**Introduction**

Modern infrastructure is increasingly using Battery Energy Storage Systems (BESS) as the energy management solution as they play a pivotal role in integrating electric vehicles, renewable energy sources and smart grids. Machine learning-based models are relied upon heavily to predict State of Charge (SoC) and State of Health (SoH) of batteries and therefore ensuring the robustness and reliability of these models has become a critical concern. These models not only manage the energy flow but also optimize battery life, predict degradation, and support real-time decision-making. However, the rise of adversarial attacks, especially poisoning attacks, poses a significant threat to the accuracy and reliability of these predictive models (Sayghe et al., 2020).

The use of estimation models has traditionally been developed under the pretence that the data processed is clean and free from manipulation, as the systems rely on these machine learning algorithms to predict battery behaviour based on battery parameters including voltage, current, and temperature (Harippriya et al., 2022). Current advancements in adversarial machine learning have revealed that poisoning attacks are potential vulnerability where malicious actors will corrupt the training data to mislead the model’s learning process (Zhang et al., 2020). The malicious examples introduced into the training data may lead to incorrect or biased predictions that degrade the model’s performance (Ali & Mahmood, 2020).

Poisoning attacks are being developed to become increasingly sophisticated which has been demonstrated across many infrastructure systems including energy maintenance platforms and power grids (Sayghe et al., 2020; Singh & Gupta, 2024). The machine learning algorithms that that optimize the energy use are exploited which creates significant disruptions to prediction of SoC and SoH respectively. With reliability and safety being paramount for energy systems even small disturbances in model prediction can lead to large failures including increased operational costs, system wide outages or unanticipated battery degradation (Rahim & Jalali, 2020). This makes it important to safeguard this models against adversarial manipulations whilst also aiming to improve accuracy. (Wang & Tang, 2021).

This study will propose a framework from defending against poisoning attacks targeting the SoC and SoH classifiers. These poisoning attacks can manipulate the model during the training phases by injecting adversarial examples which leads to erroneous predictions. This is addressed by using augmentation techniques within the dataset that incorporate adversarial perturbed data into the training process (Harippriya et al., 2022; Shejwalkar & Papernot, 2023). This will improve the ability to mitigate the effects of poisoned data, which ensures more reliable predictions coming from the model even under the presence of adversarial inputs.

The deployment and use of machine learning models within energy systems will continue to expand and therefore defending against adversarial attacks is a necessity. The study will highlight the risks posed by data poisoning attacks and provide a roadmap for creating more reliable and resilient models. By enhancing the robustness, the BESS will operate more securely even when faced with sophisticated cyber-attacks.

**Problem Statement**

Using machine learning models to optimize and manage Battery Energy Storage Systems (BESS) has increased the potential for adversarial attacks creating a need to ensure the robustness of these systems.

State of Charge (SoC) and State of Health (SoH) estimators are key components of BESS, however these machine learning-based models are vulnerable to manipulation, adversarial attacks. These attacks introduce malicious data during the training phase, causing the model to learn incorrect patterns, potentially causing fault predictions during operation.

SoC and SoH estimation models for energy systems are at risk to these attacks due to the necessity for precise estimation where incorrect predictions can lead to unexpected failures, reduce operational efficiency, battery degradation and safety hazards. The existing models are not adequately designed to defend against adversarial attacks which makes them a a potential target to APT’s and State actors.

The research conducted aims to investigate and produce defence mechanisms that can protect machine learning models for SoC and SoH estimation from adversarial attacks with the primary goal is to ensure that these models remain reliable and accurate. Incorporating and developing framework to defend against adversarial attacks will enhance and improve the overall security and robustness of BESS which will ensure safe and efficient battery management across many different applications which is a common occurrence and attack vector in today’s cyber space.

**Literature Review**   
Existing research on adversarial attacks against machine learning models utilised in energy storage systems (BESS) will be critically examined in this literature review. State of charge (SoC) and state of health (SoH) estimations utilise machine learning as an integral component making them vulnerable to adversarial attacks such as data poisoning and false data injection.

Within the review the key gaps in adversarial defence mechanisms are identified and a synthesis of the current literature is conducted. This positions the research to develop more resilient and robust models to defend against these attacks.

