Korean Chem. Eng. Res., 64(2), 105163 (2026)
https://doi.org/10.9713/kcer.2026.64.2.105163
PISSN 0304-128X, EISSN 2233-9558

Machine Learning for Deep Eutectic Solvent Density: Impact of Feature Representations and Dataset Complexity on Predictive Reliability

YoonKook Park*

Department of Biological and Chemical Engineering, Hongik University, Sejong, 30016, Korea
(Received 1 March 2026; Received in revised from 6 April 2026; Accepted 10 April 2026)

Abstract – Accurately predicting the density of deep eutectic solvents (DESs) is crucial for optimizing green separation processes. This study investigated the impact of feature representations (ChemBERTa, hybrid, and critical property models) and data partitioning (binary, ternary, and comprehensive datasets) on machine learning predictions. Evaluating RF, XGBoost, CatBoost, and ANN models revealed that tree-based ensembles were highly robust, consistently achieving $R > 0.93$ on limited datasets. Conversely, ANNs required explicit physical descriptors or massive datasets (>12,000 points) to prevent overfitting. Cross-domain validations demonstrated that extrapolating from simple to complex systems failed due to restricted thermodynamic diversity, whereas specializing from a comprehensive dataset ensured excellent transferability. These findings established that combining large, diverse datasets with ensemble algorithms or physics-informed features were essential for the reliable computational design of multicomponent DES properties.

Key words: Deep eutectic solvent, Density, Machine learning, Ensemble, Feature

1. Introduction

The environmental damage caused by the widespread use of toxic conventional solvents has driven a search for greener alternatives. Ionic liquids were once heralded for this role because of their favorable properties, but concerns about biodegradability, toxicity, and costly, complex synthesis have limited their practical uptake [1]. Deep eutectic solvents (DESs) have since emerged as a compelling substitute: they are typically cheaper, simpler to prepare, and more versatile than many conventional solvents and ionic liquids [2]. Formed from hydrogen bond acceptors (HBAs) and hydrogen bond donors (HBDs), DESs can be tuned through component selection and molar ratio, enabling applications in catalysis, extraction [3-5], and CO_2 capture [6].

Moving beyond binary mixtures, ternary DESs – composed of three constituents – provide additional levers for solvent design [7]. Adding a third constituent or introducing a cosolvent expands the tunable space for polarity, hydrogen-bonding networks, and other solvent attributes [8]. These adjustments can produce meaningful performance gains: for instance, ternary DESs have been shown to remove lignin more effectively than binary systems, demonstrating promise for biomass processing [9]. Likewise, cosolvents such as ethanol can markedly increase the solubility of otherwise poorly soluble species, extending the practical reach of DES formulations [10].

Accurate physical property data are essential for designing and optimizing DES-based separations because solubility and selectivity ultimately govern extraction efficiency. Density, while not a direct measure of solubility, strongly influences phase behavior and solute partitioning and thus affects purification outcomes and process design [11]. When considered alongside viscosity and hydrogen bonding interactions, density can substantially improve biphasic extraction performance [12]. Researchers are increasingly turning to computational approaches because the immense combinatorial space of HBA-HBD pairs and molar ratios renders exhaustive experimental characterization unfeasible. Methods ranging from empirical correlation [13] and group contribution [14] to quantitative structure-property relationship (QSPR) [15] and machine learning (ML) are now widely deployed to predict density and other essential properties, thereby expediting the discovery of novel DESs.

A range of computational strategies have been applied to predict DES properties. Classical ML methods such as support vector regression (SVR) [16] and random forest (RF) [17] have provided reliable baselines. Artificial neural network (ANN) capture complex nonlinear relationships but can be less interpretable and computationally demanding [16]. More recently, tree-based ensembles methods like XGBoost [18] and CatBoost [19] have gained favor for chemical datasets because they balance speed, accuracy, and scalability. Hybrid feature strategies that combine ChemBERTa's structural embeddings [20] with explicit process variables (*e.g.*, temperature, molar ratio) have further advanced predictive accuracy [11,21], while physics-informed models, such as the critical property-based model (CPM) leverage thermodynamic principles to reduce average absolute relative deviation (AARD) relative to purely data-driven approaches [17].

However, while numerous machine learning models have been

† To whom correspondence should be addressed.
E-mail: parky@hongik.ac.kr

proposed to predict the DES density, vast majority of these existing studies relied exclusively on conventional 1D/2D molecular descriptors and were typically confined to limited subsets of simple binary mixtures. Consequently, the transferability of these models to more complex, multicomponent systems—such as ternary DESs—remained largely unexplored in the current literature. Furthermore, there was a distinct lack of systematic investigation into how fundamentally different feature representation strategies—ranging from advanced natural language processing embeddings (*e.g.*, ChemBERTa) to physics-informed macroscopic properties (*e.g.*, critical property models) impacted predictive robustness across varying chemical domains.

