TIME-SERIES PREDICTION OF SPRAY CHARACTERISTICS USING CLASSICAL MACHINE LEARNING APPROACHES

V. P. Harshak *School of Electrical Engineering, Vellore Institute of Technology,* Vellore, India
harshak.vp2023@vitstudent.ac.in
P. Suresh Kumar *Automotive Research Centre, Vellore Institute of Technology,* Vellore, India
suresh.kumar@vit.ac.in

ABSTRACT [To be rewritten with quantified results]

Accurate prediction of spray characteristics, such as spray angle and penetration length, is essential for enhancing control and efficiency in applications like fuel injection in combustion engines, agricultural spraying, and thermal systems. This study evaluates six baseline machine learning models namely linear regression, decision tree, random forest, gradient boosting, support vector regressor, and K-Nearest neighbours on a small time-series dataset derived from spray related experiments measured via shadowgraph and Mie scattering. The dataset includes six input features and four output variables comprising spray angle and spray penetration length. An 80/20 time-aware holdout split was used for validation, and model performance was assessed using mean squared error (MSE), mean absolute error (MAE), and R^2 score. Among the models tested, gradient boosting delivered the most consistent and accurate results, achieving R^2 scores of 0.9906 for spray angle (shadowgraph), 0.9998 for spray length (shadowgraph), 0.9990 for spray angle (Mie), and 0.9989 for spray length (Mie), with low associated errors. These results highlight the reliability and generalizability of gradient boosting for predicting spray parameters, even with limited data, making it a strong candidate for real-time applications in experimental and industrial spray systems.

Keywords: machine learning, gradient boosting regressor, time-series data, spray angle, spray penetration length

1. Introduction

Spray systems are fundamental to a wide range of real-world applications, including agricultural pesticide dispersion, internal combustion engines, combustion chamber in gas turbine engines and industrial coating processes. In these systems, the ability to control and predict spray behaviour, particularly spray angle and penetration length, is essential for achieving uniform coverage, maximizing system efficiency, and minimizing material waste. Accurate modeling of

spray characteristics directly contributes to improved performance, reduced operational costs, and enhanced environmental compliance. Traditional methods for analysing spray dynamics rely on experimental visualization techniques such as shadowgraph technique and Mie scattering, or on physics-based simulations using computational fluid dynamics (CFD). While these approaches offer detailed insights, they often require specialized equipment, significant computational time, and expertise to interpret, making them less suitable for rapid decision-making or integration into real-time control systems. **A detailed write-up for the difficulties associated with the experiments and modelling is to be included with references.**

In order to overcome the concerns involved in the experimental investigation and CFD analysis, baseline machine learning (ML) models provide a data-driven means of capturing complex relationships between input conditions and spray outcomes as a viable alternative strategy. These models include Linear Regression, Decision Trees, and Gradient Boosting are known for their simplicity, interpretability, and effectiveness even with modest datasets. Their computational efficiency and adaptability make them well-suited for predictive tasks in experimental and industrial contexts [references].

Recent research has increasingly explored the use of ML techniques to model spray behavior in fluid and combustion systems. In this pursuit, artificial neural network (ANN) models were employed to predict diesel engine performance under various injection strategies using biodiesel fuels [1]. While the model demonstrated high prediction accuracy, it required significant training data and parameter tuning, making it less interpretable for practical applications. In another work [2], long short-term memory (LSTM) networks were used to model spray dynamics in compression ignition (CI) engines fueled with nanotechnology-enhanced fuels. This study showed strong predictive capability for time-dependent spray behavior across different combustion chamber geometries. However, the approach was computationally intensive and dependent on high-fidelity simulations, limiting its real-time feasibility. A more interpretable alternative was demonstrated by Khan et al. [3], where gradient boosting algorithms were applied to predict spray penetration and cone angle. The hybrid approach combined experimental data and feature-engineered inputs, delivering high accuracy and offering greater model transparency. However, the study focused primarily on geometric spray outputs rather than full multivariate prediction.

Some more literature is required similar to the reference 1,2 and 3 along with their limitations to define the novelty of our work. Table 1 provides the state-of-the-art concerning the various machine learning approaches adopted for predicting the spray characteristics.

Therefore, it is evident that while deep learning and hybrid techniques are widely adopted in these domains, they often demand large datasets and high computational resources. In order to address these issues, the present study evaluates lightweight and interpretable baseline ML models including Linear Regression, Decision Trees, Random Forest, Gradient Boosting, SVR, and KNN for predicting multiple spray characteristics from a compact time-series dataset. The novelty of the current work is to conduct a comparative evaluation of six baseline ML models for predicting spray angle and penetration length using a small, time-series dataset collected from controlled experimental runs. The goal is to assess the models' accuracy, generalization, and applicability to spray prediction tasks. By integrating experimental observations with ML techniques, this study aims to advance the understanding of spray behavior in fuel injection systems and support the development of more efficient combustion strategies. **A conceptual diagram of the spray system may be included to provide readers with physical context for the experimental setup and data structure**.

Table 1. State-of-the-art of ML techniques for spray modelling

Reference	**Model(s) Used**	**Application Domain**	**Limitation**
[1]	ANN	Diesel Engine Emissions	Requires large data and tuning
[2]	LSTM	CI Engine Spray Dynamics	High complexity, simulation-heavy
[3]	Gradient Boosting	Diesel Injector Geometry	Focused only on geometric outputs
Some more studies may be added.			

2. Materials and Methodology

The experiments were performed using advanced optical diagnostics such as shadowgraph and Mie scattering techniques to capture detailed spray characteristics under various injection conditions. The present section provides a comprehensive description of the experimental test rig,

including the fuel injection system, imaging setup, and measurement instruments. It also elaborates on the experimental protocols followed to ensure consistency and repeatability across all trials. Following data acquisition, a suite of machine learning (ML) models was developed to predict and analyse spray behaviour based on key input parameters such as injection pressure, ambient conditions, and fuel properties. This study is based on a well-structured time-series dataset comprising 565 data samples, meticulously collected through a series of controlled fuel spray experiments. This section further discusses the data pre-processing steps, feature selection strategies, and model training and validation procedures employed to build robust predictive models.

