Artificial Intelligence and Deep Learning

PowerGraph

Assignment 2: Technical Reflective Report

Ghulam Ahmed
5746597

May 5, 2026

Contents

1 Introduction

A power grid stays stable as long as no transmission line is overloaded. When a line trips (lightning strike, equipment failure, tree contact), power reroutes through the remaining lines, sometimes overloading them, causing them to trip too. That chain reaction is a cascading failure, and causes large blackouts Gjorgiev and Sansavini (2022).

Cascading failures in transmission grids are rare but catastrophic. The 2003 Northeast blackout left 50 million people without power and caused an estimated $4–10 bn in economic losses (Electricity Consumers Resource Council, 2004; ICF Consulting, 2003). The cascade signal depends on topology: power redistributes along edges by Kirchhoff's laws, so flattening the network to a tabular vector discards the relational structure geometric deep learning (DL) can use (Battaglia et al., 2018; Varbella et al., 2023). This project tests whether that architectural argument holds in practice on the PowerGraph IEEE-24 binary classification subset Varbella et al. (2024). GINe Hu et al. (2020), a Graph Isomorphism Network variant that ingests edge features (line ratings, power flows) directly through edge-aware message passing, is benchmarked against an XGBoost baseline Chen and Guestrin (2016) with multi-seed confidence intervals and operational threshold optimisation.

XGBoost reaches 0.9947 ± 0.0014 balanced accuracy (BalAcc) against GINe's 0.9890 ± 0.0052 across five seeds (Table 1), which is expected based on the tabular machine-learning (ML) literature (Shwartz-Ziv and Armon, 2022; Grinsztajn et al., 2022) once node features are pooled into a per-graph vector. GINe excels where no tree ensemble can compete: edge-level prediction at PR-AUC 0.7629 ± 0.0372 (Table 2) and cross-grid transfer under contrastive pre-training (§3.3). DL excels on architectural capabilities rather than headline accuracy, and these allow regulated deployment (Battaglia et al., 2018; Bronstein et al., 2021).

2 Problem and Dataset

2.1 Problem framing

The decision problem is binary: given pre-outage node and edge features, predict whether one or more line removals increases demand-not-served (DNS) above zero Varbella et al. (2024). A secondary edge-level task asks, for each transmission line, whether it trips during the resulting cascade given the same pre-outage state. The target user is a control-room Transmission System Operator (TSO) running real-time contingency analysis every 30 minutes (North American Electric Reliability Corporation, 2024; Nakiganda and Chatzivasileiadis, 2023). The system must return a result in under a second per contingency, supporting operator screening rather than autonomous control. Only ≈20.1% of cases are positive, and missing a cascade costs far more than a false alarm. Aggregate accuracy hides this, which is why §5 focuses on threshold tuning and PR-AUC instead. Interactive demo available at https://powergraph.gahmed.com

2.2 Dataset

This project uses the unmodified PowerGraph Varbella et al. (2024) IEEE-24 cascade subset for its operational relevance: the highest binary positive rate (≈20.1% versus ≈8.96% for IEEE-39, Figure 12) and the smallest network (24 buses, 38 lines, Figure 14), which together maximise classification signal while keeping hyperparameter search feasible. Labels come from Varbella et al. (2024), generated by running the Cascades AC simulator Gjorgiev and Sansavini (2022). Each graph is a single pre-outage state. §5's IEEE-24→IEEE-39 zero-shot transfer tests whether training on one grid generalises to another.

3 ML vs DL: Comparison and Justification

3.1 Candidate classical baseline

My classical baseline was XGBoost Chen and Guestrin (2016) with 200 trees, max-depth 6, `scale_pos_weight` set to the negative-to-positive class ratio ($n_0/n_1 \approx 3.96$) to handle the 20.16% imbalance. Each graph was reduced to a 42-dimensional vector of node and edge moments, line-loading ratios, and NetworkX graph statistics (§A.4.1). Across five seeds on IEEE-24, XGBoost reached 0.9947 ± 0.0014 BalAcc; Optuna-tuned GINe reached 0.9890 ± 0.0052. A GCN Kipf and Welling (2017) that ignores edge features collapsed to 0.6656 ± 0.0279 (Table 1), establishing that edge features are necessary. Isolating message-passing's contribution beyond edge information would require an MLP-on-edge-feature baseline §A.4.2.

3.2 How tree ensembles and GNNs differ

Tree ensembles partition feature space along axes and recover interactions by repeated splitting. Message-passing GNNs Gilmer et al. (2017) aggregate neighbourhood information and are permutation-invariant Bronstein et al. (2021), encoding graph structure into the architecture itself. Because Kirchhoff's laws explain how power redistributes after a line trip, the GNN's relational bias is closer to the data-generating process (an inductive-bias framing) than a flattened feature vector Bronstein et al. (2021).

XGBoost's aggregate accuracy advantage is expected. Shwartz-Ziv and Armon (2022) show that DL rarely beats gradient-boosted trees on tabular data. Grinsztajn et al. (2022) trace the cause to high samples-per-feature ratios and low-order interactions which is the case for IEEE-24 as thousands of graphs are summarised by 42 statistics. This is visible in the GNN+XGBoost ensemble (Figure 1): SHAP Lundberg and Lee (2017) attributes 80–85% of the ensemble's decisions to GINe-embedding dimensions, yet adding those embeddings to XGBoost lifts accuracy by only 0.035 pp. SHAP measures importance within a model; lift tests whether embeddings add information beyond hand-crafted statistics and on this small topology they do not.

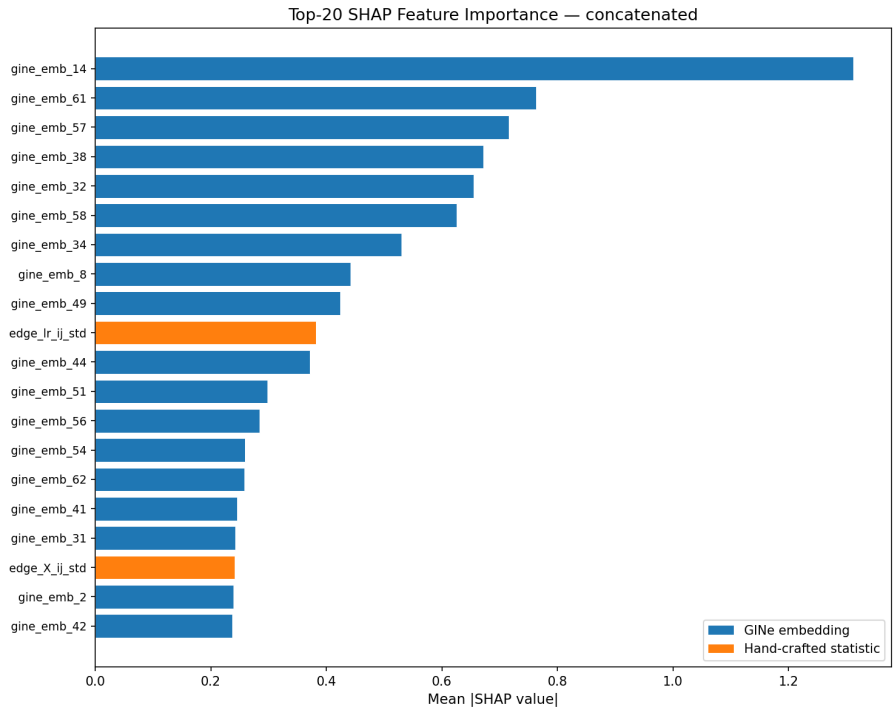

Figure 1: SHAP top-20 features for the GINe+XGBoost ensemble.

3.3 DL justification and qualifying conditions

DL is chosen for three reasons. First, edge-level prediction. GINe estimates per-edge failure probabilities at PR-AUC 0.7629 ± 0.0372 and AUROC 0.9976 ± 0.0007 once betweenness centrality (how often a line lies on rerouted power paths) and load ratio (current flow as a fraction of thermal capacity) are added (Table 2). A per-edge XGBoost head is possible in principle but requires per-topology re-engineering and loses relational coherence across simultaneous trips.

Second, cross-grid transfer. SimGRACE contrastive pre-training on all four PowerGraph topologies (IEEE-39 included without labels during SSL) followed by supervised fine-tuning on IEEE-39 was statistically indistinguishable from native training (0.9892 vs 0.9882 Bal-Acc, ± 0.005 SD), with an isolated SSL benefit of $+0.94$ pp over random-init fine-tuning. A frozen SimGRACE backbone reaches 0.7132 BalAcc against random-init's 0.5000, confirming a label-free objective recovers task-relevant structure without any target-grid labels Xia et al. (2022). Figure 2 shows a single-source (IEEE-24-only) encoder applied to IEEE-39 reaches AUROC 0.7220 zero-shot and AUROC 0.8677 after head-only fine-tuning. No tabular method offers an equivalent mechanism (Shwartz-Ziv and Armon, 2022; Grinsztajn et al., 2022). The t-SNE (Figure 3) shows embeddings coloured by grid and class; grid topology dominates both pre-trained and random-init projections, which is evidence of class-relevant structure.

