Data Science Research Project A in the School of Mathematical Sciences

Ning Ni
a1869549

April 18, 2024

Report submitted for **MATHS 7097A** at the School of Mathematical Sciences, University of Adelaide

THE UNIVERSITY of ADELAIDE

Project Area: **Hybrid Optical/Radio Frequency Communication Channel Model**
Project Supervisor: **Siu Wai Ho**

In submitting this work I am indicating that I have read the University's Academic Integrity Policy. I declare that all material in this assessment is my own work except where there is clear acknowledgement and reference to the work of others.

I give permission for this work to be reproduced and submitted to other academic staff for educational purposes.

OPTIONAL: I give permission this work to be reproduced and provided to future students as an exemplar report.

Abstract

This paper investigates the impact of weather on channel attenuation in hybrid RF/FSO communication systems, with a focus on establishing robust predictive models for attenuation across varying weather conditions. Employing ensemble algorithms, particularly Random Forest, the study delves into the relationship between weather parameters and channel attenuation, identifying key predictors for different weather scenarios. Through extensive data preprocessing, exploratory data analysis (EDA), and model development, both specific and generic models are constructed for attenuation prediction. The research underscores the importance of feature pruning and hyperparameter tuning in optimizing model complexity and performance. Notably, the streamlined generic models exhibit comparable performance to specific models while reducing computational resource and enhancing interpretability. Moreover, the research highlights the critical influence of weather factors on attenuation, revealing distinct sets of predictors for RF and FSO channels across various weather conditions. The generic models demonstrate superior predictive capabilities, particularly excelling in foggy and dusty environments with smaller sample sizes. However, specific models outperform the generic model in clear and snowy conditions for the RF channel. Through comparison with actual values, the generic models demonstrate effective and accurate prediction of channel attenuation in both RF and FSO channels. Overall, these findings offer valuable insights for mitigating weather-induced challenges in hybrid RF/FSO communication systems, thereby fostering the development of more resilient and efficient wireless connectivity solutions in diverse environmental conditions.

1 Introduction

In the realm of wireless communication systems, the integration of Radio Frequency (RF) and Free Space Optical (FSO) technologies has emerged as a promising frontier. This hybrid approach leverages the unique advantages of both RF and FSO channels to enhance data rates and link availability. However, the effectiveness of hybrid RF/FSO systems can be compromised by environmental factors such as rain, fog, or dust storms, which significantly impact channel attenuation and reduce received signal power. Understanding and mitigating these weather-induced effects are crucial for maximizing system efficacy and reliability.

This study aims to comprehensively explore weather-induced channel attenuation in hybrid RF/FSO communication systems and establish reliable channel models to predict RF and FSO channel attenuation. To achieve this, ensemble algorithms such as Random Forest [1], Adaptive Boosting (AdaBoost) [2], and Extreme Gradient Boosting (XGBoost) [3] may be employed. These algorithms will help recognize the pivotal role of weather in system performance and determine which channel is more effective in specific weather conditions.

The outcome of this research will be a suite of weather-aware attenuation models tailored to hybrid RF/FSO communication systems, which contain both special and generic models. Through rigorous evaluation and comparison, these models will illuminate the intricate interplay between weather conditions and channel attenuation, providing reliable predictions for channel attenuation in real-world scenarios. Ultimately, these findings will inform the future development of hybrid RF/FSO satellite communication systems, ushering in a new era of high-throughput wireless connectivity.

2 Background

Hybrid RF/FSO systems merge RF and FSO technologies to enhance performance, reliability, and flexibility in wireless communication networks [4]. Recognizing the unique strengths and limitations of each, RF offers resilience to atmospheric conditions but struggles with high data rates, while FSO provides high data rates but is susceptible to turbulence and obstructions [5] [6].

By combining RF and FSO, hybrid systems leverage the advantages of both. RF acts as a reliable backup when FSO is disrupted by weather or obstacles like fog and snow. Conversely, FSO boosts data rates compared to RF during periods of low turbulence, maximizing network capacity [7].

Implementation involves seamless integration of RF and FSO transceivers with intelligent switching mechanisms. Machine learning algorithms enable autonomous adaptation to environmental changes, optimizing resource allocation and mitigating turbulence impact [8] [9]. These models facilitate collaboration between RF and FSO components, enabling efficient data routing and fault tolerance.

Overall, hybrid RF/FSO systems offer improved reliability, flexibility, and performance in wireless networks. Ongoing research aims to further optimize their design for applications including telecommunications, disaster recovery, and remote sensing.

3 Methods

The data source for this research comprises real empirical data collected from a hybrid system operating in six cities globally.The methodology of this study will be divided into four main parts, including data cleaning, exploratory data analysis, model selection and evaluation metrics, and model model design. The Integrated Development Environment (IDE) utilized for this research is Visual Studio Code (version 1.86.1), and the programming language employed is Python 3.9.

3.1 Data Cleaning

The dataset studied in this research comprises 91,379 samples and 27 variables. "FSO_Att" and "RFL_Att" are the target variables. "SYNOPCode","Time",and "Frequency" are categorical variables. Others are numerical variables.

The dataset is free from any duplicated values and missing entries. This study utilize the Interquartile Range (IQR) method to identify potential outliers in some variables, as depicted in Fig. 1, yet they fall within reasonable ranges. The "RainIntensity" reaches a maximum of 90 mm/h, and the "Particulate" concentration, indicating the presence of particulate matter in the air, can peak at 1600 ug/m^3. Despite these values appearing as outliers, it's possible to occur within extreme natural environments.

3.2 Exploratory Data Analysis (EDA)

3.2.1 Target Variables Analysis

Through the analysis depicted in Fig. 2, it is observed that "FSO_Att" exhibits a bimodal right-skewed distribution within the range of 0.8 dB to 16.5 dB. On the other hand, "RFL_Att" displays a unimodal right-skewed distribution spanning from 7.8 dB to 15.9 dB.

3.2.2 Predictor Variables Analysis

The "AbsoluteHumidity" ranges from 1 g/m^3 to 25 g/m^3, with a median around 7 g/m^3. The "Particulate" is mostly distributed around 0 ug/m^3, indicating low particle content in the air and relatively clear visibility. In rare cases, it can exceed 200 ug/m^3, especially during dust storms, significantly impacting visibility. "RainIntensity" reflects mostly clear days, with occasional instances of heavy rainfall reaching up to 90 mm/h, as Fig. 3.

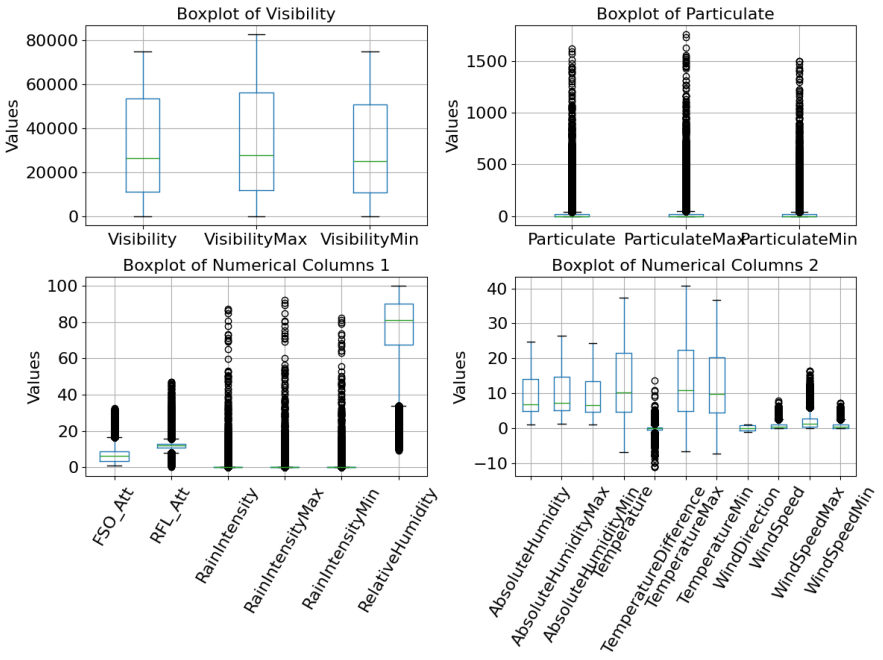

Figure 1: Outliers Analysis

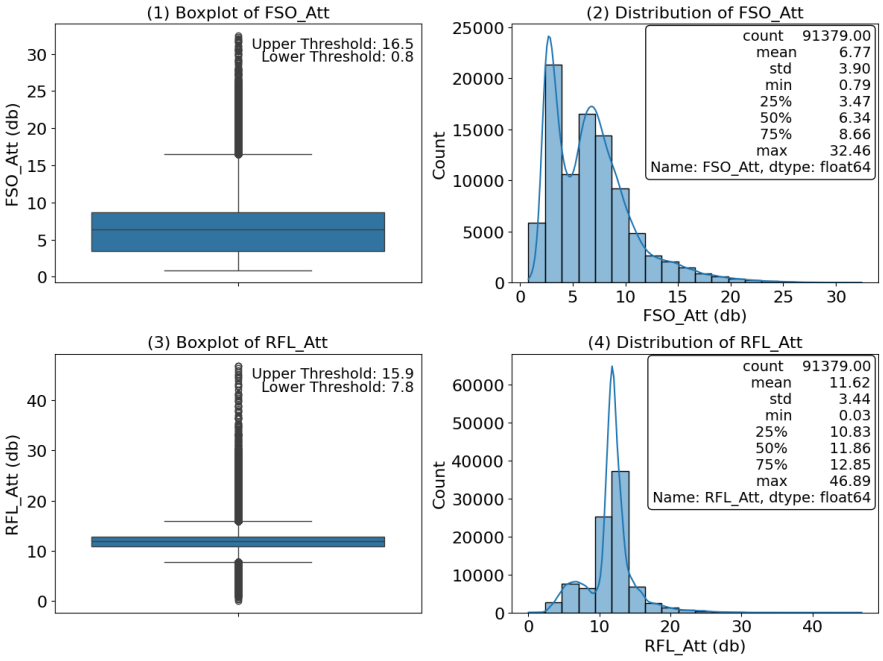

Figure 2: Target Variables Analysis

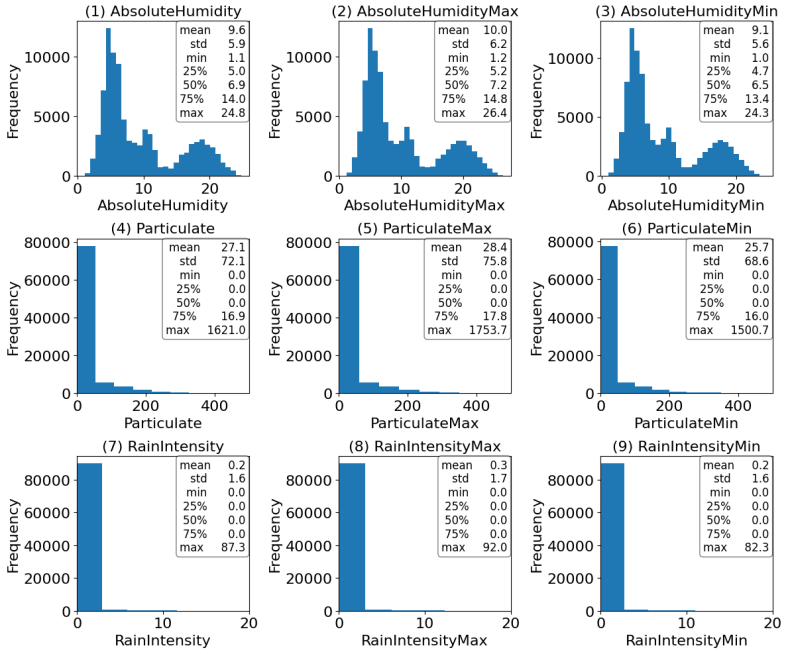

Figure 3: The Histogram of "Particulate","AbsoluteHumidity" and "RainIntensity"

Fig. 4 shows that "Temperature" ranges from 1 degree Celsius to 40 degrees Celsius, with a median around 10.3 degrees Celsius, exhibiting a distribution pattern similar to "AbsoluteHumidity". More than three-quarters of "Visibility" exceed 10,000 meters. "WindSpeed" is concentrated between 0 and 5 m/s, with occasional maximum speeds reaching 16 m/s.