2.1.Experimental setup and procedure

A detailed write-up for the experimental setup and procedure is to be included with relevant diagram, Table etc. [German team]

2.2. Dataset Description and Data pre-processing

The dataset includes six input features comprising the various parameters that affect the spray characteristics significantly. The physical parameters influencing the spray behavior are time step of the observation, chamber pressure, chamber temperature, fuel injection pressure, fuel density, fuel kinematic viscosity while the output variables are the dataset that includes the target variables representing the spray characteristics in terms of spray angle and spray length which are measured using two optical methods such as shadowgraph and Mie scattering techniques. The various input parameters and their ranges and the target variables are shown in Table 2 (provide the experimental ranges). These targets are quantitative indicators of spray structure and are treated as dependent regression variables. All samples are grouped by a Run ID, which serves as a temporal identifier for each experiment. Preserving the chronological order within each run is essential to reflect the dynamic nature of spray development over time.

Table 2. Input and output variables

Input parameters			**Output (target) parameters**
Description	**Feature Name**	**Range**	

Time step of the observation (ms)	Time	-	Spray angle (degrees) and spray length (mm)
Chamber pressure (bar)	Cham Pres	X to Y bar	
Chamber temperature (K)	Cham Temp	X to Y K	
Injection pressure (bar)	Inj pres	X to Y bar	
Fuel density (kg/m^3)	Density	Value	
Fuel kinematic viscosity (m^2/s)	Viscosity	Value	

The modeling task is framed as a multi-output regression problem, where the objective is to simultaneously predict all four spray characteristics from the given six input parameters. This formulation enables the model to exploit correlations between outputs while maintaining a compact training pipeline. Due to the time-series structure of the dataset, where samples are sequentially grouped by Run ID, the temporal order of observations is maintained throughout model development to prevent data leakage. Consequently, no data shuffling is performed prior to splitting the dataset. Each sample represents a time step from a labeled experimental run (Run ID) and records both the input conditions and corresponding spray characteristics.

The dataset comprises 565 time-sequenced samples collected across multiple experimental runs, each identified by a distinct Run ID. Each record includes six input features and four spray-related target variables, measured at specific time intervals. To preserve the chronological nature of the data, records were sorted by Run ID and Time (ms). No missing values or datatype inconsistencies were found, confirming the dataset was ready for modeling after scaling.

All input and output variables were standardized using StandardScaler, ensuring zero mean and unit variance. This step is particularly vital for distance-based models such as SVR and KNN, which are sensitive to feature magnitudes. Crucially, scaling was applied after preserving time-ordering, thereby preventing any data leakage into the modeling pipeline.

2.3. Machine learning model development [This section needs proper coherence]

This study investigates six baseline machine learning (ML) models for the prediction of spray characteristics—specifically spray angle and length—using time-series data obtained from experimental injection runs. These ML models were selected for their diversity in learning strategies, interpretability, and effectiveness in practical regression tasks. The models used in this current analysis include the following:

1) Linear Regression (LR): A simple, interpretable model that assumes a linear relationship between inputs and outputs.

2) *Decision Tree Regressor (DT):* A tree-based model capable of capturing non-linear patterns with rule-based splits.

3) *Random Forest Regressor (RF):* An ensemble of decision trees that improves robustness and reduces variance.

4) *Gradient Boosting Regressor (GB):* A boosting-based ensemble method that builds sequential trees to reduce bias.

5) *Support Vector Regressor (SVR):* A kernel-based method effective in capturing complex relationships, but sensitive to scaling.

6) *K-Nearest Neighbors (KNN):* A non-parametric model that predicts based on similarity to training data points.

[The techinque invlulved behind All the above ML model should be descrobe]

The various strategies adopted in the ML model development are summarised as follows:

2.3.2 Multi-output regression strategy

The target variable comprises four continuous outputs: spray angle and spray length each obtained from Shadowgraph and Mie techniques which make the task a multi-output regression problem. Models such as Linear Regression, Decision Tree, and Random Forest natively support multi-output learning. For models that do not (e.g., SVR, KNN, Gradient Boosting), the MultiOutputRegressor wrapper from scikit-learn was employed to enable training on multiple outputs.

2.3.3 Training setup and data handling

To evaluate model performance, the dataset was divided using a standard split of 80:20 training and testing. No data shuffling was applied to preserve the temporal sequence of the experimental runs, ensuring that the test data represents future observations relative to the training set. Feature scaling was applied using StandardScaler to normalize both input features and target variables. This transformation was implemented within a Pipeline structure in scikit-learn to prevent data leakage during training and evaluation. Scaling is especially critical for distance-based and kernel-based models like KNN and SVR.

2.3.4 Hyperparameters

Most models were trained using default hyperparameters to maintain simplicity and reproducibility and no extensive hyperparameter tuning was performed in this study. However, key parameters were adjusted where necessary to ensure functional model behavior. The Random Forest and Gradient Boosting were trained with n_estimators = 100 (default). The SVR and KNN were wrapped for multi-output training and used standard parameters (e.g., C=1.0, n_neighbors =

5). Figure 1 shows the overall modeling pipeline involved in this investigation while Table 3 provides the basic configurations of the ML models adopted in this study.

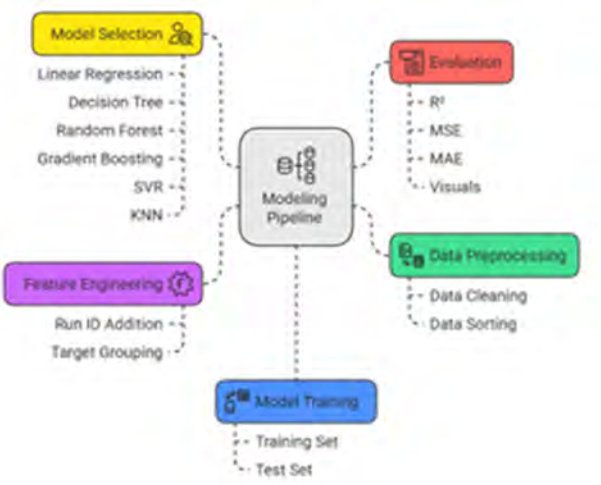


Fig. 1. Modeling Pipeline for Experimental Spray Data

Table 3. Baseline Machine Learning Models and Configuration

Model	Type	Multi-Output Handling	Key Hyperparameters
Linear Regression	Linear	Native	None (default)
Decision Tree	Tree-based	Native	random_state = 42
Random Forest	Ensemble (Bagging)	Native	n_estimators = 100, random_state = 42
Gradient Boosting	Ensemble (Boosting)	MultiOutputRegressor	n_estimators = 100, random_state = 42
SVR	Kernel-based	MultiOutputRegressor	C = 1.0, kernel = 'rbf'
K-Nearest Neighbors	Distance-based	MultiOutputRegressor	n_neighbors = 5

2.3. Performance Metrics of the Model

Model performance was evaluated using the standard regression metrics namely coefficient of determination (R^2), mean squared error (MSE) and mean absolute error (MAE) as given below:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (1)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3)$$

The R^2 values quantifies the proportion of variance in the target variable explained by the model. Values closer to 1 indicate a better fit. MSE captures the average squared difference between actual and predicted values. Lower values indicate higher prediction accuracy. MAE represents the average absolute difference between predicted and true values, expressed in the same unit as the target. Lower values denote more precise predictions.

