Imperial College London

Department of Earth Science and Engineering

MSc in Applied Computational Science and Engineering

Independent Research Project
Final Report

Explainable AI: understand the black box of predictive models with chemical engineering applications

by

Jinwei Hu

Email: jinwei.hu22@imperial.ac.uk
GitHub username: acse-jh4322
Repository: https://github.com/ese-msc-2022/irp-jh4322

Supervisors:

Dr. Sibo Cheng
Dr. Rossella Arcucci

August 2023

Abstract

In the field of chemical engineering, understanding the dynamics and probability of drop coalescence is not just an academic pursuit but a fundamental requirement for advancing process design and enhancement of their efficiencies. This research applies machine learning (ML) predictive models to decipher the intricate relationships embedded in the experimental data on drop coalescence in microfluidic device. Through the deployment of SHapley Additive exPlanations values (SHAP), critical features relevant to coalescence processes are consistently identified. Comprehensive feature ablation tests further delineate the robustness and susceptibility of each model. Furthermore, the incorporation of Local Interpretable Model-agnostic Explanations (LIME) for local interpretability offers an elucidative perspective, clarifying the intricate decision-making mechanisms inherent to each model's predictions. As a result this research not only provides the relative importance of the features for the outcome of drop encounter but also emphasises the important role of thorough model interpretation, asserting the pivotal role of model interpretability in reinforcing confidence in analytical predictions within chemical engineering.

1 Introduction

Microfluidic technologies have caused a significant paradigm shift in the manipulation and analysis of fluids across a range of fields, including chemistry, biology, and material science [21]. One key operation in this area is merging droplets in a fluid, also known as droplet coalescence. This operation has been widely studied and different methods, like using electric fields, have been developed to make it happen [44, 28, 31, 15, 14, 29, 70, 30, 33]. Microfluidics serve as a unique platform for studying coalescence under well-controlled conditions and find applications in microfluidic synthesis and analysis as well [71, 63]. Beyond academia, droplet coalescence plays an integral role in various industries applications, such as petroleum refining where it aids in separating water from crude oil [17], and in emulsion formulation, which is crucial in sectors like food processing, cosmetics, pharmaceutics, oil recovery, transport, and separation processes [47, 9, 60, 1]. Despite considerable research in this domain, the transition from theoretical understanding to real-world application is still impeded, primarily due to challenges in achieving reliable passive coalescence [19, 42]. The complexity of passive coalescence arises from the numerous conditions involved, such as fluid viscosity, droplet contact time and interfacial tension, making it difficult to achieve reliable outcomes without external interventions [39].

Therefore, accurately controlling droplet coalescence is not merely a convenience but a essential necessity, especially in fabricating complex materials used in some critical industries [66, 23]. Nonetheless, the dynamic nature of droplet coalescence, along with its potential variability under varying conditions, presents obstacles to control coalescence phenomena precisely. For instance, higher temperatures can amplify coalescence frequency, as detailed by Bera et al. [7]. Similarly, changes in fluid viscosity or flow rate can unexpectedly affect droplet sizes and coalescence occurrence, requiring real-time adjustments to maintain desired outcomes [11]. When these experimental conditions are subject to simultaneous variations, researchers encounter significant difficulties in accurately predicting coalescence outcomes, let alone deciphering the underlying reasons behind the coalescence phenomena.

In microfluidic systems, these dynamics are further complicated. The microenvironment in these devices affects droplet coalescence through an intricate balance of multiple factors. These variations could alter the chemical attributes of coalesced droplets, affecting subsequent processes and the quality of end applications [34]. Therefore, mastering the prediction and control of droplet coalescence is critical for industries that rely on specific emulsions and optimized

processes, such as drug delivery systems and material synthesis [26]. Consequently, through the research and studies performed in this field, they underscores the continued importance of understanding and predicting droplet coalescence in microfluidics, given its broad applications across diverse industries, attesting to its universal appeal and relevance [21, 7].

Traditionally, the analysis of factors influencing droplet coalescence is depended on trial-and-error methods or analytical models [62, 50]. Yet, the non-linear dynamics of droplet coalescence, caused by experimental variables such as flow rates and interfacial tension, are difficult to capture using traditional methods and observations [3]. Serving as the antithesis to conventional methodologies, Machine Learning (ML) functions as a powerful tool for modelling intricate systems and proffering predictions for dimensional data [27, 74, 45]. Owing to their ability to assimilate information from data, machine learning algorithms excel in capturing sophisticated relationships that traditional methodologies are impotent to attain [12]. In the domain of chemical engineering, machine learning algorithms are being increasingly enlisted. They provide robust support for addressing complex and critical problems, such as the capability to generate synthetic data to balance inherently imbalanced datasets which are tricky to be attained in real experiments [73]. The ambit of applications spans not only quantum chemistry research and molecular reaction kinetics, but also extends to process optimization and control, which is imperative for enhancing efficiency and safety in chemical processes [13, 59, 64]. Consequently, there is cogent reason to postulate that, by employing the initial conditions of droplet microfluidic coalescence as input, machine learning algorithms possess the capability to furnish more precise predictions and search a more conducive set of experimental parameters, thereby contributing to the enhancement of experimental accuracy.

While the advent of ML offers significant advancements in the field of chemical engineering, its application to intricate processes like drop coalescence brings forth challenges tied to model transparency. Notably, complex models, such as deep neural networks, often referred as ”black boxes” due to their opaque nature. Although their predictive capabilities are often remarkable, they rarely provide clarity on their underlying decision-making mechanisms [5, 4]. In critical domains like drop coalescence, where a comprehensive understanding of the dynamics is vital for both research and industrial applications, this opacity poses significant negative implications. The ensuing gap in transparency and interpretability can inhibit the wider adoption of these models, affecting the trust they garner among researchers and the broader public [65].

Based on this, the demand for more interpretable ML models, which do not compromise on predictive efficiency, is noticeable across myriad engineering disciplines. In response to this overarching need, attention has been shifting towards the domain of Explainable Artificial Intelligence (XAI) [18, 57]. These explainable models not only strengthen trust in the predictive outcomes of ML models but also assist researchers in discerning the pivotal factors influencing complex engineering phenomena. In the realm of microfluidic applications, Specifically, in the realm of microfluidic applications, the relevance of XAI stands out sharply. In microfluidic experiments, intricate behaviors are often observed, such as the optimization of membraneless microfluidic fuel cells, the fusion dynamics of coalescing droplets, shear-induced phase transitions, and the nuanced mechanisms of droplet breakup at T-junctions. Given that each of these applications involves a wide array of parameters and features, gaining a comprehensive understanding of key influencing factors and learning how to adeptly adjust operational parameters become increasingly indispensable. [48, 6, 68, 2]. For example, by employing XAI, researchers can gain analytic insights into the model to confirm and explain why the geometry and wetting properties of microfluidic channels are pivotal factors in droplet generation [54]. This enables a more nuanced understanding of how subtle changes in channel roughness or surface wettability can have a significant impact on droplet formation rates. As research intensifies in microfluidic processes, the clarifications brought forth by explainable models are instrumental in guiding improved experimental designs, ensuring more dependable predictions, and thereby leading to

the development of more reliable microfluidic applications.

In this work, we employ a two-pronged strategy that integrates explainability into machine learning (ML) models to investigate the coalescence of aqueous droplets in oil within microfluidic devices. The first phase involves the design and construction of a suite of ML models, specifically Random Forest, XGBoost, and Multilayer Perceptrons (MLPs), each fine-tuned through hyperparameter optimization. These models are employed to predict the coalescence behavior of droplets under varying experimental conditions.The subsequent phase is dedicated to augmenting explainability by scrutinizing the model's predictive outcomes and associated features. The goal is to offer a clear understanding of which specific attributes or conditions, such as channel geometry or flow rates, most effectively contribute to the successful coalescence of aqueous drops in oil, which can optimize resource allocation in subsequent experiments and minimize superfluous operations. This is achieved through the deployment of widely-used post-hoc explainability methodologies, such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) [32, 43, 53]. Additionally, feature ablation testing is utilized to validate the influence and relevance of each featrue on the droplet coalescence phenomena.

Our contribution lies in the novel integration of explainability within machine learning (ML) models, specifically targeting the study of droplet coalescence in microfluidic systems in chemical engineering. By employing XAI techniques, this integration enhances the trustworthiness and practical utility of our machine learning models [49]. Hence, our work does more than just provide accurate predictive models; it also offers actionable insights that are poised to catalyze advancements in both academic research and industrial applications concerning droplet coalescence in microfluidic systems.

2 Experiments and dataset

In this section, we examine the distribution of the droplet coalescence dataset which is conducted within microfluidic devices. Additionally, we provide a clear overview of the data preprocessing techniques employed.

2.1 Dataset

This study investigates an experimental tabular dataset comprising 1501 samples, inclusive of five features and one label, y. The features are:

$$[x_{\frac{D}{W}1+\frac{D}{W}2}, x_{|\frac{D}{W}1-\frac{D}{W}2|}, x_{dt}, x_{\text{flow}}, x_{\text{Heff}}]$$

, representing different conditions of the microfluidic coalescence process conducted in mineral oil. The label y is classified into two categories: "Coalescence" and "Non_coalescence", representing whether droplet coalescence occurs or not.

Each of the five features in the dataset represents a specific aspect of the microfluidic coalescence process:

1. $x_{\frac{D}{W}1+\frac{D}{W}2}$ refers to the sum of two droplet diameters (D) normalised by channel width between two walls (W). This feature is showing the size of doublet as related to channel size.