Fig. 5 indicates that "Distance" is mainly distributed at 2100 meters, 2950 meters, 3950 meters, and 4800 meters. "WindDirection" indicates predominantly northerly and northwesterly winds, while other wind directions are distributed fairly evenly. "Frequency" shows that only microwave frequencies of 73.5 GHz and 83.5 GHz are present in the RF channel. "SYNOPCode" represents the overall weather conditions of the day, where 0 indicates clear skies, accounting for the majority of cases. Following this are 5 representing drizzle and 6 for rain. Other conditions include 3 for dust, 4 for fog, 7 for snow, and 8 for showers, but these are rare occurrences. The data spans from 0 to 24 hours in time intervals of 1 hour, evenly distributed.

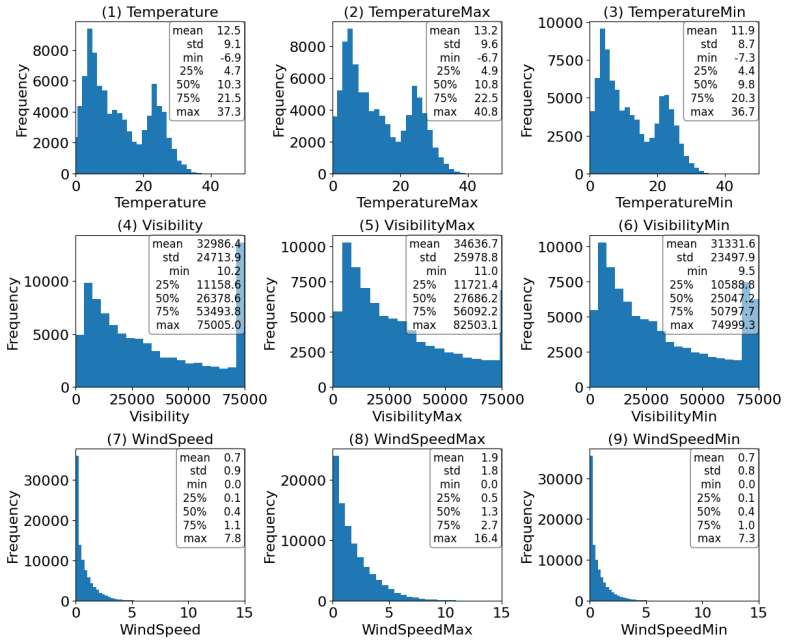

Figure 4: The Histogram of "Temperature","Visibility" and "WindSpeed"

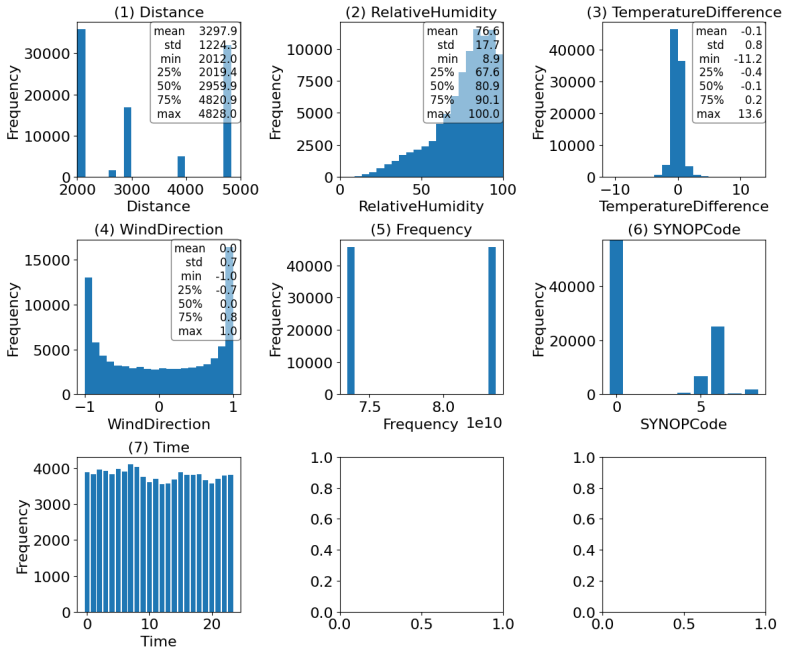

Figure 5: The Histogram of Other Features

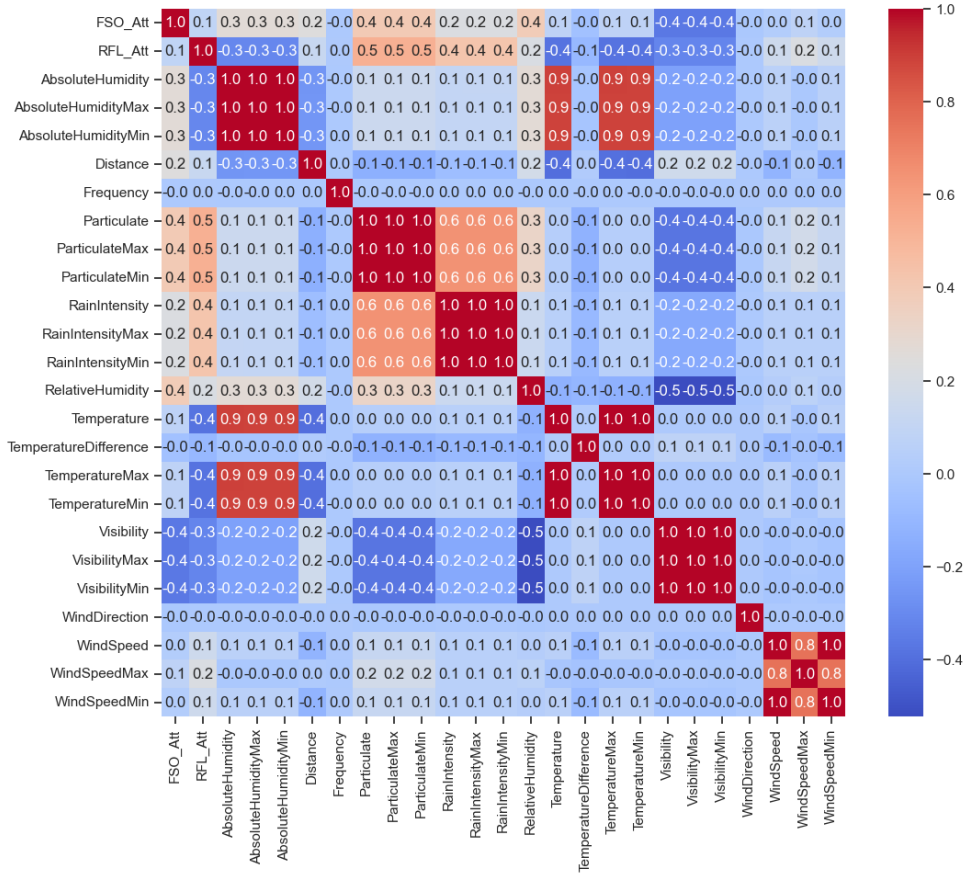

Figure 6: Correlation Heatmap of Target vs. Numerical Variables

3.2.3 The correlation analysis between Target and Predictor Variables

Combining Fig. 6 and Fig. 7, there is a weak linear correlation between FSO_Att and "Particulate" and "RelativeHumidity", while RFL_Att shows weak linear correlations with "Particulate", "RainIntensity", and "Temperature". The attenuation of both channels does not exhibit linear correlations with other numerical variables. "Temperature" has a strong linear correlation with "AbsoluteHumidity", and "Particulate" shows a clear linear correlation with "RainIntensity". Some features also exhibit strong linear correlations with their corresponding maximum and minimum values.

Through Fig. 8, it reveals the characteristics of the FSO channel and RF channel. In foggy (SYNOPCode 4), dusty (SYNOPCode 3), and snowy (SYNOPCode 7) weather conditions, the optical channel attenuation is greater than in other weather environments, while the RF channel remains relatively stable in these environments, less susceptible to inter-

