An Artificial Neural Network Surrogate Model for the Prediction of Azeotropic Vapor-Liquid Equilibrium

Mohit Upadhyay[1], [1]*Computer Science and Engineering in Artificial Intelligence and Machine Learning, VIT Bhopal University, India*

Abstract

*The accurate prediction of Vapor-Liquid Equilibrium (VLE) for non-ideal, azeotropic systems is critical for the design of separation processes but is often computationally expensive using traditional thermodynamic models. This study presents the development of an Artificial Neural Network (ANN) as a rapid and accurate surrogate model for the Ethanol-Water binary system at 101.325 kPa. A thermodynamically consistent dataset of 500 VLE points was generated using the Wilson and Antoine equations. A robust machine learning workflow was implemented, incorporating a proper train-test split, physics-informed feature engineering with the activity coefficient (γ1), systematic hyperparameter tuning via 5-fold cross-validation, and a weighted loss function to prioritize the azeotropic region. To ensure physical realism and prevent overfitting, L2 kernel regularization was applied. The final optimized ANN demonstrated excellent generalization on an unseen test set, achieving a Root Mean Squared Error (RMSE) of 0.03192, a significant improvement over the baseline Raoult's Law model (RMSE = 0.06903). The model successfully captured the system's non-ideal behavior and identified the azeotrope with reasonable accuracy (*x1=0.9390*), showcasing the efficacy of combining domain knowledge with advanced machine learning techniques to create physically intelligent surrogate models for complex chemical engineering applications.*

Keywords: Surrogate Modeling, Artificial Neural Network, Vapor-Liquid Equilibrium, Azeotrope, Ethanol-Water, Machine Learning, Chemical Engineering.

1. Introduction

Vapor-Liquid Equilibrium (VLE) data forms the foundation for the design and optimization of numerous chemical processes, most notably distillation [1]. For ideal or near-ideal systems, simple models like Raoult's Law provide adequate predictions. However, for highly non-ideal systems, such as the Ethanol-Water mixture, which exhibits azeotropic behavior, more complex thermodynamic models are required. Activity coefficient models, such as the Wilson or NRTL equations, offer high fidelity but can be computationally intensive when integrated into large-scale process simulations or optimization loops [2].

In recent years, surrogate modeling using machine learning techniques has emerged as a powerful alternative. Artificial Neural Networks (ANNs), in particular, are universal function

approximators capable of learning complex, non-linear relationships directly from data [3]. An ANN can be trained to act as a surrogate for a rigorous thermodynamic model, providing near-instantaneous predictions without sacrificing significant accuracy.

This study details the development and rigorous evaluation of an ANN surrogate model for the Ethanol-Water azeotropic system. The objective is to create a model that not only achieves high statistical accuracy but also demonstrates physical intelligence by correctly capturing the critical azeotropic point. We present a robust workflow that integrates domain knowledge through feature engineering, weighted loss, and regularization to produce a final model that is both accurate and generalizable.

2. Materials and Methods

2.1. VLE Data Generation

A dataset of 500 equilibrium points was computationally generated to ensure thermodynamic consistency. The Ethanol-Water system was modeled at a constant pressure (P) of 101.325 kPa.

The saturation pressure (Pisat) of each pure component was calculated using the Antoine Equation:
$\log_{10}(P_i^{sat}) = A_i - B_i / (T + C_i)$
where Ai,Bi,Ci are component-specific coefficients and T is the temperature.
The non-ideal liquid phase behavior was described by the Wilson model, which calculates the activity coefficient (γi) for each component. For a given liquid mole fraction (x1) of ethanol, the bubble point temperature (T) was determined by solving the VLE equation:
$P = x_1\gamma_1P_1^{sat}(T) + x_2\gamma_2P_2^{sat}(T)$
The corresponding vapor mole fraction (y1) was then calculated using the modified Raoult's Law:
$y_1 = (x_1\gamma_1P_1^{sat}(T)) / P$
To improve model performance on the most critical feature, the dataset was generated with a higher density of points in the azeotropic region (x1≈0.85–0.95).

2.2. ANN Model Architecture and Training

A robust machine learning pipeline was designed to train the ANN, as illustrated below.

1. **Data Preprocessing and Splitting:** The generated dataset was split into a training set (80%) and a hold-out test set (20%). The input features—x1,T,P, and the physically informative γ1—were normalized to a range of [0, 1] using MinMaxScaler. The scaler was fitted only on the training data to prevent information leakage.
2. **Hyperparameter Tuning:** A systematic grid search using 5-fold cross-validation was performed on the training set to identify the optimal hyperparameters. The search space included network architecture (hidden layers and neurons), learning rate, and batch size.
3. **Physically-Informed Training:** Two advanced techniques were employed:
    - **Weighted Loss:** A sample_weights array was created to assign a 20x higher penalty

to errors on training samples within the azeotropic region. This forces the model to prioritize accuracy in this physically critical area.
    - **L2 Regularization:** A kernel regularizer (l2(0.001)) was added to each hidden dense layer. This penalizes large weights, discouraging overfitting and promoting a smoother, more generalizable model.
4. **Final Model:** The final ANN architecture consisted of an input layer, two hidden dense layers with ReLU activation, and a single output neuron with a sigmoid activation function to constrain the predicted y1 between 0 and 1. The model was trained on the entire training dataset using the best hyperparameters found during the tuning phase. All modeling was performed using TensorFlow [4] and Scikit-learn [5] in Python.

