

Data Dependency in Machine Learning Models for Predicting Saturation Magnetic Flux Density of Fe-based Soft Magnetic Metallic Glasses

Chunghee Nam*

Department of Electrical and Electronic Engineering, Hannam University, Daejeon 34430, Republic of Korea

(Received 12 January 2026, Received in final form 6 February 2026, Accepted 6 February 2026)

This study developed machine learning models to accurately predict the saturation magnetic flux density (B_s), a critical property of Fe-based metallic glass soft magnetic materials and conducted an in-depth analysis of how input variable combinations and data size affect model performance. A total of 197 experimental data points were used to train Random Forest and Gradient Boosting Regressor models. The models were evaluated under three input variable conditions: compositional atomic ratio, feature descriptors derived from the Python Matminer module, and a combination of both. The results indicated that using only the atomic ratio as the input variable yielded the best prediction performance, outperforming conditions that included complex derived variables. Using Matminer features either alone or in combination with atomic ratios resulted in decreased prediction accuracy. This suggests that compositional information plays the most dominant role in determining the B_s of Fe-based metallic glasses. Consequently, it appears that certain variables extracted via Matminer acted as noise during the training process, diluting the significance of the compositional ratios and leading to performance degradation. Furthermore, to assess data dependency, experiments were conducted by increasing the test data ratio to 0.2, 0.5, and 0.8. The results confirmed that the coefficient of determination (R^2) for the test data significantly declined as the amount of training data decreased. This implies that the absolute volume of data is a critical factor in ensuring the generalization performance of machine learning models. In conclusion, developing high-performance Fe-based soft magnetic materials via machine learning requires a composition-centric feature selection strategy, alongside the essential expansion of the database through continuous experimental data collection.

Keywords : Fe-based metallic glasses, soft magnet, machine learning, feature engineering

철기반 금속유리 연자성 소재의 포화자속밀도 예측을 위한 기계학습 모델에서의 데이터 의존성

남충희*

한남대학교 전기전자공학과, 대전광역시 대덕구 한남로 70, 34430

(2026년 1월 12일 받음, 2026년 2월 6일 최종수정본 받음, 2026년 2월 6일 게재확정)

본 연구는 철기반 금속유리 연자성 소재의 핵심 물성인 포화자속밀도(B_s)를 정밀하게 예측하기 위해 기계학습 모델을 구축하고, 입력 변수의 조합과 데이터 규모가 모델 성능에 미치는 영향을 심층적으로 분석하였다. 총 197개의 실험 데이터를 활용하여 Random Forest와 Gradient Boosting Regressor 모델을 학습시켰으며, 입력 변수는 성분 원자비(compositional atomic ratio), 파이썬 모듈 Matminer 기반 특성인자, 그리고 이 두 가지를 결합한 세 가지 조건으로 분류하여 비교 평가하였다. 연구 결과, 복잡한 파생 변수를 포함하는 조건보다 원자비만을 단독으로 입력 변수로 사용했을 때 가장 우수한 예측 성능을 나타내었다. Matminer 특성인자를 단독으로 사용하거나 원자비와 결합하여 사용할 경우 오히려 예측 정확도가 감소하는 경향을 보였는데, 이는 금속유리의 B_s 결정에 있어 조성 정보가 가장 지배적인 역할을 하기 때문으로 분석된다. 즉, Matminer를 통해 추출된 일부 변수들이 모델의 학습 과정에서 불필요한 정보로 작용하여 조성 성분비의 중요도를 희석시키고 성능 저하를 유발하는 것으로 판단된다. 또한, 데이터 의존성을 확인하기 위해 테스트 데이터의 비율을 0.2, 0.5, 0.8로 증가시키며 실험한 결과, 학습에 사용되는 데이터의 양이 줄어들수록 테스트 데이터에 대한 결정계수가 뚜렷하게 감소함을 확인하였다. 이는 기계학습 모델의 일반화 성능 확보를 위해 데이터의 절대적인 수량이 결정적인 요소임을 시사한다. 결론적으로, 기계학습을 통한 고성능 철기반 연자성 소재 개발을 위

해서는 조성 중심의 효율적인 특징 선택(Feature Selection) 전략과 더불어, 지속적인 실험 데이터 수집을 통한 데이터베이스 확충이 필수적이다.