3. RESULTS AND DISCUSSION

This section presents the sample results obtained from the experimental investigations using Shadowgraph and Mie scattering techniques and a comprehensive performance analysis of six baseline machine learning models applied to predict the essential spray characteristics in terms of srpay angle and spray length. The models evaluated include Linear Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Regression (SVR), and K-Nearest Neighbors (KNN).

3.1. Experimental data

The temporal evolution of spray characteristics is depicted in Fig. 2 shows the four target variables over time (after the fuel injection) for a sample experimental run (provide the input conditions for the Figure 2). The plots confirm observable trends and justify the importance of maintaining temporal consistency in modeling. Reasons must be given for the deviation in spray angle between shadowgraph and Mie. This time-series structure plays a critical role in understanding how changes in physical inputs influence spray dynamics over successive moments. Fig. 2 also explains the temporal dynamics of spray characteristics, the evolution of all four target variables such as spray angle (Shadowgraph and Mie) and spray length (Shadowgraph/Mie), against the time for a representative experimental run (Run ID 0). As noticed from Fig. 2, the spray

angle and length exhibit typical injection patterns, with a noticeable increase in the early phase followed by stabilization or decline as time progresses. These observations emphasize the sequential nature of the data and underscore the importance of preserving temporal order during preprocessing and model evaluation. The presence of clear time-dependent trends further justifies the treatment of this task within a time-aware modeling framework.

Similar Figure can be provided for other experimental conditions to emphasis the importance of six input variables that affect the spray behaviour.

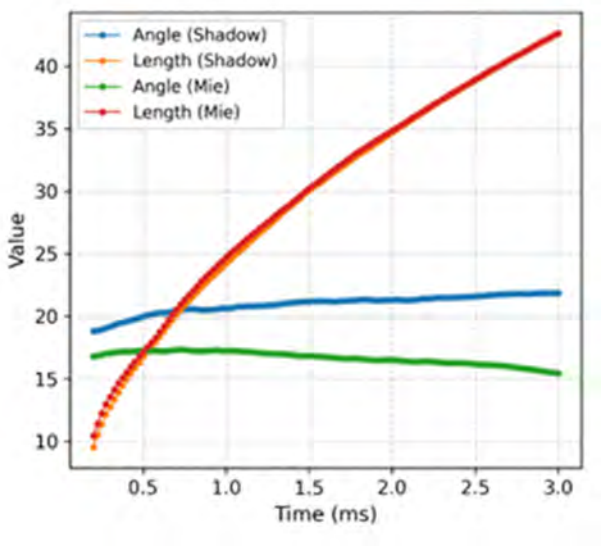


Fig. 2. Measured time dependeant spray angle and spray length (Shadowgraphy and Mie methods).

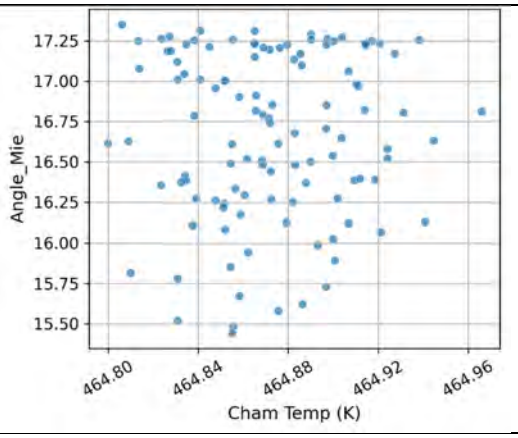
A sample data obtained from the experiments is presented in Table 4 to illustrate the structure and format of the data. Table 4 also shows the temporal variations in spray angle and spray length after the start of injection. Short notes on spray penetration and dispersion of spray (increasing spray angle with respect to the time is to be included.

Table 4: A sample data for the measured spray length and spray angle

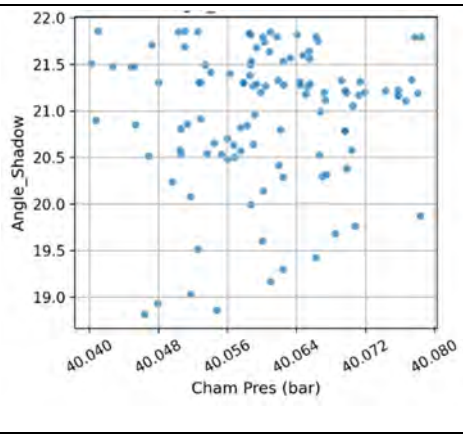
Time (ms)	Cham Pres (bar)	Cham Temp (K)	Inj pres (bar)	Density (kg/m^3)	Viscosity (m^2/s)	Angle (Shadowgraph)	Length (Shadowgraph)	Angle (Mie)	Length (Mie)	Run ID
0.2	40.046	464.86	85.38	749	$4.47*10^{-07}$	18.81	9.513	16.79	10.431	0
0.225	40.054	464.89	85.36			18.85	10.512	16.85	11.391	0
0.25	40.047	464.86	85.35			18.93	11.356	16.91	12.241	0

0.275	40.0517	464.84	85.37			19.02	12.130	17.01	12.962	0
0.3	40.061	464.90	85.39			19.16	12.775	17.06	13.549	0

To visually explore feature influence on spray characteristics, scatter plots were generated for each input-output combination, grouped by Run ID to reduce visual clutter which is shown in Fig. 3.



Scatter Plot: Chamber Temperature vs. Angle (Mie)



Scatter Plot: Chamber Pressure vs. Angle (Shadowgraph)

Fig. 3. Effect of chamber temperature and chamber pressure on spray angle

From Fig. 2, it is observed that the chamber temperature shows a strong negative trend with spray angle (Mie), aligning with the high negative correlation (–0.83) in the heatmap. Similarly, chamber pressure also displays a moderate negative relationship with spray angle (Shadowgraph), consistent with its correlation value (–0.57). [In Fig 3, the change in chamber temperature (only 0.04 K) and pressure (0.08 bar) is very minimal. How it can be justified and explained?].