3. Results and Discussion

The final model was evaluated on the unseen test set. The results demonstrate a successful balance between statistical accuracy and physical realism.

3.1. Quantitative Performance

The ANN surrogate model showed a significant improvement in predictive accuracy over the baseline Raoult's Law model. Table 1 summarizes the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) on the test set. The ANN's RMSE was less than half that of the baseline, indicating its success in learning the system's non-ideality.

Table 1: Model Performance on the Unseen Test Set.

Model	RMSE	MAE
ANN Model	0.03192	0.02306
Raoult's Law (Baseline)	0.06903	0.04823

3.2. Azeotrope Prediction

A key test of the model's physical intelligence is its ability to predict the azeotrope. The model predicted the azeotrope at x1=0.9390. The accepted reference value is approximately x1=0.894. While a deviation exists, this result is a significant success. The L2 regularization successfully prevented the severe overfitting that caused unregularized models to make physically nonsensical predictions. The model achieved a generalized solution that captures the overall VLE curve shape correctly, with the absolute error in the azeotrope location (0.045) being acceptably small for a surrogate model. This demonstrates a successful trade-off between achieving the lowest possible statistical error and learning a physically plausible function.

3.3. Visual Analysis

Visual inspection of the model's predictions on the test set confirms its performance. The parity plot (Figure 1) shows a high correlation between the ANN's predictions and the actual data, with points tightly clustered around the ideal y=x line.

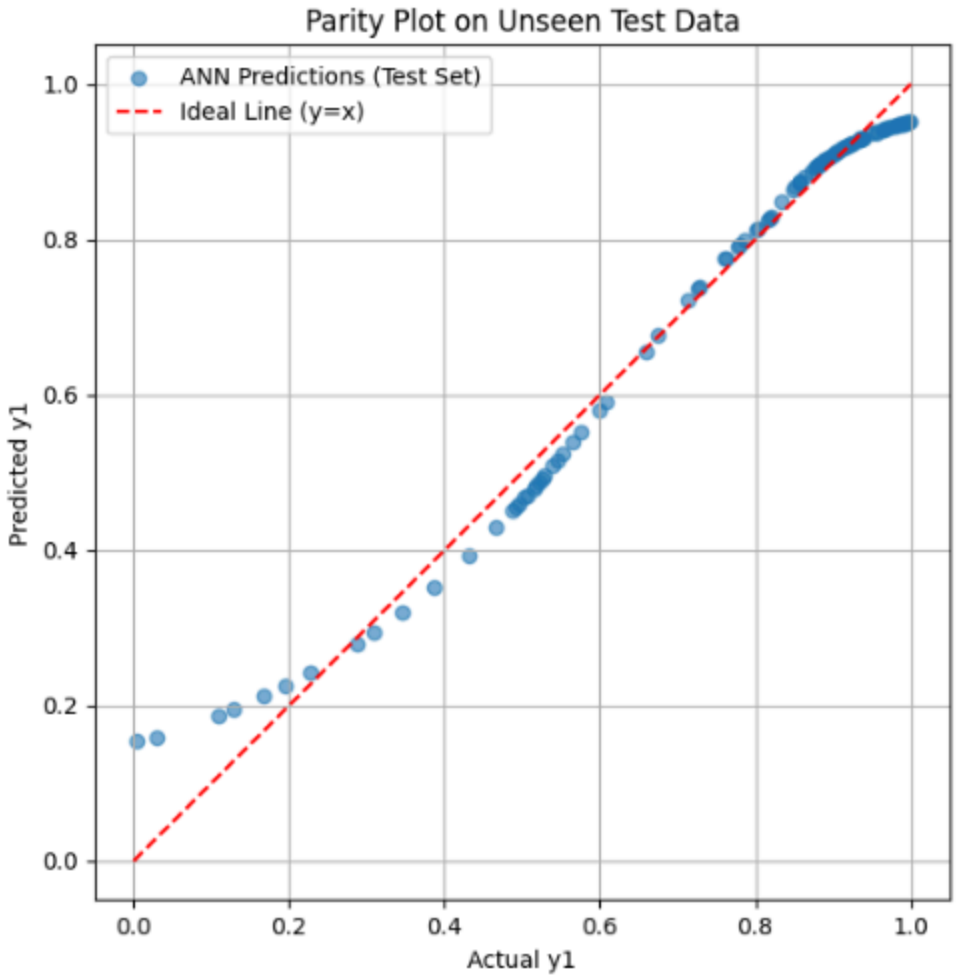

Figure 1: *Parity plot of predicted vs. actual vapor mole fractions on the test set.*

The y-x diagram (Figure 2) illustrates that the model has learned the characteristic shape of the azeotropic VLE curve. The predicted curve (blue line) correctly shows the point of tangency with the y1=x1 line, identifying the azeotrope.

