Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.11 n.56, p. 180-185. November/December, 2025.
DOI: https://doi.org/10.5935/jetia. v11i56.2444

RESEARCH ARTICLE OPEN ACCESS

ENSEMBLE LEARNING BASED MOLECULE EMBEDDING TO IMPROVE NANOFLUIDS' THERMAL CONDUCTIVITY PREDICTION

Meghazi Hadj Madani[1], Mekroussi Said[2], Maaskri Moustafa[3]

[1]Computer Science Department, University of Tiaret, Algeria.
[2] Laboratory of Industrial Technologies, Mechanical Engineering Department,University of Tiaret, Algeria.
[3]Electrical Engineering Department,University of Tiaret, Algeria

[1]http://orcid.org/0000-0002-9057-8775, [2]http://orcid.org/0000-0001-8824-3960, [3]http://orcid.org/0000-0003-1680-5033

E-mail: h.meghazi@univ-tiaret.dz, said.mekroussi@univ-tiaret.dz, moustafa.maaskri@univ-tiaret.dz

ARTICLE INFO

Article History
Received: July 24, 2025
Revised: October 20, 2025
Accepted: December 1, 2025
Published: December 31, 2025

Keywords:
Ensemble learning,
Molecular embedding,
Thermal conductivity prediction,
Nanofluids.

ABSTRACT

The thermal conductivity of nanofluids plays a critical role in numerous industrial applications, including lid-driven cavities and metallurgical lubrication technology. However, accurately predicting this property remains a persistent challenge, necessitating continued innovation. In this study, we propose a novel methodology that leverages ensemble learning with molecular embeddings to enhance the prediction of nanofluid thermal conductivity. By integrating existing correlations with experimental data, our approach generates robust predictive models that outperform state-of-the-art methods. Experiments conducted on a real-world dataset demonstrate the superior performance of the proposed framework, highlighting its potential to advance research and industrial applications in nanofluid heat transfer.

I. INTRODUCTION

Prediction of the thermal conductivity of nanofluids involves utilizing various models and algorithms. Studies have proposed models like artificial neural networks (ANN) and power law models to estimate thermal conductivity based on factors such as particle concentration, temperature, particle size, and base fluid properties, Hanief et al. [1] proposed an artificial neural network model to predict the thermal conductivity ratio of metal oxide nanofluids based on particle concentration, temperature, particle size and properties of the base fluid. Shahrivar et al.[2] created a highly accurate artificial neural network model using experimental data to forecast the thermal conductivity of nanofluids. In [3], machine learning models, particularly Gradient Boosting Regression, predict the thermal conductivity of oxide nanofluids based on particle size, concentration, and temperature variations, enhancing system efficiencies in various industries. Similarly, in [4], the thermal conductivity of nanofluids can be predicted by considering factors like nanoparticle size, type, volume fraction, and base fluid, as well as by utilizing a proposed thermal conductivity model. Theoretical models like effective medium theory, Brownian models, nanolayer models, and molecular dynamics simulations are used to predict nanofluid thermal conductivity in [5].

The study of [6] predicts TiO2/water nanofluids' thermal conductivity using machine learning models, with the LS-SVM method showing the highest prediction accuracy among the models tested.The study of [7] predicts enhanced thermal conductivity of nanofluids by incorporating boron nitride nano-barbs into ethylene glycol and propylene glycol mixtures, showing up to a 45% increase with higher concentrations. XGBoost and Gaussian process regression were utilized in [8] to predict the impact of sonication on thermal conductivity of f-MWCNT nanofluids, determining optimal conditions for enhanced stability and dispersion. In this paper [9], a physics guided machine learning approach was proposed by incorporating physics-based relations at the initial stage and into traditional loss functions for predicting the thermal conductivity of water-based nanofluids using a wide range of both experimental and simulated data. As noted in [10], the thermal conductivity of the nanofluids with SiC particles is modeled by using artificial neural network as an intelligent method, pointing up the importance of factors such as temperature and volume fraction in influencing the thermal conductivity of nanofluids, which are critical for effective modeling, revealing significant closeness of the forecasted data and corresponding experimental values. [11] features experimental and theoretical approaches for predicting the thermal conductivity of nanofluids, highlighting advancements

