

A Comprehensive Review of Robust AI Techniques for Securing Deep Learning Models in Smart Grid Systems

Yogeshwari Yawalkar, Sakshi Vinod Borkar, Sohail Sharif Shaikh, Vijay Babasaheb Amrute

D.Y. Patil Institute of Master of Computer Applications and Management, Akurdi, Pune.

Students, Department of Master of Computer Applications.

ABSTRACT

Deep learning models now support load forecasting, fault detection, state estimation and renewable-energy management in smart grids. Small, deliberately crafted changes to sensor data, known as adversarial attacks, can push these models to wrong answers and put grid reliability at risk. We reviewed twenty papers on adversarial attacks and defences in smart-grid and related energy systems, seventeen of them published from 2021 onwards, along with foundational work on FGSM, PGD and ensemble adversarial training. Attack success rates above 90% were reported against undefended detectors and adversarial training improved robustness considerably. Yet a defence trained against one attack often fails against another, most studies defend a single model and ensemble defences are rarely tested for sustainable grids. From these findings we draw four research gaps and outline a CNN–LSTM ensemble framework with adversarial training.

Keywords: Adversarial Attacks, Deep Learning, Smart Grid, Ensemble Learning, Adversarial Training, Cybersecurity

1. Introduction

A smart grid layers sensing, communication and analytics over a conventional power network so that supply and demand can be matched in real time, even with variable sources such as solar and wind connected. The voltage, current and load data it produces are far too large to inspect by hand, so utilities lean on deep learning models such as CNN, LSTM and DNN for load forecasting, fault detection, state estimation and electricity-theft detection [4], [8], [20].

Such a model can only be as trustworthy as its input. An adversarial attack makes a small, carefully designed change to an input and the model returns a wrong output even though the data looks normal [3], [5]. The Fast Gradient Sign Method (FGSM) needs just one gradient step [5]. Projected Gradient Descent (PGD) repeats that step several times and is therefore a tougher test of robustness [10]. The same idea carries over to voltage, current or load readings, for instance through false data injection [16], [19]. The effects are easy to picture (Fig. 1): a false trip, a fault that goes unnoticed or a forecast that misjudges renewable output and wastes clean energy. Protecting these models therefore supports SDG 7 (Affordable and Clean Energy), SDG 9 (Industry, Innovation and Infrastructure) and SDG 13 (Climate Action).

Several defences have been proposed, among them adversarial training [5], [10], ensemble training [18] and denoising [11]. Even so, the systematic review by Askhatuly et al. [3] finds that no single defence is robust against everything. This paper compares the existing studies, notes where they agree and where they differ and points out the gaps that make a combined CNN–LSTM ensemble with adversarial training worth studying for sustainable smart grids.

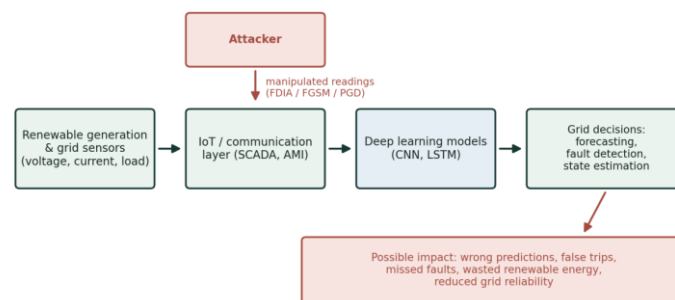


Fig. 1 Attack surface of an AI-driven smart grid

Source: Prepared by the authors; attack scenarios based on El-Toukhy et al. [4], Tusher et al. [19] and Ilahi et al. [8].

2. Problem Statement and Objectives

2.1 Problem statement

Deep learning models are used widely for forecasting, fault detection and state estimation in smart grids, yet small deliberate changes to their inputs can mislead them. Existing defences are usually tested against one attack and applied to one model. A combined CNN–LSTM ensemble with adversarial training for sustainable smart grids has not been studied properly.

2.2 Objectives

- To review twenty research papers on adversarial attacks and defences in smart-grid and related energy systems.
- To compare their methods, main findings and limitations.
- To identify the gaps in combined, multi-attack defences for sustainable smart grids.
- To propose a conceptual CNN–LSTM ensemble framework with adversarial training for future work.

