Korean Chem. Eng. Res., 58(2), 248-256 (2020)
https://doi.org/10.9713/kcer.2020.58.2.248
PISSN 0304-128X, EISSN 2233-9558

기계학습 기반의 가스폭발위험범위 예측모델에 관한 연구

정용재 · 이창준†

부경대학교 안전공학과
48513 부산광역시 남구 용소로 45
(2020년 2월 10일 접수, 2020년 3월 10일 수정본 접수, 2020년 3월 23일 채택)

A Study on Predictive Models based on the Machine Learning for Evaluating the Extent of Hazardous Zone of Explosive Gases

Yong Jae Jung and Chang Jun Lee†

Department of Safety Engineering, Pukyong National University, 45, Yongso-ro, Nam-gu, Busan, 48513, Korea
(Received 10 February 2020; Received in revised form 10 March 2020; accepted 23 March 2020)

요 약

본 연구에서는 폭발위험장소의 방폭설비 설치를 위해 필요한 가스폭발위험범위 예측모델 개발을 수행하였다. 이를 위해 12개의 가연성가스에 대한 1,200개의 폭발위험범위 데이터를 생성하였다. 가스폭발위험범위를 출력변수로 설정하였고 데이터 생성과정에서 필요한 12개의 변수를 입력변수로 설정하였다. 다중 회귀, 주성분 회귀, 인공신경망 기법을 이용해 예측모델을 개발하였다. 각각 모델의 예측 성능을 비교한 결과, 평균절대퍼센트오차(MAPE)는 각각 44.2%, 49.3%, 5.7%이고 평균제곱근오차(RMSE)는 1.389 m, 1.602 m, 0.203 m로 나타났다. 결과를 통해 인공신경망이 가장 우수한 성능을 보여주었고 가스폭발위험범위 예측을 위한 최적 모델이라는 것을 확인하였다.

Abstract – In this study, predictive models based on machine learning for evaluating the extent of hazardous zone of explosive gases are developed. They are able to provide important guidelines for installing the explosion proof apparatus. 1,200 research data sets including 12 combustible gases and their extents of hazardous zone are generated to train predictive models. The extent of hazardous zone is set to an output variable and 12 variables affecting an output are set as input variables. Multiple linear regression, principal component regression, and artificial neural network are employed to train predictive models. Mean absolute percentage errors of multiple linear regression, principal component regression, and artificial neural network are 44.2%, 49.3%, and 5.7% and root mean square errors are 1.389m, 1.602m, and 0.203 m respectively. Therefore, it can be concluded that the artificial neural network shows the best performance. This model can be easily used to evaluate the extent of hazardous zone for explosive gases.

Key words: Extent of hazardous zone of explosive gases, Multiple linear regression, Principal component regression, artificial neural network

1. 서 론

현재, 국내 화학물질 사용량은 매년 큰 폭으로 증가하고 있다. 1996년부터 2015년 사이에 화학물질 유통량은 308% 증가하였으며 석유화학산업 규모를 나타내는 에틸렌 생산량은 216% 증가하였다[1]. 이로 인하여 위험물을 대량으로 취급하는 화학 공장에서의 화재·폭발 사고 위험성도 증가하고 있다. 화재·폭발 사고의 발생빈도는 낮지만, 일단 사고가 발생하는 경우 그 피해가 사고가 발생한 사업장 내에 국한되는 것이 아니라, 인근 회사와 지역주민들에게까지 막대한 피해를 발생시킬 수 있다. 따라서, 사고 발생의 영향을 감소하기 위해서 다양한 사고완화설비에 관한 연구가 이루어지고 있다[2].

폭발위험 분위기가 생성될 수 있는 장소에서 화재·폭발이 발생하는 기본조건은 폭발 범위 내 가연성가스와 공기의 혼합과 점화원의 존재이다. 점화원은 마찰, 충격, 반응열 등의 기계적, 화학적 점화원과 정전기, 전기스파크 등의 전기적 점화원으로 분류될 수 있는데, 탄화수소계 위험물들의 점화에너지가 0.25 mJ 이하라는 것을 고려할 때 스파크를 일으킬 수 있는 모든 전기설비는 잠재적인 점화원이며 화재·폭발의 위험성은 공정지역 내의 전기설비의 수에 직접 비례한다고 볼 수 있다[3]. 전기설비가 점화원이 될 수 있는 확률을 0에

†To whom correspondence should be addressed.
E-mail: changjunlee@pknu.ac.kr
‡이 논문은 POSTECH 이인범 교수님의 정년을 기념하여 투고되었습니다.

가까운 낮은 값으로 만들어 화재·폭발을 예방하기 위한 구체적인 안전조치가 바로 전기 방폭설비이다.

2008~2017년에 발생한 중대 산업사고 65건을 분석해보면, 정전기가 점화원인 경우가 총 22건으로 34%를 차지하며, 그다음으로는 정전기로 인하여 전기스파크가 발생하여 사고가 발생하는 경우가 총 10건으로 15.4%를 차지하고 있다[4]. 또한, 같은 기간 동안 발생한 370건의 화재·폭발 사망사고 중 방폭 구조 전기설비 관련 사고가 37건으로 10%를 차지하고 있으며[4], 이는 방폭설비 미비로 인한 사고 비율이 여전히 높음을 알 수 있다.

위험물을 사용하는 공정지역에 방폭 전기설비를 설치하기 위해서는 먼저 구체적인 설치범위를 정해야 한다. 이를 위한 국내 기술표준으로는 2012년 제정되어 2017년 개정된 한국산업표준(KS C IEC 60079-10-1)과 2018년 제정된 가스시설의 폭발위험장소 종류구분 및 범위 산정에 관한 기준(KGS GC101)이 있는데, 이 기술표준들은 국제전기기술위원회 표준(IEC 60079-10-1)을 그대로 준용하여 제정되었으므로 세부내용이 거의 동일하다. 또한 이 기술표준들이 매우 복잡하고 난해하기 때문에 기술검토 단계에서 많은 인력과 시간이 필요하다. 이를 개선하기 위해 간이법(simplified methods) 및 조합법(combination methods)과 같은 간소화된 접근방법[5]이 개발되었지만, 그 정확도가 떨어지기 때문에 위험 범위를 과다하게 산정할 수 있다. 이러한 요인은 방폭설비 시공비용뿐만 아니라 추후 공정 운영 중의 유지보수 비용 증가를 유발할 수 있다.

이처럼 폭발위험범위를 산정하는 경우 정확성을 기하면 효율성이 떨어지고 효율성에 집중하면 부정확한 결과 때문에 비용이 증대되는 현실적인 문제가 있으므로, 산업현장에서 사용할 수 있는 합리적인 위험범위 산정방안을 찾기 위한 연구가 시급하다.