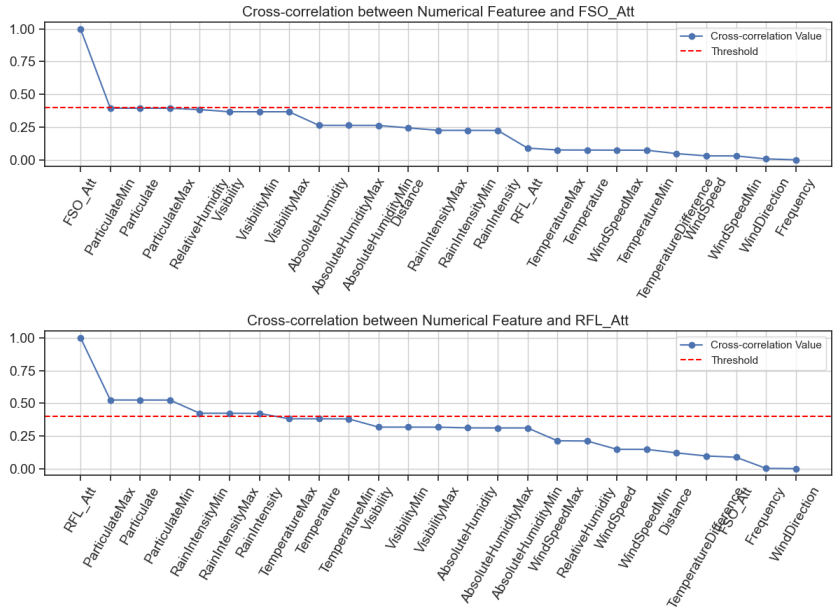

Figure 7: Cross-correlation between Numerical Feature and Attenuation

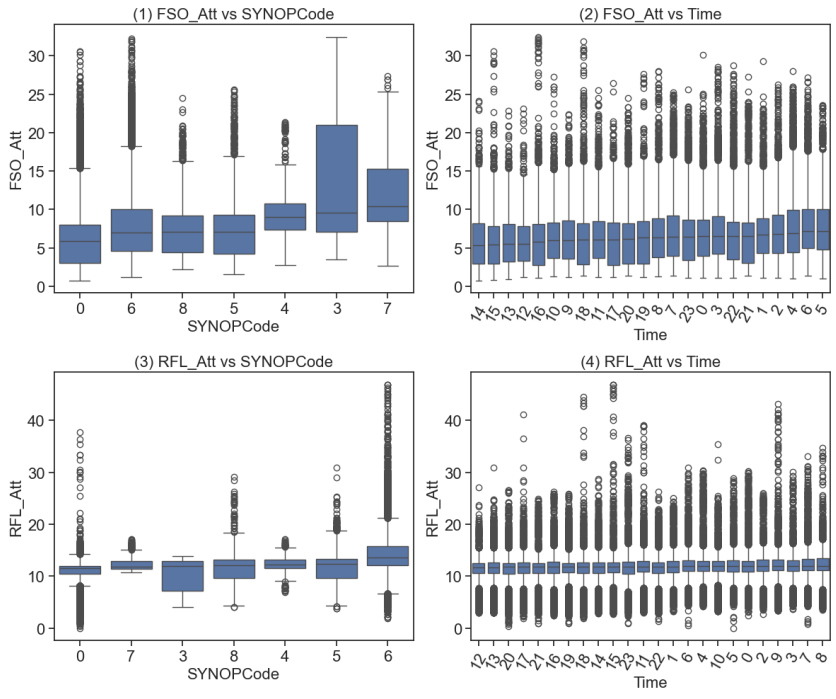

Figure 8: Target vs. Categorical Variables

ference. The FSO channel appears to have lower attenuation during the day compared to night, whereas the RF channel exhibits similar attenuation levels across different time periods without significant differences.

3.3 Model Selection and Evaluation Metrics

3.3.1 The Basic Architecture of Random Forest

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and then combines their outputs to make predictions. Each decision tree in the Random Forest is trained independently, typically using a random subset of the training data and a random subset of the features. This randomness helps to reduce overfitting and enhance robustness. During prediction, the results of all individual trees are aggregated through voting (for classification tasks) or averaging (for regression tasks) to produce the final prediction. This ensemble approach often results in a more accurate and robust model compared to a single decision tree, as showing Fig. 9 [10].

In Fig. 9, The training dataset, consisting of 250 rows and 100 columns, is randomly sampled with replacement n times. Subsequently, a decision tree is trained on each sample. Finally, during the prediction phase, the outcomes of all n trees are aggregated to yield a final decision.

3.3.2 Bagging

Bagging, or Bootstrap Aggregating, is a fundamental sampling technique in Random Forests. It operates by training each decision tree on a randomly sampled subset of the training data, with replacement, resulting in multiple bootstrap samples, as Fig. 10. In an ideal case, about 36.8 % of the total training data forms the "out-of-bag (OOB)" sample and 63.2% of that contributes each bootstrap sample. This can be shown as Eq.(1),where N is the total number of samples.

Each tree is then independently trained on one of these samples, imparting diversity as they learn different aspects of the data. During prediction, the results from all trees are aggregated, often through majority voting for classification or averaging for regression tasks. This aggregation reduces variance and mitigates overfitting, making Random Forests effective at generalizing to unseen data.

It's worth nothing that the concept of OOB is specific to each individual tree. Although a sample may be considered as "out-of-bag data" for one tree, it could also be part of the training set for another tree. As the number of trees increases, there is no concept of OOB data for

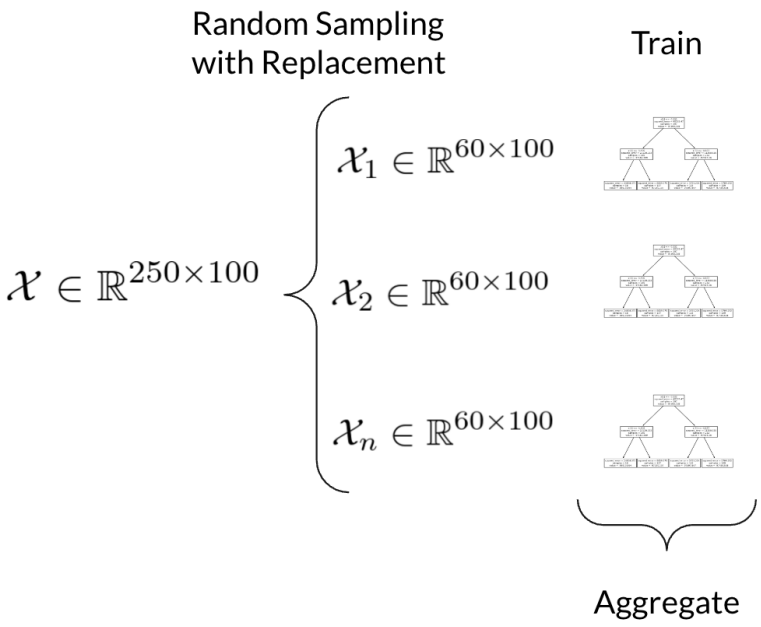

Figure 9: Basic Construction of Random Forest [10]

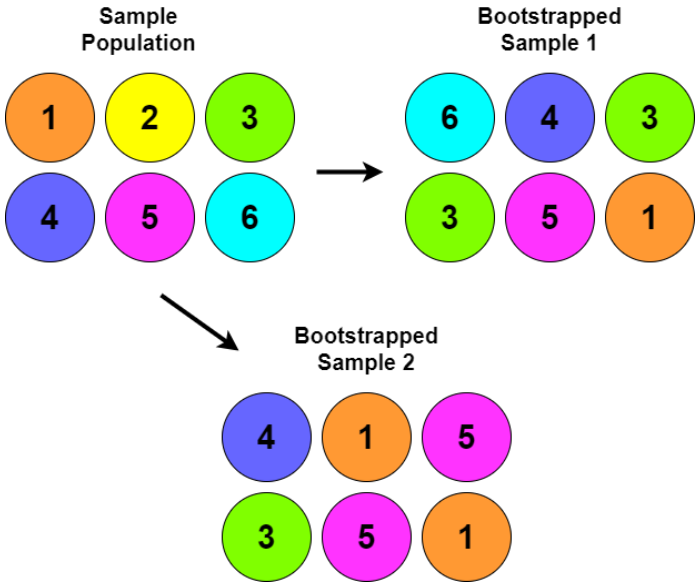

Figure 10: Bagging

the entire random forest because the entire forest is trained on the entire dataset.

$$\lim_{N\to\infty}\left(1-\frac{1}{N}\right)^{N}=e^{-1}=0.368 \tag{1}$$

3.3.3 Branching Principles

Decision trees measure impurity using either the Gini coefficient or information entropy for classification tasks, and Mean Squared Error (MSE) for regression tasks. At each split, the tree evaluates impurity for all features, selecting the feature that minimizes impurity for branching. Subsequently, at each child node, impurity is recalculated for each feature, and the process iterates, choosing the feature that minimizes impurity. With each branching layer, the overall impurity decreases, as the tree seeks to minimize impurity. The decision tree continues branching until no more features are available or the impurity metric is optimized.

However, decision trees are prone to overfitting. To mitigate this, random forests offer an effective solution. Firstly, they utilize bagging to ensure that each decision tree in the forest operates on a different subset of samples. Additionally, at each node, random forests randomly select a subset of features. It's important to note that earlier random decision forests employed the "random subspace method" [11], where each tree received a random subset of features. However, the current approach involves selecting different subsets of features for each node, while providing each tree with the complete set of features [12]. In summary, while each tree in a random forest receives the full set of features, only a random subset of features is considered at each node.

3.3.4 Evaluation Metrics

In this study, Root Mean Squared Error (RMSE) and R-square (R^2) are employed to evaluate models' performances, as Eq.(2) and Eq.(3), where $\bar{y}$ is the numerical average of y_n over all N. RMSE provides a measure of the model's error in the same units as the target variable, while R^2 evaluates the proportion of variance in the target variable explained by the model [13].

$$RMSE=\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(y_n-\hat{y}_n\right)^2} \tag{2}$$

$$R^2=1-\frac{\sum_{n=1}^{N}\left(y_n-\hat{y}_n\right)^2}{\sum_{n=1}^{N}\left(y_n-\bar{y}\right)^2} \tag{3}$$

Table 1: Data Splitting

Data set	sample number	ratio
training	63965	70%
validation	13707	15%
test	13707	15%

3.4 Model Design

3.4.1 Data Splitting

In this study, the dataset is divided into training, validation and test set, as Tab. 1.

The training set is used to build the random forest model, with each decision tree randomly selecting samples from it for training, ensuring that each tree is constructed based on a different subset of the data.

The validation set is employed to tune the model's hyperparameters, evaluating the performance of different parameter settings and selecting the best configuration to enhance the model's generalization ability.

Finally, the test set is utilized to assess the model's performance, validating its predictive capability on unseen data and providing performance metrics for real-world applications in predicting attenuation, thus evaluating the model's reliability and utility.

3.4.2 Decision Tree Model

Random Forest is a type of ensemble learning method, which combines multiple weak learners to form a strong learner. Decision tree serves as the fundamental weak learner in Random Forest. Each decision tree is trained on a random subset of the training data, and then integrated by methods such as voting or averaging to produce the final prediction.

The significance of exploring weak learners lies in two aspects:

For Random Forest to yield reliable results, individual decision trees must demonstrate certain predictive performance, achieving high accuracy on both the training and testing sets (for classification problems) or low error rates (for regression problems). Otherwise, even with multiple weak learners integrated, Random Forest cannot provide reliable predictions.

Decision trees in Random Forest should have moderate complexity, neither too simple (high bias) nor too complex (high variance). Decision trees that are too simple may lead to underfitting, while overly complex decision trees may result in overfitting. By investigating the performance of individual decision trees, suitable parameter ranges can be obtained,

Table 2: Decision Tree Model Hyperparameters

Hyperparameter	Range
splitter	"best" or "random"
criterion	"gini" or "entropy"
max_depth	*range(10,41,2)
min_samples_leaf	[1,10,50,100]
min_samples_split	[2,11,51,101]

Table 3: Decision Tree Model Metrics

Data set	RMSE	R^2
training_FSO	1.24	90%
validation_FSO	1.19	90.5%
training_RFL	0.77	95%
validation_RFL	0.73	95.5%

such as the maximum depth of the decision tree (max_depth), the minimum number of samples required to be at a leaf node (min_samples_leaf), and the minimum number of samples required to split an internal node (min_samples_split). This provides an approximate parameter range for Random Forest hyperparameter tuning, reducing the learning time for hyperparameter adjustment.

Based on the value ranges provided in Tab. 2 for training the model and tuning hyperparameters, Fig. 11 and Fig. 12 demonstrate that as the max_depth increases, the model's RMSE decreases and R^2 increases. Once max_depth reaches approximately 20, the model's performance stabilizes. Additionally, the model achieves its best performance when min_samples_leaf is set to 10, and further increasing this value does not yield improvements in model performance. In the experiment, parameters such as splitter, criterion, and min_samples_split have minimal influence on model performance. Consequently, the optimized parameters are max_depth set to 22 and min_samples_leaf set to 10. This parameter combination can serve as a reference for training subsequent Random Forest models.