The injection pressure, though not strongly correlated with any target (e.g., –0.13 with spray length (Shadowgraph)), was still visualized for completeness. To avoid visual clutter, plots were generated for individual Run IDs, offering representative insights across different runs. [where is the diagram for this statement concerning the injection pressure?]

3.2. Correlation Analysis

To investigate linear relationships between features and spray characteristics, a Pearson correlation heatmap was generated using all 565 samples. The following are the notable insights obtained fromt his analysis:

- Injection Pressure shows a moderately strong negative correlation with spray angle (Mie) (–0.63) and spray Length (Mie) (–0.17), suggesting an inverse but limited influence on Mie-based measurements.

- Chamber Temperature has a strong negative correlation (–0.83) with spray angle (Mie), indicating that temperature plays a key role in angular dispersion under Mie conditions.
- Chamber Pressure demonstrates a moderate negative correlation (–0.57) with spray angle (Shadowgraph), while showing little to no correlation with other targets.

To understand how these relationships vary across experiments, per-run heatmaps were also generated (e.g., Run ID 0), revealing minor fluctuations in correlation strength as shown in Fig. 4.

.

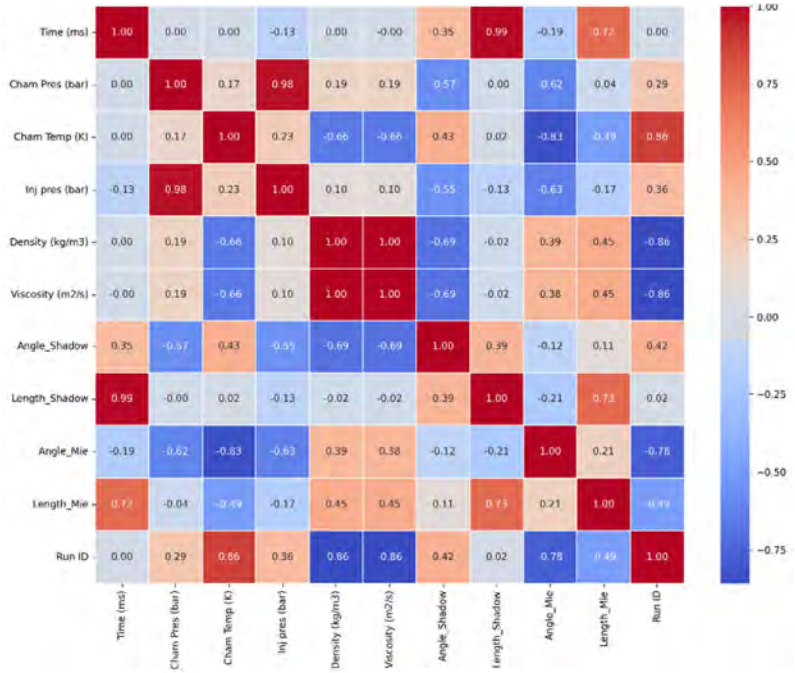


Fig. 4: Correlation Heatmap of All Features and Targets [I don't understand this. This is to be explained clearly]

3.3. Model Validation

This study evaluated six baseline machine learning models for predicting spray characteristics—specifically spray angle and length measured through shadowgraph and Mie scattering techniques. The results demonstrated that models such as Gradient Boosting and K-Nearest Neighbors consistently outperformed others in terms of R^2, MSE, and MAE across all four target variables. Table 5 summarizes the performance of all six models across the four output targets. The results reveal that Gradient Boosting and K-Nearest Neighbors (KNN) consistently outperform other models, achieving high R^2 scores and low MSE and MAE values for both angle and length predictions.

Table 5. Performance Comparison of ML Models Across all Targets

Model	Target	MSE	MAE	R^2
Linear Regression	Angle (Shadowgraph)	0.222281	0.35953	0.86318

	Length (Shadowgraph)	1.525607	1.03754	0.98019
	Angle (Mie Scattering)	0.069239	0.17934	0.98547
	Length (Mie Scattering)	11.52545	2.95737	0.78911
Decision Tree	Angle (Shadowgraph)	0.048168	0.0997	0.97035
	Length (Shadowgraph)	0.291097	0.44944	0.99622
	Angle (Mie Scattering)	0.002664	0.03478	0.99944
	Length (Mie Scattering)	0.207541	0.33865	0.9962
Random Forest	Angle (Shadowgraph)	0.017672	0.08738	0.98912
	Length (Shadowgraph)	0.067421	0.21122	0.99913
	Angle (Mie Scattering)	0.007448	0.05538	0.99844
	Length (Mie Scattering)	0.053384	0.17425	0.99902
Gradient Boosting	Angle (Shadowgraph)	0.0152	0.08944	0.99064
	Length (Shadowgraph)	0.01423	0.09214	0.99982
	Angle (Mie Scattering)	0.004665	0.04953	0.99902
	Length (Mie Scattering)	0.062559	0.19408	0.99886
Support Vector Regressor	Angle (Shadowgraph)	0.086478	0.16792	0.94677
	Length (Shadowgraph)	1.885788	0.77892	0.97552

	Angle (Mie Scattering)	0.030635	0.10773	0.99357
	Length (Mie Scattering)	2.010679	0.85452	0.96321
K-Nearest Neighbors	Angle (Shadowgraph)	0.003159	0.0365	0.99806
	Length (Shadowgraph)	0.179146	0.32767	0.99767
	Angle (Mie Scattering)	0.001776	0.02911	0.99963
	Length (Mie Scattering)	0.129028	0.26095	0.99764

Table 5 revelas that the Gradient Boosting achieved near-perfect R^2 scores (above 0.99) for all targets, reflecting both its robustness and its ability to capture nonlinear dependencies in the data. Therefore, among the six baseline models, Gradient Boosting delivered the most consistent and accurate results, achieving minimal errors across all four spray targets. Its ensemble structure effectively captured nonlinear spray dynamics. K-Nearest Neighbors also performed well, particularly for angular predictions, indicating local similarity patterns. In contrast, Linear Regression struggled with Length (Mie), highlighting its limitations with nonlinear behavior. SVR and Decision Trees showed moderate accuracy but were sensitive to specific targets or underfitting. These results underscore the effectiveness of ensemble and local methods in modeling time-dependent spray characteristics.