폭발위험장소 설정 기준에 관한 다양한 선행연구가 있다. Jung 등 [6]은 폭발위험장소 설정 기준 개정에 대해 이전과 최신기준을 비교, 분석하여 최신기준과 관련하여 향후 추가로 수행할 필요가 있는 연구내용을 제시하였으며, Choi [7]는 폭발위험장소 설정 기준 관련 제한사항과 보완대책을 제시한 바 있으며, Bozek [8]은 국제전기기술위원회 표준(IEC 60079-10-1)을 적용한 사례연구를 통해 가연성 물질의 방출 특성, 작업 환경 등을 동시에 고려해야 함을 제시하였다. 응용소프트웨어를 활용하여 공학적 해석을 시도한 연구도 있는데, Souza 등 [9]은 ANSYS 소프트웨어로 메탄가스 누출 및 확산에 대한 전산유체역학(CFD:Computational Fluid Dynamics) 시뮬레이션을 수행하여, CFD가 수작업으로 계산한 결과에 비해 높은 정확도를 보여줌을 확인하였다. Miranda 등 [10]은 천연가스 보일러실 폭발위험범위 사례연구에서 FLUENT 응용소프트웨어를 사용한 CFD 모델링 결과와 IEC 60079-10-1 표준방법에 따른 위험범위 산정결과를 분석하여 CFD 사용이 더 유용한 결과를 제시할 수 있음을 보여주었다. Shrivastava 등 [11]은 3차원 모델링을 활용한 위험범위 설정방법이 기존의 2차원 도면을 사용하는 방법에 비해 기술기준을 더 엄격하게 적용하는 것이라 하였으며, 사례연구를 통해 보다 높은 정확도의 연구결과를 보여주었다.

현재 주목받고 있는 전산유체역학을 이용한 연구의 경우, 매우 정확하지만 다양한 결과를 얻기 위해서는 많은 시간과 노력이 필요하다는 단점이 있다. 실제 산업현장에서 전산유체역학 기법을 이용하여 폭발위험범위를 산정하는 것이 매우 어렵기 때문에, 더욱 쉽게 적용할 수 있으면서도 높은 정확도를 갖는 모델의 개발이 시급하다.

다양한 통계 기법을 이용하여 위험범위를 예측한 연구를 분석해보면, Jung과 Lee [12]는 액체 누출 및 확산 메커니즘 연구로 산출한 데이터를 민감도 분석 및 다중회귀분석과 같은 통계적 기법을 이용하여 인화성액체 폭발위험범위를 산정할 때 활용할 수 있는 증발률 추정 모델을 제시하였다. Zohdirad 등 [13]은 PHAST 소프트웨어로 수소 누출·확산 모델링 데이터를 수집한 후 요인분석과 회귀분석을 실시하여 누출압력과 누출구멍 크기를 입력변수로 하는 위험범위 예측식을 도출하였다. 최근에는 다양한 분야에서 머신러닝 기법을 이용하여 예측모델을 개발하는 연구가 활발히 이루어지고 있다. 예를 들면 환경 분야에서 기상데이터로 인공신경망과 서포트벡터머신 분석을 수행하여 미세먼지 농도를 예측한 연구[14], 공정제어 분야에서 주성분분석과 서포트벡터머신으로 공정이상 진단 관련 연구[15,16], 토목공학 분야에서 다중회귀분석과 주성분회귀로 교량 빅데이터를 분석하여 교통량에 따른 변위 추정모델을 도출한 연구[17] 등 실용적인 연구가 많이 수행되었다. 본 연구에서는 기계학습 회귀기법을 포함한 다양한 회귀분석 기법을 이용하여 정확성도 유지하는 한편, 기존의 모델들보다 쉽고 효율적으로 적용이 가능한 폭발위험범위 모델을 제시하고자 한다.

2, 3장에서는 본 연구를 위해 필요한 데이터를 생성하기 위해 적용한 한국산업표준(KS C IEC 60079-10-1)에 대해 소개하고, 회귀모델을 만들기 위한 통계기법에 대해 설명하고자 한다. 그리고 4장에서는 다양한 모델들의 성능을 평가하고 어느 정도 정확도가 있는지 검증하고자 한다.

2. 연구내용

2-1. 폭발위험범위 산정방법 선정

앞서 언급한 폭발위험범위 산정에 관한 기술표준 외에도 미국방화협회(NFPA)나 미국석유협회(API) 기술표준도 해외에서 널리 사용되고 있으며 국내에서도 예전에는 화학플랜트 설계단계에서 적용된 사례가 있다. 그러나 산업안전보건법이나 고압가스안전관리법 등의 국내 법규에서 국제전기기술위원회 표준을 기반으로 제정된 국내 기술표준들을 적용하도록 강제하고 있어, 현재 국내에서는 이 기술표준들이 다른 표준들에 비해 우선순위가 높다. 그러므로 이번 연구에서는 한국산업표준(KS C IEC 60079-10-1)에 따라 폭발위험범위를 산정하여 데이터를 생성하고자 한다.

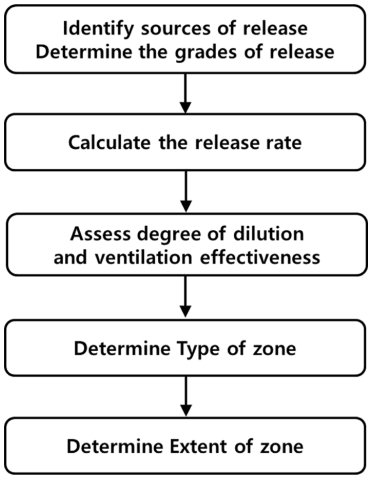

Fig. 1. The hazardous area classification step.

Table 1. Gas release rate formulas [5]

Case	Formula
Choke	$C_d S p \sqrt{\gamma \frac{M}{ZRT}\left[\frac{2}{\gamma+1}\right]^{(\gamma+1)/(\gamma-1)}}$
Non Choke	$C_d S p \sqrt{\frac{M}{ZRT}\frac{2\gamma}{\gamma-1}\left[1-\left(\frac{p_a}{p}\right)\right]^{(\gamma-1)/\gamma}\left(\frac{p_a}{p}\right)^{1/\gamma}}$

Fig. 1에서는 폭발위험범위를 산정하는 절차를 단계적으로 보여주고 있다. 먼저 누출원 평가를 위해 공정에서 사용될 위험물질 물성 및 운전조건이 포함된 데이터 목록을 작성한 후 Table 1의 수식에 따라 누출량(kg/s)을 계산한다[5]. 본 연구에서는 가연성가스에 대한 위험범위 예측모델을 제안하려는 목적이 있으므로 가스 누출량 계산식만을 반영하였다. Table 1에서, C_d는 배출계수, S는 누출구멍 크기(mm^2), M은 분자량(kg/kmol), Z는 압축인자, R은 가스상수(8,314 J/kmolK), γ은 비열비이다. 또한 p_a는 대기압(bar), p는 누출압력(bar), T는 누출온도(ºC)이며, 계산식에는 대기압과 누출압력 단위를 pa로 누출온도는 절대온도(K) 단위로 변환하여 입력하였다.

환기평가 단계에서는 희석등급과 환기유효성을 결정하여야 한다. 희석등급을 결정하기 위해 먼저 누출특성과 환기 속도를 산정해야 하는데, 누출특성(m^3/sec)은 누출량을 가스밀도(kg/m^3), 폭발하한계 및 누출물 혼합 정도를 고려한 안전계수로 나누어 구한다. 여기서 가스밀도는 대기온도(ºC)에 따라 달라질 수 있으므로, 가스밀도를 계산할 때 대기온도를 반드시 반영하여야 위험범위를 보다 정확하게 산정할 수 있다. 환기 속도(m/s)는 옥외 풍속을 적용하거나 옥내 환기량을 건물 수직 단면적으로 나누어서 산정한다[5,12].