The specific results are recorded in Tab. 3, indicating that the Decision Tree with the optimized parameters performs well in predicting both "FSO_Att" and "RFL_Att", which is validated by tests on the validation set, showing the model's robust generalization capabilities.

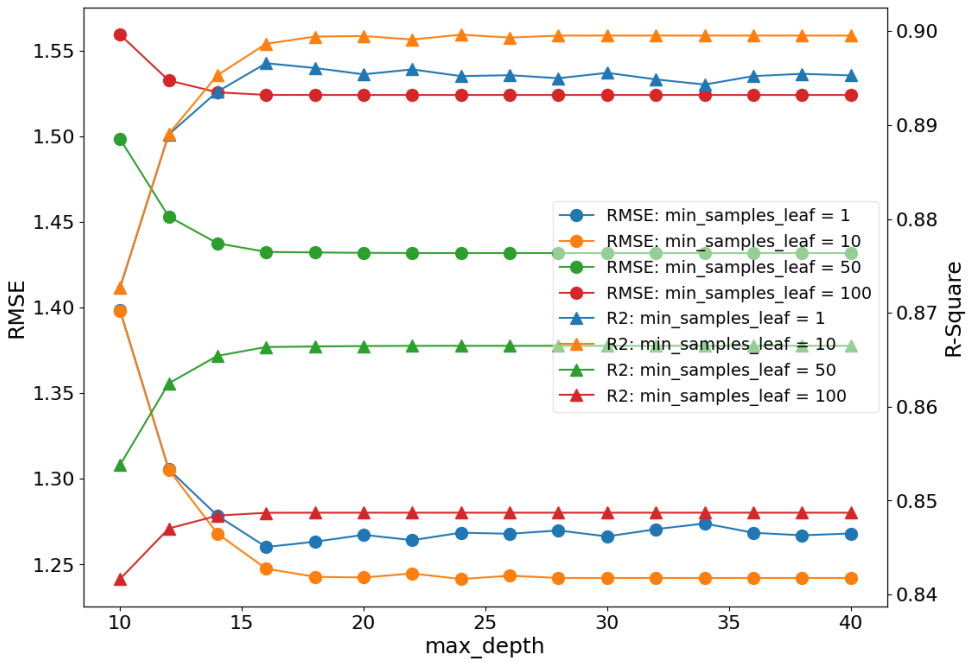

Figure 11: Learning Curve of Decision Tree Regression Model on FSO Training Set

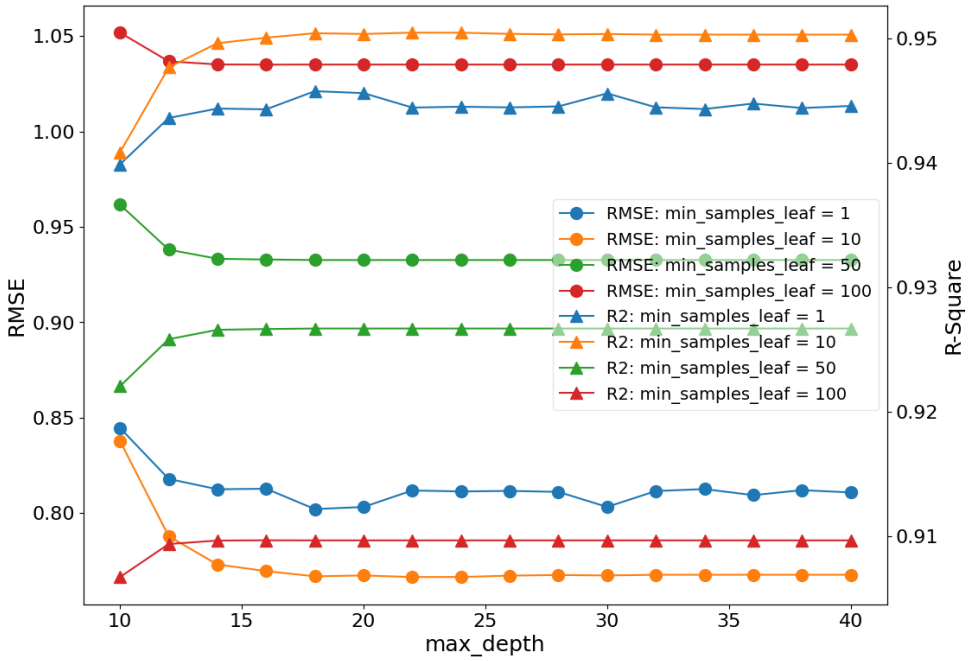

Figure 12: Learning Curve of Decision Tree Regression Model on RFL Training Set

Table 4: Hyperparameters Tuning of RF Model for FSO and RFL

Hyperparameter	Coarse Tune	Fine Tune(FSO)	Fine Tune(RFL)
n_estimators	range(10,301,20)	[120,130,140]	range(100,141,10)
max_depth	range(5,31,5)	range(27,33,1)	range(23,28,1)
min_samples_leaf	[4,5,6,7]	[1,2,3,4]	[1,2,3,4]
min_samples_split	[5,6,7,8]	[2,3,4,5]	[2,3,4,5]

3.5 Random Forest Model Establishment

During the coarse tuning of hyperparameters for FSO prediction, Fig. 13 and Fig. 14 demonstrate a clear trend that as the value of n_estimators increases, the RMSE decreases and R^2 increases slightly across both the training and validation sets. However, beyond a certain point, specifically when n_estimators reaches 130, further increments fail to enhance the model's performance in either dataset.

Regarding max_depth, optimal performance is achieved when it is set to 30. However, the improvement in model performance when max_depth is 30 compared to when it's 20 is marginal. Based on these observations, the fine tuning ranges will be outlined in Tab. 4.

In the fine-tuning process, as illustrated in Fig. 15, it is evident that the model's performance improves in the training set as the number of n_estimators and max_depth increase. However, in the validation set, the optimal configuration occurs when n_estimators is set to 130 and max_depth is 30. Beyond this point, further increases lead to a decline in model performance, indicating overfitting. Additionally, it is observed in Fig. 16 that reducing min_samples_leaf continues to enhance model performance until this parameter reaches a value of 1.

In the coarse tuning phase of the RFL model, as Fig. 17 and Fig. 18 it was determined that the optimal hyperparameter configuration occurred when n_estimators was set to 100, max_depth to 25, and min_samples_leaf to 4, resulting in the best model performance.

During the fine-tuning process of the RFL model, depicted in Fig. 19, it became apparent that the model's performance in the training set improved with increasing values of n_estimators and max_depth. Yet, in the validation set, the optimal parameters were identified as n_estimators being 110 and max_depth being 28. Further escalation of these parameters led to a decline in performance, signaling overfitting. Additionally, Fig. 20 illustrated that reducing min_samples_leaf improved model performance, with the optimal value transitioning from 4 to 1.

Finally, Tab. 5 presents the optimal hyperparameters for both models, while Tab. 6 displays the performance of both models on the training, validation, and test datasets.

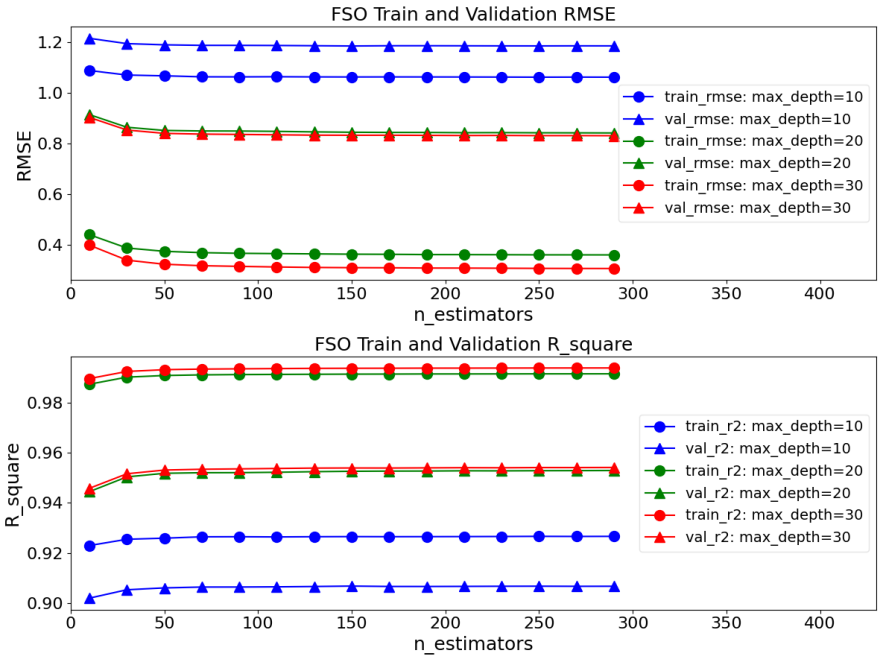

Figure 13: FSO Coarse Learning Curve of Random Forest Regression Model with Respect to n_estimators and max_depth

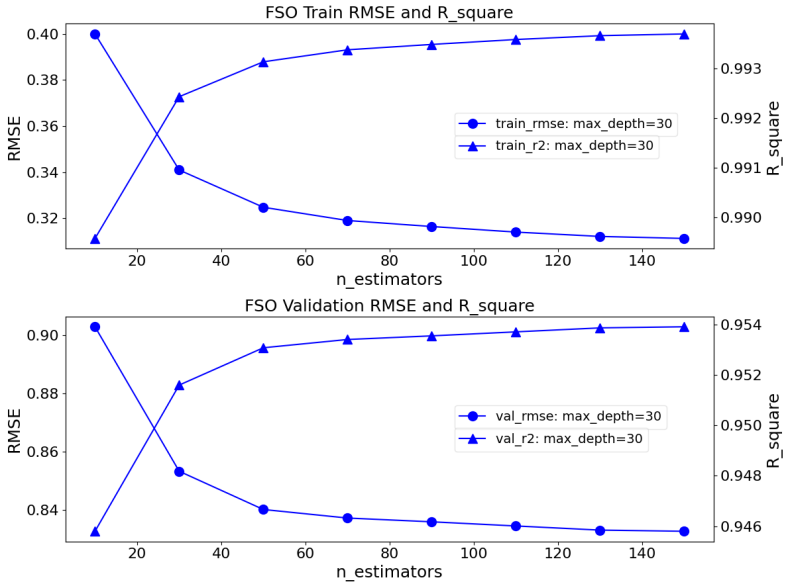

Figure 14: FSO Coarse Learning Curve of Random Forest Regression Model with Respect to n_estimators and max_depth = 30

