



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

공학석사 학위논문

리튬 이온 배터리의 SOH 추정 정확도  
향상을 위한 시계열 데이터 생성 및  
평가 방안 연구

A Study on Generation and Evaluation of  
Time-Series Data to Improve Lithium-Ion  
Battery State-of-Health Estimation

울산대학교 대학원

전기전자컴퓨터공학과

설수진

리튬 이온 배터리의 SOH 추정 정확도  
향상을 위한 시계열 데이터 생성 및  
평가 방안 연구

지도교수 김병우

이 논문을 공학석사학위 논문으로 제출함

2023년 08월

울산대학교 대학원

전기전자컴퓨터공학과

설수진

설 수 진의 공학석사학위 논문을 인준함

심사위원장 임 동 국 인

심사위원 김 병 우 인

심사위원 곽 수 진 인

울 산 대 학 교 대 학 원

2023년 08월

## 감사의 글

대학원에 입학한 지 엇그제 같은데 졸업논문을 마무리 지으며 감사의 글을 쓰니 감회가 새롭습니다. 부족한 제가 나날이 성장할 수 있었던 이유는 제 주위의 많은 사람들이 저를 지지해 주시고 응원해주신 덕분이라고 생각합니다. 이 자리를 빌려 주위의 모든 분들께 감사의 인사 전해드립니다.

먼저 2년간의 석사과정 동안 지도해주신 김병우 교수님께 감사의 마음 전해드립니다. 교수님의 도움으로 학문에서의 성취와 더불어 다양한 경험과 성과를 얻을 수 있었습니다. 교수님이 주신 가르침이 있었기에 제 삶에 대한 방향성이 잡을 수 있었고, 저의 길을 향해 나아갈 수 있었습니다. 또한, 바쁘신 와중에도 저의 부족한 석사학위논문을 심사해주신 임동국 교수님, 광수진 부문장님께 감사의 말씀을 드립니다.

석사과정 동안 함께 생활하며 지낸 자동차 전자제어 연구실의 선배, 후배들에게도 감사의 마음 전달해 드립니다. 누구보다 많은 시간들을 함께 하면서 가르쳐주고 힘이 되어주어서 많은 것들을 배울 수 있었습니다. 선배들의 조언과 후배들의 응원이 있었기에 석사 과정의 마지막에 올 수 있었던 것 같습니다. 늘 질문을 던져 스스로 생각하여 몸소 배울 수 있도록 해준 재우 오빠, 힘들 지칠 때 항상 옆에 있어준 경은 언니, 서툴고 부족한 부분을 채워나갈 수 있게 도와준 정은 언니, 항상 같이 웃고 떠들며 서로에게 힘이 되어준 팽도 오빠, 나의 첫 번째 사수 동희 오빠, 유일한 동기로서 같이 성장한 민식이, 스스로없이 장난도 잘 받아주는 정수, 어떤 일이든 최선을 다해서 해내는 차니까지, 있었기에 많은 성장을 이뤄낼 수 있었습니다. 그리고 자동차 전자 제어 연구실에 올 수 있도록 큰 역할을 해준 종규 오빠와 연구실에 방문해서 아낌없는 조언해주는 졸업생 선배님들까지 계신 우리 연구실의 모든 분들께 감사의 인사드립니다.

또한, 아낌없는 지지를 해주신 부모님, 친구같이 편안한 언니, 막내지만 든든한 동생 재욱이, 늘 반갑게 꼬리 흔들며 반겨주는 은행이까지. 가족에게도 감사의 마음 전합니다.

그리고 늘 응원을 보내주는 모든 분들께 감사의 말 전합니다. 한 명 한명 이름을 적지 못하였지만 주위의 모든 사람들에게 감사의 인사를 전해드립니다.

2023. 06. 30.

설수진 올림

# 리튬 이온 배터리의 SOH 추정 정확도 향상을 위한 시계열 데이터 생성 및 평가 방안 연구

설 수 진

울산대학교대학원 전기전자컴퓨터공학과

## 요 약

최근 화석 연료 차량의 규제가 강화됨에 따라 친환경 에너지원인 리튬 이온 배터리를 활용한 xEV 시장이 확대되고 있다. 배터리의 사용에 있어 안전성과 신뢰성 확보는 필수적이며, 그에 따라 배터리의 건강 상태(SOH, State Of Health) 추정이 중요하다. SOH 추정 방법에는 모델 기반과 데이터 기반이 있다. 그중 데이터 기반으로 SOH를 추정하기 위해서는 많은 양과 높은 품질의 배터리 데이터셋을 필요로 한다. 실측이 어려운 데이터셋 확보 방법으로는 데이터 증강 알고리즘이 있다. 데이터 증강 알고리즘을 통해 생성된 시계열 데이터셋의 품질 평가는 간접적인 평가가 주로 사용되며, 이는 학습 환경에 따라 결과가 유동적이라는 한계가 있다. 따라서 본 논문에서는 SOH 추정 정확도 향상을 위하여 생성된 배터리 데이터셋의 품질을 정량적으로 평가하기 위한 방법으로 선형회귀의 상관계수 변화율과 클러스터링 지표인 Dunn Index, 실루엣 계수를 제안한다. 또한, 제안한 평가 방법을 통해 신뢰성이 확보된 많은 양과 높은 품질의 데이터셋을 활용하여 LSTM 및 GRU를 통해 SOH 추정 정확도를 비교하였다.

주요어 : 리튬이온배터리(lithium-ion battery), GAN(Generative Adversarial Networks), TimeGAN(Time-series Generative Adversarial Networks), 데이터셋(dataset), 합성 데이터(synthetic data), 선형 회귀(linear regression), Dunn Index, 실루엣 계수(silhouette coefficient) SOH 추정(SOH estimation), 딥러닝(deep learning), RNN(Recurrent Neural Network), LSTM(Long Short Term Memory), GRU(Gate Recurrent Unit)

# 목 차

목 차 .....	i
표 및 그림 목차 .....	iii
<b>I. 서론</b> .....	<b>1</b>
1. 연구 배경 및 필요성 .....	1
2. 연구 이론 및 내용 .....	4
2.1. GAN(Generative Adversial Network) .....	4
2.2. t-SNE .....	5
2.3. RNN(Recurrent Neural Network) .....	6
<b>II. 제안 알고리즘</b> .....	<b>7</b>
1. 연구 개요 .....	7
2. 리튬이온 배터리 데이터 .....	8
3. 데이터 전처리 .....	9
4. TimeGAN 기반의 데이터 증강 .....	11
4.1. Embedding & Recovery .....	11
4.2. Generator & Discriminator .....	12
5. 합성 배터리 데이터셋 평가 .....	13
5.1. 선형회귀의 상관계수 변화율 .....	14
5.2. Dunn Index .....	15
5.3. 실루엣 계수 .....	16
6. SOH 추정알고리즘 .....	17
6.1. LSTM .....	17
6.2. GRU .....	19
7. SOH 추정 정확도 평가 .....	21
<b>III. 결과 및 분석</b> .....	<b>22</b>
1. 합성 데이터셋 평가 .....	22
1.1. 선형회귀의 상관계수 변화율 .....	25
1.2. Dunn Index .....	26

1.3. 실루엣 계수 .....	26
2. SOH 추정 정확도 평가 .....	28
2.1. LSTM .....	29
2.2. GRU .....	30
IV. 결론 .....	32
참고문헌 .....	34
ABSTRACT .....	38

# 표 및 그림 목차

## I. 표 목차

[표 1] 리튬 이온 배터리 데이터셋 .....	8
[표 2] 합성 배터리 데이터셋 평가 결과 .....	27
[표 3] SOH 추정 정확도 향상률 비교 결과 .....	31

## II. 그림 목차

[그림 1] GAN 구조 .....	4
[그림 2] 가우시안 분포 vs t 분포 .....	5
[그림 3] RNN 구조 .....	6
[그림 4] 연구 개요 .....	7
[그림 5] 배터리 데이터 측정 사이클 .....	8
[그림 6] 3차원 변환 샘플링 .....	10
[그림 7] TimeGAN 구조 .....	12
[그림 8] LSTM 구조 .....	18
[그림 9] LSTM 레이어 .....	18
[그림 10] GRU 구조 .....	20
[그림 11] GRU 레이어 .....	20
[그림 12] RW9의 학습 횟수에 따른 t-SNE 변화 .....	23
[그림 13] RW10의 학습 횟수에 따른 t-SNE 변화 .....	23
[그림 14] RW11의 학습 횟수에 따른 t-SNE 변화 .....	24
[그림 15] RW12의 학습 횟수에 따른 t-SNE 변화 .....	24
[그림 16] 선형회귀 변화율(Iteration 10000회) .....	25
[그림 17] LSTM의 RMSE boxplot .....	29
[그림 18] GRU의 RMSE boxplot .....	30

# I. 서론

## 1. 연구 배경 및 필요성

최근 환경 문제가 대두되면서 자동차 산업에서도 화석 연료 차량에 대한 규제가 강화되고 있으며, xEV(전기자동차, 수소연료자동차, 하이브리드자동차 등)의 시장이 확대되고 있다. xEV 핵심부품 중 하나인 배터리는 상태나 관리 방법에 따라 시스템의 효율성 또는 안전성에 영향을 미친다. 따라서 배터리의 안전성과 신뢰성의 확보를 위해 배터리의 상태(SOx) 모니터링의 필요성이 대두되고 있다.

SOH는 용량 대비 특정 시점에서 완전히 충전된 배터리가 방전할 수 있는 최대 용량의 백분율로 정의한다. 이는 배터리의 노화 상태를 나타내는 중요한 지표로써, SOH 추정 방법에 대하여 활발히 연구되고 있다. SOH 추정 방법은 크게 두 가지 방법으로 구분되며, 모델 기반 추정 방법[1-3]과 데이터 기반 추정 방법[4-6]이 있다.

모델 기반의 SOH 추정은 배터리의 물리/화학적 원리를 분석하여 수학적으로 모델링하고, 그 모델을 기반으로 추정 필터를 적용하는 방법이다. [1]는 테브난 배터리 모델 기반의 KF(Kalman Filter)를 통해 배터리의 SOH를 추정하였고, [2]는 UPF(Unscented Particle Filter) 알고리즘을 통해 배터리의 SOH를 추정하였다. [3]는 Dual EKF를 통해 배터리의 SOH를 추정하였다. 하지만 모델 기반의 SOH 추정은 비선형적으로 복잡하게 변하는 내부의 상태를 선형화하여 모델링 해야 하는 한계가 있다.

데이터 기반의 SOH 추정 연구는 배터리 내부의 복잡한 상태변화를 고려하지 않아도 데이터의 특성을 반영하여 SOH를 추정할 수 있다. [4]는 GRU를 기반으로 한 DSAT-GRU 알고리즘을 이용하여, [5]는 CNN-Transformer를 이용하여 SOH를 추정했다. [6]는 LSTM과 NN을 결합한 AST-LSTM NN을 이용하여

SOH 추정 및 RUL 예측을 했다. SOH 추정 정확도 향상을 위해서 데이터 기반의 다양한 알고리즘이 연구되고 있으며 이러한 데이터 기반의 SOH 추정 방법은 데이터의 특성에 따라 좋은 성능을 내는 알고리즘이 다르며, 많은 양과 높은 품질의 데이터셋을 필요로 한다.

