Hybrid Nanofluid Density Prediction Using Regression Models

Deepak Binkam
Towson University

***Abstract*—Hybrid nanofluids have important thermal properties that are critical in technologies such as automotives, engine cooling, heat exchangers, and similar thermal management technology. Accurate density prediction of these nanofluids is essential for optimizing the design and safety of this technology. This paper analyzes an experimental dataset of 436 samples, sourced from Kaggle [1], to predict hybrid nanofluid density. This paper includes an overview of the dataset, preprocessing and feature engineering (such as the creation of polynomial features), and regression modeling using Linear Regression, Decision Tree Regression, and Random Forest Regression. In addition, various visualizations are presented to reveal insights. Finally, alternative approaches and future iterations are discussed that may further improve these models.**

***Index Terms*—Hybrid Nanofluids, Density Prediction, Data Preprocessing, Feature Engineering, Regression Models.**

I. Introduction

Hybrid nanofluids are produced by dispersing two or more types of nanoparticles into a base fluid, resulting in improved thermal properties. Their use in heat exchangers, automotive cooling systems, and electronic thermal management creates the need for precise density prediction, as density directly impacts fluid dynamics and heat transfer efficiency. This study presents a comprehensive approach to predicting hybrid nanofluid density using regression models. The experimental dataset from Kaggle [1] is examined, preprocessing and feature engineering methods are detailed, and visualizations are included. The performance of the regression models are then evaluated.

II. Dataset Description

The dataset, obtained from Kaggle [1], comprises 436 experimental samples extracted from peer-reviewed literature. The dataset includes several input parameters and a continuous target variable representing the nanofluid density.

A. Input Parameters

- **Base Fluid:** A categorical variable that indicates the type of base fluid used in the mixture.
- **Nano Particle:** A categorical variable that indicates the blend of nanoparticles in the mixture.
- **Temperature (°C):** A continuous variable representing the operating temperature.
- **Volume Concentration (ϕ):** A continuous variable representing the volume fraction of nanoparticles.
- **Density of Base Fluid (ρ_{bf}):** The density of the base fluid (in g/cm^3).
- **Density of Nano Particle 1(ρ_{np1}):** The density of the primary nanoparticle (in g/cm^3).
- **Density of Nano Particle 2(ρ_{np2}) :** The density of the second nanoparticle (in g/cm^3).
- **Volume Mixture of Particle 1:** The volume percentage of nanoparticle 1 in the mixture.
- **Volume Mixture of Particle 2:** The volume percentage of nanoparticle 2 in the mixture.

B. Output Parameter

- **Density (ρ):** The overall density of the hybrid nanofluid, which is the target variable.

C. Derived Features

To improve predictive performance, two features were engineered:

- **Density Ratio:** Computed as ρ/ρ_{bf}, it normalizes the overall density by the density of the base fluid.
- **Particle1 Ratio:** This feature is defined as

$$\frac{\text{Volume Mixture of Particle 1}}{\text{Volume Mixture of Particle 1} + \text{Volume Mixture of Particle 2}}$$

which indicates the relative contribution of the primary nanoparticle over both particles.

III. Data Preprocessing and Feature Engineering

A series of preprocessing and feature engineering steps were performed to prepare the data for modeling. Since the dataset is peer-reviewed, outliers are assumed to show real experimental variability and are kept.

A. Outlier Retention and Skewness

Outlier analysis indicated significant outlier counts in certain density features. Since these outliers likely capture real conditions, no outlier removal was performed. Similarly, the skewness was computed for all numerical features. Some density-related variables exhibit high skewness, potentially affecting model assumptions. Although transformations like the Box-Cox and log1p transformation were considered, because of the peer reviewed data, no transformations were applied.

B. Polynomial Feature Generation

Degree-2 polynomial features were generated from Temperature and Volume Concentration to capture potential non-linear interactions. The resulting features include:

- $(\text{Temperature})^2$,
- Temperature $\times$ ϕ, and

- $(\phi)^2$.

These additional features help regression models to account for more complex relationships in the data.

C. Scaling and Encoding

Numerical variables were scaled using a RobustScaler. Categorical variables (Base Fluid and Nano Particle) were one-hot encoded, with the first category dropped.

IV. Exploratory Data Visualizations

A variety of visualizations were produced to gain insights into the data for feature engineering, a few are discussed:

A. Correlation Visualizations

Two correlation heatmaps were created:

- An annotated heatmap using `seaborn`, with correlation values.
- A secondary heatmap using `matplotlib`'s `imshow` function.