2. $x_{|\frac{D}{W}1-\frac{D}{W}2|}$ is the absolute value of difference between two normalised diameters. This is an indicator of the disparity in sizes of droplets.

3. x_{dt} represent a temporal element in the experiment, i.e. the time interval between successive entrance of droplets into chamber.

4. x_{flow} refer to the total flow rate in the each of input into chamber.

5. x_{Heff} is the effective height of the channel explained in experimental section.

The experimental dataset in this study is a balanced dataset, comprising 782 instances of "Coalescence" and 719 instances of "Non-Coalescence". For robust evaluation of the ML models, we partition the original dataset into two subsets - a training dataset and a testing dataset which are stratified by approximately 1.09 and the details are showed in Table. 1.

	Coalescence	**Non-coalescence**	**Balance Ratio (BR)**	**Total**
Total dataset	782	719	1.09	1501
Training dataset	625	575	1.09	1200
Testing dataset	157	144	1.09	301

Table 1: Distribution of instances about dataset split

This partitioning is performed using a shuffle strategy, meticulously maintaining the original label distribution ratio, thus ensuring the same stratification. Furthermore, it is unnecessary to create a separate validation dataset because k-fold cross-validation is utilized in the training process. This strategy can assess the generalization ability of predictive models and prevent overfitting during training [20, 8].

The distribution of all features for Coalescence and Non-coalescence are displayed in Figure. 1. The initial five plots in Figure. 1 illustrate the distribution of the five features within the dataset, specifically categorized under the labels "Coalescence" and "Non-Coalescence." These distributions are evaluated using the KDE method, with the solid lines in each plot representing the estimated distribution trends accordingly.

Furthermore, the sixth plot demonstrates the lucid distribution ratio of instances for this binary classification task. It attests to a near-equal distribution of instances, with "Coalescence" constituting 52.1% and "Non-Coalescence" making up 47.9% of the data. This near parity prove it is a well-balanced dataset, which ensures a fair representation of both classes, thereby eschewing biases and facilitating an objective evaluation of the proposed machine learning models.

2.2 Data preprocessing

The raw tabular data, pertaining to the coalescence phenomena of aqueous droplets on an mineral oil, need preprocessing to ensure its amenability for the ensuing analytical phase. A key preprocessing step involves the execution of Min-Max normalisation, thereby rescaling data to a predefined range of $[0, 1]$. This prudent normalisation technique functions not merely to mitigate potential discrepancies in the scale across different features, but rather it can conscientiously preserves the relative relationships amongst individual sample points in the feature space [56]. The corresponding normalisation strategy adheres to the mathematical formulation as depicted in equation (1)

$$x_{\text{scaled}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}, \quad x \in [x_{\frac{D}{W}1+\frac{D}{W}2}, x_{|\frac{D}{W}1-\frac{D}{W}2|}, x_{dt}, x_{\text{flow}}, x_{\text{Heff}}] \tag{1}$$

where x_{max} and x_{min} are the maximum and minimum value of the feature x. x_{scaled} is the scaled results after normalizing.

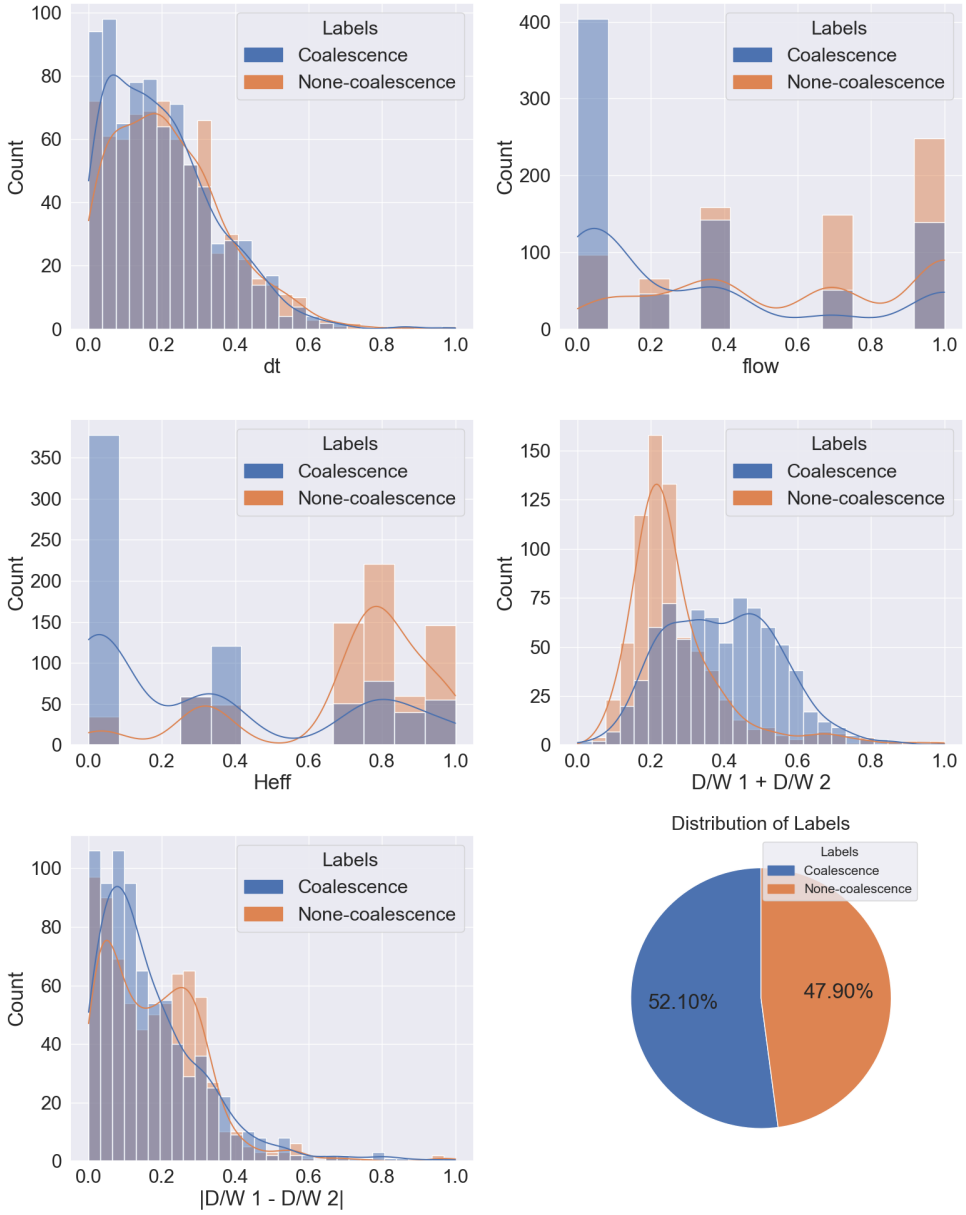


Figure 1: Feature Distribution Comparison between Coalescence and Non-coalescence

3 Methodology

In this section, the construction and evaluation of machine learning models involved in the first phase, as well as methods falling under the domain of XAI, are elaborated in details. The contents specifically encompass two kinds of tree-based models, a type of Deep Neural Networks (DNNs) known as MLPs, hyperparameter space search methods, along with SHAP and LIME explanations. The details of predictive models, hyperparameter space search methods and performance metrics are shown in Appendix. A.

3.1 SHapley Additive exPlanations(SHAP)

SHAP values serve to interpret the influence of features on a specific prediction by computing the average marginal contribution of each feature across all conceivable permutations. This method is based on cooperative game theory and is utilized for interpreting the outcomes of machine learning model [36, 41]. In the context of SHAP, features are treated as "players" that "contribute" to the prediction. The accumulated contribution of all features constitutes the ultimate prediction results of the model [22]. The subsequent equation (2) formally defines this process.

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x_i^{'} \tag{2}$$

where $f(x)$ refers to the outcome of the original predictive model, defined as a function of the input vector x, which can be a high-dimensional feature vector. $g(x')$, on the other hand, signifies an interpretable surrogate model that approximates $f(x)$ but is expressed in terms of x', a simplified or lower-dimensional version of x. The transformation between x and x' is often achieved through feature selection or dimensionality reduction techniques, making g more straightforward to interpret than f. ϕ_0 acts as a base value from which the contributions of individual features are added or subtracted and ϕ_i is the SHAP value for the i-th feature. For any given sample, the feature possessing the larger absolute Shapley value wields a more substantial influence on the prediction result for that sample. The magnitude and sign of these Shapley values give insight into how significantly and in what direction each feature influences a given prediction [58]. Specifically, a positive Shapley value suggests that the corresponding feature contributes towards an increase in the model's predicted value. Conversely, a negative Shapley value implies that the feature has a reducing effect on the model's prediction.

Thus, the Shapley value not only measures the magnitude of a feature's impact on the model's prediction but also denotes the nature of this impact—whether it is beneficial (positive) or detrimental (negative) [37, 38]. The Shapley value is calculated as equation (3):

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{3}$$

where:

- ϕ_i is the Shapley value, representing the contribution of feature i to the prediction.
- N is the total set of features, which corresponds to all the elements in the feature vector x.
- S is a subset of N that includes selected features from the original feature vector x. The subset S does not contain the elements of feature represented by i.

- $f(S)$ is the predictive function with features in S.
- The term $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ is the weighting factor, representing the number of permutations that include feature i.