Battery Energy Storage Systems (BESS)

Overview

Modern infrastructure uses Battery Energy Storage Systems (BESS) as a critical component for managing energy storage and distribution in renewable energy sources and electric vehicles. Estimation of the State of Charge (SoC) and State of Health (SoH) is a part of BESS, and it is important that it is accurate due to the key role in optimising battery performance, prolonging battery life, and ensuring operation safety. SoC is the available energy in a battery in comparison to its full charge whereas SoH indicates the condition of the battery and is expressed as a percentage. These metrics are important for predicting the charging requirements, expected life span and usage patterns of the battery which is necessary for battery management.

Estimation is normally conducting through Machine Learning (ML) models due to their ability to process large amounts of sensor data providing real time insights into battery behaviour. The estimation models provide more accurate and timely predictions of battery health and charge levels as they detect subtle trends which traditional approaches may overlook. (Attia et al., 2019). These models are vulnerable to adversarial attacks such as data poisoning which can cause the model to make inaccurate predictions. (Sayghe et al., 2020; Harippriya et al., 2022).

Sayghe et al. (2020) discusses the input data can be distorted by data poisoning attacks used estimations, which leads to incorrect decisions regarding battery management. The integrity of the input data is modified which can cause significant disruptions as the integration of the model, is under the assumption of clean data. These interruptions can come under the forms of premature battery degradation or increased operational costs (Rahim & Jalali, 2020). The current literature and studies emphasize developing robust defences to safeguard BESS models from adversarial attacks but there is a significant gap in the current literature with regards to the absence of defending SoC and SoH estimation models. (Wang & Tang, 2021). This demonstrates the need for further development with techniques such as incorporating adversarial modified data within the training process to improve the resilience of these models against such attacks. (Harippriya et al., 2022).

Core Technologies:

The ML algorithms within BESS are used for time-series data which means that continuous monitoring of the battery parameters is occurring. The common algorithms within this context include deep learning models (DL) such as Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), Convolutional Neural Networks (CNN), And hybrid models which combine these architectures to achieve strengths of them both.

Generally, LSTM is used for SoC and SoH estimation as it is effective when dealing with time-series predictions due to its ability retain information over long periods whilst simultaneously addressing the vanishing gradient problem (Liu et al., 2019). The accuracy of estimation model prediction was improved by utilising LSTM as the learn the long-term dependencies of the time series data which is important for detecting various trends.

For models where pattern recognition within the BESS data is important CNNs are utilised due to their strength when dealing with multi-dimensional sensor data. Feature extraction is also a strength of CNN’s which can be used to enhance estimation accuracy by identifying subtle changes that indicate degradation. On their own however CNN’s may miss temporal dependencies and are not as efficient with time series data as LSTMs. The limitations of CNN’s can be improved with hybrid models such as CNN-LSTMs which combine both strengths of feature extraction and temporal handling of data (Liu et al., 2019). Generally, these models have been found to be the best performing for SoC and SoH estimation as they leverage the strengths effectively. As estimation models grow more complex the risk of adversarial attack increases targeting either the feature extraction or the sequence learning phases (Attia et al., 2019). Therefore, ensuring the robustness of these advanced models is a critical area of ongoing research.

Adversarial Attacks on Machine Learning Models

Overview

Machine learning models can be manipulated using adversarial attacks to deceive the model into making incorrect estimations. These attacks can be classed as several different types such as **data poisoning, evasion attacks**, and **false data injection attacks (FDIA).**

When adversarial data is inserted during the interference stage an adversarial attack is occurring which causes the machine learning model to make incorrect predictions based on the interference data. (Khan & Ghafoor, 2024). Data poisoning targets the training data with malicious injections into the training set that result in incorrect patterns which affect the ability to make correct predictions. In the context of BESS these poisoning attacks can lead to unreliable battery management and unexpected failures during operation. (Zhang et al., 2020). The poisoned models can overestimate battery capacity or fail to predict battery degradation which may result in operational and management decisions based on incorrect data (Sayghe et al., 2020). An estimator trained on poisoned data might fail to predict when the battery is nearing critical degradation, this can lead to reduced operational efficiency or system outages (Liu et al., 2019). FDIA attacks are specifically designed to manipulate the flow of data within a system and are a subset of poisoning attack. The data introduced can mimic normal sensor data which is dangerous as it can make detection quite difficult. (Rahim & Jalali, 2020). These attacks can also cause multiple type of battery failures or operational disruptions and therefore are a critical vulnerability for BESS.