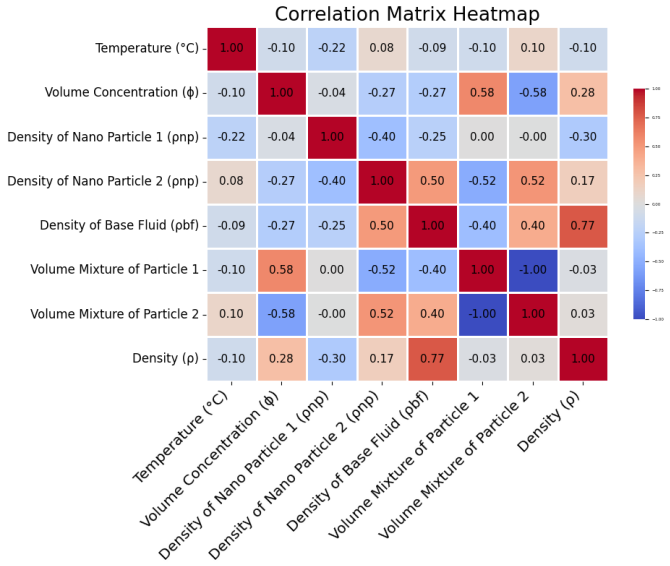

Fig. 1: The heatmap shown using seaborn shows some moderately linear relationships, but a lot are non-linear.

B. Scatter Plots

Scatter plots were generated to explore pairwise relationships:

- Temperature vs. Density, revealing an expected negative relationship.
- Volume Concentration vs. Density, indicating a moderate positive association.
- Density of Base Fluid vs. Overall Density, demonstrating a strong positive correlation.

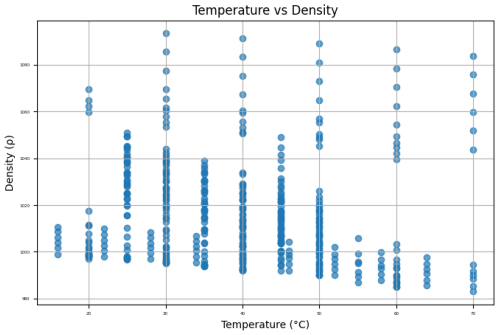

Fig. 2: This scatterplot of temprature vs density shows a negative correlation.

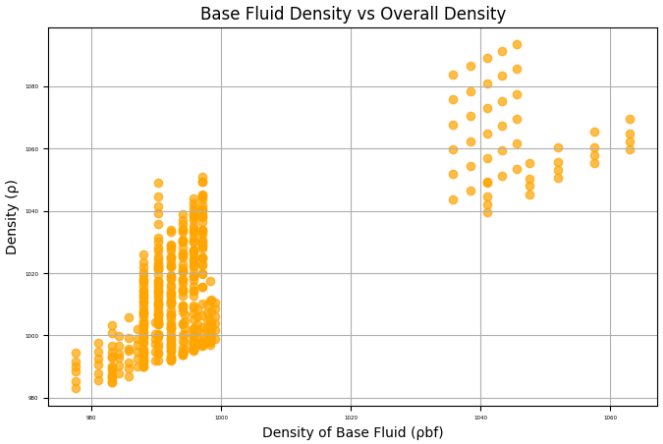

Fig. 3: The scatter of Base vs Overall Density shows a positive correlation.

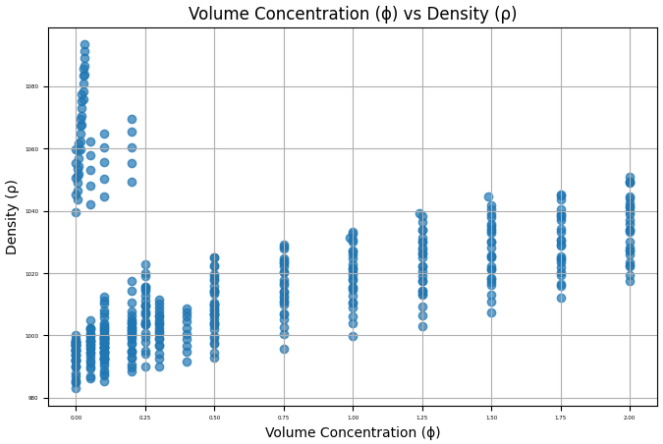

Fig. 4: This scatter of Volume vs Density shows they are lightly positively correlated.

C. Histograms

Histograms were used to examine distributions of key variables:

- A histogram of the target variable (Density) provided an overview of its distribution.
- A histogram of Volume Concentration assessed its spread and symmetry.