데이터셋의 중요성이 강조되면서 데이터 증강 알고리즘에 대한 연구가 활발하게 이루어지고 있다. 이미지 분야에서 DCGAN[7], WGAN[8], LSGAN[9] 등의 알고리즘이 개발되어 왔고, 현재에는 시계열 분야에서도 C-RNN-GAN[10], RCGAN[11], TimeGAN[12] 등의 알고리즘이 연구되고 있다. C-RNN-GAN[10]은 GAN 아키텍처를 시퀀스 데이터에 직접 적용하였고, RCGAN[11]은 유사하지만 이전 출력에 대한 의존도를 낮추는 사소한 아키텍처 차이를 도입하였다. TimeGAN[12]은 시계열 데이터를 일반화하고, Latent space를 통해 데이터의 차원을 축소하여 학습을 용이하게 하였다.

또한 TimeGAN을 통해 많은 양과 높은 품질의 데이터를 생성하여 추정 및 분류 알고리즘의 성능을 향상시키는 연구가 진행되고 있다. [13]는 수집이 어려운 타이어 마모 데이터를 생성하여 데이터의 예측 정확도를 높이며, [14]는 조건에 따라 다른 특성을 가지는 난방 시스템 데이터를 보완하여 예측 정확도를 높인다. 또한, [15]는 고장 분류 정확도를 향상시키기 위하여 데이터 증강 알고리즘을 통해 베어링 고장 데이터의 불균형을 해소한다. 따라서 배터리와 같이 제한적인 양을 가진 오픈 소스 데이터셋에 대해서는 데이터셋을 보완한 추정 및 분류 정확도 향상 연구가 필요하다.

생성된 학습 데이터셋에 비정상적인 특징을 갖는 데이터가 많은 경우, 학습모델 정확도가 저하되기 때문에 많은 양과 높은 품질의 데이터셋을 확보가 요구된다. 이미지 분야에서 합성 데이터를 평가하기 위해서는 데이터 증강 알고리즘의 학습 손실률[8]과 분류 알고리즘의 정확도[9-10] 비교를 통한 간접적인 평가 방법을 활

용한다. 또한, 생성된 데이터의 품질을 시각적으로 판단하기 위하여 평가자의 주관[8] 또는 다수의 의견이 반영된[9] 정성적 평가 방법을 사용한다. 최근에는 이미지의 품질과 다양성을 반영한 IS 지표[16], 이미지의 중복성, 특성, 유사성을 반영한 CID index[17] 등으로 검증하는 정량적 평가 방법도 연구되고 있다. 이러한 지표는 같은 라벨(label)의 이미지의 중복성, 분포 등을 평가하기 때문에 시계열 데이터에는 적합하지 않다.

합성 시계열 데이터의 평가를 위해서도 데이터 증강 알고리즘의 손실률[12]과 분류 알고리즘의 정확도[11] 비교를 통한 간접적인 평가 방법을 사용한다. 또한, 시계열 데이터의 특성을 반영한 t-SNE[18], PCA[19] 등의 저차원 그래프 또는 샘플 단위 데이터의 그래프를 시각적으로 비교하는 정성적인 평가 방법도 사용하고 있다.[12] 하지만 이러한 평가 방법은 평가자의 주관에 개입되어 객관적인 평가가 어려우며, 적용 알고리즘 및 학습 환경에 따라 결과의 기준이 달라 비교가 어렵다. 또한 그 결과가 유동적으로 변한다는 한계가 있다. 따라서 시계열 데이터의 신뢰성을 직접적-정량적으로 비교할 수 있는 방법에 대한 연구가 필요하다.

본 논문에서는 학습에 영향을 끼치는 중한 요소인 데이터셋을 시계열 데이터 증강 알고리즘인 TimeGAN으로 보완하여 합성 시계열 데이터를 제안하는 정량적 평가 방법을 통해 직접적으로 데이터셋을 평가하고, 평가된 데이터를 통하여 배터리 SOH 추정 정확도를 향상시키고자 한다. 배터리 데이터는 실측이 어렵기 때문에 TimeGAN을 통해 생성한다. 생성된 데이터의 품질 확보하기 위한 방법으로 선형회귀의 상관계수 변화율, Dunn Index[20], 실루엣 계수[21] 지표를 통해 합성 시계열 데이터를 직접적이고 정량적인 평가 방법을 제안하고자 한다. 또한, 높은 품질로 평가된 합성 배터리 데이터셋을 활용하여 순환 신경망을 기반으로 한 LSTM[22]과 GRU[23]를 통해 SOH 추정 정확도 향상률을 비교한다. 이를 통해 본 논문에서 제안한 데이터셋 생성 및 평가를 통한 SOH 추정 정확도 향상 방안의 유용성을 확인하고자 한다.

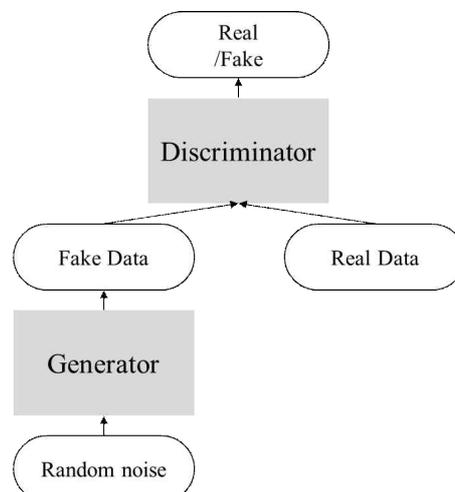
## 2. 연구 이론 및 내용

### 2.1. GAN(Generative Adversial Network)

데이터 증강의 대표적인 알고리즘으로 GAN[24]이 있으며, 이를 기반으로 이미지 분야(DCGAN[7], WGAN[8], LSGAN[9]등)와 시계열 분야(C-RNN-GAN[10], RCGAN[11], TimeGAN[12] 등)로 다양한 알고리즘이 파생되었다. 기본 GAN 알고리즘은 [그림 1]과 같이 생성자와 판별자로 구성되어 있으며, 생성자는 입력으로 랜덤 노이즈를 사용하여 합성 데이터를 생성하고, 판별자는 원본 데이터와 합성 데이터를 분류하는 역할을 한다. 식(1)은 생성자와 판별자의 역할을 기반으로 GAN 알고리즘은 원본 데이터와 유사한 합성 데이터를 생성하는 것을 목표로 한다. 이 때,  $D$ 와  $G$ 는 각각 판별자와 생성자를 나타내며,  $V(D, G)$ 는 판별자와 생성자의 가치 함수,  $E$ 는 기댓값을 나타낸다.

$$\min_G \max_D V(D, G) = E_{x \sim P_{data}(x)} [\log D(x)] + E_{z \sim P_Z(z)} [\log(1 - D(G(z)))] \quad (1)$$

판별자가 원본 데이터를 True로, 합성 데이터를 False로 판단할 때, 가치 함수가 최댓값을 가진다. GAN 알고리즘은 가치 함수가 최대가 되는 방향으로 학습한다.



[그림 1] GAN 구조

## 2.2. t-SNE

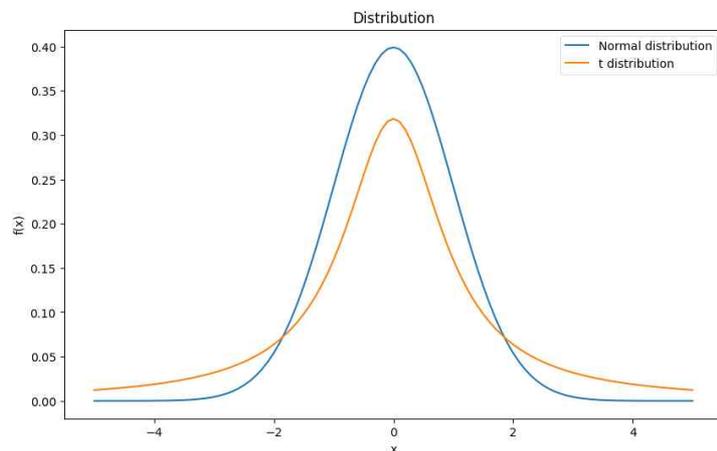
t-SNE[18]는 고차원의 데이터를 저차원으로 표현하는 차원 축소 기법 중 하나로, SNE에서 사용되는 가우시안 분포의 단점을 보완한 기법이다.

SNE[25]는 고차원 데이터를 저차원으로 축소하면서 거리가 가까운 데이터의 거리 정보를 우선하여 거리 정보를 확률로 보존하며, 거리에 따른 확률은 식(2)과 같은 가우시안 분포를 따르는 특징을 가지고 있다. 하지만, 가우시안 분포는 확률값의 변화가 커서 특정 거리 이상에서는 학습이 반영되지 않는 한계가 있다.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

이러한 한계를 극복하기 위해서 정규분포보다 예측 범위가 넓은 t 분포를 사용한 t-SNE 차원 축소 기법이 사용되고 있다. 여기서 t 분포는 식(3)과 같이 정의되며, 표준정규분포보다 분산이 크므로 [그림 2]과 같이 양쪽 꼬리가 두껍게 나타나며 예측 범위가 넓음을 나타낸다.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad (3)$$



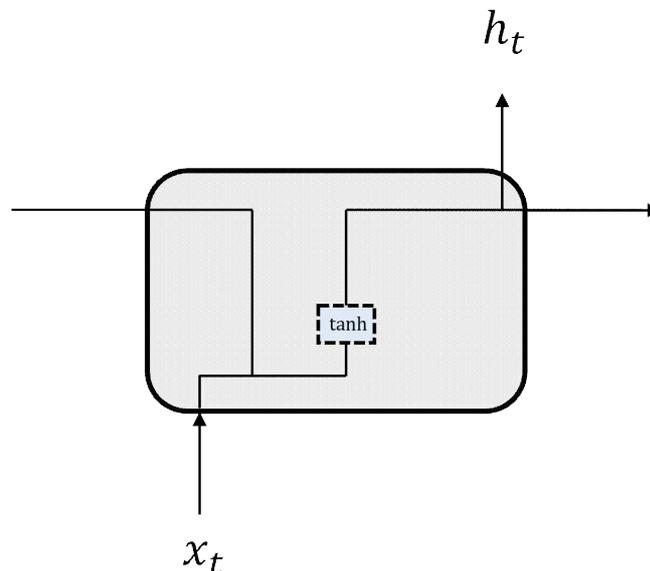
[그림 2] 가우시안 분포 vs t 분포

## 2.3. RNN(Recurrent Neural Network)

순환 신경망의 대표적인 알고리즘인 RNN[26]은 일반적인 단방향 신경망과 달리 순환적 구조를 가진 인공신경망의 한 종류이다. RNN 구조는 [그림 3]과 같으며, 시간에 따라 변하는 특성을 가진 데이터들을 모델링 할 수 있는 특징이 있다. 식(4)은 순환신경망의 상태 정보는 현재의 입력과 이전의 상태 정보를 통해 결정됨을 나타낸다.

$$h_t = \tanh(W_h * [h_{t-1}, x_t] + b_h) \quad (4)$$

RNN은 시퀀스 단위로 입력과 출력을 처리하며, 입출력의 시퀀스 길이를 다르게 설계할 수 있어 SOH 추정[27], 언어 번역[28], 스팸 메일 분류[29] 등 다양한 분야에 사용되고 있다. 가장 단순한 형태의 Vanilla RNN은 시퀀스가 길어질수록 기울기 소실로 인하여 앞쪽의 정보를 충분히 반영하지 못하는 현상이 발생하는 장기 의존성 문제가 있으며, 이러한 한계를 보완한 알고리즘으로는 LSTM[22]과 GRU[23]가 있다.