3. Review Methodology

We selected twenty papers that deal with adversarial robustness of deep learning in smart-grid or closely related energy systems. Seventeen were published in 2021 or later and three foundational works on FGSM [5], PGD [10] and ensemble adversarial training [18] are included because the later studies build on them. Every paper was compared on its topic, method, main finding and limitation. We then grouped the recurring results and weaknesses (Sections 4 and 5) and derived the research gaps in Section 8 from that comparison.

4. Literature Review

The papers are discussed in four groups: foundational work, surveys, detection of false data, theft and intrusion and attacks on forecasting, estimation and protection tasks. Fig. 2 shows the families of defence from the latest systematic review and Table 1 lists the core attacks and defences.

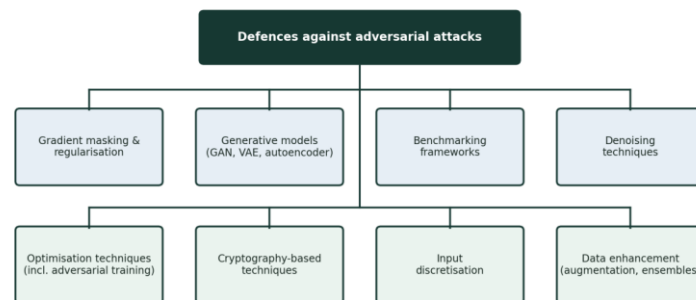


Fig. 2 Families of defence against adversarial attacks

Source: Adapted from the taxonomy of Askhatuly et al. [3], IEEE Access, 2025.

4.1 Foundations of adversarial robustness

Goodfellow et al. [5] showed that tiny perturbations along the sign of the loss gradient (FGSM) make a model misclassify with high confidence and that training on such examples makes it more robust. FGSM is only a single step, though, so it is weaker than iterative attacks. Madry et al. [10] treated robustness as a min-max optimisation problem. Their PGD-based training resisted a wide range of attacks on image benchmarks, but it was expensive to train. Tramèr et al. [18] noticed that ordinary adversarial training copes badly with black-box attacks transferred from other models and proposed ensemble adversarial training, which uses perturbations crafted on several models; they tested it on images only. Askhatuly et al. [3] reviewed 132 studies and found that adversarial training and certified robustness dominate, both with scalability limits, while hybrid and ensemble defences are under-evaluated. For time series, Karim et al. [9] used adversarial transformation networks with a surrogate model and all 42 UCR datasets they tested were susceptible (as summarised in [3]); their data, however, is general time series rather than grid data.

4.2 Surveys on smart-grid and related systems

Hao and Tao [7] surveyed evasion and poisoning attacks on deep learning in smart grids and concluded that the threat is long-term and needs ongoing attention; the survey offers no new defence and no experiments. Sánchez et al. [14] sorted attacks on learning-based grid models into categories, observed that experimental code is rarely shared and suggested adding explainable AI to adversarial models, backed by an initial experiment on substations. Ilahi et al. [8] covered state, action, reward and model attacks on deep reinforcement learning. Most defences, they found, work only against the attacks they were built for and ensemble defences for reinforcement learning remain an open problem.

4.3 Detection of false data, theft and intrusion

Abdulaal et al. [1] combined deep and ensemble learning to spot false readings from smart meters in real time, though only for advanced metering infrastructure (AMI). El-Toukhy et al. [4] attacked electricity-theft detectors built on a deep Q-network with FFNN, CNN and GRU models. FGSM and reinforcement-learning attacks succeeded 92.92–99.96% of the time and adversarial training brought that down to 1.80–9.20%, but a defence trained on one attack did not protect against the other. The study covers theft only, uses a single dataset of 130 customers and relies on reinforcement-learning detectors. Takiddin et al. [15] built a robust data-driven detector against evasion attacks that make theft readings look normal, again on metering data. Acharya et al. [2] proposed ADVIS-G, an intrusion detection system for DNP3 sessions converted into images. Under the momentum iterative method (MIM) the baseline F1-score fell to 0.5; adversarial training lifted it to 0.7, autoencoder blocking to about 0.9 and using both was most resilient. Only one DNP3-related dataset was used. Hamedani et al. [6] combined game theory with an MLP defender and found that the attacker's gain dropped sharply compared with a traditional state estimator, but the result depends on the scenario and on one MLP. Sampedro et al. [13] built a new attack dataset with adversarial perturbation, backdoor and denial-of-service classes. Accuracy ranged from 29.10% to 73.9%, with a fine-tuned DNN doing best and a single DNN was used rather than an ensemble.