주제어 : 철기반 금속유리, 연자성, 기계학습, 특성인자

I. 서 론

철-기반의 연자성 금속유리 소재(Fe-based metallic glass soft magnets)는 Inoue 박사에 의해서 소개된 이후로 비정질 특성에 의해서 나타나는 우수한 연자성 특성과 내마모성과 같은 우수한 기계적 물성으로 인해서 연구가 활발히 진행되고 있다[1]. 철-기반 연자성 금속유리 소재가 연자성 소재로서의 응용성을 높이기 위해서 요구되는 주요한 특성으로는 포화자속밀도(saturation magnetic flux density: B_s)가 있다. 이 값은 규소강(silicon steel)의 특성과 유사하거나 능가하는 물성이 필요하다[2]. 높은 포화자속 밀도를 가지는 철-기반 금속유리 소재를 개발하기 위해서는 금속유리의 고유한 비정질적인 구조적 특성을 가져야 한다. 금속유리는 장거리 규칙성(long-range ordering)이 없기 때문에 결정 이방성이 없을 뿐 아니라, 결정에서 나타나는 격자의 결합 등이 존재하지 않기 때문에 자화반전 과정에서 자구벽의 이동을 방해하는 원인이 현저히 감소하여 보자력이 낮은 값을 갖는다[2]. 즉 아주 작은 외부 자장에 의해서도 쉽게 자화가 되므로 높은 투자율과 함께 보자력이 낮은 특성을 가지므로 연자성 소재로서 장점을 가지고 있다. 하지만 비정질의 금속유리 특성을 가지기 위해서는 소재 제작과정에서 급속냉각이 필요하기 때문에 일정 이상의 크기를 가지는 소재를 제작하기가 어렵다. 그러므로 우수한 금속유리 형성(glass forming ability: GFA) 원소를 사용하여 냉각속도를 줄여서 급속냉각과 같은 복잡한 공정을 해결하기 위한 조성 선택의 연구가 많이 진행되었다. 하지만 우수한 금속유리 형성 원소로 알려진 P, Si, C 등은 비자성 또는 반자성 특성을 가지며, 희토류 원소는 전이금속과 반강자성 결합을 하기 때문에 포화자속 밀도를 높이기에는 한계를 가진다[3,4]. 이러한 상충되는 특성은 신소재 개발과정에서 일반적으로 나타나는 현상이기 때문에 상충되는 특성에 대한 적절한 균형을 찾는 원소와 조성을 탐색하는 것은 중요한 해결책으로 남아 있다.

철-기반 금속유리 소재의 복잡한 조성을 설계하는 연구에서 데이터 기반 기계학습을 이용한 연구가 최근 많이 보고되고 있다. 예로서, 실험 데이터를 기반으로 하여 포화자속 밀도를 예측하는 기계학습 모델을 구축하여 새로운 소재를 탐색하는 방법이 소개되었다[3]. 또한 우수한 금속유리 형성과 관련된 특성을 예측하는 연구[5], 열-중성자 흡수 성능을 예측하는 연구[6], 설명가능한 기계학습 모델을 이용하여 포화자속 밀도 예측 및 새로운 소재 개발 등과 관련된 연구가 있

다[7]. 신소재 개발과정에서 기계학습을 이용하는 연구는 다방면에서 활발히 진행되고 있지만, 계산 데이터가 아닌 실험 데이터를 사용할 경우에 데이터 수가 많지 않기 때문에 기계학습 모델의 일반화 성능이 현저히 떨어지는 문제가 있다. 또한 적은 수의 데이터로 기계학습 모델을 적용할 경우 과적합(overfitting)의 문제 또한 해결해야 할 과제로 남아있다. 철-기반 금속유리는 비정질의 구조적 특성으로 인해서 밀도 범함수이론(Density Function Theory: DFT)과 같은 물리계산을 통한 물성을 확인하기 어렵기 때문에 실험결과에서 얻은 데이터가 중요하다.

본 연구에서는 실험데이터로 구성되어진 철-기반 금속유리의 화학식량과 목표값인 포화자속 밀도로 이루어진 197개의 소규모 데이터를 이용하여 목표값을 예측하는 기계학습 모델을 구축하였다[3]. 데이터 규모의 의존성과 철-기반 금속유리 합금의 성분기반 원소함량과 개별 원소특성에서 얻어지는 특성인자 계산 값들이 최종 예측 성능에 어떤 영향을 미치는지 각각 확인하였다. 첫째, 입력데이터로서 원소함량만 사용할 경우, 원소특성에서 얻어지는 특성인자를 사용할 경우, 두가지를 혼합하여 사용할 경우 예측성능이 어떻게 달라지는지 확인하였다. 둘째, 원소함량만 사용할 경우 가장 좋은 회귀 예측성능을 얻을 수 있었으며, 이를 이용하여 학습데이터와 테스트데이터의 비율을 8:2, 5:5, 2:8 로 나누어서 비교함으로써 학습데이터 규모의 중요성을 확인하였다.