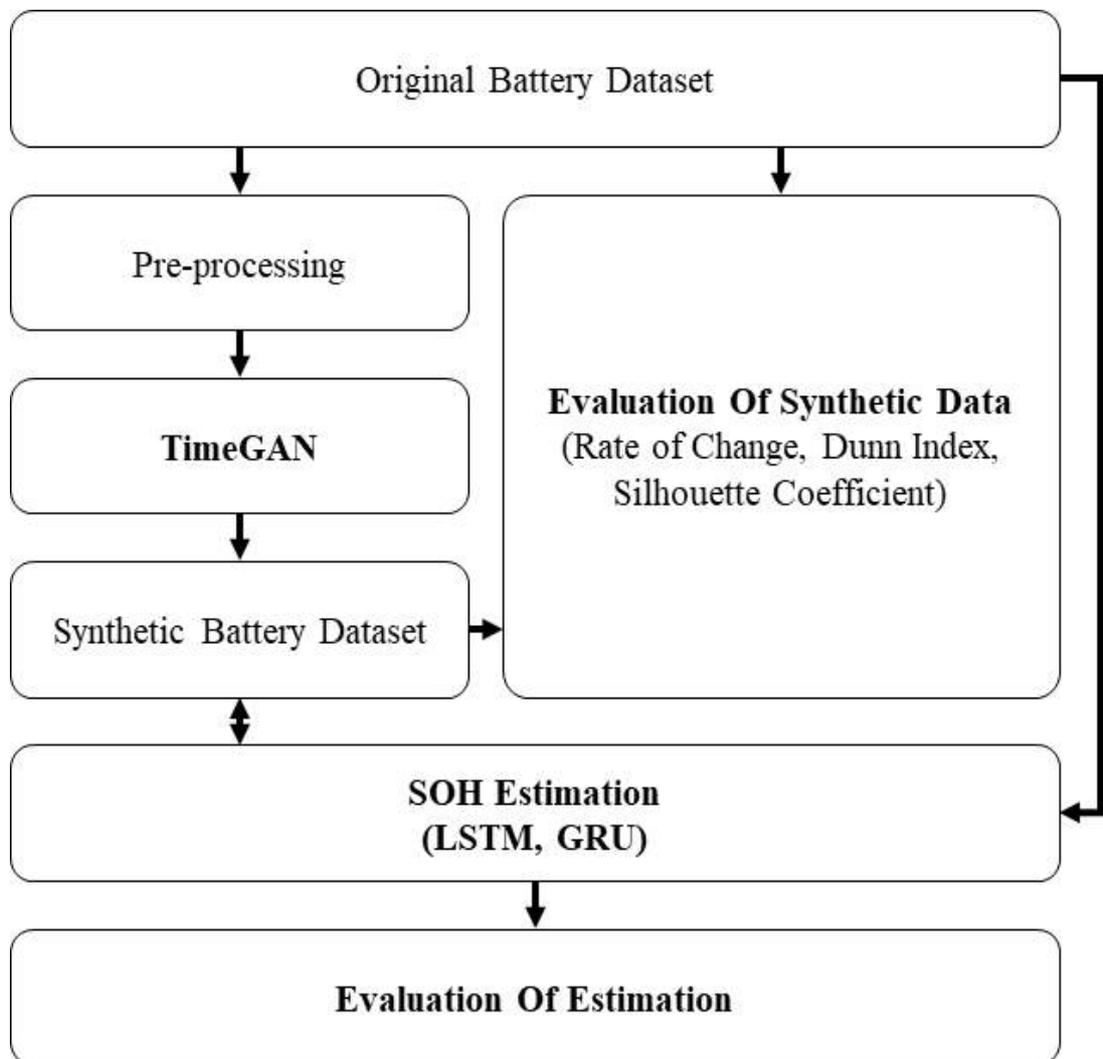


[그림 3] RNN 구조

## II. 제안 알고리즘

### 1. 연구 개요

연구 개요는 [그림 4]에 제시되어 있다. 첫 번째, TimeGAN[12]을 이용하여 합성 배터리 데이터를 생성한다. 두 번째, 정량적 평가 기법을 통해 합성 배터리 데이터셋의 품질을 평가한다. 세 번째, 입력을 두 가지로 나누어 LSTM과 GRU 알고리즘을 사용하여 SOH를 추정한다. 학습 시 입력 데이터셋은 두 가지로, 하나는 원본 데이터셋이며, 또 다른 하나는 원본 데이터셋에 합성 데이터가 추가된 데이터셋이다. SOH 추정의 정확도 비교를 위하여 RMSE 지표로 평가를 진행한다.



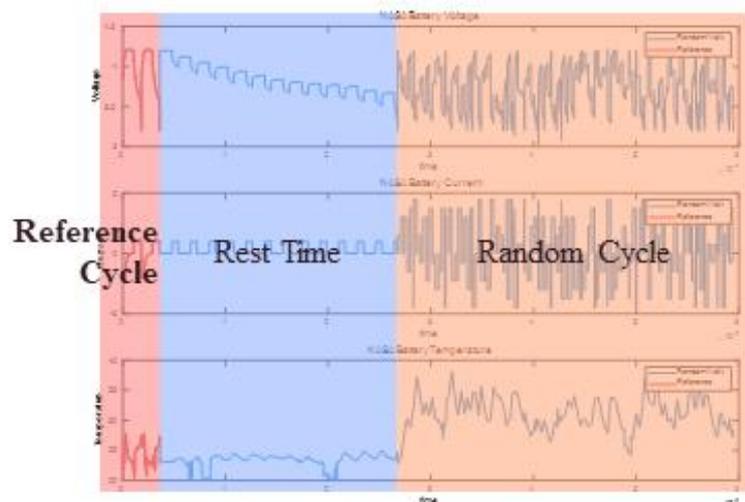
[그림 4] 연구 개요

## 2. 리튬이온 배터리 데이터

본 논문에서 사용한 데이터는 NASA Randomized Battery Dataset[30]이며, 배터리는 LG 18650 배터리로 사용된 배터리 데이터의 특성은 [표 1]와 같다. 해당 데이터셋은 28개의 셀을 7개의 그룹으로 구분된다. 각 그룹에 따라 테스트 조건(온도, 랜덤 충방전 조건)을 가변 하여 데이터를 측정하였으며, 측정 사이클은 [그림 5]과 같다. 해당 데이터셋은 전압 전류, 온도 등의 데이터로 구성되어 있다. 그 중 실온에서 랜덤하게 충전과 방전을 수행한 RW9, RW10, RW11, RW12 셀 데이터를 활용하였으며, 전체 사이클 중에서 최대 커패시터 용량을 측정한 기준 충방전 데이터를 활용하였다.

[표 1] 리튬 이온 배터리 데이터셋

battery properties	18650 LIBs
chemistry	18650 lithium cobalt oxide vs. graphite
nominal capacity	2.10Ah
capacity range	0.80~2.10Ah
voltage range	3.2~4.2V



[그림 5] 배터리 데이터 측정 사이클

### 3. 데이터 전처리

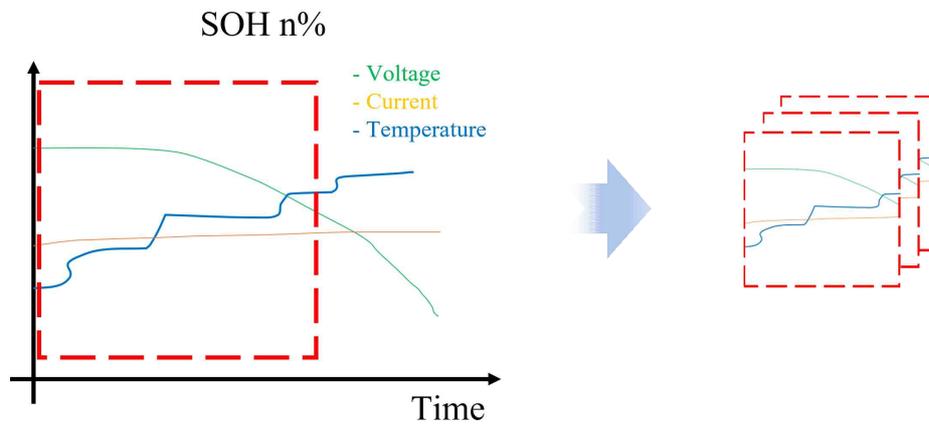
배터리 데이터셋을 시계열 데이터 증강 알고리즘 학습의 입력으로 사용하기 위하여 3단계의 전처리를 실시하였다.

첫 번째, 빠른 최적화 속도와 학습의 정확도 향상을 위해 정규화를 하였다. 정규화를 통해 입력 데이터(온도, 전류, 전압, SOH)의 스케일에 치우치지 않도록 한다. 정규화가 이루어지지 않는 경우, 스케일이 큰 파라미터의 오차가 많이 반영된다는 한계점이 존재한다. 본 논문에서는 최대-최소 정규화 기법을 활용하였으며, 식(5)과 같다. 이 때,  $x_{new}$ 는 정규화가 완료된 데이터이고,  $x$ 는 원본 데이터이다.  $x_{max}, x_{min}$ 는 각각 원본 데이터의 최대, 최솟값을 나타낸다.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (5)$$

두 번째, TimeGAN은 3차원 데이터를 학습하므로 배터리 데이터셋을 3차원으로 변환하기 위해 샘플링을 수행하였다. 3차원 데이터는 [그림 6]과 같이 하나의 시퀀스 내에 전압, 전류, 온도, SOH 데이터를 포함한다. 3차원 샘플링을 통하여 TimeGAN 학습 시 배터리 데이터의 다양한 특성을 반영할 수 있다.

세 번째, 학습 데이터와 테스트 데이터를 일정 비율로 나누었다. 배터리 데이터의 60%는 제안하는 알고리즘의 학습 데이터로 사용하고, 40%는 제안 알고리즘을 검증하기 위한 테스트 데이터로 사용하였다.



[그림 6] 3차원 변환 샘플링

## 4. TimeGAN 기반의 데이터 증강

GAN 알고리즘을 기반으로 시계열 데이터의 시간적 특성을 반영하는 TimeGAN 알고리즘이 제안되었다. TimeGAN 구조는 [그림 7]과 같이 GAN 구조에 오토 인코더 아키텍처가 추가되어 작은 차원에서 시간 역학(temporal dynamics)을 학습할 수 있다. 오토 인코더에서는 임베딩과 리커버리 함수를 통해 데이터를 축소 및 복원하며, GAN에서는 축소된 차원에서 데이터 생성 및 판별 함수를 통해 원본과 유사한 합성 배터리 데이터셋을 생성한다.

### 4.1. Embedding & Recovery

[그림 5]에서의 오토 인코더는 임베딩 함수와 리커버리 함수로 구성되어 있으며, 데이터의 차원을 축소하거나 복원한다.