3.4. Comparison between Measured and Predicted values

To visually assess the temporal performance of Gradient Boosting, a comparison is made between predicted and actual values over the time axis for each target using the test set and are shown in Figs. 5–8. From the Figs. 5-8, it is noticed that the predicted curves closely track the true spray dynamics across time, accurately capturing both the rising and stabilizing phases of angle and length. These plots confirm the model's ability to reflect real-world spray fluctuations and validate its reliability for time-series predictions. [Results are to discussed elaborately with a proper supporting references from the literature]

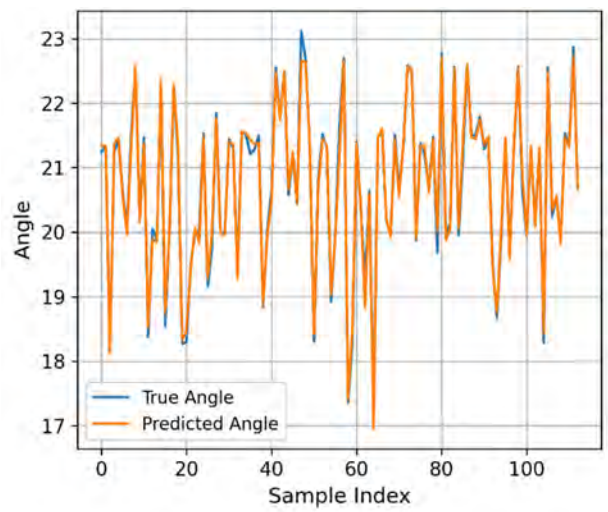


Fig. 5. Predicted vs. Actual Values for Angle (Shadowgraph) using Gradient Boosting

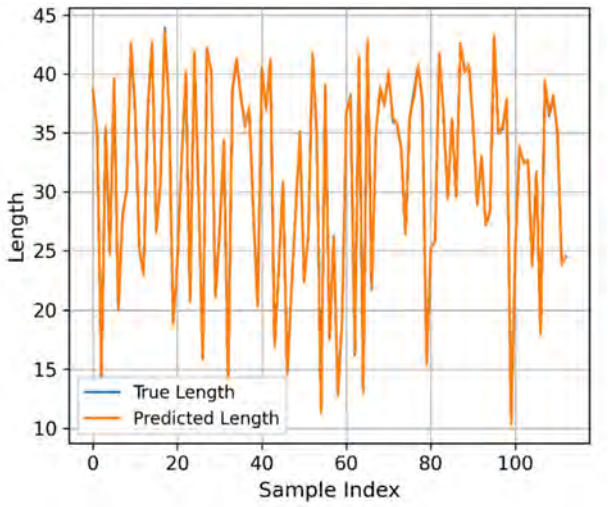


Fig. 6. Predicted vs. Actual Values for Length (Shadowgraph) using Gradient Boosting

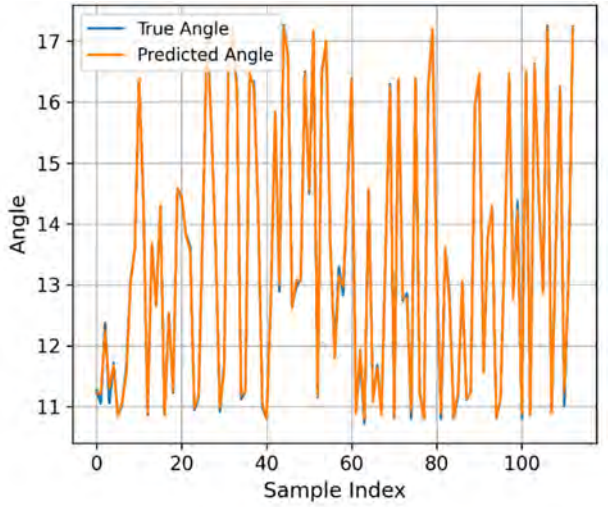


Fig. 7. Predicted vs. Actual Values for Angle (Mie) using Gradient Boosting

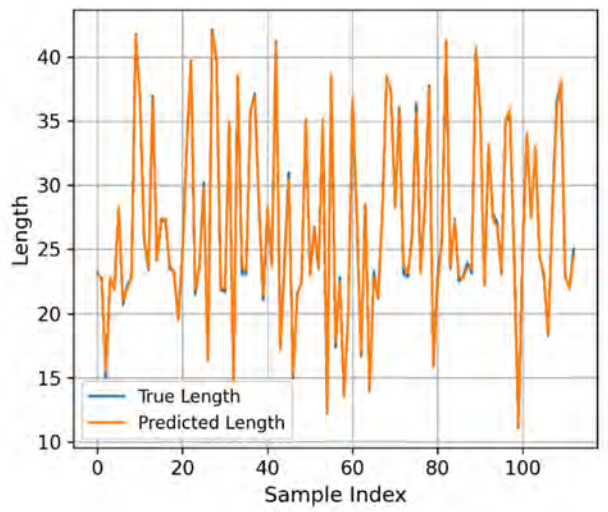


Fig. 8. Predicted vs. Actual Values for Length (Mie) using Gradient Boosting

[Place the figure legends uniformly]

The comparative regression plots for all the developed models (Figs. 9–12) further illustrate the performance gaps among baseline approaches. Gradient Boosting consistently produced tightly clustered predictions along the identity line across all targets, confirming its strong generalization and low residual variance. KNN also showed high alignment, particularly in angle prediction, despite requiring minimal hyperparameter tuning. Random Forest demonstrated good predictive accuracy, though with slightly more spread in length predictions compared to Gradient Boosting. SVR and Decision Tree models exhibited greater scatter in their predictions, reflecting sensitivity to target variance and potential underfitting. Linear

Regression, while interpretable, showed broader dispersion for nonlinear targets such as Length (Mie), underscoring its limitations in capturing complex spray dynamics. Collectively, these plots reinforce the superiority of ensemble-based models for accurate, multi-output spray prediction.