II. 실험 방법

Fig. 1은 전체적인 연구방법을 개략적으로 보여주고 있다. 철-기반 금속유리 소재의 조성기반 원소함량(atomic ratio)은 0~1 사이 값으로 이루어져 있다. 합금에 포함된 원소는 27개이며 Table 1에 나타나 있다. 합금에 포함된 원소의 성분 비율로 계산한 특성인자는 파이썬 모듈 Matminer를 통해서 계산되었으며 5개의 특성은 원자번호(atomic number), 원자량(atomic weight), 녹는점(melting temperature), 전기음성도(electronegativity), 공유결합 반지름(covalent radius)으로서 Table 1에 나타나 있다[8,9]. 원소의 특성을 합금에 포함된 원소의 성분 비율로 계산할 때 평균값(mean)과 표준편차(standard deviation)를 각 원소특성을 바탕으로 계산하였으므로 총 10개의 특성인자를 입력데이터로 사용하였다. 다음으로 전체 197개 데이터는 학습데이터와 테스트 데이터를 8:2, 5:5, 2:8의 3가지 방법으로 나누어 학습데이터로 학습한 모델

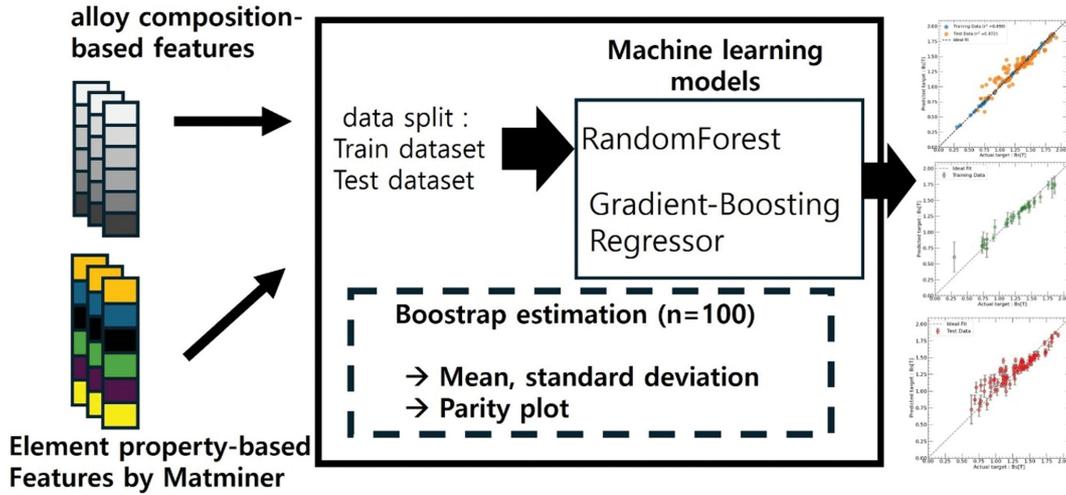


Fig. 1. (Color online) Schematic diagram of the machine learning framework employed in this study.

Table I. The constituent elements of the Fe-based metallic glass alloys used for data input and the element-based descriptors obtained by Matminer.

Elements in alloys	‘B’, ‘C’, ‘Al’, ‘Si’, ‘P’, ‘Cr’, ‘Fe’, ‘Co’, ‘Ni’, ‘Cu’, ‘Ga’, ‘Y’, ‘Zr’, ‘Nb’, ‘Mo’, ‘Sn’, ‘Sb’, ‘Pr’, ‘Nd’, ‘Sm’, ‘Gd’, ‘Tb’, ‘Dy’, ‘Er’, ‘Tm’, ‘Hf’, ‘Ta’
Element property	‘Atomic-Number’, ‘Atomic-Weight’, ‘Melting-temperature’, ‘Electronegativity’, ‘Covalent Radius’

