Journal of The Korea Convergence Society
Vol. 13. No. 3, pp. 23-31, 2022

e-ISSN 2713-6353
DOI : https://doi.org/10.15207/JKCS.2022.13.03.023

PCA 및 변수 중요도를 활용한 냉동컨테이너 고장 탐지 방법론 비교 연구

이승현[1], 박성호[1], 이승재[1], 이희원[1], 유성열[2], 이강배[3*]
[1]동아대학교 경영정보학과 학생, [2]부산카톨릭대학교 경영정보학과 교수, [3]동아대학교 경영정보학과 교수

A Comparative Study on the Methodology of Failure Detection of Reefer Containers Using PCA and Feature Importance

Seunghyun Lee[1], Sungho Park[1], Seungjae Lee[1], Huiwon Lee[1], Sungyeol Yu[2], Kangbae Lee[3*]
[1]Student, Dept. of MIS, Donga University
[2]Professor, Dept. of MIS, Catholic University of Pusan
[3]Professor, Dept. of MIS, Donga University

요 약 본 연구는 H해운사에서 제공받은 Starcool사의 실제 냉동 컨테이너 운영데이터를 분석하였다. H사의 현장 전문가와 인터뷰를 통해 4가지 고장 알람 중 Critical 및 Fatal Alarm만 고장으로 정의하였고, 냉동 컨테이너 특성상 모든 변수를 사용하는 것은 비용측면에서 비효율을 초래하는 것을 확인하였다. 이에 본 연구는 특성 중요도 및 PCA 기법을 통한 냉동 컨테이너 고장 탐지 방법을 제시한다. 모델의 성능 향상을 위해 XGBoost, LGBoost 등과 같은 트리계열 모델을 통해 변수 중요도(Feature Importance)를 기반으로 변수 선택(Feature selcetion)을 하고 선택되지 않은 변수는 PCA를 사용하여 전체 변수의 차원을 축소시켜 각 모델별로 지도학습을 수행한다. 부스팅 기반의 XGBoost, LGBoost 기법은 본 연구에서 제안하는 모델의 결과가 62개의 모든 변수를 사용한 지도 학습의 결과보다 재현율(Recall)이 각각 0.36, 0.39씩 향상되는 되는 결과를 보였다.

주제어 : 냉동 컨테이너, 고장 탐지, 머신러닝, PCA, 변수 선택, 변수중요도

Abstract This study analyzed the actual frozen container operation data of Starcool provided by H Shipping. Through interviews with H's field experts, only Critical and Fatal Alarms among the four failure alarms were defined as failures, and it was confirmed that using all variables due to the nature of frozen containers resulted in cost inefficiency. Therefore, this study proposes a method for detecting failure of frozen containers through characteristic importance and PCA techniques. To improve the performance of the model, we select variables based on feature importance through tree series models such as XGBoost and LGBoost, and use PCA to reduce the dimension of the entire variables for each model. The boosting-based XGBoost and LGBoost techniques showed that the results of the model proposed in this study improved the reproduction rate by 0.36 and 0.39 respectively compared to the results of supervised learning using all 62 variables.

Key Words : Refrigerated container, Fault detection, Machine learning, PCA, Feature selection, Feature importance

*This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-02091,Development and Commercialization of IoT-based refrigerated container real-time monitoring and BigData / AI-based failure predictive service platform to strengthen competitiveness of shipping & logistics company)

*Corresponding Author : Kangbae Lee(kanglee@donga.ac.kr)

Received December 10, 2021 Revised January 3, 2022
Accepted March 20, 2022 Published March 28, 2022

1. 서론

컨테이너 운송시장에서 냉동 컨테이너는 빠르게 성장하는 시장 중 하나다. 인구증가, 소득수준 향상 등과 같은 다양한 요인으로 인해 신선한 식품, 의약품 식물 등과 같은 제품에 대한 수요는 꾸준히 증가할 것으로 예상된다[1]. 세계 해상 냉동 운송량은 점차적으로 증가하고 있으며 2021년에는 1억 3000만 톤에 이를 것으로 예상된다. 냉동 컨테이너로 운송되고 있는 냉장・냉동 화물은 다양한 환경 조건 및 해양 요인의 영향으로 인해 냉동 컨테이너에 고장이 발생하여 화물이 손상될 수 있다[2]. 냉동 컨테이너로 운송되는 화물 중 일부분인 육류, 수산물, 과일, 채소류 등은 온도나 습도 등에 민감하여 부패되기 쉬운 성질을 갖고 있기 때문에 화물이 손상되지 않도록 각별한 주의가 필요하다.

냉동 컨테이너의 고장에는 경미한 것부터 심각한 것까지 다양한 고장 유형이 있지만 대부분의 고장은 냉동장치 고장에서 일어난다. 예를 들어, 압축기에서 고장이 발생하게 되면 공급되는 공기의 온도가 설정 값보다 낮아지게 되고, 화물의 온도가 상승하여 화물에 손상이 생기게 된다[3]. 컨테이너가 고장나게 된다면 불필요한 시간낭비를 포함하여 손상화물의 검사비용, 손상화물의 폐기비용 등 불필요한 경제적 비용을 증가시킨다. 따라서 냉동 컨테이너의 고장을 감지하는 것이 중요하다[4].

