Machine Learning Project Topics

Format and expectations. 2–3 students per team. Each project must deliver *(i)* open, reproducible code and benchmarks, and *(ii)* a paper-style report (8–12 pages) covering theory, implementation details, and applications. Topics balance theory & practice, and reference A* venues (NeurIPS, ICML, ICLR, AAAI, CVPR/ICCV) from 2023–2025 where possible.

1 Explainable AI (XAI): Tuning Models for Faithful Attributions

Idea. Implement and evaluate methods that improve the faithfulness and discriminativeness of post-hoc explanations. Reproduce Distractor Erasure Tuning (DiET) and the ICLR 2025 "How to Probe" framework; benchmark across backbones and datasets (e.g., ImageNet, CIFAR, GLUE).
Scope. (i) Formalize metrics (sufficiency/necessity, deletion/insertion, pointing game), (ii) compare standard attribution (GradCAM/IG) vs. DiET-style training, (iii) analyze sensitivity to pre-training and probing design.
Starting points. [1, 2].

2 Bayesian Deep Learning: Posterior Geometry and Flatness-aware BNNs

Idea. Study how posterior geometry (modes, symmetries, flatness) affects uncertainty quality. Reproduce Flat-Seeking BNNs and the symmetry-aware posterior exploration; run vision (CIFAR/ImageNet) and tabular tasks.
Scope. (i) Implement sharpness-aware variational posteriors, (ii) visualize/post-process symmetries, (iii) compare uncertainty metrics (ECE, NLL, OOD).
Starting points. [3, 4].

3 Graph Transformers at Scale: Global, Sparse and Hybrid Attention

Idea. Build and benchmark scalable graph transformers: GOAT (approx. global attention) vs. Exphormer (expander-graph sparsity) and a recent global-to-local variant. Evaluate on long-range graph benchmarks (OGB, ZINC, Peptides).
Scope. (i) Complexity vs. accuracy tradeoffs; (ii) homophily/heterophily robustness; (iii) ablations on positional encodings and global-token designs.
Starting points. [5, 6, 7].

4 Label-aware Graph Classification: Distillation for Graph-level Tasks

Idea. Investigate label-attentive distillation to align node embeddings with graph labels to improve graph-level classification. Extend LAD-GNN to new backbones; study calibration and sample efficiency.
Starting points. [8].

5 Pretrained LMs under Resource Constraints: 4-bit Fine-tuning and Adapters

Idea. Systematic study of QLoRA-style low-memory finetuning on several LLMs and tasks (instruction-following, classification). Compare full fine-tuning, LoRA, and QLoRA across GPUs, quantizers (NF4), and datasets.
Starting points. [9].

6 Time-series Forecasting: Are (Cross-)Attentions Necessary?

Idea. Reassess Transformer design for TSF: PatchTST vs. linear baselines (DLinear) vs. the NeurIPS 2024 CATS model that removes self-attention. Evaluate long-horizon multivariate forecasting under strict protocol.
Starting points. [10, 11, 12].

7 Autoencoders for Anomaly Detection: From Reconstruction to Semi-supervised Signals

Idea. Beyond pure reconstruction error, exploit few labeled anomalies to separate manifolds and improve detection. Compare VAE/AE/AMA-style methods on time-series (MSL/SMAP) and vision (MVTec).
Starting points. [13, 14].

8 Adversarial Robustness via Diffusion Priors: Purification and Counter-attacks

Idea. Implement score-based test-time optimization (ScoreOpt) and evaluate under robust, recent attack suites including ones tailored to diffusion purification (e.g., DiffAttack). Provide strong AutoAttack/BPDA/EOT evaluations.
Starting points. [15, 16, 17].

9 Personalized Federated Learning under Heterogeneity and Budgets

Idea. Study client heterogeneity (data & resources) using pFedGate's sparse personalization. Evaluate convergence/accuracy/efficiency vs. baselines (Per-FedAvg, FedProx, FedSpa).
Starting points. [18].

10 Causal Representation Learning from Multiple Domains

Idea. Learn identifiable causal latents from unpaired multi-domain observations and compare with weak-invariance methods. Apply to synthetic (linear SCM) and real multi-view datasets.
Starting points. [19, 20].

11 Generative Models in Practice: Diffusion vs. GANs vs. VAEs

Idea. A rigorous empirical and compute-aware comparison of diffusion models, GANs, and VAEs on conditional generation. Include FID/KID/IS, compute/FPS, and controllability; analyze classifier guidance and negative prompts.
Starting points. [21, 22].

12 Fair Representation Learning with Certificates + Bias Audits

Idea. Combine algorithmic fairness with practical certificates (upper bounds on downstream unfairness). Reproduce FARE and conduct a bias audit on a vision or recsys pipeline; report fairness–utility trade-offs.
Starting points. [23, 24].