에 대해서 테스트 데이터를 적용하여 성능지표 기반으로 예측성능을 비교하였다. 학습에 사용된 모델은 앙상블 모델로서 랜덤포레스트모델(RandomForest: R 모델)과 그래디언트 부스트 회귀모델(Gradient boosting regressor: G 모델)을 사용하였다. 각 모델의 하이퍼파라미터 튜닝은 그리드-탐색법(grid-search) 방법을 적용하여 최적화하였다. 입력데이터와 학습데이터 수가 달라질 때마다 각 모델의 최적화를 진행하였다. 학습된 모델을 통해서 테스트 데이터에 대한 평가를 통하여 결정계수(r^2 score) 성능지표를 통하여 예측성능을 비교하였다. 추가적으로, 데이터수가 많지 않기 때문에 부트스트랩(bootstrap sampling) 평가를 통해서 예측 성능에 대한 불확실성을 예측 평균과 표준편차를 통해서 확인하였다. 부트스트랩 방법에서 사용된 횟수는 100번이다.

III. 결과 및 토의

Fig. 2는 회귀예측 성능의 결과를 학습데이터와 테스트 데이터에 대해서 각각 보여준다. Fig. 2(a)~(c)는 R 모델에 대한 결과이며, Fig. 2(d)~(f)는 G 회귀모델에 대한 결과로서 학습데이터와 테스트 데이터 8:2로 분할한 후에 3가지 다른 입력데이터를 사용한 경우이다. Fig. 2(a), (d)는 합금원소의 함량만 입력데이터로 사용한 경우이며 테스트 데이터에 대한 예측성능인 결정계수 값은 R 모델은 0.933이며 G 모델은

0.935이다. Fig. 2(b), (e)는 원소특성을 바탕으로 계산한 특성인자를 사용한 것이며 R 모델의 결정계수는 0.756, G 모델은 0.710이다. Fig. 2(c), (f)는 두가지 모두를 입력데이터로 사용한 것으로 R 모델의 결정계수는 0.926, G 모델의 결정계수는 0.928 이다. 이를 통해서 특성인자가 많다고 해서 모델의 예측성능이 향상되는 것이 아님을 알 수 있다. 그러므로 적절한 특성인자를 선택하는 것이 중요하며, 중요성이 떨어지는 특성인자는 다른 주요 특성인자의 영향을 약화시키는 효과를 나타냄을 알 수 있다. 학습데이터에 대한 회귀 성능을 보면 R 모델에 비해서 G 모델이 과대적합이 있음을 알 수 있다. 그러므로 두 모델 가운데 일반화 성능을 고려하면 R 모델이 포화자속 밀도 예측에 더욱 적합함을 알 수 있다.

Fig. 3은 위에서 학습한 모델을 통해서 테스트 데이터에 대한 부트스트랩 평가의 결과를 보여준다. Fig. 2의 방법과 마찬가지로 Fig. 3(a), (d), 3(b), (e), 3(c), (f)는 특성인자를 각각 다르게 사용한 것이다. Matminer 파이썬 모듈을 통해서 원소의 특성으로 얻어진 10개의 특성인자를 입력데이터를 사용한 경우가 불분확실성이 가장 높다는 것을 Fig. 3(b), (e)를 통해서 알 수 있다. 주목할만한 점은 Fig. 3(a), (d)와 같이 원소의 함량만 사용한 것 보다 Fig. 3(c), (f)와 같이 두가지 특성인자를 함께 입력데이터로 사용한 경우가 비록 결정계수 값은 낮지만, 대부분의 테스트 데이터에 대해서 표준편차 값이 감소하는 것을 알 수 있다. 이것은 특성인자가 많이 사용