그리고 기술기준에서 제시한 Fig. 2에서 수평축의 누출특성 및 수직축의 환기 속도에 해당하는 각각의 값에 대한 교차점을 찾아서 어느 부분에 속하는지 확인하여 고희석, 중희석, 저희석 여부를 판단한다[5,12]. 환기 유효성 등급은 우수, 양호, 미흡으로 구분하며, 평가자의 정성적이고 경험적인 판단에 따라결정된다.

마지막으로 폭발위험장소로 설정된 경우 Fig. 3을 이용해 가로축의 누출특성과 그래프 직선의 교차점에 해당하는 세로축 값으로부터 위험범위를 산출할 수 있다[5,12].

Fig. 2와 3은 국제전기기술위원회(IEC)에서 실시하였던 전산유체역학 시뮬레이션(CFD) 데이터를 바탕으로 작성되어 기술표준에 반영된 것이다. 현재까지 Fig. 2의 희석등급을 분류하기 위한 기술적 근거나 위험범위를 구하는 Fig. 3의 그래프 함수가 공개되어 있지 않아, Fig. 2와 3의 차트 그림을 보고 직접 읽어서 희석등급과 위험범위를 각각 결정하여야 한다. 또한, Fig. 3에서 'Heavy gas', 'Diffusive', 'Jet'로 가스의 종류를 나누는 뚜렷한 기준이 없으며 평가자의 경험에 의존해야 한다.

첫 번째 단계인 누출원 평가에서 고려해야 하는 주요 누출원으로는 펌프와 압축기 등 회전기기 밀봉부, 배관 및 밸브의 피팅류 개스킷 접속부, 안전밸브와 통기관 말단이 있다. 위험물을 대량 취급하는 화학공정 설비에는 이러한 누출원이 셀 수 없이 많으며, 개별 누출

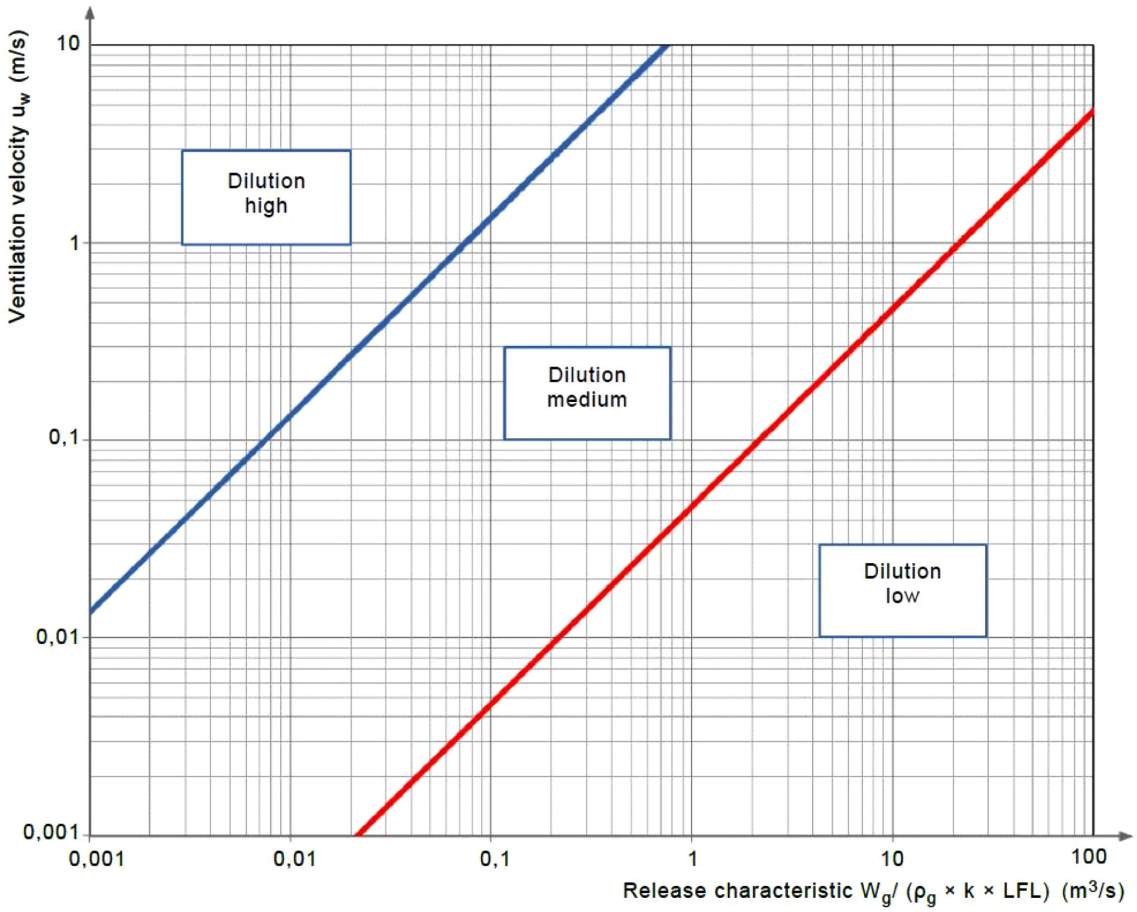

Fig. 2. The chart for assessing the degree of dilution [5].