Adversarial attacks exploit BESS reliance on ML algorithms for which minor disturbances in the accuracy of estimations can lead to large failures and unexpected consequences. Robustness for this system is high priority as adversarial misclassifications may not be detected until problems and damage has already occurred to the battery or the system has experienced a failure. The consequences of perturbed data can be catastrophic especially when BESS are integrated into critical infrastructure and electrical vehicle networks. Having misleading predictions can lead to wide scale failures, safety hazards and even increased operational costs which demonstrates the importance of developing defences to these attacks.

Vulnerabilities to Adversarial Attacks:

Machine learning models used for **State of Charge (SoC)** and **State of Health (SoH)** estimation in **Battery Energy Storage Systems (BESS)** are particularly vulnerable to **adversarial attacks** due to their reliance on accurate **time-series data**. These models process sensor data to make predictions, and even small manipulations in the input data can lead to significant prediction errors, ultimately compromising the system's performance and safety.

**Adversarial attacks** exploit the sensitivity of machine learning models by introducing manipulated data into the input stream. For time-series data, this might involve subtle changes in the patterns ofvoltage, current, or temperature measurements that the model interprets over time. These small perturbations are often undetectable by traditional anomaly detection systems, yet they can dramatically affect the model’s prediction accuracy, leading to incorrect SoC or SoH estimations (Anderson & Thompson, 2021). For example, a minor shift in voltage readings could cause the model to overestimate the SoC, leading to overcharging and potential overheating of the battery.

The **temporal nature** of the data adds an additional layer of complexity, as machine learning models like **LSTM** and **CNN** depend on patterns and dependencies across multiple time steps to make predictions. An adversarial attacker can manipulate specific points within the time-series data to disrupt these dependencies, causing the model to fail in recognising the true state of the battery. These attacks can result in cascading prediction errors that not only affect the immediate decision-making process but also accumulate over time, worsening the model's overall accuracy and reliability (Severson et al., 2020).

Even minor **input manipulation** can lead to significant disruptions. For example, a slight modification in the temperature sensor data, designed to mimic normal fluctuations, could trick the model into underestimating the battery’s degradation. Over time, this could result in premature battery failure due to undetected overuse, reducing the lifespan of the battery and potentially leading to costly system failures (Rahim & Jalali, 2020).

Moreover, **false data injection attacks (FDIA)**, as previously discussed, are particularly dangerous in BESS because they can subtly alter time-series data in ways that are difficult to detect. By injecting false sensor readings that mimic legitimate data, an attacker can cause the machine learning model to make consistent and damaging misclassifications without triggering immediate alerts (Zhang et al., 2020). These vulnerabilities highlight the critical need for robust defences, especially in systems that depend heavily on time-series data for real-time decision-making.

Robustness and Defence Mechanisms

Defensive Mechanisms:

In existing literature there are several defence mechanisms proposed to increase resilience for estimation models against adversarial attacks. The idea behind these defences is to increase reliability and accuracy of the models even under adversarial conditions.

Robust training involves retraining machine learning models using adversarial perturbed data which in doing so makes the model more resilient to attacks during operation. The retraining process allows the model to recognise and adapt to adversarial conditions which reduces the risk of misclassification during attacks. Overall, this training makes the model more robust by making it less sensitive to perturbations in the data. (Islam et al., 2022). During this process data augmentation is also utilised on the training set with adversarial examples to improve the model’s ability to generalise to unseen inputs. This is done with the aim to enhance robustness but also ensure accuracy under different operating conditions. (Harippriya et al., 2022).

To defend against adversarial attacks anomaly detection is also another key defensive mechanism to identify false inputs. Anomaly detection systems monitor the data in real time and flag any inputs that deviate from the expected range or flag any unusual patterns. Anomaly detection is crucial as an early warning system to alert operators to potential adversarial manipulations before substantial damage is sustained. (Wang & Tang, 2021).

Hybrid models are used to combine traditional methods to increase system resilience and leverage the strength of both ML and statistical approaches. This is done to combine both high pattern detection accuracy and ensuring robustness against adversarial attacks. These models have been used to provide multiple layers of defence against different attack vectors against sublet threat actors whilst also identifying outliers and false positives. (Islam et al., 2022).