오토 인코더에서 임베딩 함수는 데이터의 시간적 특징을 축소된 차원의 Latent space에 매핑(mapping) 하는 역할을 한다. 기존 차원에서 데이터의 정적 특징 공간( $s$ )과 시간적 특징 공간( $x_t$ )을 Latent space에서의 정적 특징 공간( $h_s$ )과 시간적 특징 공간( $h_t$ )로 나타낼 수 있으며, 식(6), (7)과 같다.

$$h_s = e_s(s) \quad (6)$$

$$h_t = e_X(h_s, h_{t-1}, x_t) \quad (7)$$

리커버리 함수는 데이터의 학습을 위해 축소된 데이터를 기존의 차원으로 복원하는 역할을 한다. Latent space로 표현된 정적 특징 공간( $h_s$ )과, 시간적 특징 공간( $h_t$ )을 기존 차원의 정적 특징 공간( $s$ ), 시간적 특징 공간( $x$ )로 복원한다. 리커버리 함수는 식(8), (9)과 같다.

$$\tilde{s} = r_s(h_s) \quad (8)$$

$$\tilde{x}_t = r_X(h_t) \quad (9)$$

## 4.2. Generator & Discriminator

[그림 5]에서의 GAN은 생성자 함수와 판별자 함수를 통해, 차원이 축소된 Latent space에서 데이터를 생성 및 판별한다.

생성자 함수는 랜덤 노이즈를 입력으로 받아 합성 데이터를 생성한다. Latent space에서의 정적 특징 공간은 랜덤 노이즈를 입력으로 가지고, 시간적 특징 공간은 Latent space의 합성 정적 특징 공간, 이전 시점의 합성 시간적 특징 공간, 기존 차원의 시간적 특징 공간을 입력으로 한다. 식은 식(10), (11)과 같다.

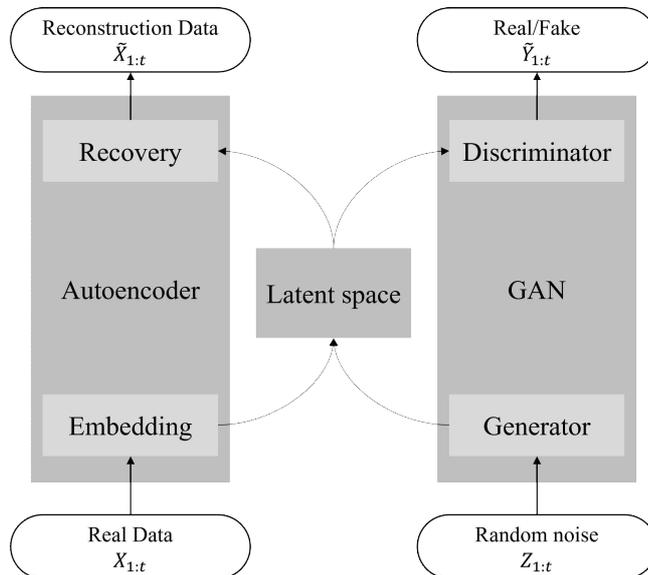
$$\hat{h}_s = g_s(z_s) \quad (10)$$

$$\hat{h}_t = g_X(\hat{h}_s, \hat{h}_{t-1}, x_t) \quad (11)$$

판별자는 Latent space에서 합성 데이터와 원본 데이터를 판단하며, 식(12), (13)과 같다.

$$\tilde{y}_s = d_s(\tilde{h}_s) \quad (12)$$

$$\tilde{y}_t = d_X(-\vec{u}_t, \vec{u}_t) \quad (13)$$



[그림 7] TimeGAN 구조

## 5. 합성 배터리 데이터셋 평가

합성 데이터셋을 평가하기 위해서 기존 연구[7-12]에서는 데이터의 시각적인 평가가 주를 이루고 있으며, 데이터를 정량적으로 표현하기 위하여 추정 및 분류 알고리즘을 통하여 정확도 비교를 평가 방법으로 활용하고 있다.

하지만 시각적인 평가는 평가자의 주관적이며 평가자에 따라 다른 해석 결과를 가질 수 있다는 한계가 있으며, 다양한 조건에 따른 합성 데이터의 우수성을 확인하기에는 부족하다. 또한, 추정 및 분류 알고리즘을 통해 평가를 진행하는 것은 학습 환경에 따라 그 결과가 변하여 객관적인 비교가 어렵다는 한계가 있다. 따라서 생성된 데이터셋의 품질을 평가하는 방법에 대한 연구의 필요성이 대두되고 있다.

따라서 합성 배터리 데이터의 품질 확보를 위해 직접적-정량적인 평가 방법을 제안하고자 한다. t-SNE를 이용하여 데이터의 특성을 가장 잘 나타낼 수 있는 이차원 그래프로 나타내고, 세 가지의 평가 지표(선형회귀의 상관계수 변화율, Dunn Index, 실루엣 계수)를 통해 정량적으로 데이터셋의 품질을 평가한다.

## 5.1. 선형회귀의 상관계수 변화율

선형회귀는 주어진 데이터로부터 두 변수 간의 선형 상관관계를 모델링 하는 회귀 분석 기법으로써, 주어진 데이터를 가장 잘 설명할 수 있는 상관 계수를 찾아 변수 간의 관계를 설명하는 분석 기법이다. 선형회귀식은 주로 목표 변수 추정 및 예측에 사용되며, 결정 계수를 통해 선형회귀식의 신뢰도를 평가한다. 결정 계수란 선형회귀 모델이 주어진 데이터를 얼마나 잘 설명할 수 있는가를 정량적으로 나타내는 지표로 식(14)과 같다. 이 때,  $y_i$ 는 실제 데이터,  $\hat{y}_i$ 는 선형회귀에 의해 추정한 데이터,  $\bar{y}$ 는  $y_i$  또는  $\hat{y}_i$ 의 평균값을 의미한다.

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (14)$$

본 논문에서는 선형회귀에서 상관계수 변화율을 계산하여 데이터의 품질을 평가하고자 한다. 이차원 그래프에서 원본 데이터셋과 합성 데이터셋의 분포를 선형회귀를 통해 나타내고, 결정 계수를 통하여 상관계수의 신뢰도를 확보한다. 신뢰도가 확보된 원본 데이터셋의 상관계수와 합성 데이터를 추가한 데이터셋의 상관계수를 식(15)를 통해 데이터의 품질을 평가한다. 원본 데이터셋과 합성 데이터셋이 유사한 경우, 상관계수의 변화가 작으므로 변화율의 수치가 작다. 원본 데이터셋과 합성 데이터셋이 상이한 경우, 상관계수 변화율은 크게 나타난다.

$$\text{변화율}(\%) = \frac{r_{ori} - r_{syn}}{r_{ori}} \times 100 \quad (15)$$

이 때,  $r_{ori}$ 는 원본 데이터셋의 상관계수를 의미하며,  $r_{syn}$ 는 원본 데이터에 합성 데이터가 더해진 데이터셋의 상관계수를 의미한다.

## 5.2. Dunn Index

클러스터링 평가 지표 중 Dunn Index는 문서의 키워드 분류[19], 유효성 평가 [20] 등 클러스터링 분야에서 군집화를 평가하는 대표적인 지표이다[16]. Dunn Index는 같은 군집 내 속해 있는 두 개체 간 최대 거리와 서로 다른 군집에 속해 있는 두 개체 간의 최소 거리의 차이를 비율로 나타내며, 식(16)과 같다. 이 때,  $\Delta_k$ 는 같은 군집 내 개체 간의 Euclidean 거리를 나타내며,  $\delta(C_i, C_j)$ 는 서로 다른 군에 속한 개체 간의 Euclidean 거리를 나타낸다.

$$DI_K = \frac{\min \delta(C_i, C_j)}{\max \Delta_k} \quad (16)$$

같은 군집 내의 개체 간 거리가 가깝고 서로 다른 군집에 속한 개체 간의 거리가 커진 경우, Dunn Index는 큰 값을 나타내고 이는 클러스터링이 잘 이루어졌다고 평가한다. 이와 반대로 Dunn Index가 작은 경우, 같은 군집 내의 개체 간 거리보다 다른 군집의 개체 간의 거리가 더 작으므로 클러스터링이 잘 이루어지지 않았다고 평가한다. 즉, 두 군집이 유사한 특성의 구분이 어렵다고 판단한다.

따라서 본 논문에서는 데이터의 품질을 평가하기 위하여, 축소된 차원에서의 데이터를 나타내고 Dunn Index를 사용하여 원본 데이터셋과 합성 데이터셋 간의 유사도를 평가하고자 한다. Dunn Index가 작은 경우, 특성의 구분이 어려우므로 원본 데이터셋과 합성 데이터셋의 특성이 유사하다고 판단한다. Dunn Index가 큰 경우, 원본 데이터셋과 합성 데이터셋의 특성이 상이하여 클러스터링이 가능하며, 두 특성이 다르다는 것을 나타낸다. 따라서 품질이 좋은 합성 데이터셋은 원본 데이터셋과 유사한 특성을 지녀야 하며, Dunn Index는 작은 값을 가진다.

### 5.3. 실루엣 계수

Dunn Index와 더불어 클러스터링의 평가 지표로 실루엣 계수가 있다.[17] 실루엣 계수는 한 군집의 군 내 유사도와 인접한 다른 군 간 유사도를 비교하며, 식 (17)과 같다. 이 때,  $a(i)$ 는 군집 내 다른 데이터와의 거리 평균을 의미하며,  $b(i)$ 는 다른 군집의 데이터와의 거리 평균을 의미한다.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (17)$$

실루엣 계수는 각 개체가 적절한 군집으로 속한 경우, 1에 가까운 값을 가지며 그렇지 않은 경우 -1에 가까운 값을 가진다. 같은 군집 내의 개체 간의 거리가 짧고 데이터 특성이 명확한 경우, 1에 가까운 값을 가진다. 서로 다른 군집 간의 유사도는 가깝지만 같은 군집 내에서 유사도가 떨어지는 경우 -1에 가까운 값을 가진다. 서로 다른 군집 간 개체의 거리와 같은 군집 내의 개체 간 거리가 비슷한 경우, 0에 가까운 값을 가진다. 거리의 최대 최솟값을 활용하는 Dunn Index와 달리 모든 개체에 대한 거리의 평균값을 통해 클러스터링을 평가한다.

본 논문에서는 원본 데이터셋과 합성 데이터셋 간의 실루엣 계수를 비교하여 데이터의 품질을 평가한다. 원본 데이터셋과 합성 데이터셋의 특성이 유사한 경우, 0에 가까운 값을 가진다. 원본 데이터셋과 합성 데이터셋의 특성이 다르거나 같은 데이터셋에서의 유사도가 떨어지는 경우, 실루엣 계수의 절댓값은 1에 가까운 값을 가진다. 따라서 원본 데이터셋과 유사한 품질의 합성 데이터셋은 실루엣 계수의 크기가 0에 가까운 값을 가진다.

## 6. SOH 추정 알고리즘

데이터 증강 알고리즘을 통해 생성된 데이터의 평가를 통해 신뢰성이 있는 배터리 데이터셋을 확보하여 SOH 추정 정확도를 비교하였다. SOH를 추정하기 위하여 대표적인 순환 신경망인 LSTM과 GRU를 활용하였다.

### 6.1. LSTM

LSTM(Long Short Term Memory)은 순환 신경망의 기울기 소실에 따른 장기 의존성 문제를 해결하기 위해 고안된 알고리즘으로, 구조는 [그림 8]과 같다.