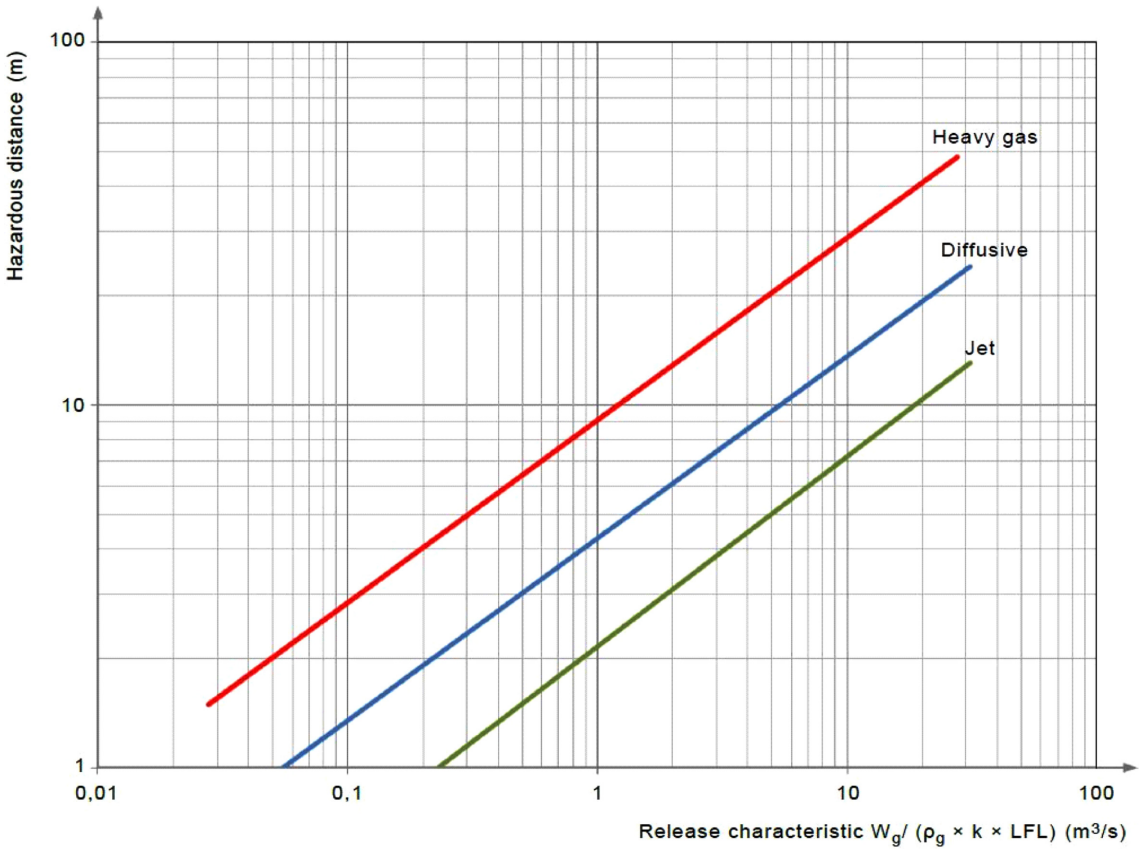

Fig. 3. The Chart for estimating hazardous area extent [5].

원마다 이제까지 기술한 방법으로 일일이 폭발위험범위를 산정하는 것은 많은 인력과 시간을 요구하는 일이다.

2-2. 연구대상 물질

본 연구에서 사용할 가연성가스를 선정하기 위해 2019년 한국석유화학공업협회 석유화학 편람[18]과 2014년 환경부 화학물질 통계조사 자료[19]를 토대로 국내 생산량 및 유통량 상위를 차지하는 가연성가스 12종을 Table 2과 같이 선정하였다. 선정된 가연성가스의 물성치 데이터는 한국가스안전공사[20] 및 미국 해양대기관리국(NOAA)[21]에서 공개한 자료를 이용하였다.

2-3. 변수 선정 및 데이터 생성

본 연구의 목적은 기계학습 회귀기법을 포함한 다양한 통계기법을 이용하여 폭발위험범위 예측모델을 만들고자 하는 것이다. 이를 위해 출력변수와 출력변수에 영향을 미치는 입력변수를 지정해야 한다. 폭발위험범위를 출력변수로 지정하였으며, 2장에서 설명한 위험범위 산정방법으로부터 입력변수를 선정할 수 있다. 누출원 평가단계에서 설비 운전조건이 반영된 누출압력과 누출온도가 출력변수에 영향을 미치는 입력변수이다. 누출량 계산 단계에서는 앞서 제시한 Table 1의 계산식에 반영되는 배출계수, 누출구멍 크기 외에 누출물질의 고유 물성치인 분자량, 압축인자, 비열비를 입력변수로

Table 2. Physical properties of combustible gas [20,21]

Gas	Molecular formula	Molar mass (g/mol)	Density (kg/m^3)	L.E.L (vol%)
Hydrogen	H_2	2.02	0.082	4.0
1,3Butadiene	C_4H_6	54.1	2.428	2.0
1Butene	C_4H_8	56.1	2.366	1.6
Methylbutene	CH_2	70.1	2.993	1.5
Ammonia	NH_3	17.03	0.707	15
Butane	C_4H_{10}	58.12	2.110	1.8
Ethane	C_2H_6	30.07	1.242	3.0
Ethylene	C_2H_4	28.05	1.261	3.1
Hydrogen Sulfide	H_2S	34.08	1.406	4.3
Methane	CH_4	16.04	0.717	5.0
Propane	C_3H_8	44.09	1.868	2.2
Propylene	C_3H_6	42.08	1.786	2.0

볼 수 있다. 그러나 계산식에 반영되었던 대기압은 1기압으로 항상 일정한 수치가 반영되므로 입력변수에서 제외하였다. 환기평가 단계에서는 폭발하한계, 대기온도, 가스밀도 및 안전계수와 함께 환기속도가 필요하다. 이렇게 고려한 내용을 종합하여 본 연구에서는 출력변수로 폭발위험범위(Y)를 지정하고, 입력변수에는 분자량(X1), 폭발하한계(X2), 대기온도(X3), 운전압력(X4), 운전온도(X5), 누출공 크기(X6), 배출계수(X7), 안전계수(X8), 환기속도(X9), 가스밀도(X10), 비열비(X11), 압축인자(X12)를 지정하여 출력변수 1개, 입력변수 12개로 모델링을 진행하고자 한다.

연구 대상인 12종의 가연성가스에 대해 각 가스당 100차례씩 위험범위를 산정하여 총 1,200개의 폭발위험범위 데이터를 생성하였다. 데이터 생성과정에서 분자량, 폭발하한계 및 비열비에 대해 물질별 고유물성치를 적용하였고 압축인자는 누출압력과 온도, 가스밀도는 대기 온도와 분자량에 따라 산정된 수치가 반영되었다. 그 외 데이터 생성 시 입력변수의 값은 무작위추출을 통해 계산에 반영하였다.

12개의 입력변수는 각각 단위와 척도가 모두 다르거나 비열비나 압축인자와 같은 무차원수도 있어 데이터 비교에 어려움이 있다. 이러한 문제를 해결하기 위해, 식 (1)과 같이 표준점수(Z-Score)로 구한 후 평균 0, 표준편차 1인 데이터 분포로 변환한 표준화를 실시하였다.

$$Z = \frac{X-\mu}{\sigma} \tag{1}$$

where, X:Data, μ:Mean, σ:Standard deviation

일반적으로 예측모델을 만드는 경우 학습(Training)과 검증(Validation)의 단계를 거쳐야 한다. 학습 과정에서는 주어진 데이터를 이용하여 모델을 생성하는 단계이며, 이 과정에서 모델 수립에 필요한 최적의 파라미터를 결정한다. 검증 단계에서는 학습 과정에서 생성된 모델이 제대로 성능을 발휘하는지 평가하는 과정이며, 과적합(Over-fitting)이나 과소 적합(Under-fitting)의 발생 여부를 평가하고, 예측모델의 정확도가 충분한지를 검증한다. 본 연구에서는 1,200개의 데이터 중 학습 및 검증 데이터 중 절반씩 무작위로 선정하였다.

3. 연구방법

3-1. 연구방법 선정

일반적으로 회귀분석은 출력변수를 입력변수의 함수로 출력변수를 예측하는 모델을 통칭하는 것으로 모델이 간단할 뿐 아니라 해석이 쉽고 안정적이어서 통계 분야에서 널리 사용되고 있는 방법이다. 하지만, 입력변수의 수가 많은 다변량 자료를 분석하는 데 있어서 다중공선성(collinearity) 문제가 발생할 수 도 있으며, 정확한 예측모델을 만드는데 어려움이 있다. 이와 같은 문제점이 예상되므로, 본 연구에서는 가장 단순한 다중 회귀분석과 입력 변수 수를 줄일 수 있는 기법인 주성분 회귀, 그리고 비선형 모델을 만드는 데 장점을 가지고 있는 인공신경망 모델을 생성하여 가장 좋은 성능을 보여주는 회귀모델을 최적모델로 선정하기로 하였다.

