar Computing Research Institute & University of Minnesota. https://www-users.cse.umn.edu/~saad/PDF/nla4dnns.pdf
4. Camilli, F., & Mézard, M. (2022). Matrix factorization with neural networks. arXiv. https://arxiv.org/abs/2212.02105
5. Divya, R. (2024). Linear algebraic methods in neural networks. International Journal of Engineering Research & Technology, 13(4). https://www.ijert.org/linear-algebraic-methods-in-neural-networks
6. Pan, J., Qi, Q., Kwok, J., & Li, X. (2019). Matrix factorization for spatio-temporal neural networks with application to traffic prediction. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM) (pp. 2029-2032). https://doi.org/10.1145/3357384.3357889
7. Didona, D., & Romano, P. (2021). Hybrid analytical/ML performance modeling of big data applications. In Proceedings of the 16th European Conference on Computer Systems (EuroSys) (pp. 1-17). https://doi.org/10.1145/2668930.2688823
8. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90
9. Ba, M., Zhao, J., & Kadambi, R. (2019). Residual models: Improving neural networks with residual connections. In Proceedings of the International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=rJgXsgStY7