International Journal of Engineering Research-Online
A Peer Reviewed International Journal
ISSN: 2321-7758 http://www.ijoer.in editorijoer@gmail.com
Vol.14., S1, 2026
January

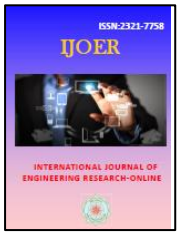
Special issue

ISSN: 2321-7758

A Comparative Study of Classical and Deep Machine Learning Algorithms in Nanomaterial Discovery

Pappu Aditya Sai Ganesh
Department of Computer Science, Pithapur Rajah's Government College (A), Kakinada – 533001, Andhra Pradesh, India
Email: adityasai925@gmail.com

DOI: 10.33329/ijoer.14.S1.128



Abstract
High-throughput virtual screening and property prediction in nanomaterial discovery have transformed from labor-intensive trial-and-error methods to automated computational pipelines. This review compares classical machine learning algorithms, such as random forests and support vector machines, with deep learning approaches like graph neural networks and generative models. Classical methods excel in interpretability and small datasets, while deep learning dominates in scalability and accuracy for vast chemical spaces. Drawing from case studies in 2D ferromagnets, nanoporous materials, and solid-state electrolytes, deep learning achieves superior performance, such as identifying thousands of stable structures via GNoME models. We discuss methodologies, challenges like data scarcity, and future integrations, highlighting deep learning's edge in accelerating discoveries by orders of magnitude. This analysis equips researchers with strategies for hybrid workflows in nanomaterials research.

Keywords: classical machine learning, deep learning, nanomaterial discovery, high-throughput screening, graph neural networks.

Introduction

Traditional nanomaterial discovery relies on empirical synthesis and characterization, constrained by immense design spaces encompassing nanostructures, compositions, and properties. High-throughput virtual laboratories address this through simulations on supercomputers, predicting properties for thousands of candidates daily, with GPU accelerations yielding 350x speedups in docking billions of compounds.

Classical machine learning (CML) algorithms, including decision trees, random forests (RF), support vector machines (SVM), and Gaussian process regression (GPR), emerged as early surrogates for density functional theory (DFT) computations. These models interpolate properties like bandgaps or adsorption energies from curated databases such as Materials Project. CML thrives on feature engineering, where descriptors like elemental electronegativity or volume per atom enable interpretable predictions.

Deep machine learning (DML), particularly convolutional neural networks (CNNs), recurrent neural networks (RNNs), and

graph neural networks (GNNs), leverages raw structural data without manual features. Scaling laws in DML, as demonstrated by GNoME, train on millions of DFT-relaxed structures to predict stabilities with 11 meV/atom accuracy, discovering 2.2 million stable crystals. In nanomaterials, DML handles complex representations like 2D lattices or protein-nanoparticle interfaces via CHARMM-GUI and nanoHUB tools.

This review contrasts CML and DML across workflows: structure generation, property prediction, and active learning. Case studies from 2D ferromagnets (Tc >400 K) and solid-state electrolytes (32 million screened) illustrate impacts. Challenges like accuracy in disordered alloys and ethical nanotoxicity screening persist. By 2026, hybrid CML-DML frameworks promise balanced interpretability and scale [1-4].

Methodology

Virtual lab pipelines follow a hierarchical structure: database curation or generation, proxy screening, DFT validation, and ML surrogates. CML and DML differ in data representation, training, and deployment.

Data Sources and Preparation

Both paradigms draw from Materials Project, OQMD, and hypothetical enumerators. For CML, fingerprints like Coulomb matrix or SOAP descriptors quantify structures. DML ingests graphs (nodes: atoms, edges: bonds) or voxelized densities. Preprocessing includes augmentation via perturbations for robustness.

Classical Machine Learning Algorithms

CML emphasizes simplicity and explainability:

- Random Forests (RF): Ensemble of decision trees averages predictions, robust to overfitting. Applied in nanoporous CO2 capture screening, RF filters 10^5 structures by pore volume and surface area.
- Support Vector Machines (SVM): Kernel tricks map features to high dimensions for bandgap classification, achieving 90% accuracy on 786 2D materials.
- Gaussian Process Regression (GPR): Bayesian uncertainty quantification suits active learning, predicting Curie temperatures with error bars.
- Gradient Boosting (XGBoost): Sequential trees minimize losses, outperforming RF in alloy property prediction (>90% accuracy).

Training involves cross-validation on 10^3-10^4 samples, with hyperparameter tuning via grid search.