in understanding that show no significant change upon nanoparticle addition. In the findings of [12], a deep forest model was employed to predict the thermal conductivity of hybrid nanofluids, and the results showed that the model's predictive performance was better than those of artificial neural network (ANN), LSTM, support vector regression (SVR), random forest (RF), and various empirical equations. In this paper [13], the thermal conductivity and dynamic viscosity of five kinds of graphene oxide nanofluids, with a mass fraction of 0.002% to 0.01%, were studied in the temperature range of 25 – 50°C. [14] In this article, statistical analysis method such as cluster analysis was used to group the effective thermal conductivity models of nanofluids and the recommended GPI for the model ranges from -6.4197 to 2.5742.

As noted in [15], the transfer learning-based models' results are compared with those from baseline models which are trained only on experimental data using a goodness of fit metric (R) and it is found that the transfer learning models perform better with R values of ∼0.93 as opposed to ∼0.83 from the baseline models. The authors of [16] investigated the thermal properties of MXene-based nanofluids and found that MXene provided a 30.6% improvement in the effective thermal conductivity of the 0.5 wt% MXene/water compared to the pure water base fluid. In this paper [17], theoretical and mathematical models have been compared to predict the thermal conductivity of nanofluids and the experimental data have been collected from literature and compared with various models, results showed that mathematical models are effective when volume fraction value is less than 0.01. [18] In this paper, two appropriate semi-experimental models based on none-linear regression over 800 extracted experimental data to predict the thermal conductivity coefficient of nanofluids were presented. [19] In this article, a review of the effect of alignment, electric field, and green nanofluid on thermal conductivity of metal oxide nano-nodes is presented and discussed.

[20] In this article, the thermal conductivity of 11 nanofluids, for a total of 239 experimental points, was analyzed in detail, and a new semi-empirical, scaled equation was also proposed. [21] In this paper, a neural regression model for predicting carbon nanotube nanofluids (knf) is proposed, which takes into account four influencing factors, including carbon-notube diameter, volume fraction, temperature and base fluid thermal conductivity (kf). [22] In this article, the authors used locally weighted linear regression (LWLR) to predict the thermal conductivity ratio of hybrid nanofluids in water, ethylene glycol and various volume percentages of the base fluid. [23] In this paper, a high-accuracy regression equation is developed for the prediction of thermal conductivity of graphene nanoplatelet-water nanofluids, based on the temperature (15-60 °C), nanoparticle weight fraction (0.025-0.1 wt.), and graphene nanoparticle specific surface area (300-750 m^2/g). While these studies demonstrate significant advancements in predicting nanofluid properties, the investigation of integrating molecular embedding techniques with ensemble learning remains an emerging area.

II. PROPOSED APPROACH

In this study, we present a novel approach leveraging ensemble learning combined with molecular embeddings to enhance the prediction of thermal conductivity in nanofluids. Molecule embeddings are used to address limitations in previous methodologies, which often relied on simple encoding techniques, such as one-hot encoding. These traditional methods fail to capture the inherent structure and relationships within molecular data. As noted in [10], one-hot encoding might represent the molecule "**CuO**" as "**1000**" in one experiment and "**0100**" in another, result- ing in inconsistent and non-representative encodings. In contrast, **molecule embeddings** provide a well-defined vectorial representation that encapsulates molecular properties and relationships, enabling more accurate learning (See figure 1). To maximize the potential of this representation, we adopted an **ensemble learning approach**, which combines the predictions of multiple models to create a more robust and generalized solution. Ensemble learning leverages the strengths of individual models while mitigating their weaknesses, thereby enhancing prediction accuracy and reducing overfitting.