Cell state를 활용하여 현재 시점 이전에 입력된 정보를 전달하고, 세 개의 게이트(Forget gate, Input gate, Output gate)를 통해 부여된 가중치를 통해 학습을 진행한다. Forget gate에서는 과거 데이터의 중요도를 판단하여 가중치를 부여한다. Input gate에서는 현재 시점의 입력과 현재 시점 이전의 hidden state 데이터를 얼마나 반영할 것인지 결정한다. Output gate는 hidden state로 전달할 때의 가중치를 결정한다.

Forget gate의 경우, 식(18)과 같이  $h_{t-1}$ 과  $x_t$ 의 값에 적절한 가중치를 부여하여 Cell state에 전달한다. 유의미한 데이터의 경우 1, 그렇지 않은 경우 0에 가까운 가중치를 부여한다.

$$f_t = \sigma(W_f^*[h_{t-1}, x_t] + b_f) \quad (18)$$

Input gate에서는  $i_t$ 와  $\tilde{C}_t$ 를 통해 cell state를 업데이트하여 새로운 Cell state  $C_t$ 를 생성하며 식(19), (20), (21)과 같다.

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \quad (19)$$

$$\tilde{C}_t = \tanh(W_c * [h_{t-1}, x_t] + b_c) \quad (20)$$

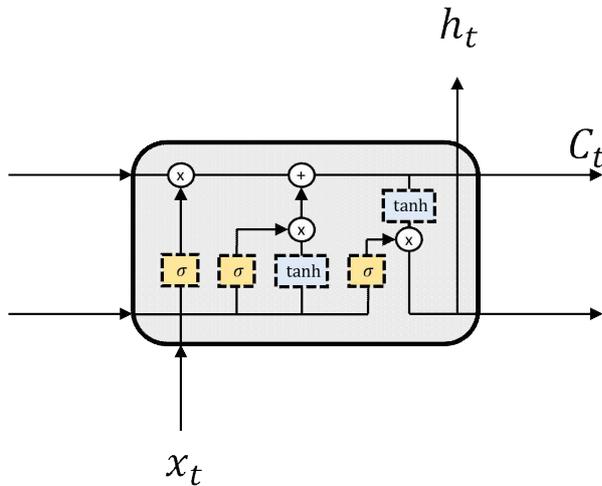
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (21)$$

Output Gate에서는  $C_t$ 의 출력을 결정하며, 식(22), (23)과 같다.

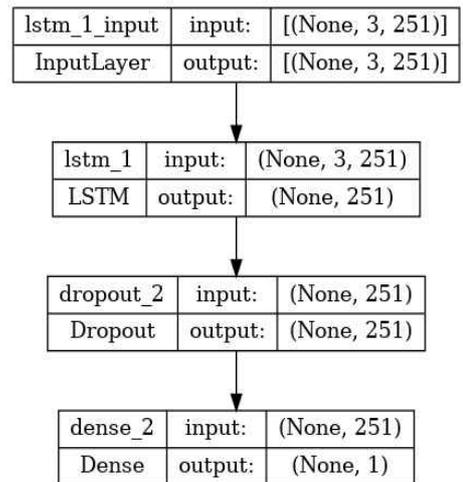
$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o) \quad (22)$$

$$h_t = o_t * \tanh(C_t) \quad (23)$$

SOH 추정을 위해 사용한 LSTM의 레이어는 [그림 9]과 같으며 Input layer와 LSTM layer, Dropout, Dense layer로 구성된다.



[그림 8] LSTM 구조



[그림 9] LSTM 레이어

## 6.2. GRU

GRU(Gated Recurrent Unit)는 LSTM을 간소화하고자 제안된 알고리즘이다. LSTM의 Cell state와 hidden state를 통합하고, Forget gate와 Input gate를 통합하여 [그림 10]과 같은 구조를 가진다. GRU는 Update gate와 Reset gate로 간단한 구조를 가지고, LSTM에 비하여 연산량이 작은 것이 특징이다.

Reset gate에서는 이전 시점의 hidden state  $h_{t-1}$ 와 현시점의 입력  $x_t$ 의 가중치를 결정하며, 식(24), (25)과 같다.

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r) \quad (24)$$

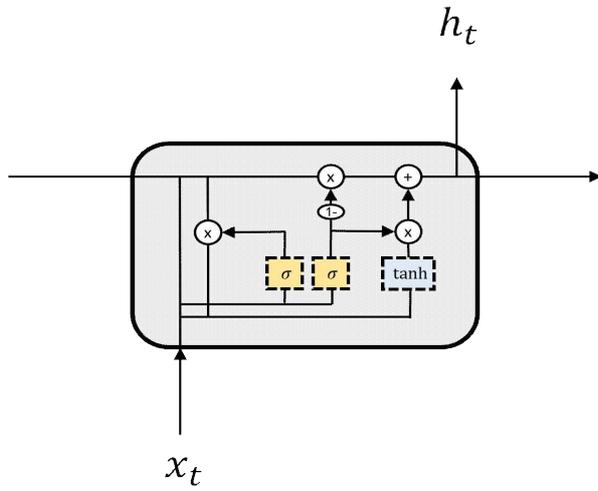
$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z) \quad (25)$$

Update gate는 과거와 현재 정보의 반영 비율을 판단하여 가중치를 부여하며, 식(26), (27)과 같다.

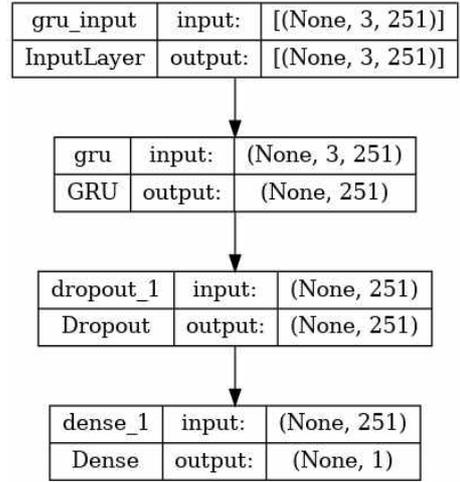
$$\tilde{h}_t = \tanh(W_c[r_t * h_{t-1}, x_t] + b_c) \quad (26)$$

$$h_t = z_t * h_{t-1} + (1 - z_t) * \tilde{h}_t \quad (27)$$

SOH 추정을 위해 사용한 GRU 구성은 [그림 11]과 같으며 Input layer와 GRU layer, Dropout, Dense layer로 구성된다.



[그림 10] GRU 구조



[그림 11] GRU 레이어

## 7. SOH 추정 정확도 평가

SOH 추정 정확도를 평가하기 위한 지표로 RMSE를 사용하였다. RMSE는 추정 정확도 평가의 대표적인 지표로, 실제 값과 예측값의 차이를 식(28), (29)과 같이 나타내며, 그 값이 작을수록 오차가 작아져 정확도가 높다. 여기서  $y_i$ 와  $\hat{y}_i$ 는 각각 실제 값과 예측값이며,  $n$ 은 데이터의 수를 나타낸다.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (28)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (29)$$

### III. 결과 및 분석

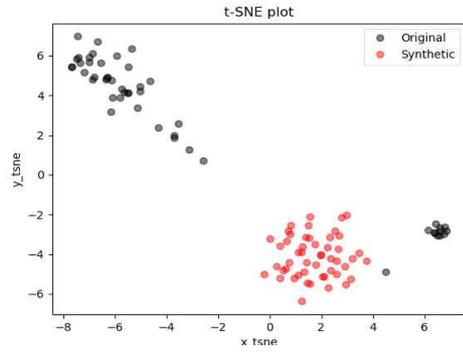
데이터 생성 및 평가를 통한 SOH 추정 정확도 향상 방법에 대하여 평가는 두 가지 측면으로 진행하였다. 첫 번째, TimeGAN을 통해 생성된 합성 데이터셋과 원본 데이터셋의 유사성을 평가하기 위하여 기존의 평가 방법인 학습 손실과 제안한 평가 방법인 상관계수 변화율, Dunn Index, 실루엣 계수를 비교하였다, 두 번째, 데이터셋 확보에 따른 SOH 추정 정확도 향상률을 비교하기 위하여 합성 배터리 데이터셋을 사용하여 LSTM, GRU 알고리즘을 통해 SOH를 추정하였다.

#### 1. 합성 데이터셋 평가

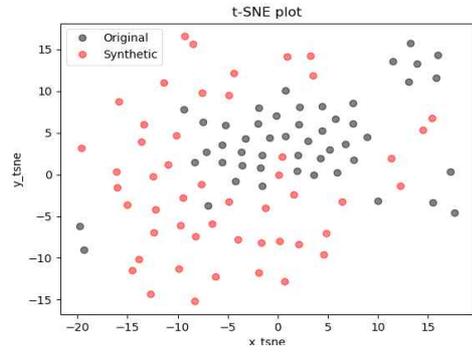
원본 데이터셋과 TimeGAN을 통해 생성된 합성 데이터셋을 t-SNE를 통해 이차원 그래프에 데이터를 나타내고, 원본 데이터셋과 합성 데이터셋의 선형회귀의 상관계수 변화율, Dunn Index, 실루엣 계수를 적용하였다.

차원이 축소된 시계열 데이터는 [그림 13-16]과 같으며, t-SNE에서의 검은색 점(black point)은 원본(original) 데이터를 의미하며, 빨간색 점(red point)은 합성(synthetic) 데이터를 의미한다.

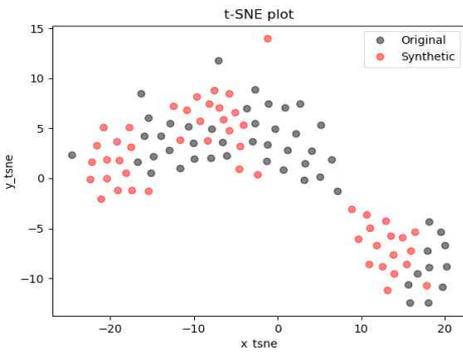
[그림 13-16]과 같은 표현된 차원 축소 그래프에서 기존의 정성적 평가 방법으로 데이터의 유사도 평가가 일부 가능하지만, 시각적으로 차이가 명확하지 않은 경우[그림 15](c),(d), 객관적으로 유사도를 확인하기 어렵다는 한계가 있다. 제안하는 정량적인 평가 방법은 합성 데이터셋의 유사도 객관적으로 평가할 수 있다. 일반적으로 데이터 증강 알고리즘의 학습 횟수가 증가함에 따라 학습 손실이 줄어들고 합성 데이터셋은 원본 데이터셋의 특성을 더 반영한다. 따라서, 제안한 평가 지표를 통해 원본 데이터셋과 학습 횟수(10, 100, 1000, 10000번)에 따른 합성 데이터셋의 유사도를 정량적으로 비교하여 평가한다.