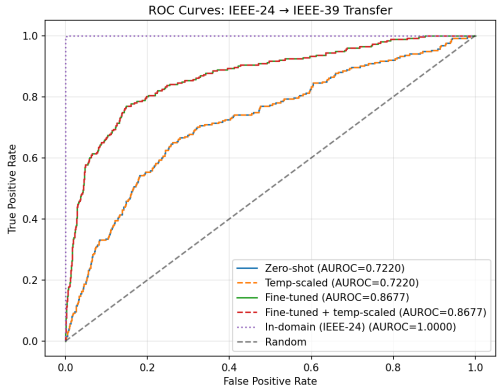

Figure 2: ROC curves for IEEE-24 → IEEE-39 single-source transfer. The IEEE-24 in-domain ceiling and random-classifier baseline are shown for reference. Temperature-scaled variants overlap the un-scaled curves (ROC is threshold-independent).

Figure 3: SimGRACE t-SNE coloured by grid (top) and class (bottom). Grid topology dominates both panels; class separation is not strongly visible in 2-D projection

Third, directionality. Power flows along directed paths, so standard GINe discards relational signal by treating edges as undirected. Wrapping GINe with a directed aggregation layer Rossi et al. (2024) reaches 0.9919 ± 0.0011 BalAcc with $\sim 5\times$ lower seed variance. The variance reduction matters more than the accuracy gain, reflecting a stronger inductive bias that reduces sensitivity to random initialisation. The mixing parameter converging to $\alpha \approx 0.5$ confirms that line-trip influence propagates both upstream and downstream (§A.4.3).

On clean IEEE-24, XGBoost matches GINe on accuracy at far lower training cost Okoyomon and Goebel (2026). The choice is a capability-set comparison instead of aggregate accuracy, i.e. GNN when per-edge outputs, cross-grid transfer, or distributional robustness matter, and XGBoost otherwise.

4 Design Decisions and Implementation

4.1 Architecture choice and rationale

I chose GINe as the production model because it matches the problem structure. Whether a line trips depends mostly on edge features, i.e. line ratings, power flows, reactance and GINe Hu et al. (2020) reads these directly through edge-aware message passing (§A.4.4), whereas GCN cannot. I validated this choice empirically against alternatives: GCN Kipf and Welling (2017), GAT Veličković et al. (2018), GIN, GINe and the hybrid graph transformer GPS Rampášek et al. (2022). GAT's attention added variance without lifting accuracy on IEEE-24; vanilla GIN, lacking edge-aware updates, is similar to GCN ablation. The IEEE-24 versus IEEE-39 results are the most informative evidence. GPS narrowly beat GINe on IEEE-24 (0.9908 ± 0.0019 vs 0.9890 ± 0.0052 BalAcc, one-third the seed variance) but underperformed by 0.75pp on IEEE-39 with substantially higher variance. Multi-head attention pays off on small dense topologies but needs per-grid tuning at scale (Xu et al., 2019; Bronstein et al., 2021).

4.2 Training setup

Every training choice is based on a specific property of the data. I used focal loss Lin et al. (2017) because of the 20% positive rate, with α capped from inverse class frequency (§A.4.5). I rejected weighted cross-entropy after seeing unstable gradients at the $>100\times$ weight ratios it produced at the tails. I used Adam Kingma and Ba (2015) with the PowerGraph paper's $lr = 10^{-3}$ as a Bayesian search anchor. Optuna returned $hidden\_dim = 64$, $num\_layers = 4$ and `mean_max` pooling instead of the paper's defaults of 32/3/max. I early-stopped on validation BalAcc (operational target) instead of loss. Determinism was enforced across NumPy, PyTorch, and CUDA.

4.3 Evaluation strategy

The evaluation design reflects three principled choices. First, a stratified 85/5/10 train/validation/test split matching Varbella et al. (2024), with the test set held back until the final stage. Second, every headline metric is reported as the mean ± standard deviation across five seeds (23, 34, 456, 789, 999), so each comparison shows seed-level uncertainty rather than a single point. I acknowledge Errica et al. (2020) that fixed splits across seeds capture initialisation variance but not data-split variance, which nested cross-validation would. Third, the metrics are strategically chosen, and not based on convention. PR-AUC takes priority over ROC-AUC because ROC overstates performance under imbalance Saito and Rehmsmeier (2015). I optimised the decision threshold for asymmetric error costs: Youden's J Youden (1950) at $\tau = 0.68$ raises precision by 3.4pp over the default 0.5 at the cost of just 0.55pp in miss rate which is the key trade-off for an operator. I report calibrated uncertainty two ways: MC Dropout Gal and Ghahramani (2016) entropy-gated selective prediction (model-centric: how uncertain is this prediction?) and split conformal prediction sets (Vovk, 2013; Angelopoulos and Bates, 2023). Both address Article 14's human-in-the-loop requirement; see §5.1 for results.

4.4 Role of AI tools

AI tools (Claude Code, GitHub Copilot) generated boilerplate, e.g. MLflow logging, plot helpers, dataset loaders and helped with formatting. Every cited paper was verified at the primary source and every metric comes from the `scripts/` pipeline. No AI tool generated numerical results. I overrode AI suggestions twice: rejecting weighted cross-entropy after observing gradient instability (§4.2) and early-stopping on BalAcc rather than validation loss.

5 Results and Interpretation

5.1 Results and baseline delta

Graph-level metrics on IEEE-24 are near-saturated (Table 1). GINe reached PR-AUC 0.9980 ± 0.0017 across five seeds. Edge-level prediction (Table 2) is more informative, with PR-AUC 0.7629 ± 0.0372. These per-line outputs can not be produced by any flat-feature model. XGBoost on 42 hand-engineered statistics slightly outperformed GINe at graph level (PR-AUC 0.9989 ± 0.0008), as Shwartz-Ziv and Armon (2022) and Grinsztajn et al. (2022) predict for tabular settings of this size (§3.2). I retained GINe because per-edge outputs, calibrated uncertainty, and cross-grid transfer matter for deployment and are not given by XGBoost, even though graph-level numbers tie.

Threshold optimisation (Figure 4) shifted the dominant error from false alarms to missed cascades: $\tau = 0.684$ lifted precision from 0.94 to 0.97 at the cost of 0.55 pp recall. That trade-off is ultimately a policy decision about what error costs the operator can accept.

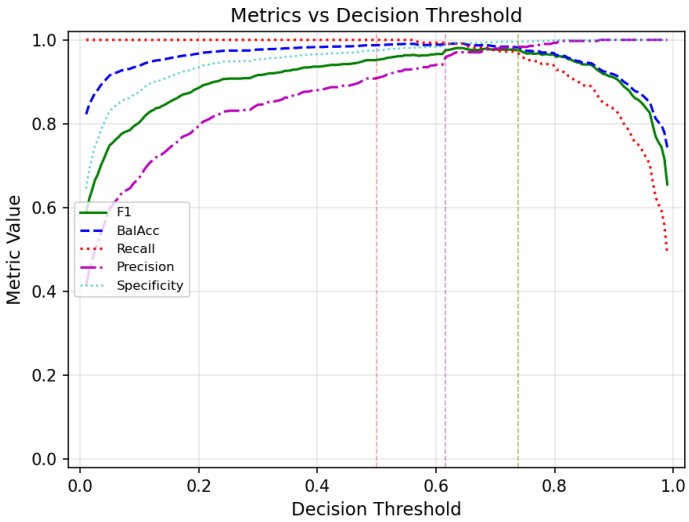

Figure 4: Threshold optimisation.

GINe outputs are calibrated for selective prediction two ways. MC Dropout Gal and Ghahramani (2016) entropy gating across stochastic forward passes reaches 0.9995 BalAcc at 50% coverage (Figure 5) (§A.4.6).

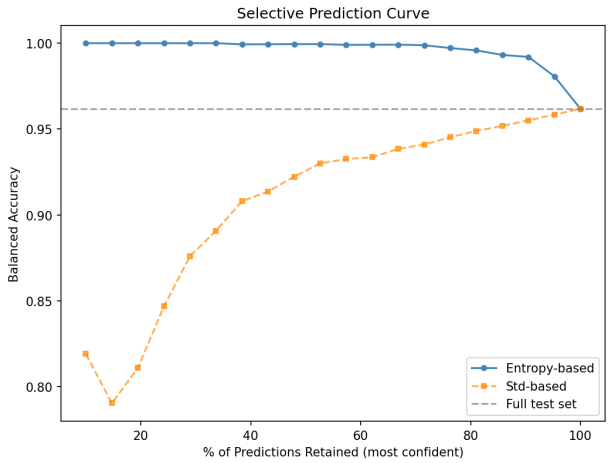

Figure 5: MC-Dropout entropy gating.