This framework integrates diverse algorithms, enabling the exploitation of molecule embeddings' rich information effectively. Using this approach, we evaluated four (04) models: H2O AutoML [24], XGBoost Tree Ensemble, Gradient Boosted Trees and Regression Tree, by utilizing molecule embeddings as the primary input, which represents the key originality of this paper in the field. To enhance the predictive capability, we combined the molecule embeddings with additional features, including the volume fraction (ϕ), temperature (T), and particle size. The AutoML process involved training the specified models and compiling them into a single table. These models were then tested on the test set, with their predictions compared against the actual values. Multiple performance metrics were calculated to assess the accuracy of the predictions, and the model demonstrating the highest R^2 score was selected as the best-performing model. This comprehensive evaluation highlights the effectiveness of our method in advancing the prediction of thermal conductivity in nanofluids.

III. EXPERIMENTS

III. 1 DATASET AND EXPERIMENTAL ENVIRONMENT

For this study, we utilized the same dataset as in [15], which comprises molecular properties and thermal conductivity values for a range of nanofluids. The dataset includes diverse combinations of nanoparticles (e.g., CuO, Al_2O_3, TiO_2) and base fluids (e.g., **W**ater, **E**thylene **G**lycol) to provide comprehensive coverage of nanofluid compositions. The dataset comprises **1,015** observations, characterizing different nanofluid compositions through six attributes: particle type, fluid composition, volume fraction (ϕ), temperature (T), particle size, and thermal conductivity ratio (k ratio). The volume fraction (ϕ) varies between **0** and **0.1828**, with an average of **0.0398** and a standard deviation of **0.0331**, indicating a range of particle concentrations. The recorded temperatures range from **4.00°C** to **97.29°C**, with a mean of **37.15°C**, reflecting a diverse thermal environment.

Particle sizes span from $\mathbf{5.00 \times 10^{-9}}$ **m** to $\mathbf{9.80 \times 10^{-8}}$ **m**, with an average of $\mathbf{3.59 \times 10^{-8}}$ **m** (≈ **36 nm**), suggesting a predominant nanoparticle regime (See figure 2). The thermal conductivity ratio (k ratio), a key property for assessing heat transfer efficiency, exhibits values between **1.00** and **1.69**, with a mean of **1.1579** and a standard deviation of **0.115**, indicating a moderate variation in thermal enhancement. The dataset includes multiple particle- fluid combinations, allowing for comparative analysis of their thermophysical behaviors. Given the range of values, this dataset provides a solid foundation for investigating the influence of particle size, concentration, and temperature on the thermal performance of nanofluids (See figure 3).

Both the nanoparticles and base fluids in the dataset were first transformed into their respective **SMILES** (**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem) notations, a standardized textual representation of molecular structures.

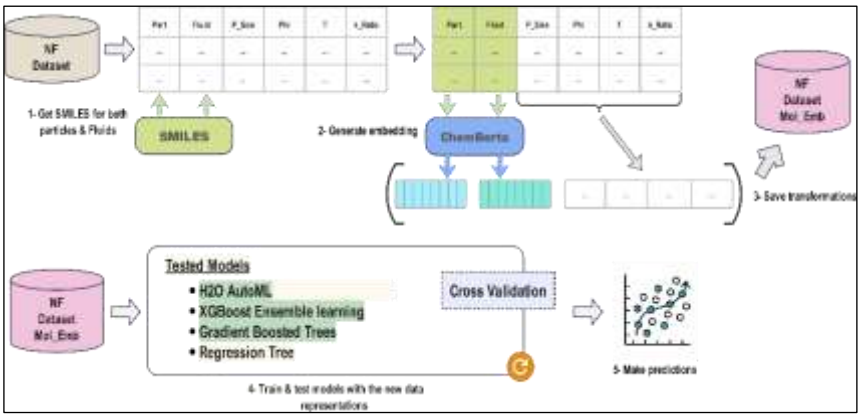

Figure 1: Workflow of the Proposed Approach for Nanofluid Thermal Conductivity Modeling.
Source: Authors, (2025).