Deliverables for each project

- Reproducible code (Python) with configuration files and experiment runners; baseline re-implementations where needed.
- Benchmarks against strong SOTA or standard baselines; report mean±std over multiple seeds; include compute budget.
- Paper-style report with: problem statement, related work, methods, experimental setup, results/ablations, limitations/ethics checklist.

References

[1] R. Yeh, S. Zhang, Y. Liu, et al. *Discriminative Feature Attributions: Bridging Post Hoc Explainability and Model Robustness via Distractor Erasure Tuning (DiET)*. NeurIPS 2023. Link.

[2] S. Gaire, S. Agarwal, V. Prabhu, et al. *How to Probe: Simple Yet Effective Techniques for Improved Post-hoc Explanations*. ICLR 2025. PDF.

[3] V.-A. Nguyen, T.-L. Vuong, H. Phan, et al. *Flat Seeking Bayesian Neural Networks*. NeurIPS 2023. PDF.

[4] O. Laurent, M. Baradel, J. Aubert, et al. *A Symmetry-Aware Exploration of Bayesian Neural Network Posteriors*. ICLR 2024. Link.

[5] D. Kong, H. Liu, P. Konda, et al. *GOAT: A Global Transformer on Large-scale Graphs*. ICML 2023. Link.

[6] H. Shirzad, A. Velingker, B. Venkatachalam, D. Sutherland, A. K. Sinop. *Exphormer: Sparse Transformers for Graphs*. ICML 2023. PDF.

[7] Y. Zhang, Q. Zhang, Y. Wang, et al. *G2LFormer: Global-to-Local Attention Scheme in Graph Transformers*. 2025 (preprint). PDF.

[8] X. Hong, W. Li, C. Wang, et al. *Label Attentive Distillation for GNN-Based Graph Classification*. AAAI 2024. Link.

[9] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer. *QLoRA: Efficient Finetuning of Quantized LLMs*. NeurIPS 2023. Link.

[10] Y. Nie, N. H. Nguyen, J. Kalagnanam. *A Time Series is Worth 64 Words: Long-term Forecasting with Transformers (PatchTST)*. ICLR 2023. PDF.

[11] A. Zeng, M. Chen, L. Zhang, et al. *Are Transformers Effective for Time Series Forecasting?* AAAI 2023. Link.

[12] D. Kim, J. Park, J. Lee, H. Kim. *Are Self-Attentions Effective for Time Series Forecasting? (CATS)*. NeurIPS 2024. PDF.

[13] F. Angiulli, F. Fassetti, L. Ferragina. *Reconstruction Error-based Anomaly Detection with Few Outlying Examples*. 2023. arXiv.

[14] S. N. Khan, A. A. Alhudhaif, et al. *Variational Autoencoder for Anomaly Detection: A Review*. 2024. arXiv.

[15] B. Zhang, Z. Lin, W. Sun, et al. *Enhancing Adversarial Robustness via Score-Based Optimization*. NeurIPS 2023. PDF.

[16] Z. Li, Y. Li, Z. Zhang, et al. *DiffAttack: Evasion Attacks Against Diffusion-Based Adversarial Purification*. NeurIPS 2023. Link.

[17] S. Lee, J. Kim, H. Kim, et al. *Robust Evaluation of Diffusion-Based Adversarial Purification*. ICCV 2023. PDF.

[18] D. Chen, L. Yao, D. Gao, B. Ding, Y. Li. *Efficient Personalized Federated Learning via Sparse Model-Adaptation (pFedGate)*. ICML 2023. PDF.

[19] N. Sturma, C. Squires, M. Drton, C. Uhler. *Unpaired Multi-Domain Causal Representation Learning*. NeurIPS 2023. PDF.

[20] K. Ahuja, A. Mansouri, Y. Wang. *Multi-Domain Causal Representation Learning via Weak Distributional Invariances*. AISTATS 2024 (PMLR). PDF.

[21] P. Dhariwal, A. Q. Nichol. *Diffusion Models Beat GANs on Image Synthesis*. NeurIPS 2021. PDF.

[22] M. T. Mirza, S. T. M. Rawan, et al. *Comparative Analysis of Generative Models: Enhancing Image Synthesis with VAE, GAN, and Diffusion*. 2024. arXiv.

[23] N. Jovanovic, M. Balunovic, D. I. Dimitrov, M. Vechev. *FARE: Provably Fair Representation Learning with Practical Certificates*. 2023. PDF.

[24] J. Buolamwini, T. Gebru. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. PMLR 2018. PDF.