Split conformal prediction (Vovk, 2013; Angelopoulos and Bates, 2023) returns prediction sets with operational structure: a singleton commits to one label, a pair flags uncertainty for operator review, and an empty set escalates an out-of-distribution input (§A.4.6). Empirical coverage tracks the target across $\alpha \in [0.01, 0.20]$ with a 99.0% singleton rate at $\alpha = 0.05$ (Figures 6, 7). Both methods address the core requirements of Article 14(4)(d) by making the auto-clear/escalate boundary a regulator-tunable parameter rather than an arbitrary confidence cutoff.

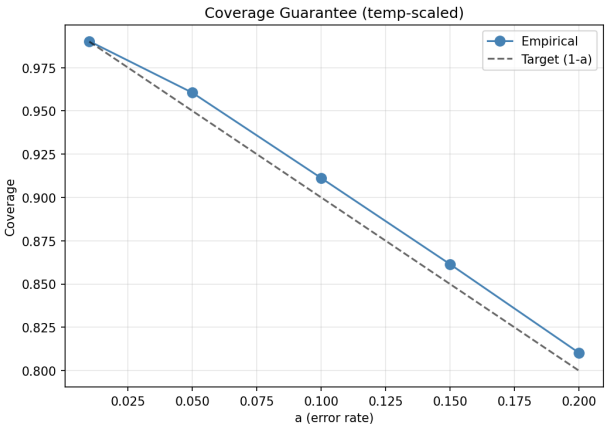

Figure 6: Empirical coverage meets or exceeds the $1-\alpha$ target across $\alpha \in [0.01, 0.20]$; 96.0% at $\alpha = 0.05$.

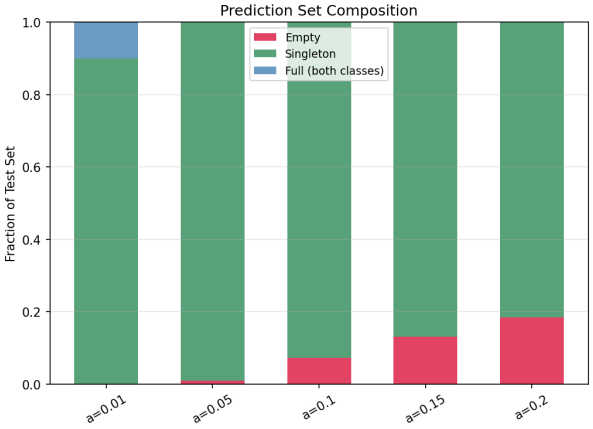

Figure 7: Prediction set composition by α: singletons reach 99.0% at $\alpha = 0.05$. Pairs (visible at $\alpha = 0.01$, ~10%) and empty sets (rising with α) both route to the operator.

5.2 Error analysis and interpretability

Errors cluster around specific failure modes. The multi-class head over-predicted cascades: 1,966 safe graphs (Category D, no DNS) were labelled as cascades (Category A, high DNS), with B and C as intermediate DNS levels which caused the gap between weighted F1 (0.83) and macro F1 (0.75). At edge level, binary-F1 was around 0.42 (network-centric mode) because the typical cascade trips only one or two of ~38 edges, so ranking was near-perfect (AUROC 0.998) but precision-at-k is capped by extreme imbalance, consistent with what Kazim et al. (2025) report for network-centric features.

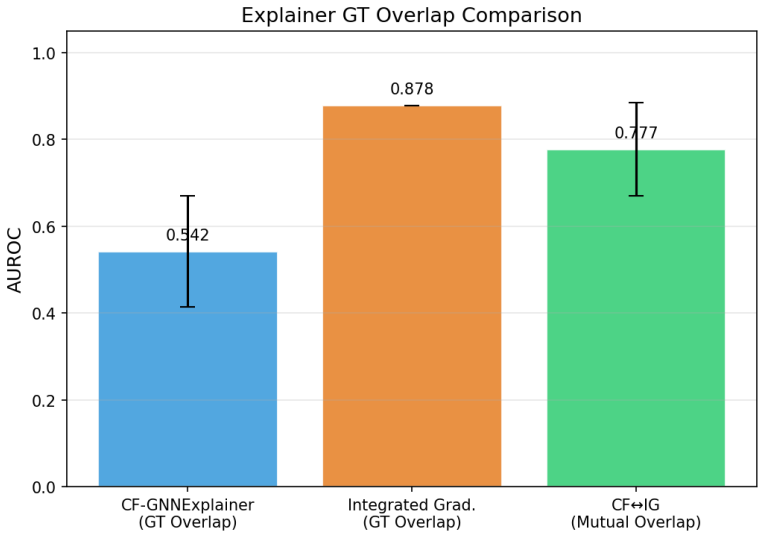

Figure 8: CF-GNNExplainer vs Integrated Gradients overlap with ground-truth `exp.mat`

Integrated Gradients Sundararajan et al. (2017) recovered the ground-truth cascading edges with AUROC 0.878±0.146 against PowerGraph's `exp.mat` masks (Figure 8). The same explanations scored Fidelity$^+ = -0.729$: removing the top-attributed edges did not change predictions, so the attributions are decorative rather than causal. CF-GNNExplainer Lucic et al. (2022) flipped the prediction (the definition of a 'valid' counterfactual) on only four of 200 confident positives, a 2% validity rate. These support Rudin (2019) argument that post-hoc explanations are unreliable for individual decisions. Post-hoc explainable AI (XAI) suits aggregate model audit, where averaging across cases smooths attribution noise. Per-decision operator justification (Articles 13–14) requires interpretable-by-design architectures, e.g., prototype-based GNNs. Conformal abstention (§5.1) provides per-decision routing meanwhile.

6 Implications

6.1 Individual and organisational impact

Missed cascades carry higher costs than false alarms: a false negative triggers emergency response and load shedding, while a false positive costs only review time. This influences threshold selection: even at the precision-favouring Youden's J point ($\tau = 0.684$, §5.1), the miss rate stays below 1%. The TSO control-room operator remains accountable for missed cascades as the system is only a facilitator. The classifier's training scope covers line-overload cascade events. April 2025 Iberian Peninsula cascade due to voltage instability, traced by ENTSO-E to inadequate reactive-power control and market-design barriers (Albustami and Taha, 2025; European Network of Transmission System Operators for Electricity, 2026; Morão, 2026), falls outside that scope.

6.2 Ethical and societal considerations

Under EU AI Act Regulation (EU) 2024/1689 European Parliament and Council of the European Union (2024) Annex III §2, this system is high-risk. Auditable artefacts addressing Articles 9, 10, 13, 14 and 15 are listed in §A.4.7. The biggest Article 9 finding is what I call a *robustness inversion*: at $\varepsilon = 0.05$ Gaussian feature noise (Figure 9), GINe loses only 1.3pp BalAcc while XGBoost loses 19.6pp, reversing the clean-data accuracy results Ghamizi et al. (2024). However, 20% targeted edge dropout (Figure 10) costs GINe 28.5pp of BalAcc (0.939 $\rightarrow$ 0.654, a 30% relative drop), so the model tolerates noise but not structural change.

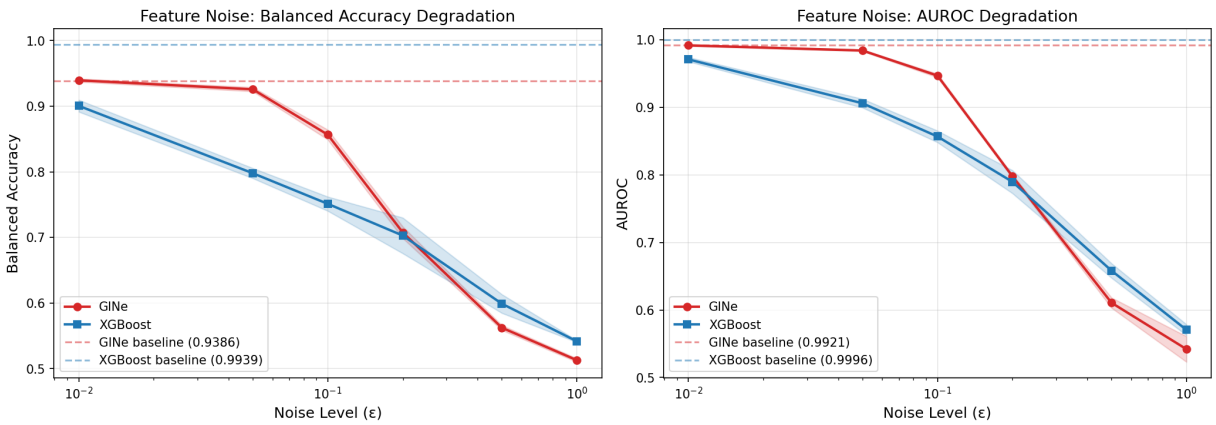

Figure 9: Feature-noise robustness across $\varepsilon \in [0.01, 1.0]$, mean over 3 seeds with ±1 std bands. The performance gap between the two models closes around $\varepsilon \approx 0.2$, suggesting a threshold for safe deployment.