3-2. 다중 회귀분석(MLR: Multiple Linear Regression)

본 연구에서는 하나의 출력변수에 12개의 입력변수를 포함하는 다중 회귀분석을 모델링 방법으로 적용하였다.

출력변수 Y에 대해 n개의 입력변수(X_1, X_2, …, X_n)가 있을 때 다중 회귀모델은 식(2)과 같으며, 회귀계수(β_1, β_2, …, β_n)는 식 (3)과 같이 계산되는 잔차제곱합을 최소화하는 최소제곱법으로 추정될 수 있다.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon \tag{2}$$

$$\sum_{i=1}^{p} (Y_i - \beta_0 - \beta_1 X_1 - \cdots - \beta_n X_n)^2 \tag{3}$$

여기서 p는 데이터 개수이다. 또한 ε은 오차이며, 평균 0, 분산 σ^2에 정규분포를 따른다.

3-3. 주성분 회귀분석(PCR : Principal Component Regression)

입력변수가 12개인 본 연구의 데이터와 같이 변수의 개수가 많은 다변량 자료인 경우 데이터 분석에 제한이 있을 수 있으므로 데이터 압축이 필요하다. 즉, 복잡한 데이터를 간단히 하여 해석과 설명을 쉽게 할 수 있도록 차원을 축소하거나 요약해야 할 필요도 있는 것이다. 이를 위한 주성분 분석에서는 다변량 변수들의 공분산 행렬이나 상관행렬을 이용하여 입력변수들의 선형결합을 통해 서로 상관되어 있지 않은 새로운 변수인 주성분(Principal Component)들을 산출한다. 이 중 분산이 최대가 되도록 설정된 주성분을 중심으로 입력변수의 정보를 가장 많이 가지고 있는 주성분들을 선정한다. 그리고 주성분 분석을 통해 선정된 주성분들을 입력변수로 반영하여 출력변수를 예측하는 회귀분석 모델을 만든다. 이러한 과정으로 회귀모델을 개발한다면 기존 입력변수들의 정보를 대부분 보유하면서도 상관관계가 높은 변수 간 동일한 주성분으로 뭉치는 효과와 함께 각 주성분 간의 독립성을 통해 다중공선성 문제를 해결할 수 있는 장점이 있다[22]. 주성분 회귀분석에 대한 자세한 내용은 Yang과 Park [22]의 연구를 참고하면 된다.

3-4. 인공신경망(ANN : Artificial Neural Network)

인공신경망은 입력층, 출력층 그리고 입력층과 출력층 사이의 은닉층으로 구성되어 있다. 입력층은 입력변수를 입력하는 노드로 구성되어 있으며 은닉층은 입력층으로 전달된 정보를 결합해서 출력층에 전달하거나 다른 은닉층에 정보를 전달하는 층이다[23]. 그리고 출력층 노드에서는 입력변수 입력값에 대한 예측결과를 보여준다[23]. 인공신경망의 실행은 순전파와 역전파 과정으로 구분된다. 각 노드의 입력 값에 연결강도(가중치)를 곱하고 여기서 산출된 값을 다시 활성화함수에 적용하여 출력값을 결정하는 과정을 순전파(Forward Propagation)라 한다. 그리고 순전파를 통해 결정된 출력값과 목표값을 비교하여 오차를 계산한 후 오차를 역으로 전파하여 연결강도를 조정하는 과정을 오차 역전파(Back-Propagation)라 한다. 이와 같은 과정을 반복 수행하여 출력층 오차가 목표값에 근접하는 연결강도로 결정하는 알고리즘이 많이 사용된다[24].

4. 연구결과

4-1. MLR 모델

2.3절에서 확인한 바와 같이 폭발위험범위(Y)를 출력변수로 설정하고 분자량(X1), 폭발하한계(X2), 대기온도(X3), 운전압력(X4), 운전온도(X5), 누출공 크기(X6), 배출계수(X7), 안전계수(X8), 환

기속도(X9), 가스밀도(X10), 비열비(X11), 압축인자(X12)를 입력변수로 설정하여 MLR 모델 분석을 실시하였다.

분산분석 결과 회귀모델의 F-value는 147.873, p-value는 0.000 ($p<0.05$)으로 나타났으며, 출력변수와 입력변수의 상관계수는 0.867로 높은 상관관계를 보였다. 결정계수(R^2)는 0.746으로서 출력변수의 변동에 대해 MLR 모델을 통해 74.6%를 설명할 수 있음을 보여주었고 평균제곱근오차(RMSE:Root Mean Square Errors)는 1.389 m, 평균절대퍼센트오차(MAPE:Mean Absolute Percentage Errors)는 44.2%의 예측성능을 보였다.

4-2. PCR 모델

입력변수 사이에 강한 상관성이 존재할 경우, 회귀모델의 일반화 성능이 매우 제한될 수 있어 다중공선성 여부를 검증하여야 한다. 구체적인 검증방법으로 변량의 팽창정도를 의미하는 분산팽창계수(V.I.F:Variance Inflation Factor)가 10 이상이 되거나 공차한계(Tolerance)가 0.1 이하이면 다중공선성이 있다고 판단할 수 있다[26]. MLR 모델 분석결과 Table 3에서는 분자량(X1)과 가스밀도(X10)가 VIF 10을 초과하고 공차한계가 0.1 이하임을 보여주었다. 이와 같은 MLR 모델의 다중공선성 문제를 해결하기 위해 상관성이 높은 입력변수 중 일부를 제거하여 회귀분석을 할 수도 있다. 그러나 변수 선정 및 제거과정이 다소 임의적이며 데이터 손실을 수반한다는 점에서 합리적인 공학 분석으로 보기 어렵다.

3.3절에서 설명한 바와 같이 주성분을 입력변수로 회귀분석을 한다면 데이터 손실을 최소화하면서도 주성분 간의 독립성으로 다중공선성을 해소할 수 있는 장점이 있다. 먼저 주성분 분석이 가능한지 아닌지를 확인하기 위해 KMO (Kaiser Meyer Olkin) 측도를 산출하고 Bartlett 구형성 검정을 하였다. Kaiser는 KMO 측도가 0.5이상이고, Bartlett 검증결과에서 $p<0.05$ 이면 주성분 분석에 적합하다고 제시하였다[27]. 본 연구에서는 KMO 측도가 0.579, 유의확률은 $p<0.000$으로 나타나 주성분 분석의 적합함을 보여주었다.

주성분 분석을 통해 12개의 입력변수를 이보다 적은 수의 주성분으로 차원을 축소하여 데이터를 압축하기 위해서 사용하여야 할 주성분의 개수를 결정하여야 한다. 이와 관련하여 Kaiser는 고유값이 1보다 큰 주성분을 선택할 것을 제시하였고[28], Jolliffe는 고유값이 0.7보다 큰 주성분과 누적기여율(Cumulative proportion) 80~90% 이상으로 주성분 개수를 결정할 것을 제안하였다[29]. 또한, Catell은 스크리 검정(Scree test)결과 Scree plot에서 고유값이 수평을 유지하기 전 단계로 주성분의 수를 선택할 것을 제시하였다[30]. 주성분 분석결과는 Table 4와 같다.