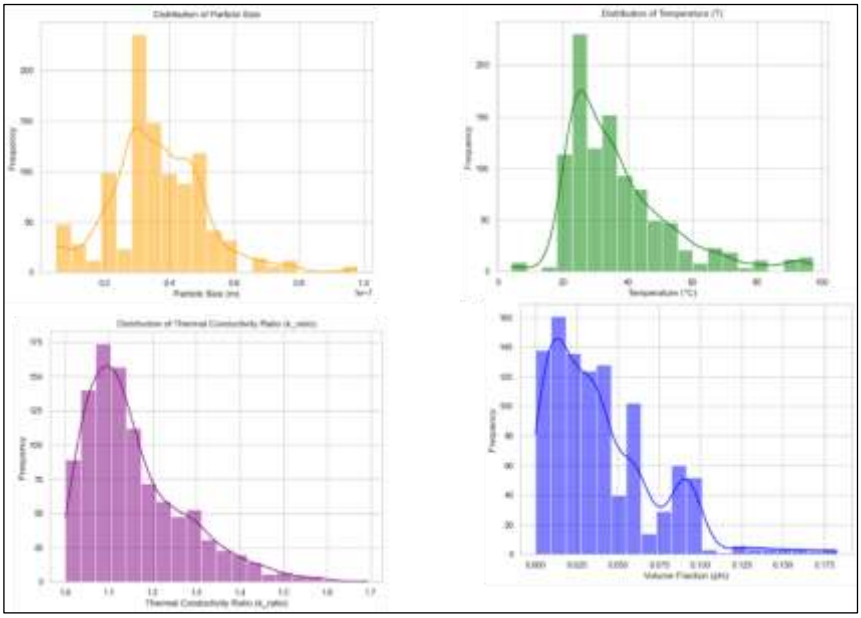

Figure 2: Descriptive statistics of the distribution of thermal conductivity ratio, temperature, particle size and volume fraction from the dataset.
Source: Authors, (2025).

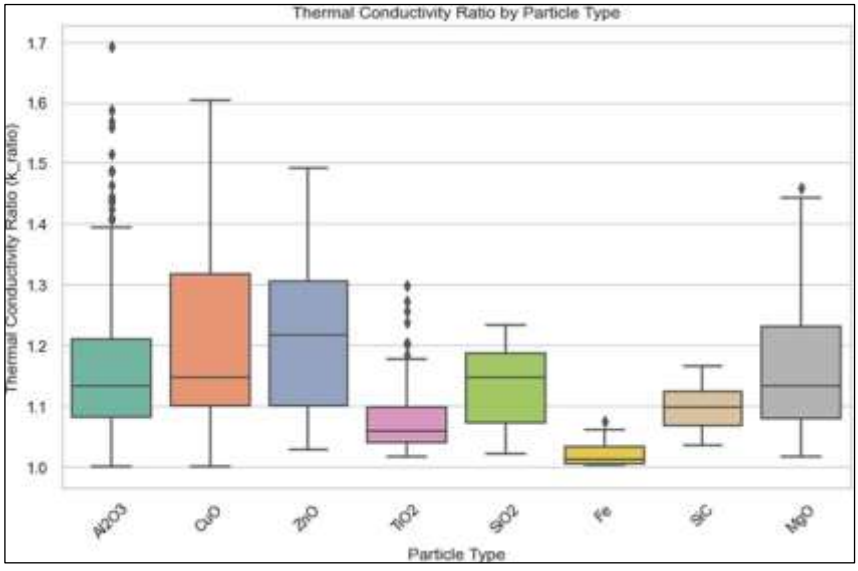

Figure 3: Thermal conductivity ratio by particle type.
Source: Authors, (2025).