Challenges:

As the complexity of BESS increase the vulnerability to adversarial threats also increases and their evolving nature makes them difficult to mitigate or even detect. Attackers are always trying different attack vectors because whilst models can be trained to recognise specific adversarial inputs attackers can generate new was to perturb data to avoid detection. This means that defending against these attacks is a persistent challenge. (Soltani & Lavaei, 2021).

Another challenge identified in recent literature is the difficulty in generalising defensive mechanisms across BESS. Due to the wide scale of BESS the SoC and SoH estimation methods vary depending on the operational context of the systems specific requirements. In some systems different defensive measures may not be as effective as they are in other systems leading to vulnerabilities in the broader application of these defences. (Harippriya et al., 2022). This makes it a challenge to effectively create a solution that works for all BESS. Hybrid attacks that combine false data injection and data poisoning are becoming more common meaning that existing countermeasures may not be able to stop more advanced attacks (Wang & Tang, 2021).

Gaps in Literature and Future Trends

There are gaps in current studies that have been identified despite the significant advancements in the application of ML for BESS. These gaps lie in the long-term impacts of adversarial attacks and how to improve the resilience of SoC and SoH predication against poisoning and false data injection attacks. The degradation of a model’s performance under continuous adversarial pressure also needs further investigation to help improve the reliability of BESS within important infrastructure (Williams & Brown, 2020).

Current defence may also be a limitation of recent implementations due to the model specific nature as they are designed to protect specific ML architectures against attacks. Different strategies such as robust training or anomaly detection do not generalise well across different ML models or types of BESS. The lack of generalisability makes it harder to secure BESS applications across the board (Harippriya et al., 2022). There is also not a lot of current studies focussing on adversarial attacks during the operational phase and how to defend and detect them. When live operational data is attack this can compromise system reliability with a small window for detection or intervention. (Williams & Brown, 2020).