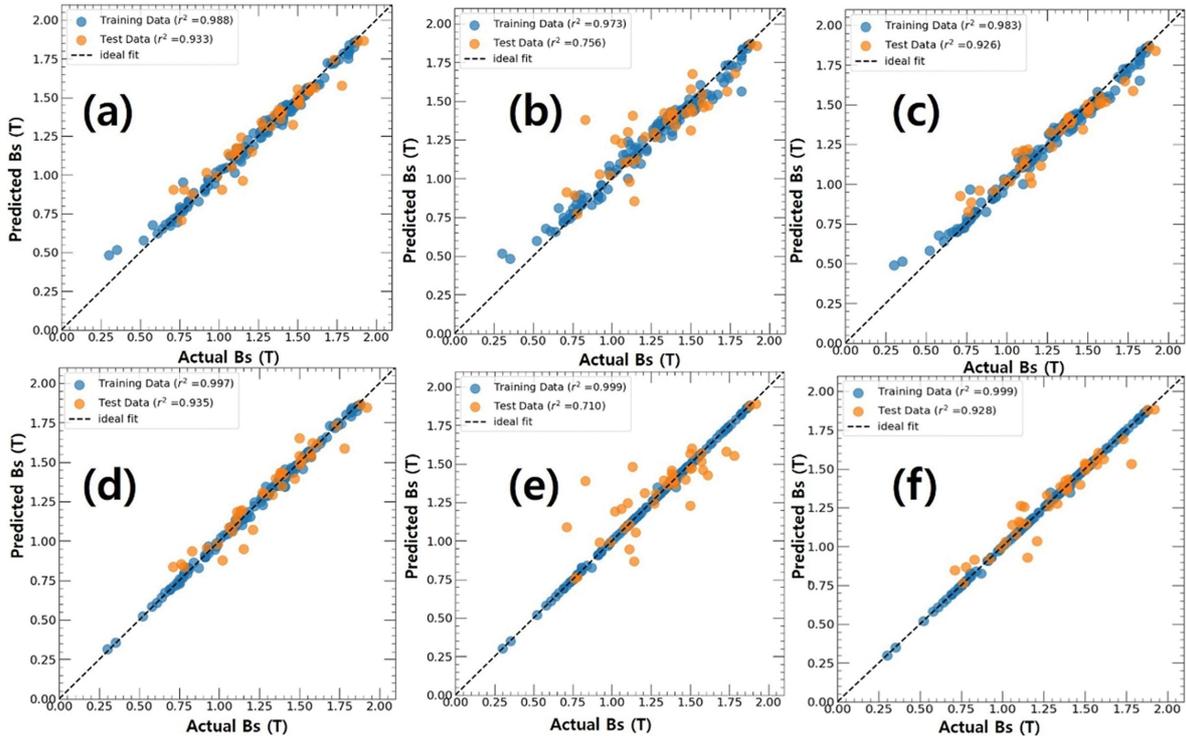


Fig. 2. (Color online) Parity plots of the actual vs predicted B_s values for (a)–(c) the RandomForest, (d)–(f) Gradient-boosting regressor models. The input data used for each model are as follows: (a), (d) Alloy composition-based atomic ratios; (b), (e) Elemental properties; (c), (f) Combined features.