For the fluids, we directly used the generated embeddings in the case of water alone or ethy- lene glycol alone, but in the case where we have **EG/W mixtures**, we made the proportions of each come into play. For example, if we have the **EG/W mixture** at **60:40**, the corresponding embedding is: **(0.6*Embeddings of the SMILES of EG) + (0.4*Embeddings of the SMILES of water)**. The dataset was preprocessed to address missing values and ensure consistency in features. Molecule embeddings were scaled and normalized to facilitate effective learning across models. For performance evaluation, the dataset was split into training (75%), validation (15%), and test (10%) subsets using a stratified sampling technique to preserve the distribution of thermal conductivity values. The experimental environment was configured on a platform equipped with a GPU NVIDIA RTX3060Ti, CPU Intel 12900k and 32GB of RAM using Python 3.10 and TensorFlow 2.18. The experiments were conducted under controlled hyperparameter tuning protocols using grid search and cross-validation techniques to optimize model performance.

III. 2 EXPERIMENTAL RESULTS AND DISCUSSION

We evaluated the performance of four (04) machine learning models. The primary evaluation metric was the **R^2 score**, which measures the accuracy and robustness of the models in predicting thermal conductivity.

H2O AutoML (using molecule embeddings): This model achieved the highest R^2 score **of 0.9923**, showcasing its superior ability to utilize molecule embeddings for highly accurate predictions.

XGBoost Tree Ensemble (using molecule embeddings): With an R^2 score of **0.9867**, this model demonstrated excellent performance, only slightly trailing H2O AutoML.

Gradient Boosted Trees (using molecule embeddings): Achieving an R^2 score of **0.9851**, this model performed strongly but ranked just below H2O AutoML and XGBoost Tree Ensemble.

Regression Tree (using molecule embeddings): This model delivered a solid R^2 score of **0.9732**, indicating effective use of molecule embeddings but falling short of the top three models.

H2O AutoML (using OneHot encoding): While competitive with an R^2 score of **0.93**, this model was notably outperformed by those leveraging molecule embeddings, emphasizing the value of molecular embedding representation in improving prediction accuracy.

The results of all evaluated models are summarized in Table 1, providing a comprehensive comparison of their performance metrics

Table 1: Performance Comparison of Machine Learning Models for Thermal Conductivity Prediction.

Models	***R^2 values***
H2O AutoML (using Molecule embedding)	**0.9923**
XGBoost Tree Ensemble (using Molecule embedding)	0.9867
Gradient Boosted Trees (using Molecule embedding)	0.9851
Regression Tree (using Molecule embedding)	0.9732
AutoML(using OneHot)	0.9300
Ensemble Learning [15](Train+Test Data)	0.8300
Transfer Learning [15] (Train+Test Data)	0.9300

Source: Authors, (2025).

The H2O AutoML model, using molecule embeddings and identified as our best-performing approach, achieved an outstanding R^2 score of **0.9923**, significantly surpassing the results reported in [15]. The models in [15], namely "*Ensemble Learning*" and "*Transfer Learning*", both utilizing one-hot encoding, achieved R^2 scores of **0.83** and **0.93**, respectively. Compared to these models, our approach demonstrates an R^2 improvement of **19.55%** over Ensemble Learning and **6.7%** over Transfer Learning, underscoring the added value of molecular embeddings as input features. Figure 4 illustrates and summarizes the results of the tested models. In the proposed approach, we utilized ChemBERTa embeddings on a relatively small dataset. These embeddings transformed both the particle and fluid columns into 384 columns each, which in itself presented a challenge. Instead of working with a dataset of (**6 columns * 1015 rows**), we worked on a dataset of (**772 columns * 1015 rows**), yet we achieved impressive performance.