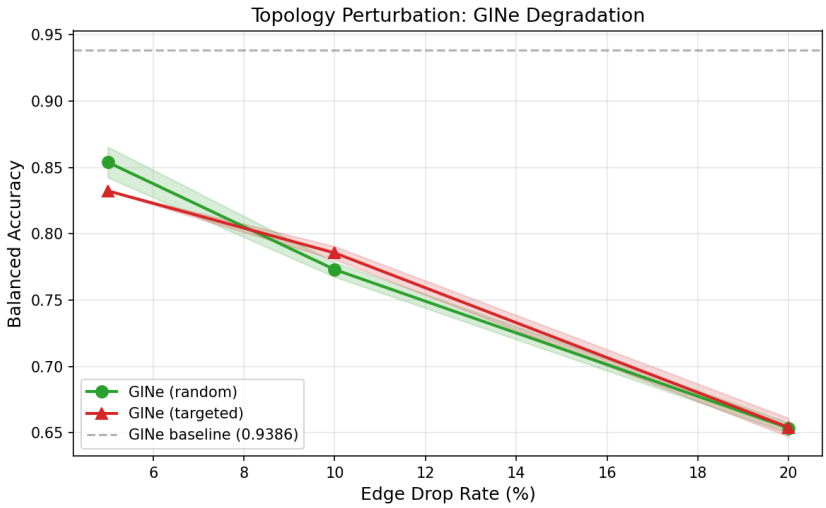

Figure 10: BalAcc under random and targeted (highest-loaded edges first, ranked by $|P_{ij}|/\text{rating}_{ij}$) deletion, mean over 3 seeds with ± 1 std bands. Targeted removal is more damaging at low drop rates (5–10%), where a few heavily-loaded edges carry disproportionate signal; by 20% the curves converge as overall connectivity loss dominates.

Article 73 governs this model's failure mode: a missed cascade would likely qualify as a "serious and irreversible disruption" of critical infrastructure, triggering a 2-day notification duty European Parliament and Council of the European Union (2024) and post-incident regulatory scrutiny. The MLflow run IDs and conformal calibration timestamps exist to be read backwards from a real failure, not forwards from a checklist. Compliance does not, however, resolve the underlying problem. Rudin (2019) argues that post-hoc methods cannot reliably explain individual decisions, and the Fidelity$^+ = -0.729$ result (§5.2) confirms it on this dataset. Producing auditable transparency artefacts is not the same as producing epistemically adequate ones, and Article 13's wording does not distinguish the two.

6.3 Environmental cost

Training emitted ∼6.3–7.6 kgCO_2e (∼50–60 kWh at GB's 2025 grid intensity of 126 gCO_2/kWh Evans and Nam (2026); derivation in §A.4.8), small relative to typical deep learning training (Patterson et al., 2021; Strubell et al., 2019). `codecarbon` Courty et al. (2024) would better estimate emissions. Inference cost dominates at deployment: a TSO screening contingencies on a per-minute cycle accumulates forward passes that exceed one-off training energy within weeks Patterson et al. (2021). The ONNX INT8 model's 99.5% BalAcc parity (§6.2) therefore makes quantised inference the main way to reduce deployment emissions, rather than training-side optimisation.

7 Emerging Trends and Future Work

The transformative trend is graph SSL as the foundation for infrastructure-scale pre-trained models. SimGRACE's Gaussian-noise encoder-weight perturbation Xia et al. (2022) matters because standard graph augmentations destroy electrical validity: dropping an edge in a power graph forces Kirchhoff's laws to redistribute current globally, producing a physically different grid rather than a perturbation of the same one. Encoder-side noise leaves the input unchanged (§A.4.9). The cross-grid transfer result is paradigmatic (Figures 2, 3): contrastive pre-training on four PowerGraph topologies matched native fine-tuning within seed-level variance and the

frozen backbone reached 0.71 BalAcc against random-init's chance-level 0.50 (§3.3). A GridFM-class model (Hamann et al., 2024; Liu et al., 2025) pre-trained on diverse topologies would invert today's per-grid cold-start methods, replacing the ∼3,600 MATPOWER simulations my learning curve required (Figure 11) with hundreds of fine-tuning labels per new grid. Concurrent physics-informed pre-training results Sevak et al. (2026) show the approach is already viable.

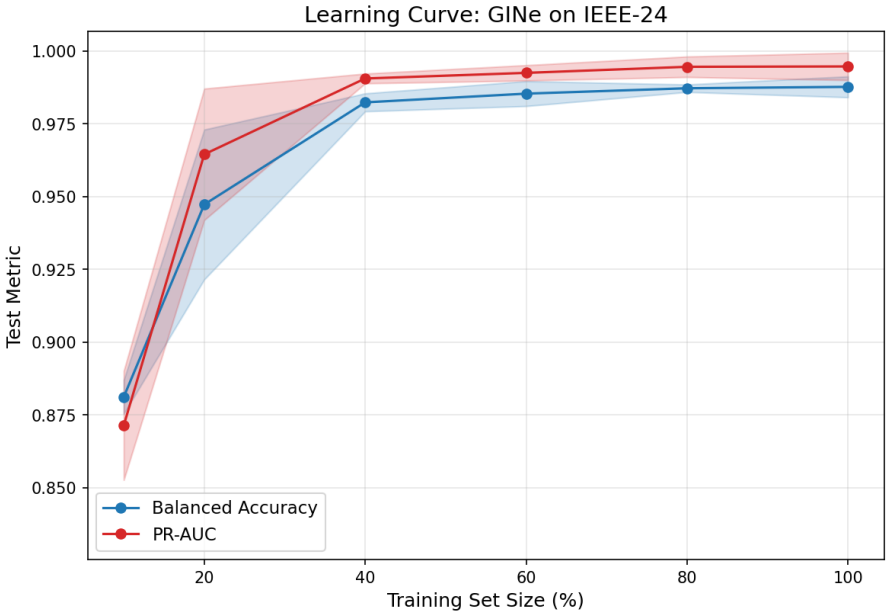

Figure 11: Learning curve: ∼3,600 MATPOWER simulations are required to saturate from scratch on IEEE-24

Recent work on agentic AI moves in the opposite direction. Autonomous LLM-orchestrated systems are prevalent in 2026, but their attractive properties are disqualifying for grid operations: open-ended autonomy collides with Article 14's human-oversight mandate (European Parliament and Council of the European Union, 2024; Gardhouse et al., 2026), and errors in an agent loop compound across steps rather than staying contained to a single inference Rabanser et al. (2026). This project's conformal abstention is the counter-mechanism: rather than constrain agency post-hoc, it provides a formal coverage guarantee that defines when the system must defer to the operator.

Distribution-free uncertainty is a deployment requirement that split conformal already meets at the calibration step (§5.1; Vovk, 2013; Angelopoulos and Bates, 2023). The next step is conformal-aware training Stutz et al. (2022): the loss directly penalises oversized prediction sets subject to a coverage constraint, collapsing train-then-calibrate into a single objective in which the model learns representations that produce tight coverage rather than only minimising cross-entropy.

SSL and conformal prediction are mutually reinforcing: one supplies transferable representations across grids, the other provides the coverage guarantee required for regulated deployment. The question for grid reliability is not what systems can do but what they should not do.

8 Conclusion

XGBoost wins on headline accuracy. The case for deep learning here is built on capabilities tree ensembles cannot reach at all: per-edge failure probabilities, transfer to unseen grid topologies after self-supervised pre-training, and prediction sets with finite-sample coverage. Whether those capabilities matter depends on deployment.

The contribution I want to emphasise is methodological. The rubric used here, i.e. inductive-bias framing, robustness inversion, and capability-set comparison offers a way to defend deep learning in safety-critical domains where aggregate accuracy is the wrong target. The specific model is replaceable.

Three limitations qualify these claims. The binding one (the deployment-breaking limit) is the gap between simulator-generated labels and PMU field measurements, which any real deployment will have to bridge. The second is that a single synthetic IEEE-24 topology drives both training and evaluation; although SimGRACE shows the gap closes once pre-training spans multiple grids, that scaling has yet to be demonstrated at the size deployment would require. The third is temporal: each graph encodes a single pre-outage state, so the dynamics of the cascade itself fall outside the model's scope until a framework such as PI-GN-JODE Sevak et al. (2026) is integrated. The April 2025 Iberian Peninsula cascade Albustami and Taha (2025) tested all three: a real grid, voltage instability rather than line overload, and dynamics that unfolded over minutes.

Infrastructure foundation models trained against conformal-aware objectives would address all three concerns. Each limit becomes an incremental extension of the same project instead of a separate project restarted from scratch.