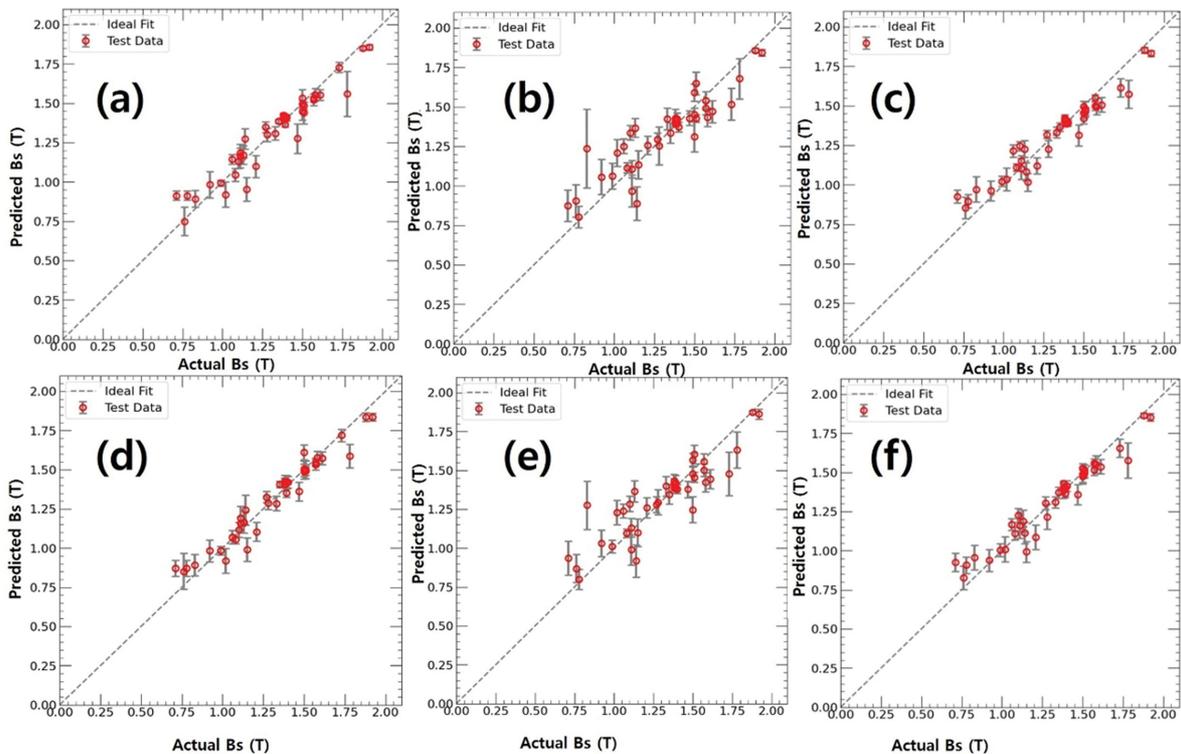


Fig. 3. (Color online) Parity plots of the bootstrap evaluation results on the test dataset. The models were trained using different input features: (a), (d) alloy composition-based atomic ratios, (b), (e) element property-based features derived from Matminer, and (c), (f) a combination of both feature sets. Error bars represent the standard deviation obtained from 100 bootstrap iterations.

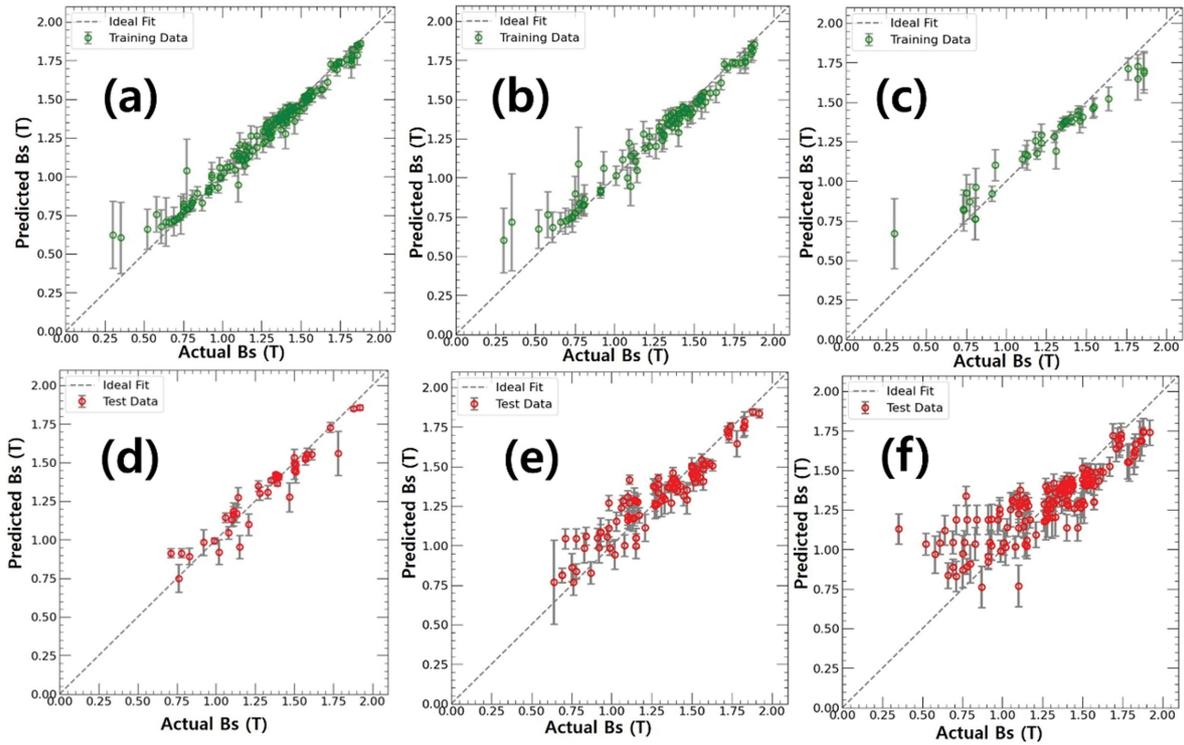


Fig. 4. (Color online) Parity plots showing the effect of data split ratios on prediction performance using the Random Forest model. The datasets were split into training (green, top row) and test (red, bottom row) sets with ratios of (a), (d) 8:2, (b), (e) 5:5, and (c), (f) 2:8. Error bars represent the uncertainty derived from bootstrap estimation.