냉동시스템 고장에 관한 다수의 연구들이 진행되고 있다. 하지만 냉동 컨테이너 고장 탐지에 관한 연구는 많이 존재하지 않는다. 또한 냉동시스템에 대한 연구는 실제 데이터가 아닌 시뮬레이션 데이터를 활용한 연구가 다수이다[13-19]. 실험용 데이터는 실제로는 테스트하기 어려운 다양한 문제들을 모든 특성값을 사용하여 분석할 수 있지만, 데이터의 특성이 많아지면 차원의 저주에 빠질 수 있는데, 차원의 저주란 분석하고자 하는 데이터의 특성이 많으면 성능에 오히려 부정적인 영향을 끼치는 문제이다. 차원의 저주 문제는 PCA(Principal Component Analysis)나 LDA(Linear Discriminant Analysis)등과 같은 고차원의 데이터를 저차원의 데이터로 변환시켜주는 알고리즘을 통해 해결할 수 있다[5].

본 연구에서는 실제 냉동 컨테이너 운영 데이터를 활용하여 지도학습을 기반으로 한 고장 탐지 방법을 제시한다. 냉동 컨테이너 전문가와의 인터뷰를 통해 Critical Alarm과 Fatal Alarm을 고장으로 정의하였다. XGBoost, Decision Tree, Random Forest, LGBoost와 같은 트리계열 모델로 변수중요도(Feature importance)를 기반으로 Feature selection을 수행했다. 그리고 주요변수로 선택되지 않은 변수는 차원의 저주 문제를 해결하기 위해 PCA를 통해 차원을 저차원으로 축소시켰다. 각 모델별로 Feature selection과 PCA를 통한 데이터 조합을 이용하여 지도학습을 이용한 고장 탐지 방법을 개발하였다. 본 연구에서 제안한 방법은 XGBoost, LGBoost와 Random Forest 모델로 62개의 모든 변수를 사용하여 지도학습을 수행한 것 보다 성능이 향상되었다.

본 연구의 구성은 다음과 같다. 1장은 서론, 2장에서는 냉동시스템 및 냉동 컨테이너 고장에 관한 선행연구를 설명한다. 3장에서는 분석에 사용한 데이터와 데이터 처리방법에 대해 설명하고, 본 연구에서 제안하는 분석 모델을 소개하였다. 4장에서 분석 결과를 종합하여 비교하고, 5장은 결론으로 마무리 한다.

2. 관련 연구

2.1 기계 학습

인공지능의 한 분야인 기계학습은 컴퓨터 파워 부족 등의 현실적인 이유로 침체기를 겪다가 컴퓨터 파워 향상과 머신러닝 방법론의 발전으로 다양한 분야에서 연구되고 있다.

데이터 기반의 분석을 진행하는 기계학습은 학습에 사용할 데이터에 정답이 포함되어 있는 지도 학습과 정답이 포함되어 있지 않은 비지도 학습으로 구분할 수 있다[6,7]. 본 연구에서는 기계 학습 중 지도 학습 기법인 XGBoost(Extreme Gradient Boosting; XGBoost), DT(Decision Tree), RF(Random Forest), LGBoost (Light Gradient Boosting; LGBoost)와 함께 비지도 학습 PCA를 사용하였다. XGBoost 모델은 2016년 Chen과 Cuestri이 제안한 트리 임베딩을 기반으로 부스팅 기법 학습 모델이다. 분류 정확도가 낮은 여러 트리 모델을 결합하여 보다 정확한 모델을 구축하여 정확도를 향상시키는 학습 방법으로 학습과 분류가 빠른 장점이 있다[8]. DT(Decision Tree)는 데이터들의 속성을 기반으로 분할 기준을 판단하고, 분할 기준에 따라 트리 형태로 모델링하는 방법이다. 구현이 간단하고 분석 결과를 시각적으로 이해하기 쉽기 때문에 해석이 용이하여 일반적으로 많이 사용되는 분류 방법이다[9]. RF

(Random Forest)는 2001년 Breiman에 의해 제안된 결정 트리와 배깅 알고리즘(Bagging Algorithm)을 기반으로 하는 앙상블 머신 러닝 방법이다. 회귀 및 분류 문제에 적용할 수 있을뿐만 아니라 정확도가 높고, 변수 선택에서도 좋은 성능을 보인다[10]. LGBoost모델은 부스팅 기반의 모형으로 기울기 기반 단측 표본추출법인 GOSS(Gradient-base One-Side Sampling)와 배타적 변수 묶음인 EFB(Exclusive Feature Bundling)를 적용해 트리를 분할함으로써 속도가 빠르고 큰 사이즈의 데이터를 다룰 수 있으며 정확도도 우수하다[11]. PCA는 1901년 Karl Pearson이 도입하였으며 고차원의 데이터를 저차원의 데이터로 차원을 축소시키는 방법이다. 중복이나 노이즈가 많은 데이터에서 관련성이 높은 정보를 추출할 때 사용하는 방법이다. 원래 데이터의 차원을 축소하면서 그 데이터의 특징을 최대한 손실 없이 살리는 장점이 있다[12]. 본 연구에서 실제 냉동 컨테이너의 특성은 62가지이며 그 중 고장 탐지에 주요한 변수를 뽑고, 주요하지 않은 변수들은 PCA로 차원을 축소하여 데이터의 특징을 최대한 살려주었다.