This study compared two feature representations - ChemBERTa embeddings and hybrid features including temperature and mole ratio – across multiple algorithms (RF, XGBoost, CatBoost, ANN) to evaluate their influence on DES density prediction. Datasets were organized into three groups: binary DESs (D_2), ternary DESs and binary DESs with cosolvent (D_3) and a comprehensive collection of over 12,000 density records (D_{total}). By systematically evaluating model performance across these partitions and feature sets, author aimed to clarify the impact of data type and representation on predictive reliability and to inform the development of practical computational tools for DES property estimation.

2. Databank Description

All density measurements used for DES density prediction were taken at atmospheric pressure. The dataset contained two DES categories: two-component systems (D_2 in Table 1) comprising one hydrogen bond donor (HBD) and one hydrogen bond acceptor (HBA), three-component systems (D_3 in Table 1), which included either one HBA with two HBDs or two components plus a cosolvent (*e.g.*, ethanol) [10]. A total of 691 DES formulations, involving 324 different chemicals, were identified in the literature to construct density database used in this study. The mole fractions spanned wider ranges, with maximum values of 0.917 for HBAs, 0.950 for HBDs, and 0.9819 for cosolvents. The minimum mole fractions were 0.0021 (HBA), 0.0106 (HBD), and 0.016 (cosolvent).

DES can be classified into five categories, distinguished by the type of complexing agent employed, including organic salts, non-ionic compounds, metal halides, and their hydrates [22]. In the present study, the majority of systems (57%) belong to Type III, while approximately 25% fall under Type V. Nevertheless, representatives of Types I, II, and IV were also incorporated into the dataset.

Table 1. Comparison of the dataset used in this study

Dataset	Number of data points	Temperature Range (K)	Density range ($g \cdot cm^{-3}$)
D_2	1,808	283.15-363.15	0.806-1.7012
D_3	1,792	283.15-363.15	0.80358-1.988
D_{total}	12,041	278.15-413.15	0.80358-1.988

2-1. Data preparation

A comprehensive literature search yielded over 12,000 data entries (D_{total} in Table 1) spanning binary and ternary DES systems. The dataset included HBAs such as choline chloride, betaine, tetraalkylammonium halides, trimethylphosphonium bromide, and zinc chloride, and HBDs including alcohols, diols, and carboxylic acids, sampled at various molar ratios across a temperature range of 278.15 K to 413.15 K under atmospheric pressure. Although water in prepared DESs can influence density, residual water content was neglected unless its molar fraction exceeded 0.047. When multiple density values were reported for the same DES at identical temperature and molar ratios, the mean value was used.

To assess the influence of DES type on density, binary (D_2) and ternary (D_3) subsets were extracted from the full dataset (D_{total}); extraction required at least two components to match exactly between the ternary and binary systems and other conditions, such as, temperature, to be similar. Each dataset was then randomly split into training and test sets: approximately 80% of the data were used for model training and the remaining 20% for independent evaluation.

The dataset was systematically organized according to distinct experimental conditions. Within this study, 371 unique conditions were identified for DES composed of choline chloride and ethylene glycol, of which 27 conditions included multiple reported measurements. For example, at 293.15 K and 1:2 ratio of choline chloride to ethylene glycol, two density values have been reported: 1.1193 $g \times cm^{-3}$ [23] and 1.1171 $g \times cm^{-3}$ [24]. When such measurements exhibited close agreement – as in this case, where the discrepancy was less than 0.2% – the arithmetic mean was adopted. Conversely, when a measurement deviated substantially from the others (*e.g.*, 1.12 vs. 1.46 $g \times cm^{-3}$ under identical conditions), it was designated as a potential outlier. In such cases, the original source was reexamined to identify possible errors in the SMILES notation, temperature specification, or unit conversion. If the inconsistency cannot be resolved, the anomalous data point was excluded to ensure that the model was not trained on spurious or physically implausible information.

2-2. Machine learning models

Supervised learning methods predominated in DES density prediction. Feature representations employed in the literature included group contribution descriptors, COSMO-based features, and cheminformatics encodings generated by toolkits such as RDKit and pretrained language models like ChemBERTa. Typical supervised learning models encompassed a wide array of regression algorithms, among which ensemble learning approaches such as RF and XGBoost, alongside MLP, exhibited outstanding performance. Their success stemmed from a strong ability to capture nonlinear relationships within small- to medium-sized data sets. For instance, in a study conducted by Wu *et al.* [25] mainstream supervised learning algorithms were systematically evaluated using the LazyRegressor toolkit for predicting DES densities and melting points. The results demonstrated that XGBoost consistently outperformed other models across

Table 2. Method of training and test in each prediction

Input output	Binary (D_2)	Ternary (D_3)	Total (D_{total})
Binary (D_2)	80/20 split	Binary to Ternary	Binary to Total
Ternary (D_3)	Ternary to Binary	80/20 split	Ternary to Total
Total (D_{total})	Total to Binary	Total to Ternary	80/20 split

multiple evaluation metrics, underscoring the power of ensemble strategies.

2-3. Data partitioning and validation protocols

This study evaluated model performance across three distinct scenarios, as outlined in Table 2, to rigorously assess both internal consistency and cross-domain generalizability.