될 경우 과대적합의 결과로 나타날 수 있음을 보여준다. 즉 입력데이터는 합금의 원소함량만을 사용하는 것이 포화자속 밀도 예측에 가장 적절함을 알 수 있다.

Fig. 4는 부트스트랩 평가를 통해서 학습데이터와 테스트 데이터의 분할 비율에 따른 데이터 수 규모에 의존하는 특성을 나타낸다. 앞서 확인한 것처럼 본 연구의 주어진 데이터에서는 R 모델과 원소함량만을 사용하는 것이 가장 적절함을 확인하였다. 그러므로 이러한 조건을 기준으로 학습데이터와 테스트 데이터의 분할 비율을 Fig. 4(a), (d)의 8:2, Fig. 4(b), (e)의 5:5, Fig. 4(c), (f)의 2:8로 나누어 각각의 경우에 대해서 회귀성능을 학습데이터((a)-(c))와 테스트 데이터((d)-(f))로 시각화하여 나타낸 것이다. 학습데이터가 적어질 수록(8:2 → 5:5 → 2:8) 회귀 예측 모델의 성능이 감소하기 때문에 테스트 데이터에 대한 평가결과 또한 크게 감소함을 Fig. 4를 통해서 확인할 수 있다. Fig. 4(a), (d)에서 보는 것처럼 8:2로 분할 된 경우에는 예측성능이 우수함을 알 수 있다. 하지만 5:5로 분할 된 경우, 포화자속 밀도가 낮은 데이터에 대해서 학습이 잘 이루어지지 않음을 알 수 있다. 그 결과로서 테스트 데이터에 대한 불확실성이 커지는 것을 Fig. 4(e)를 통해서 확인할 수 있다. Fig. 4(c)의 2:8로 분할 된 경우에는 포화자속 밀도가 0.5 T 이하인 데이터가 하나뿐 이

기 때문에 테스트 데이터에 대한 예측값이 실제값과 많이 차 이나는 것을 알 수 있으며, 불확실성도 높아지는 것을 Fig. 4(f)를 통해서 알 수 있다.

기계학습에서 사용하는 데이터규모는 예측성능 뿐 아니라 일반화 성능에도 큰 영향을 미친다. 이를 확인하기 위해서 t-SNE(t-distributed Stochastic Neighbor Embedding) 방법을 통해서 서로 다른 비율로 분할된 학습데이터와 테스트 데이터에 대해서 차원 축소된 공간에서의 데이터의 상관관계를 확인하였다[10]. Fig. 5는 전체 데이터 197개에 대해서 학습데이터와 테스트 데이터를 8:2, 5:5, 2:8로 분할하여 합금의 원소함량만을 사용한 입력데이터를 t-SNE 방법으로 차원 축소하여 목표값인 포화자속 밀도의 값으로 색을 입혀 표현한 것이다. 학습데이터는 닫힌 원형모양(closed circle symbol)로 표시하였으며, 테스트 데이터는 열린 원형모양(open circle symbol)로 나타내어 구분하였다. Fig. 5(a)에서 보는 것처럼 8:2로 분할된 경우에는 대부분의 학습데이터가 테스트 데이터의 조성공간에 겹쳐지기 때문에 예측성능이 높은 것으로 이해할 수 있다. 그러나, 학습데이터의 규모가 줄어들 경우 Fig. 5(b), (c)와 같이 학습되지 않은 조성 공간에 대한 테스트 데이터가 부분적으로 존재함을 확인할 수 있다. Fig. 5(b), (c)에서 붉은색 화살표로 표시된 영역은 학습되지 않은 조성공