In terms of global interpretability, SHAP provides an aggregated view across all samples, allowing researchers to discern the overall importance of each feature in the model. By analysing the distribution of SHAP values for a particular feature, people can visualize not only the magnitude of its importance but also the direction of its effect on model predictions. Features with higher absolute SHAP values are typically more influential, and their consistent positive or negative values indicate a systematic increase or decrease in the model's prediction, respectively.Consequently, SHAP's global interpretability improves the identification of potential feature interactions and non-linear dependencies [72]. The details of black-box model's and explainable model's interactions are shown in Figure. 2.

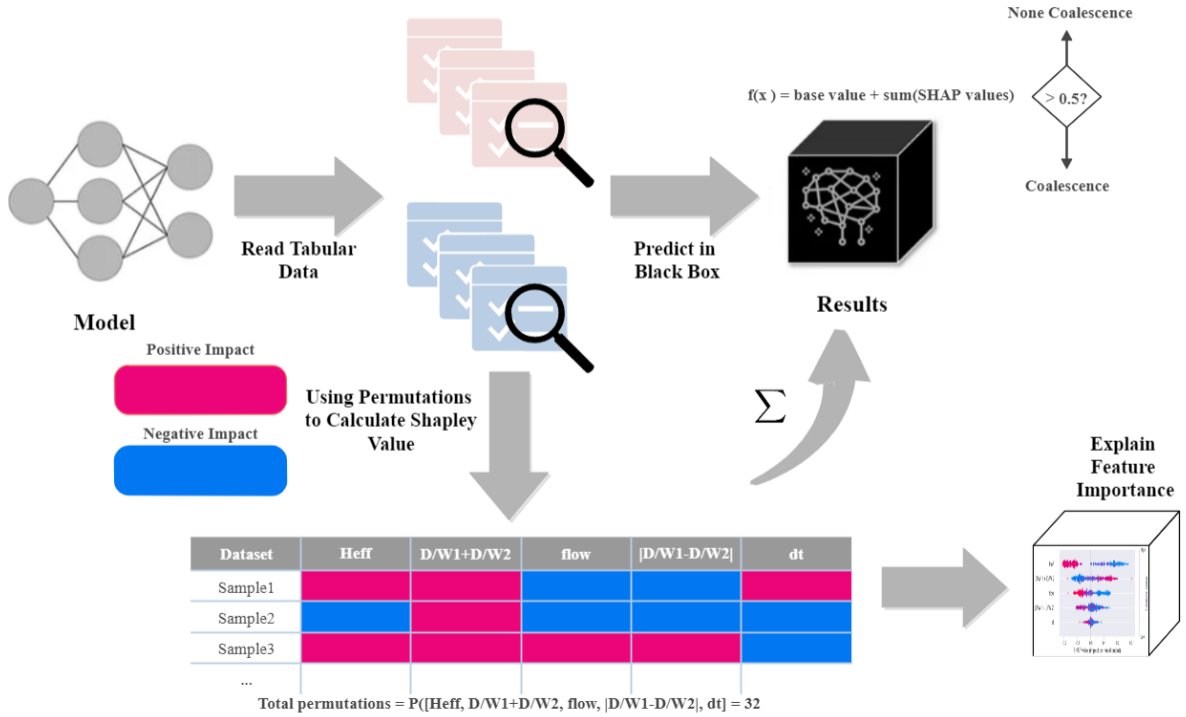


Figure 2: Visual Representation of SHapley Additive exPlanations (SHAP) in Chemical Applications

3.2 Local Interpretable Model-agnostic Explanations(LIME)

LIME is another explanatory method designed to clarify the predictions provided by any classifier or regressor in an understandable and faithful way [52]. It achieves this objective by approximating the model with a local surrogate model that is inherently interpretable, thus improving comprehension of the model's decisions in the vicinity of the instance under consideration [51]. It's worth noting that although both SHAP and LIME are designed to enhance the interpretability of machine learning models, they employ distinct approaches to achieve this objective. SHAP provides a global interpretability method with a theoretical foundation grounded in cooperative game theory. In contrast, LIME focuses on local interpretability by approximating the model's behavior near each individual prediction, typically by constructing a local linear surrogate model to explain each prediction. The mathematical formulation of LIME is as equation (4) [52]:

$$\xi(x) = \arg\min_{z \in Z} L(f, z, \pi_x) + \Omega(z) \tag{4}$$

where:

- f denotes the original, potentially non-linear, prediction model under scrutiny.
- x is the instance of the dataset for which the explanation is being computed.
- Z generally refers to the family of models that are considered "interpretable" and can act as local surrogate models to approximate the behavior of the more complex model f. In the specific context of this study, which focuses on drop coalescence classification tasks, 'LimeTabularExplainer' is utilized. Consequently, Z is restricted to linear models. Specifically, each $z \in Z$ is a logistic regression model that is trained to provide a faithful local approximation of f's decision-making process in a localized region surrounding the data instance x.
- π_x is the proximity measure between the instance x and the data instances used to learn the explanation model. In this implementation, the measure metric is the Euclidean distance.
- $\xi(x)$ represents the explanation model for the instance x. Essentially, $\xi(x)$ is the optimized local linear surrogate model z that best approximates the original model f within a predefined local neighborhood around x, according to the minimization of the loss function $L(f, z, \pi_x)$ and the complexity term $\Omega(z)$.
- $L(f, z, \pi_x)$ is a measure of how unfaithfully z approximates f in the vicinity of instance x, defined by π_x.
- $\Omega(z)$ is a measure of complexity of the explanation model z, which aims to keep the explanation as simple as possible.

LIME learns the explanation model z by minimizing the loss function $L(f, z, \pi_x)$ and the complexity measure $\Omega(z)$, effectively ensuring that z is locally faithful to f and is interpretable [52]. The Figure. 3 intuitively illustrates schematic representation of LIME's methodology, i.e. exploiting local linearity within a complex, globally non-linear dataset. Despite the global distribution of data points depicts a non-linear relationship, through zooming in on a specific subset or a selected local region of the dataset, a simplified linear relationship can be observed. The transition from non-linear complexity to linear simplicity underlines the versatility of LIME in deciphering the decision boundaries set by complex models.

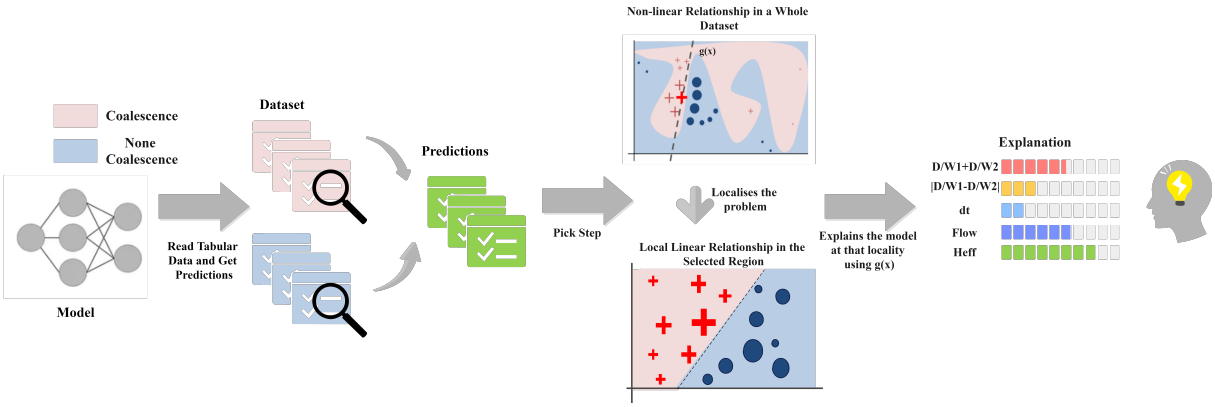


Figure 3: Visual Representation of Local Interpretable Model-agnostic Explanations (LIME) in Chemical Applications

By focusing on the local view, rather than the global view, LIME can generate reasonable explanations that align with the simpler linear relationship, which are more comprehensible to humans and can keep local fidelity.

4 Results and Analysis

In this section, we instantiate the three models mentioned in section. 3 and train them on the preprocessed dataset to calculate the accuracy scores for both validation and testing. Subsequently, we apply XAI techniques to conduct ablation tests which shown in Appendix. B.1, aiming to comparatively analyze which features exert a substantial impact on the microfluidic coalescence of aqueous droplets in oil. The overall observations and discussions are shown in Appendix. B.2.