2.2 선행 연구

Zhimin Du는 HVAC system에서 반환수온, 공급공기 온도, 냉각수 밸브 등 공기 조화기에서 발생하는 이상을 감지하기 위해 시뮬레이션 데이터로 이중 인공신경망(Dual ANN)알고리즘을 사용하여 오경보, 경보 누락 및 감지 시간의 감지 효율성을 크게 향상시켰다[13].

Tran, D. A. T.는 원심 냉각기 시스템의 결함 감지 및 진단을 하기 위해 ASHRAE RP-1043 시뮬레이션 데이터를 사용하여 차동 진화 방식에 기반한 LSSVR 모형을 제안하였고, 고장별로 냉매 과충전 2.5%, 비응축성 가스 0.5%, 냉매 누출 8.15%, 응축기 2.26%씩 고장진단율을 상승시켰다[14].

Dan Li는 냉각기 결함 감지 및 진단 문제를 다중 클래스 분류 문제로 공식화 하는 2단계 LDA 기반 데이터 기반 전략을 제시하였는데, ASHRAE RP-1043 시뮬레이션 데이터를 사용하여 7개의 일반적인 결함을 성공적으로 감지 및 진단하였다[15].

Ronggeng Huang은 연관 분류 알고리즘을 기반으로 한 원심 냉각기 고장 진단 모델을 제안하였는데, ASHRAE RP-1043 시뮬레이션 데이터를 사용하여 평균 86.3%의 Accuracy로 7가지의 결함을 식별해냈다[16].

Li G는 원심 냉각기 결함 진단을 위해 PCA-R-SVDD 모델을 제안하였고, ASHRAE-RP-1043 시뮬레이션 데이터로 평가한 결과로 PCA, SVDD 및 PCA-PC-SCDD 3가지 기존 방법에 비해 결함 감지에서 개선된 결과를 도출했다[17].

Yan, R은 HVAC시스템의 AHU(Air Handling Unit)의 오류 감지 및 진단을 하기 위해 의사 결정 트리기반 모델 제안하였고, ASHRAE-RP-1312 시뮬레이션 데이터를 사용하여 F1-Score값이 평균 0.97인 우수한 진단 성능을 달성하였다[18].

Lee, K는 ASHRAE-RP-1043 시뮬레이션 데이터를 사용하여 심각도 수준을 고려한 SVM과 LGBM 방법론을 기반으로 고장 유형 진단 알고리즘을 제안하였는데, 다른 고장 유형 진단 성능을 유지하면서 냉매 과충전과 냉매부족 고향 진단 성능을 향상 시키는 결과를 도출하였다[19].

냉동 컨테이너 고장과 관련된 선행연구들이 많이 존재하지 않아 냉동 컨테이너와 유사한 시스템 구조인 냉동기 시스템 관련 선행연구를 참고하였다. 선행연구에서 사용한 RP-1043 데이터는 총 65개의 변수로 이루어져 있고, 7가지의 고장에 따라 4단계의 심각도를 포함하고 있다. 전체 데이터를 사용할 경우, 고장 원인 진단의 성능이 낮음을 확인할 수 있었다. 선행연구에서는 해당 문제를 해결하기 위해 PCA 기법을 통한 차원 축소나 고장에 영향을 주는 변수를 선택하는 방법을 사용하였다. 본 연구에서는 전체 특성을 사용하지 않고 고장에 영향을 주는 주요변수를 추출하였고, 주요하지 않은 변수는 차원의 저주 문제를 해결하기 위해 PCA를 통해 차원을 축소하여 고장 진단을 하였다.

3. 데이터 및 연구 프로세스

3.1 데이터

본 연구에서는 H해운사에서 제공받은 Starcool사의 실제 냉동 컨테이너 운영 데이터를 활용하였다. 총 111개의 컨테이너에서 온도, 전압, 증발기[1] 등의 다중 센서로부터 1시간 간격으로 수집된 데이터이다. 고장이 발생했을 경우 고장이 발생한 시점에 고장 데이터가 생성생성된 고장 데이터는 고장에 대한 정보를 제외한 모든 특성값이 기록되지 않고 수집이 된다. 수집된 된다. 데이터는

1) 2019년 Starcool 매뉴얼북 참조

총 299,576개이며 그 중에 고장은 31,850개가 존재한다. 수집된 데이터는 컨테이너의 정보를 포함한 총 67개의 특성을 갖고 있으며 아래의 Table 1과 같다.