2-3-1. Internal validation (The diagonal)

The first scenario, represented by the diagonal entries in Table 2, assessed model performance within a single domain (*e.g.*, training and testing on D_2, D_3, and D_{total}). Following standard QSAR best practices recommended by Tropsha [26], these datasets were partitioned into an 80% training set and a 20% independent test set. This approach is typically applied to benchmark ML algorithms and feature sets under interpolation conditions.

2-3-2. Generalization (Upper-right triangle)

The second scenario focused on the model's ability to extrapolate from simpler systems to more complex ones. Here, a simpler dataset (*e.g.*, D_2 or D_3) served as the training source, while a more complex dataset (*e.g.*, D_3 or D_{total}) was used for testing. In accordance with Hawkins [27], who argued that a final model should be trained on all available data to maximize predictive power before external application, these models utilized 100% of the source dataset for training rather than retaining a validation split.

2-3-3. Specialization (Lower-left triangle)

The third scenario evaluated whether a model trained on a comprehensive, complex dataset (*e.g.*, D_{total}) can accurately specialize to simpler dataset (*e.g.*, D_2). This effectively tested if the broader physical laws learned from diverse mixtures hold true when restricted to specific sub-domains. Similar to the robustness checks performed by Mjalli [28], this strategy validated if the underlying physics captured by the model remained consistent across different chemical combinations and complexity levels.

2-4. Performance evaluation methods

To comprehensively evaluate the model performance, author employed multiple metrics: the coefficient of determination, R^2, root-mean-square error (RMSE), and AARD%. An R^2 value approaching 1 indicated stronger explanatory power, whereas RMSE values denoted greater prediction accuracy and precision. A smaller AARD% reflected reduced systematic bias in the predictions. Together, these metrics provided a balanced assessment of both fit and error, facilitating robust model selection and optimization [21,25]. The mathematical definitions were given in Eqs. (1) - (3), where $r_{exp,i}$ and $r_{pred,i}$ denoted the experimental and predicted values, respectively, r_{exp} was the mean of the experimental values, and N was the number of data.

$$R^2 = 1 - \frac{\sum \left(\rho_{exp,i} - \rho_{pred,i}\right)^2}{\sum \left(\rho_{exp,i} - \rho_{exp}\right)^2} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum \left(\rho_{exp,i} - \rho_{pred,i}\right)^2}{N}} \tag{2}$$

$$AARD(\%) = \frac{100}{N} \sum \frac{\left|\rho_{exp,i} - \rho_{pred,i}\right|}{\rho_{exp,i}} \tag{3}$$

3. Results and Discussion

3-1. Performances of the models with internal validation (The diagonal)

Table 3 shows that the model performance on internal validation with ChemBERTa, hybrid, and CPM features was generally excellent for all methods, though the ANN showed notable variance depending on the feature type. With pure ChemBERTa embeddings, the ANN

Table 3. Model performance on internal validation (diagonal)

Train → Test	Feature Set	Model	R^2	AARD (%)	RMSE	Absolute RD distribution (%)			
						< 1	1~5	5~10	> 10
$D_2 \rightarrow D_2$	ChemBERTa	CatBoost	0.9827	0.4597	0.0146	97.62	2.38	0.00	0.00
$D_2 \rightarrow D_2$	Hybrid	CatBoost	0.9801	0.4862	0.0157	97.79	2.21	0.00	0.00
$D_2 \rightarrow D_2$	CPM	XGBoost	0.9771	0.4837	0.0168	99.00	1.00	0.00	0.00
$D_3 \rightarrow D_3$	ChemBERTa	CatBoost	0.9951	0.1961	0.0117	99.78	0.22	0.00	0.00
$D_3 \rightarrow D_3$	Hybrid	CatBoost	0.9967	0.2201	0.0097	99.83	0.17	0.00	0.00
$D_3 \rightarrow D_3$	CPM	XGBoost	0.9991	0.1733	0.0050	100.00	0.00	0.00	0.00
$D_{total} \rightarrow D_{total}$	ChemBERTa	RF	0.9858	0.5119	0.0183	96.80	2.81	0.32	0.07
$D_{total} \rightarrow D_{total}$	Hybrid	XGBoost	0.9909	0.3864	0.0147	94.19	5.59	0.17	0.04
$D_{total} \rightarrow D_{total}$	CPM	RF	0.9891	0.3779	0.0161	97.64	1.89	0.32	0.15

