Hybrid Nanofluid Density Prediction Using Clustering

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 study clusters 436 hybrid-nanofluid experiments using K-Means, Agglomerative, and DBSCAN. After feature engineering, an ablation test was done that brought to light two redundant ratio features, K-Means and Agglomerative achieved a silhouette of 0.632 with four clusters, while a tuned DBSCAN ($\epsilon = 2.5$, minPts=10) reached 0.350. These clusters reveal temperature-concentration groups based on nanoparticle composition.**

***Index Terms*—Hybrid Nanofluids, Density Prediction, Data Preprocessing, Feature Engineering, Clustering 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 clustering. The experimental dataset from Kaggle [1] is examined, preprocessing and feature engineering methods are detailed, and visualizations are included. The performance of the clustering 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
- $(\phi)^2$.

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.

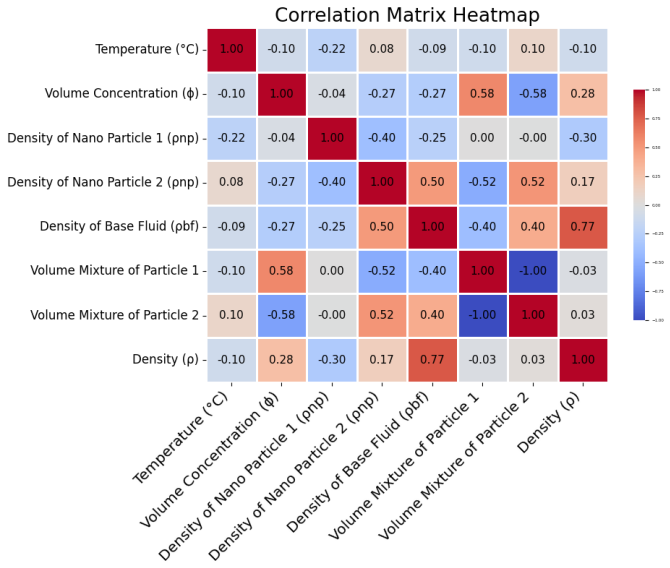

Fig. 1: The heatmap shown using seaborn shows some moderately linear relationships, but a lot are non-linear.

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.

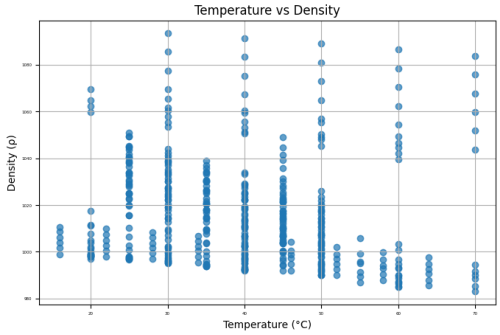

Fig. 2: This scatterplot of temprature vs density shows a negative correlation.

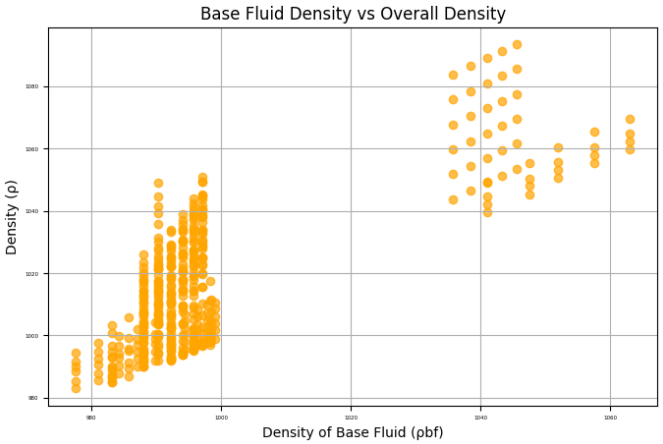

Fig. 3: The scatter of Base vs Overall Density shows a positive correlation.

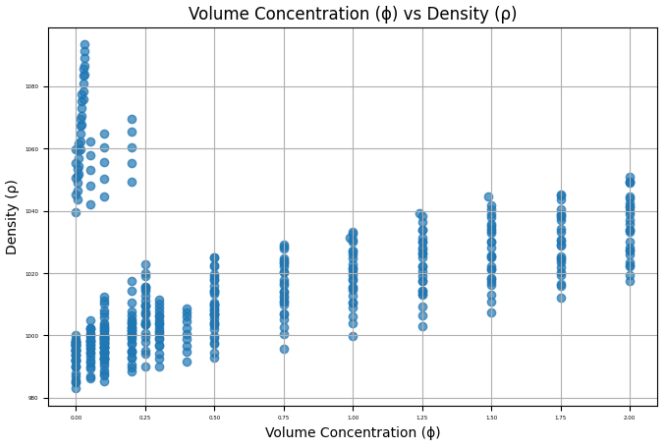

Fig. 4: This scatter of Volume vs Density shows they are lightly positively correlated.