Table 1. Data Features

NO	Feature	Full Name	Unit
1.	Time	-	-
2.	Container	Container Number	-
3.	Type	Collected data types	-
4.	Tset	Temperature Set Point	℃
...	...	...	...
64	Hevap Avg	Evaporator Heater Average Status	%
65	Mpump On Avg	Motor pump on Average	%
66	Battery pack voltage	Battery pack voltage	V
67	Mpump service runtime	Motor pump service runtime	Hour

3.2 고장 데이터

위의 3.1절에서 수집한 전체 데이터는 총 299,576개이다. 그중 고장 데이터는 31,850개로 약 10.6%로 정의하였다. 고장 데이터에는 고장의 시작을 알리는 Active Alarm과 고장의 끝을 알리는 Inactive Alarm이 존재하는데 본 연구에서는 Active Alarm만 고장으로 취급하고, Inactive Alarm은 정상으로 판단하고 분석하였다.

현업 담당자와 인터뷰를 통해 고장 데이터에 대한 2가지 사실을 확인하였다. 첫 번째로 고장에는 심각도가 존재하며 그 중 Critical Alarm과 Fatal Alarm이 발생했을 때 냉동 컨테이너에 심각한 문제를 발생시킬 수 있다는 것을 확인하였다. 이에 본 연구에서는 Critical Alarm과 Fatal Alarm만 고장으로 정의하였으며, Log Alarm과 Warning Alarm은 정상으로 판단하고 분석하였다.

두 번째로 심각한 고장이 발생할 때는 이전 데이터에도 고장 징후가 나타날 수 있다는 것을 확인하였다. 그래서 고장이 발생하기 직전의 데이터에도 고장이라는 Label을 달아주어 부족한 고장의 수를 늘려주었다. 아래의 Table 2는 전처리 후 분석에 사용한 총 데이터의 수를 나타낸다.

Table 2. The number of data used for analysis

Data Type	Before	After
Normal Data	267,726	295,853
Abnormal Data	31,850	3,723
Total Data	299,576	299,576

3.3 데이터 전처리

67개의 변수 중 컨테이너의 정보를 포함한 Time, Container Number, Type, Event ID, State의 5개의 변수는 분석에 불필요하기 때문에 제외시키고, 62개의 변수를 분석에 사용하였다. 수집된 데이터는 정상 데이터와 고장 데이터 모두 Null값이 존재하기 때문에 본 연구에서는 각 변수 특성에 맞게 결측치를 아래의 Table 3와 같이 보간하였다.

Table 3. Interpolate Method

Feature	method
Setting Value	Use the previous value
Numerical Value	Linear interpolation
Categorical Value	Use a new category & Use the previous value

Tset, RH_set, CO2 set, O2 set과 같은 설정값 변수들은 사전에 설정한 값들이기 때문에 앞의 값을 사용하여 보간하였고, 수치형 변수들은 선형 보간법을 사용하여 보간하였다. 선형 보간법(Linear interpolation)이란 끝 점의 값이 주어졌을 때 그 사이에 위치한 값을 추정하기 위하여 직선 거리에 따라 선형적으로 계산하여 결측치를 보간하는 방법이다. 데이터와 결측치 사이의 값들은 증가 혹은 감소를 할 수 있기 때문에 해당 방법을 사용하였다. 보간 후 변수들의 크기와 범위가 다르기 때문에 표준화 변환을 하였다. 그리고 범주형 변수는 이전의 값을 사용한 보간법과 카테고리에는 'CA', 'CA DEF', 'DEF' 등 다양한 값들이 존재하기 때문에 카테고리 값들이 각각의 의미가 있다고 생각하여 새로운 카테고리 값을 만들어 레이블 인코딩을 진행하였다.

3.4 변수 선택(Feature selection)

변수 선택(Feature selection)은 데이터를 효율적으로 설명하면서 관련없는 변수의 영향력을 줄여 종속변수의 예측 결과 성능을 향상시키는 입력 변수의 하위 집합을 선택하는 것이다[20]. 본 연구에서는 4개의 트리계열

모델로부터 각각의 변수 중요도(Feature importance)를 통하여 모델별로 어떠한 변수가 고장을 판별하는데 주된 영향을 주는지 모델별로 상위 15개의 변수를 선택하여 분석하였다.

3.5 최종 분석 프로세스

본 연구는 실제 냉동 컨테이너 운영 데이터를 활용한 지도학습 기반의 고장 탐지 방법을 제시한다. 본 연구의 최종 분석 프로세스는 Fig 1과 같이 진행된다. 분석에 사용한 데이터는 위의 3.2에서 설명한 것처럼 고장 데이터를 전처리 후에 3.3에서 설명한 보간법을 데이터 전체에 적용하였다. 그리고 Train과 Test를 8 대 2의 비율로 나눠준 뒤 표준화를 적용하였다. 그 후 트리계열 분류 알고리즘인 XGBoost, DT, RF, LGBoost를 통해 각 모델별 고장 탐지 주요변수를 변수 중요도(Feature importance)를 기반으로 뽑아낸다. 현장 전문가와의 인터뷰를 통해 주요 변수만 사용하는 것이 비용측면에서 효율적인 것을 확인하였다. 따라서 본 논문에서는 특성 중요도와 PCA 기법을 활용하여 데이터의 차원을 축소하였다. 그 후 모델별로 고장 탐지 주요변수와 PCA변수를 합쳐서 트리계열 분류 알고리즘을 사용하여 고장 탐지 분류모델을 개발한다. 그리고 지도학습, PCA, 주요변수로만 지도학습을 한 결과와 비교 및 분석한다.

