

Study on Failure Classification of Missile Seekers Using Inspection Data from Production and Manufacturing Phases

Ye-Eun Jeong* · Kihyun Kim* · Seong-Mok Kim* · Youn-Ho Lee** · Ji-Won Kim** · Hwa-Young Yong** ·
Jae-Woo Jung** · Jung-Won Park*** · Yong Soo Kim***[†]

*Department of Industrial & Systems Engineering, Kyonggi University Graduate School

**PGM Integrated Product Support R&D, LIG Nex1

***Department of Industrial & Systems Engineering, Kyonggi University

생산 및 제조 단계의 검사 데이터를 이용한 유도탄 탐색기의 고장 분류 연구

정예은* · 김기현* · 김성목* · 이연호** · 김지원** · 용화영** · 정재우** · 박정원*** · 김용수***[†]

*경기대학교 일반대학원 산업시스템공학과

**LIG넥스원 PGM IPS연구소

***경기대학교 산업시스템공학과

This study introduces a novel approach for identifying potential failure risks in missile manufacturing by leveraging Quality Inspection Management (QIM) data to address the challenges presented by a dataset comprising 666 variables and data imbalances. The utilization of the SMOTE for data augmentation and Lasso Regression for dimensionality reduction, followed by the application of a Random Forest model, results in a 99.40% accuracy rate in classifying missiles with a high likelihood of failure. Such measures enable the preemptive identification of missiles at a heightened risk of failure, thereby mitigating the risk of field failures and enhancing missile life. The integration of Lasso Regression and Random Forest is employed to pinpoint critical variables and test items that significantly impact failure, with a particular emphasis on variables related to performance and connection resistance. Moreover, the research highlights the potential for broadening the scope of data-driven decision-making within quality control systems, including the refinement of maintenance strategies and the adjustment of control limits for essential test items.

Keywords : Missile, Failure Classification, Data Augmentation, Dimensionality Reduction, Machine Learning

1. 서론

현대 군사 작전에서 핵심적인 역할을 수행하며 중요성

이 지속적으로 증대되고 있는 유도탄은 탐색기, 관성측정기, 구동장치 및 추진기관 등 다양한 전자, 기계 화공품목들이 집약적으로 통합된 시스템이다[7]. 이 중 탐색기는 목표물을 탐지 및 추적하는 핵심적인 역할을 수행한다. 유도탄의 핵심부품인 탐색기의 고장은 가장 빈번하며 치명적이므로 본 연구에서는 유도탄 내 탐색기 고장을 고장이라 정의하였다. 유도탄의 생산 및 제조 단계의 데이터를

Received 11 March 2024; Finally Revised 3 April 2024;

Accepted 11 April 2024

[†] Corresponding Author : kimys@kgu.ac.kr

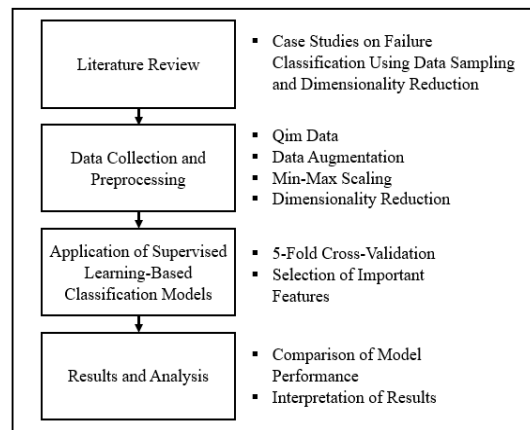
이용하여 탐색기 고장을 식별함으로써 유도탄의 요구되는 수명 이상을 만족할 수 있도록 하고자 한다.

유도탄은 장기 저장 후 1회 운용되는 일회용 장치(One-Shot Device)로 전술적 및 전략적 목적에 따라 다양하게 활용된다. 이러한 유도탄은 여러 구성품이 복잡한 형태로 탑재되어 있으며 이 중 핵심부품인 탐색기의 고장 또는 오작동은 심각한 결과를 야기할 수 있으므로 유도탄의 탐색기 고장을 예방하는 연구의 중요성이 강조된다. 유도탄의 긴 수명주기 동안 요구 수명 이상을 유지하고 고장을 예방하기 위한 연구가 활발히 진행되고 있으나, 대다수의 연구는 유도탄의 상태 진단 및 모니터링을 위한 절차 및 방안 제시에 관한 연구[14]나 고장 데이터 기반 신뢰성 예측에 관한 연구[20]에 집중되어 있다. 이에 반해 다양한 산업 분야에서는 생산 및 제조 단계의 데이터를 활용하여 품질 측면에서 잠재적 고장을 사전에 식별하는 연구가 활발히 진행되고 있다. 유도탄 분야에서는 생산 이후 필드에서 수집된 데이터를 기반으로 이상 징후를 분석하거나 결함을 탐지하는 연구가 존재하나, 생산 시점에서의 데이터를 활용한 연구는 부족한 실정이다.

10년 이상의 장기 수명을 갖는 OO 유도탄은 공정 단계에서 관리도를 통해 시험 SPEC인 공정관리 한계를 만족한 후 군에 배치되며, 배치 0년 후 매년 랜덤 샘플링 검사를 수행한다. 그러나 관리 한계를 만족하여 장기적인 수명을 가질 것이라 판단되어 군에 배치되었으나 군배치 후 탐색기 고장이 발생한 유도탄들이 존재한다. 탐색기 고장이 발생한 유도탄들에 대하여 생산 단계에서의 데이터 특징이 존재하는지 또는 생산 단계 데이터로 고장이 발생할 유도탄의 분류가 가능한지 검토해 볼 필요가 있다. 또한 일반적으로 유도탄의 경우, 배치 이후 획득되는 데이터양이 절대적으로 부족하므로 생산 시 데이터의 적극적 활용이 요구된다. 따라서 본 연구에서는 특정 기업 내 품질관리시스템에서 생산 및 제조단계의 제품 또는 공정 품질의 평가 및 관리에 사용하는 Quality Inspection Management(QIM) 데이터를 활용하여 향후 탐색기 고장 발생 가능성이 높은 유도탄을 생산 및 제조단계에서 머신러닝 모델을 활용하여 분류하고자 한다. 유도탄의 고장 데이터는 정상 데이터에 비해 상대적으로 적으며, QIM 데이터는 변수가 666개로 데이터의 차원이 매우 크다. 이러한 데이터 불균형은 학습과정에서 부정적인 영향을 미칠 수 있으며, 관측치보다 변수의 수가 많아지는 고차원 상황에서는 차원의 저주(Curse of Dimensionality)가 발생할 수 있다[6]. 차원의 저주는 차원이 증가함에 따라 특성 공간 차원이 함께 증가하여 함수의 최적화 문제를 복잡하게 만드는 현상이다. 본 연구에서는 이러한 문제점을 해결하기 위해 소수의 고장 데이터와 다차원 QIM 데이터를 사용하여 데이터 증강 및 차원축소 기법을 통해 데이터를 전처리한 후 머신러닝 기반의 분류모델로 생산 단계에서