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.

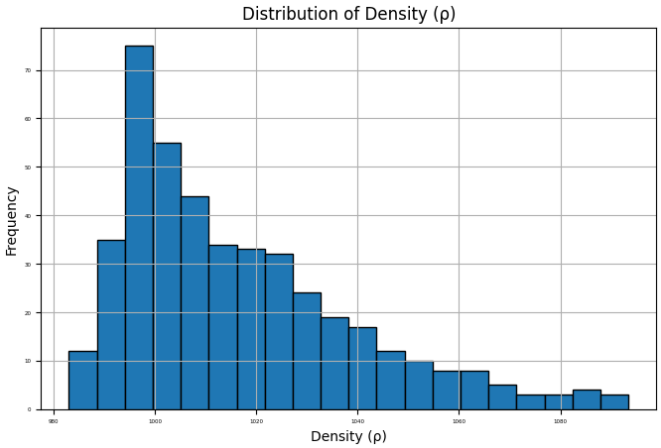

Fig. 5: This histogram shows that Density is right skewed.

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.

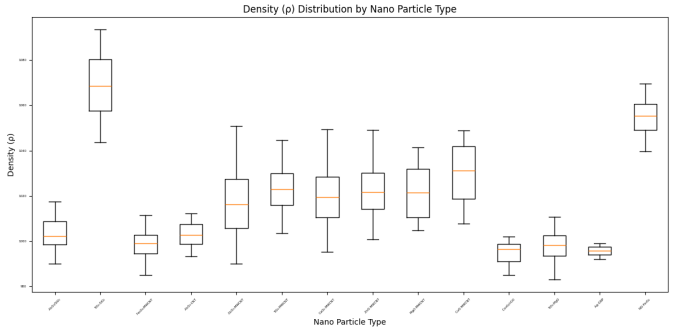

Fig. 6: A boxplot of each nano particle type based on density. This boxplot makes it easier to see variability in density per nano particle.

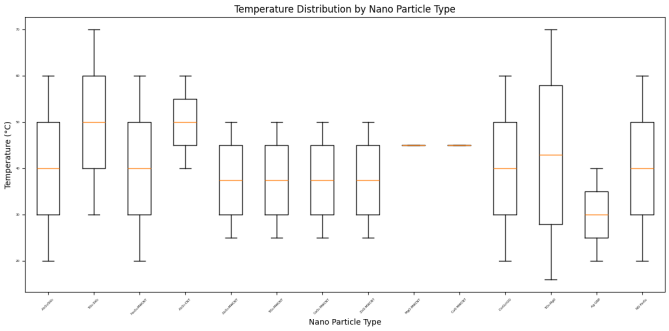

Fig. 7: A boxplot of each nano particle type based on temperature. The data shows particles MgO-MWCNT and CuO-MWCNT have low temperature variabliity, and TiO2-MgO has the most.

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.

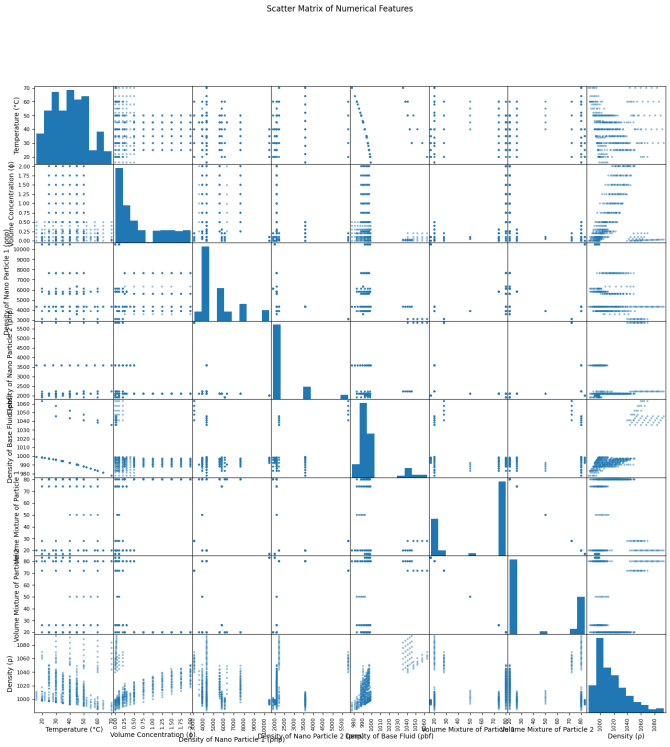

Fig. 8: This scatter matrix more easily allows the identification of any kind of linear or pair-wise relationships.