Table 4에서 보는 바와 같이 Scree plot 기준을 적용하면 주성분(Principal Component) 중 고유값이 큰 주성분 1~3을 선택하여야 하는데 누적기여율이 53.057%에 불과하여 주성분 회귀분석에서 손실되는 데이터가 많다. 주성분 1~5는 고유값이 1보다 크지만 누적기여율은 70.627%로 여전히 데이터 손실량이 많은 편이다. 주성분 6과 7은 고유값이 1 미만이지만 1에 거의 근접하는 수치를 보여주고 있으며, 주성분 1~7을 선택하면 누적기여율이 86.125%까지 높아져 데이터 손실을 최소화할 수 있다. 그러므로 본 연구에서는 PC 1~7까지 7개의 주성분을 입력변수로 정하여 회귀분석을 실시하였다.

Table 5의 분석결과 분산분석에서 주성분 회귀모델의 F-value는 178.094, p-value는 0.000($p<0.05$)으로 나타났으며, 출력변수와 입력변수의 상관계수는 0.823의 높은 수준의 상관관계를 보여주었다. 또한, 각 주성분의 VIF가 10 미만, 공차한계가 0.1 이상으로 나타나 주성분 간의 다중공선성이 없는 것으로 확인되었다. R^2 값은 0.674로 출력변수의 변동에 대해 PCR 모델을 통해 67.4%를 설명할 수 있음을 보여주었고 RMSE는 1.602 m, MAPE는 49.3%의 예측성능을 보였다.

4-3. ANN 모델

이번 연구에서는 12개 변수를 신경망의 입력변수로 적용하고 위험범위를 출력변수로 적용하였다. 입력층과 출력층 노드가 정해져 있는 상태에서 신경망을 학습시키기 위해 은닉층의 노드 수, 활성

Table 3. The results of MLR

	b	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
Coefficient	-0.017	0.380	-0.227	-0.035	0.314	0.093	0.350	0.107	-0.104	-0.007	-0.500	-0.321	-0.276
t-value	-0.824	0.737	-8.680	-0.733	11.621	3.585	16.667	5.103	-4.987	-0.338	-0.972	-5.779	-8.551
p-value (<0.05)	0.410	0.462	0.000	0.464	0.000	0.000	0.000	0.000	0.000	0.735	0.332	0.000	0.000
Tolerance	-	0.002	0.634	0.188	0.555	0.609	0.931	0.977	0.972	0.916	0.002	0.134	0.357
V.I.F.	-	637.553	1.578	5.314	1.801	1.643	1.074	1.023	1.028	1.092	635.654	7.438	2.798

Table 4. The results of principal component analysis

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Initial Eigenvalues	3.731	1.498	1.138	1.062	1.047	0.938	0.921	0.822	0.568	0.189	0.085	0.001
% Variance	31.094	12.479	9.484	8.848	8.722	7.820	7.679	6.848	4.737	1.572	0.710	0.007
% Cumulative	31.094	43.573	53.057	61.905	70.627	78.447	86.125	92.973	97.710	99.283	99.993	100.000

Table 5. The results of PCR

	b	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Coefficient	-0.014	0.172	0.299	-0.066	0.031	0.099	0.046	-0.138
t-value	-0.592	27.817	19.278	-3.286	1.441	4.553	1.759	-5.570
p-value (<0.05)	0.554	0.000	0.000	0.001	0.150	0.000	0.079	0.000
Tolerance	-	0.996	0.996	0.996	0.997	0.996	0.999	0.996
V.I.F.	-	1.004	1.004	1.004	1.003	1.004	1.001	1.004

화함수, 초기 학습률 등의 하이퍼 매개변수(Hyper parameter)를 설정하여야 한다. 하이퍼 매개변수들은 학습에 큰 영향을 미치지만, 연구자가 임의로 적용하는 값이므로 최적값이 어떤 것인지 단정할 수 없어 경험과 시행착오를 통해 결정된다.

하이퍼 매개변수 중 은닉층과 은닉노드의 개수는 인공신경망의 용량을 규정하며, 너무 많으면 과적합(Over-fitting), 너무 적으면 과소 적합(Under-fitting)이 발생할 수 있으므로 적절하게 선정되어야 한다. Hornik의 보편적 근사 정리(Universal approximation theorem)와 Cybenko의 정리(Cybenko's theorem)에서는 하나의 은닉층을 갖는 인공신경망은 충분한 수의 은닉노드가 주어진다면 임의의 연속인 다변수 함수를 원하는 정도의 정확도로 근사할 수 있음을 보여주었다[31,32]. 은닉층이 2개인 경우 은닉층이 1개인 경우에 비해 함수의 형태가 매끄럽지 못해서 목적함수를 최소화시키는 가중치를 찾기 어려우므로 은닉층을 여러 개 두지 않는다[23]. 그러나 1개의 은닉층에서 노드 수가 지나치게 많아지면 은닉층의 수를 2개로 하여 신경망을 작성하는 것도 바람직하다고 볼 수 있다[23]. 은닉층 노드 수와 관련하여 Stephen은 은닉층이 1개인 신경망에서 [(입력노드 수+1)×은닉노드 수+(은닉노드 수+1)×출력노드 수]의 가중치가 있으며, 경험적으로 가중치 개수의 10배에 해당하는 데이터가 필요하다고 하였다[33]. 또한 Bengio는 은닉층 노드 수는 입력층 노드 수보다 많은 것이 좋으며[34], Berry 등은 은닉층 노드 수는 입력층 노드 수의 2배 미만으로 설정되어야 한다고 하였다[35]. 이와 같이 기존의 연구결과를 종합해볼 때 2개 이하의 은닉층과 13~23개의 노드로 구성된 신경망 구조가 적절할 것으로 판단된다.

활성화 함수로 시그모이드 함수나 쌍곡탄젠트 함수를 적용했을 때, 출력층에서 멀어질수록 출력층의 오차가 제대로 반영되지 못하여 기울기(Gradient)가 점점 작아져 결국 0에 가까워지는 기울기 소멸(Gradient vanishing) 문제가 발생할 수 있다. 이는 여러 개의 은닉층으로 구성된 심층신경망에서 역전파 알고리즘으로 학습하는 과정에서 발생하며, 이를 방지하기 위해 최근의 딥러닝 학습에서는 활성화함수로 정류기저함수(ReLU)를 많이 적용하고 있다[36]. 그러나 본 연구에서는 1~2개 은닉층으로 구성된 얕은 신경망 구조를 채택하여 기울기 소멸 발생 가능성이 매우 작으므로, 은닉층에 쌍곡탄젠트 함수, 출력층에 시그모이드 함수를 활성화 함수로 적용하였다.

역전파 과정 중 연결강도를 최적화하는 방법으로 기울기 하강법(Gradient descent)을 적용하는데, 여기서 학습률을 통해 연결강도의 조절 정도를 적절히 조정하여야 한다. 만약 학습률이 너무 높으면 빠르게 학습이 진행될 수 있지만 최적값을 찾지 못하여 학습되지 않거나, 학습률이 너무 낮으면 연결강도 변화량이 미세하여 학습시간이 매우 길어지기 때문이다. 학습률에 관해서는 기존의 연구에서 구체적인 범위를 제시한 사례가 있는데 Bengio는 0.01~1[34]을 제시하였고, Lee 등은 0.05~0.75[23], Stephen은 0.1~0.4[33]를 제시하였다. 본 연구에서는 사례 연구 중에서 학습률의 편차가 제일 작은 Stephen[33]의 연구 결과를 반영하였으며 0.1에서 0.4의 중간값인 0.25를 반영하였다. 또한, 모멘텀은 경사하강법에서 국소최적해를 피하고 오차함수(Error Faction)의 수렴을 도와주는 역할을 하며 0~1 사이의 상수인데, Oh는 0.5, 0.9, 0.99를 제시하였고[36] Stephen은 0.9를 제시하였다[33]. 본 연구에서는 Stephen[33]의 연구 결과를 반영하여 모멘텀은 0.9를 사용하였다.