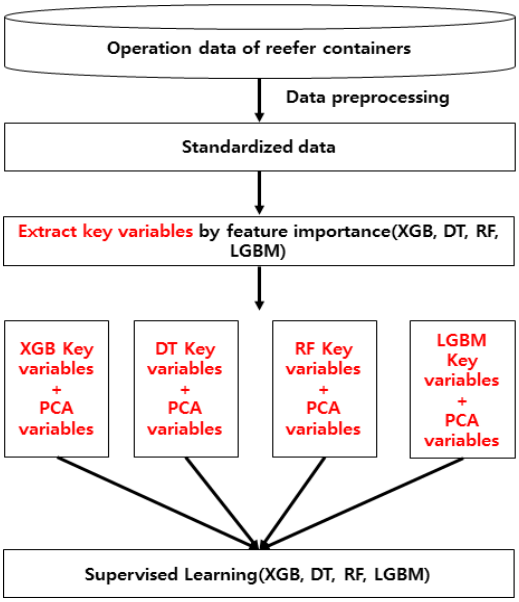

Fig. 1. Research model

3.6 성능지표

위의 3.3의 Table 3에서 본 연구에서 사용한 데이터는 클래스 불균형(Class imbalance)문제를 확인할 수 있다. 불균형 데이터를 사용할 경우 정확도(accuracy)만 가지고는 성능을 판단하기 어렵다. 그 이유는 소수 클래스를 완전히 무시하고 다수 클래스의 성능만 가지고도 좋은 결과를 얻을 수 있기 때문이다. 그렇기 때문에 본 연구에서는 정확도(Accuracy) 뿐 만 아니라 불균형 데이터의 성능 평가지표로 쓰이는 정밀도(Precision), 재현율(Recall), F1-Score를 추가적인 성능지표로 사용하였다. 정확도(Accuracy)는 전체 데이터 중 맞게 예측한 비율을 나타내고, 정밀도(Precision)는 Positive(양성)으로 예측한 것 중 실제 Positive(양성)인 것의 비율을 나타낸다. 재현율(Recall)은 실제 Positive(양성)인 것 중에 Positive(양성)으로 예측한 것의 비율을 나타낸다. 재현율(Recall)은 실제 고장을 고장으로 예측한 비율을 나타내는 성능지표이기 때문에 본 연구에서 가장 중요한 성능지표이다. F1-Score는 정밀도(Precision)와 재현율(Recall)의 조화평균을 의미한다.

Table 4. Performance indicators

Index	Formula
Accuracy	(TP+TN) / (TP+TN+FP+FN)
Precision	TP / (TP+FP)
Recall	TP / (TP+FN)
F1-Score	2*Precision*Recall/ Precision+Recall

4. 분석 결과

4.1 모델별 지도학습 결과

아래의 Table 5는 62개의 변수로 모델별로 지도학습만 수행한 결과이다. 분석에 사용한 데이터는 불균형 데이터이기 때문에 Accuracy는 모든 모델에서 높게 나왔다. DT와 RF 모델은 각각 0.75, 0.70의 재현율(Recall)을 기록하였지만, XGB와 LGBM은 상대적으로 낮은 0.26, 0.33의 재현율(Recall)을 기록하였다.

Table 5. Supervised learning results

Model	Precision	Recall	F1-Score	Accuracy
XGB	0.91	0.26	0.40	0.9904
DT	0.73	0.75	0.74	0.9934
RF	0.82	0.70	0.76	0.9944
LGBM	0.44	0.33	0.38	0.9864

4.2 모델별 PCA 결과

아래의 Table 6는 PCA로 차원을 축소하고 지도학습을 수행한 결과이다. 분산설명력을 90%로 하였을 때, 62개의 변수에서 21개의 변수로 축소된다. 지도학습 결과에 비해 XGBoost와 LGBoost의 재현율(Recall)이 크게 향상되어 F1-Score가 상승했다. 하지만 DT와 RF의 재현율(Recall)은 낮아진 것을 확인할 수 있다.

Table 6. PCA results

Model	Precision	Recall	F1-Score	Accuracy
XGB	0.85	0.52	0.65	0.9928
DT	0.66	0.66	0.66	0.9915
RF	0.84	0.63	0.72	0.9939
LGBM	0.76	0.63	0.69	0.9929

4.3 각 모델별 주요변수 지도학습 결과

4.3.1 각 모델별 상위 15개 주요변수

각 모델별 Feature importance를 통해 상위 15개의 주요변수를 뽑은 결과 Table 7과 같다. 상위 15개의 주요변수 추출방법은 Han, J의 선행연구를 참고하였다[21]. 또한 현장 전문가와의 인터뷰를 통해 주요 변수만 사용하는 것이 비용측면에서 효율적인 것을 확인하였다. 각 모델별로 고장 탐지에 주요한 변수가 다른 것을 확인할 수 있다.