4.4 Forecasting, state estimation, protection and fault prediction

Xu and Teng [20] designed an optimised availability attack on piecewise-linear load-forecasting networks. Adversarial training cut the error on attacked samples by more than 50% while raising clean-data error by only about 1% (as summarised in [3]); only forecasting models were examined. Mahmoudnezhad et al. [11] paired a denoising autoencoder with a GAN (SMDAEGAN) and improved forecasting resilience under false data injection for short and long horizons, without an ensemble or adversarial training. Tian et al. [16] showed that neural-network state estimators can be manipulated and that adversarial training and protection-based defences improve robustness, although the defences were studied separately. Tian et al. [17] found CNN-based power-quality recognition vulnerable to adversarial inputs and proposed a defence, but the work is specific to that application and one CNN family. Saber et al. [12] ran FGSM against four detectors in differential relays. It fooled the MLP in 97.24% of cases, the LSTM in 49.77%, the CNN in 35.26% and the ResNet in 9.78% and adversarial training improved robustness without hurting fault detection; the study used FGSM only and one simulated microgrid. Tusher et al. [19] studied wind-turbine fault prediction under cyber-attack with RF, XGBoost, LSTM and an autoencoder-LSTM. Attacks cut accuracy by up to 15% and recall by 28%, adversarial training gave the best robustness and XGBoost was the most resilient model, though each model was defended on its own.

Table 1. Core attacks and defences recurring in the reviewed literature

Method	Type	Core idea	Reference
FGSM	Single-step, gradient-based attack	Moves the input one step in the direction of the sign of the loss gradient; fast and also used to create training samples.	Goodfellow et al. [5]
PGD	Iterative, gradient-based attack	Repeats small gradient steps and projects the result back into an allowed perturbation limit; a stronger test than FGSM.	Madry et al. [10]
Ensemble adversarial training	Defence	Trains with perturbations crafted on other models so the model resists transferred attacks.	Tramèr et al. [18]
False data injection (FDIA)	Data manipulation	Attacker alters measurements so that detection, estimation or forecasting models are misled.	Studied in [4], [11], [16], [19]

Source: Compiled by the authors from Goodfellow et al. [5], Madry et al. [10], Tramèr et al. [18] and the smart-grid studies [4], [11], [16], [19].

5. Comparison of the Reviewed Studies

One result shows up in nearly every study: models that do well on clean data fall apart under attack. El-Toukhy et al. [4] report attack success rates of 92.92–99.96% against theft detectors, Tusher et al. [19] saw accuracy drop by up to 15% and recall by up to 28% in wind-turbine fault prediction and Saber et al. [12] fooled an MLP-based relay detector with 97.24% of FGSM samples. Adversarial training is the usual remedy and it worked in [4], [12], [16], [19] and [20]. Acharya et al. [2] got the highest resilience under stronger attacks by adding an autoencoder-based blocking model on top of it. Askhatuly et al. [3] likewise name adversarial training as one of the two dominant defence families, with a heavy computational cost.

How far a defence generalises varies. The cross-attack experiments in [4] (Fig. 3) are the clearest evidence: a detector hardened against one attack still showed success rates of 62.0–97.8% against the other and only training on both attacks pushed them down to 0.6–10.7%. Tramèr et al. [18] saw something similar in image models and the broader reviews [3], [8] conclude that no defence is universally robust. Most smart-grid studies also depend on one model family, such as an MLP [6], a CNN [17] or a DDQN-based detector [4]. Tusher et al. [19] compare several models but defend each one separately.