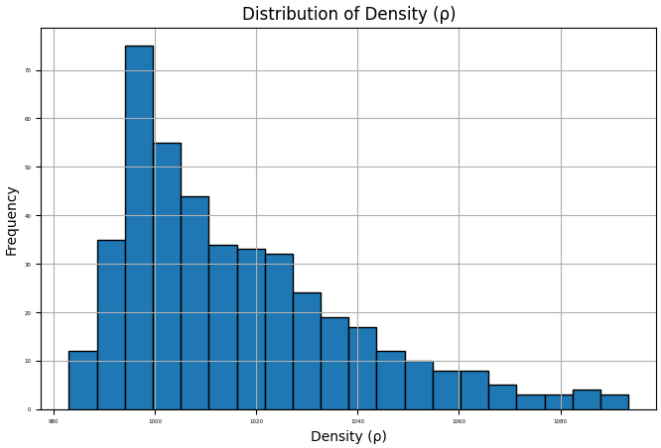

Fig. 5: This histogram shows that Density is right skewed.

D. Boxplots

Boxplots were constructed to compare distributions across categorical groups:

- A boxplot of Density by Nano Particle Type highlighted differences between groups.
- A boxplot of Temperature by Nano Particle Type.

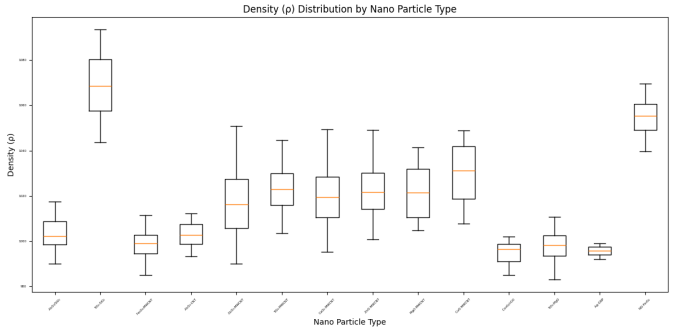

Fig. 6: A boxplot of each nano particle type based on density. This boxplot makes it easier to see variability in density per nano particle.

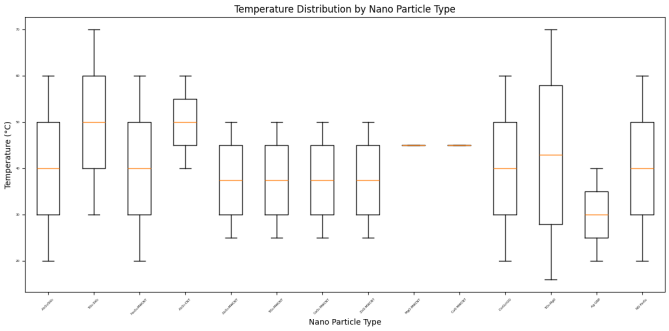

Fig. 7: A boxplot of each nano particle type based on temperature. The data shows particles MgO-MWCNT and CuO-MWCNT have low temperature variabliity, and TiO2-MgO has the most.

E. Scatter Matrix and Bar Chart

Additional visualizations included:

- A scatter matrix (pair plot) for numerical features to observe pairwise trends and potential non-linear relationships.
- A bar chart illustrating Mean Density by Nano Particle Type, offering insight into the influence of nanoparticle composition.

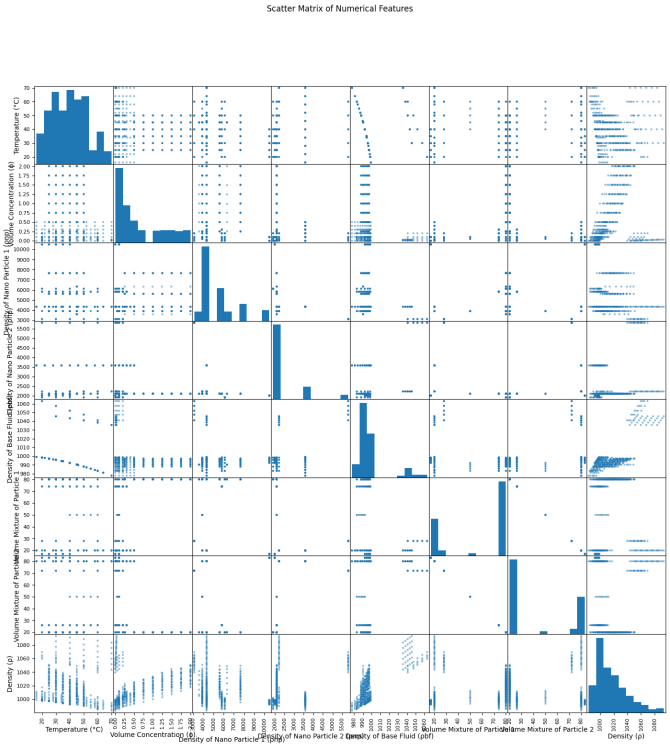

Fig. 8: This scatter matrix more easily allows the identification of any kind of linear or pair-wise relationships.