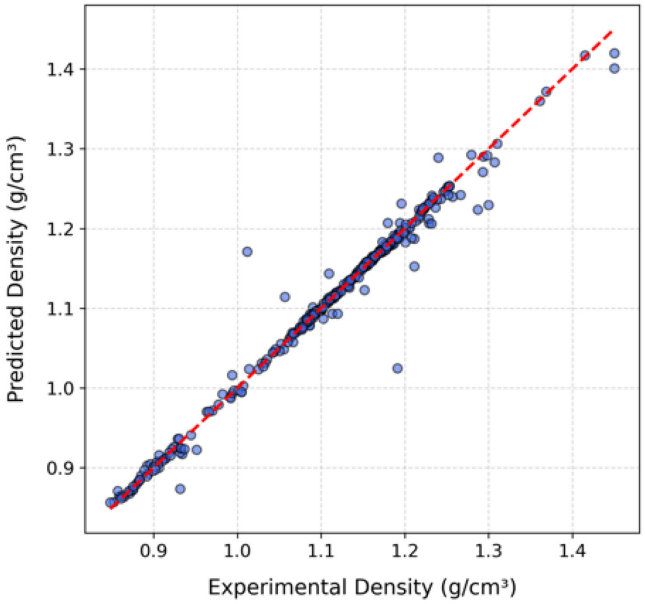

Fig. 1. Parity plots comparing experimental and predicted density for internal validation subsets (D_2, D_3, and D_{total}). Predictions were generated using the XGBoost algorithm trained on critical property model (CPM) features.

stood out as the weakest performer, yielding relatively lower R^2 values during the internal validation process. Meanwhile, the ensemble tree algorithms – including RF, XGBoost, and CatBoost – consistently achieved superior R^2 values exceeding 0.93 across the board. However, when utilizing hybrid embeddings and CPM features, the ANN's performance improved significantly, reaching R^2 values well above 0.90 for all respective subsets.

To visually confirm the high predictive accuracy of the ensemble methods, Fig. 1 presents parity plots comparing experimental and predicted densities for the internal validation subsets using the XGBoost model with CPM features. The tight clustering of data points along the diagonal ($y = x$) line strongly corroborated the excellent R^2 values, demonstrating its higher correlation with the experimental measurements without systematic bias.

The absolute RD distribution illustrated the differences between calculated densities and experimental data under internal validation (diagonal cases). Analysis of the validation sets demonstrated that nearly all cases ($D_2 \rightarrow D_2$, $D_3 \rightarrow D_3$, $D_{total} \rightarrow D_{total}$) produced errors below 5%. In one instance, all deviations were less than 1% ($D_3 \rightarrow D_3$) when employing the CPM feature set with the XGBoost algorithm. These high-accuracy distributions indicated that approximately 2,000 datasets were sufficiently robust to predict DES density with high reliability, with ChemBERTa and hybrid feature sets consistently achieving over 97% of results within the < 1% error range.

This initial underperformance by the ANN with high-dimensional ChemBERTa embeddings may be explained by the fact that neural networks usually require thousands of data points to optimize their weights effectively without overfitting. This extensive data requirement also explained why the ANN's R^2 value increased noticeably as the number of data points grew from around 2,000 to over 12,000 in the combined D_{total} dataset. The vast improvement seen with hybrid and embeddings suggested that explicitly providing condensed physical descriptors helped the ANN compensate for smaller dataset sizes. In contrast, tree-based methods such as RF, XGBoost, and CatBoost were invariant to monotonic transformations [18]. Owing to this mathematical property, these models were naturally insulated against scaling issue, making them much more robust to raw physical data. As a result, they can effectively map the complex property space of chemical mixtures without needing the massive data volumes that ANNs demand. Therefore, the findings in Table 3 clearly highlighted the practical advantages of using ensemble tree methods over ANNs for datasets of this size, unless domain-specific physical features were explicitly provided.

3-2. Performances of the models with generalization (Upper-right triangle)

Table 4 presents the model performance results for generalization, revealing that several methods struggled significantly when trained on the simpler dataset, often yielding negative R^2 values across all three features types. A negative R^2 value indicated that the predictive method performed worse than a simple horizontal line representing the mean of the target variable. This poor performance can primarily be attributed to the simpler dataset lacking the structural and thermodynamic diversity required to train complex ML algorithms effectively.

This lack of diversity was visually evident in Fig. 2, which illustrates the principal component analysis (PCA) of the chemical space defined by the CPM descriptors. The ternary dataset (D_3) covered a noticeably broader and more complex region of the principal component space than the binary dataset (D_2). When models were trained on highly restricted chemical spaces of D_2, they failed to learn the underlying physical laws and merely memorize the limited training examples. Consequently, when evaluated on external test sets of D_3 or D_{total}, these over-fitted models produced highly inaccurate predictions,

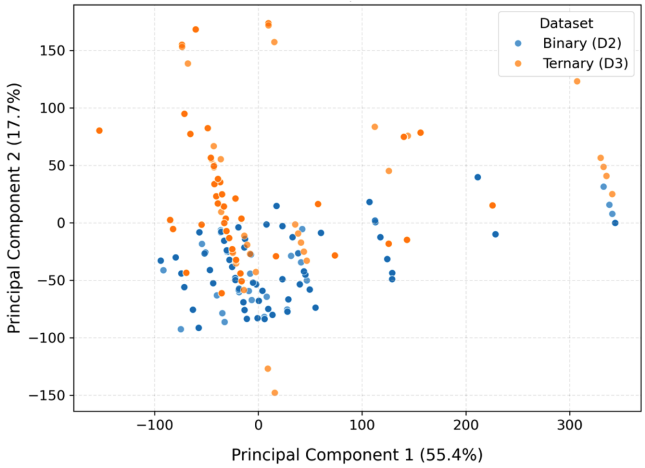

Fig. 2. Principal component analysis (PCA) of the chemical space for binary (D_2) and ternary (D_3) deep eutectic solvents based on critical property model (CPM) descriptors.