This highlights the scalability and effectiveness of our methodology even with high-dimensional data. Additionally, the integration of molecular representations, alongside other critical parameters such as temperature (T), particle size, and volume fraction (ϕ), significantly enhanced predictive accuracy. The AutoML process leveraging these embeddings achieved the best R^2 score, which was improved by **1.3%** following a transformation of particle size values, ***multiplying*** them by $\mathbf{10^8}$ to address the adverse effects of working with values on the order of $\mathbf{10^{-8}}$. These results emphasize the superiority of molecular embeddings over one-hot encoding, as well as the robustness and originality of our methodology in modeling the thermal conductivity of nanofluids. This comprehensive evaluation not only highlights the effectiveness of our approach but also showcases its potential to advance research in this domain.

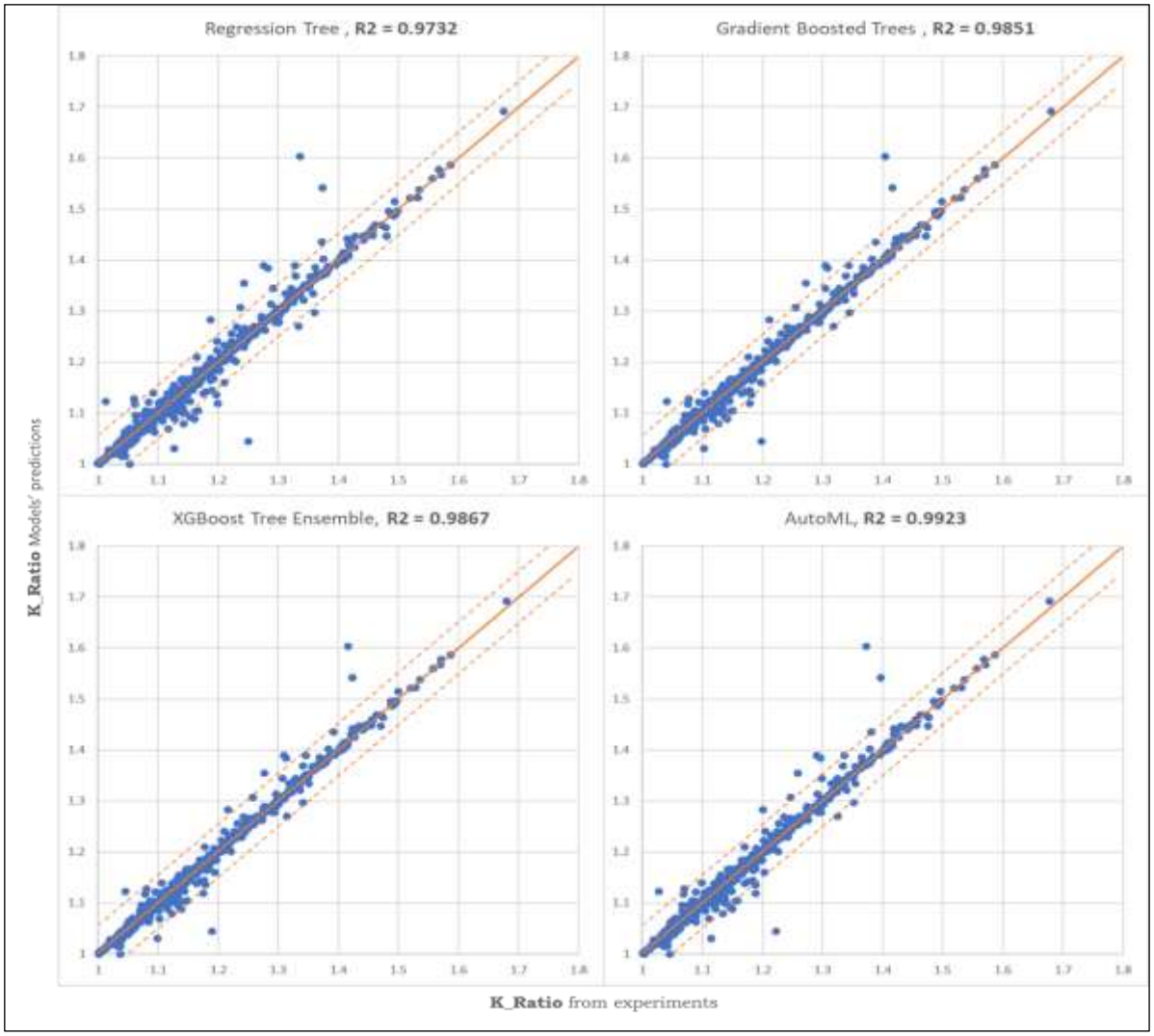

Figure 4: Predictive Performance of Tested Models Using Molecular Embeddings with R^2 as the Performance Metric.
Source: Authors, (2025).

V. CONCLUSIONS

In this study, we demonstrated the effectiveness of leveraging molecular embeddings as input features in predictive models for nanofluid thermal conductivity. By employing an AutoML framework, particularly H2O AutoML, and integrating molecular representations alongside key parameters such as temperature, particle size, and volume fraction, we achieved a state-of-the-art R^2 score of **0.9923**. This surpasses the performance of existing models from the literature, including Ensemble Learning and Transfer Learning approaches, by **19.55%** and **6.7%**, respectively. The incorporation of molecular embeddings enabled a deeper representation of the underlying chemical properties, significantly improving predictive accuracy compared to traditional one-hot encoding techniques. These results emphasize the potential of molecular embeddings as a transformative tool for enhancing predictive modeling in nanofluid research.