고장이 발생할 가능성이 있는 유도탄을 식별하고 모델의 고장 결정기준이 되는 주요 변수를 추출하여 추후 검사항목 관리에 용이성을 높이고 유도탄의 품질을 향상시키고자 한다.

본 연구는 유도탄 품질 향상을 목표로 데이터 불균형 문제를 해결하고 차원 축소 기법을 비교하여 탐색기 고장을 분류하고 중요 변수를 추출하고자 한다. 본 연구의 절차는 <Figure 1>과 같다. 제2장에서는 관련 문헌 연구를 통해 데이터 샘플링 및 차원축소 기법이 산업에서 적용된 사례를 파악한다. 제3장에서는 선정된 샘플링 및 차원축소 기법, 지도학습 기반 분류 모델에 대한 방법론을 소개한다. 제4장에서는 데이터 소개 및 실제 데이터를 바탕으로 여러 전처리 방법론을 적용한 후 분류모델을 사용하여 탐색기의 고장을 분류하고 성능을 비교 분석하여 중요 변수를 도출한다. 이를 토대로 제5장에서 결론을 제시한다.



<Figure 1> Research Process

2. 관련 문헌 연구

QIM 데이터를 분석하기 위해 데이터 샘플링, 특징 선택 및 추출 등 전처리 수행을 통해 데이터의 양을 조절하고 차원을 축소해야 한다. 여러 산업 분야에서 차원축소 및 데이터 증강과 고장 분류에 다양한 기법이 적용된 연구사례를 <Table 1>에 정리하였다. Kim et al.[10]은 솔레노이드 펌프의 5가지 조건에 대한 진동 데이터로부터 21개의 통계적 특징을 추출하였다. 추출된 특징에 대하여 Spearman 상관관계 테스트를 통해 상관관계가 적은 특징을 선택한 후 다층 퍼셉트론을 통해 솔레노이드 펌프의 고장을 분류하였다. Chae et al.[2]은 화력발전소에서 센서를 통해 획득된 데이터에서 주요 특징을 선택하기 위해 fastICA 알고리즘을 통해 차원을 20개로 축소한 후 다변량 관리도를 이용하여 이상을 감지하여 초기에 고장을 발견함으로써 설비의 가용

성을 높였다. Si et al.[21]은 선형회귀를 통한 Stepwise 기법을 통해 98개의 특징에서 32개의 특징을 선택한 후 해당 연구에서 제안된 적대적 생성 방식의 모델을 통해 사출기의 고장을 분류하였다. Lee and Kim[13]은 SHAP(SHapley Additive exPlanations)를 통해 특징을 선택한 후 TadGAN (Generative Adversarial Networks)을 통해 광업 센서데이터의 이상 구간을 탐지하였으며 일반적인 방법론에 비해 성능이 향상된 것을 확인하였다. Lee et al.[12]은 62개의 특징에 대하여 XGBoost(Extreme Gradient Boosting;), DT(Decision Tree), RF(Random Forest) 등과 같은 트리 계열 모델을 통해 변수 중요도를 기반으로 변수를 선택하고 선택되지 않은 변수들은 PCA(Principle Component Analysis)를 통해 전체 변수의 차원을 축소시켜 상위 15개의 특징으로 모델별 지도 학습을 수행한 결과 기존보다 성능이 향상되었다.

또한, 제조 공정에서는 양품보다 불량품이 소수이고, 이에 따라 불균형하게 데이터가 수집되므로 데이터의 특성을 고려하여 표본을 추출해야 한다[17]. 불균형 데이터에서는 오버 샘플링(OverSampling), 언더 샘플링(UnderSampling) 및 ROSE(Random Over Sampling Example)[16] 등과 같은 다양한 샘플링 방법이 제안되어 왔다. 이 중 가장 일반적으로 사용되는 알고리즘은 오버 샘플링 방법[15]이며, 이외에 데이터 불균형 문제를 해결하고 고장을 분류하는 연구가 활발히 진행되어 왔다. Kim et al.[11]은 적층 복합재 시편들로부터 획득된 처프 입력 신호와 응답 신호를 상태공간 모델을 이용하여 데이터를 증강한 후 딥러닝 모델을 통해 적층 복합재의 고장을 분류

하였다. Seo et al.[19]은 정압기 설비에서 자체 수집되는 유량 및 유압 데이터를 통해 이상 상태를 조기에 판단하기 위해 SMOTE(Synthetic Minority Oversampling Technique)를 사용하여 오버샘플링을 적용하였고 Gradient Boosting, 1D-CNN(Convolutional Neural Networks), LSTM(Long Short-Term Memory)을 활용하여 정압기의 고장을 분류한 결과 Gradient Boosting이 정확도 99.97%로 가장 우수한 성능을 보였다. Wei et al.[23]은 베어링 결함 진단을 위해 다중 클래스 LS-SVM(Least Squares Support Vector Machine) 기반 새로운 다중 클래스 불균형 결함 진단 방법에 관한 연구로 소수 샘플의 중요도 정렬을 기반으로 이진 불균형 문제를 해결하기 위해 새로운 오버샘플링 방법인 SCOTE(Sample-Characteristic Oversampling Technique)를 제안하였다. Choi et al.[5]은 전동기 고장 진단에 부족한 고장 데이터 수를 증대시키기 위해 GAN과 믹스업 알고리즘 비교하였으며 CNN으로 고장을 분류한 결과, 믹스업 알고리즘이 GAN 기반 전이학습에 비해 분산이 증가하여 더 높은 모델 분류성능을 보였다.

