VII International Conference on Mathematics and its Applications in Science and Engineering (ICMASE 2026)

ICMASE 2026

SYNTHETIC TABULAR DATA GENERATION UNDER PHYSICAL CONSTRAINTS FOR PHASE CLASSIFICATION AND ALLOY RECOMMENDATION IN DATA-SCARCE HEA STUDIES

Xabier Ballesteros Vallejo[2,], Ignacio Fidalgo Astorquia[1*], Nerea Gómez-Larrakoetxea[1], Fernando Boto[1]

[1] *Department of Computing, Electronics and Communication Technologies, University of Deusto, Avenida de las Universidades 24, 48007 Bilbao, Spain*

[2] *DeustoTech-Deusto Institute of Technology, University of Deusto, Avenida de las Universidades 24, 48007 Bilbao, Spain*

ABSTRACT

We present a mathematically grounded workflow to alleviate data scarcity in High-Entropy Alloys (HEAs) by generating, filtering, and exploiting synthetic tabular data for downstream tasks in phase classification and candidate recommendation. The approach integrates four families of generators—CTGAN, TVAE, Gaussian copulas, and SMOTE—evaluated through a triple protocol that balances statistical fidelity (SDMetrics column-shape and pair-trend scores), predictive utility via TSTR (Train Synthetic, Test Real), and novelty/privacy through distance-to-closest-record (DCR). We cast the amount and mixture of synthetic data as an optimization problem, applying Bayesian search (Optuna) with early stopping to identify generator proportions that maximize weighted F1 over Random Forest and XGBoost while preserving plausibility.

At the sample level, we introduce an individual quality score combining a KDE-based likelihood (fidelity), outlier penalties (Isolation Forest), and a target "sensible novelty" term shaped on DCR. Percentile-based filtering shifts solutions toward a Pareto-efficient frontier between fidelity and predictive performance. Beyond augmentation, we deploy a TVAE-driven exploratory stage for alloy recommendation: generate a virtual composition library under compositional constraints, screen with a calibrated phase classifier, score by a target novelty function that favors moderate distance from the real support, and select diverse representatives via clustering. A lightweight thermodynamic post-filter using $\Delta Hmix$, $\Delta Smix$, δ, Ω, and VEC further prioritizes physically plausible candidates.

On a curated HEA dataset, the best configuration reaches a mean F1 of 0.814 in the full-data scenario and 0.777 when starting from 25% of real data. Under exploratory generation, 3,288 high-confidence BCC/FCC candidates are proposed, of which 1,555 pass the thermodynamic screen. Validation across external datasets indicates that gains are largest when real data are scarce and structurally coherent; improvements attenuate on noisier bibliographic corpora, underscoring data quality as a moderator of augmentation benefits. The contribution is a reproducible, optimization-driven pipeline that formalizes the augmentation–fidelity trade-off, integrates domain constraints, and supports early-stage discovery via diverse, plausibly novel compositions.

***K*eywords** Synthetic tabular data · CTGAN · TVAE · Gaussian copulas · SMOTE · Bayesian optimization · TSTR · DCR · HEAs · phase classification · alloy recommendation · thermodynamic screening

References

[1] Georganas, E., et al., On the Evaluation of Synthetic Tabular Data: Likelihoods, Distances, and Downstream Utility, Proc. NeurIPS Datasets and Benchmarks, 2023.

[*] *Corresponding Author's E-mail: abc@hbv.edu.tr*

[2] Kumar, I., et al., Evaluating and Characterizing Synthetic Tabular Data, arXiv:2109.11352, 2021.

[3] Miracle, D.B., Senkov, O.N., A critical review of high entropy alloys and related concepts, Acta Materialia, 122: 448–511, 2017.

[4] Wen, C., et al., Machine learning assisted design of high entropy alloys with desired property, Acta Materialia, 170: 109–117, 2019.

[5] Yang, X., Zhang, Y., Prediction of high-entropy stabilized solid-solution in multi-component alloys, Materials Chemistry and Physics, 132(2–3): 233–238, 2012.

[6] Xu, L., et al., Modeling Tabular Data using Conditional GAN, Advances in Neural Information Processing Systems (NeurIPS), 32, 2019.

[7] Chawla, N.V., et al., SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, 16: 321–357, 2002.

[8] Yale, A., et al., Generation and evaluation of privacy-preserving synthetic health data, Neurocomputing, 416: 244–255, 2020.

[9] Settles, B., Active Learning, Synthesis Lectures on Artificial Intelligence and Machine Learning, 6(1): 1–114, 2012.

[10] Akiba, T., et al., Optuna: A Next-generation Hyperparameter Optimization Framework, Proc. KDD, 2623–2631, 2019.