By combining these advanced representations with automated machine learning, we offer a robust and scalable methodology that surpasses conventional techniques. Our work not only establishes a new benchmark for predicting nanofluid thermal conductivity but also lays the foundation for future research integrating molecular representations in related domains. This approach holds significant promise for advancing industrial applications where precise thermal conductivity predictions are critical, fostering further innovation in heat transfer technologies. While this study leveraged molecular embeddings, the field offers several techniques, such as graph-based embeddings and SMILES-based embeddings, that remain underexplored in the context of predictive modeling for thermal conductivity and related properties. A deeper exploration of these techniques will help identify the most suitable methods for different types of chemical data and physical properties.

VIII. REFERENCES

[1] M. Hanief, Q. Irfan, and M. Parvez, "Modeling and prediction of thermal conductivity ratio of metal-oxide based nano-fluids using artificial neural network and power law," Chemical and Process Engineering, pp. 159–163, 2022.

[2] I. Shahrivar, A. Ghafouri, Z. Niazi, and A. Khoshoei, "Development of a neural architecture to predict the thermal conductivity of nanofluids," Journal of the Brazilian Society of Mechanical Sciences and Engineering, vol. 45, no. 12, p. 646, 2023.

[3] H. M. Singh and D. P. Sharma, "Machine learning model to predict the efficiency of thermal conductivity of oxide nanofluids," Waves in Random and Complex Media, pp. 1–20, 2023.

[4] J. Qin, Y. Tao, Q. Liu, Z. Li, Z. Zhu, and N. He, “Experimental and Theoretical Studies of Different Parameters on the Thermal Conductivity of Nanofluids,” Micromachines (Basel), vol. 14, no. 5, p. 964, 2023.

[5] R. Fulmer and S. Vafaei, “Modelling and Mechanisms of Nanofluid Thermal Conductivity,” 2022.

[6] J. Li, W. Deng, S. Qing, Y. Liu, H. Zhang, and M. Zheng, “Prediction and Optimization of the Thermal Properties of TiO 2/Water Nanofluids in the Framework of a Machine Learning Approach.,” Fluid Dynamics & Materials Processing, vol. 19, no. 8, 2023.

[7] A. O. Maselugbo, B. L. Sadiku, and J. R. Alston, “Thermal conductivity of ethylene glycol and propylene glycol nanofluids with boron nitride nano-barbs,” Nanoscale, vol. 15, no. 18, pp. 8406–8415, 2023.