**Bibliography:**

1. Sayghe, A., Zhao, J., & Konstantinou, C. (2020). Evasion Attacks with Adversarial Deep Learning Against Power System State Estimation. In 2020 IEEE Power & Energy Society General Meeting (PESGM) (pp. 1-5). IEEE.
2. Harippriya, S., Vigneswaran, E.E., & Jayanthy, S. (2022). Battery Management System to Estimate Battery Aging using Deep Learning and Machine Learning Algorithms. Journal of Physics: Conference Series, 2325(1), 012004. IOP Publishing.
3. Zhang, X., Liu, Y., & Yu, H. (2020). Online Data Poisoning Attacks. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020) (pp. 1-12).
4. Ali, A., & Mahmood, A. (2020). Adversarial Poisoning Attacks on Reinforcement Learning-Driven Energy Pricing. Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS 2020) (pp. 1-10).
5. Singh, C., & Gupta, R. (2024). Accurate Joint Detection of False Data Injection Attacks on Islanded PV Output Power and SoH Estimation. International Conference on Smart Grid Renewable Energy (SGRE24) (pp. 1-6).
6. Rahim, P., & Jalali, M. (2020). Intelligent Microgrid Control Using Adversarial Machine Learning Approaches. IEEE Transactions on Industrial Informatics, 16(4), 2321-2330.
7. Wang, H., & Tang, H. (2021). Cybersecurity Threats in Battery Management Systems: A Comprehensive Review. Journal of Energy Storage, 35(6), 3010-3020.
8. Severson, K.A., Attia, P.M., Jin, N., Perkins, N., Herring, P.K., & Braatz, R.D. (2020). Data-driven Prediction of Battery Cycle Life Before Capacity Degradation. Nature Energy, 5(5), 307-315.
9. Liu, L., Wang, S., & Zhang, Y. (2019). Application of Machine Learning Algorithms for Battery Capacity and Cycle Life Prediction. Energies, 12(2), 382.
10. Anderson, M.T., & Thompson, L.M. (2021). Attacks on Power System Load Forecasting Models Using Adversarial Machine Learning. IEEE Transactions on Power Systems, 36(6), 5654-5663.
11. Zhang, W., & Gao, X. (2020). Machine Learning-Based Prediction of Battery SoC and SoH in Electric Vehicles. IEEE Transactions on Industrial Electronics, 67(9), 7849-7858.
12. Islam, S.N., Sharma, D.K., & Tehranipoor, M. (2022). Adversarial Learning in Industrial Control Systems: Challenges and Solutions. IEEE Transactions on Dependable and Secure Computing, 19(5), 2500-2512.
13. Williams, A., & Brown, R. (2020). Vulnerability of Industrial Control Systems to Adversarial Machine Learning Attacks. Computers & Security, 97, 101916.
14. Soltani, S., & Lavaei, J. (2021). A Survey on Adversarial Machine Learning in Energy Systems. Renewable and Sustainable Energy Reviews, 141, 110755.
15. Liu, Y., & Zhang, Y. (2020). Improving Robustness of Machine Learning in Battery State Estimation. Journal of Applied Energy, 269, 114918.
16. Zhang, Y., & Shen, Y. (2020). Adversarial Defence Strategies in Battery Management Systems. Energy Conversion and Management, 211, 112754.
17. Severson, K.A., Jin, N., & Braatz, R.D. (2020). Battery Cycle Life Prediction Using Machine Learning and Early-Cycle Data. IEEE Transactions on Energy Conversion, 35(4), 1846-1856.
18. Li, Z., Wang, S., & Cheng, X. (2021). Adversarial Attacks on BMS in Smart Grids. Energy Systems, 12(2), 349-360.
19. Nguyen, P., & Roberts, A. (2022). Adversarial Machine Learning for Battery Health Monitoring in Electric Vehicles. IEEE Transactions on Vehicular Technology, 71(8), 8811-8822.
20. Carter, S., & Zhang, H. (2022). Adversarial Robustness in Battery Management Using Hybrid Models. IEEE Transactions on Smart Systems, 7(3), 1240-1250.
21. Ferdowsi, E., Abdolhosseinzadeh, S., & Saleem, H. (2021). Data-driven State of Health Modeling of Battery Energy Storage Systems Providing Grid Services. IEEE Access, 9, 118353-118365.
22. Liu, L., Wang, S., & Zhang, Y. (2019). Application of Machine Learning Algorithms for Battery Capacity and Cycle Life Prediction. Energies, 12(2), 382.
23. Sayghe, A., Zhao, J., & Konstantinou, C. (2020). Evasion Attacks with Adversarial Deep Learning Against Power System State Estimation. In 2020 IEEE Power & Energy Society General Meeting (PESGM) (pp. 1-5). IEEE.
24. Zhang, W., & Gao, X. (2020). Machine Learning-Based Prediction of Battery SoC and SoH in Electric Vehicles. IEEE Transactions on Industrial Electronics, 67(9), 7849-7858.
25. Islam, S.N., Sharma, D.K., & Tehranipoor, M. (2022). Adversarial Learning in Industrial Control Systems: Challenges and Solutions. IEEE Transactions on Dependable and Secure Computing, 19(5), 2500-2512.
26. Williams, A., & Brown, R. (2020). Vulnerability of Industrial Control Systems to Adversarial Machine Learning Attacks. Computers & Security, 97, 101916.
27. Zhang, X., Liu, Y., & Yu, H. (2020). Online Data Poisoning Attacks. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020) (pp. 1-12).

In-Text References (Numbered):

1. (Sayghe et al., 2020)
2. (Harippriya et al., 2022)
3. (Zhang et al., 2020)
4. (Ali & Mahmood, 2020)
5. (Singh & Gupta, 2024)
6. (Rahim & Jalali, 2020)
7. (Wang & Tang, 2021)
8. (Severson et al., 2020)
9. (Liu et al., 2019)
10. (Anderson & Thompson, 2021)
11. (Zhang & Gao, 2020)
12. (Islam et al., 2022)
13. (Williams & Brown, 2020)
14. (Soltani & Lavaei, 2021)
15. (Liu & Zhang, 2020)
16. (Zhang & Shen, 2020)
17. (Severson et al., 2020)
18. (Li et al., 2021)
19. (Nguyen & Roberts, 2022)
20. (Carter & Zhang, 2022)
21. (Ferdowsi et al., 2021)
22. (Liu et al., 2019)
23. (Sayghe et al., 2020)
24. (Zhang & Gao, 2020)
25. (Islam et al., 2022)
26. (Williams & Brown, 2020)
27. (Zhang et al., 2020)