leading to the observed negative R^2 metrics. These findings aligned closely with similar literature results, where researchers have noted that limited thermodynamic datasets often caused sophisticated algorithms to underperform compared to simple linear regressions [29,30]. Previous studies similarly confirmed that when the chemical diversity of a dataset was too simple, the generalization capability of advanced tree-based and neural network models drastically deteriorated [29].

The absolute RD distribution presented in Table 4 highlights model performance when generalized across datasets, such as training on D_2 and testing on D_3. Unlike internal validation, these generalization cases exhibited substantially higher deviations, with many errors exceeding 10% (*e.g.*, $D_2 \rightarrow D_3$ with CPM features reaching 71%). Nevertheless, the $D_2 \rightarrow D_{total}$ case with the CPM features and the XGBoost algorithm retained 38% of predictions within a < 1% deviation. These findings suggested that while internal accuracy was strong, greater dataset diversity was required to achieve comparable predictive reliability for unseen DES categories, with ChemBERTa and hybrid features generally outperforming CPM in maintaining < 5% error distributions.

3-3. Performances of the models with specialization (Lower-left triangle)

Table 5 details the predictive performance when models were trained on datasets containing more component data, yet, some scenarios still surprisingly resulted in negative R^2 values (such as extrapolating from D_3 to D_2). The occurrence of negative R^2 values in this context was likely due to the fact that while the system complexity increased (*e.g.*, ternary mixtures), the training data size remained relatively small at around 2,000 points. Complex multi-component systems introduced high-dimensional interactions that required extensive data to model accurately; thus, 2,000 data points were simply insufficient to capture these intricate relationships robustly.

Conversely, when the models were trained on the expansive D_{total} dataset comprising over 12,000 points, they yielded significantly greater performance across the board for all embedding types. The large

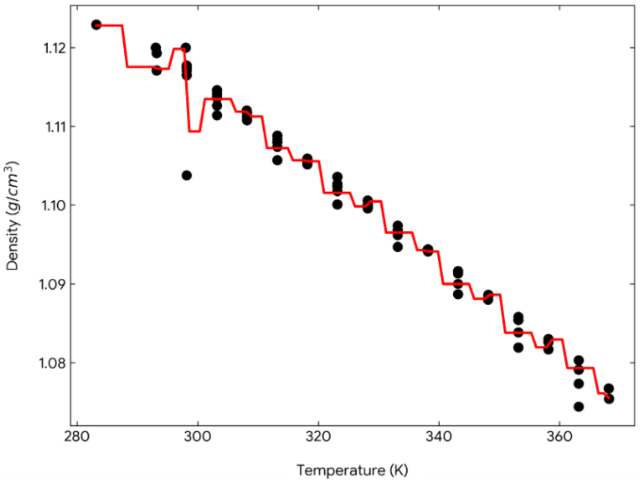

Fig. 3. Density versus temperature profile for a representative deep eutectic solvent composed of choline chloride (1) and ethylene glycol (2) at a constant molar fraction of x1 = 0.333. Predictions were generated using the CatBoost algorithm trained on critical property model (CPM) features.

Table 4. Model performance on generalization (upper-right triangle).

Train → Test	Feature Set	Model	R^2	AARD (%)	RMSE	Absolute RD distribution (%)			
						< 1	1~5	5~10	> 10
$D_2 \rightarrow D_3$	ChemBERTa	ANN	0.2263	8.3699	0.1523	14.45	42.13	18.75	24.67
$D_2 \rightarrow D_3$	Hybrid	CatBoost	0.1151	8.9036	0.1629	12.11	36.38	18.42	33.09
$D_2 \rightarrow D_3$	CPM	RF	0.0564	8.4272	0.1682	2.62	10.94	14.90	71.54
$D_2 \rightarrow D_{total}$	ChemBERTa	CatBoost	0.3203	7.3392	0.1243	31.00	30.49	21.49	17.02
$D_2 \rightarrow D_{total}$	Hybrid	CatBoost	0.5471	4.6556	0.1015	30.65	31.31	22.66	15.38
$D_2 \rightarrow D_{total}$	CPM	XGBoost	0.5511	4.5717	0.1010	38.99	29.67	17.20	14.14
$D_3 \rightarrow D_{total}$	ChemBERTa	CatBoost	0.3329	8.3550	0.1232	22.89	26.61	23.79	26.71
$D_3 \rightarrow D_{total}$	Hybrid	SVR	0.3960	7.2479	0.1172	15.92	21.24	17.36	45.48
$D_3 \rightarrow D_{total}$	CPM	CatBoost	0.5561	6.6116	0.1005	23.58	29.79	18.01	28.62