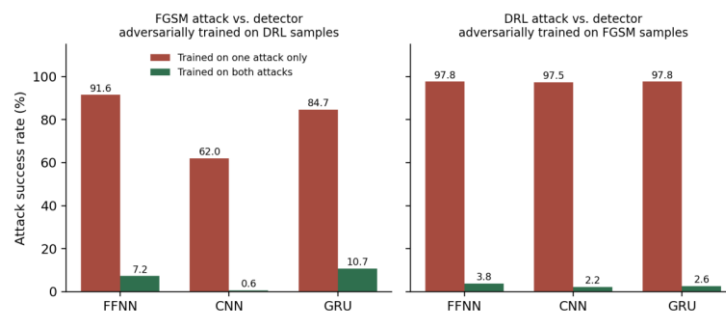


Fig. 3 Attack success rate when a defence is trained on one attack versus both attacks

Source: Data redrawn from Tables 7 and 8 of El-Toukhy et al. [4], IEEE Access, vol. 11, 2023.

Architecture matters too. In the relay study of Saber et al. [12] (Fig. 4), the share of successful adversarial samples differed widely between models. Since different architectures fail on different samples, an ensemble is a sensible way to reduce dependence on any one model, which is the thinking behind ensemble adversarial training [18]. Only that idea and the hybrid direction mentioned in [3] point toward combined defences and neither has been applied to CNN–LSTM models in sustainable smart grids.

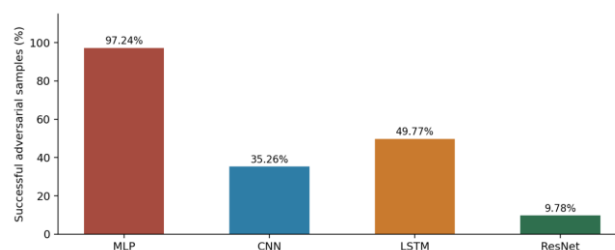


Fig. 4 Share of FGSM adversarial samples that succeeded against four model types

Source: Data redrawn from Table II of Saber et al. [12], arXiv:2506.19302, 2025.

6. Applications Where Robustness Matters

In each application covered by the reviewed studies, a wrong prediction has a physical or financial cost:

- **Load forecasting** [11], [20]: wrong forecasts lead to wrong dispatch decisions.
- **Wind-turbine and renewable fault prediction** [19]: manipulated SCADA/IoT data can hide faults or trigger needless shutdowns.
- **State estimation** [16]: a core control-centre function built on grid measurements.
- **Electricity-theft detection and metering security** [4], [15]: attackers try to make false readings look normal.
- **Relay protection** [12]: telling real faults from false data injection so a healthy line is not tripped.
- **Power-quality recognition** [17]: classifying voltage and current disturbances with CNN models.
- **Attack classification and intrusion detection** [1], [2], [6]: separating normal from malicious activity in grid communication and measurements.

7. Benefits and Challenges

7.1 Benefits of robust AI in smart grids

- Predictions stay reliable when sensor data is manipulated, so there are fewer false trips and missed faults.
- Dependable forecasting and fault prediction help integrate solar and wind power and avoid wasted renewable energy.
- Using several models removes a single point of failure and models that have seen attack examples are better prepared for new ones.
- Greater trust in AI-based decisions supports resilient infrastructure and SDG 7, SDG 9 and SDG 13.

7.2 Challenges

- **Computational cost:** adversarial training, especially with iterative attacks such as PGD, takes more time and resources [3], [10].
- **Accuracy–robustness trade-off:** some defences reduce accuracy on clean data, so both have to be measured [3].
- **Attack diversity:** a defence trained on one attack may fail against another [4], [18].
- **Scarce attack data:** public grid datasets rarely hold labelled attacks; El-Toukhy et al. [4] generated malicious samples artificially and Sampedro et al. [13] built a new attack dataset.
- **Ensemble complexity:** combining models adds complexity and can affect real-time performance [8].
- **Time-series data:** image-style attacks and defences may not transfer directly to voltage, current and load readings [3].

8. Research Gap

Deep learning, adversarial training and ensemble learning have each proved useful, but their combined use for sustainable smart grids has not been studied adequately. We identify four gaps.

- **Gap 1 – Combined defence:** adversarial training [4], [16], [19], [20] and ensemble ideas [18] are usually studied separately; a unified CNN–LSTM ensemble with adversarial training for smart grids is missing.
- **Gap 2 – Multi-attack generalisation:** defences are mostly tested against one attack type, although [4] shows this does not guarantee protection against another.
- **Gap 3 – Single-model dependence:** many studies defend one model, such as an MLP [6], a CNN [17], a DNN [13] or a DDQN detector [4], which leaves a single point of failure, while [12] shows that CNN and LSTM models fail differently.
- **Gap 4 – Sustainable-energy focus:** most robustness research targets images or generic ML [3]; fewer studies link it to renewable-energy-enabled grids and time-series grid data.