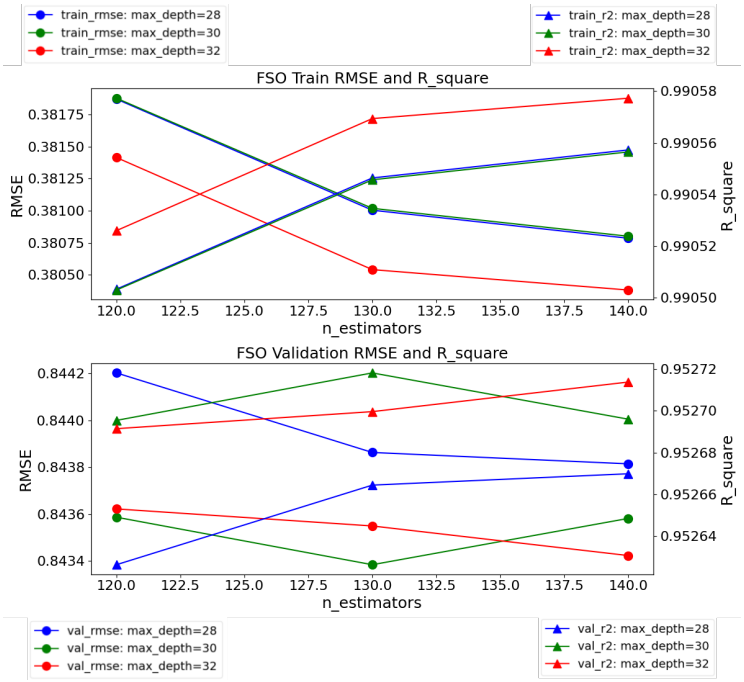

Figure 15: FSO Fine Learning Curve of Random Forest Regression Model with Respect to n_estimators and max_depth

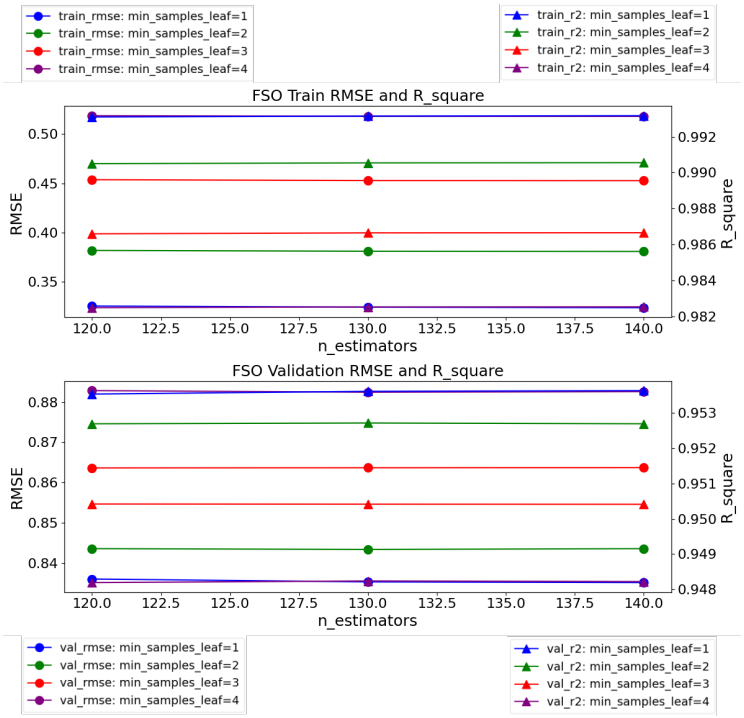

Figure 16: FSO Fine Learning Curve of Random Forest Regression Model with Respect to n_estimators and min_sample_leaf

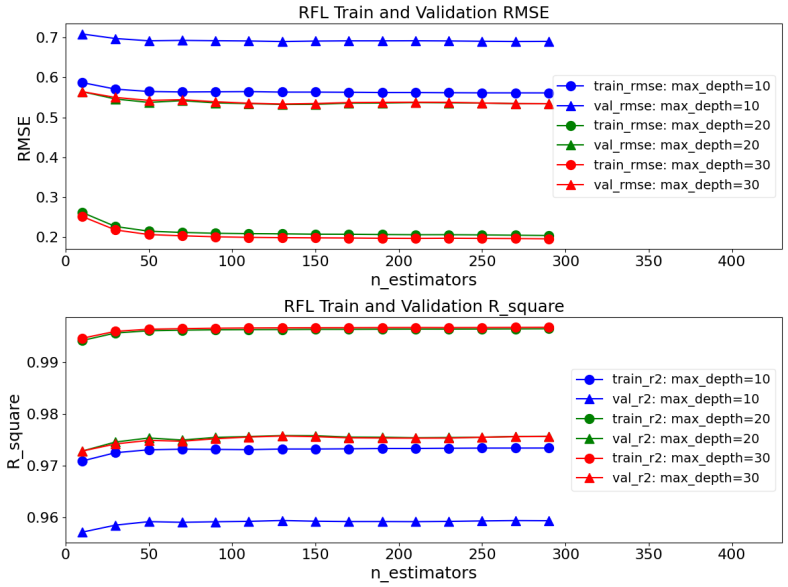

Figure 17: RFL Coarse Learning Curve of Random Forest Regression Model with Respect to n_estimators and max_depth

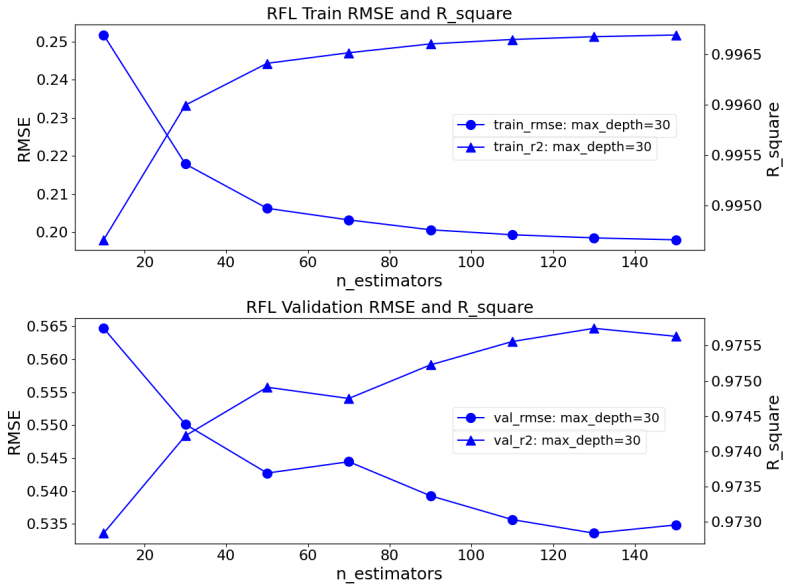

Figure 18: RFL Coarse Learning Curve of Random Forest Regression Model with Respect to n_estimators and max_depth = 30

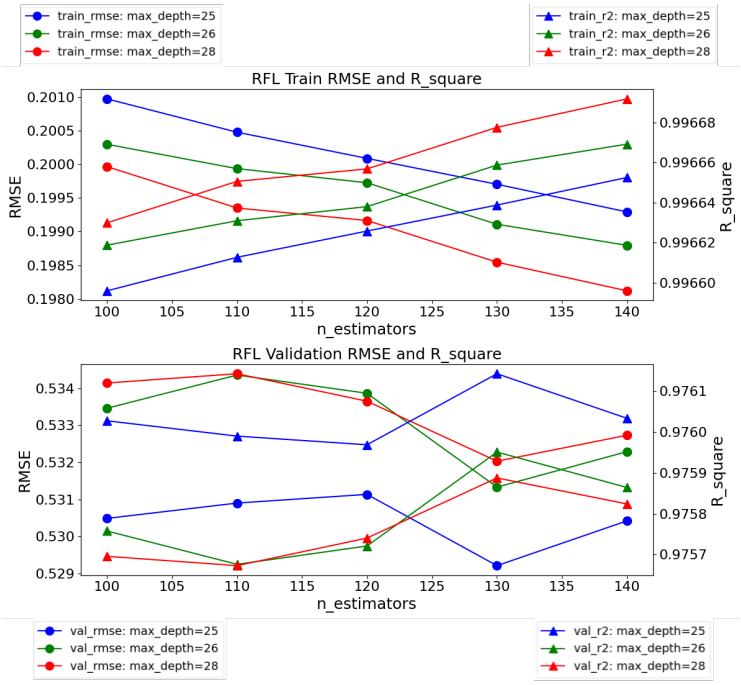

Figure 19: RFL Fine Learning Curve of Random Forest Regression Model with Respect to n_estimators and max_depth

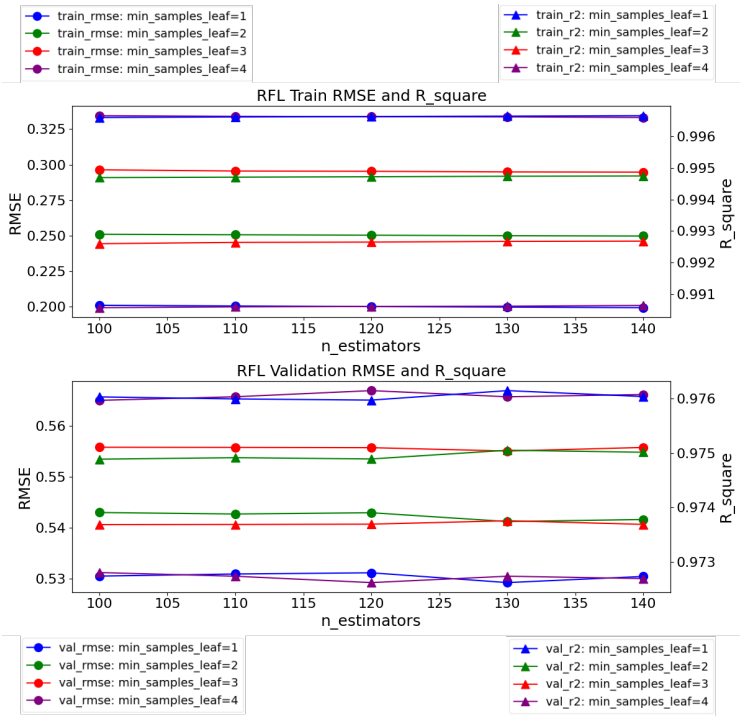

Figure 20: RFL Fine Learning Curve of Random Forest Regression Model with Respect to n_estimators and min_samples_leaf

Table 5: Optimal Hyperparameters for Models

Hyperparameter	FSO Model	RFL Model
n_estimators	130	110
max_depth	30	28
min_samples_leaf	1	1
min_samples_split	2	2

Table 6: Metrics of Optimal Models

Metrics	FSO Model	RFL Model
training RMSE	0.31	0.20
validation RMSE	0.83	0.54
testing RMSE	0.78	0.50
training R^2	99.4%	99.7%
validation R^2	95.4%	97.6%
testing R^2	95.9%	97.9%

3.5.1 Features Importance and Feature Pruning

As discussed in Section 3.3.3, during Random Forest training, values are collected to measure how, on average, the node split decreases the impurity or MSE. The average over all trees in the forest yields the measure of feature importance [14]. This method is implemented in scikit-learn's Random Forest [15].The computed importances is the relative values since they are normalized. One significant advantage of this method is its computational speed—all necessary values are computed during Random Forest training. However, in the case of correlated features, it may select one feature over another, potentially leading to incorrect conclusions.