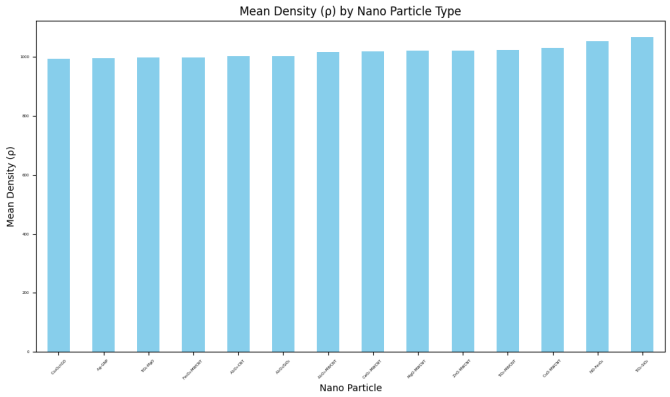

Fig. 9: This Bar graph shows the average mean of each nano particle, showing the order from least to most dense.

V. Clustering Methodology

While the previous paper with regression models quantify how input features predict nanofluid density, using clustering reveals natural groupings in the experimental data, identifying groups of temperature, concentration, and nanoparticle composition that behave similarly. This can be useful to guide targeted experiments. Variance analysis showed Density Ratio (variance $\approx$ 2 × 10-4) and Particle1 Ratio were making the models worse, and so these columns were excluded from the final data. Three clustering models were evaluated:

Three clustering models were evaluated on the processed dataset.

A. Clustering Overview

After feature scaling and encoding, clustering was applied to the 28-dimensional feature space, after excluding the Density Ratio and Particle1 Ratio engineered features. The goal was to group nanofluid samples based on similar thermophysical properties without using the target density label.

B. K-Means

K-Means clustering partitions data into k groups by minimizing variance in clusters. Models were trained for $k = 2$ to $k = 10$. Silhouette scores and Davies-Bouldin Indices were computed to determine the optimal number of clusters, which was found to be $k = 4$.

C. Agglomerative

Agglomerative clustering is a hierarchical method that builds nested clusters via a bottom-up approach. Using Ward linkage and Euclidean distance, it was also trained for $k = 2$ to $k = 10$. $k = 4$ also had the best silhouette performance.

D. DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) identifies clusters based on dense regions of points. First it was run with $\epsilon = 1.5$, minPts = 5, which produced clusters that were fragmented and did not connect well. A grid search was conducted over $\epsilon \in [1.5, 2.0, 2.5, 3.0]$ and minPts $\in \{5, 8, 10\}$, revealing $\epsilon = 2.5$, minPts=10 was the best.

E. Hyperparameter Search

For K-Means and Agglomerative, silhouette coefficients were maximized across k values. For DBSCAN, the model was selected based on maximizing the silhouette score after excluding noise.

F. Validation Metrics

Two metrics were computed for the cluster outputs:

- Silhouette Coefficient: measures how similar an object is to its own cluster versus other clusters.
- Davies-Bouldin Index: measures intra-cluster compactness and separation, and finds the average worst case similarity.

Principal Component Analysis (PCA) was used to project the clusters into two dimensions for visualization.

VI. Results and Discussion

A. Clustering Results

Table I summarizes the best silhouette scores achieved by each model.

TABLE I: Clustering Model Performance

Model	Number of Clusters	Silhouette Score
K-Means	4	0.632
Agglomerative	4	0.632
DBSCAN	5	0.350

K-Means and Agglomerative models produced four compact and well-separated clusters. Tuned DBSCAN produced five clusters and classified some samples as noise.

B. Feature Ablation Impact

Feature ablation, where Density Ratio and Particle1 Ratio were removed, led to an improvement in silhouette score from 0.623 to 0.632, confirming that these engineered features were negatively affecting clustering.

C. Cluster Interpretation

Clusters separated samples according to temperature and volume concentration bands. Cluster 0 corresponded to fluids of low-temperature, low-concentration. Cluster 3 contained high-temperature, high-concentration fluids. The DBSCAN noise occurred in the extreme temperatures.

D. PCA Visualization

Figures 10 and 11 show PCA projections of the Agglomerative and K-Means clusters shows that there is clear separation, and proper clusters. Figure 12 shows the sparser clustering of DBSCAN.