9. Proposed Methodology

To address these gaps, we propose the framework in Fig. 5. Baseline CNN and LSTM models are first evaluated on clean smart-grid data, then attacked with FGSM and PGD, retrained with adversarial examples, combined into an ensemble and finally compared on both clean and attacked data. This is a proposed direction and no experimental results are claimed in this paper.

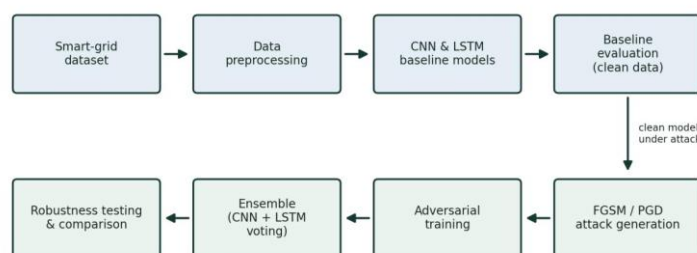


Fig. 5 Proposed conceptual framework for the robust CNN–LSTM ensemble

Source: Prepared by the authors; attack methods from Goodfellow et al. [5] and Madry et al. [10], ensemble idea informed by Tramèr et al. [18].

9.1 Planned steps

- **Data:** time-series smart-grid readings (voltage, current and load from IoT/SCADA/AMI sensors) from a public dataset still to be selected. Public grid datasets rarely contain labelled attacks [4], [13], so attacked samples will be generated artificially.
- **Preprocessing:** cleaning, normalisation, windowing of the time series and a train/validation/test split.

- **Baseline models:** CNN and LSTM models trained on clean data.
- **Attack generation:** FGSM [5] and PGD [10] applied to the test inputs.
- **Defence:** adversarial training of each model with attacked samples, then a CNN + LSTM ensemble with voting, informed by ensemble adversarial training [18].
- **Tools:** a Python-based deep learning implementation.

9.2 Evaluation metrics

- Accuracy, precision, recall and F1-score on clean and attacked data.
- Attack success rate, to compare the baseline, adversarially trained and ensemble models.
- Confusion matrix and ROC curve for a visual comparison of the models.

10. Future Scope

- Implement and test the CNN–LSTM ensemble with adversarial training on a public smart-grid dataset.
- Go beyond FGSM and PGD to other data-manipulation attacks relevant to grids, such as scaling or pulse attacks [19].
- Compare the ensemble with single-model CNN and LSTM baselines on clean and attacked data to measure the accuracy–robustness trade-off and the extra computational cost [3].
- Test against transferred black-box attacks [18] and against training on several attack types together, since protection against one attack does not guarantee protection against another [4].
- Add autoencoder-based blocking or denoising layers [2], [11] on top of ensemble adversarial training for a layered defence.
- Combine robust models with federated learning and explainable AI for privacy and trust.
- Study lightweight defences suited to real-time and edge deployment and develop standard benchmarks for adversarial robustness in smart-grid time-series data [3].

11. Conclusion

This review looked at twenty papers on adversarial attacks and defences relevant to deep learning in smart grids. Deep learning is widely used for forecasting, fault detection, state estimation and theft detection and it helps integrate renewable energy. The studies show that these models can be fooled by small manipulations of their input, with attack success rates above 90% against undefended detectors [4]. Adversarial training is the most common defence and improves robustness considerably, but it is costly and a defence trained against one attack may not hold against another [4]. Ensemble and hybrid defences look promising but are under-evaluated [3] and most smart-grid work relies on a single model. The gaps identified here support developing a combined CNN–LSTM ensemble with adversarial training, tested against several attacks while keeping clean-data accuracy high. Such a framework could contribute to more secure and reliable AI in sustainable smart-grid systems.