Table 7. Key variables for each model

No.	XGB	DT	RF	LGBM
1	Tusda1	Energy	Energy	Pdis
2	Energy	O2	Mcond_avg	Umax
3	Psuc	Vhg_avg	Tcargo	RH
4	Tset	Tset	Tusda1	Net_freq
5	SH_act	Umax	Vhg_avg	Tcargo
6	Vhg_avg	Tpump	Umax	Battery_pack_voltage
7	Net_freq	Net_freq	TO	Tret
8	Umax	Tcargo	Tusda3	Tint
9	Vexp_avg	Tsup	Tpump	Tfc
10	TC	Tevap	Tsup	TC
11	Hevap	Psuc	Tret	Tamb
12	I2	Tret	Tusda2	SH_ref
13	Tret	Mcond_avg	Tsuc	Psuc
14	Tret_instant	SH_ref	Tret_instant	Mevap_avg
15	Tcargo	TC	Battery_pack_voltage	Tpump

4.3.2 각 모델별 주요변수 지도학습 결과

아래의 Table 8은 주요변수만으로 4개의 데이터 셋을 만들어 각 데이터를 기계학습 모델로 분석한 결과이다. XGBoost와 LGBoost는 DT의 주요변수를 기반으로 분석 하였을 때, 지도학습과 PCA만 적용한 결과보다 재현율(Recall)이 각각 0.62, 0.71로 상승하는 좋은 성능을 보였다. 하지만 DT와 RF 모델은 XGB의 주요변수를 기반으로 분석하였을 때, 지도학습만 수행한 결과와 재현율(Recall)이에서 큰 차이를 내지 않았다.

Table 8. key variables results

Model	Precision	Recall	F1-Score	Accuracy	importance
XGB	0.83	0.62	0.71	0.9936	DT
DT	0.72	0.74	0.73	0.9931	XGB
RF	0.82	0.71	0.76	0.9944	XGB
LGBM	0.76	0.71	0.74	0.9936	DT

4.4 각 모델별 주요변수 + PCA 변수 지도학습 결과

위의 4.3.1의 Table 7에서 설명한 각 모델별 상위 15개 주요변수를 제외한 나머지 47개는 분산설명력 90%로 PCA를 수행하여 차원을 축소시키고 주요변수와 함께 지도학습을 수행하였다. 각 모델별로 주요변수 15개를 포함하여 XGBoost는 33개, DT는 32개, RF는 32개, LGBoost는 32개의 특성으로 축소되었다.

Table 9. key variables + PCA variables results

Model	Precision	Recall	F1-Score	Accuracy	Feature importance
XGB	0.84	0.62	0.71	0.9937	XGB(33)
DT	0.71	0.74	0.72	0.9931	RF(32)
RF	0.83	0.72	0.77	0.9947	RF(32)
LGBM	0.78	0.72	0.75	0.9936	RF(32)

위의 Table 9는 4개의 데이터로 각각 지도학습을 수행하였을 때, 가장 높은 재현율(Recall)을 나타낸 데이터와 결과를 보여준다. 각 모델별로 XGBoost는 XGB 기반의 주요변수와 PCA 변수를 사용하였을 때 0.62, DT는 RF 기반의 주요변수와 PCA 변수를 사용하였을 때 0.74, RF는 RF 기반의 주요변수와 PCA 변수를 사용하였을 때0.72, LGBoost는 RF 기반의 주요변수와 PCA 변수를 사용하였을 때 0.72의 가장 높은 재현율(Recall)을 확인하였다.

4.5 최종 결과

4.5.1 최종 분석 결과

본 연구에서 연구한 4가지 방법들의 성능을 비교하였을 때 XGBoost와 LGBoost 모델은 본 연구에서 제안하는 모델을 이용하였을 때 가장 많은 성능 향상을 보였다. 그리고 XGB는 XGB기반의 데이터를 사용하였을 때 가장 좋은 성능을 보였고, LGBM은 RF의 기반의 데이터를 사용하였을 때 가장 좋은 성능을 보였다. DT와 RF의 경우 성능에 큰 차이는 없지만, PCA를 이용하여 차원을 축소한 후 지도학습을 하면 재현율(Recall)이 감소하는 것을 확인하였다. 아래의 Table 10은 정밀도(Precision), 재현율(Recall), F1-Score를 사용한 최종 결과를 정리한 표이다.

Table 10. Final results

Model	Method	Precision	Recall	F1-Score	Feature importance
XGB	S.L	0.91	0.26	0.40	-
	PCA	0.85	0.52	0.65	-
	Key	0.83	0.62	0.71	DT
	Key+PCA	0.84	0.62	0.71	XGB
DT	S.L	0.73	0.75	0.74	-
	PCA	0.66	0.66	0.66	-
	Key	0.72	0.74	0.73	XGB
	Key+PCA	0.71	0.74	0.72	RF
RF	S.L	0.82	0.70	0.76	-
	PCA	0.84	0.63	0.72	-
	Key	0.82	0.71	0.76	XGB
	Key+PCA	0.83	0.72	0.77	RF
LGBM	S.L	0.44	0.33	0.38	-
	PCA	0.76	0.63	0.69	-
	Key	0.76	0.71	0.74	DT
	Key+PCA	0.78	0.72	0.75	RF