3. 연구방법론

본 장에서는 데이터 불균형 및 다차원 문제를 해결하기 위해 샘플링 기법과 차원 축소 기법을 검토하고 고장 분류 및 중요 변수 추출에 적합한 트리 계열 모델인 Random Forest, XGBoost, Gradient Boosting을 설명한다.

<Table 1> Applications of Data Processing Techniques in Fault Classification: A Literature Review

Item	Case Study	Author	Year
Gas Pressure Regulator	Failures in pressure regulators were classified using Gradient Boosting after data augmentation with SMOTE, achieving the highest accuracy of 99.97%.	Seo et al. [19]	2020
Pump	Vibration data from solenoid pumps were analyzed with Spearman's test for dimension reduction and classified for failures using a multilayer perceptron.	Kim et al. [10]	2021
Bearing	The imbalance data problem in bearings was solved and failures were classified using the proposed oversampling method SCOTE and Multi-Class LS-SVM.	Wei et al. [23]	
Thermal Power Plant	In a thermal power plant, 20 features obtained were reduced using the fastICA algorithm, and failures were classified through control charts.	Chae et al. [2]	
Silica Concentrate	Features were selected using SHAP, and anomalies were detected using TadGAN before classifying failures with machine learning models.	Lee and Kim [13]	2022
Refrigerated Container	In tree-based models, the top 15 features were selected based on variable importance from 62 features for failure classification.	Lee et al. [12]	
Laminated Composite Material	Failure in laminated composite specimens was classified using deep learning models after data augmentation through state space models with chop input and response signals.	Kim et al. [11]	
Injection Molding Machine	From 98 features, 32 were selected using the Stepwise technique through linear regression, and failures in injection molding machines were classified using the proposed adversarial generation method.	Si et al. [21]	2023
Electric Motor	The mixup algorithm was used to balance the ratio of normal and failure data in electric motors, and failures were classified using CNN.	Choi et al. [5]	

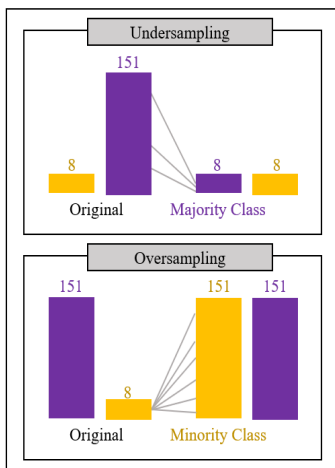
3.1 샘플링

불균형을 수치적으로 검증하기 위해 Hixiang et al.[9]은 다수 범주와 소수 범주 간 비율인 IR(Imbalance Ratio, 불균형률)을 계산하고 아래 <Table 2>와 같이 비율에 따른 불균형 정도를 제안하였다.

<Table 2> Degree of Imbalance

Notation	Imbalance
IR<4	Low imbalance
4<IR<10	Intermediate Imbalance
IR>10	High Imbalance

본 연구에서 사용된 데이터셋의 IR은 18.87로 10을 넘어가므로 불균형 정도가 매우 심하다고 볼 수 있다. 따라서 데이터의 불균형을 해결하기 위해 샘플링 기법을 통해 각 범주 간의 데이터를 균형있게 맞춰 성능을 비교할 필요가 있으며, 이에 대한 방법론에는 <Figure 2>와 같은 언더샘플링과 오버샘플링 등의 기법이 존재한다. IR이 작은 경우에는 언더샘플링, IR이 큰 경우에는 오버샘플링 방법을 적용하는 것이 최적의 성능을 보였으므로[9] 본 연구에는 오버샘플링이 더 적합하다고 판단된다.



<Figure 2> Comparison of Undersampling and Oversampling

오버샘플링과 언더샘플링은 각각 장단점을 지니며, 활용되는 데이터에 따라 선택이 가능하다. 데이터가 클래스마다 불균형한 분포를 가지는 경우, 오버샘플링은 소수 범주의 데이터를 증강시켜 각 클래스 간의 균형을 맞추는 기법으로 데이터의 손실은 발생하지 않으나, 데이터의 품질에 따라 이상치의 데이터를 증강하거나 과적합이 발생하는 문제가 존재한다. 반면에 언더샘플링은 다수 범주의 일부를 추출하여, 소수 범주의 데이터와 비율을 맞추는 기

법이다. 이는 임의로 데이터를 추출하는 기법으로 데이터의 손실이 발생하며, 데이터의 편향이 존재할 가능성을 배제할 수 없다. 또한, 머신러닝은 다양한 데이터의 입력이 필요하므로 인위적인 손실이 발생하는 언더샘플링을 통한 데이터의 증강은 본 연구에서 배제하기로 한다. 오버샘플링 기법에서 데이터를 생성하는 대표적인 알고리즘에는 머신러닝을 활용한 SMOTE 알고리즘이 존재한다.