신경망의 학습방식에는 배치, 온라인, 미니배치가 있으며, 배치는

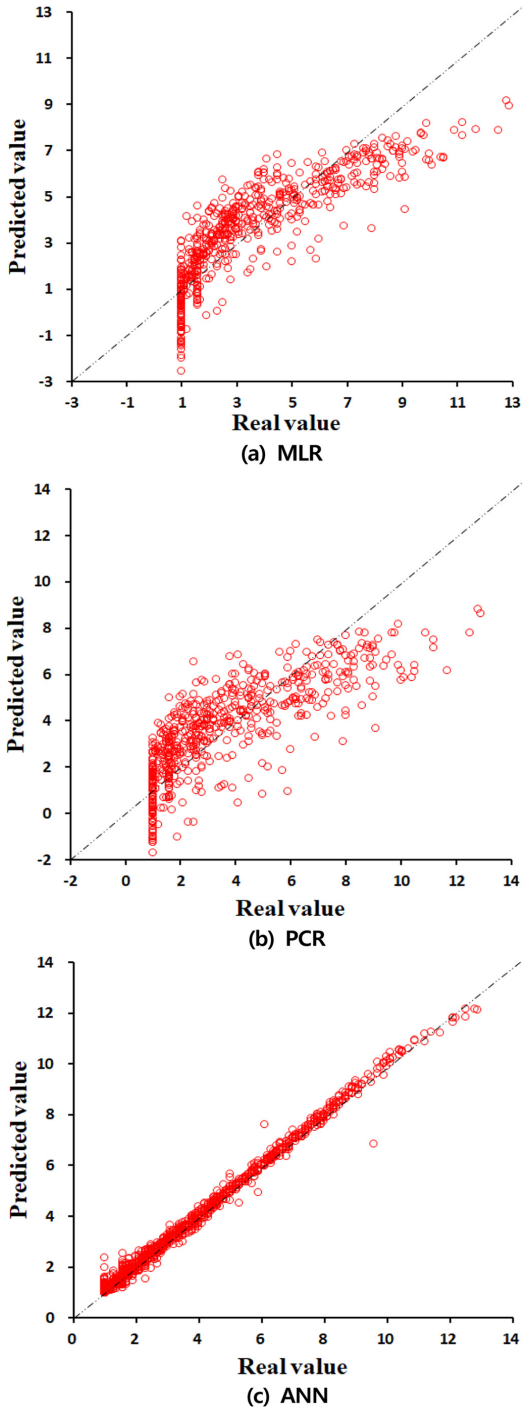

Fig. 4. The Comparison between the predictive and real values of hazardous extent.

학습데이터 세트를 메모리에 한꺼번에 올리는 방식이고 온라인은 학습데이터 행렬에서 하나의 행을 훈련할 때마다 넣으면서 학습을 하고 그때마다 신경망을 조정하여 업데이트하는 방식이다. 미니배치는 배치와 온라인을 결합한 방식으로써 다른 방식보다 신경망의 일반화 능력을 향상시켜 주며, 최근의 기계학습에서는 미니배치를 표준으로 여겨 널리 사용하고 있다[36]. 본 연구에서는 미니배치 방식을 적용하였다.

신경망 구조에서 은닉층과 노드 수를 변화시키면서 성능평가를 시행하여 RMSE와 MAPE가 가장 작게 나오는 구조의 신경망을 폭

Table 6. Decision of ANN structure

Number of Hidden layer	Number of Nodes	RMSE (m)	MAPE (%)	R^2 (%)
1	13	0.465	14.1	96.8
	16	0.434	12.6	97.2
	19	0.454	13.2	98.3
	23	0.212	6.6	99.3
2	13	0.203	5.7	99.4
	16	0.364	10.2	98.0
	19	0.441	12.5	97.1
	23	0.354	8.7	98.1

Table 7. Comparison of predictive performance

	MLR	PCR	ANN
RMSE (m)	1.389	1.602	0.203
MAPE (%)	44.2	49.3	5.7
R^2 (%)	74.6	67.4	99.4

발위험범위 예측 모델링에 적용하였다.

성능평가 결과는 Table 6에서 보는 바와 같으며, 2개의 은닉층에 각 은닉층 마다 13개의 노드로 구성된 신경망에서 가장 좋은 예측 성능을 보여주었다. 하나의 은닉층에서 노드 수가 증가할수록 예측 성능이 좋아졌으며, 2개의 은닉층에서는 노드 수가 증가할수록 예측 성능이 저하되는 경향을 보여주었다. 이는 신경망 구조가 복잡해질수록 과적합으로 인해 모델의 일반화 성능이 저하되는 것이 주요 원인이라 판단된다.

4-4. 예측 모델별 성능 비교분석

폭발위험범위 예측 성능평가 결과를 Table 7에 나타내었다. 모델의 정확도 평가를 위한 MAPE는 ANN 모델에서 가장 적게 나타났고 R^2 값은 ANN 모델에서 가장 높게 나타났다. 특히 ANN 모델에서는 이번 연구에서 수행한 모든 신경망 구조에서 96%를 초과하는 R^2 값을 보여주어 MLR 모델과 PCR 모델보다 훨씬 높은 수준의 예측 성능을 나타냈다.

5. 결 론

이번 논문에서는 폭발위험장소 내 전기 방폭설비 설치범위 합리화를 위한 가스폭발위험범위 예측모델 최적화 연구를 수행하였다. 이를 위해 화학 공정에서 많이 사용하는 12개의 가연성가스에 관해 폭발위험범위 사례연구를 하여 1,200개의 연구데이터를 생성하였고 위험범위를 출력변수로, 데이터 생성과정에서 반영한 12개 입력수치를 입력변수로 선정하였다.

본 연구에서는 3가지 회귀모델을 이용하여 모델을 수립하였다. 다중 회귀분석(MLR)을 적용하여 모델링을 한 결과, R^2 값은 74.6%, RMSE는 1.389 m, MAPE는 44.2%의 예측 성능을 보였다. 주성분 회귀분석(PCR)을 이용하여 모델링을 한 결과, R^2 값은 67.4%, RMSE는 1.602 m, MAPE는 49.3%를 나타내었고, 그 다음으로 ANN을 이용하여 모델링을 한 결과, R^2 값은 99.4%, RMSE는 0.203 m, MAPE는 5.7%를 보여주었다.

이를 통해 ANN이 가장 우수한 성능을 보여주고 있음을 확인하였다. 이러한 원인으로는 MLR의 경우 모델을 학습할 때 일부 입력변수 간에 다중공선성의 가능성도 존재하기 때문이며 PCR의 경우는 데이터 차원을 축소하는 과정에서 데이터 손실이 발생하기 때문에 정확도가 떨어질 수 있다는 문제점이 있다.