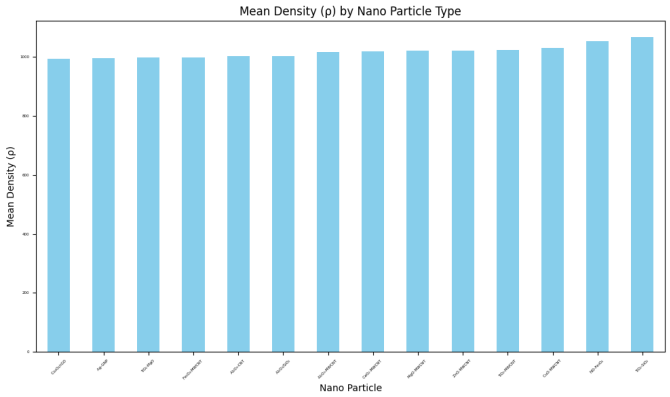

Fig. 9: This Bar graph shows the average mean of each nano particle, showing the order from least to most dense.

V. Regression Modeling and Experimental Evaluation

The preprocessed dataset was partitioned into 80% training and 20% testing sets. Three regression models were evaluated:

A. Linear Regression

Linear Regression models the relationship between the features and the target variable as a linear combination of the features. This method estimates coefficients that minimize the sum of squared errors. It is more simple and interpretable, which makes it a useful baseline model.

B. Decision Tree Regression

Decision Tree Regression recursively partitions the feature space into smaller regions based on input variables, using decision rules derived from the data. It is good at capturing non-linear relationships without requiring transformations beforehand, making it suitable for complex relationships like the one in this paper.

C. Random Forest Regression

Random Forest Regression is an ensemble method that builds multiple decision trees using samples of the data and averages their predictions. This approach reduces overfitting and enhances generalization. However, the 5-fold cross-validation suggests that further hyperparameter tuning is needed.

D. Evaluation Metrics and Results

The performance of each model was evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), and the R^2 score. A 5-fold cross-validation was also performed for the Random Forest. The results are summarized in Table I.

TABLE I: Regression Model Evaluation Metrics

Model	Test MSE	Test MAE	Test R^2	5-Fold CV MSE
Linear Regression	0.02	0.10	1.00	–
Decision Tree	5.42	1.73	0.99	–
Random Forest	4.59	1.39	0.99	186.97

VI. Results Discussion

Table I summarizes the evaluation metrics of the regression models. The Linear Regression model had a Test MSE of 0.02, Test MAE of 0.10, and Test R^2 of 1.00. This suggests that the relationship between the predictors and the target variable could be highly linear. However, it is likely that the model has over simplified the representation of nano particle physics.

The Decision Tree Regression model produced slightly higher errors (Test MSE of 5.42, Test MAE of 1.73, and a Test R^2 of 0.99). This performance indicates that while the decision tree is able to capture non-linear interactions, it introduces a small increase in prediction error compared to the linear model.

The Random Forest Regression model also demonstrated comparable performance on the test set, with a Test MSE of 4.59, Test MAE of 1.39, and Test R^2 of 0.99. However, the 5-fold cross-validation revealed an average MSE of 186.97, showing high variability across different data splits. This means there could be a potential overfitting issue in the Random Forest model, indicating that further hyperparameter tuning (like adjusting the number of trees, maximum tree depth, or minimum samples) could enhance its ability to generalize.

Overall, the evaluation reveals a trade-off between model simplicity and the ability to capture complex, non-linear interactions. While Linear Regression's performance appears perfect on the test set, its simplicity likely masks the underlying complexities of the data. Conversely, ensemble methods like Random Forest can capture more intricate relationships, but are sensitive to parameter settings as seen by the high variability in cross-validation. Future iterations should incorporate hyperparameter optimization and additional analysis to ensure models are robust and generalizable in their performance.

VII. Conclusion

A comprehensive study has been conducted to predict hybrid nanofluid density using regression models. Based on an experimental, peer-reviewed dataset from Kaggle, detailed exploratory analysis, preprocessing, and feature engineering were performed. The retention of outliers preserved genuine experimental variability, while polynomial feature generation captured non-linear relationships. The evaluation of Linear Regression, Decision Tree, and Random Forest models indicated robust performance on the test set, but the variability seen in the cross-validation test suggests that further model tuning is needed. This future work can include an extensive grid or random search to optimize model parameters for ensemble models. Other models, such as Support Vector Regression and Gradient Boosting, can also be evaluated to compare performance.

References

[1] A. I4A Lab, "Nanofluid Density Prediction," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/ai4a-lab/nanofluid-density-prediction