Table 5. Model performance on specialization (lower-right triangle)

Train → Test	Feature Set	Model	R^2	AARD (%)	RMSE	Absolute RD distribution (%)			
						< 1	1~5	5~10	> 10
$D_3 \rightarrow D_2$	ChemBERTa	CatBoost	0.3289	6.0176	0.0891	14.05	47.01	22.84	16.10
$D_3 \rightarrow D_2$	Hybrid	CatBoost	0.3534	5.5734	0.0875	19.41	37.89	25.17	17.53
$D_3 \rightarrow D_2$	CPM	CatBoost	0.4575	5.6044	0.0801	17.53	41.21	20.58	20.68
$D_{total} \rightarrow D_2$	ChemBERTa	XGBoost	0.9853	0.6154	0.0182	31.65	24.18	19.47	24.70
$D_{total} \rightarrow D_2$	Hybrid	XGBoost	0.9890	0.4129	0.0157	33.74	27.04	19.08	20.14
$D_{total} \rightarrow D_2$	CPM	RF	0.9870	0.3885	0.0171	37.48	29.25	16.40	16.87
$D_{total} \rightarrow D_3$	ChemBERTa	CatBoost	0.9969	0.5408	0.0096	21.21	30.47	23.35	24.97
$D_{total} \rightarrow D_3$	Hybrid	XGBoost	0.9993	0.2268	0.0470	23.08	22.27	17.49	37.16
$D_{total} \rightarrow D_3$	CPM	RF	0.9983	0.3139	0.0071	19.77	24.96	22.93	32.34

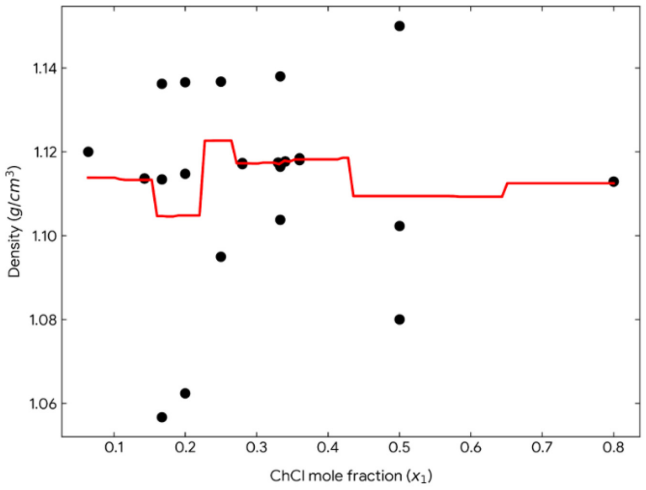

Fig. 4. Density versus composition profile for a representative deep eutectic solvent composed of choline chloride (1) and ethylene glycol (2) at 298.15 K. Predictions were generated using the CatBoost algorithm trained on critical property model (CPM) features.

volume of data in D_{total} allowed the algorithms to effectively learn comprehensive thermodynamic behaviors, drastically improving transferability to specific subsets like D_2 and D_3. The robust physical foundation of these global models is further demonstrated in Fig. 3, which plots the predicted versus experimental density across a range of temperature for a representative DES mixture. The CatBoost model accurately reproduced the linear thermodynamic trend of decreasing density with increasing temperature, confirming that models trained on D_{total} successfully internalized fundamental phase behaviors rather than merely interpolating noise. Fig. 4 illustrates the relationship between chemical composition and the density of a DES composed of choline chloride and ethylene glycol. A solid line represented the model's predicted density, showing how the algorithm smoothen the transition across the full range of mole fractions. This comparison highlighted the high precision of the predictive model, as the line closely aligned with the experimental observations throughout the concentration range.

Interestingly, the transfer performance from D_{total} to D_3 was notable better than the transfer from D_{total} to D_2. This better performance on D_3 can be explained by the fact that the ternary dataset (D_3) encompassed a more complex and inclusive thermodynamic space, making it easier for a comprehensively trained model to interpolate with it. Meanwhile, predicting the simpler binary subset (D_2) from a globally trained model might introduced slight deviations, as the model attempted to apply complex ternary interaction rules to simpler binary systems. This phenomenon was consistent with similar literature results, which demonstrated that models trained on highly complex, multi-component mixture often showed excellent predictive power when interpolating within complex spaces [31]. Other published thermodynamic ML studies have also concluded that expanding dataset size to tens of thousands of points was the most reliable way to overcome poor generalization and negative R^2 values in multi-component fluid predictions [30,32].

The absolute RD distribution for specialization examined deviations when models trained on broader datasets were applied to specific subsets, such as $D_3 \rightarrow D_2$. In these cases, a substantial proportion of predictions remained within acceptable ranges, with the hybrid model for $D_3 \rightarrow D_2$ achieving 57% of results below 5% error. Notably, the CPM feature set with the RF algorithm reached 37% of predictions within < 1% deviation, underscoring strong localized accuracy. Although approximately 20% of $D_3 \rightarrow D_2$ predictions exceeded 10% deviation, the inclusion of ~10,000 additional data points in D_{total} appeared to stabilize the RD distribution, enabling reliable DES density predictions across the combined dataset.