3.1.1 SMOTE[3]

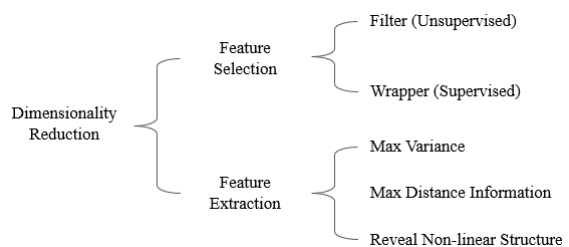
SMOTE 알고리즘은 오버샘플링 기법에서 일반적으로 널리 사용되며, 기존의 오버샘플링과 같이 기존 데이터를 랜덤으로 복원 및 추출하는 것이 아니라 서로 이웃된 범주의 데이터들을 서로 보간하여 데이터를 증강하는 기법이다. 이 기법은 소수 범주 샘플에 대해 해당 샘플의 K-최근접 이웃(K-Nearest Neighbors) 중 하나를 선택하여 x_i 로 정의한다. 그 후, 두 샘플 x_i, \hat{x}_i 사이의 연속적인 공간에서 임의의 데이터 포인트를 생성하여 새로운 샘플 x_{new} 으로 증강한다. 증강된 샘플은 식 (1)과 같이 계산된다.

$$x_{new} = x_i + \lambda(\hat{x}_i - x_i) \tag{1}$$

수식에서 λ 는 연속 균등 분포(Continuous Uniform Distribution)를 의미하며, 0과 1사이의 값을 랜덤으로 곱해 새로운 샘플이 x_i, \hat{x}_i 사이 어디에 존재하는지를 결정한다. 이러한 방식으로 생성된 데이터는 원본 데이터의 분포를 모방하며 소수 범주의 데이터를 가상으로 증강한다. 이는 분류 문제에서 클래스의 불균형이 심한 경우, 성능개선에 기여할 수 있는 장점이 있다.

3.2 차원축소 기법

차원축소는 고차원 데이터의 차원을 축소하여 새로운 차원의 데이터를 생성하는 과정으로 <Figure 3>과 같이 불필요한 변수 제거를 목적으로 하는 변수 선택(Feature selection)과 변수의 효과적인 요약을 위한 변수 추출(Feature extraction)로 구분된다.



<Figure 3> Classification of Dimension Reduction

본 연구에서는 유도탄의 666개 검사항목에 대한 측정값의 데이터 특성을 효과적으로 요약하고 통합하기 위해 원본 데이터의 정보를 새로운 형태로 통합하는 변수 추출 기법이 아닌 데이터의 원본 특성을 유지하고 설명력이 더 높은 변수 선택 방법을 적용하고자 한다. 아래는 대표적인 차원 축소 기법 중 변수추출의 PCA와 변수선택의 Lasso Regression에 대한 설명이다.

3.2.1 주성분 분석(PCA)[18]

PCA는 Pearson[18]에 의해 처음 기술되었으며, 다변량 데이터에서 선형 및 직교 변환을 사용하여 연관 관계가 있는 표본들을 저차원 공간의 표본들로 변환한다. 데이터의 분산을 최대화하는 주축을 찾아 고차원 데이터를 저차원 공간으로 투영함으로써 데이터를 정제하고 크기를 축소하는데 사용된다.

3.2.2 Lasso Regression[22]

Lasso Regression은 능형 회귀 (ridge regression)의 장점인 회귀계수 축소를 통해 예측 정확도를 높이고, 동시에 다중공선성이 높은 독립변수의 회귀계수 값을 쉽게 0으로 만드는 변수 선택의 기능이 있어 해석력을 높여준다. Lasso regression의 추정량은 식 (3)과 같이 구할 수 있다.

$$\arg \min_{\beta_1 \dots \beta_p} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (3)$$

여기서 β_0, \dots, β_p 의 값은 회귀계수, x_{ij} 는 설명변수, y_i 는 종속변수, n 은 관측값의 수, p 는 변수의 수, λ 는 정규화 매개변수로, 모델의 복잡성과 회귀 계수의 크기에 대한 패널티의 강도를 조절한다.

3.3 트리 기반 분류 모델

차원 축소를 통해 정제된 변수들을 활용하여 트리 기반 모델을 이용한 분류를 통해 고장이 발생할 가능성이 있는 유도탄을 공정 단계에서 식별하고자 한다. 트리 기반 모델은 변수 중요도를 명확하게 제시할 수 있는 강점을 지니고 있으므로 중요 변수의 도출이 가능하다. 본 연구에서는 트리 기반 분류 모델 중 다수의 결정 트리를 결합하여 분류 정확도를 높이고 과적합 방지 및 일반화 성능에 이점이 있는 Random Forest를 적용하고자 한다.

3.3.1 Random Forest[1]

Random Forest는 결정 트리의 앙상블 기법으로, 다수의 결정 트리를 생성하고 그 예측을 결합하여 최종 결정을

내리는 방법이다. Random Forest는 각 트리 훈련 시 무작위로 선택된 데이터 샘플 및 특성을 사용함으로써 모델의 분산을 줄이고, 과적합을 방지한다. 랜덤 포레스트는 분류와 회귀 문제 모두에 사용될 수 있으며, 높은 정확도와 강력한 일반화 능력을 제공한다. 랜덤 포레스트 모델의 예측은 개별 결정 트리의 예측을 평균내거나 다수결로 결정하여 계산된다. 분류 문제의 경우, 가장 많은 투표를 받은 클래스가 최종 예측 결과가 된다.

3.3.2 Gradient Boosting[8]

Gradient Boosting은 Friedman[8]에 의해 제안된 방법으로, 약한 학습기(주로 의사결정나무)를 순차적으로 학습시키면서 이전 학습기의 오차를 줄이는 방향으로 모델을 개선해 나가는 기법이다.

$$F_m(x) = F_{m-1}(x) + \rho_m h_m(x) \quad (4)$$

식 (4)에서 $F_m(x)$ 는 m 번째 단계에서의 모델을 나타내며, h_m 은 m 번째 약한 학습기의 예측, ρ_m 은 최적화할 파라미터로 m 번째 학습기의 기여도를 조절하는 학습률을 의미한다.