4.5.2 성능 비교

재현율(Recall)은 실제 고장인 것 중에 고장을 예측한 비율로서 고장 탐지에서는 가장 중요한 성능지표이다. 위의 Table 11은 본 연구에서 제시한 방법 중 XGBoost와 LGBoost모델에 사용한 방법에 대한 재현율(Recall)을 비교한 표이다. 부스팅 기반의 XGBoost와 LGBoost 모델의 재현율(Recall)이 크게 향상되었다. 부스팅은 여러 개의 의사결정 트리를 사용하지만 단순히 결과를 평균 내는 것이 아니라 결과를 보고 오답에 대한 가중치를 부여한다. 그리고 가중치가 적용된 오답에 대해서는 정답이 될 수 있도록 결과를 만들고 해당 결과에 대한 다른 오답을 찾아 다시 똑같은 작업을 반복적으로 진행한다. 이에 본 연구에서 제안하는 모델 중 부스팅 기반의 XGBoost와 LGBoost의 성능이 향상된 결과를 보였다.

Table 11. Model performance comparison

Model	S.L	The proposed method
XGB	0.26	0.62
LGBM	0.33	0.72

5. 결 론

본 연구에서는 H해운사에서 제공받은 Starcool사의 실제 냉동 컨테이너 운영 데이터를 대상으로 머신러닝 기반의 고장 탐지 주요변수를 활용한 고장 탐지 모델 설계를 제안하였다. 고장을 사전에 진단하는 것은 모든 산업에서 중요한 부분 중 하나이고, 고장 발생에 대한 사전 조치를 취할 수 있게 한다. 본 연구에서는 트리계열 분류 기법인 XGBoost, DT, RF, LGBoost를 이용하여 고장 탐지에 주요한 변수를 뽑고, PCA와 결합하여 모델의 성능을 비교하였다. 모든 변수를 사용하여 분석하는 것은 차원의 저주로 인한 성능 저하와 컴퓨터 메모리 측면에서 비효율적이기 때문에 변수 선택과 PCA기법은 모델의 성능을 높이기 위해 필요한 기술 중 하나이다. 본 연구에서 제안하는 방법 중 부스팅 기반의 XGBoost와 LGBoost 모델은 XGB 기반의 주요변수+PCA 데이터와 RF 기반의 주요변수+PCA 데이터를 사용하여 분석하였을 때 실제 고장을 고장이라고 예측할 수 있는 Recall(재현율)이 62개의 모든 변수를 사용하여 지도학습을 했을 때 보다 각각 0.36, 0.39씩 상승한 것을 확인하였다. 본 연구에서는 모든 특성을 사용하지 않고, 데이터에 포함된 모든 특성에서 주요 특성을 추출하여 고장을 탐지하는 방법을 제시하였다. 본 연구를 통해 모든 데이터를 사용하기 어려운 다른 산업 분야에서 고장 진단 분석 연구에 활용할 수 있을 것으로 기대된다.

REFERENCES

[1] B. Castelein, H. Geerlings & R. Van Duin. (2020). The reefer container market and academic research: A review study. *Journal of Cleaner*

Production, 256, 120654.
DOI : 10.1016/j.jclepro.2020.120654

[2] A. Kan, T. Wang, D. Zhu & D. Cao. (2021). The characteristics of cargo temperature rising in reefer container under refrigeration-failure condition. *International Journal of Refrigeration, 123*, 1-8.
DOI : 10.1016/j.ijrefrig.2020.12.007

[3] N. Hoffmann, R. Stahlbock & S. Voß. (2020). A decision model on the repair and maintenance of shipping containers. *Journal of Shipping and Trade, 5(1)*, 1-21.
DOI : 10.1186/s41072-020-00070-2

[4] S. K. Park, Y. G. Park & Y. R. Shin. (2012). A Study on the Improvement of Damage to Reefer Container Cargo. *Journal of Navigation and Port Research, 36(10)*, 803-810.
DOI : 10.5394/KINPR.2012.36.10.803

[5] G. S. Gim, H. S. Shon, K. H. Ryu & S. H. Lee. (2013). Performance of PCA Algorithm for Multivariate Data Analysis. *In Proceedings of the Korea Information Processing Society Conference* (pp. 1264-1266). Korea Information Processing Society.
DOI : 10.3745/PKIPS.Y2017M11A.1264

[6] K. B. Lee, S. H. Park, S. H. Sung & D. M. Park. (2019). A Study on the Prediction of CNC Tool Wear Using Machine Learning Technique. *Journal of the Korea Convergence Society, 10(11)*, 15-21.
DOI : 10.15207/JKCS.2019.10.11.015

[7] Y. D. Yun, Y. W. Yang, H. S. Ji & H. S. Lim. (2017). Development of Smart Senior Classification Model based on Activity Profile Using Machine Learning Method. *Journal of Cleaner Production, 8(1)*, 25-34.
DOI : 10.15207/JKCS.2017.8.1.025

[8] T. Chen & C. Guestrin. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785-794.
DOI : 10.1145/2939672.2939785

[9] A. Priyam, G. R. Abhijeeta, A. Rathee & S. Srivastava. (2013). Comparative analysis of decision tree classification algorithms. *International Journal of current engineering and technology, 3(2)*, 334-337.