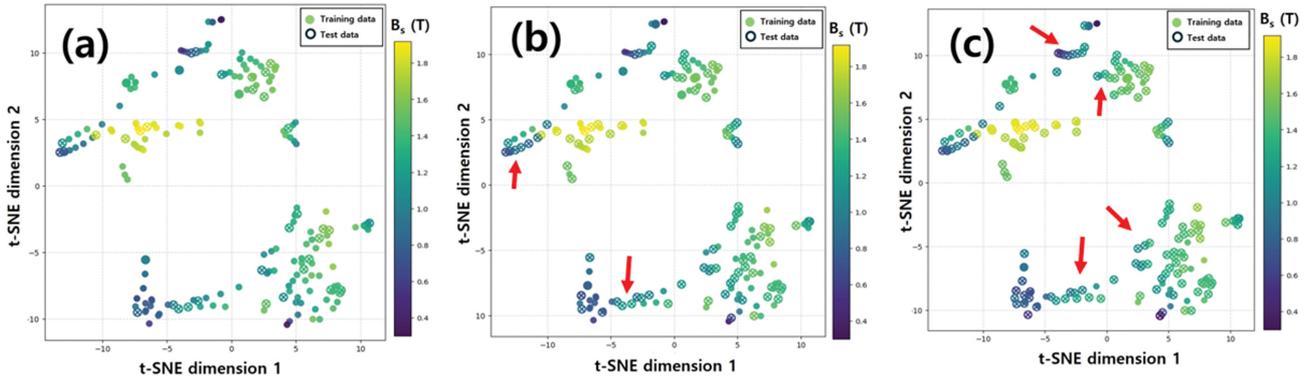


Fig. 5. (Color online) t-SNE visualization of the data distribution in the compositional space for different split ratios: (a) 8:2, (b) 5:5, and (c) 2:8. Closed and open circles represent training and test data, respectively, colored by saturation flux density (B_s). Red arrows indicate regions where the test data are in compositional spaces not covered by the training data, highlighting the lack of generalization with smaller training sets.

간에 테스트 데이터가 분포하고 있는 것을 나타낸다. Fig. 5(b)와 같이 5:5로 분할된 경우에는 두 군데 영역에서 학습되지 않은 테스트 데이터가 분포함을 알 수 있고, Fig. 5(c)와 같이 2:8로 분할된 경우 네 군데 이상이 학습되지 않은 조성 공간에 대한 테스트 데이터가 분포함을 알 수 있다. 즉 학습되지 않은 조성공간 영역에 테스트 데이터가 있을 경우 기계 학습 모델은 불확실성이 커질 수밖에 없으므로, 데이터 규모에 큰 영향을 받는 것을 확인할 수 있다.

IV. 결 론

본 연구는 197개의 철-기반 연자성 금속유리 실험 데이터를 기반으로 포화자속 밀도를 예측하는 기계학습 모델을 분석하였다. 랜덤 포레스트(R 모델)와 그래디언트 부스팅 모델(G 모델)을 비교한 결과, 조성기반 원소의 기본적 특성을 바탕으로 얻어진 복잡한 물리적 특성인자보다 원소 함량(atomic ratio)만을 입력 변수로 사용할 때 가장 높은 예측 성능을 보였다. 또한 차원축소법인 t-SNE 분석을 통해 학습 데이터 비중이 줄어들수록 학습되지 않은 조성 영역이 늘어나 예측 불확실성이 커짐을 확인하였으며, 일반화 성능 측면에서 랜덤 포레스트 모델이 더 적합함을 알 수 있었다.

감사의 글

이 논문은 2025 학년도 한남대학교 학술연구비 지원에 의하여 연구되었음.

References

- [1] A. Inoue, Y. Shinohara, and J. S. Gook, *Mater. Trans., JIM* **36**, 1427 (1995).
- [2] C. Suryanarayana and A. Inoue, *Int. Mater. Rev.* **58**, 131 (2013).
- [3] Z. Lu, X. Chen, X. Liu, D. Lin, Y. Wu, Y. Zhang, H. Wang, S. Jiang, H. Li, X. Wang, and Z. Lu, *npj Comput. Mater.* **6**, 187 (2020).
- [4] L. Shi, X. Hu, Y. Li, G. Yuan, and K. Yao, *Intermetallics* **131**, 107116 (2021).
- [5] X. Li, G. Shan, and C. H. Shek, *J. Mater. Sci. Technol.* **103**, 113 (2022).
- [6] J. Gao, J. Hou, Y. Wu, B. Ji, D. Wang, K. Qiu, J. You, and J. Wang, *J. Alloy. Compd.* **1010**, 177595 (2025).
- [7] Q. Bi, J. Liu, J. Li, H. Chen, S. Chen, and X. Han, *J. Alloy. Compd.* **1016**, 178858 (2025).
- [8] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, *Comput. Mater. Sci.* **68**, 314 (2013).
- [9] L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster, and A. Jain, *Comput. Mater. Sci.* **152**, 60 (2018).
- [10] R. Furushima, Y. Nakashima, Y. Maruyama, Y. Zhou, K. Hirao, T. Ohji, and M. Fukushima, *J. Am. Ceram. Soc.* **108**, e20173 (2025).