[8] Z. Said, P. Sharma, B. J. Bora, and A. K. Pandey, “Sonication impact on thermal conductivity of f-MWCNT nanofluids using XGBoost and Gaussian process regression,” J Taiwan Inst Chem Eng, vol. 145, p. 104818, 2023.

[9] B. Bhaumik, S. Changdar, and S. De, “An expert model based on physics-aware neural network for the prediction of thermal conductivity of nanofluids,” J Heat Transfer, vol. 144, no. 10, p. 103501, 2022.

[10] R. M. Shahzad, H. F. Fard, I. Mahariq, M. E. H. Assad, and M. A. AlShabi, “Thermal conductivity prediction of nanofluids containing SiC particles by using artificial neural network,” in Energy Harvesting and Storage: Materials, Devices, and Applications XII, 2022, pp. 37–41.

[11] T. M. Koller, F. E. Berger Bioucas, and A. P. Fröba, “Thermal Conductivity of Nanofluids–Experiments, Models, and their Advancements,” 2022.

[12] D. He, C. Li, Y. Chen, and X. Li, “Thermal conductivity prediction of hybrid nanofluids based on deep forest model,” Heat Transf Res.

[13] X. Mei, X. Sha, D. Jing, and L. Ma, “Thermal conductivity and rheology of graphene oxide nanofluids and a modified predi cation model,” Applied Sciences, vol. 12, no. 7, p. 3567, 2022.

[14] H. Salhi and N. Chafai, “Evaluation of the thermal conductivity of nanofluids using statistical analysis methods,” Nanoscience and Technology: An International Journal, vol. 13, no. 4, 2022.

[15] S. S. Pai and A. Banthiya, “Transfer-learning-based surrogate model for thermal conductivity of nanofluids,” arXiv preprint arXiv:2201.00435, 2022.

[16] M. Mao et al., “Ti3C2Tx MXene nanofluids with enhanced thermal conductivity,” Chemical Thermodynamics and Thermal Analysis, vol. 8, p. 100077, 2022.

[17] B. K. Dandoutiya and A. Kumar, “Comparison of mathematical models to estimate the thermal conductivity of TiO2-water based nanofluid: A review,” Thermal Science, vol. 26, no. 1 Part B, pp. 579–591, 2022.

[18] I. Shahrivar, Z. Niazi, A. Khoshoei, and A. Ghafouri, “A semi-experimental model to predict the thermal conductivity coefficient of nanofluids,” Heat and Mass Transfer, pp. 1–9, 2022.

[19] H. Yasmin, S. O. Giwa, S. Noor, and M. Sharifpur, “Thermal conductivity enhancement of metal oxide nanofluids: A critical review,” Nanomaterials, vol. 13, no. 3, p. 597, 2023.

[20] G. Coccia, S. Tomassetti, and G. Di Nicola, “Thermal conductivity of nanofluids: A review of the existing correlations and a scaled semi-empirical equation,” Renewable and Sustainable Energy Reviews, vol. 151, p. 111573, 2021.

[21] H. Zou et al., “A neural regression model for predicting thermal conductivity of cnt nanofluids with multiple base fluids,” Journal of Thermal Science, vol. 30, pp. 1908–1916, 2021.

[22] R. Pourrajab, I. Ahmadianfar, M. Jamei, and M. Behbahani, “A meticulous intelligent approach to predict thermal conductivity ratio of hybrid nanofluids for heat transfer applications,” J Therm Anal Calorim, vol. 146, no. 2, pp. 611–628, 2021.

[23] E. B. Elcioglu, “A High-Accuracy Thermal conductivity model for water-based graphene nanoplatelet nanofluids,” Energies (Basel), vol. 14, no. 16, p. 5178, 2021.

[24] E. LeDell and S. Poirier, “H2o automl: Scalable automatic machine learning,” in Proceedings of the AutoML Workshop at ICML, 2020.