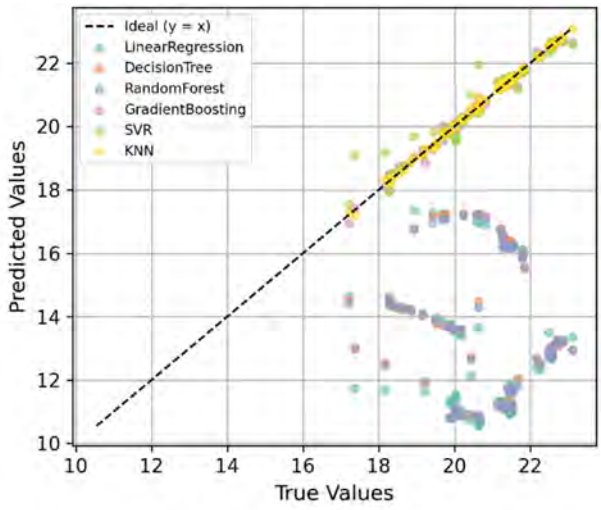


Fig. 9. True vs. Predicted Values for Angle (Shadowgraph) Across All Models

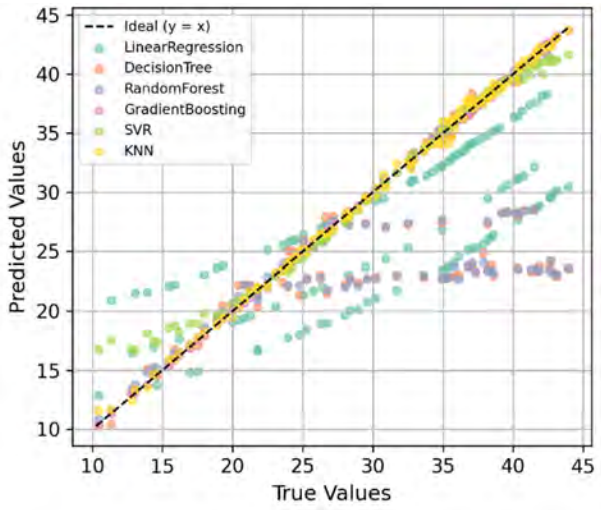


Fig. 10. True vs. Predicted Values for Length (Shadowgraph) Across All Models

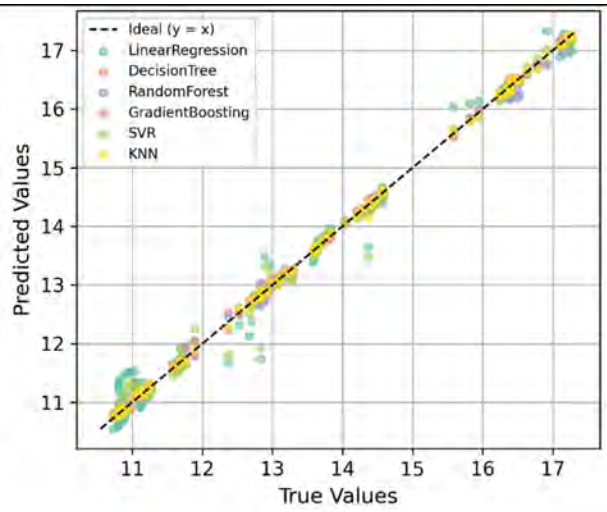


Fig. 11. True vs. Predicted Values for Angle (Mie) Across All Models

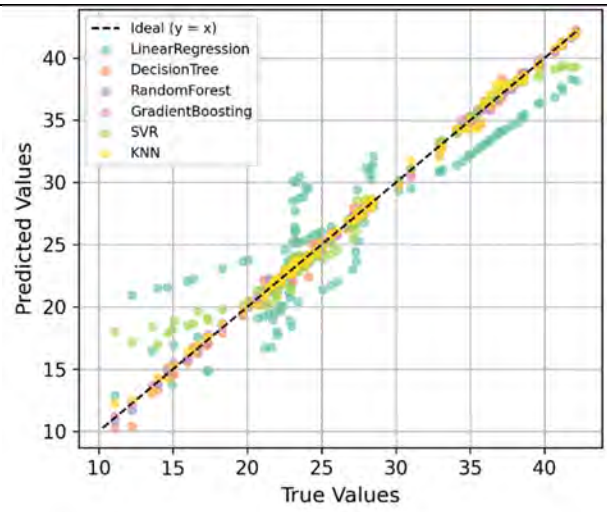


Fig. 12. True vs. Predicted Values for Length (Mie) Across All Models

3.5. Residual Analysis (Error Distribution)

Further, residual distributions were evaluated across all four targets to assess model reliability. Spray angle (Mie) is shown as a representative case in Figs. 13–15. Fig. 13 (KDE plots ???): shows Gradient Boosting and KNN show narrow, cantered residuals, indicating low bias. SVR and Decision Tree exhibit wider, asymmetric distributions. Fig. 14 (histogram + KDE) depicts the ensemble models which consists of concentrated error frequencies; others display broader, less stable patterns. Fig. 15 (Residuals vs. Predicted) reveals that the top models maintain consistent spread around zero, while underperformers

reveal scatter and variance. These insights align with quantitative metrics and confirm that ensemble models offer better error control, especially for complex targets. [This section is not impressive. Detailed explanation would improve the quality of the manuscript.]

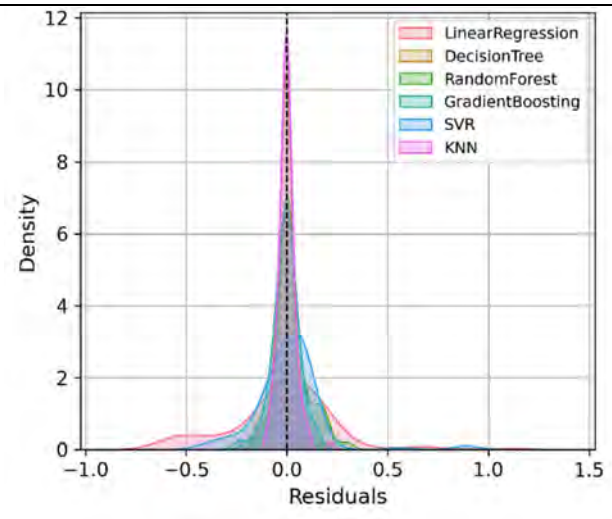


Fig. 13. KDE Residual Plots For Angle (Mie) Across All Models.