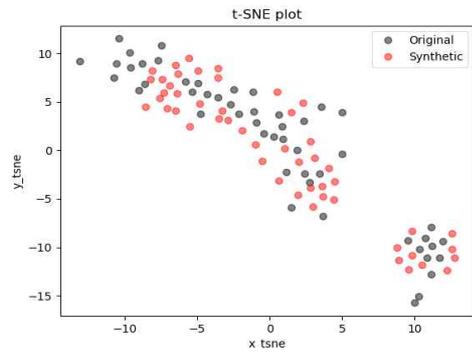
(a) Iteration 10



(b) Iteration 100

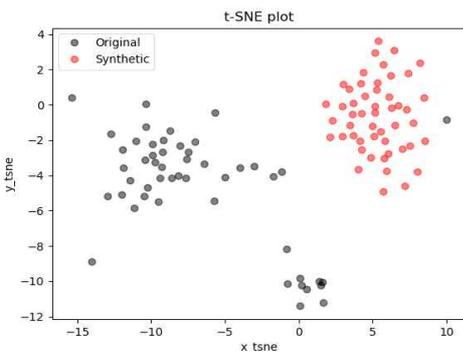


(c) Iteration 1000

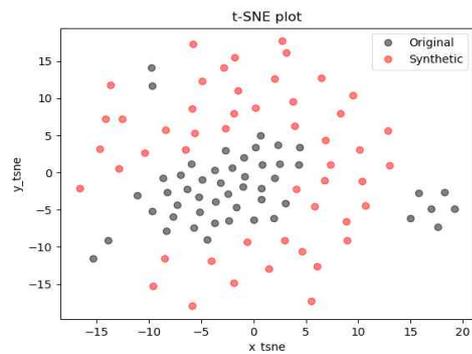


(d) Iteration 10000

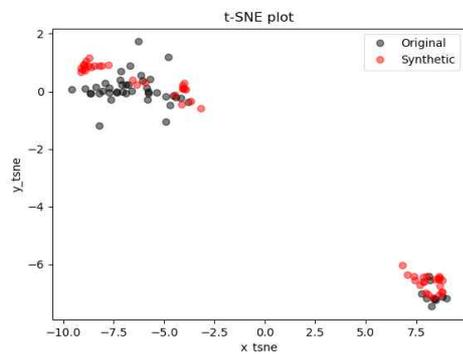
[그림 12] RW9의 학습 횟수에 따른 t-SNE 변화



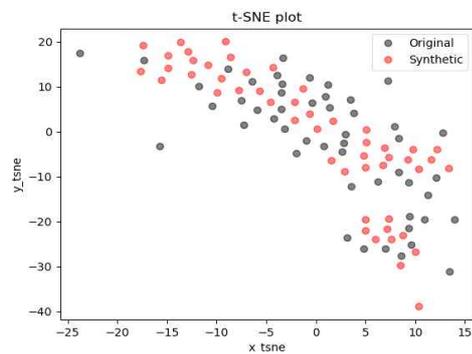
(a) Iteration 10



(b) Iteration 100

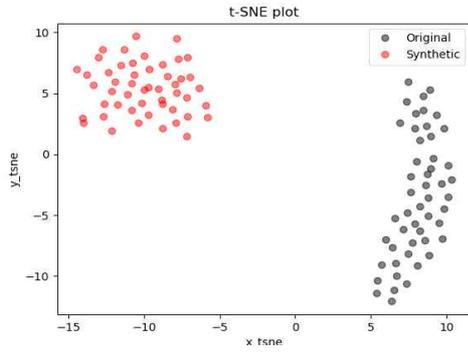


(c) Iteration 1000

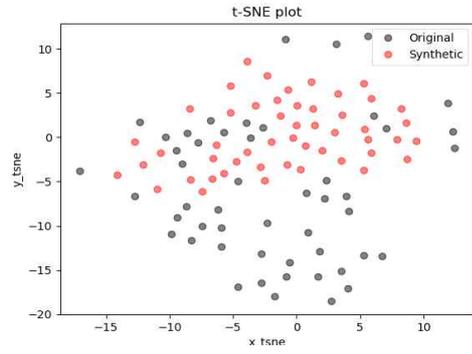


(d) Iteration 10000

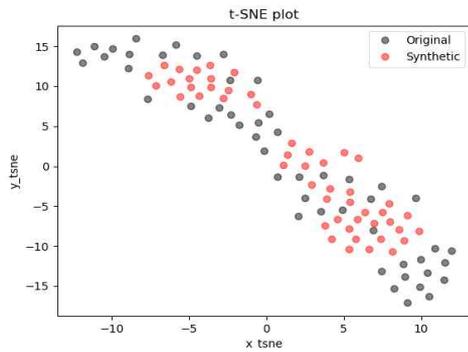
[그림 13] RW10의 학습 횟수에 따른 t-SNE 변화



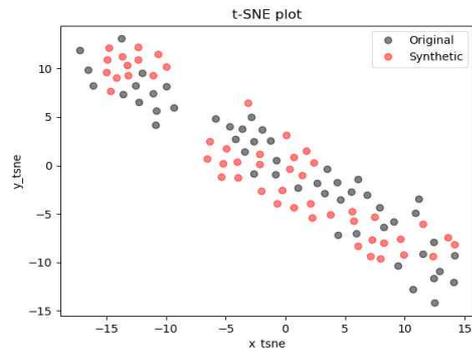
(a) Iteration 10



(b) Iteration 100

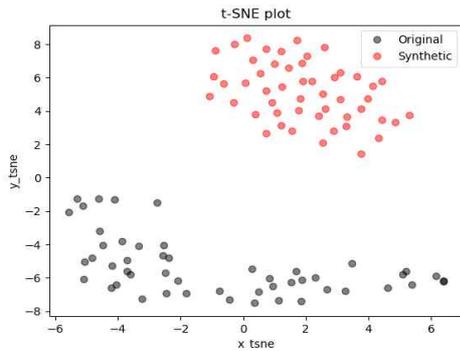


(c) Iteration 1000

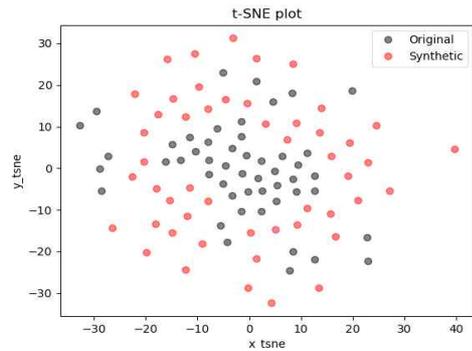


(d) Iteration 10000

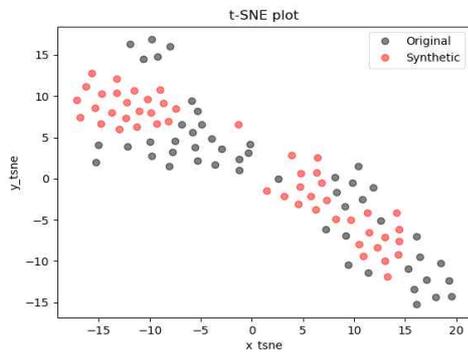
[그림 14] RW11의 학습 횟수에 따른 t-SNE 변화



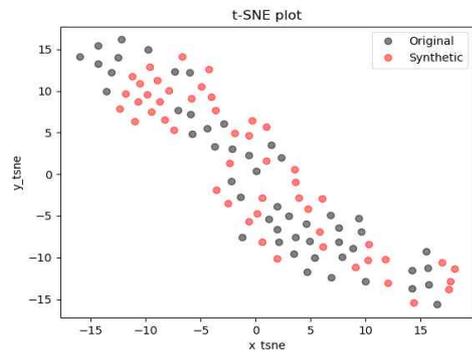
(a) Iteration 10



(b) Iteration 100



(c) Iteration 1000



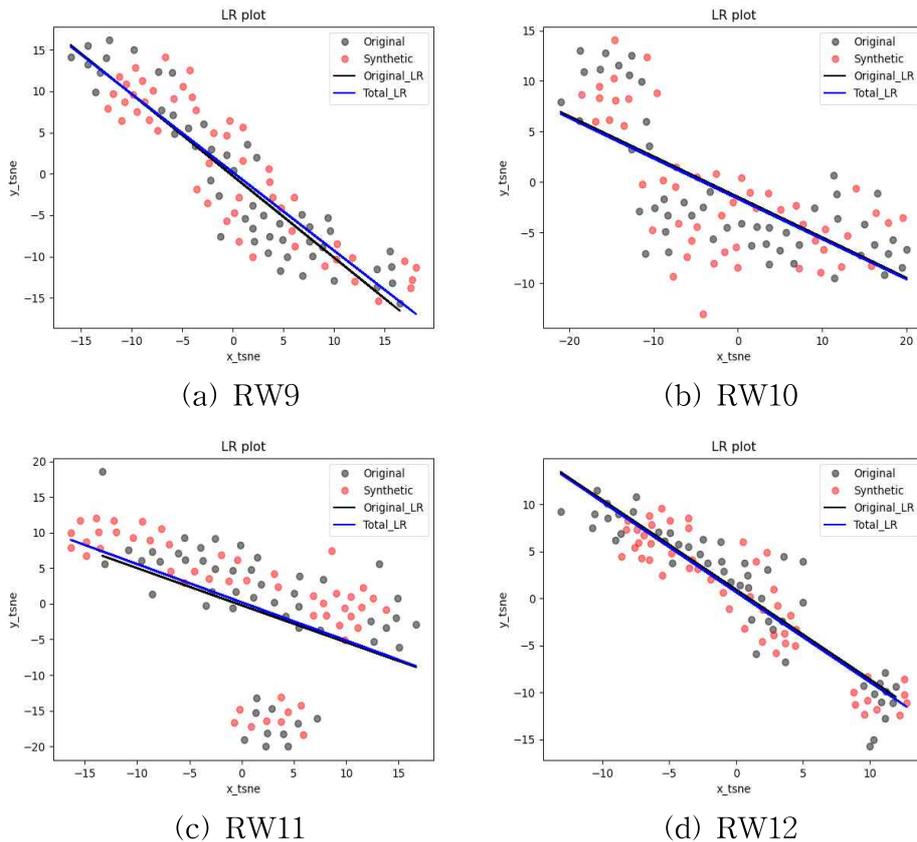
(d) Iteration 10000

[그림 15] RW12의 학습 횟수에 따른 t-SNE 변화

## 1.1. 선형회귀의 상관계수 변화율

데이터의 선형 관계를 비교한 상관계수의 변화율을 통해 원본 데이터셋과 합성 데이터셋 간의 유사성을 판단한다. 먼저, 결정 계수를 통하여 선형회귀식이 해당 데이터를 충분히 설명할 수 있는지 확인하고, 결정 계수가 충분한 경우 상관계수의 변화율을 통해 데이터의 유사성을 비교하였다. 결정 계수가 작은 경우에는 선형회귀식이 해당 데이터셋을 충분히 설명할 수 없어 해당 평가 방법을 적용하기 어려우며, 이러한 경우에는 Dunn Index와 실루엣 계수와 같은 클러스터링 지표를 활용하여야 한다.

[그림 16]에서 그래프에서 검은색 실선은 원본 데이터셋의 선형회귀식이며, 파란색 실선은 원본 데이터셋에 합성 데이터셋이 추가된 데이터셋의 선형회귀식이다.