Another method is Permutation Importance [16] that is to conduct a "destructive test" on each feature in the model by randomly altering the order of its values, and then observing the change in model performance to evaluate the contribution of the feature to the model predictions. First, the trained model is used to predict the test data, and the prediction results are recorded as a baseline. Next, for each feature to be evaluated, the order of its values is randomly shuffled, and the model's predictions on the shuffled data are recalculated. The change in model performance on the shuffled data is then computed. The importance of a feature is defined as the degree of change in model performance. If the model performance decreases significantly after shuffling the feature values, it indicates that the feature has a significant impact on the model's performance and thus has high importance; conversely, if the performance change is small, it suggests that the feature has low importance. Permutation Importance

provides intuitive interpretability and can offer useful information even in the presence of feature correlations or nonlinear relationships.

The above mentioned methods provide a sorted order of features based on their importance scores. Wrapper methods [17], like Recursive Feature Elimination (RFE), go beyond this by selecting features based on actual model performance. RFE evaluates subsets of features using Random Forests, by iteratively removing features and assessing their impact on model performance. This process involves training the model on the entire feature set and iteratively eliminating the least important features based on their derived feature importances. This iterative procedure continues until the desired number of features is achieved. By directly evaluating their influence on model performance, wrapper methods offer a more comprehensive understanding of feature importance.

Embedded methods [18], on the other hand, incorporate feature selection as an integral part of the model training process. In Random Forests, feature importance is inherently embedded within the model construction, as the splitting criteria at each node are based on feature importance measures.

Embedded methods can select or delete groups of features based on certain thresholds of feature importance scores. Using embedded methods enables the rapid identification of features with significant predictive power. Subsequently, the selected features can be further refined using wrapper methods like RFE. This combined approach maximizes the efficiency of embedded methods and the precision of wrapper methods, allowing for more effective feature selection and modeling, particularly in high-dimensional datasets.

Given that this research dataset comprises 25 features, employing wrapper methods alone should suffice for eliminating redundant features.

Fig. 21 and Fig. 22 respectively illustrate the features importance of FSO and RFL channels. This study employs the wrapper method to prune features.

The process of feature pruning is iteratively removing features based on their feature importance scores, starting from the least important to the most important, as depicted in Fig. 21 and Fig. 22. Subsequently, the performance of the models is observed. If the models' performance does not exhibit significant changes, pruning occurs. During this process, consideration is given to the interactions between features and their impact on model performance.

The objective is to identify the most relevant subset of features capable of effectively predicting FSO and RFL attenuation while simultaneously minimizing overfitting and computational overhead.

Fig. 23 and Fig. 24 demonstrate the impact of feature pruning on both

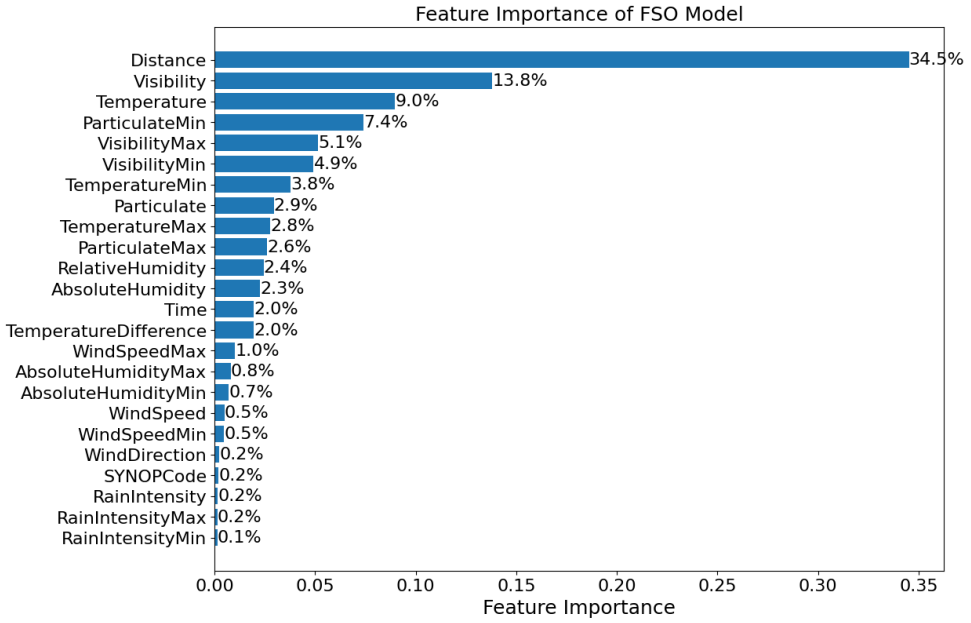

Figure 21: Feature Importance of FSO Model

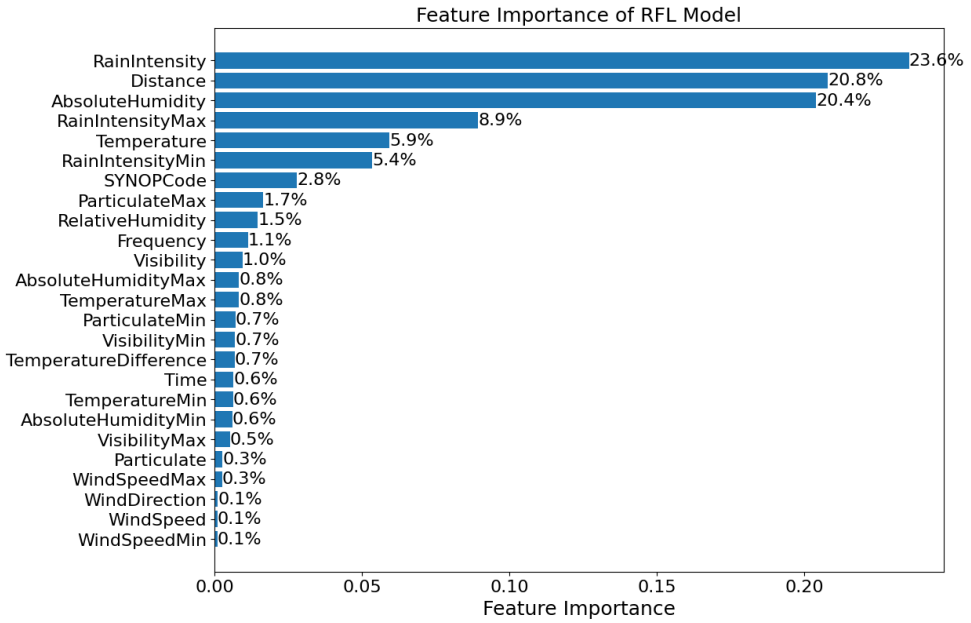

Figure 22: Feature Importance of RFL Model

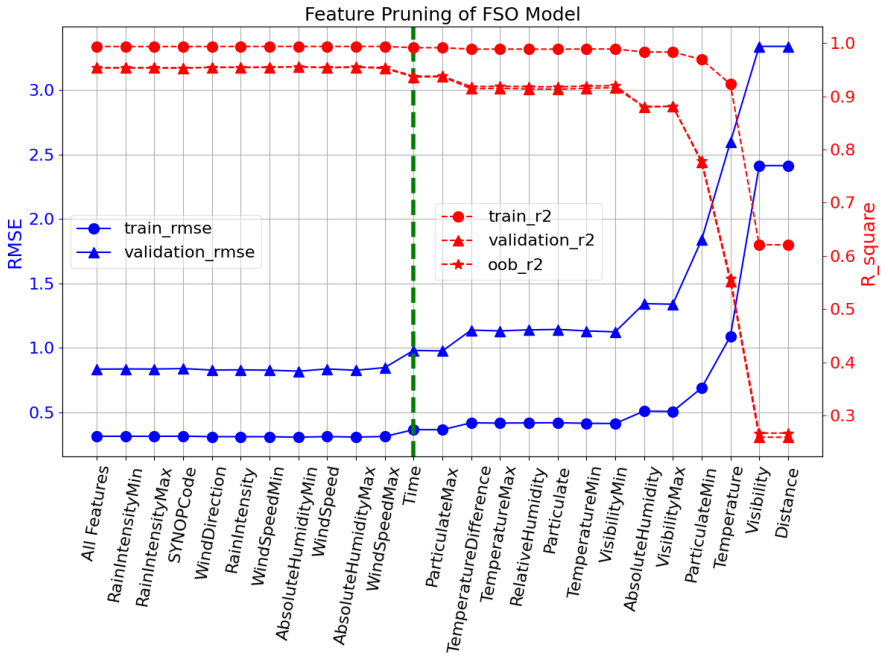

Figure 23: Feature Importance of FSO Model

models. Without compromising predictive performance, the complexity of features is significantly reduced, from 27 features to approximately 14 for the FSO Model or 11 features for the RFL Model, as indicated by the green line in figures.

Utilizing the pruned features, further hyperparameter tuning is conducted for both models. The coarse tuning and fine tuning processes remain consistent with those outlined in Section 3.5.

Following the descriptions provided in Fig. 25 and Fig. 26, the hyperparameters are determined considering the trade-off between model complexity and performance. The results are presented in Tab. 7. Both of n_estimators and max_depth in optimal hyperparameter with feature pruning are a little larger than that of optimal hyperparameter without feature pruning in Tab. 5.

Tab. 8 displays the evaluation metrics for the two models with feature pruning. Comparing with Tab. 6, the performances of two models with feature pruning are almost some as models without feature pruning.

Through feature pruning and hyperparameter tuning, the overall complexity of the two random forest models are significantly reduced. This helps improve the interpretability and generalization ability of the models while reducing the risk of overfitting.