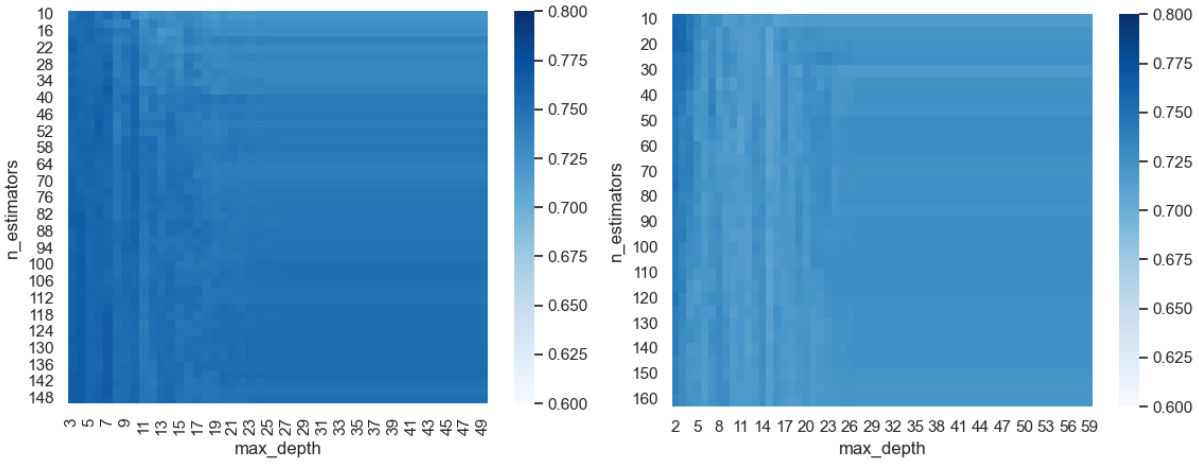
4.1 Implementation details

In the modelling phase, the raw microfluidic droplet coalescence dataset is preprocessed following the pipelines delineated in Section. 2.2. Due to this preprocessing, the dataset is divided into two balanced subsets with same balance ratio: a training set and a test set. Three distinct machine learning algorithms, Random Forest, XGBoost, and a Multi-Layer Perceptron (MLP) are then optimised through Grid Search method. This method is strategically employed to identify an optimal set of hyperparameters in predefined hyperparameter space, ensuring the most favourable performance on the extant dataset. Throughout the whole training process, a five-fold cross validation technique is consistently applied. Both the F1-Score and accuracy are adopted as the principal performance metrics, instrumental in refitting the model. After a series of calculated iterations adjusting the hyperparameters, the optimal combination that yields the peak validation accuracy is selected for the final model's architecture.

The key hyperparameters of random forest and XGBoost are **n_estimators** and **max_depth**. In this study, we investigate the n_estimators parameter in the range of 10-151 and the max_depth parameter in the range of 3-50 for the random forest model. For the XGBoost model, due to its mechanism of calculating the next layer through weights, we set the range for n_estimators as 10-160 and for max_depth as 2-60. Finally, we set the optimal parameters for both the random forest and XGBoost models are [7, 145] and [2, 15] respectively. As shown in Figure. 4, the specific details of hyperparamter tuning are displayed.

In the presented visualisations, the deep blue regions signify areas where the model achieves higher validation accuracy. Comparing the visualization heatmaps of the two models, it is evident that the Random Forest model exhibits a more extensive deep blue region. This prominence of deeper shades in the Random Forest heatmap underscores its superior adaptability on the droplet coalescence dataset in comparison to XGBoost. The broader coverage of this high-accuracy zone suggests that Random Forest might be inherently more suited for the intricacies and nuances of this particular dataset because its bagging strategy can result in the model more resilient to overfitting than using boosting strategy.

In our quest to optimise the MLPs, a deep learning model by its nature, we recognise the necessity of fine-tuning an extensive array of hyperparameters. This requirement arises due to the inherent complexity of the MLPs model, as compared to more traditional machine learning counterparts. Consequently, we explore a diverse set of hyperparameters to achieve optimal performance in Table. 4 which shown in Appendix. C. The hyperparameters that are ultimately selected for the final implementation are denoted in bold within Table. 4.



(a) Random Forest

(b) XGBoost

Figure 4: Validation heatmaps for the tuning hyperparameters of tree-based predictive models.

4.2 Predictive Results

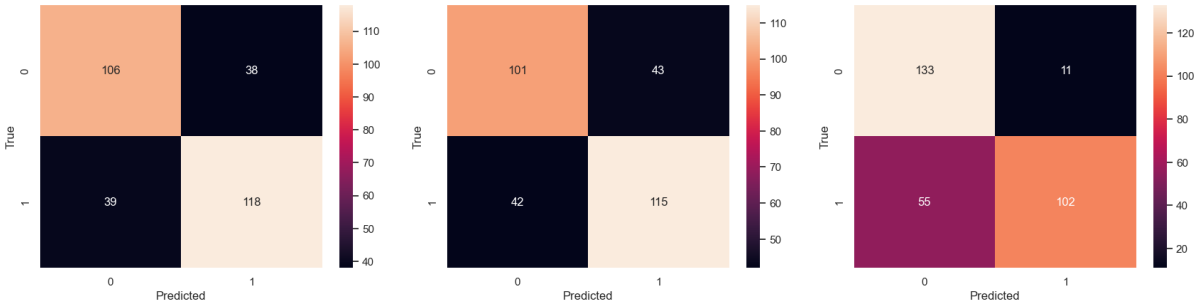
To assess the predictive performance of the considered machine learning models on coalescence events after training with 5-fold cross-validation, we present the predictive results and their confusion matrices on testing dataset. The predictive results for all three models are shown in Table. 2.

Table 2: Model Performance Metrics for Hyperparameter Tuning

Model	Precision(%)	Recall(%)	F1-Score(%)	Accuracy(%)
Random Forest	75.64	**75.16**	75.40	74.42
XGBoost	72.78	73.25	73.02	71.76
MLP	**90.27**	64.97	**75.56**	**78.07**

From the highlighted figures in Table. 2, it's evident that the Multilayer Perceptron (MLP) consistently outperforms the tree-based models in metrics like precision, accuracy, and F1-score on the testing dataset. This superior performance underscores MLP's effectiveness for the droplet coalescence dataset. It indicates that, for this specific dataset, the MLP is more adept at generalising its learnings from the training data to unseen samples. The intricacies of the data might be captured better by the MLP model structure than by the tree-based counterparts. Within the context of this study, 'coalescence' is designated as the positive class (1) whilst 'non-coalescence' serves as the negative class (0). Accordingly, Figure 5 presents the respective metrics from the confusion matrix, offering an intuitive insight into the predictive capabilities of the models.

Upon evaluating the classification between coalescence and non-coalescence events using three models, distinct performance metrics are observed. The Random Forest model displayed in Figure. 5a yields a recall of 75% and a precision of 76%, reflecting its capability to identify coalescence events with a balanced accuracy, as further evidenced by its F1 score. The XGBoost model exhibited in Figure. 5b exhibits a recall of 73% with a precision of 73%, indicating a consistent balance in its predictions. In contrast, in Figure. 5c, the MLP model displays a recall of 66% but achieves a notably high precision of 90%. This indicates that while it may not capture all coalescence events, its predictions are predominantly accurate when classifying an event as coalescence (Highest precision among three models).

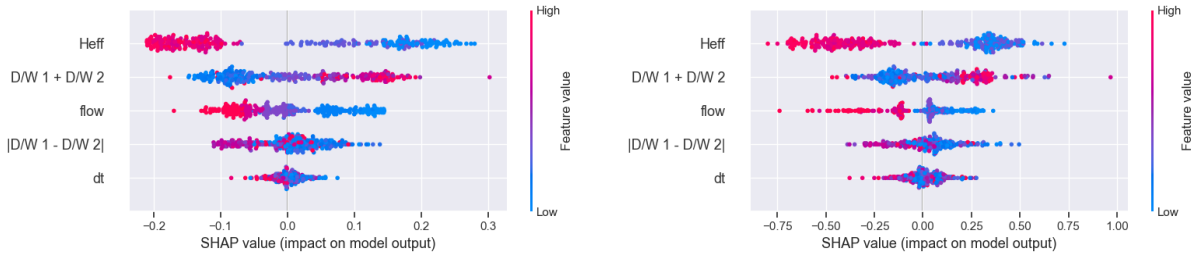


(a) Random Forest (b) XGBoost (c) MLP

Figure 5: Confusion Metrics of Three Predictive Models.

4.3 Global Intrepretability

The SHAP summary plots in Figure. 6 provide global interpretability for both the Random Forest and XGBoost models, revealing the influence of features on predictions. These images show the list of important features ranked from most significant to least significant (top to bottom).The abscissa, denoted as the X-axis, represents the impact on the label 1 (coalescence), where positive SHAP values connote a beneficial impact, whereas negative values signify an detrimental effect. The color bar indicates the quantity level of the original feature value, with red indicating high feature values, blue dots marking low feature values, and purple dots representing medium feature values.



(a) Random Forest (b) XGBoost

Figure 6: SHAP Plot for Tree-based Models

In both plots, the feature importance rank are analysed as 'Heff', '$\frac{D}{W}1 + \frac{D}{W}2$', 'flow', '$|\frac{D}{W}1 - \frac{D}{W}2|$', and 'dt'. Furthermore, higher values of '$\frac{D}{W}1 + \frac{D}{W}2$' correspond to positive SHAP values for both models, indicating that as this ratio increases, the model's prediction is more likely to lean towards the positive class (Coalescence). Conversely, for the remaining features, lower values generally lead to positive SHAP values, suggesting an inverse relationship with positive predictions.

A pellucid observation is the clustering of SHAP values for '$|\frac{D}{W}1 - \frac{D}{W}2|$' and 'dt' around original point (X-axis = 0). This indicates that these features might have a neutral or varied impact on the predictions, reflecting their potential limited discriminative power in these models.

The consistency in patterns across both the Random Forest and XGBoost models can be attributed to their shared tree-based structure. Both models employ a methodology of recursively splitting data based on feature thresholds, possibly leading to similar patterns in data. The ensemble nature of both models, which derive predictions from multiple decision trees, could also contribute to this similarity.