References

1. Abdulaal, M. J., Ibrahim, M. I., Mahmoud, M. M. E. A., Khalid, J., Aljohani, A. J., Milyani, A. H., & Abusorrah, A. M. (2022). Real-time detection of false readings in smart grid AMI using deep and ensemble learning. *IEEE Access*, 10, 47541–47556.
2. Acharya, R., Al Sardy, L., Muhammad, M., & German, R. (2026). ADVIS-G: An adversarially defended intrusion detection system for smart grids using deep learning. *Künstliche Intelligenz*. <https://doi.org/10.1007/s13218-026-00905-3>
3. Askhatuly, A., Berdysheva, D., Berdyshev, A., Adamova, A., & Yedilkhan, D. (2025). Adversarial attacks and defense mechanisms in machine learning: A structured review of methods, domains and open challenges. *IEEE Access*, 13, 185145–185168. <https://doi.org/10.1109/ACCESS.2025.3624409>
4. El-Toukhy, A. T., Mahmoud, M. M. E. A., Bondok, A. H., Fouda, M. M., & Alsabaan, M. (2023). Countering evasion attacks for smart grid reinforcement learning-based detectors. *IEEE Access*, 11, 97373–97390. <https://doi.org/10.1109/ACCESS.2023.3312376>

5. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv. <https://arxiv.org/abs/1412.6572>
6. Hamedani, K., Liu, L., Jagannath, J., & Yi, Y. (2021). Adversarial classification of the attacks on smart grids using game theory and deep learning. arXiv. <https://arxiv.org/abs/2106.03209>
7. Hao, J., & Tao, Y. (2022). Adversarial attacks on deep learning models in smart grids. *Energy Reports*, 8, 123–129. <https://doi.org/10.1016/j.egy.2021.11.026>
8. Ilahi, I., Usama, M., Qadir, J., Janjua, M. U., Al-Fuqaha, A., Hoang, D. T., & Niyato, D. (2022). Challenges and countermeasures for adversarial attacks on deep reinforcement learning. *IEEE Transactions on Artificial Intelligence*, 3(2), 90–109. <https://doi.org/10.1109/TAI.2021.3111139>
9. Karim, F., Majumdar, S., & Darabi, H. (2021). Adversarial attacks on time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10), 3309–3320.
10. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv. <https://arxiv.org/abs/1706.06083>
11. Mahmoudnezhad, et al. (2024). Electric load forecasting under false data injection attacks via denoising deep learning and generative adversarial networks. *IET Generation, Transmission & Distribution*. <https://doi.org/10.1049/gtd2.13273>
12. Saber, A. M., Maheshwari, A., Youssef, A., & Kundur, D. (2025). Adversarial attacks on deep learning-based false data injection detection in differential relays. arXiv. <https://arxiv.org/abs/2506.19302>
13. Sampedro, G. A., Ojo, S., Krichen, M., Alamro, M. A., Mihoub, A., & Karovic, V. (2024). Defending AI models against adversarial attacks in smart grids using deep learning. *IEEE Access*, 12, 157408–157417. <https://doi.org/10.1109/ACCESS.2024.3473531>
14. Sánchez, G., Elbez, G., & Hagenmeyer, V. (2024). Attacking learning-based models in smart grids: Current challenges and new frontiers. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems* (pp. 589–595). <https://doi.org/10.1145/3632775.3661984>
15. Takiddin, A., Ismail, M., & Serpedin, E. (2023). Robust data-driven detection of electricity theft adversarial evasion attacks in smart grids. *IEEE Transactions on Smart Grid*, 14(1), 663–676.
16. Tian, J., et al. (2022). Adversarial attack and defense methods for neural network based state estimation in smart grid. *IET Renewable Power Generation*, 16(16), 3507–3518. <https://doi.org/10.1049/rpg2.12334>
17. Tian, J., Wang, B., Li, J., & Wang, Z. (2022). Adversarial attacks and defense for CNN based power quality recognition in smart grid. *IEEE Transactions on Network Science and Engineering*, 9(2), 807–819.
18. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2017). Ensemble adversarial training: Attacks and defenses. arXiv. <https://arxiv.org/abs/1705.07204>
19. Tusher, A. S., Rahman, M. A., Islam, M. R., Hossain, M. A., Anwar, A., & Hossain, M. J. (2025). Cyber-attack resilient data-driven approaches for early fault prediction system for wind turbines. *Sustainable Energy Technologies and Assessments*, 84.
20. Xu, W., & Teng, F. (2023). Availability adversarial attack and countermeasures for deep learning-based load forecasting. arXiv. <https://arxiv.org/abs/2301.01832>