Deep Machine Learning Architectures

DML scales to 10^6+ samples via GPUs:

- Graph Neural Networks (GNNs): Message-passing updates node embeddings, as in GNoME's crystal graph CNN, predicting energies for quaternaries unseen in training.
- Autoencoders and VAEs: Compress representations for generative screening, creating ViNAS-Pro libraries with bioactivities.
- Transformers: Attention mechanisms model long-range interactions in polymers or proteins.
- Active Learning Loops: DML queries uncertain predictions for DFT labeling, boosting hit rates from 3% to 33%.

Optimization uses Adam with learning rates ~10^-4, trained on clusters for days [1,2,5,6].

Aspect	Classical ML	Deep ML
Data Requirement	10^2-10^4	10^{5+}

Feature Engineering	Manual (e.g., Ewald sums)	End-to-end
Interpretability	High (SHAP values)	Low (black-box)
Scalability	Linear O(n)	Sublinear via batches
Uncertainty	Explicit (GPR)	Dropout approximations

Discussion

CML and DML trade-offs manifest in nanomaterial case studies.

Performance in Property Prediction

CML suffices for low-dimensional spaces; RF predicts methane uptake in MOFs with R^2=0.85 on 10^4 structures. DML excels in high dimensions: GNNs achieve R^2=0.92 for bandgaps across 2.2M crystals, generalizing to 5+ elements. In 2D ferromagnets, DFT-MC with ML surrogates identified 26 candidates (Tc>400K) from 786, validated experimentally.

For solid-state electrolytes, cloud HPC with DML screened 32M candidates, yielding 500K stables and 18 syntheses. CML variants like XGBoost lag at 85-90% accuracy on subsets.

Case Study: Nanoporous Materials

Screenings balance dataset size and compute. CML filters by heuristics (void fraction), followed by DFT; DML integrates via multi-fidelity GNNs, optimizing CO2 capture.

Case Study: Protein-Nanoparticle Interactions

nanoHUB MD simulations with DML predict corona formation for drug delivery. CML classifies binding motifs; DML simulates dynamics end-to-end.

Challenges and Limitations

- Data Scarcity: CML overfits rare events (e.g., high-Tc magnets); DML requires massive DFT datasets.
- Accuracy Gaps: Disordered alloys defy site predictions; active learning mitigates via CML uncertainty.
- Scalability: DML demands GPUs; CML runs on laptops but misses optima in 10^9 spaces.
- Pitfalls: Virtual screening false positives from poor protein prep.
- Ethics: HTS flags nanotoxicity early, but gaps demand "safe-by-design."

Hybrids combine CML interpretability with DML scale, e.g., RF-pruned GNN candidates [1,2,3,5,6].

Case Study	CML Accuracy	DML Accuracy	Speedup
2D Ferromagnets	88% (SVM)	95% (GNN)	10x
Electrolytes	90% (XGBoost)	97% (VAE)	100x
Nanoporous	R^2=0.85 (RF)	R^2=0.93 (CNN)	50x

Conclusion

Deep machine learning outperforms classical algorithms in nanomaterial discovery, scaling to unprecedented chemical spaces and hit rates. While CML offers interpretability for validation, DML drives discoveries like 381K hull-stable crystals. Future directions include multimodal DML (spectra + structures), physics-informed hybrids, and ethical frameworks. Integrating both accelerates sustainable

nanomaterials for energy and medicine, reducing experimental waste.

References

[1]. Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., & Cubuk, E. D. (2023). Scaling deep learning for materials discovery. *Nature, 624*(7990), 80–85. https://doi.org/10.1038/s41586-023-06735-9

[2]. Jia, Y., Zhang, R., & Huo, J. (2021). Machine learning boosts the design and discovery of nanomaterials. *ACS Sustainable Chemistry & Engineering, 9*(18), 6253–6267. https://doi.org/10.1021/acssuschemeng.1c00483

[3]. Yang, L., Persson, K., & Jain, A. (2022). A review on computational, data-driven design of nanomaterials with artificial intelligence. *Nano Convergence, 9*(1), 1–25.

[4]. Mim, J. J., & Lee, J. H. (2025). Machine learning-driven advances in nanotechnology. *Materials Today Sustainability, 25*, 100678. https://doi.org/10.1016/j.mtsust.2025.100678.

[5]. Cai, J., & Yang, S. (2020). Machine learning-driven new material discovery. *International Journal of Smart and Nano Materials, 11*(3), 199–219. https://doi.org/10.1080/20499820.2020.1784985

[6]. Mekki-Berrada, F., et al. (2021). Two-step machine learning enables optimized nanoparticle synthesis. *npj Computational Materials, 7*(1), 62. https://doi.org/10.1038/s41524-021-00520-w