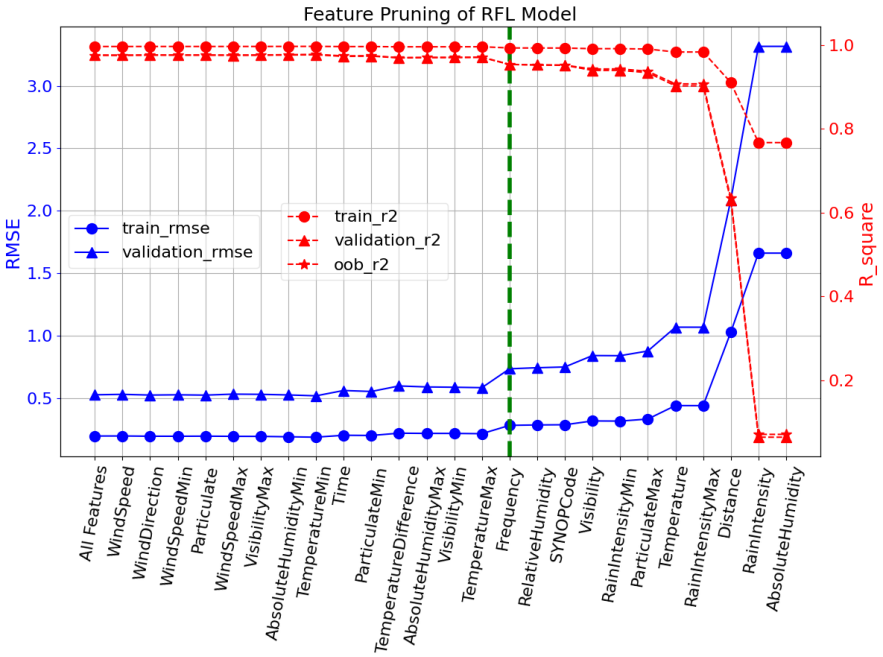

Figure 24: Feature Importance of RFL Model

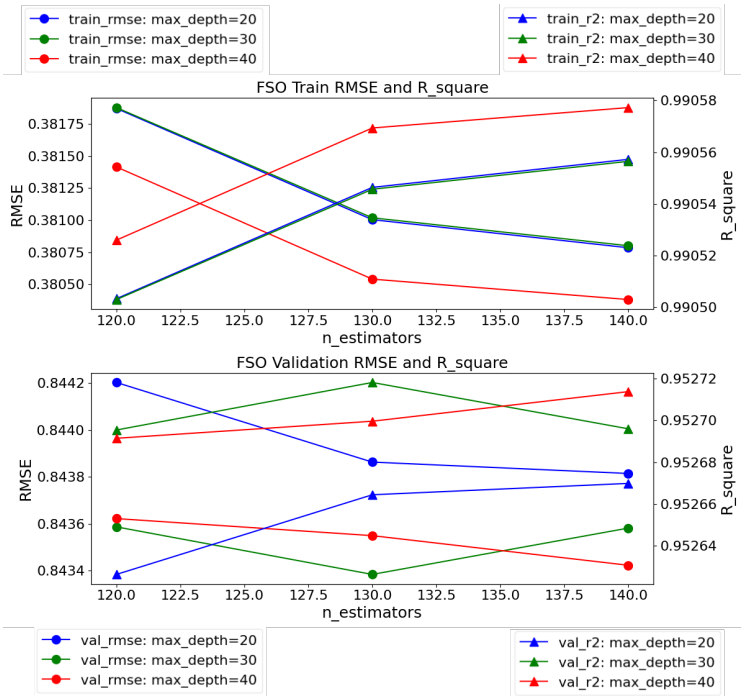

Figure 25: Hyperparameters Tuning for FSO Model with Features after Pruning

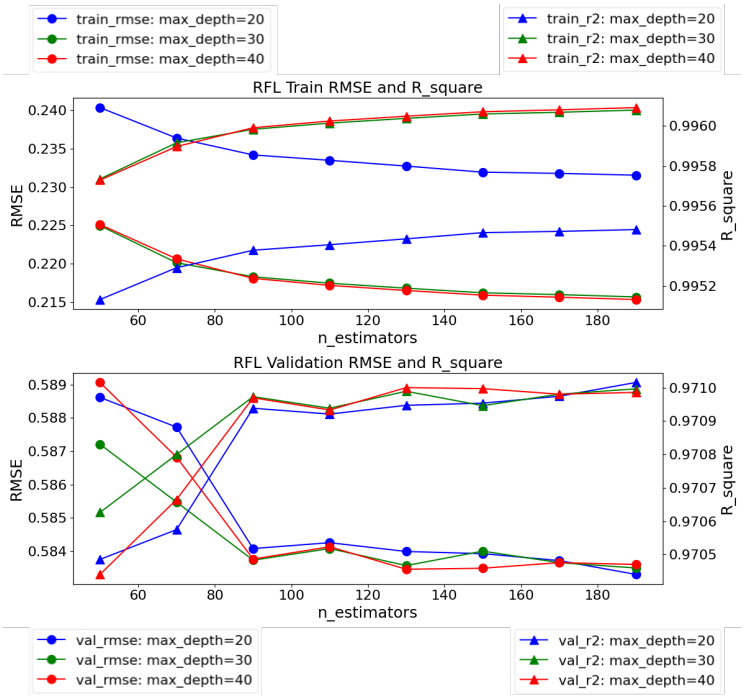

Figure 26: Hyperparameters Tuning for RFL Model with Features after Pruning

Table 7: Optimal Hyperparameters for Models with Feature Pruning

Hyperparameter	FSO Model	RFL Model
n_estimators	150	120
max_depth	34	32
min_samples_leaf	1	1
min_samples_split	2	2

Table 8: Metrics of Optimal Models with Feature Pruning

Metrics	FSO Model	RFL Model
training RMSE	0.31	0.22
validation RMSE	0.84	0.58
testing RMSE	0.78	0.54
training R^2	99.3%	99.6%
validation R^2	95.26%	97.1%
testing R^2	95.9%	97.5%

4 Results

4.1 The Important Predictors in Special Models and Generic Model

In the original dataset, the data will be divided into subsets based on the values of the feature "SYNOPCode", representing different weather conditions. Each subset of data corresponding to specific weather conditions will be used to build a random forest model separately. This process will establish seven distinct models to predict the impact of weather on signal attenuation for specific weather scenarios. The "SYNOPCode" is shown as Tab. 9.

Table 9: The Value of SYNOPCode Corresponding to Weather

SYNOPCode	0	3	4	5	6	7	8
Weather	Clear	Dust	Fog	Drizzle	rain	snow	showers

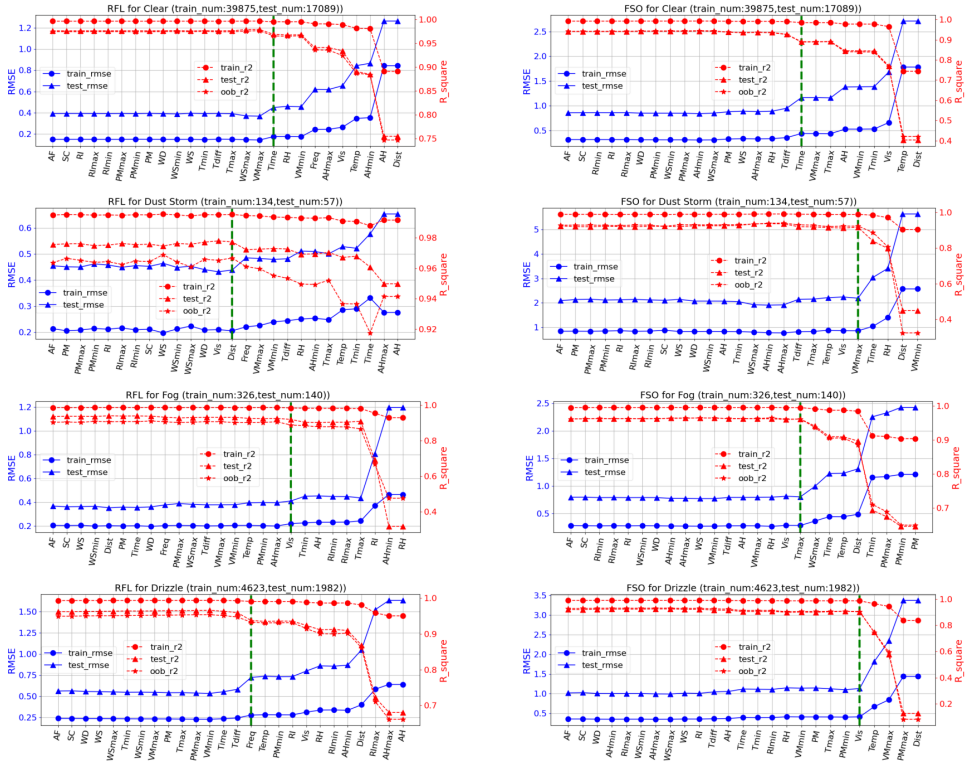

Figure 27: Predictor Importance Graph for Clear, Dust, Fog, Drizzle

Fig. 27 and Fig. 28 depict the performance of each specific model based on different weather conditions. Each specific model has its own

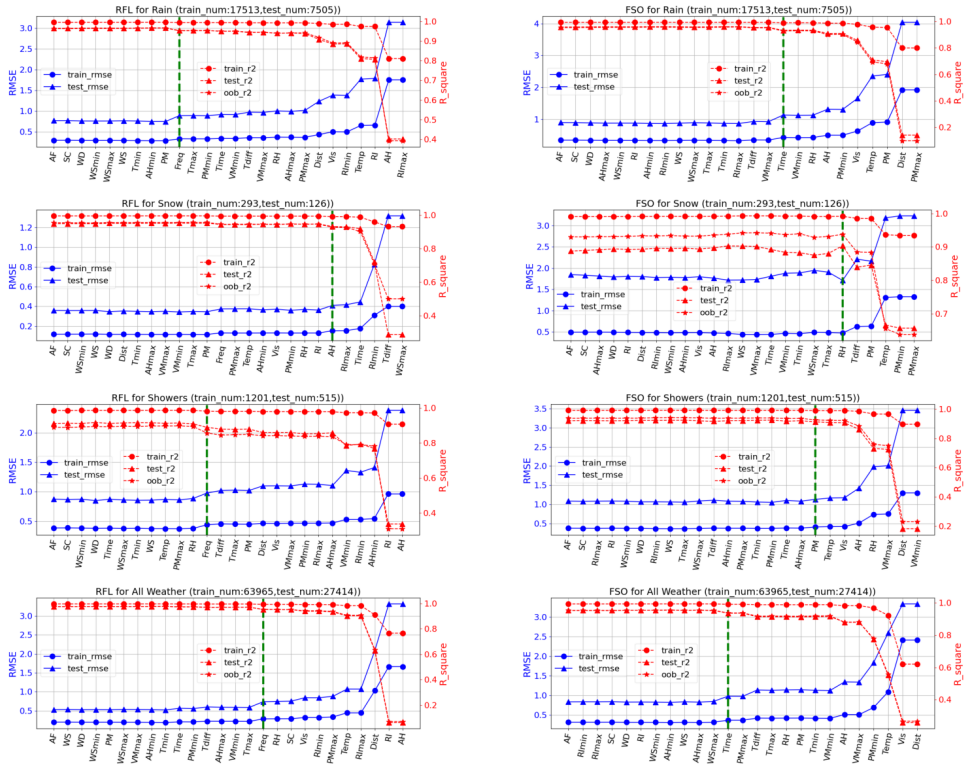

Figure 28: Predictor Importance Graph for Rain, Snow, Showers and All Weather

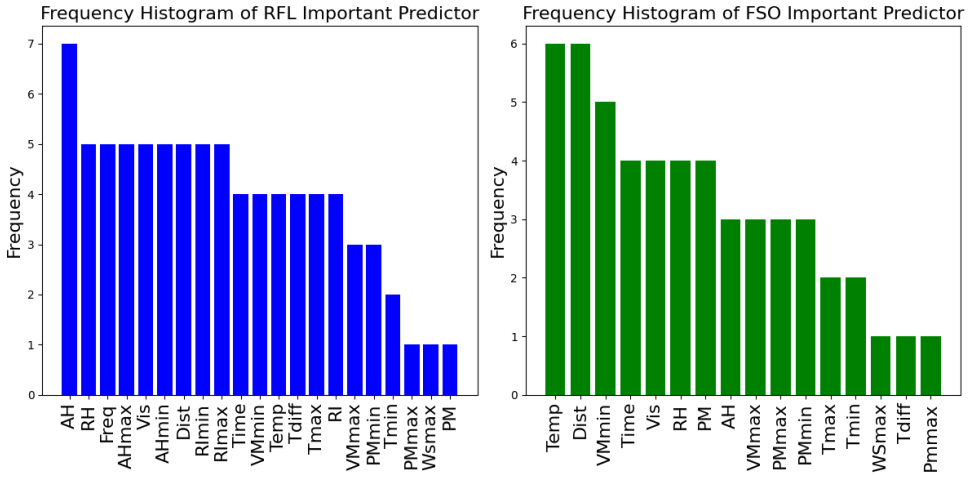

Figure 29: Important Predictor Bar Chart of All Specific Models

set of important predictors, indicating that certain features are more effective in predicting attenuation under specific weather conditions.