3.3.3 XGBoost[4]

XG부스트 회귀 모델은 Chen and Guestrin[4]에 의해 제시된 고성능 Gradient Boosting 라이브러리로, 의사결정나무로 구성된 다른 트리 기반 앙상블 모델과 달리 CART(Classification and Regression Tree) 모델을 기반으로 구성되어 있다. 또한 정규화와 병렬 처리를 통해 기존의 Gradient Boosting 방법보다 빠른 학습 속도와 높은 예측 정확도를 제공한다.

$$y'_i = \sum_{k=1}^K f_k(x_i) \quad (5)$$

$$\mathcal{O}_j = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (6)$$

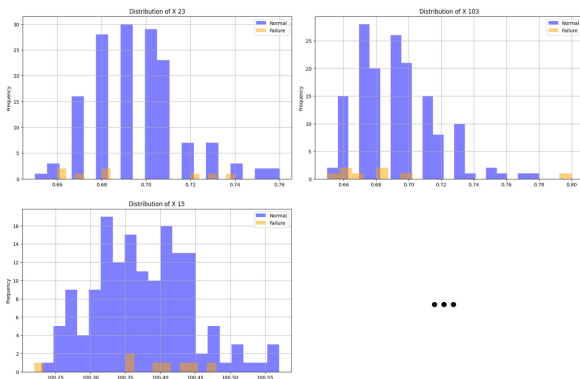
식 (6)에서 $l(y_i, \hat{y}_i)$ 은 i 번째 데이터에 대한 손실 함수이며, $\Omega(f_k)$ 는 k 번째 트리 f_k 의 복잡성을 측정하는 규제 항 (regularization term)을 의미한다. 이 규제 항은 모델의 복잡도에 따른 패널티를 부여하여 과적합을 방지할 수 있으며, 트리의 복잡성은 일반적으로 트리 내 노드의 수, 트리의 깊이 등을 기반으로 계산된다.

4. 실험 및 결과

본 장에서는 연구에 사용된 데이터셋을 소개하고 증강된 데이터에 대해 다양한 차원 축소 및 탐색기 고장 분류 모델을 적용한 결과를 비교분석 한 뒤 주요 변수를 도출하고자 한다.

4.1 연구 데이터

본 연구에서 사용한 생산 및 제조 단계에서의 QIM 데이터는 유도탄의 특성을 평가할 수 있는 다양한 시험 항목들과 그에 대한 측정값으로 구성되어 있다. QIM 데이터를 통해 고장을 분류하고 탐색기 고장에 영향을 미치는 변수를 도출하고자 한다. 전체 159발의 유도탄 중 8발의 탐색기 고장이 발생하였다. 데이터는 159발의 유도탄에 대해 수행된 666개의 검사항목에 따른 총 105,894개의 측정값으로 구성되어 있다. 검사항목 일부를 시각화한 그래프는 <Figure 4>와 같다. 666개의 시험 항목은 유도탄의 관성항법성능, 연결저항, 탐색성능, 구동성능, 발사절차, 날개펼침성능, 자체점검 등 여러 검사 항목 범주로 분류되며 그에 대한 구체적인 생산 단계에서의 부품 시험 항목으로 구성되어 있다. 구체적인 시험 항목은 보안상의 이유로 X1, ..., X666와 같이 표기하고자 한다.



<Figure 4> Comparison of Histograms between Normal Missiles and Guided Missiles that Will Fail

4.2 데이터 전처리

본 연구에서는 소수 클래스 데이터를 보간하여 새로운 샘플을 생성하고 클래스 간 불균형을 줄이는 효과적인 방법인 SMOTE 기법을 적용하여 오버샘플링을 진행하였다. 오버샘플링 된 데이터가 기존 원본 데이터의 분포를 잘 따르는지 검증하기 위해 통계적 검정 방법인 콜모고로프-스미르노프(Kolmogorov-Smirnov, KS) 테스트와 만-휘트

니 U(Mann-Whitney U) 테스트를 사용하였다. KS 테스트를 통해 데이터 분포의 전체적인 형태를 확인하고, Mann-Whitney U 테스트로 중앙값 차이를 평가하였다. KS 테스트는 두 누적 분포 함수(CDF) 간 최대 차이를 측정한다. 두 표본 집단에서 두 CDF 사이 최대 차이인 KS 통계량 D 의 식은 (7)과 같다. $F_n(x)$ 는 첫 번째 표본의 CDF, $G_m(x)$ 는 두 번째 표본의 CDF이다.

$$D = \max_x |F_n(x) - G_m(x)| \tag{7}$$

Mann-Whitney U 테스트는 두 표본의 순위 합을 기반으로 계산되어 두 표본 간 순위 차이 측정치인 U 를 통해 두 집단 간 순위 분포가 얼마나 다른지를 나타내며 이를 통해 중앙값의 차이를 검정한다.

데이터 분포의 전반적인 형태를 평가하기 위해 KS 테스트와 Mann-Whitney U 테스트를 적용한 결과, 두 통계적 검정에서 모두 P-value가 0.05 이상으로 나타났다. 이는 유의 수준 0.05에서 기존 데이터 분포와 증강된 데이터 분포가 통계적으로 유의미한 차이가 없음을 의미한다. 따라서 증강된 데이터가 기존 데이터의 분포와 동질함을 확인하였으며 이는 데이터 증강 과정이 원본 데이터 특성을 잘 보존하였음을 시사한다.

<Table 3> Results of the Test

KS Test		Mann-Whitney U Test	
Test Item	P-value	Test Item	P-value
X1	0.6918	X1	0.6045
X2	0.3761	X2	0.9259
...
X666	0.9616	X666	0.9971

유도탄 QIM 데이터의 경우 다양한 검사 항목을 포함하므로 측정값의 범위 및 단위가 각각 다르게 나타난다. 지도학습 기반 분류 모델에 적용하기 전 이러한 문제를 해결하기 위해 모든 변수를 동일한 범위 내로 조정하는 정규화 중 Min-Max Scaling을 적용하였다. 정규화는 식 (8)과 같이 수행하였으며, 666개의 변수 중 결측값이 존재하거나 모든 열의 값이 동일한 경우 해당 변수를 제거하였으며 데이터의 일관성 및 분석 정확성을 높였다. 전처리 결과, 646개의 변수에서 20개가 제거되어 646개가 되었다.

$$x_{mi}^* = \frac{x_{mi} - \min(x_m)}{\max(x_m) - \min(x_m)} \tag{8}$$

4.3 차원축소

증강 및 전처리된 데이터 세트를 활용하여 차원 축소를 진행하고자 한다. 변수의 효과적인 요약은 목적으로 하는 변수 추출 기법인 PCA, LDA, SVD와 변수의 제거를 목적으로 하는 변수 선택 기법인 RFE, Lasso Regression, Elastic Net을 포함한 다양한 기법들을 적용하여 차원을 축소 한 결과는 아래 <Table 4>와 같다.