References

Albustami, A. A. and Taha, A. F. (2025), 'The iberian blackout: A black swan or a gray rhino? a thorough power system analysis'. Preprint.
URL: *https://arxiv.org/abs/2511.17433*

Angelopoulos, A. N. and Bates, S. (2023), 'Conformal prediction: A gentle introduction', *Found. Trends Mach. Learn.* **16**(4), 494–591.
URL: *https://doi.org/10.1561/2200000101*

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V. F., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gülçehre, Ç., Song, H. F., Ballard, A. J., Gilmer, J., Dahl, G. E., Vaswani, A., Allen, K. R., Nash, C., Langston, V., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M. M., Vinyals, O., Li, Y. and Pascanu, R. (2018), 'Relational inductive biases, deep learning, and graph networks', *CoRR* **abs/1806.01261**.
URL: *http://arxiv.org/abs/1806.01261*

Bronstein, M. M., Bruna, J., Cohen, T. and Veličković, P. (2021), 'Geometric deep learning: Grids, groups, graphs, geodesics, and gauges', *CoRR* **abs/2104.13478**.
URL: *http://arxiv.org/abs/2104.13478*

Chen, T. and Guestrin, C. (2016), XGBoost: A scalable tree boosting system, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, pp. 785–794.

Courty, B., Schmidt, V., Luccioni, S., Goyal-Kamal, MarionCoutarel, Feld, B., Lecourt, J., LiamConnell, Saboni, A., Inimaz, supatomic, Léval, M., Blanche, L., Cruveiller, A., ouminasara, Zhao, F., Joshi, A., Bogroff, A., de Lavoreille, H., Laskaris, N., Abati, E., Blank, D., Wang, Z., Catovic, A., Alencon, M., Stechly, M., Bauer, C., de Araújo, L. O. N., JPW and MinervaBooks (2024), 'mlco2/codecarbon: v2.4.1'.
URL: *https://doi.org/10.5281/zenodo.11171501*

Electricity Consumers Resource Council (2004), The economic impacts of the August 2003 Blackout, Technical report, Electricity Consumers Resource Council (ELCON), Washington, DC, USA.

Errica, F., Podda, M., Bacciu, D. and Micheli, A. (2020), A fair comparison of graph neural networks for graph classification, *in* '8th International Conference on Learning Representations (ICLR)', Addis Ababa, Ethiopia.
URL: *https://openreview.net/forum?id=HygDF6NFPB*

European Network of Transmission System Operators for Electricity (2026), Grid incident in Spain and Portugal on 28 April 2025: Investigation expert panel — final report, Technical report, ENTSO-E.
URL: *https://www.entsoe.eu/publications/blackout/28-april-2025-iberian-blackout/*

European Parliament and Council of the European Union (2024), 'Regulation (EU) 2024/1689 of the european parliament and of the council of 13 June 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act)', Official Journal of the European Union, L series.
URL: *https://eur-lex.europa.eu/eli/reg/2024/1689/oj*

Evans, S. and Nam, H. W. (2026), 'Analysis: UK renewables enjoy record year in 2025 – but gas power still rises'.
URL: *https://www.carbonbrief.org/analysis-uk-renewables-enjoy-record-year-in-2025-but-gas-power-still-rises/*

Gal, Y. and Ghahramani, Z. (2016), Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, *in* 'Proceedings of the 33rd International Conference on Machine Learning (ICML)', Vol. 48 of *Proceedings of Machine Learning Research*, PMLR, pp. 1050–1059.

Gardhouse, K., Oueslati, A. and Kolt, N. (2026), 'Regulating AI agents'. Preprint.
URL: *https://arxiv.org/abs/2603.23471*

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H. and Crawford, K. (2021), 'Datasheets for datasets', *Communications of the ACM* **64**(12), 86–92.

Ghamizi, S., Bojchevski, A., Ma, A. and Cao, J. (2024), 'Safepowergraph: Safety-aware evaluation of graph neural networks for transmission power grids'.
URL: *https://arxiv.org/abs/2407.12421*

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. and Dahl, G. E. (2017), Neural message passing for quantum chemistry, *in* 'Proceedings of the 34th International Conference on Machine Learning (ICML)', Vol. 70 of *Proceedings of Machine Learning Research*, PMLR, pp. 1263–1272.

Gjorgiev, B. and Sansavini, G. (2022), 'Identifying and assessing power system vulnerabilities to transmission asset outages via cascading failure analysis', *Reliability Engineering & System Safety* **217**, 108085.

Grinsztajn, L., Oyallon, E. and Varoquaux, G. (2022), Why do tree-based models still outperform deep learning on typical tabular data?, *in* 'Advances in Neural Information Processing Systems 35 (NeurIPS), Datasets and Benchmarks Track'.

Hamann, H. F., Brunschwiler, T., Gjorgiev, B., Martins, L. S. A., Puech, A., Varbella, A., Weiss, J., Bernabe-Moreno, J., Massé, A. B., Choi, S., Foster, I., Hodge, B.-M., Jain, R., Kim, K., Mai, V., Mirallès, F., Montigny, M. D., Ramos-Leaños, O., Suprême, H., Xie, L., Youssef, E.-N. S., Zinflou, A., Belyi, A. J., Bessa, R. J., Bhattarai, B. P., Schmude, J. and Sobolevsky, S. (2024), 'A perspective on foundation models for the electric power grid'.
URL: *https://arxiv.org/abs/2407.09434*

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V. S. and Leskovec, J. (2020), Strategies for pre-training graph neural networks, *in* '8th International Conference on Learning Representations (ICLR)', Addis Ababa, Ethiopia.
URL: *https://openreview.net/forum?id=HJlWWJSFDH*

ICF Consulting (2003), The economic cost of the blackout: An issue paper on the northeastern blackout, August 14, 2003, Technical report, ICF Consulting.

Kazim, M., Pirim, H., Le, C., Le, T. and Yadav, O. (2025), 'Edge-level explainable graph neural networks with network centric features for transmission line failure prediction in power grids', *Sustainable Energy, Grids and Networks* **44**, 101969.

Kingma, D. P. and Ba, J. (2015), Adam: A method for stochastic optimization, *in* '3rd International Conference on Learning Representations (ICLR)', San Diego, CA, USA.

Kipf, T. N. and Welling, M. (2017), Semi-supervised classification with graph convolutional networks, *in* '5th International Conference on Learning Representations (ICLR)', Toulon, France.
URL: *https://openreview.net/forum?id=SJU4ayYgl*

Lin, T.-Y., Goyal, P., Girshick, R. B., He, K. and Dollár, P. (2017), Focal loss for dense object detection, *in* 'IEEE International Conference on Computer Vision (ICCV)', IEEE Computer Society, pp. 2980–2988.

Liu, J., Yang, C., Lu, Z., Chen, J., Li, Y., Zhang, M., Bai, T., Fang, Y., Sun, L., Yu, P. S. and Shi, C. (2025), 'Graph foundation models: Concepts, opportunities and challenges', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**(6), 5023–5044.
URL: *http://dx.doi.org/10.1109/TPAMI.2025.3548729*

Lucic, A., ter Hoeve, M. A., Tolomei, G., de Rijke, M. and Silvestri, F. (2022), CF-GNNExplainer: Counterfactual explanations for graph neural networks, *in* 'Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)', Vol. 151 of *Proceedings of Machine Learning Research*, PMLR, pp. 4499–4511.

Lundberg, S. M. and Lee, S.-I. (2017), A unified approach to interpreting model predictions, *in* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds, 'Advances in Neural Information Processing Systems', Vol. 30, Curran Associates, Inc.
URL: *https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf*

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D. and Gebru, T. (2019), Model cards for model reporting, *in* 'Proceedings of the Conference on Fairness, Accountability, and Transparency', ACM, pp. 220–229.
URL: *http://dx.doi.org/10.1145/3287560.3287596*

Morão, H. (2026), The 2025 Iberian Peninsula blackout: Lessons for modern power systems and policy implications, Working Papers REM 2026/0413, ISEG – Lisbon School of Economics and Management, REM, Universidade de Lisboa.
URL: *https://ideas.repec.org/p/ise/remwps/wp04132026.html*

Nakiganda, A. M. and Chatzivasileiadis, S. (2023), 'Graph neural networks for fast contingency analysis of power systems', *CoRR* **abs/2310.04213**.
URL: *https://arxiv.org/abs/2310.04213*

North American Electric Reliability Corporation (2024), 'Reliability standard TOP-001-6: Transmission operations', NERC Reliability Standard, Requirement R13: Real-time Assessments at least once every 30 minutes.
URL: *https://www.nerc.com/pa/Stand/Reliability%20Standards/TOP-001-6.pdf*

Okoyomon, E. and Goebel, C. (2026), 'Boost-rpf: Boosted sequential trees for radial power flow'. Preprint.
URL: *https://arxiv.org/abs/2603.21977*

Patterson, D. A., Gonzalez, J., Le, Q. V., Liang, C., Munguia, L.-M., Rothchild, D., So, D. R., Texier, M. and Dean, J. (2021), 'Carbon emissions and large neural network training', *CoRR* **abs/2104.10350**.
URL: *http://arxiv.org/abs/2104.10350*

Rabanser, S., Kapoor, S., Kirgis, P., Liu, K., Utpala, S. and Narayanan, A. (2026), 'Towards a science of AI agent reliability'. Preprint.
URL: *https://arxiv.org/abs/2602.16666*

Rampášek, L., Galkin, M., Dwivedi, V. P., Luu, A. T., Wolf, G. and Beaini, D. (2022), Recipe for a general, powerful, scalable graph transformer, *in* 'Advances in Neural Information Processing Systems 35 (NeurIPS)', Curran Associates, Inc.
URL: *https://arxiv.org/abs/2205.12454*

Rossi, E., Charpentier, B., Giovanni, F. D., Frasca, F., Günnemann, S. and Bronstein, M. M. (2024), Edge directionality improves learning on heterophilic graphs, *in* S. Villar and B. Chamberlain, eds, 'Proceedings of the Second Learning on Graphs Conference', Vol. 231 of *Proceedings of Machine Learning Research*, PMLR, pp. 25:1–25:27.
URL: *https://proceedings.mlr.press/v231/rossi24a.html*

Rudin, C. (2019), 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence* **1**(5), 206–215.