4. Conclusions

The study compared machine learning models and feature sets for predicting the density of deep eutectic solvents (DESs). Results demonstrated that ensemble tree-based method (RF, XGBoost, and CatBoost) consistently outperformed artificial neural networks (ANNs) on smaller datasets and handle raw physical data without requiring extensive scaling. ANNs struggled with high-dimensional ChemBERTa embeddings on limited data, but accuracy improved when condensed physical descriptors (hybrid or CPM) were added. Cross-domain tests showed models trained on simple or small datasets generalized poorly (negative R^2) due to insufficient thermodynamic diversity, whereas training on a comprehensive dataset containing over 12,000 records yielded excellent predictive performance across all subsets. Ultimately, maximizing dataset size and diversity, alongside utilizing tree-based algorithms or physics-informed features, provided the most reliable framework for accelerating DES discovery and property simulation. Furthermore, while the present investigation centered on density, the proposed methodology framework – particularly the integration of physics-informed critical property models with robust tree-based ensembles – can be readily extended to accurately predict other vital thermophysical properties of DESs, such as viscosity, surface tension, and electrical conductivity.

Acknowledgements

This work was supported by 2026 Hongik University Research Fund.

Data availability: Data will be made available on request.

References

1. Costa, S. P. F., Azevedo, A. M. O., Pinto, P. C. A. G. and Saraiva, M. L. M. F. S., "Environmental Impact of Ionic Liquids: Recent Advances in (eco)toxicology and (bio)degra-dability," *ChemSusChem*, **10**, 2321-2347(2017).
2. Płotka-Wasylka, J., Guardia, M., Andruch, V. and Vilková, M., "Deep Eutectic Solvents vs. Ionic Liquids: Similarities and Differences," *MicroChem. J.*, **159**, 105539(2020).