When compared with tree-based models, the SHAP analysis for the MLP in Figure. 7 demonstrates both similarities and distinctions. A primary similarity among them is the unanimous recognition of 'Heff' as the most influential feature. Yet, a notable departure surfaces in the ranking of other features. Specifically, in the MLP's assessment, the 'flow' feature ascends to the second position, nudging '$\frac{D}{W}1 + \frac{D}{W}2$', which held this rank in the tree-based models, to a different position. This reshuffling indicates variations in the way the MLP and tree-based models weigh the importance of these features.

Figure 7: SHAP Plot for MLP

Moreover, the overall trend in the features' influence remains consistent. Both kinds of ML models indicate that a higher value of '$\frac{D}{W}1 + \frac{D}{W}2$' is associated with a positive impact on the output. Conversely, an increase in other feature values predominantly leads to a negative influence. Such consistent trends across different model types underscore the robust patterns intrinsic to the data. Nonetheless, the nuanced disparities also demonstrate unique internal workings and sensitivities inherent to different model architectures. Simultaneously, MLP delineates the influence of SHAP values more distinctly compared to the other two models. This observation indicate that the MLPs model is more proficient in segregating and distinguishing the contributions of individual features.

4.4 Local Intrepretability

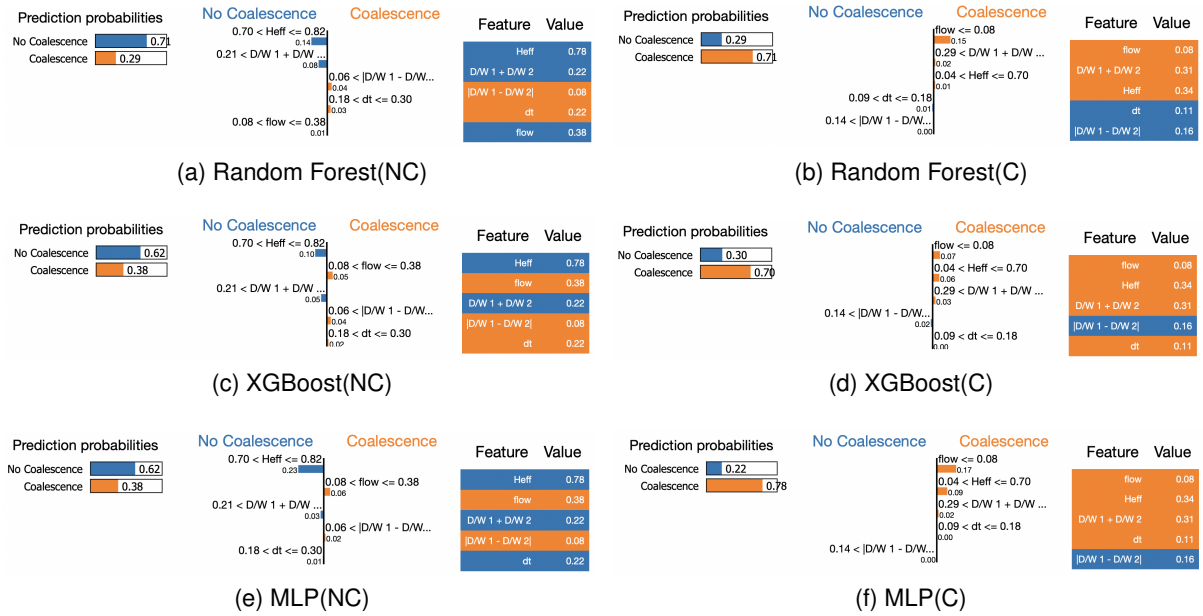
To gain a more nuanced understanding of our models' decision-making processes for specific instances, we employ LIME, an approach which is famous for its ability to provide local model-agnostic explanations. LIME functions by generating a perturbed dataset around a chosen instance and learning a locally interpretable model on this new dataset. The reason is that although our machine learning models (such as MLPs) might be complex globally in this drop coalescence dataset, they might behave linearly in the vicinity of specific instances, thus allowing for clear interpretability and easy understanding.

Upon reviewing the LIME analyses for the two instances labelled 'Non-Coalescence' and 'Coalescence', several observations can be discussed in Figure. 8. For the 'Non-Coalescence' instance, all three models - Random Forest, XGBoost, and MLP - produce predictions that align well with the true label, yielding probabilities of 0.71, 0.62, and 0.62 respectively. Although there are slight differences in how each model arrives at its prediction, there is a consistent emphasis on the features 'Heff', '$\frac{D}{W}1 + \frac{D}{W}2$', and 'flow' across all models.

In drawing parallels with previous SHAP analyses, it's noteworthy that the features 'Heff', '$\frac{D}{W}1 + \frac{D}{W}2$', and 'flow' are similarly emphasised as significant contributors to prediction outcomes. This recurrent emphasis across different analytic methods substantiates the robustness of these features in the decision-making process. Moreover, the feature ablation tests conducted earlier reveal that when these features are individually removed from the model, a marked degradation

in prediction accuracy is observed. This aligns seamlessly with the feature importance indicated by both LIME and SHAP analyses.

The consistency in feature importance suggests a stable and strong relationship between these features and the predicted outcomes, and this matches the findings from the previous SHAP analysis. For the 'Coalescence' instance, the models again provide predictions that closely match the true label, with probabilities of 0.71, 0.70, and 0.78. Again, the features 'Heff', '$\frac{D}{W}1 + \frac{D}{W}2$', and 'flow' are highlighted as key determinants in the decision-making process across all models.



(a) Random Forest(NC) (b) Random Forest(C)

(c) XGBoost(NC) (d) XGBoost(C)

(e) MLP(NC) (f) MLP(C)

Figure 8: LIME Plot for Two instances. NC means the actual label for instance is Non-Coalescence, and C means the actual label for instance is Coalescence

Another point of note is the similarity in predicted probabilities across the three models. This consistent performance across different predictive model architectures indicates that, despite potential differences in how they process the dataset, their conclusions are relatively uniform. It's worth recalling that the SHAP and feature ablation tests produced similar overarching themes, strengthening this claim. Such alignment across models suggests that the findings are not an outcome of a particular model but are likely reflective of the underlying data patterns.

5 Conclusion

In this study, an in-depth analysis of predictive models and their inherent relationships with critical features is undertaken. Three predominant ML models — Random Forest, XGBoost, and Multi-Layer Perceptron (MLP) — are utilised to elucidate patterns in the data relating to 'Coalescence' and 'Non-Coalescence'. Feature importance is initially evaluated through SHAP values, identifying 'Heff,' '$\frac{D}{W}1+\frac{D}{W}2$', and 'flow' as pivotal determinants across the models. Subsequently, feature ablation testing elucidates the sensitivity and robustness of each model when these critical features are absent. Complementing this, local interpretability through LIME not only corroborates the overarching importance of the identified features but also offers specific insights into each model's inferential logic, thereby reinforcing the global interpretability insights obtained from SHAP. Collectively, this multifaceted approach integrates global and local interpretability

with feature ablation, enhancing a profound understanding of the decision-making mechanics within the chosen models and amplified the trustworthiness of their predictions.

Consequently, this clear examination establishes a robust template for future endeavors in predictive analytics applicable to chemical engineering disciplines. It thereby serves as an precedent, underscoring the imperative for an analytical strategy that transcends rudimentary predictive outcomes to scrutinize the intricate mechanics of decision-making and feature contributions. In delineating this approach, the study facilitates the development of increasingly transparent, accountable, and verifiable machine learning applications for complex engineering systems. This accentuates the indispensable role of model interpretability in executing data-driven decisions with significant real-world implications.

References

[1] Ola Aarøen, Enrico Riccardi, and Marit Sletmoen. Exploring the effects of approach velocity on depletion force and coalescence in oil-in-water emulsions. *RSC advances*, 11(15):8730–8740, 2021.

[2] Shahriar Afkhami, AM Leshansky, and Y Renardy. Numerical investigation of elongated drops in a microfluidic t-junction. *Physics of Fluids*, 23(2), 2011.

[3] Ilke Akartuna, Donald M Aubrecht, Thomas E Kodger, and David A Weitz. Chemically induced coalescence in droplet-based microfluidics. *Lab on a Chip*, 15(4):1140–1144, 2015.

[4] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, 11(11):5088, 2021.

[5] Christina B Azodi, Jiliang Tang, and Shin-Han Shiu. Opening the black box: interpretable machine learning for geneticists. *Trends in genetics*, 36(6):442–455, 2020.

[6] Charles N Baroud, Francois Gallaire, and Rémi Dangla. Dynamics of microfluidic droplets. *Lab on a Chip*, 10(16):2032–2045, 2010.

[7] Bijoy Bera, Rama Khazal, and Karin Schroën. Coalescence dynamics in oil-in-water emulsions at elevated temperatures. *Scientific reports*, 11(1):10990, 2021.

[8] Daniel Berrar et al. Cross-validation., 2019.

[9] Claire Berton-Carabin and Karin Schroën. Towards new food emulsions: Designing the interface and beyond. *Current Opinion in Food Science*, 27:74–81, 2019.

[10] Rishabh Choudhary and Hemant Kumar Gianey. Comprehensive review on supervised machine learning algorithms. In *2017 International Conference on Machine Learning and Data Science (MLDS)*, pages 37–43. IEEE, 2017.

[11] Pasquale Dell'Aversana, Jayanth R Banavar, and Joel Koplik. Suppression of coalescence by shear and temperature gradients. *Physics of Fluids*, 8(1):15–28, 1996.

[12] Renze Dong, Hongze Leng, Juan Zhao, Junqiang Song, and Shutian Liang. A framework for four-dimensional variational data assimilation based on machine learning. *Entropy*, 24(2):264, 2022.