[그림 16] 선형회귀 변화율(Iteration 10000회)

## 1.2. Dunn Index

Dunn Index는 그 값이 작을수록 원본 데이터셋의 품질을 유사하게 모사한 것으로 판단한다. 학습 횟수(Iteration)가 10, 100, 1000, 10000으로 증가할수록 Dunn Index 값은 감소하여 0에 가까운 값을 가지는 것을 확인할 수 있으며, 이는 원본 데이터셋과 합성 데이터셋이 유사하다는 것을 [표 2]와 같이 정량적으로 나타낸다. Dunn Index는 군 간의 유사도 비교를 위하여 극 값을 사용하므로, [그림 15] ©와 같이 동일한 데이터셋 내에서 데이터의 분포가 넓은 경우에도 Dunn Index가 작은 값이 나오는 특징이 있다.

## 1.3. 실루엣 계수

실루엣 계수는 크기가 0에 가까울수록 원본 데이터의 품질을 유사하게 모사한 것으로 판단하고, 1에 가까울수록 서로 데이터셋 간의 특성이 다르다고 판단한다. 제안된 지표를 통해 평가한 결과는 [표 2]와 같으며, 대부분의 셀 데이터셋에 대하여 학습 횟수가 증가하여 학습 손실이 감소하여 데이터가 유사해짐에 따라 실루엣 계수가 0에 더 가까워짐을 확인하였다. 실루엣 계수는 최대 최소 거리를 사용하는 Dunn Index와 달리 평균 거리를 사용하여 데이터의 유사도를 비교한다.

제안한 데이터셋 평가 방법을 통해 합성 배터리 데이터셋의 품질을 정량적으로 확인하였고, 그 결과는 [표 2]와 같다. 모든 셀에서 학습 횟수가 10000회인 합성 데이터셋이 모든 평가 방법에서 높은 유사도를 나타내었고, 각 셀에서 10000회 합성 데이터셋을 활용하여 SOH 추정 정확도 향상률을 비교하였다.

[표 2] 합성 배터리 데이터셋 평가 결과

Data	Evaluation Method	Iteration				
		10	100	1000	10000	
rw9	기존 평가	학습손실	0.4900	0.3900	0.1800	<b>0.1200</b>
		상관계수 변화율	Not used	Not used	0.2203	<b>0.0027</b>
	제안 평가	Dunn index	0.4899	0.3127	0.3030	<b>0.2540</b>
		실루엣 계수	0.3356	0.1053	0.0308	<b>0.0073</b>
rw10	기존 평가	학습손실	0.4950	0.2850	0.2550	<b>0.1450</b>
		상관계수 변화율	Not used	Not used	0.0325	<b>0.0291</b>
	제안 평가	Dunn index	0.5586	<b>0.2718</b>	0.3533	0.2719
		실루엣 계수	0.1064	0.0381	0.0889	<b>0.0089</b>
rw11	기존 평가	학습손실	0.4850	0.4700	0.2200	<b>0.1050</b>
		상관계수 변화율	Not used	Not used	0.0382	<b>0.0320</b>
	제안 평가	Dunn index	0.7071	0.3223	0.2722	<b>0.2639</b>
		실루엣 계수	0.5836	0.2172	0.0199	<b>0.0084</b>
rw12	기존 평가	학습손실	0.4900	0.28500	0.3200	<b>0.1540</b>
		상관계수 변화율	Not used	Not used	0.1444	<b>0.0414</b>
	제안 평가	Dunn index	0.9768	0.6826	<b>0.2900</b>	0.2938
		실루엣 계수	0.6826	0.0267	0.0295	<b>0.0156</b>

## 2. SOH 추정 정확도 평가

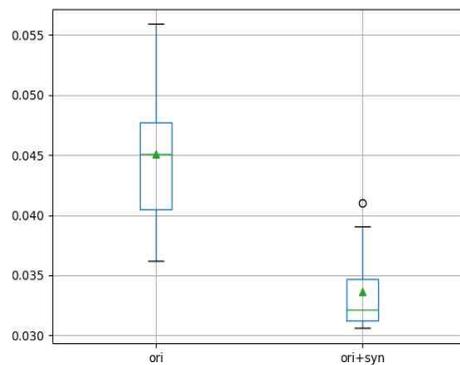
TimeGAN을 통해 생성된 합성 배터리 데이터셋이 SOH 추정 정확도 향상에 미치는 영향을 확인하기 위하여 각 셀마다 10000번 학습을 통해 높은 신뢰도를 가지는 합성 배터리 데이터셋을 활용하였으며, 추정 알고리즘은 LSTM과 GRU를 적용하였다. 테스트 결과의 신뢰성을 얻고자 각 셀에 대하여 10번의 반복 테스트를 진행하였다. 하나의 셀은 원본 데이터셋, 합성 데이터가 추가된 데이터셋 두 가지의 입력과, LSTM, GRU 두 가지의 추정 알고리즘 학습을 통해 정확도를 RMSE로 비교하였다. 테스트 10회 반복 후 RMSE의 평균값을 통하여 정확도 향상률을, RMSE의 IQR을 통하여 학습의 안정도를 비교하였다.

[그림 18,18]의 Boxplot에서 녹색 실선은 RMSE 결과의 중앙값을 의미하며, 녹색 삼각형은 평균값을 의미한다. 파란색 박스는 데이터의 IQR로 데이터의 25%~75% 사이의 범위를 의미하여, 이상치를 제외한 데이터의 흩어짐 정도를 나타낸다. IQR은 분산과 유사한 의미를 지니지만 특이 값의 영향을 거의 받지 않는다는 특징이 있다.

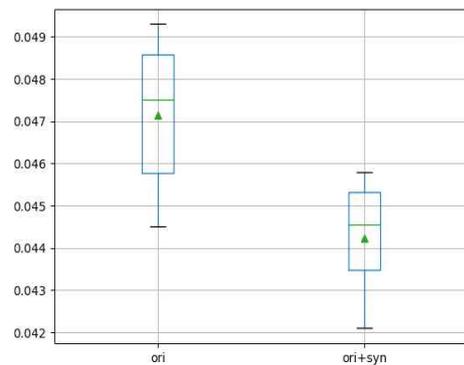
LSTM과 GRU를 이용하여 배터리 SOH 추정 정확도와 향상률에 대한 결과는 [그림 17], [그림 18]과 [표 3]와 같다.

## 2.1. LSTM

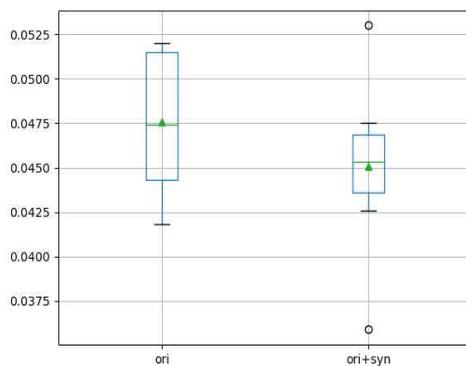
LSTM을 이용한 SOH 추정 결과는 [그림 17]과 같다. 모든 셀에서 RMSE의 평균값 및 중앙값은 감소하였고, 이는 TimeGAN을 통해 생성된 합성 데이터셋이 SOH 추정 정확도를 향상시킬 수 있음을 보여준다. 평균값을 기준으로 RW9 데이터에서 25.43%로 가장 높은 정확도 향상률을 보였고, RW11 데이터에서 5.29%로 가장 작은 정확도 향상률을 보였다. 그뿐만 아니라 RW9, RW10, RW11 셀의 추정 정확도에서는 IQR이 향상되었으며 이는 제안하는 방법이 SOH 추정 결과의 안정도에도 영향을 미치는 것을 확인하였다.



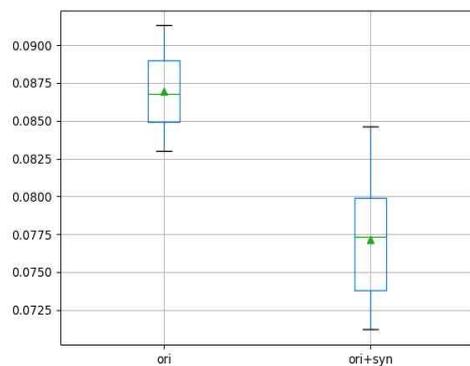
(a) RW9



(b) RW10



(c) RW11

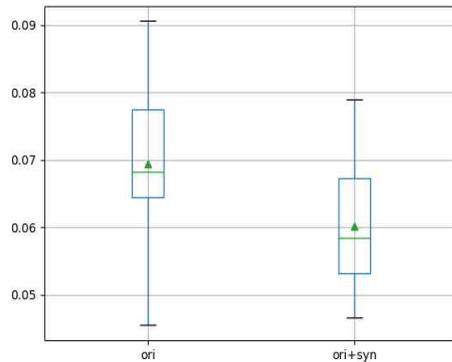


(d) RW12

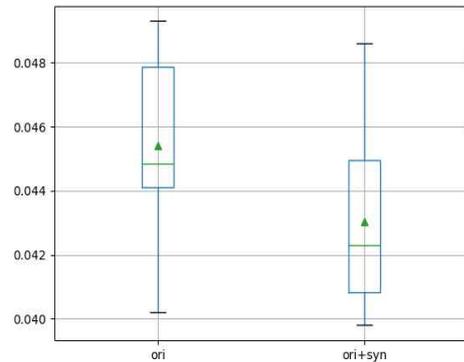
[그림 17] LSTM의 RMSE boxplot

## 2.2. GRU

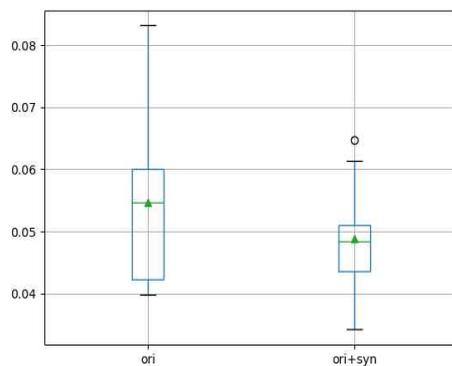
다음으로, GRU를 이용한 배터리 SOH 추정 결과는 [그림 18] 같다. 모든 셀에 대하여 RMSE의 평균값 및 중앙값은 감소하여 TimeGAN을 통해 생성된 합성 데이터셋이 SOH 추정 정확도 향상시킬 수 있음을 확인하였다. 특히, 평균값을 기준으로 RW12 데이터에 18.11%로 가장 높은 정확도 향상률을 보였으며, RW10 데이터에 대해서는 5.22%로 가장 작은 정확도 향상률을 보였다. 또한, RW9, RW10 셀에서는 IQR이 비슷하나 RW12, RW11 셀에서는 IQR이 향상됨을 보여주므로 학습의 안정도의 향상에 영향을 미치는 것을 확인하였다.