3. Jung, M. and Park, Y., "Thermophysical Properties and Liquid-liquid Equilibria of Pseudoternary Systems {toluene + n-heptane + deep Eutectic Solvents Based on Levulinic Acid}," *J. Chem. Eng. Data*, **67**, 416-427(2022).
4. Shang, M., Wang, P., Cheng, Y., Zhou, Y. and Lei, Z., "Separation of Azeotropes in the Dimethoxymethane Production Using Deep Eutectic Solvents: Liquid-liquid Extraction Experiments and Molecular Insights," *Fuel*, **406**, 137184(2026).
5. Wang, D., Wang, Y., Ma, R., Ma, J., Xu, M., Ai, L., Leng, C., Ma, Q., Jia, D., Wang, L. and Guo, N., "Efficient and Green Separation of Phenolics by Halogen Free Deep Eutectic Solvents," *J. Mol. Liquids*, **417**, 126602(2025).
6. Alnajjar, A. and Onaizi, S., "CO_2 Capture and Conversion Using Deep Eutectics Solvents: Recent Progress and Outlooks," *J. Mol. Liquids*, **421**, 126832(2025).
7. Abranches, D. O., Silva, L. P., Martins, M. A. R., Pinho, S. P. and Coutinho, J. A. P., "Understanding the Formation of Deep Eutectic Solvents: Betaine as a Universal Hydrogen Bond Acceptor", *ChemSusChem*, **13**, 4916-4921(2020).
8. Mulyono, S., Hizaddin, H. F., Alnashef, I. M., Hashim, M. A., Fakeeha, A. F. and Hadj-Kali, M. K., "Separation of BTEX Aromatics from n-octane Using a (tetrabutylammonia bromide + sulfolane) Deep Eutectic Solvent – experiments and COSCO-RS Prediction," *RSC Adv.*, **4**, 17597(2014).
9. Ji, Q., Yu, X., Yagoub, A. Chen, L. and Zhou, C., "Efficient Removal of Lignin from Vegetable Waste by Ultrasonic and Microwaves-assisted Treatment with Ternary Deep Eutectic Solvent," *Industrial Crops and Products*, **149**, 112357(2020).
10. Park, Y., The Solubility of Deep Eutectic Solvents Derived from Allyltriphenylphosphonium Bromide in Supercritical Carbon Dioxide in the Presence of Ethanol as a Cosolvent, *Korean J. Chem. Eng.*, **41**, 2091-2097(2024).
11. Boublia, A., Lemaoui, T., Almustafa, G., Darwish, A. S., Benguerba, Y., Banat, F. and AlNashef, I. M., "Critical Properties of Ternary Deep Eutectic Solvents Using Group Contribution with Extended Lee–Kesler Mixing Rules," *ACS Omega*, **8**, 13177(2023).
12. Almustafa, G., Darwish, A. S., Lemaoui, T., O'Conner, M. J., Amin, S., Arafat, H. A. and AlNashef, I., Liquification of 2,2,4-trimethyl-1,3-pentanediol Into Hydrophobic Eutectic Mixtures: A Multi-criteria Design for Eco-efficient Boron Recovery," *Chem. Eng. J.*, **426**, 131342(2021).
13. Park, Y., Liquid-liquid Equilibria of Cylcohexene-cyclohexane with Betaine-glyceol DES: Experiments and Correlation," *Korean Chem. Eng. Res.*, **63**, 334-340(2025).
14. Di Pietro, T., Cesari, L. and Mutelet, F., Group Contribution Models for Densities and Heat Capacities of Deep Eutectic Solvents," *Fluid Phase Eq.*, **572**, 113854(2023).
15. Hu, J., Peng, D., Huang, X., Wang, N., Liu, B., Di, D., Liu, J., Qu, Q., Pei, D., "COSMO-SAC and QSPR Combined Models: A Flexible and Reliable Strategy for Screening the Extraction Efficiency of Deep Eutectic Solvents," *Sep. Purif. Technol.*, **315**, 123699(2023).
16. Were, K., Bui, D. T., Dick, O. B. and Singh, B. R., "A Comparative Assessment of Support Vector Regression, Artificial Neural Networks, and Random Forests for Predicting and Mapping Soil Organic Carbon Stocks Across an Afromontrane Landscape," *Ecol. Ind.*, **52**, 394(2015).
17. Wang, Y. X., Hou, X. J., Zeng, J., Wu, K. J. and He, Y. C., Random Forest Models to Predict the Densities and Surface Tensions of Deep Eutectic Solvents," *AIChE J.*, **69**, e18095(2023).
18. Chen, T. and Guestrin, C., "XGBoost: A Scalable Tree Boosting System, In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16), San Francisco, CA, USA, August 13–17, 2016; pp. 785–794.
19. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. and Gulin A., "CatBoost: Unbiased Boosting with Categorical Features," In Proceedings of the 32nd International conference on neural information processing systems, pp. 6639-6649.
20. Chithraranda, S., Grand, G. and Ramsundar, B., "ChemBERTa: Large-scale Self-supervised Pretraining for Molecular Property Prediction," arXiv:2010.09885(2020).
21. Wu, T., Song, J., Hu, Y., Qiu, Q., Tang, J., Peng, W., Fu, X., Shi, C. and Lin, X., "Methodological Roadmap for Machine Learning in Deep Eutectic Solvent Research: A Framework-driven Review and Perspective," *Ind. Eng. Chem. Res.*, **64**, 16443-16465(2025).
22. Omar, K. A. and Sadeghi, R., "Database of Deep Eutectic Solvents and Their Physical Properties: A Review," *J. Mol. Liquids*, **384**, 121899(2023).
23. Gajardo-Parra, N. F., Controneo-Figueroa, V. P., Aravena, P., Vesovic, V. and Canales, R. I., "Viscosity of Choline Chloride-based Deep Eutectic Solvents: Experiments and Modeling," *J. Chem. Eng. Data*, **65**, 5581-5591(2020).
24. Yadav, A., Kar, J. R., Verma, M., Naqvi, S. and Pandey, S., "Densities of Aqueous Mixtures of (choline chloride+ethylene glycol) and (choline chloride+malonic acid) Deep Eutectic Solvents in Temperature Range 283.15-363.15K," *Thermo. Acta*, **600**, 95-101 (2015).
25. Wu, T., Zhan, P., Chen, W., Lin, M., Qiu, Q., Hu, Y., Song, J. and Lin, X., "ChemBERTa Embeddings and Ensemble Learning for Prediction of Density and Melting Point of Deep Eutectic Solvents with Hybrid Features," *Comput. Chem. Eng.*, **196**, 109065 (2025).
26. Tropsha, A., "Best Practices for QSAR Model Development, Validation, and Assessment," *Mol. Inf.*, **29**, 476-488(2010).
27. Hawkins, D. M., "The Problem of Overfitting," *J. Chem. Inf. Comput. Sci.*, **44**, 1-12(2004).
28. Mjalli, F. S., "Mass Connectivity Index-based Density Prediction of Deep Eutectic Solvents," *Fluid Phase Equilibria*, **409**, 312-317(2016).
29. Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K. and Pande, V., "MoleculeNet: a Benchmark for Molecular Machine Learning," *Chem. Sci.*, **9**, 513-530 (2018).
30. Goh, G. B., Hodas, N. O. and Vishnu, A., "Deep Learning for Computational Chemistry," *J. Comput. Chem.*, **38**, 1291-1307(2017).
31. Liu, Y., Hong, W. and Cao, B., "Machine Learning for Predicting Thermodynamic Properties of Pure Fluids and Mixtures," *Energy*, **188**, 116091(2019).
32. Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelly, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K. and Barzilay, R., Analyzing Learned Molecular Representations for Property Prediction," *J. Chem. Inf. Model.*, **59**, 3370-3388(2019).

Authors

YoonKook Park: Professor, Department of Bio and Chemical Engineering, Hongik University, Sejong 30016, Korea; parky@hongik.ac.kr