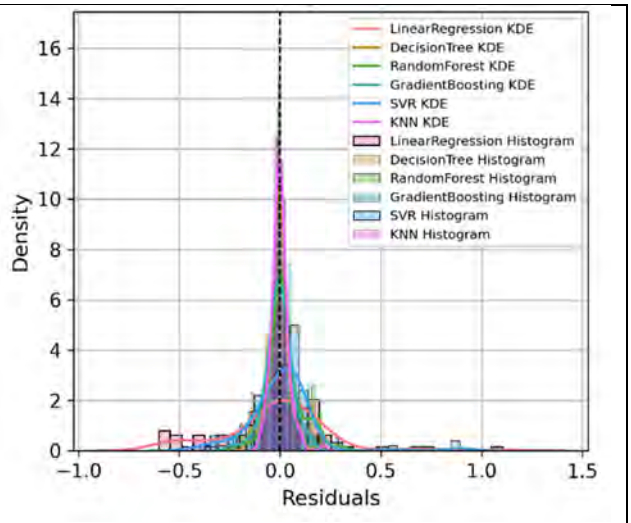


Fig. 14. Histogram And KDE Overlays Of Angle (Mie) Residuals Across All Models

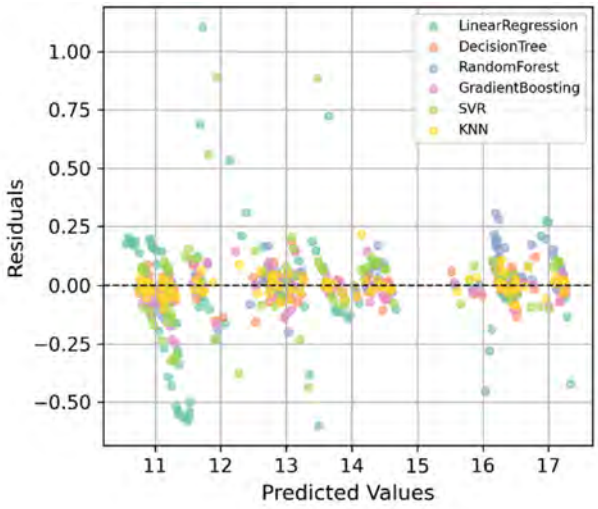


Fig. 15. Residuals vs. predicted values for Angle (Mie) Across All Models

One of the key strengths of this study lies in the effectiveness of relatively simple, interpretable models on a small, real-world time-series dataset. Models like Random Forest and XGBoost offer intuitive interpretability through feature importance scores, which can be invaluable in practical applications. In

addition, these models require significantly less training time and computational resources compared to more complex techniques, making them suitable for rapid prototyping or deployment in resource-constrained environments.

Despite the encouraging results, several limitations must be acknowledged. First, the dataset is small and limited to 565 sequential samples. This restricts the model's ability to generalize to unseen or highly variable operating conditions. Second, evaluation was based on a single 80/20 train-test split without cross-validation. While this preserved the temporal integrity of the sequence, it does not account for variability across different data segments or runs. Lastly, although feature importance plots indicate consistent influence from parameters such as injection pressure and chamber temperature, these relationships may shift under different experimental setups.

To support interpretability, Fig. 17 shows the feature importance ranking derived from the trained Random Forest model for predicting Angle (Mie). It highlights injection pressure and chamber temperature as the most influential features, aligning well with physical expectations of spray formation dynamics.

In conclusion, the use of baseline models offers a practical and explainable approach to modeling spray characteristics in small, experimental datasets. With further data and expanded evaluation methods, their predictive capabilities can be validated across broader industrial and research scenarios.

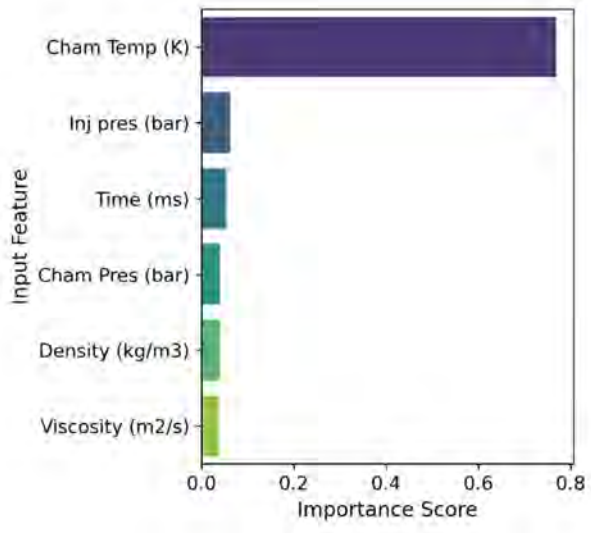


Fig. 16. Feature importance for predicting Angle (Mie) using Random Forest. Chamber temperature and injection pressure are the most influential inputs.

3.6. Comparison of ML models with Artificial Neural Network (ANN) model

To assess the performance limits of baseline models, an additional ~~experimental~~ comparison was conducted using an Artificial Neural Network (ANN). The ANN was developed in MATLAB with a feedforward architecture consisting of two hidden layers, each comprising 7 neurons. It was trained using the Levenberg–Marquardt algorithm and configured for multi-output regression with all four target variables predicted simultaneously. The network was trained multiple times, and the best-performing iteration was selected based on test set metrics.

Results showed that the ANN achieved slightly better predictive accuracy than Gradient Boosting across most targets, with higher R^2 scores and lower MAE/MSE scores in certain cases. This suggests that neural networks can offer improved performance when sufficient tuning and repeated training are applied. However, given the simplicity, interpretability, and faster training time of Gradient Boosting, it remains a practical and robust choice—especially in data-limited settings. Table 6 shows the comparison of error metrics (in terms of MAE, MSE, R^2) between Gradient Boosting and ANN model.

Table 6. Comparison of error metrics between ANN and Gradient Boosting for each spray characteristic target on the test set.

Model	Target	MSE	MAE	R2
Gradient Boosting	Angle (Shadowgraph)	0.0152	0.08944	0.99064
	Length (Shadowgraph)	0.01423	0.09214	0.99982
	Angle (Mie Scattering)	0.00467	0.0495	0.999
	Length (Mie Scattering)	0.06256	0.1941	0.9989
Artificial Neural Network (ANN)	Angle (Shadowgraph)	0.0104	0.0845	0.9953
	Length (Shadowgraph)	0.0056	0.0624	0.9999
	Angle (Mie Scattering)	0.0039	0.0505	0.999
	Length (Mie Scattering)	0.0528	0.156	0.9989

The regression plots (Figs. 17 - 20) are included for reference, with the ANN serving as a performance benchmark rather than a core model under evaluation [Need suitable explanation for these figures]. Fig 22 shows the KDE plots of residuals for all four target variables predicted by the ANN model. [Only figure is provided without any explanation – needs suitable write up].