[13] Pavlo O Dral. Quantum chemistry in the age of machine learning. *The journal of physical chemistry letters*, 11(6):2336–2347, 2020.

[14] Marcin Dudek, Diana Fernandes, Eirik Helno Herø, and Gisle Øye. Microfluidic method for determining drop-drop coalescence and contact times in flow. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 586:124265, 2020.

[15] Marcin Dudek, Kelly Muijlwijk, Karin Schroën, and Gisle Øye. The effect of dissolved gas on coalescence of oil drops studied with microfluidics. *Journal of colloid and interface science*, 528:166–173, 2018.

[16] Dhivya Elavarasan and Durai Raj Vincent. Reinforced xgboost machine learning model for sustainable intelligent agrarian applications. *Journal of Intelligent & Fuzzy Systems*, 39(5):7605–7620, 2020.

[17] John S Eow and Mojtaba Ghadiri. Electrostatic enhancement of coalescence of water droplets in oil: a review of the technology. *Chemical Engineering Journal*, 85(2-3):357–368, 2002.

[18] Bahareh Esteki, Mahmood Masoomi, Mohammad Moosazadeh, and ChangKyoo Yoo. Data-driven prediction of janus/core–shell morphology in polymer particles: A machine-learning approach. *Langmuir*, 39(14):4943–4958, 2023.

[19] Shen Feng, LI Yi, Liu Zhao-Miao, CAO Ren-Tuo, and Wang Gui-Ren. Advances in micro-droplets coalescence using microfluidics. *Chinese Journal of Analytical Chemistry*, 43(12):1942–1954, 2015.

[20] Michael N Fienen and Nathaniel G Plant. A cross-validation package driving netica with python. *Environmental Modelling & Software*, 63:14–23, 2015.

[21] Edgar A Galan, Haoran Zhao, Xukang Wang, Qionghai Dai, Wilhelm TS Huck, and Shaohua Ma. Intelligent microfluidics: The convergence of machine learning and microfluidics in materials science and biomedicine. *Matter*, 3(6):1893–1922, 2020.

[22] María Vega García and José L Aznarte. Shapley additive explanations for no2 forecasting. *Ecological Informatics*, 56:101039, 2020.

[23] Unnati Garg, Swati Chauhan, Upendra Nagaich, and Neha Jain. Current advances in chitosan nanoparticles based drug delivery and targeting. *Advanced pharmaceutical bulletin*, 9(2):195, 2019.

[24] Robin Genuer, Jean-Michel Poggi, Christine Tuleau-Malot, and Nathalie Villa-Vialaneix. Random forests for big data. *Big Data Research*, 9:28–46, 2017.

[25] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data?, 2022.

[26] Thao Minh Ho, Aysan Razzaghi, Arun Ramachandran, and Kirsi S Mikkonen. Emulsion characterization via microfluidic devices: A review on interfacial tension and stability to coalescence. *Advances in Colloid and Interface Science*, 299:102541, 2022.

[27] Zurki Ibrahim, Pinar Tulay, and Jazuli Abdullahi. Multi-region machine learning-based novel ensemble approaches for predicting covid-19 pandemic in africa. *Environmental Science and Pollution Research*, 30(2):3621–3643, 2023.

[28] Bibin M Jose and Thomas Cubaud. Droplet arrangement and coalescence in diverging/-converging microchannels. *Microfluidics and nanofluidics*, 12:687–696, 2012.

[29] Nina M Kovalchuk, Marten Reichow, Thomas Frommweiler, Daniele Vigolo, and Mark JH Simmons. Mass transfer accompanying coalescence of surfactant-laden and surfactant-free drop in a microfluidic channel. *Langmuir*, 35(28):9184–9193, 2019.

[30] Nina M Kovalchuk and Mark JH Simmons. Review of the role of surfactant dynamics in drop microfluidics. *Advances in Colloid and Interface Science*, page 102844, 2023.

[31] NM Kovalchuk, J Chowdhury, Zoe Schofield, Daniele Vigolo, and MJH Simmons. Study of drop coalescence and mixing in microchannel using ghost particle velocimetry. *Chemical Engineering Research and Design*, 132:881–889, 2018.

[32] Soonki Kwon and Younghoon Lee. Explainability-based mix-up approach for text data augmentation. *ACM Transactions on Knowledge Discovery from Data*, 17(1):1–14, 2023.

[33] Thomas Leary, Mohsen Yeganeh, and Charles Maldarelli. Microfluidic study of the electrocoalescence of aqueous droplets in crude oil. *ACS omega*, 5(13):7348–7360, 2020.

[34] Tao Li, Junjun Wang, Fenglong Wang, Lishu Zhang, Yanyan Jiang, Hamidreza Arandiyan, and Hui Li. The effect of surface wettability and coalescence dynamics in catalytic performance and catalyst preparation: a review. *ChemCatChem*, 11(6):1576–1586, 2019.

[35] Petro Liashchynskyi and Pavlo Liashchynskyi. Grid search, random search, genetic algorithm: a big comparison for nas. *arXiv preprint arXiv:1912.06059*, 2019.

[36] Kang Lin and Yuzhuo Gao. Model interpretability of financial fraud detection by group shap. *Expert Systems with Applications*, 210:118354, 2022.

[37] Jing-Jing Liu and Jian-Chao Liu. Permeability predictions for tight sandstone reservoir using explainable machine learning and particle swarm optimization. *Geofluids*, 2022:1–15, 2022.

[38] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[39] Pengcheng Ma, Di Liang, Chunying Zhu, Taotao Fu, and Youguang Ma. An effective method to facile coalescence of microdroplet in the symmetrical t-junction with expanded convergence. *Chemical Engineering Science*, 213:115389, 2020.

[40] Polina Mamoshina, Armando Vieira, Evgeny Putin, and Alex Zhavoronkov. Applications of deep learning in biomedicine. *Molecular pharmaceutics*, 13(5):1445–1454, 2016.

[41] Michael L Martini, Sean N Neifert, Eric K Oermann, Jeffrey T Gilligan, Robert J Rothrock, Frank J Yuk, Jonathan S Gal, Dominic A Nistal, and John M Caridi. Application of cooperative game theory principles to interpret machine learning models of nonhome discharge following spine surgery. *Spine*, 46(12):803–812, 2021.

[42] Linas Mazutis and Andrew D Griffiths. Selective droplet coalescence using microfluidic systems. *Lab on a Chip*, 12(10):1800–1806, 2012.

[43] Branka Hadji Misheva, Joerg Osterrieder, Ali Hirsa, Onkar Kulkarni, and Stephen Fung Lin. Explainable ai in credit risk management. *arXiv preprint arXiv:2103.00949*, 2021.

[44] Kelly Muijlwijk, Ivanna Colijn, Herditya Harsono, Thomas Krebs, Claire Berton-Carabin, and Karin Schroën. Coalescence of protein-stabilised emulsions studied with microfluidics. *Food Hydrocolloids*, 70:96–104, 2017.

[45] Konstantia Nathanael, Sibo Cheng, Nina M Kovalchuk, Rossella Arcucci, and Mark JH Simmons. Optimization of microfluidic synthesis of silver nanoparticles: a generic approach using machine learning. *Chemical Engineering Research and Design*, 193:65–74, 2023.

[46] Randa Natras, Benedikt Soja, and Michael Schmidt. Ensemble machine learning of random forest, adaboost and xgboost for vertical total electron content forecasting. *Remote Sensing*, 14(15):3547, 2022.

[47] Eugénie D Ngouémazong, Stefanie Christiaens, Avi Shpigelman, Ann Van Loey, and Marc Hendrickx. The emulsifying and emulsion-stabilizing properties of pectin: A review. *Comprehensive Reviews in Food Science and Food Safety*, 14(6):705–718, 2015.

[48] Dang Dinh Nguyen, Muhammad Tanveer, Hang-Nga Mai, Thinh Quy Duc Pham, Haroon Khan, Cheol Woo Park, and Gyu Man Kim. Guiding the optimization of membraneless microfluidic fuel cells via explainable artificial intelligence: Comparative analyses of multiple machine learning models and investigation of key operating parameters. *Fuel*, 349:128742, 2023.

[49] Cosmas Ifeanyi Nwakanma, Love Allen Chijioke Ahakonye, Judith Nkechinyere Njoku, Jacinta Chioma Odirichukwu, Stanley Adiele Okolie, Chinebuli Uzondu, Christiana Chidimma Ndubuisi Nweke, and Dong-Seong Kim. Explainable artificial intelligence (xai) for intrusion detection and mitigation in intelligent connected vehicles: A review. *Applied Sciences*, 13(3):1252, 2023.

[50] Sadegh Poozesh and Ecevit Bilgili. Scale-up of pharmaceutical spray drying using scale-up rules: A review. *International Journal of Pharmaceutics*, 562:271–292, 2019.

[51] Juan A Recio-García, Belén Díaz-Agudo, and Victor Pino-Castilla. Cbr-lime: a case-based reasoning approach to provide specific local interpretable model-agnostic explanations. In *Case-Based Reasoning Research and Development: 28th International Conference, ICCBR 2020, Salamanca, Spain, June 8–12, 2020, Proceedings 28*, pages 179–194. Springer, 2020.

[52] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[53] Ahmed Salih, Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E Petersen, Gloria Menegaz, and Karim Lekadir. Commentary on explainable artificial intelligence methods: Shap and lime. *arXiv preprint arXiv:2305.02012*, 2023.