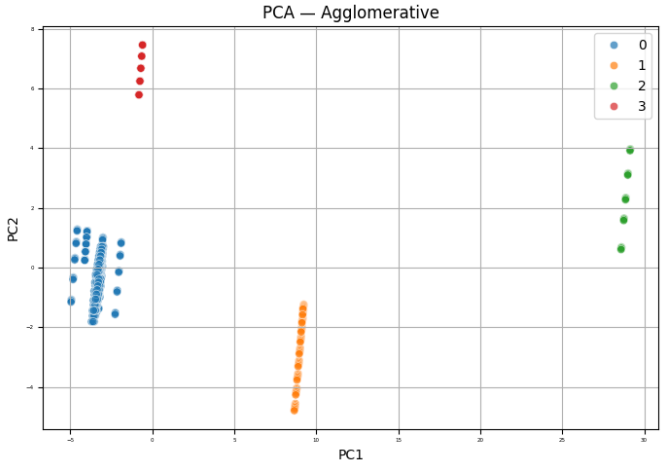

Fig. 10: PCA Projection of Agglomerative Clusters shows that similar to K-Means, it has grouped the data well. There is clear separation.

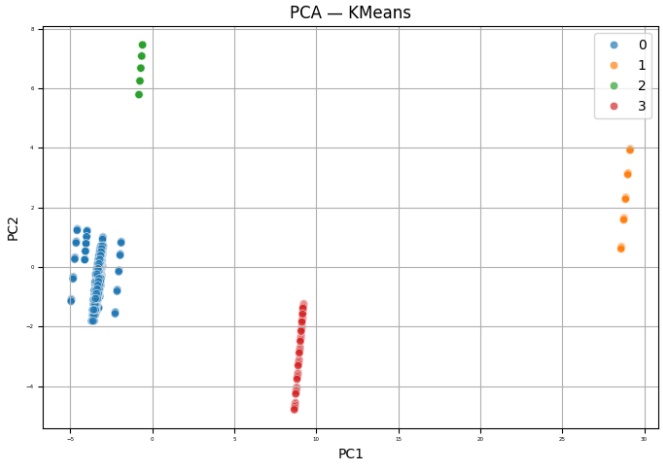

Fig. 11: PCA Projection of K-Means Clusters shows a similar result to Agglomerative, it has proper clusters.

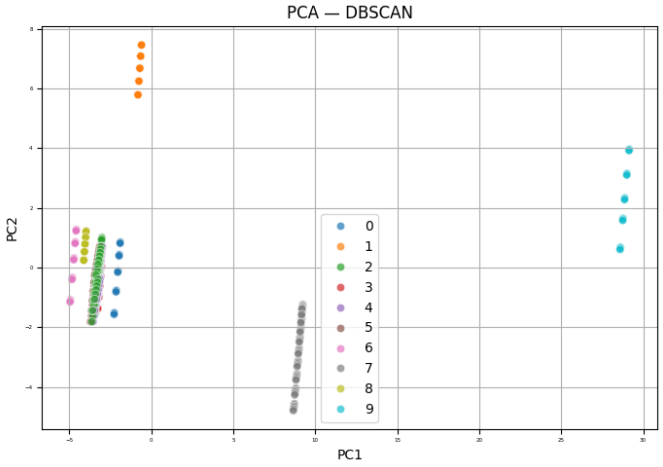

Fig. 12: PCA Projection of DBSCAN Clusters shows that it is struggling to place data properly into 10 clusters, as seen on the lower left side of the plot. There is a cluster of red (3) data directly under another cluster of green (2) data.

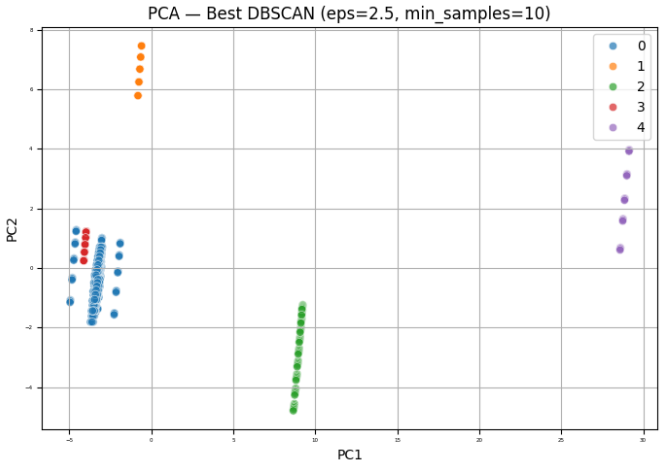

Fig. 13: This PCA Projection of the best DBSCAN Clusters shows an improvement to the base DBSCAN. There are now 5 clusters and it is much cleaner.

VII. Conclusion

In this study, clustering was applied to a hybrid nanofluid dataset to explore underlying density clusters. After preprocessing, feature engineering, and feature ablation, K-Means and Agglomerative clustering models each achieved a silhouette score of 0.632 with four proper clusters. A tuned DBSCAN found five clusters with a silhouette score of 0.350. Feature ablation confirmed that Density Ratio and Particle1 Ratio made clustering performance worse. Future work can use HDBSCAN to compete with DBSCAN and find a better number of clusters.

References

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