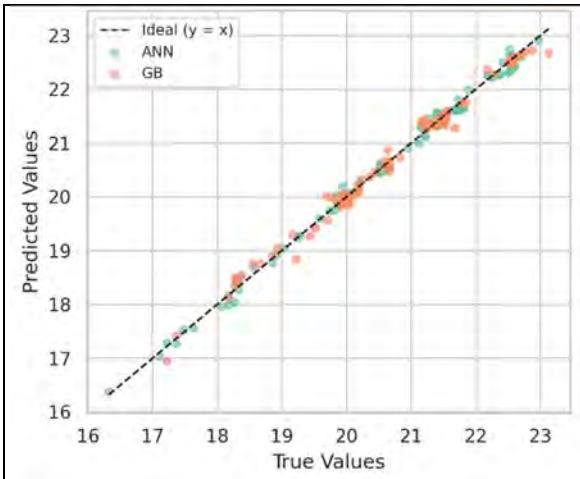


Fig. 17. True vs Predicted values for Spray Length (Shadowgraph) using ANN and Gradient Boosting

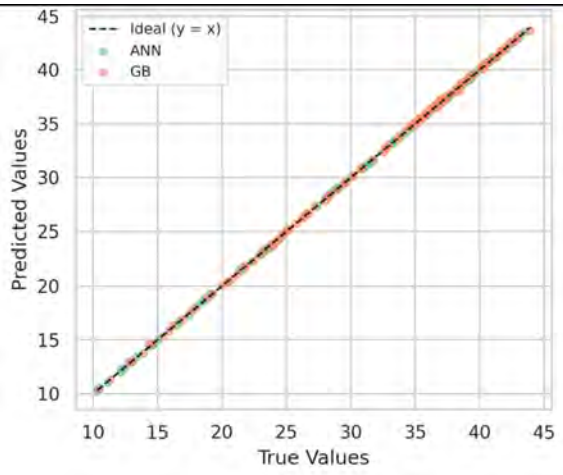


Fig. 18. True vs Predicted values for Spray Angle (Mie) using ANN and Gradient Boosting

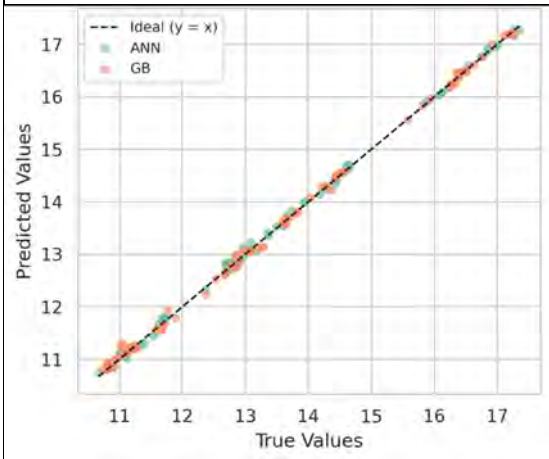


Fig. 19. True vs Predicted values for Spray Length (Mie) using ANN and Gradient Boosting.

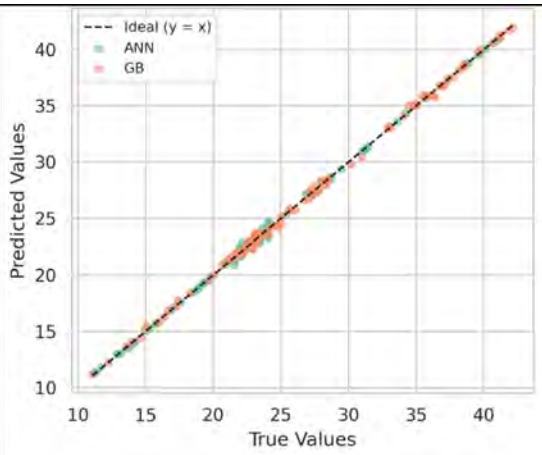


Fig. 20. KDE plots of residuals for all four target variables predicted by the ANN model

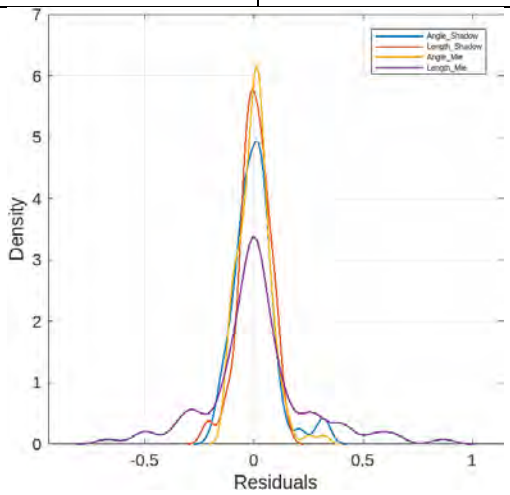


Fig. 22. KDE plots of residuals for all four target variables predicted by the ANN model

4. CONCLUSION [to be enhanced with the outcome and findings from this work with quantified results]

This study evaluated the effectiveness of six baseline machine learning models in predicting spray characteristics namely spray angle and spray length measured through shadowgraph and Mie methods, using a small, time-series experimental dataset. The modeling pipeline included careful data preprocessing, scaling, and an 80/20 holdout evaluation without data shuffling to preserve temporal integrity. Among the tested models, Gradient Boosting and K-Nearest Neighbors demonstrated strong performance, achieving high R^2 scores and low error values across multiple targets. [quantified results in terms of R^2, MSE and MAE are to be provided]

The results affirm that baseline models such as Random Forest, XGBoost, and Linear Regression offer a practical and interpretable approach to spray prediction, especially in resource-constrained or small-data scenarios. Future work may focus on expanding the dataset to enhance model generalizability, incorporating domain-specific features, and applying uncertainty estimation techniques to improve the reliability and transparency of model predictions in real-world applications.

References [not sufficient]

[1] D. Babu, V. Thangarasu, and A. Ramanathan, “Artificial neural network approach on forecasting diesel engine characteristics fuelled with waste frying oil biodiesel,” Applied Energy, vol. 263, p. 114612, Feb. 2020, doi: 10.1016/j.apenergy.2020.114612.

[2] H. Venu et al., “Nanotechnology and LSTM machine learning algorithms in advanced fuel spray dynamics in CI engines with different bowl geometries,” Scientific Reports, vol. 15, no. 1, Jan. 2025, doi: 10.1038/s41598-024-83211-y.

[3] S. Khan, M. Masood, M. Medina, and F. Alzahrani, “Advancing fuel spray characterization: A machine learning approach for directly injected gasoline fuel sprays,” Fuel, vol. 371, p. 131980, Jun. 2024, doi: 10.1016/j.fuel.2024.131980.