[54] Ralf Seemann, Martin Brinkmann, Thomas Pfohl, and Stephan Herminghaus. Droplet based microfluidics. *Reports on progress in physics*, 75(1):016601, 2011.

[55] Mohsen Shahhosseini, Rafael A Martinez-Feria, Guiping Hu, and Sotirios V Archontoulis. Maize yield and nitrate loss prediction with machine learning algorithms. *Environmental Research Letters*, 14(12):124026, 2019.

[56] Dalwinder Singh and Birmohan Singh. Feature wise normalization: An effective way of normalizing data. *Pattern Recognition*, 122:108307, 2022.

[57] Abhishek Sivaram and Venkat Venkatasubramanian. Xai-meg: Combining symbolic ai and machine learning to generate first-principles models and causal explanations. *AIChE Journal*, 68(6):e17687, 2022.

[58] Roland Stirnberg, Jan Cermak, Simone Kotthaus, Martial Haeffelin, Hendrik Andersen, Julia Fuchs, Miae Kim, Jean-Eudes Petit, and Olivier Favez. Meteorology-driven variability of air pollution (pm 1) revealed with explainable machine learning. *Atmospheric Chemistry and Physics*, 21(5):3919–3948, 2021.

[59] Sina Stocker, Gábor Csányi, Karsten Reuter, and Johannes T Margraf. Machine learning in chemical reaction space. *Nature communications*, 11(1):5505, 2020.

[60] Xiaofei Sun, Yanyu Zhang, Guangpeng Chen, and Zhiyong Gai. Application of nanoparticles in enhanced oil recovery: a critical review of recent progress. *Energies*, 10(3):345, 2017.

[61] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S. Emer. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12):2295–2329, 2017.

[62] Priyanka Talukdar and Arindam Dey. Hydraulic failures of earthen dams and embankments. *Innovative Infrastructure Solutions*, 4:1–20, 2019.

[63] Eujin Um, Dae-Sik Lee, Hyeon-Bong Pyo, and Je-Kyun Park. Continuous generation of hydrogel beads and encapsulation of biological materials using a microfluidic droplet-merging channel. *Microfluidics and Nanofluidics*, 5:541–549, 2008.

[64] Venkat Venkatasubramanian. The promise of artificial intelligence in chemical engineering: Is it here, finally? *AIChE Journal*, 65(2):466–478, 2019.

[65] ES Vorm and David JY Combs. Integrating transparency, trust, and acceptance: The intelligent systems technology acceptance model (istam). *International Journal of Human–Computer Interaction*, 38(18-20):1828–1845, 2022.

[66] Jing-Tao Wang, Juan Wang, and Jun-Jie Han. Fabrication of advanced particles and particle-based materials assisted by droplet-based microfluidics. *small*, 7(13):1728–1754, 2011.

[67] Zhijin Wang, Xiufeng Liu, Yaohui Huang, Peisong Zhang, and Yonggang Fu. A multivariate time series graph neural network for district heat load forecasting. *Energy*, page 127911, 2023.

[68] Yining Wu, Taotao Fu, Chunying Zhu, Xiaoda Wang, Youguang Ma, and Huai Z Li. Shear-induced tail breakup of droplets (bubbles) flowing in a straight microfluidic channel. *Chemical Engineering Science*, 135:61–66, 2015.

[69] Min Xue, Huaming Wu, and Ruidong Li. Dnn migration in iots: Emerging technologies, current challenges and open research directions. *IEEE Consumer Electronics Magazine*, 2022.

[70] Haozhe Yi, Taotao Fu, Chunying Zhu, and Youguang Ma. Local deformation and coalescence between two equal-sized droplets in a cross-focused microchannel. *Chemical Engineering Journal*, 430:133087, 2022.

[71] Bo Zheng and Rustem F Ismagilov. A microfluidic approach for screening submicroliter volumes against multiple reagents by using preformed arrays of nanoliter plugs in a three-phase liquid/liquid/gas flow. *Angewandte Chemie International Edition*, 44(17):2520–2523, 2005.

[72] Xinzhi Zhou, Haijia Wen, Ziwei Li, Hui Zhang, and Wengang Zhang. An interpretable model for the susceptibility of rainfall-induced shallow landslides based on shap and xgboost. *Geocarto International*, 37(26):13419–13450, 2022.

[73] Kewei Zhu, Sibo Cheng, Nina Kovalchuk, Mark Simmons, Yi-Ke Guo, Omar K Matar, and Rossella Arcucci. Analyzing drop coalescence in microfluidic devices with a deep learning generative model. *Physical Chemistry Chemical Physics*, 2023.

[74] Yilin Zhuang, Sibo Cheng, Nina Kovalchuk, Mark Simmons, Omar K Matar, Yi-Ke Guo, and Rossella Arcucci. Ensemble latent assimilation with deep learning surrogate model:

application to drop interaction in a microfluidics device. *Lab on a Chip*, 22(17):3187–3202, 2022.

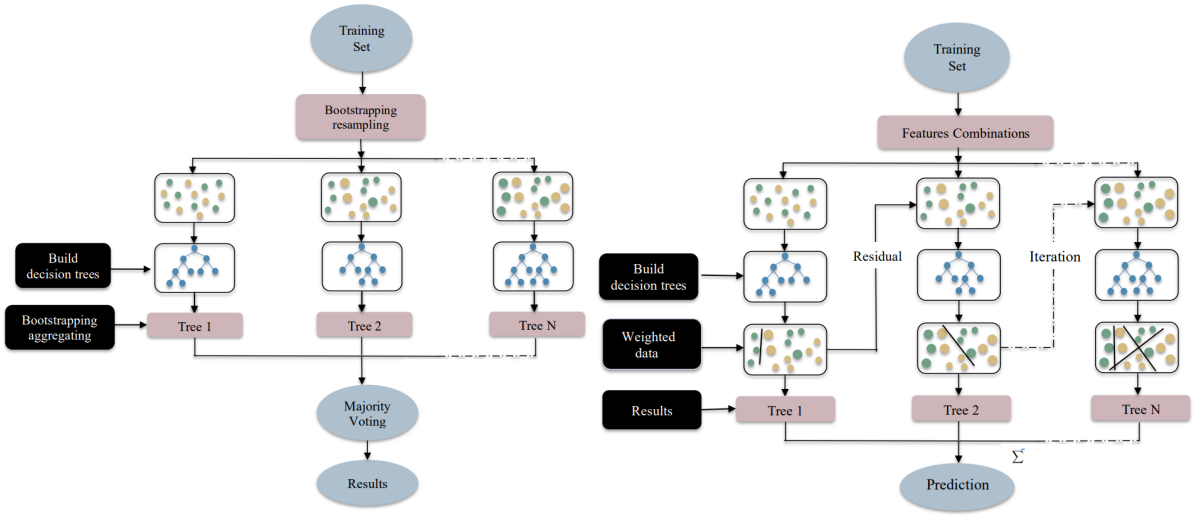
Appendices

A Preliminaries

A.1 Predictive Models

A.1.1 Random Forest and XGBoost:

Random Forest and XGBoost are both ensemble machine learning algorithms that utilize decision trees as their fundamental building blocks, albeit with different methods [46]. Random Forest creates an series of decision trees, each constructed independently through the utilization of bootstrapped samples from the dataset in a concurrent process [24]. Conversely, XGBoost constructs trees in a sequential manner, whereby each successive tree seeks to ameliorate the errors perpetrated by its predecessor [16]. Consequently, XGBoost frequently attains better performance; however, it is more computationally demanding and necessitates meticulous tuning of hyperparameters in comparison to random forest. In contrast, random forest is typically more flexible in training, exhibits robustness, and often delivers satisfactory performance with default configurations [55]. Moreover, due to these two tree-based methods always show a powerful ability to process tabular data in small or medium-sized datasets, they are applied in this research with the expectation that they can demonstrate better performance than neural networks [25]. The intuitive representation of these two models are shown in Figure. 9.



(a) Random Forest (b) XGBoost

Figure 9: Visualization of tree-based predictive models

A.1.2 Deep Neural Networks and Multilayer Perceptron:

Deep Neural Networks (DNNs), inclusive of the widely utilised Multilayer Perceptron (MLPs), are marked by their layered structure and copious count of artificial neurons. The depth of these networks, coupled with the non-linear activation functions employed within their hidden layers, equips them with the capability to discern and replicate highly complex patterns within data, thus demonstrating their proficiency in modelling non-linear relationships [40]. However, their large parameter space renders them computationally demanding [67]. Nevertheless, their capacity to adeptly handle various types of data has fostered their wide application across diverse tasks, such as machine translation, sentiment analysis, image and speech recognition,

anomaly detection, text classification, as well as various regression and classification problems [61, 69].

A.2 Grid Search Method

Grid search is a technique for exploring a predefined set of hyperparameters to optimize a machine learning algorithm. [35]. It can carefully traverse multiple combinations of hyperparameter configurations, performing cross-validation to determine the configuration that yields the best performance. In this study, the Grid Search method is employed to optimize both tree-based models and MLPs, utilizing 5-fold cross-validation for model training.