<Table 4> Dimensionality Reduction Results

Method	Model	Dimension
	Original	646
Feature extraction	PCA	66
	LDA	1
	SVD	300
Feature selection	RFE	30
	Lasso Regression	88
	Elastic Net	311

<Table 5>는 변수추출의 대표적인 기법인 PCA를 적용하여 차원축소 한 결과이다. PCA와 같은 변수 추출 방법을 통해서도 주요 변수를 도출할 수 있으나 이는 데이터의 분산을 최대화하는 선형조합으로 새로운 형태의 변수가 생성되어 원본 변수의 해석이 어렵다. <Table 5>와 같이 PCA를 통한 차원축소는 646개의 기존 변수들의 조합으로 구성되어 66개의 변수로 축소되므로 하나의 주성분이 갖는 중요도를 식별하는 과정에서도 상당한 해석이 요구되며 원본 데이터에 대한 직관적인 해석이 어렵다. 유도탄 품질 및 수명 향상을 위해 탐색기 고장에 크게 영향을 미치는 시험 항목을 도출하여 추후 특정 항목의 관리한계 수정 및 공정 개선을 위해서는 직관적인 변수 해석이 중요하다. 따라서 탐색기 고장 분류에 주요 변수를 파악하기 위해서는 변수 추출 보다는 변수 선택이 원본 특성 유지 및 설명력에 있어 더 적합하다고 판단되어 분류 모델 성능이 유사하다면 변수 선택 방법을 적용하고자 한다.

<Table 5> PCA Principal Component

Dimension	X1	X2	X3	...	X646
1	-0.0089	-0.0053	-0.0094	...	-0.0055
2	-0.052	-0.0367	-0.0263	...	0.0033
3	-0.035	-0.0445	-0.0438	...	0.0069
...
66	0.0278	0.0386	0.0213	...	-0.0104

4.4 기계학습 모델링

4.3절에서 여러 기법을 통해 차원 축소된 데이터를 활용

하여 고장 분류 후 변수 중요도를 파악하여 탐색기 고장에 대한 주요 변수를 도출하기 위해 Random Forest를 포함한 여러 트리 계열 모델을 적용하고자 한다. 또한 회귀 모델 기반 Logistic Regression, 거리 기반 모델 KNN(K-Nearest Neighbors) 등의 방법론도 함께 성능을 비교하고자 한다.

고장 분류 머신러닝에 적용한 하이퍼파라미터는 그리드 서치(Grid Search) 기법을 적용하여 최적 값을 선정하였다. 그리드 서치 기법은 여러 하이퍼파라미터 조합 중 최적의 성능을 달성하는 조합을 찾는 방법이다. 그리드 서치 방법으로 하이퍼파라미터를 조정하고 5-fold 교차검증으로 성능을 확인하여 성능이 가장 높게 나오는 하이퍼파라미터를 적용하였다. 5-fold 교차검증은 데이터셋을 5개의 독립적인 부분으로 나눠 각각을 검증 데이터를 사용하고 나머지를 훈련 데이터로 사용하는 방법으로 과적합을 방지하고 모델의 일반화 성능을 객관적으로 평가할 수 있다. 랜덤포레스트의 경우 앙상블 내 트리개수는 200개, 트리 최대 깊이는 제한을 두지 않고 내부 노드 분할을 위한 샘플 수는 2개로 설정하였다. XGBoost의 경우 학습률은 0.1, 트리개수는 200개로 설정하였다. Gradient Boosting의 경우 학습률은 0.1, 트리 최대 깊이는 3, 트리개수는 200개로 정하여 실험에 사용하였다. 로지스틱 회귀분석은 규제 강도를 결정하는 역수인 c를 10으로 설정하였으며 최적화 문제를 해결하는 알고리즘으로는 liblinear를 사용하였다. KNN의 거리측정방법은 유클리디안 거리를 사용하였으며 가까운 이웃의 수 k를 5로 정하여 실험에 사용하였다.

4.5 성능평가

연구 결과를 정량적으로 비교하기 위해 모델 성능 평가 시 전반적 성능 파악이 가능한 Accuracy와 불균형 데이터셋에서 모델 성능을 정밀하게 평가할 수 있는 F1-Score을 사용하고자 한다. 다양한 차원 축소 기법과 분류 모델들의 5-fold 교차검증에서 얻은 결과의 성능 비교를 통해 최적의 모델 구성 방안을 도출하고자 한다.

<Table 6> Confusion Matrix

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Accuracy란 모델이 올바르게 분류한 항목의 비율로 계산된다.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{9}$$

F1-Score는 재현율과 정밀도의 조화 평균으로, 두 지표의 균형을 나타내는 값을 제공한다. 이는 특히 불균형한 데이터셋에서 유용한 지표로 활용된다.

$$F_1\text{-Score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (10)$$

증강된 데이터셋에 다양한 차원 축소 기법을 적용하여 5-fold 교차검증을 통해 여러 머신러닝 모델의 분류성능을 평가 및 비교한 결과는 <Table 7>과 같다. 실험 결과 대부분 높은 성능을 보였으며 차원 변수 추출 기법은 PCA, SVD를 적용한 후 Random Forest, XGBoost를 적용한 모델이 99.70%로 높은 성능을 보였으며 변수 선택 기법은 Lasso Regression를 통한 차원 축소 후 Random Forest 적용한 모델이 정확도 및 F1-Score에서 99.40%로 가장 높은 성능을 보였다. 차원 축소 후 대부분의 모델에서 전반적으로 정확도 및 F1-score가 향상 되었다. LDA-KNN의 경우 정확도가 58.08%에서 91.04%로 F1-Score가 70.59%에서 91.66%로 가장 큰 효과를 보였다. 필드에서 고장이 발생한 유도탄의 생산 단계 데이터 확인 시, 모든 탄이 공정관리 한계 내에 위치하나 탐색기 고장이 발생한 유도탄들의 생산 단계 데이터에서의 특징을 학습하였을 때 생산 단계 데이터로 고장을 사전에 분류할 수 있음을 확인하였다. 다음 장에서는 어떠한 시험 항목 변수들이 가장 영향을 미치는지 도출하고자 한다.