Saito, T. and Rehmsmeier, M. (2015), 'The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets', *PLOS ONE* **10**(3), 1–21.
URL: *https://doi.org/10.1371/journal.pone.0118432*

Sevak, B., Jadhav, S. and Bui, V.-H. (2026), 'Physics-informed graph neural jump odes for cascading failure prediction in power grids'. Preprint.
URL: *https://arxiv.org/abs/2603.20838*

Shwartz-Ziv, R. and Armon, A. (2022), 'Tabular data: Deep learning is not all you need', *Inf. Fusion* **81**(C), 84–90.
URL: *https://doi.org/10.1016/j.inffus.2021.11.011*

Strubell, E., Ganesh, A. and McCallum, A. (2019), Energy and policy considerations for deep learning in NLP, *in* A. Korhonen, D. Traum and L. Màrquez, eds, 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 3645–3650.
URL: *https://aclanthology.org/P19-1355/*

Stutz, D., Dvijotham, K. D., Cemgil, A. T. and Doucet, A. (2022), Learning optimal conformal classifiers, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=t8O-4LKFVx*

Sundararajan, M., Taly, A. and Yan, Q. (2017), Axiomatic attribution for deep networks, *in* D. Precup and Y. W. Teh, eds, 'Proceedings of the 34th International Conference on Machine Learning', Vol. 70 of *Proceedings of Machine Learning Research*, PMLR, pp. 3319–3328.
URL: *https://proceedings.mlr.press/v70/sundararajan17a.html*

Varbella, A., Amara, K., Gjorgiev, B., El-Assady, M. and Sansavini, G. (2024), Powergraph: A power grid benchmark dataset for graph neural networks, *in* 'The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track'.
URL: *https://openreview.net/forum?id=qWTfCO4HvT*

Varbella, A., Gjorgiev, B. and Sansavini, G. (2023), 'Geometric deep learning for online prediction of cascading failures in power grids', *Reliability Engineering & System Safety* **237**, 109341.
URL: *https://www.sciencedirect.com/science/article/pii/S0951832023002557*

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P. and Bengio, Y. (2018), Graph attention networks, *in* '6th International Conference on Learning Representations (ICLR)', Vancouver, BC, Canada.
URL: *https://openreview.net/forum?id=rJXMpikCZ*

Vovk, V. (2013), 'Conditional validity of inductive conformal predictors', *Machine Learning* **92**, 349–376.

Xia, J., Wu, L., Chen, J., Hu, B. and Li, S. Z. (2022), SimGRACE: A simple framework for graph contrastive learning without data augmentation, *in* 'Proceedings of the ACM Web Conference (WWW)', ACM, pp. 1070–1079.

Xu, K., Hu, W., Leskovec, J. and Jegelka, S. (2019), How powerful are graph neural networks?, *in* '7th International Conference on Learning Representations (ICLR)', New Orleans, LA, USA.
URL: *https://openreview.net/forum?id=ryGs6iA5Km*

Youden, W. J. (1950), 'Index for rating diagnostic tests', *Cancer* **3**(1), 32–35.

A Appendix

A.1 Supplementary Tables

Table 1: Graph-level results on IEEE-24 (mean ± std across 5 seeds). GINe is the production model; GCN is the edge-feature-blind ablation; XGBoost and Random Forest are classical baselines on 42 hand-engineered statistics.

Model	BalAcc	PR-AUC	AUROC	Macro-F1
GINe (production)	0.9890 ± 0.0052	0.9980 ± 0.0017	0.9995 ± 0.0004	0.9771 ± 0.0101
XGBoost	0.9947 ± 0.0014	0.9989 ± 0.0008	0.9997 ± 0.0002	0.9937 ± 0.0017
Random Forest	0.9639 ± 0.0028	0.9863 ± 0.0021	0.9959 ± 0.0006	0.9373 ± 0.0058
Plain Transformer	0.8760 ± 0.0244	0.8934 ± 0.0504	0.9624 ± 0.0173	0.7791 ± 0.0361
GCN (no edge feat.)	0.6656 ± 0.0279	0.5065 ± 0.0171	0.8084 ± 0.0125	0.5085 ± 0.0579

Table 2: Edge-level prediction with GINe (mean ± std across 5 seeds). The network-centric mode adds betweenness centrality and load ratio as edge features. Binary-F1 is bounded by extreme positive-edge sparsity (∼1–2 of 38 edges trip per cascade).

Mode	PR-AUC	AUROC	Macro-F1	Binary-F1
Standard	0.7574 ± 0.0286	0.9972 ± 0.0009	0.6923 ± 0.0231	0.3924 ± 0.0449
Network-centric	0.7629 ± 0.0372	0.9976 ± 0.0007	0.7061 ± 0.0198	0.4191 ± 0.0382

A.2 Supplementary Figures

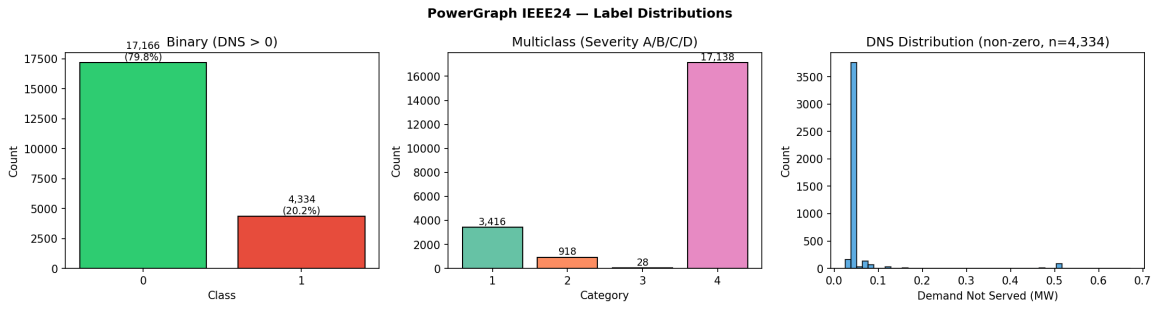

Figure 12: IEEE-24 binary positive rate ≈20.1%; multi-class A/B/C/D distribution.

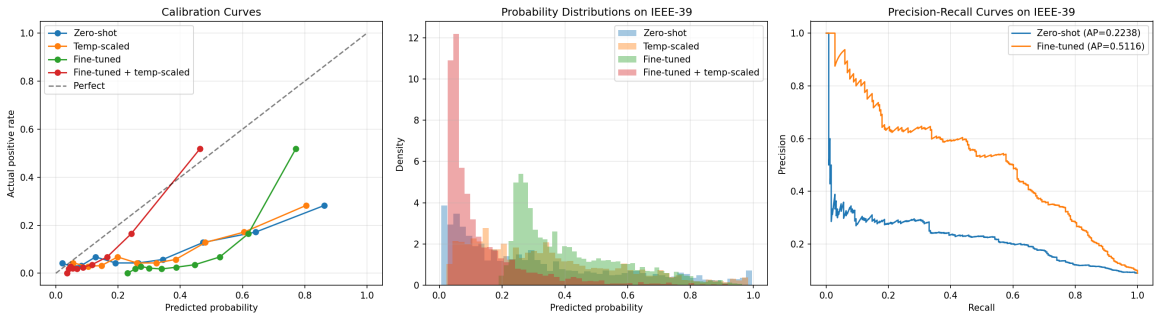

Figure 13: Cross-grid calibration: temperature scaling and Platt rescaling restore reliability under topology shift; companion to Figure 2.