[10] R. Tian, F. Chen & S. Dong. (2021). Compound Fault Diagnosis of Stator Interturn Short Circuit and Air Gap Eccentricity Based on Random Forest and XGBoost. *Mathematical Problems in Engineering, 2021*.
DOI : 10.1155/2021/2149048

[11] M. J. Oh, E. S. Choi, K. W. Roh, J. S. Kim & W. S. Jo. (2021). A Study on the Design of Supervised and Unsupervised Learning Models for Fault and Anomaly Detection in Manufacturing Facilities. *The Korean Journal of BigData, 6(1)*, 23-35.
DOI : 10.5394/KINPR.2012.36.10.803

[12] S. Cateni, M. Vannucci, M. Vannocci & V. Colla. (2012). Variable selection and feature extraction through artificial intelligence techniques. *Multivariate Analysis in Management, Engineering and the Science, 103-118*.
DOI : 10.5772/53862

[13] Z. Du, B. Fan, X. Jin & J. Chi. (2014). Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis. *Building and Environment, 73*, 1-11.
DOI : 10.1016/j.buildenv.2013.11.021

[14] D. A. T. Tran, Y. Chen, H. L. Ao & H. N. T. Cam. (2016). An enhanced chiller FDD strategy based on the combination of the LSSVR-DE model and EWMA control charts. *International Journal of Refrigeration, 72*, 81-96.
DOI : 10.1016/j.ijrefrig.2016.07.024

[15] D. Li, G. Hu & C. J. Spanos. (2016). A data-driven strategy for detection and diagnosis of building chiller faults using linear discriminant analysis. *Energy and Buildings, 128*, 519-529.
DOI : 10.1016/j.enbuild.2016.07.014

[16] R, Huang et al. (2018). An effective fault diagnosis method for centrifugal chillers using associative classification. *Applied Thermal Engineering, 136*, 633-642.
DOI : 10.1016/j.applthermaleng.2018.03.041

[17] G. LI et al. (2016). An improved fault detection method for incipient centrifugal chiller faults using the PCA-R-SVDD algorithm. *Energy and Buildings, 116*, 104-113.
DOI : 10.1016/j.enbuild.2015.12.045

[18] R. Yan, Z. Ma, Y. Zhao & G. Kokogiannakis. (2016). A decision tree based data-driven diagnostic strategy for air handling units. *Energy and Buildings, 133*, 37-45.
DOI : 10.1016/j.enbuild.2016.09.039

[19] K. B. Lee, S. H. Park, H. W. Lee, S. J. Lee & S. H. Lee. (2021). A study on the 3-step classification algorithm for the diagnosis and classification of refrigeration system failures and their types. *Journal of the Korea Convergence Society, 12(8)*, 31-37.

DOI : 10.15207/JKCS.2021.12.8.031

[20] G. Chandrashekar& F. Sahin. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40(1)*, 16-28.
DOI : 10.1016/j.compeleceng.2013.11.024

[21] H. J. Han, D. K. Ko & H. C. Choe. (2019). Prediction and Analyzing Factor Affection Financial Stress of Household Using Machine Learning: Application of XGBoost. *Journal of Consumer Studies, 30(2), 21-43.*
DOI : 10.35736/JCS.30.2.2

이 승 현(Seung-hyun Lee) [학생회원]

· 2021년 2월 : 동아대학교 경영정보학과(학사)
· 2021년 3월 ~ 현재 : 동아대학교 경영정보학과 석사과정
· 관심분야 : 머신러닝, 딥러닝
· E-Mail : hyunwow1263@naver.com

박 성 호(Sungho Park) [정회원]

· 2017년 2월 : 동아대학교(학사)
· 2019년 2월 : 동아대학교(석사)
· 2019년 2월 ~ 현재 : 동아대학교 경영정보학과 박사과정
· 관심분야 : 머신러닝, 딥러닝
· E-Mail : psh2975@donga.ac.kr

이 승 재(Seung-jae Lee) [학생회원]

· 2021년 2월 : 동아대학교 경영정보학과(학사)
· 2021년 3월 ~ 현재 : 동아대학교 경영정보학과 석사과정
· 관심분야 : 머신러닝, 딥러닝
· E-Mail: sj2170497@donga.ac.kr

이 희 원(Hui-Won Lee) [학생회원]

· 2021년 2월 : 동아대학교 경영정보학과(학사)
· 2021년 3월 ~ 현재 : 동아대학교 경영정보학과 석사과정
· 관심분야 : 머신러닝, 딥러닝
· E-Mail : slcw@naver.com

유 성 열(Sungyeol Yu) [정회원]

· 1997년 ~ 2000년 : LGCNS
· 2002년 3월 ~ 현재 : 부산가톨릭대학교 경영정보학과 교수
· 관심분야 : 인공지능, 빅데이터, 기계학습
· E-Mail : syyu@cup.ac.kr

이 강 배(Kangbae Lee) [정회원]

· 1991년 3월 ~ 1995년 8월 : 한국과학기술원 산업공학(박사)
· 2001년 3월 ~ 2004년 8월 : 부산가톨릭대학교 경영정보학과 교수
· 2008년 2월 ~ 현재 : 동아대학교 경영정보학과 교수
· 관심분야 : 머신러닝, 딥러닝
· E-Mail : kanglee@daonga.ac.kr