4.6 최적 모형 선정

변수 추출 및 변수 선택 기법의 성능을 비교 분석한 결과, 두 접근 방식에서 모두 유사한 높은 성능을 보이므로 원본 변수 데이터 해석에 용이한 변수 선택 방법의 모델을 최종 모델로 선정하고자 한다. 따라서 본 연구에서는 변수 선택 기법 중 가장 높은 성능을 보인 Lasso Regression-

Random Forest를 최적 모형으로 선정하였다. Lasso Regression-Random Forest를 통해 고장을 분류하고 분류 기준에 대한 변수 중요도를 파악한 결과는 <Table 8>과 같다. 보안상의 이유로 66개의 시험 항목은 X633, X571, X246, ..., X133으로 표기하였다.

변수 중요도 분석 결과, 구동성능, 날개펼침성능, 연결저항, 탐색성능, 관성항법 성능에 관한 변수들 중 가장 높은 영향을 미치는 변수는 특정 부품의 탐색 성능에 대한 시험 측정값으로 확인되었다. 그 외에도 발사절차, 연결저항 등에 대한 특정항목의 시험 측정값이 중요도가 높게 도출되었다. 가장 중요도가 높은 변수는 탐색성능 관련 시험 항목이었으나 가장 많은 개수를 차지하는 것은 연결저항과 관련된 특정 시험 항목이었다.

이와 같이 생산 단계 데이터로 고장이 발생할 가능성이 높은 유도탄을 분류하는 모델에서 중요도가 높은 시험 항목 변수를 도출하였다. 이를 통해 탐색기 고장과 관련된 시험 항목의 중요도와 관련 시험 항목 관리한계 재설정에 도움을 줄 수 있으며 추후 품질관리시스템 내 데이터 기반 의사결정 범위를 확장시키고 관리의 용이성 향상을 기대할 수 있을 것이다. 또한 이를 통해 품질 향상과 필드에서 고장이 발생할 유도탄을 생산 단계에서 사전에 검토함으로써 필드에서 고장을 방지하여 유도탄의 수명을 향상시킬 수 있을 것이라 기대할 수 있다.

<Table 8> Variable Importance for Classification Criteria

Ranking	Importance	Test Item Number	Test Item
1	0.0866	X633	Exploration performance
2	0.4197	X571	Launch procedure
3	0.04139	X246	Connection resistance
...
88	0.00045	X133	Connection resistance

<Table 7> Experimental Results

	Model	Dimension	Classification Model									
			Random Forest		XGBoost		Gradient Boosting		Logistic Regression		KNN	
			Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
	Original	646	0.9552	0.9565	0.9641	0.9639	0.9696	0.9705	0.9279	0.9348	0.5808	0.7059
Feature extraction	PCA	66	0.9970	0.9970	0.9970	0.9969	0.9850	0.9846	0.9369	0.9422	0.6824	0.7618
	LDA	1	0.9338	0.9405	0.9338	0.9405	0.9338	0.9405	0.9189	0.9278	0.9104	0.9166
	SVD	300	0.9970	0.9970	0.9970	0.9969	0.9730	0.9732	0.9279	0.9348	0.5808	0.7059
Feature selection	RFE	30	0.9850	0.9845	0.9580	0.9584	0.9462	0.9508	0.9069	0.9133	0.8799	0.8965
	Lasso Regression	88	0.9940	0.9937	0.9491	0.9506	0.9699	0.9714	0.9729	0.9741	0.7766	0.8238
	Elastic Net	311	0.9880	0.9874	0.9696	0.9705	0.9641	0.9661	0.9549	0.9578	0.6060	0.7173

5. 결론

본 연구는 유도탄의 품질 및 수명 향상을 목표로 하며, 유도탄의 탐색기 고장을 예방하기 위해 생산 및 제조 단계에서의 QIM 데이터를 활용하여 트리 계열 분류 모델을 통해 생산 단계에서 탐색기 고장이 발생할 유도탄을 사전에 분류하고 탐색기 고장 결정에 중요한 변수를 도출하였다.

고장 데이터의 양이 제한적이며 차원의 수가 높은 상황을 고려하여 오버샘플링과 차원 축소 기법을 적용한 전처리 과정을 수행하였다. 변수 추출과 변수 선택 기법이 유사한 높은 성능을 보였으므로 데이터 특성 파악 및 해석과 추후 관리 측면에 있어 더 적합한 변수 선택 방법을 선택하였다. 최종 선택된 모델은 Lasso Regression을 통한 차원 축소 후 Random Forest로 고장을 분류하는 모델이며 99.40%의 정확도로 높은 성능을 보였다. Lasso Regression-Random Forest를 통해 주요 변수를 파악하여 탐색기 고장에 영향을 미치는 생산 단계에서의 시험 항목을 도출하였다. 이러한 접근 방식을 통해 탐색기 고장이 발생할 유도탄을 생산 단계에서 선별함으로써 고장률을 감소시키고 제품의 수명을 연장할 수 있는 가능성을 확인하였다. 또한 향후 관련 시험 항목의 관리한계 범위 조정 및 유지보수 전략 개선 등과 같은 품질관리시스템 내 데이터 기반 의사결정 범위 확장에 기여할 수 있을 것으로 기대된다. 본 연구에서는 생산 단계에서 탐색기 고장의 발생 가능성이 높은 유도탄을 분류하고 분류된 유도탄에 영향을 미치는 검사항목들을 도출하였다. 향후 연구에서는 표본 검사된 정상 유도탄에 대하여 도출된 검사항목의 열화 추세와 지역별 온도 및 습도가 열화에 미치는 영향 등을 파악하여 유도탄의 신뢰성을 향상시킬 수 있는 연구를 진행하고자 한다.