A.3 Dashboard

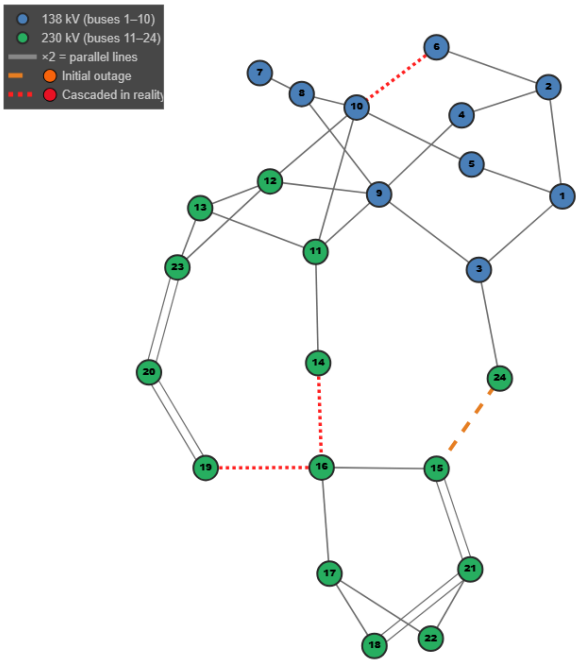

Figure 14: IEEE-24 reliability test system topology: 24 buses, 38 transmission lines. Some buses have double transmission lines.

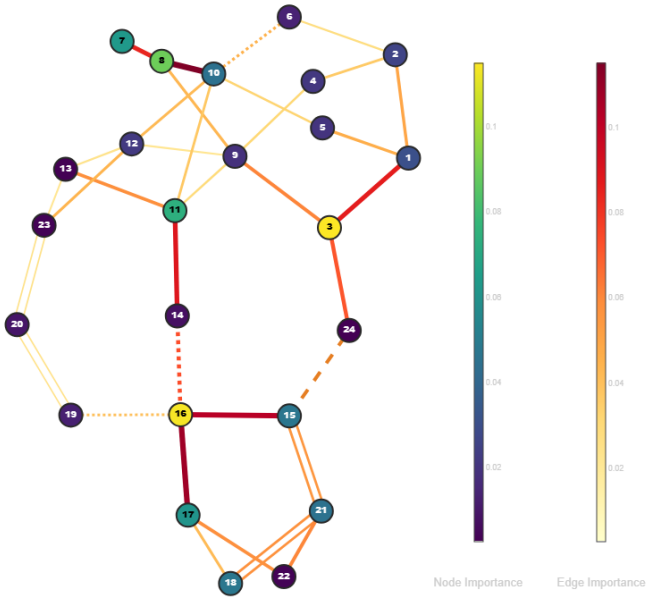

Figure 15: Topology with node and edge importance highlighted

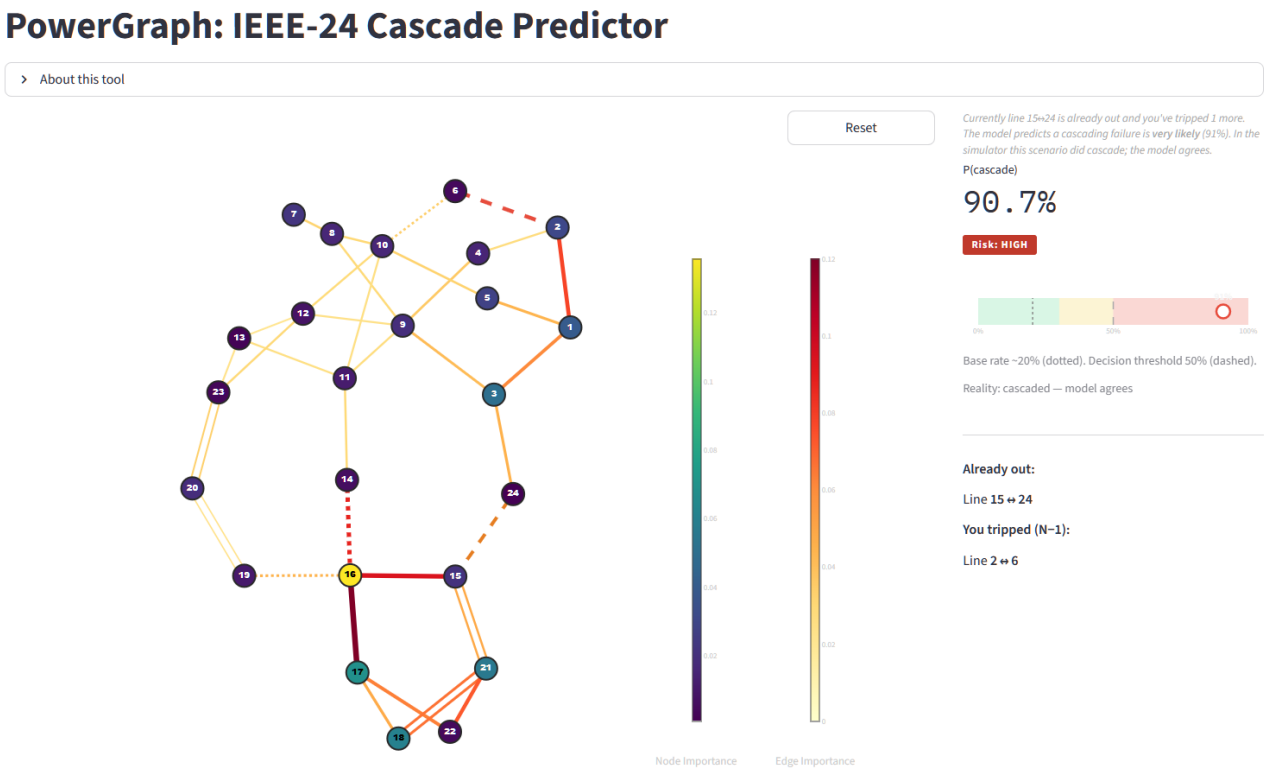

Figure 16: Streamlit Dashboard. Live demo: https://powergraph.gahmed.com.

A.4 Methodological Details

A.4.1 XGBoost Feature Vector

Table 3: 42 features used by the XGBoost baseline. Node and edge attributes are collapsed to graph-level scalars via four moments (mean, std, min, max). The line-loading ratio $\eta_e = |P_e|/r_e$, where r_e is the line thermal rating in MVA, captures thermal stress and is reported both as moments and as threshold-exceedance counts. There are four graph-structural scalars.

Group	Attributes	Moments	Count
Node	P_{net}, S_{net}, V	mean, std, min, max	12
Edge	P_{flow}, Q_{flow}, X, rating r, dir	mean, std, min, max	20
Load ratio η_e=$\lvert P_e\rvert/r_e$: mean, std, max, p_{95}, $\sum_e \mathbf{1}[\eta_e{>}0.8]$, $\sum_e \mathbf{1}[\eta_e{>}0.9]$			6
Graph structure: degree mean/std/max, density $2\lvert E\rvert/(N(N{-}1))$ (undirected)			4
Total			**42**

`scale_pos_weight` $= n_0/n_1 \approx 3.96$ (n_0=17,166, n_1=4,334; positive rate 20.16%).

A.4.2 MLP-on-Edge-Features Baseline Analysis

An MLP-on-edge-features baseline would pool the 5-dimensional per-edge feature vectors (active power flow P, reactive flow Q, reactance X, line rating r, direction d) to a graph-level vector via mean and max aggregation, then train a 2–3 layer MLP classifier. This pooling step discards the *assignment* of feature values to specific edges. Two graphs with identical edge-feature distributions but different topologies (for instance, one where the overloaded line lies on the main transfer path versus one where it lies on a peripheral branch) produce the same pooled vector and thus receive the same prediction. GINe avoids this by propagating edge features through the adjacency structure before pooling, preserving which edge is connected to which. Because Kirchhoff's laws make that relational information the primary signal (power reroutes along specific paths, not uniformly), the MLP's pooling step discards exactly what matters. The GCN ablation (BalAcc 0.6656 ± 0.0279 against GINe's 0.9890 ± 0.0052, Table 1) provides indirect evidence: the GCN reads topology but ignores edge features; an MLP-on-edge-features baseline reads edge features but ignores topology. Both ablations are predicted to underperform GINe, and the GCN result confirms the direction of effect.

A.4.3 Directed GINe: Mixing Parameter

The directed aggregation layer Rossi et al. (2024) runs two independent GINe message-passing channels, one over forward edges, one over backward edges, and combines their node embeddings as

$$\mathbf{h}_v = \alpha\, \mathbf{h}_v^{\text{fwd}} + (1-\alpha)\, \mathbf{h}_v^{\text{bwd}},$$

where $\alpha \in [0,1]$ is a single learned scalar shared across all nodes and layers. $\alpha = 1$ recovers a purely forward model; $\alpha = 0$ a purely backward one; $\alpha = 0.5$ weights both equally. Across five seeds the learned value consistently converged to $\alpha \approx 0.5$, meaning the model settled on balanced bidirectionality rather than a directional preference. Physically, this makes sense: a line trip increases loading on downstream lines (forward propagation of rerouted flow) and reduces generation margin on upstream buses (backward propagation), so both directions carry predictive signal.