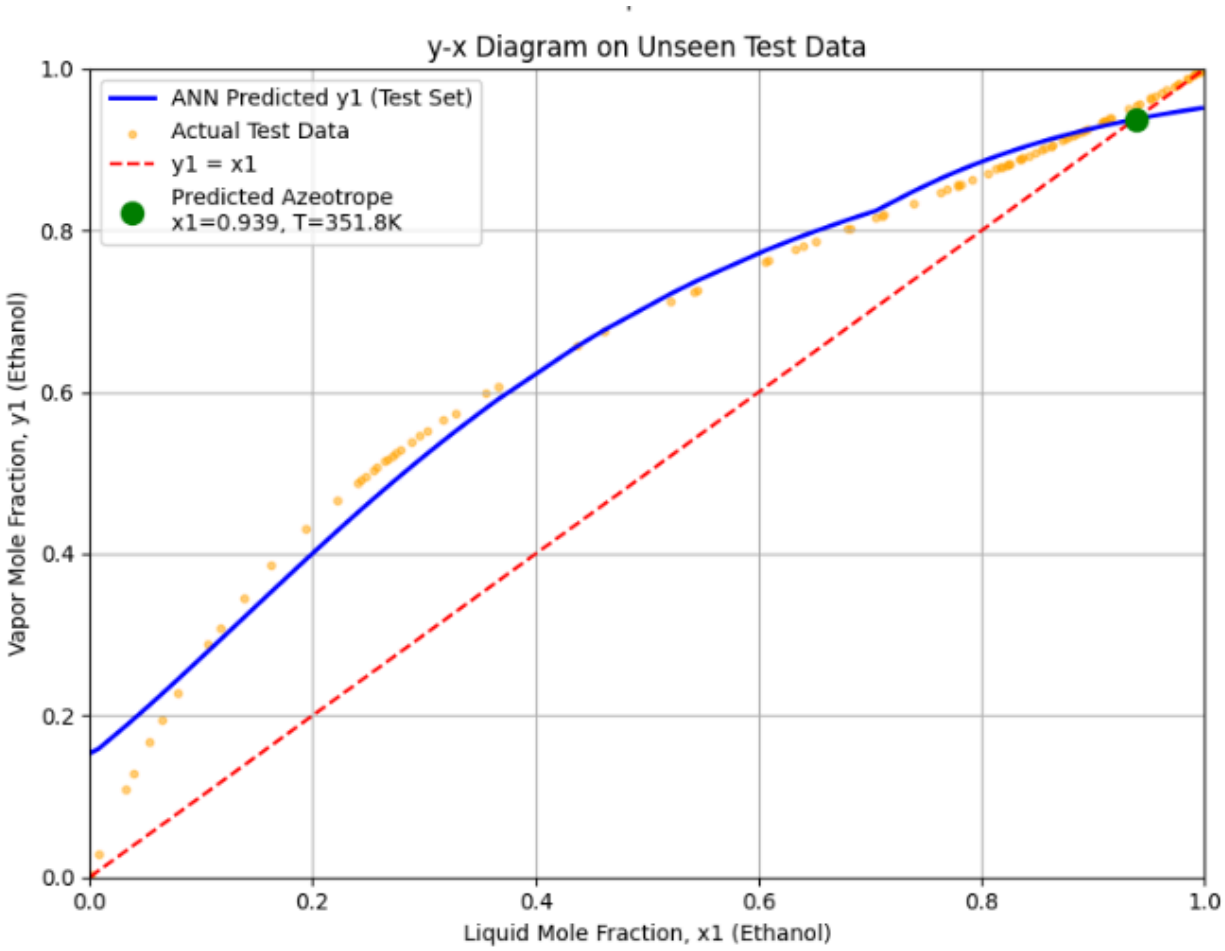

Figure 2: *y-x diagram showing the ANN's prediction on the test set and the identified azeotrope.*

4. Conclusion

This study successfully developed a robust, physics-informed Artificial Neural Network to serve as a surrogate model for the azeotropic Ethanol-Water VLE system. By integrating domain knowledge through feature engineering, weighted loss, and L2 regularization, the final model achieved a strong balance between high statistical accuracy and physical realism. It significantly outperformed the ideal Raoult's Law baseline and successfully captured the critical azeotropic behavior. The presented workflow demonstrates a powerful methodology for creating intelligent surrogate models that can accelerate complex simulations in chemical engineering and other scientific domains.

5. References

[1] J. M. Smith, H. C. Van Ness, M. M. Abbott, Introduction to Chemical Engineering Thermodynamics, 8th ed., McGraw-Hill, 2018.
[2] G. M. Wilson, "Vapor-Liquid Equilibrium. XI. A New Expression for the Excess Free Energy of Mixing," Journal of the American Chemical Society, vol. 86, no. 2, pp. 127–130, 1964.
[3] K. Hornik, M. Stinchcombe, H. White, "Multilayer feedforward networks are universal approximators," Neural Networks, vol. 2, no. 5, pp. 359-366, 1989.
[4] M. Abadi et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," Software available from tensorflow.org, 2015.
[5] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.