Acknowledgement

이 논문은 2022년 정부(방위사업청)의 재원으로 국방기술진흥연구소의 지원을 받아 수행된 연구임(KRIT-CT-22-081, 무기체계 CBM+ 특화연구센터).

References

- [1] Breiman, L., Random Forests, *Machine Learning*, 2001, Vol. 45, No. 1, pp. 5-32.
- [2] Chae, S.G., Kim, G.R., Bae, A.-Y., and Bae, S.J., Failure Diagnosis and Prediction for a Thermal Power Plant Generator using fastICA, *Journal of Applied Reliability*, 2021, Vol. 21, No. 4, pp. 341-351.
- [3] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W. P., SMOTE: synthetic minority over-sampling technique, *JAIR*, 2002, Vol. 16, pp. 321-357.
- [4] Chen, T. and Guestrin, C., XGBoost: A scalable tree boosting system, in *Proc. of 22nd ACM SIGKDD Int. Conf. on Knowl. Discovery and Data Mining*, Aug. 2016, pp. 785-794.
- [5] Choi, E.-J., Han, J.-H., and Hong, S.-K., Performance Improvement of Transfer Learning-based Motor Fault Diagnosis Using Mix-up Data Considering Data Imbalance, *KIEE*, 2023, Vol. 72, No. 2, pp. 255-262.
- [6] Donoho, D.L., High-dimensional Data Analysis: The Curses and Blessings of Dimensionality, *AMS Conference on Math Challenges of the 21st Century*, 2000 pp. 1-32.
- [7] Fleeman, E.L., *Tactical Missile Design*, American Institute of Aeronautics and Astronautics Reston, 2nd Ed., 2006.
- [8] Friedman, J.H., Greedy function approximation: A gradient boosting machine, *Ann. of Statist.*, 2001, Vol. 29, No. 5, pp. 1189-1232.
- [9] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G., Learning from class-imbalanced data: Review of methods and applications, *Expert systems with applications*, 2017, Vol. 73, No. 1, pp. 220-239.
- [10] Kim, S., Ejike, A.U., and Hur, J., A Study on Fault Classification of Solenoid Pumps based on Multi-Layer Perceptron, *Journal of Applied Reliability*, 2021, Vol. 21, No. 1, pp. 12-19.
- [11] Kim, S.J., Khan, A., Sohn, J.W., and Kim, H.S., Deep Learning based Fault Diagnosis in Laminated Composites using Augmented Data, *Spring Conference of the Korean Society of Mechanical Engineers*, 2022, Busan, Korea, pp. 374-375.
- [12] Lee, S., Park, S., Lee, S., Lee, H., Yu, S., and Lee, K., A Comparative Study on the Methodology of Failure Detection of Reefer Containers Using PCA and Feature Importance, *Journal of The Korea Convergence Society*, 2022, Vol. 13, No. 3, pp. 23-31.
- [13] Lee, S.H. and Kim, Y.S., A Pre-processing Process Using TadGAN-based Time-series Anomaly Detection, *Journal of Korean Society for Quality Management*, 2022, Vol. 50, No. 3, pp. 459-471.
- [14] Lee, Y., Yong, H., Jung, J., and Kim, J., Development of Dormant Missile Health Monitoring Methodology based on Environmental Data, *Journal of Applied Reliability*, 2022, Vol. 22, No. 3, pp. 219-228.
- [15] Mao, W., He, L., Yan, Y., and Wang, J., Online

Sequential Prediction of Bearings Imbalanced Fault Diagnosis by Extreme Learning Machine, *MSSP*, 2017, Vol. 83, pp. 450-473.

- [16] Menardi, G. and Torelli, N., Training and assessing classification rules with imbalanced data, *Data Mining and Knowledge Discovery*, 2014, Vol. 28, No. 1, pp. 92-122.
- [17] Mourtzis, D., Vlachou, E., and Milas, N., Industrial Big Data as a Result of IoT Adoption in Manufacturing, *Procedia CIRP*, 2016, Vol. 55, pp. 290-295.
- [18] Pearson, K., On Lines and Planes of Closest Fit to Systems of Points in Space, The London, *Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1901, Vol. 2, No. 11, pp. 559-572.
- [19] Seo, C.-Y., Suh, Y.-J., and Kim, D.-J., Study on Fault Detection of a Gas Pressure Regulator Based on Machine Learning Algorithms, *JKSCI*, 2020, Vol. 25, No. 4, pp. 19-27.
- [20] Seo, Y., Lee, K., Lee, Y., and Kim, J., Reliability Prediction Based on Field Failure Data of Guided Missile, *Journal of Applied Reliability*, 2018, Vol. 18, No. 3, pp. 250-259.
- [21] Si, J., Jeong, J., Jeong, M., and Kim, S., Anomaly

Detection of Injection Molding using Statistics-based Feature Selection and Generative Adversarial Learning, *JKIIT*, 2023, Vol. 21, No. 4, pp. 11-20.

- [22] Tibshirani, R., Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1996, Vol. 58, No. 1, pp. 267-288.
- [23] Wei, J., Huang, H., Yao, L., Hu, Y., Fan, Q., and Huang, D., New imbalanced bearing fault diagnosis method based on Sample-characteristic Oversampling TechniquE (SCOTE) and multi-class LS-SVM, *Applied Soft Computing*, 2021, Vol. 101, p. 107043.

ORCID

- Ye-Eun Jeong | <https://orcid.org/0009-0002-5280-2315>
- Kihyun Kim | <https://orcid.org/0009-0008-1887-1679>
- Seong-Mok Kim | <https://orcid.org/0009-0007-2600-4404>
- Youn-Ho Lee | <https://orcid.org/0009-0009-0398-3149>
- Ji-Won Kim | <https://orcid.org/0009-0009-5480-1886>
- Hwa-Young Yong | <https://orcid.org/0009-0005-9016-4336>
- Jae-Woo Jung | <https://orcid.org/0009-0008-9992-8947>
- Jung-Won Park | <https://orcid.org/0009-0004-2652-6382>
- Yong Soo Kim | <https://orcid.org/0000-0003-3362-4496>