A.3 Performance Metrics

The confusion matrix serves as a visualization tool for evaluating the performance of machine learning models which are mentioned in section. A.1. A confusion matrix, in its simplest form, is a two-dimensional matrix that visualizes the performance of a supervised learning algorithm. For binary classification task, it has four entries:

- True Positives (TP): Occurrences in which both the actual outcome and the model's prediction are positive.
- True Negatives (TN): Occurrences in which both the actual outcome and the model's prediction are negative.
- False Positives (FP): Occurrences where the model erroneously classifies a negative instance as positive.
- False Negatives (FN): Occurrences where the model erroneously classifies a positive instance as negative.

These entries can be used to calculate various performance metrics [10]. The primary metric, accuracy, gauges the proportion of correct predictions. However, precision, recall, and the F1 score are also crucial for a more nuanced understanding of the model's performance. Precision focuses on the correctness of positive predictions, while recall assesses how well the model captures all actual positive cases. The F1 score harmonizes these two metrics, offering a balance between them. Four performance metrics are defined as equation (5) - (8):

$$\text{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

$$\text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

B Additional Analysis

B.1 Feature Ablation Testing

Feature ablation is used to empirically determine the impact of specific features on ML model's performance. By systematically removing or altering a feature and then evaluating the model's performance, we can gain insights into the true significance of that feature for the model's predictions. It essentially provides a way to validate feature importances derived from XAI techniques. The feature ablation testing results are shown in Table. 3.

Table 3: Feature Ablation Testing (Evaluated By Accuracy)

Model	Baseline	w/o Heff	w/o $\frac{D}{W}1 + \frac{D}{W}2$	w/o flow	w/o $\lvert\frac{D}{W}1 - \frac{D}{W}2\rvert$	w/o dt
Random Forest	**74.42%**	72.09%	70.10%	74.09%	72.09%	70.10%
XGBoost	71.76%	66.11%	72.43%	72.09%	72.43%	**74.09%**
MLP	**78.07**%	69.77%	75.08%	76.08%	76.74%	75.75%

For the Random Forest model, the SHAP analysis placed 'Heff' as the primary influential feature. However, the results from the feature ablation suggest a modest performance drop when this feature is removed. This difference between the expected and observed impact of 'Heff' imply the inherent robustness of Random Forests. The ensemble nature of Random Forests, consisting of a multitude of decision trees, might allow it to adapt to the absence of even a critical feature. Each individual tree captures different facets of the dataset, and collectively, they may maintain performance levels even when a significant feature is missing. This resilience may cause a divergence between the perceived importance from techniques like SHAP and the empirical results from feature ablation. Nevertheless, it's notable that removing 'dt' leads to a more considerable performance decrease, from 74.42% to 70.10%. This discrepancy between the SHAP analysis and the feature ablation results suggests that 'dt' has a more intricate role than previously assumed.

In contrast, the XGBoost model exhibits a unique trend compared to the other models. While the removal of most features enhances its performance relative to the baseline, 'Heff' stands as a clear exception. The omission of 'Heff' incurs a significant drop in performance, highlighting its importance, which is consistent with the SHAP rankings. The improved performance upon the removal of '$\frac{D}{W}1 + \frac{D}{W}2$' and 'flow', despite their indicated significance by SHAP, implies the possibility that these features, in the presence of other variables, introduce complexity or ambiguity which the XGBoost model finds hard to navigate. In simpler terms, the model might achieve clearer and more accurate decision boundaries when these features are absent. Even more intriguing is the noticable enhancement in performance upon excluding 'dt'. This could suggest that 'dt', within the context of the XGBoost model, may be contributing a level of noise or might be entangled with other features in a manner that hampers the model's predictive clarity. Despite sharing a tree-based structure with Random Forest, XGBoost demonstrates a distinct sensitivity pattern when features are omitted. This variance can be attributed to XGBoost's boosting mechanism, where trees are constructed sequentially to correct the preceding trees' errors. In such a setting, the absence of potentially confounding features can lead to clearer error correction paths, ultimately boosting performance. Moreover, according to Table. 3 and Figure. 6b, it's evident that XGBoost, although structurally similar to Random Forest as a tree-based model, reacts more sensitively to the omission of vital features. This distinction could be attributed to the boosting mechanism of XGBoost, which sequentially constructs trees to correct the errors of the preceding outputs. Consequently, each tree is more reliant on crucial features to correct the previous errors and enhance the model's predictive capacity.

Upon examining the MLP's results, one notes a distinct pattern that diverges from the tree-

based models. First and foremost, the removal of 'Heff' leads to a significant drop in accuracy, which stands in alignment with the SHAP results, denoting its importance. Nevertheless, an observation shows when assessing the interchange between '$\frac{D}{W}1 + \frac{D}{W}2$' and 'flow'. While their SHAP rankings have a switch, the performance variance when either is excluded from the model is rather moderate. This suggests that, within the MLP structure, these features may exhibit an interconnected influence, possibly sharing redundant information or compensating for one another when absent according to all three models' results. Lastly, the exclusion of 'dt' also presents an abnormal scenario. Although SHAP results suggest its relative insignificance, the model's accuracy actually declines when this feature is neglected. This counters the expectation based on the SHAP analysis, suggesting that 'dt' have a hidden or non-linear contribution that isn't fully captured by the importance ranking. This finding can also be proved by random forest's results and indicates the complex interplay and hidden dependencies that might exist between 'dt' and other features within the densely connected MLPs framework.

B.2 Observation and Discussion

This section synthesizes key insights from our XAI analysis on three machine learning models: Random Forest, XGBoost, and MLP, focusing on feature importance and model sensitivity to feature ablation.

B.2.1 SHAP Feature Rankings:

- **Random Forest**: Heff $>$ $\frac{D}{W}1 + \frac{D}{W}2 >$ flow $> |\frac{D}{W}1 - \frac{D}{W}2| >$ dt
- **XGBoost**: Heff $>$ $\frac{D}{W}1 + \frac{D}{W}2 >$ flow $> |\frac{D}{W}1 - \frac{D}{W}2| >$ dt
- **MLP**: Heff $>$ flow $> \frac{D}{W}1 + \frac{D}{W}2 > |\frac{D}{W}1 - \frac{D}{W}2| >$ dt

B.2.2 General Observations:

- 'Heff' consistently appears as the most crucial feature across all models and interpretability methods, aligning with its top SHAP ranking in each model.
- '$\frac{D}{W}1 + \frac{D}{W}2$' and 'flow' are also frequently significant but their importance ranking varies between models, as indicated by SHAP.
- The importance of 'dt' is consistently lowest across all models in SHAP rankings, but its actual effect, particularly in the MLP model, suggests more complex relationships.
- Different models react differently to feature omission despite having similar SHAP rankings, highlighting their unique sensitivities and structural differences.
- There is a strong alignment between the features that are significant globally (via SHAP and feature ablation) and locally (via LIME), suggesting the robustness of these features.

B.2.3 Discussion:

- **Feature Robustness**: 'Heff' consistently maintains its top SHAP ranking and shows significant impact in both local (LIME) and global (feature ablation) interpretability analyses, confirming its critical role. Similarly, the features '$\frac{D}{W}1 + \frac{D}{W}2$' and 'flow' are not only statistically significant but also practically significant, making substantial contributions in both local and global interpretability analyses.
- **Model Sensitivity**:Although Random Forest and XGBoost share similar SHAP rankings, they exhibit different resilience to feature omission, highlighting the nuanced differences

between their tree-based architectures. Specifically, Random Forest's ensemble mechanism lends it robustness to the absence of critical features, while XGBoost's boosting technique makes it more sensitive to such omissions.

- **Hidden Dependencies in MLP**: The lowest SHAP ranking of 'dt' in all models contrasts with its actual impact on model performance, hinting at complex, hidden dependencies. This discrepancy in MLP suggests that 'dt' may interact non-linearly with other features, a behavior more easily captured by the densely connected architecture of MLPs. Therefore, 'dt"s role is more nuanced than linear methods like SHAP can reveal, reflecting the intricate feature interdependencies inherent to neural networks
- **Congruence Across Methods**: The congruence between LIME, SHAP, and feature ablation tests speaks to the reliability and validity of these critical features in the dataset. The consistency across different analytical methods and predictive models reaffirms that these features hold not just statistical but also practical significance in the phenomena being studied.
- **Local vs Global Interpretability**: The consistency between LIME and SHAP rankings for significant features suggests that these models, although complex, may behave linearly or at least predictably in the vicinity of specific instances. This adds a layer of trust to the predictive power and interpretability of these models.

Overall, the MLP model emerges as the most effective in predicting microfluidic droplet coalescence, outperforming its tree-based counterparts, Random Forest and XGBoost, in both accuracy and resilience to feature ablation. While the tree-based models share similar SHAP rankings, they display unique sensitivities to feature exclusion due to their internal architectures. MLP, however, stands out for its superior ability to capture complex feature interdependencies, as evidenced by its performance when the lowest-ranking feature 'dt' is removed. This indicates a more nuanced and adept predictive framework, confirming MLP's suitability for this specific task.

C MLP Experimental Hyper-parameters

This section details the optimal hyperparameters for our Multilayer Perceptron (MLP) model, as discussed through grid search techniques.

Table 4: Hyperparameters Tuning for MLP

Hyperparameter	**Options**
Activation Functions	**relu**, tanh, sigmoid, linear
Optimisers	**adam**, sgd, rmsprop
Learning Rates	**0.001**, 0.01, 0.1
L2 Rates	0.0, 0.001, **0.01**, 0.1
Dropout Rates	0.0, **0.1**, 0.2, 0.3
Epochs	10, 50, 100, 150, 200, 250, 300, 400, 450, 500, 800, 1000, 1500, 2000, **2500**, 3000