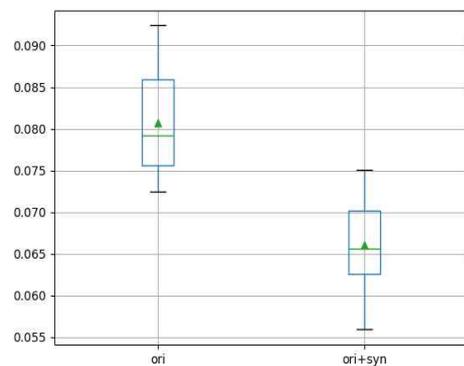
(a) RW9



(b) RW10



(d) RW12



(c) RW11

[그림 18] GRU의 RMSE boxplot

[표 3] SOH 추정 정확도 향상률 비교 결과

model	RMSE	RW9		RW10		RW11		RW12	
		ori	ori +syn	ori	ori +syn	ori	ori +syn	ori	ori +syn
LSTM	mean	0.0451	0.0336	0.0471	0.0442	0.0476	0.0451	0.0870	0.0772
	향상률	<b>25.41%</b>		6.17%		<b>5.29%</b>		11.28%	
	median	0.0451	0.0322	0.0475	0.0446	0.0474	0.0454	0.0868	0.0773
	향상률	<b>28.71%</b>		6.21%		<b>4.32%</b>		10.94%	
	IQR	0.0072	0.0035	0.0028	0.0019	0.0072	0.0032	0.004	0.0061
GRU	mean	0.0695	0.0602	0.0454	0.043	0.0547	0.0488	0.0807	0.0661
	향상률	13.32%		<b>5.22%</b>		10.68%		<b>18.11%</b>	
	median	0.0683	0.0585	0.0449	0.0423	0.0547	0.0484	0.0792	0.0656
	향상률	14.36%		<b>5.69%</b>		11.53%		<b>17.17%</b>	
	IQR	0.0131	0.0141	0.0038	0.0041	0.0177	0.0075	0.0104	0.0077

[표 3]과 같이 TimeGAN을 통해 데이터셋이 확보됨에 따라 학습의 정확도는 모든 셀과 알고리즘에 대하여 전반적으로 향상됨을 확인하였다. 추정 알고리즘 및 학습 환경에 따라 정확도는 유동적으로 변할 수 있기에 추정 알고리즘에 따라 가장 높거나 낮은 정확도를 보이는 셀은 달랐다. 이를 통해 데이터의 특성에 따라 학습에 적합한 추정 알고리즘은 다를 수 있지만 데이터 증강 알고리즘을 통한 데이터셋의 확보를 통해서 SOH 추정 정확도를 향상시킬 수 있음을 확인하였다. 또한, 모든 셀에서 유의미한 합성 데이터가 증가함에 따라 SOH 추정 정확도는 향상되었고, SOH 추정 결과의 안정도에도 영향을 미치는 것을 확인하였다.

## IV. 결론

본 논문에서는 TimeGAN을 통해 배터리 데이터셋의 생성 및 평가와 이를 활용한 SOH 추정 정확도 향상에 대하여 연구를 진행하였다.

배터리 데이터셋의 생성을 위하여 TimeGAN을 적용하였고, 높은 품질을 확보하기 위해 생성된 합성 데이터셋과 원본 데이터셋 간의 유사도를 확인하였다. 생성된 데이터셋을 t-SNE를 통해 2차원 그래프로 표현하고, 3가지 평가 지표(선형 회귀의 상관계수 변화율, 클러스터링의 Dunn Index와 실루엣 계수)를 이용하여 원본 데이터셋과의 유사성을 직접적-정량적으로 평가하는 방법을 제안하였고 제안한 평가 방법을 통해 원본 데이터셋과의 유사성을 확인하였다.

그 결과, 선형회귀식의 상관계수 변화율을 지표로 활용함에 있어 결정 계수가 충분한 값을 가져야 하는 특징이 있으며, 선형적인 데이터의 유사도 평가를 통하여 학습 횟수에 따른 증강 데이터의 특성을 확인하였다. Dunn Index는 학습 횟수가 증가함에 따라 Dunn Index가 감소하여 데이터의 유사성을 판단할 수 있다. 실루엣 계수는 일부 데이터가 군 내에서 멀리 분포하는 경우에도 정확한 평가가 가능함을 확인하였다. 따라서, 제안한 세 가지의 직접적인 평가 방법은 기존의 간접적인 평가 방법인 학습 손실과 유사한 결과를 얻음으로써 제안한 방법의 유용성을 확인하였다. 세 가지의 정량적인 지표 중 가장 높은 신뢰성을 갖는 지표는 실루엣 계수임을 확인하였다.

또한, 제안한 평가 방법을 통해 원본 데이터셋과의 유사성이 확인된 합성 데이터셋이 SOH 추정 알고리즘 정확도 향상에 미치는 영향을 비교하였다. 원본 데이터셋을 학습한 SOH 추정 정확도와 합성 데이터가 추가된 데이터셋을 학습한 SOH 추정 정확도의 평균 및 중앙값을 비교하여 정확도 향상률과 IQR을 이용하여 학습의 안정도에 미치는 영향을 비교하였다.

그 결과, 품질이 확인된 합성 배터리 데이터셋을 원본 데이터셋에 추가하여 SOH 추정을 하였을 때, 모든 셀에 대하여 배터리 SOH 추정 정확도가 향상됨을 확인할 수 있었다. 높은 품질의 합성 데이터를 추가하여 SOH 추정 알고리즘 학습에 활용함으로써 정확도 향상에 도움이 됨을 확인하였다. 또한, 합성 데이터 추가를 통해 SOH 학습의 안정도를 향상시킬 수 있었다. 많은 양의 합성 데이터를 통해 SOH 추정 알고리즘을 학습에 활용함으로써, 과적합 또는 과소적합의 문제가 해결되어 학습의 결과가 안정적으로 나타남을 확인하였다.

## 참고문헌

- [1] Topan, Paris Ali, et al. "State of Charge (SOC) and State of Health (SOH) estimation on lithium polymer battery via Kalman filter." 2016 2nd International Conference on Science and Technology-Computer (ICST). IEEE, 2016.
- [2] Liu, Datong, et al. "An on-line state of health estimation of lithium-ion battery using unscented particle filter." *Ieee Access* 6 (2018): 40990-41001.
- [3] Dai, Haifeng, et al. "Online cell SOC estimation of Li-ion battery packs using a dual time-scale Kalman filtering for EV applications." *Applied Energy* 95 (2012): 227-237.
- [4] Cui, Shengmin, and Inwhee Joe. "A dynamic spatial-temporal attention-based GRU model with healthy features for state-of-health estimation of lithium-ion batteries." *Ieee Access* 9 (2021): 27374-27388.
- [5] Gu, Xinyu, et al. "A novel state-of-health estimation for the lithium-ion battery using a convolutional neural network and transformer model." *Energy* 262 (2023): 125501.
- [6] Li, Penghua, et al. "State-of-health estimation and remaining useful life prediction for the lithium-ion battery based on a variant long short term memory neural network." *Journal of power sources* 459 (2020): 228069.
- [7] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." *arXiv preprint arXiv:1511.06434* (2015).
- [8] Arjovsky, Martin, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks." *International conference on machine learning*. PMLR, 2017.

- [9] Mao, Xudong, et al. "Least squares generative adversarial networks." Proceedings of the IEEE international conference on computer vision. 2017.
- [10] Mogren, Olof. "C-RNN-GAN: Continuous recurrent neural networks with adversarial training." arXiv preprint arXiv:1611.09904 (2016).
- [11] Esteban, Cristóbal, Stephanie L. Hyland, and Gunnar Rätsch. "Real-valued (medical) time series generation with recurrent conditional gans." arXiv preprint arXiv:1706.02633 (2017).
- [12] Yoon, Jinsung, Daniel Jarrett, and Mihaela Van der Schaar. "Time-series generative adversarial networks." Advances in neural information processing systems 32 (2019).
- [13] Shangguan, Anqi, et al. "Train wheel degradation generation and prediction based on the time series generation adversarial network." Reliability Engineering & System Safety 229 (2023): 108816.
- [14] Zhang, Yunfei, et al. "Data augmentation for improving heating load prediction of heating substation based on TimeGAN." Energy 260 (2022): 124919.
- [15] Li, Jun, Yongbao Liu, and Qijie Li. "Generative adversarial network and transfer-learning-based fault detection for rotating machinery with imbalanced data condition." Measurement Science and Technology 33.4 (2022): 045103.
- [16] Salimans, Tim, et al. "Improved techniques for training gans." Advances in neural information processing systems 29 (2016).
- [17] Guan, Shuyue, and Murray Loew. "Evaluation of generative adversarial network performance based on direct analysis of generated images." 2019

- IEEE Applied Imagery Pattern Recognition Workshop (AIPR). IEEE, 2019.
- [18] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9.11 (2008).
- [19] Bryant, Fred B., and Paul R. Yarnold. "Principal-components analysis and exploratory and confirmatory factor analysis." (1995).
- [20] Dunn, Joseph C. "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters." (1973): 32-57.
- [21] Rousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics* 20 (1987): 53-65.
- [22] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.
- [23] Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." *arXiv preprint arXiv:1412.3555* (2014).
- [24] Goodfellow, I. et al. "Generative adversarial nets". *Adv. Neural Inf. Process. Syst.* 27, 2672 - 2680 (2014).
- [25] Hinton, Geoffrey E., and Sam Roweis. "Stochastic neighbor embedding." *Advances in neural information processing systems* 15 (2002).
- [26] Mikolov, Tomas, et al. "Recurrent neural network based language model." *Interspeech*. Vol. 2. No. 3. 2010.
- [27] Sui, Xin, et al. "A review of non-probabilistic machine learning-based state of health estimation techniques for Lithium-ion battery." *Applied Energy* 300 (2021): 117346.
- [28] Yoon, Seunghyun, et al. "Efficient transfer learning schemes for personalized language modeling using recurrent neural network." *arXiv*

preprint arXiv:1701.03578 (2017).

- [29] Alauthman, M. O. H. A. M. M. A. D. "Botnet spam e-mail detection using deep recurrent neural network." *Int. J* 8.5 (2020): 1979–1986.
- [30] Bole, Brian, Chetan S. Kulkarni, and Matthew Daigle. "Adaptation of an electrochemistry-based li-ion battery model to account for deterioration observed under randomized use." *Annual Conference of the PHM Society*. Vol. 6. No. 1. 2014.

## ABSTRACT

# A Study on Generation and Evaluation of Time-Series Data to Improve Lithium-Ion Battery State-of-Health Estimation

SU JIN SEOL

Graduate school of Electrical, Electronic and computer  
Engineering  
University of Ulsan  
Ulsan, Korea

With the recent tightening of regulations on fossil fuel vehicles, the xEV market, which utilizes lithium-ion batteries as an environmentally friendly energy source, is expanding. Ensuring the safety and reliability of batteries is essential in their use, and therefore, estimating the State Of Health (SOH) of batteries is crucial. There are two methods for estimating SOH: model-based and data-driven. Among them, the data-driven approach requires a large quantity and high-quality battery dataset. One method to acquire datasets when actual measurements are challenging is data augmentation algorithms. The quality evaluation of time-series datasets generated through data augmentation algorithms mainly relies on indirect assessments, which have limitations in producing variable results depending on the learning environment. Therefore, this paper proposes a method to quantitatively evaluate the quality of generated battery datasets for improving the accuracy of SOH estimation. The proposed evaluation method utilizes the rate of change in the correlation coefficient of linear regression, as well as clustering metrics such as the Dunn Index and the

silhouette coefficient. Furthermore, using the proposed evaluation method, the paper compares the accuracy of SOH estimation between LSTM and GRU models by leveraging a large quantity of high-quality datasets with ensured reliability.

Key Words : Lithium-ion battery, GAN(Generative Adversarial Networks), TimeGAN(Time-series Generative Adversarial Networks), dataset, synthetic data, linear regression, Dunn Index, silhouette coefficient SOH estimation, deep learning, RNN(Recurrent Neural Network), LSTM(Long Short Term Memory), GRU(Gate Recurrent Unit)