Specifically, when the weather is rainy or showers, the RFL model requires additional features to maintain prediction performance at a stable level. The RMSE in rainy or shower conditions is higher compared to other weather conditions, indicating a significant impact of rainy weather on the RFL channel. In rainy weather, the RMSE of FSO channel attenuation resembles that of RFL channels. However, the attenuation in the FSO channel is typically higher than in other weather conditions, particularly in dusty weather, where the FSO model performs poorly.

Fig. 29 illustrates the frequency of important predictors across all specific models. In the RFL channel, Absolute Humidity is utilized in all the special models and features related to Rain Intensity, Relative Humidity, Frequency, Distance, and Visibility are commonly utilized for prediction in various weather conditions, while in the FSO channel, features related to Distance, Temperature, and Visibility are commonly applied for prediction in different weather environments. The consistency between the frequency of important predictors in specific models and the important predictors identified in the generic model demonstrates the reliability of building prediction models based on these important predictors.

4.2 The Comparison of Results between Special Models and Generic Model

As shown Fig. 30, the performance of the specific RFL model is slightly better under clear and snowy conditions. This is because half of the training data is collected under clear conditions, resulting in the special RFL model being able to better explore the relationship between features and predictive attenuation in the RFL channel. On the other hand, the most important predictor of Maximum Wind Speed under snowy conditions is only relevant for predicting under snow conditions and do not contribute to predictions in other weather conditions. Under other weather conditions, the performance of generic RFL model is better than special RFL model.

In the FSO channel, the generic model outperforms the specific model under dusty and snowy conditions. This is due to the small size of the training set in such weather conditions, making the specific model unreliable. However, the predicted results are good in the generic model under dust and fog conditions, possibly due to data from other environments contributing to predicting attenuation under these conditions. In other weather conditions, the performance of generic FSO model is better than special FSO models.

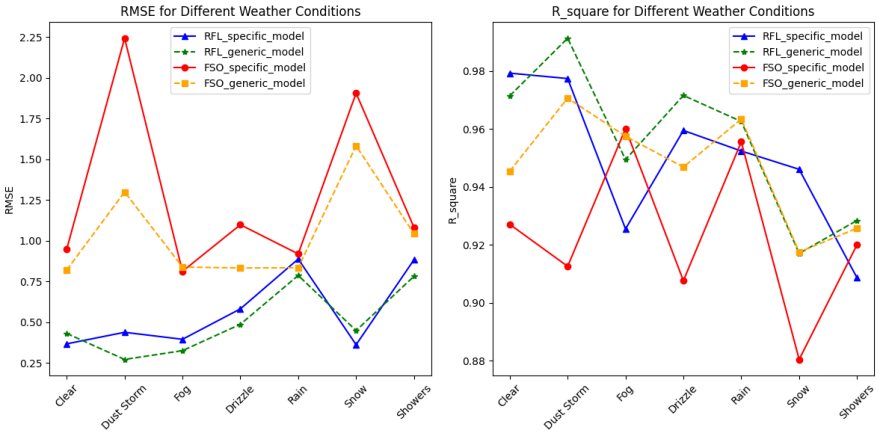

Figure 30: Comparison of Specific and Generic Model

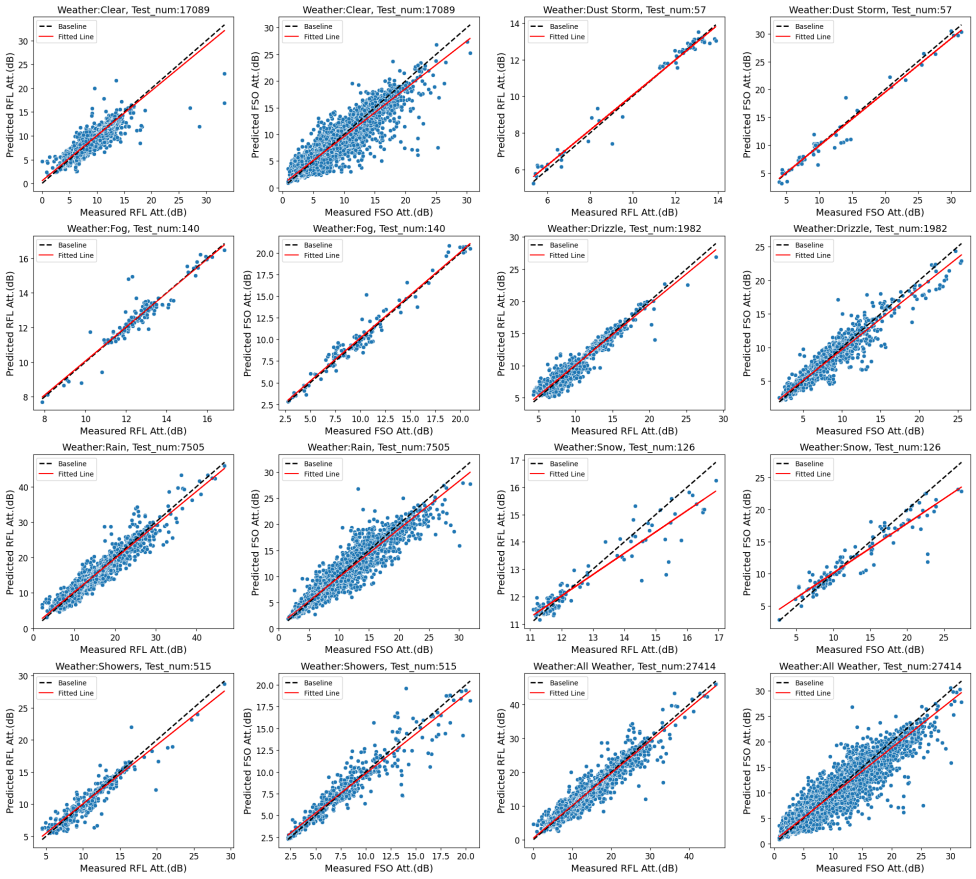

Figure 31: Comparison between the Measured and Predicted Attenuation

4.3 Predicted Results under Various Weather Conditions Using Generic Model

The comparison between the measured and predicted attenuation is shown as Fig. 31. The scatter plot shows points closely clustered around the baseline, indicating a high degree of linear correlation between predicted and actual values. The even distribution of points on both sides of the baseline suggests that the generic model maintains balanced predictive capabilities across various attenuation ranges, without significant bias.

5 Conclusion

This study comprehensively explored weather-induced channel attenuation in hybrid RF/FSO communication systems, employing ensemble algorithms to establish predictive models for attenuation prediction. Through extensive data preprocessing, EDA, and model establishment, this research developed both specific and generic models capable of accurately predicting channel attenuation under various weather conditions.

The analysis emphasized the significance of feature pruning and hyperparameter tuning in optimizing model complexity and performance. Through these processes, the generic random forest model has been streamlined, reducing its reliance on 25 features to 14 for the FSO channel and 11 for the RFL channel. Despite removing some redundant features, the performance in both channel models remains basically unchanged. This reduction in complexity has significantly alleviated computational overhead while simultaneously enhancing the interpretability and generalization ability of the models.

Furthermore, the investigation revealed the pivotal role of weather factors in influencing channel attenuation. By building separate models for specific weather conditions and a generic model encompassing diverse environments, the study demonstrated that in the RFL channel, features related to Absolute Humidity, Rain Intensity, Relative Humidity, Frequency, Distance, and Visibility are frequently employed for prediction across diverse weather conditions. Conversely, in the FSO channel, features associated with Distance, Temperature, and Visibility are commonly utilized for prediction across various weather environments.

Additionally, the generic model excels in predicting channel attenuation across diverse weather conditions, particularly providing more precise predictions for fog and dust smaller sample sizes compared to the specific model. However, in clear and snowy conditions, the specific models outperforms the generic model in the RFL channel.

Finally, the comparison of predicted values and actual values demonstrates the generic models in both channels can predict the channel attenuation effectively and accurately.

Overall, the research provides valuable insights into mitigating the impact of weather on hybrid RF/FSO communication systems, facilitating the development of more robust and efficient wireless connectivity solutions.

6 Appendix

The code can be accessed in github:
https://github.com/hahawang1986/Hybrid-Optical-Radio-Frequency-Communication-Channel-Model.git

References

[1] Antonios Lionis, Konstantinos Peppas, Hector E Nistazakis, Andreas Tsigopoulos, Keith Cohn, and Athanassios Zagouras. Using machine learning algorithms for accurate received optical power prediction of an fso link over a maritime environment. In *Photonics*, volume 8, page 212. MDPI, 2021.

[2] Cao Ying, Miao Qi-Guang, Liu Jia-Chen, and Gao Lin. Advance and prospects of adaboost algorithm. *Acta Automatica Sinica*, 39(6):745–758, 2013.

[3] Yi Feng, Linlan Liu, and Jian Shu. A link quality prediction method for wireless sensor networks based on xgboost. *IEEE Access*, 7:155229–155241, 2019.

[4] Syed Agha Hassnain Mohsan, Muhammad Asghar Khan, and Hussain Amjad. Hybrid fso/rf networks: A review of practical constraints, applications and challenges. *Optical Switching and Networking*, 47:100697, 2023.

[5] Mostafa Zaman Chowdhury, Moh Khalid Hasan, Md Shahjalal, Md Tanvir Hossan, and Yeong Min Jang. Optical wireless hybrid networks: Trends, opportunities, challenges, and research directions. *IEEE Communications Surveys & Tutorials*, 22(2):930–966, 2020.

[6] P Series. Propagation data and prediction methods required for the design of earth-space telecommunication systems. *Recommendation ITU-R*, pages 618–12, 2015.

[7] P Series. Propagation data required for the design of terrestrial free-space optical links. *Recommendation ITU-R*, 2012.

[8] Antonios Lionis, Konstantinos Peppas, Hector E Nistazakis, Andreas Tsigopoulos, Keith Cohn, and Athanassios Zagouras. Using machine learning algorithms for accurate received optical power prediction of an fso link over a maritime environment. In *Photonics*, volume 8, page 212. MDPI, 2021.

[9] Kappala Vinod Kiran, Subhanesh Perinbaraj, Jayashree Pradhan, Pradeep Kumar Mallick, Ashok Kumar Turuk, and Santos Kumar Das. Machine learning aided switching scheme for hybrid fso/rf transmission. *Intelligent Decision Technologies*, 14(4):529–536, 2020.

[10] Wikipedia contributors Harry585. Random forest bagging illustration, 2023. [Online; accessed 1-Apr-2024].

[11] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

[12] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[13] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7:e623, 2021.

[14] Leo Breiman and Adele Cutler. Manual–setting up, using, and understanding random forests v4. 0. 2003. *URL https://www. stat. berkeley. edu/˜ breiman/Using_random_forests_v4. 0. pdf*, 2011.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[16] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[17] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[18] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.