이번 연구결과를 통해 ANN을 이용한 폭발위험범위 예측모델이 정확도가 높은 최적모델이라는 것을 확인하였다. 본 연구에서 제안한 모델은 가연성가스를 사용하는 화학 공장 및 제조업 사업장에서 폭발위험범위를 산정하는 경우 쉽고 효율적으로 사용할 수 있다. 이를 통해, 폭발위험범위 산정에 필요한 시간과 비용을 크게 저감할 수 있을 것으로 기대한다.

References

1. Lee, H. S. and Yim, J. P., "A Study on Prevention Measure Establishment Through Cause Analysis of Chemical Accidents," *Journal of the Korean Society of Safety*, **32**(3), 21-27(2017).
2. Lee, K. O., Park, J. Y. and Lee, C. J., "Evaluation of a Mitigation System for Leakage Accidents Using Mathematical Modeling," *The Korean Institute of Chemical Engineers*, **35**(2), 348-354(2018).
3. Crowl, D. A. and Louvar, J. F., Chemical Safety Process : Fundamentals with applications, 3rd ed., *Prentice Hall*, USA(2011).
4. Byun, J. H., Lee, S. J. and Jeong, K, H., "A Study on the Technical and Institutional Management Plan to Maintain Explosion Proof Equipment and Apparatus," 2019-OSHRI-1648, *Occupational Safety &Health Research Institute*, Korea(2019).
5. KS C IEC 60079-10-1 : "Explosive atmospheres Part 10-1 : Classification of Areas-Explosive Gas Atmospheres," *Korean Industrial Standards*, Korea(2017).
6. Jung, Y. J. and Lee, C. J., "A Study on Gas Explosion Hazardous Ranges for International Electrotechnical Commission Technical Standards," *Journal of the Korean Society of Safety*, **33**(3), 39-45 (2018).
7. Choi, J. Y., "An Analysis on the Main Amendment of Hazardous Area Classification in Korea and a Study on Its Limitation," *Korean Journal of Hazardous Materials*, **6**(1), 8-17(2018).
8. Bozek, A., "Application of IEC 60079-10-1 Edition 2.0 For Hazardous Area Classification," *Petroleum and Chemical Industry Technical Conference*, September, Calgary, AB(2017).
9. Souza, A. O. and Luiz, A. M., "CFD Predictions for Hazardous Area Classification," *Chinese Journal of Chemical Engineering*, **27**(1), 21-31(2019).
10. Miranda, J. T. and Camacho, E. M., "Comparative Study of the Methodologies Based on Standard UNE 60079/10/1 and Computational Fluid Dynamics (CFD) to Determine Zonal Reach of Gas-generated Atex Explosive Atmospheres," *Journal of Loss Preven-*

tion in the Process Industries, **26**, 839-850(2013).
11. Shrivastava, V. and Mohan, G., "An Innovative Approach to Hazardous Area Classification-Three Dimentional (3D) Modelling of Hazardous Area," *Petroleum and Chemical Industry Technical Conference*, September, Philadelphia(2016).
12. Jung, Y. J. and Lee, C. J., "A Study on the Estimation Model of Liquid Evaporation Rate for Classification of Flammable Liquid Explosion Hazardous Area," *Journal of the Korean Society of Safety*, **33**(4), 21-29(2018).
13. Zohdirad, H., Ebadi, T. and Givehchi, S., "Predictive Modeling of Hazard Radius for Refinery Predictive Modeling of Hazard Radius for Refinery," *International Journal of Hydrogen Energy*, **41**(26), 11491-11496(2016).
14. Cho, K. W., Kang, C. G. and Oh, C. H., "Conformity Assessment of Machine Learning Algorithm for Particulate Matter Prediction," *Journal of the Korea Institute of Information and Communication Engineering*, **23**(1), 20-26(2019).
15. Park, J. Y. and Lee, C. J., "Principal Component Analysis Based Method for Effective Fault Diagnosis," *Journal of the Korean Society of Safety*, **29**(4), 73-77(2014).
16. Lee, C. J., Song, S. O. and Yoon, E. S., "The Monitoring of Chemical Process using The Support Vector Machine," *Korean Chemical Engineering Research*, **42**(5), 538-544(2004).
17. Park, J. H., Shin, S. W. and Kim, S. Y., "Traffic Volume Dependent Displacement Estimation Model for Gwangan Bridge Using Monitoring Big Data," *Journal of the Korean Society of Civil Engineers*, **38**(2), 183-191(2018).
18. Korea Petrochemical Industry Association, Petrochemical Handbook, Seoul(2019).
19. Ministry of Environment, Chemical Statistics Survey, Sejong(2014).
20. http://www.kgs.or.kr/kgsmain/SearchAction.do?method=mat1List&windowId=0305000101.html.
21. https://cameochemicals.noaa.gov/search/simple.html.
22. Yang, W. S. and Park, H. M., "Improving Polynomial Regression Using Principal Components Regression With the Example of the Numerical Inversion of Probability Generating Function," *Journal of the Korea Contents Association*, **15**(1), 475-481(2015).
23. Lee, T. R., Koo, J. Y., Park, H. J., Lee, K. H. and Choi, D. W., Data mining, *Korea National Open University*, Seoul(2004).
24. Park, J. H., Shin, S. W., and Kim, S. Y., "Modeling on Expansion Behavior of Gwangan Bridge using Machine Learning Techniques and Structural Monitoring Data," *Journal of the Korean Society of Safety*, **33**(6), 42-49(2018).
25. Oh, C. S., Artificial Neural Networks for Deep Learnig, *Naeha Inc.*, Korea(2016).
26. Kwon, S. H., Lee, J. W. and Chung, G. H., "Snow Damages Estimation using Artificial Neural Network and Multiple Regression Analysis," *Journal of the Korean Society of Hazard Mitigation*, **17**(2), 315-325(2017).
27. Kaiser, H. F., "An Index of Factorial Simplicity," *Psychometrika*, **39**(1), 31-36(1974).
28. Kaiser, H. F., "The Application of Electronic Computers to Factor Analysis," *Educational and Psychological Measurement*, **20**, 141-151(1960).
29. Jolliffe, I. T., "Discarding Variables in a Principal Component Analysis. I: Artificial Data," *Journal of the Royal Statistical Society Series C*, Royal Statistical Society, **21**(2), 160-173(1972).
30. Catell, R. R., "The Scree Test for Number of Factors," *Multivariate Behavioral Research*, **1**, 245-276(1966).
31. Hornik, K., "Multilayer Feedforward Networks are Universal Approximators," *Neural Networks*, **2**, 359-366(1989).
32. Cybenko, G., "Approximation by Superpositions of a Sigmoidal Function," *Mathmatics of Control, Signals and Systems*, **2**, 303-314(1989).
33. Stephen, M., Machine Learning:An Algorithmic Perspective, 2nd ed., *Jpub Inc*., Korea(2018).
34. Bengio, Y., "Practical Recommendations for Gradient-Based Training of Deep Architectures," Neural Networks:Tricks of the Trade 2nd ed., 437-478, *Springer*, Heidelberg(2012).
35. Berry, M. J. A. and Linoff, G., Data mining technique, *Wiley*, New York(1997).
36. Oh, I. S., Machine Learning, *Hanbit Academy Inc.*, Seoul(2017).