A.4.4 GINe Message-Passing Equations

The GINe update rule at layer k is

$$\mathbf{h}_v^{(k)} = \mathrm{MLP}_k\Big((1+\varepsilon)\,\mathbf{h}_v^{(k-1)} + \sum_{u\in\mathcal{N}(v)} \mathrm{ReLU}\big(\mathbf{h}_u^{(k-1)} + \mathbf{e}_{uv}\big)\Big), \tag{1}$$

where $\mathbf{e}_{uv} \in \mathbb{R}^5$ is the edge-feature vector (active power flow P, reactive flow Q, reactance X, line rating r, direction d), ε is a learnable scalar, the inner $\mathrm{ReLU}(\cdot)$ matches the canonical PyTorch Geometric `GINEConv` formulation, and each MLP_k is a two-layer `Linear` $\to$ `ReLU` $\to$ `Linear` network. Each layer applies BatchNorm1d, ReLU, and Dropout(p=0.126) after the MLP, in that order. After K=4 layers the graph embedding is formed by mean-max pooling:

$$\mathbf{h}_G = \mathrm{MEAN}\big\{\mathbf{h}_v^{(K)}\big\} \,\|\, \mathrm{MAX}\big\{\mathbf{h}_v^{(K)}\big\} \;\in \mathbb{R}^{128}, \tag{2}$$

yielding a 128-dimensional vector ($2 \times \mathit{hidden\_dim}$). The GCN ablation omits $\mathbf{e}_{uv}$ entirely, using PyTorch Geometric's `GCNConv` which implements the renormalised propagation rule of Kipf and Welling (2017): $\mathbf{h}_v^{(k)} = \sigma\Big(\mathbf{W}^{(k)} \sum_{u\in\mathcal{N}(v)\cup\{v\}} \mathbf{h}_u^{(k-1)} / \sqrt{\hat{d}_u\,\hat{d}_v}\Big)$, where $\hat{d}_v = 1 + |\mathcal{N}(v)|$ accounts for the inserted self-loop and $\mathbf{W}^{(k)}$ is the per-layer learnable transform; that single omission of edge features drives BalAcc from 0.989 to 0.666.

A.4.5 Focal Loss Parameterisation

The focal loss Lin et al. (2017) for a binary prediction with per-sample probability p_t is

$$\mathrm{FL}(p_t) = -\alpha_t\,(1-p_t)^{\gamma}\,\log p_t. \tag{3}$$

Class weights are set to inverse class frequency: $\alpha_c = n_{\text{total}}/(2\,n_c)$. To prevent the >100× weight ratios that caused gradient instability with plain weighted cross-entropy, the ratio $\max(\boldsymbol{\alpha})/\min(\boldsymbol{\alpha})$ is hard-capped at `alpha_cap`. Optuna found $\gamma = 2.02$ (search range [0.5, 3.0]) and `alpha_cap` = 7.56 (range [5, 20]). Implementation: $\mathbf{p}_t = \exp\big(-\mathrm{CE}(\mathbf{logits}, \mathbf{y}, \text{weight}{=}\boldsymbol{\alpha})\big)$, then $\mathcal{L} = \mathrm{mean}\big((1-\mathbf{p}_t)^{\gamma}\cdot \mathrm{CE}\big)$.

A.4.6 Uncertainty Quantification: MC Dropout and Conformal Prediction

MC Dropout Gal and Ghahramani (2016) runs stochastic forward passes with dropout active at inference. High predictive entropy across passes flags low-confidence cases for abstention: the model declines to issue a prediction on uncertain graphs and routes them to the operator, trading coverage (the fraction of cases automatically decided) for accuracy on the remainder. The selective-prediction curve in Figure 5 sweeps this trade-off; the operator-facing parameter is the entropy threshold, set to satisfy an acceptable miss-rate.

Split conformal prediction (Vovk, 2013; Angelopoulos and Bates, 2023) returns a set of plausible labels rather than a single prediction. The set size carries operational meaning: a singleton commits to one label, a pair defers when both labels remain plausible, and an empty set escalates to the operator because neither label clears the calibrated threshold. The method guarantees the true label is contained in the predicted set with probability at least $1-\alpha$, and this guarantee holds without distributional assumptions, requiring only that calibration and test data be exchangeable. Abstention rate is therefore set by a single tunable α rather than an arbitrary confidence cutoff.

Both methods feed the same operator workflow but answer different questions: MC Dropout asks "how uncertain is the model on this case", conformal asks "which labels can I rule out at this confidence level". Both address the key requirements of Article 14(4)(d) by exposing a regulator-tunable parameter that governs when the system defers to the operator.

A.4.7 Compliance Artefacts

The high-risk classification under EU AI Act Annex III §2 European Parliament and Council of the European Union (2024) obliges the producer to maintain a per-system register of auditable artefacts addressing Articles 9 (risk management), 10 (data and data governance), 13 (transparency), 14 (human oversight) and 15 (accuracy, robustness, cybersecurity). The artefacts produced by this project are listed in Table 4; each maps to one or more articles and points to the corresponding `scripts/` or MLflow output that generates it. The list is not a substitute for the substantive critique in §6.2: every artefact below is necessary for compliance and none is sufficient for epistemic adequacy.

Table 4: Compliance artefacts produced by this project, mapped to EU AI Act articles.

Artefact	Article(s)
Pandera schema validation	10
Dataset card Gebru et al. (2021)	10, 13
Model card Mitchell et al. (2019)	13
MC-Dropout selective prediction	14
Conformal escalation Angelopoulos and Bates (2023)	14, 15
FP32/INT8 parity check	15
Seed-variance intervals (5 seeds)	15
Feature-noise robustness curve	15

A.4.8 Energy Estimation

Training was estimated at ∼120–150 GPU-hours on an RTX PRO 5000 Blackwell (300 W TDP × 70% utilisation + 150 W system overhead ≈ 360 W average draw), giving ∼43–54 kWh. `codecarbon` Courty et al. (2024) was not run, so this is a back-of-envelope estimate rather than a measured trace; a measured run would supersede it. Conversion to CO_2e uses Great Britain's 2025 generation-based grid intensity of 126 gCO_2/kWh Evans and Nam (2026).

A.4.9 SimGRACE Pre-training Details

SimGRACE Xia et al. (2022) creates two views of the same graph by perturbing encoder weights rather than the input graph. For each parameter tensor $\boldsymbol{\theta}$, a perturbed copy is sampled as

$$\boldsymbol{\theta}' = \boldsymbol{\theta} + \boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}\big(\mathbf{0},\ \sigma_p^2 \mathbf{I}\big), \quad \sigma_p = \sigma_{\text{scale}} \cdot \text{std}(\boldsymbol{\theta}), \tag{4}$$

with $\sigma_{\text{scale}} = 0.1$ (a multiplicative coefficient on the per-tensor weight standard deviation, equivalent to η in the SimGRACE paper's original notation but renamed here to free η for the line-loading role used elsewhere in this appendix). The two views $\mathbf{z} = f_{\boldsymbol{\theta}}(G)$ and $\mathbf{z}' = f_{\boldsymbol{\theta}'}(G)$ are then aligned via the NT-Xent loss. Let $\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_{2B} = (\mathbf{z}_1, \ldots, \mathbf{z}_B, \mathbf{z}'_1, \ldots, \mathbf{z}'_B)$ be the concatenated batch of both views and let $p(i)$ index the positive partner of view i (so $p(i){=}i{+}B$ for $i \le B$ and $p(i){=}i{-}B$ otherwise). The loss is

$$\mathcal{L}_{\text{NT-Xent}} = -\frac{1}{2B} \sum_{i=1}^{2B} \log \frac{\exp\big(\text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_{p(i)})/\tau\big)}{\sum_{k=1}^{2B} \mathbf{1}_{[k \neq i]}\ \exp\big(\text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_k)/\tau\big)}, \tag{5}$$

where $\tau = 0.2$, $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v} / (\|\mathbf{u}\|\,\|\mathbf{v}\|)$ is cosine similarity, embeddings are ℓ_2-normalised before computing it, and the indicator $\mathbf{1}_{[k \neq i]}$ excludes only the anchor itself (so the positive partner remains in the denominator, giving $2B - 1$ contrastive terms per anchor). This preserves

electrical validity: node masking or edge dropping changes the input topology and violates Kirchhoff constraints globally (power redistributes across every remaining line); encoder perturbation leaves the input graph unchanged.

Pre-training used all four PowerGraph topologies combined (IEEE-24, IEEE-39, UK, IEEE-118; $\approx$236 k graphs), 100 epochs, batch size 512, learning rate 2×10^{-3} with 10-epoch linear warmup followed by cosine decay, and a two-layer projection head of output dimension 128.